Modeling Mosquito Abundance Using Forest Fire Data: A Practical Evaluation of Machine Learning Pipeline Techniques

Raymond Lin[1], Haley Oba[2], Krish Desai[3], Ashmit Dewan[4], Zarar Haider[5]

[1]West Windsor Plainsboro High School South, [2]Palo Alto High School, [3]Harold M. Brathwaite Secondary School, [4]West Windsor Plainsboro High School North, [5]Trinity School

Abstract

Mosquitoes have been a major health concern for decades, and with climate change expanding their range, their threat to public health is increasing. In response, mosquito abundance prediction machine learning models have been researched in numerous locations. Our research builds on this and seeks to explore novel methods such as using natural disaster data, optimizing hyperparameters through Bayesian Search, and inspecting models using Partial Dependence (PDP) and Individual Condition Expectation (ICE) plots. Based on previous work, we selected four base ecological variables. We then acquired variations of these base variables and assessed their effectiveness by training Random Forest Regressors (RFR) using the variables' variations instead of the base variable. Out of all the variations, only minimum daily temperature proved better than its base variable (mean daily temperature). Our final model used the best variable variations and our custom forest fire index. We optimized all our models using Bayesian Search, which we found to be more effective than Grid Search. Our final RFR model had a root mean squared error (RMSE) of 3.94 when predicting the test set. To see whether forest fire index had any impact on accuracy, we used Drop-column variable importance, the purest way of calculating variable importance. We found that forest fire marginally increased accuracy, which is the best case scenario for rare-occurrence data, where most of the values are 0. Using PDP and ICE plots, we found that our model was able to synthesize accurate relationships between variables like temperature and mosquito abundance that reflect field and lab findings. Further research should be done on machine learning model inspection and its use cases. Within mosquito research, further work can explore other novel datasets, like forest fires, to form a more comprehensive understanding of mosquito abundance.

Keywords: machine learning, mosquito abundance, forest fire, model inspection, feature optimization

Research Questions

How significant is forest fire data for predicting mosquito abundance?

What features or feature variations are most useful for predicting mosquito abundance, and in what way does each feature contribute to the model?

What machine learning techniques can be used to create an accurate model of mosquito abundance?

Introduction and Review of Literature

Mosquito and Forest Fire Background

Mosquitos carry many deadly diseases, including malaria, dengue fever, yellow fever, and West Nile virus, which kill hundreds of thousands of people each year (World Health Organization [WHO], 2020). Forest fires are attributed to burning vegetation and disrupting water and air quality (WHO, n.d.), revealing a significant consequence for nearby mosquito populations. For example, increased water temperatures and lack of canopy shade after a natural marsh fire in East Texas caused a two-month absence of mosquito larvae (Janousek & Olson, 1994). Drought exacerbated by climate change combined with fire suppression practices resulted in record numbers of severe forest fires; counterintuitively, fire suppression encourages the expansion of coniferous tree forests, which are densely packed and highly flammable, into historically non-forested areas, thus increasing the frequency, range, and severity of forest fires (NASA Earth Observatory, 2022; United States Department of Agriculture [USDA] Forest Service, 2016; Alberta Government 2012). Globally, forest fires are increasing at an alarming rate, with the United Nations (2022) reporting that the total amount of catastrophic forest fires could increase by 50% by 2100. By exploring the relationship between forest fires and mosquitoes, our study takes a step into understanding how mosquitos are being affected by a transforming environment.

Machine Learning Background

Machine learning is widely used to predict mosquito abundance because it can perform highly accurate predictions on large amounts of data in a timely fashion. In particular, the random forest method can predict variable significance and can model complex relationships between variables (Cutler et al., 2007). Kwon et al. (2015) calculated the relative importance of meteorological variables on mosquito abundance in South Korea by using random forest, which had the highest prediction accuracy compared to the support vector machine and classification and regression tree models.

Rainfall, specific humidity, normalized difference vegetation index (NDVI), and temperature are critical factors in determining mosquito abundance: above-average rainfall and 30-80% relative humidity are strongly associated with a high abundance of mosquitoes, NDVI is used to show suitable habitats for mosquitoes by estimating the spatial and temporal dynamics of vegetation types, and partial dependence plotting has been used to find that temperature has a clear impact on *Culex pipiens*, typically requiring a minimum temperature of 11ºC for presence in North America (Madzokere et al., 2020; Kofidou et al., 2021; Arora et al., 2022). Thus, we decided on these four ecological factors for the two base abundance models, and we added the novel factor, burn area, to one of our models. We then compared the random forest regression models to determine the importance of fire data on mosquito abundance predictions.

Research Methods

Area of Interest (AOI)

The large amount of wildfires that have occurred in Shasta County over the past 12 years, alongside Shasta MVCD's consistent mosquito population data is what led to Shasta County being our AOI for this study. The Shasta-Cascade region had one of the highest annual *Culex pipiens* counts in California (Barker et al., 2010).

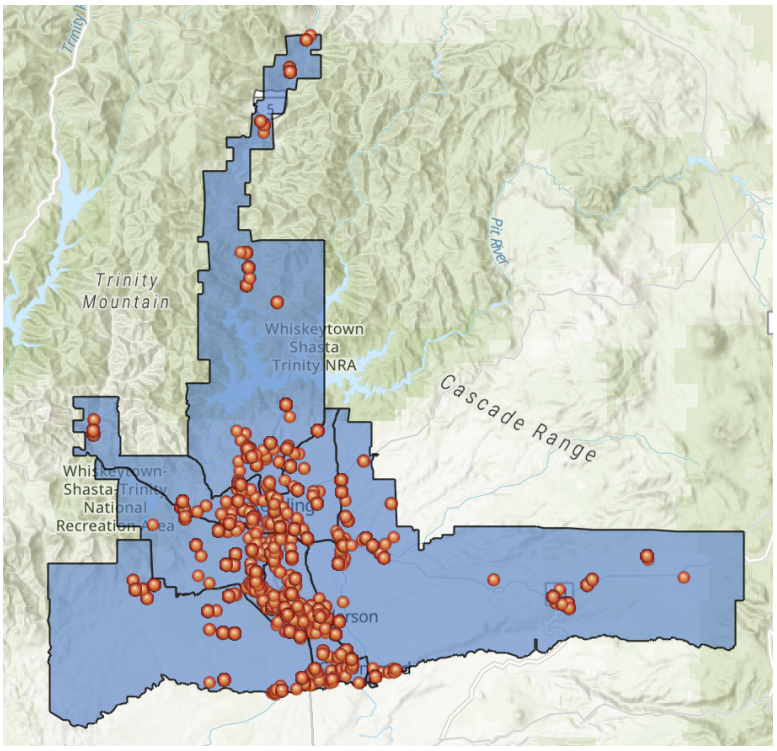
In 2021 alone, more than 2.23 million acres of California were burned, making it the most wildfire-prone state (National Interagency Coordination Center [NICC], 2021). Therefore, we determined California as a potential AOI. Furthermore, the Shasta Mosquito and Vector Control District (Shasta MVCD) has been monitoring mosquito populations and vector-borne diseases in the county multiple times per week for the past few decades in gravid and CO2-baited Encephalitis Vector Survey (EVS) traps, providing us with an abundance of mosquito data from 2010-2022 (Shasta MVCD, n.d.). Shasta County has the typical California Mediterranean climate: warm, dry summers and cold, wet winters, with an average annual precipitation of around 165 cm (United States Climate Data, 2022; Kauffman et al., 2003).

The county is surrounded by forests, urbanized around Redding, and Shasta Mountain is the headwater of the Sacramento River, the largest river in California (Zanaga et al., 2021). Based on citizen scientist data reported along the Pacific Crest Trail just south of Mount Shasta, around 43% of land cover consists of evergreen or deciduous trees (Global Learning and Observations to Benefit the Environment [GLOBE], 2022).

Severe heat and drought due to climate change and overgrown forests from decades of fire suppression caused a recent surge in California's destructive wildfires (NASA Earth Observatory, 2021). Between 2010 and 2022, Shasta County endured 175 weeks of wildfires (Earth Engine Data Catalog, 2022). During these 175 weeks, the 2012 Bagley fire, the 2018 Carr Fire, and the 2021 Fawn Fire are the most significant in terms of acres burnt with each burning 46,011, 229,000, and 8,578 acres respectively (California Department of Forestry and Fire Protection, 2022).

Figure 1

Shasta MVCD Surveillance Area with all mosquito observations from 2010-2022 plotted as red dots.



Data Collection

Time Frame

When selecting our time series, the mosquito data from Shasta County served as a constraint. Though the data was plentiful, it was not consistently recorded on daily timesteps and data recorded before 2010 was incomplete. Therefore, we decided to use a weekly timescale based on the CDC's epidemiological (EPI) weeks, which has been used before by Schneider et al. (2021) and serves as the standard for governmental epidemiological reporting, and limit our time series from 2010 onwards. Consequently, our timeframe ranges from 2010-01-03, the first EPI week of 2010, to 2022-07-10, the most recent EPI week reflected in the Shasta mosquito dataset.

Acquiring Ecological Data

We used Google Earth Engine and various satellites to acquire ecological datasets. We extracted minimum and maximum relative humidity, as well as mean specific humidity data from the University of Idaho Gridded Surface Meteorological Dataset (GRIDMET); total precipitation and minimum, maximum, and mean temperature from the Daily Spatial Climate Dataset (PRISM); NDVI values from MODIS Terra Daily NDVI Dataset provided by Google; and fire burn data from MOD14A1.006 provided by the NASA LP DAAC at the USGS EROS Center.

NDVI specifically stands for Normalized Difference Vegetation Index, wherein its values range from +1.0 to -1.0. It is derived from visible light range (VIS) and near-infrared light range (NIR) data collected by satellites, in our case MODIS Terra. This works because healthy, green vegetation generally absorbs most visual light, reflecting more infrared light, whereas unhealthy vegetation or barren ground reflects more visual light and less infrared light. Its formula is a normalized difference between NIR and VIS:

$$NDVI = \frac{NIR-VIS}{NIR+VIS}$$

A value of 0.1 or less indicates no vegetation, typically indicative of a barren area of sand or snow. Conversely, high values, from roughly 0.6 to 0.9, indicate green leaves, which normally correspond with dense vegetation. Values in between these two ranges indicate an area with sparse vegetation, suggestive of shrubs or grasslands. NDVI data analysis is also particularly useful in identifying changes in vegetation due to wildfires, accounting for most long-term effects of fire while fire data itself accounts for short-term impacts.

Using the MOD14A1.006 dataset provided by the NASA LP DAAC at the USGS EROS Center, we created our own fire metric in order to accurately compare between "no fire" and different levels of fire. This dataset has various bands, such as the maximum fire radiative power and the position of the fire pixel within the image. However, we choose to use the FireMask band which describes the confidence of fire. The FireMask band returns a binary number that corresponds to the confidence level of the fire, whether there is a cloud or unknown land in the area, or represents that there's no fire at all. Because FireMask can show attributes like clouds or unknown land, we created a filter that only looks at the fire confidence (if it's non-null) and returns a null value for all other possible attributes. Fire confidence (low, medium, high) is a bitmask that returns a 7, 8, or 9, while any other attributes are represented as a null (or a 0). This poses a problem because it creates a large disparity between a low confidence fire (7) and no fire at all. A low confidence fire is not far from no fire at all, and because our machine learning model numerically takes the difference into account, this caused a problem. To solve this, we used an equation that reduced the fire values that were returned (7, 8, 9) to 1, 2, or 3, respectively. From the dataset, we exported the average confidence level of fires each day, and the total fire pixels in the county each day. From this, we derived a value that represented the sum of the confidence levels of each fire pixel in the county each day.

One of the key common values that are used to predict mosquito abundance is humidity. There were two types of humidity dataset values available to us, specific and relative. The main difference is that specific humidity is absolute while relative humidity changes based on air conditions.

For each dataset, we first filtered the data to only include pixels in our AOI, Shasta County, and our Time Frame of Interest (TFOI), 2010-2022. This was done by only fetching the data in our TFOI, and then only using pixels that intersected the Shasta County Borders geometry. Then, in order to turn the data, which is daily, per-pixel values, into weekly, entire-area averages, we preprocessed using Google Earth Engine and Google Sheets. To get a daily value for the whole of Shasta County, we took the value of each pixel in our AOI and averaged their values. This was done using Google Earth Engine's "reduceRegion" function. Then, to get weekly averages, we exported from Google Earth Engine into Google Sheets, where we used an averaging formula to create weekly averages.

Acquiring Mosquito Abundance Data

Accurate and ample *Culex pipiens* and *Culex tarsalis* abundance data for this study was provided by Shasta MVCD. In nearly all years of our TFOI, data was recorded from March to October. Their gravid and 60 CO2-baited EVS traps are collected multiple times per week to gain an accurate understanding of mosquito populations. These traps were primarily located in urbanized areas around Shasta's cities, Redding and Anderson, and a few were located in the mountains.

Acquiring GLOBE Land Cover Data

Land cover data was acquired through the GLOBE Advanced Data Access Tool using the following filters: 1/24/2010-7/10/22 for the date range, California for the state, Oregon State University GEOGRAPHY for the team, and Land Cover as the protocol. We then exported the CSV file and plotted the points as a hosted feature layer in an ArcGIS map and looked at the Land Cover pictures and descriptions there.

Data Analysis

Feature names:

Precip: Precipitation (mm)
Temp: Mean Temperature (℃)
NDVI: Normalized Difference Vegetation Index
Hum: Specific Humidity (%)
Fire: Fire index
Mosquito: Mosquito Abundance (#)

Feature variation names:

Temp_Min: Temperature calculated from daily minimums (℃)
Temp_Max: Temperature calculated from daily maximums (℃)
Hum_Min: Relative humidity calculated from daily minimums (%)
Hum_Max: Relative humidity calculated from daily maximums (%)

Data Statistics

After extracting and averaging the data into a workable format, we looked at the statistics of each feature shown in Table 1. Precipitation (mm), fire index, and mosquito abundance all display high variability, with standard deviations significantly larger than their means. On top of that, all three have many high outliers and are significantly right-skewed, as shown by how their maxes are far larger than their 75% percentiles. This makes sense since fires and precipitation are rare occurrence events, although precipitation is still far more frequent, and mosquitoes come in waves because of their short lifetime and dependency on good conditions to oviposit (CARBAJO AE et al., 2006).

Table 1. Mean, standard deviation, and 0th, 25th, 50th, 75th, and 100th percentiles for each feature

	mean	std	min	25%	50%	75%	max
Precip (mm)	3.09	5.23	0.00	0.00192	0.336	4.05	32.2
Temp (℃)	13.3	7.46	-0.445	7.22	12.2	20.0	28.4
NDVI	0.417	0.169	-0.00197	0.283	0.451	0.566	0.658
Hum	0.00508	0.00152	0.00171	0.0039	0.00485	0.00619	0.00907
Temp_Min (℃)	6.71	6.01	-5.99	1.77	5.53	11.9	19.9
Temp_Max (℃)	20.0	9.05	3.19	12.3	19.0	28.2	37.2
Relative Hum_Min	34.0	15.6	6.28	22.0	31.4	43.8	83.7
Relative Hum_Max	71.9	15.5	32.3	61.0	73.6	83.9	99.0
Fire	4.46	24.5	-0.857	0.00	0.00	0.286	286
Mosquito (#)	5.82	7.61	0.00	0.00	2.85	7.96	53.5

We also visualized all our data in line plots to find general patterns and validate our data with climate reports from Shasta. Figure 2 shows every feature plotted over all years. It is evident from Figure 2 that precipitation mostly happens in Shasta's winter months. This is crucial because most mosquito abundance models rely heavily on short-term precipitation since it creates oviposition habitats, but in our case precipitation does not line up with the summer months when mosquitos are active, making it far less useful than usual (Cleckner HL et al., 2011). Fire also seems peculiar with one sharp peak at around week 30 of the year. This is the largest and most severe fire in our time frame, the 2018 Carr Fire. Apart from precipitation and fire, the other features seem to be as expected, with temperature, humidity, NDVI, and mosquito abundance all peaking moderately simultaneously, which reflects the positive correlation that other studies have found (Cleckner HL et al., 2011).

Figure 2 (a)

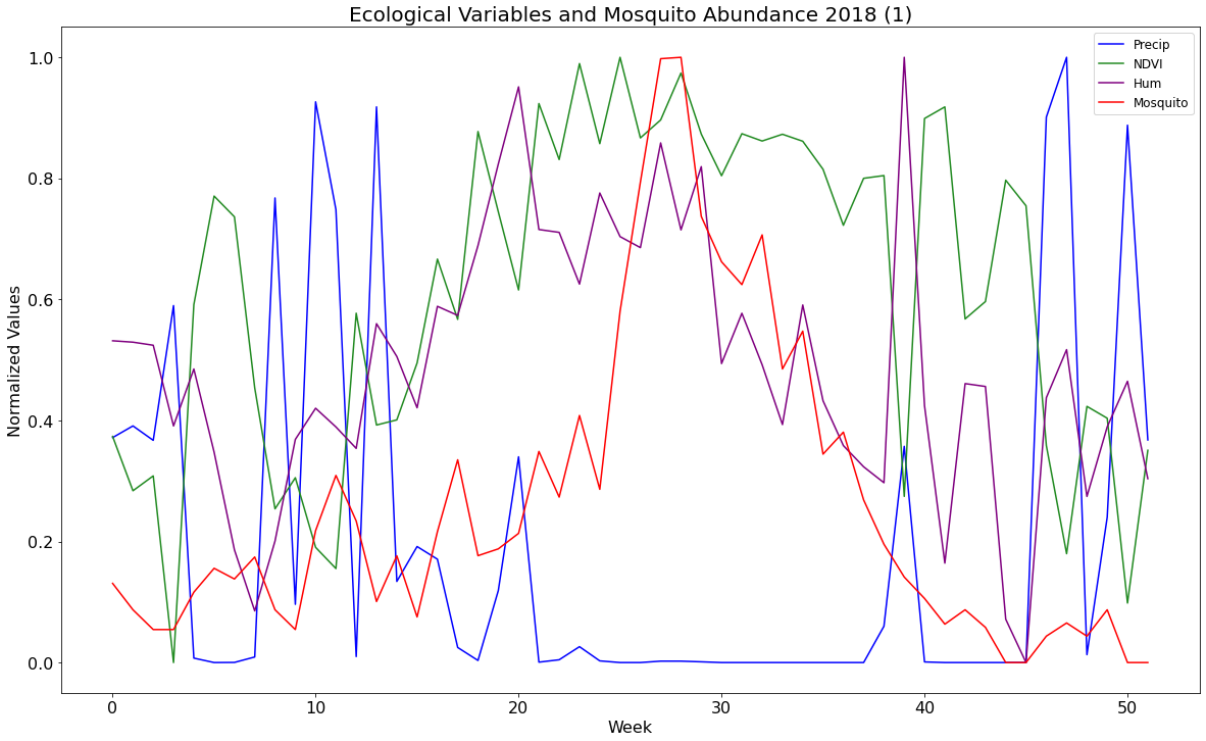
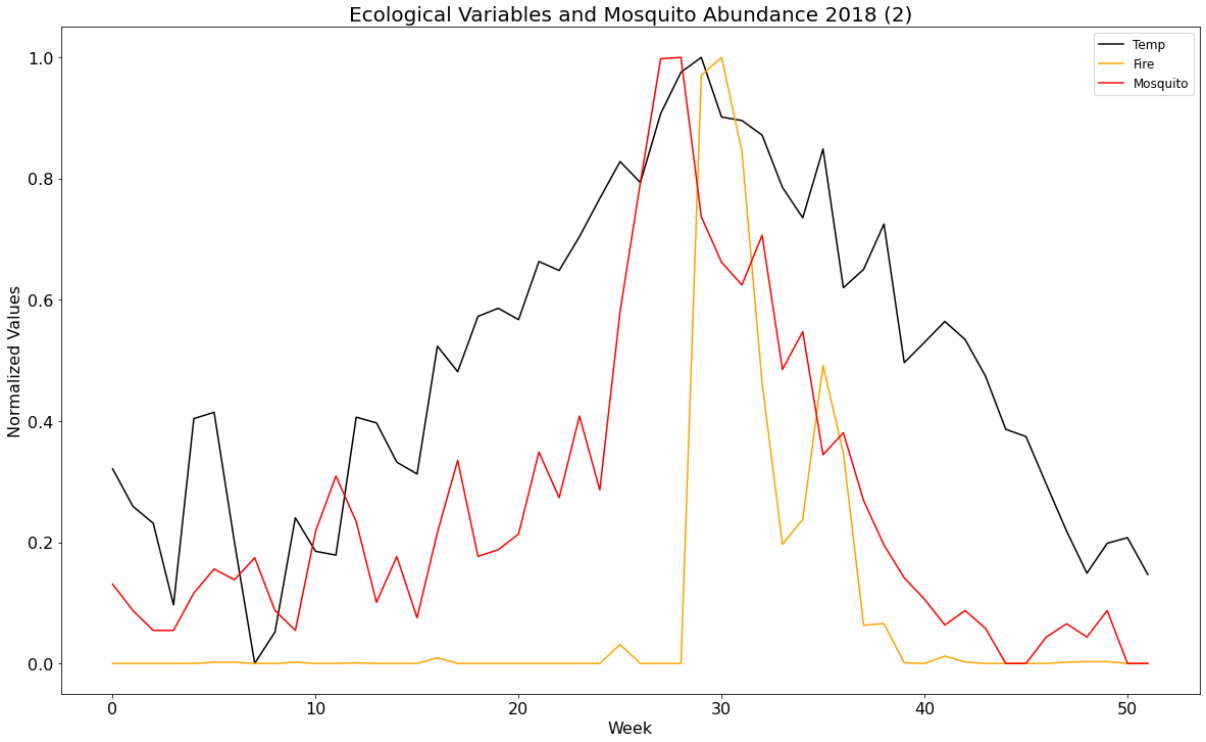


Figure 2 (b)



The features in Figure 2 are from the daily mean datasets of each feature, but many previous studies have found variations like the daily minimum or maximum to be more helpful (Lee KY et al, 2017). To evaluate feature variations in Figure 3 (a), we plotted all three temperature variations along with mosquito abundance and did the same with the humidity variations in Figure 3 (b). It should be

noted that Hum_Min and Hum_Max values are relative humidity, whereas Hum is specific humidity. There is not much to be gleaned by the human eye from Figure 3 (a), all three variations seem very similar, rising and falling together as expected. However, in Figure 3 (b), while the local peaks happen at the same time between all three variations, the Hum_Min and Hum_Max tend to be higher in the winter and lower in the summer whereas the Hum is higher in the summer and lower in the winter. This is because relative humidity measures how saturated the air is compared to its maximum saturation, and maximum saturation rises with temperature, which causes much larger denominators during hotter weeks. On the other hand, specific humidity is the direct proportion of water vapor mass to total air mass, which does not take into consideration variations in air condition.

Figure 3 (a)

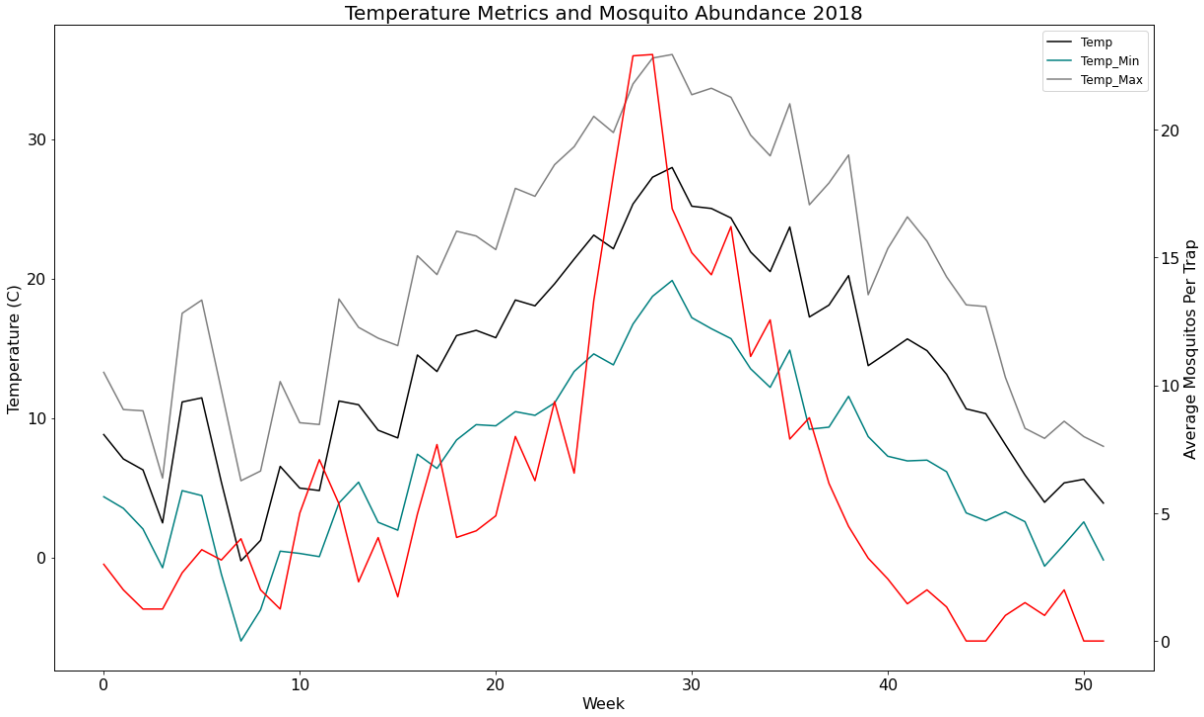
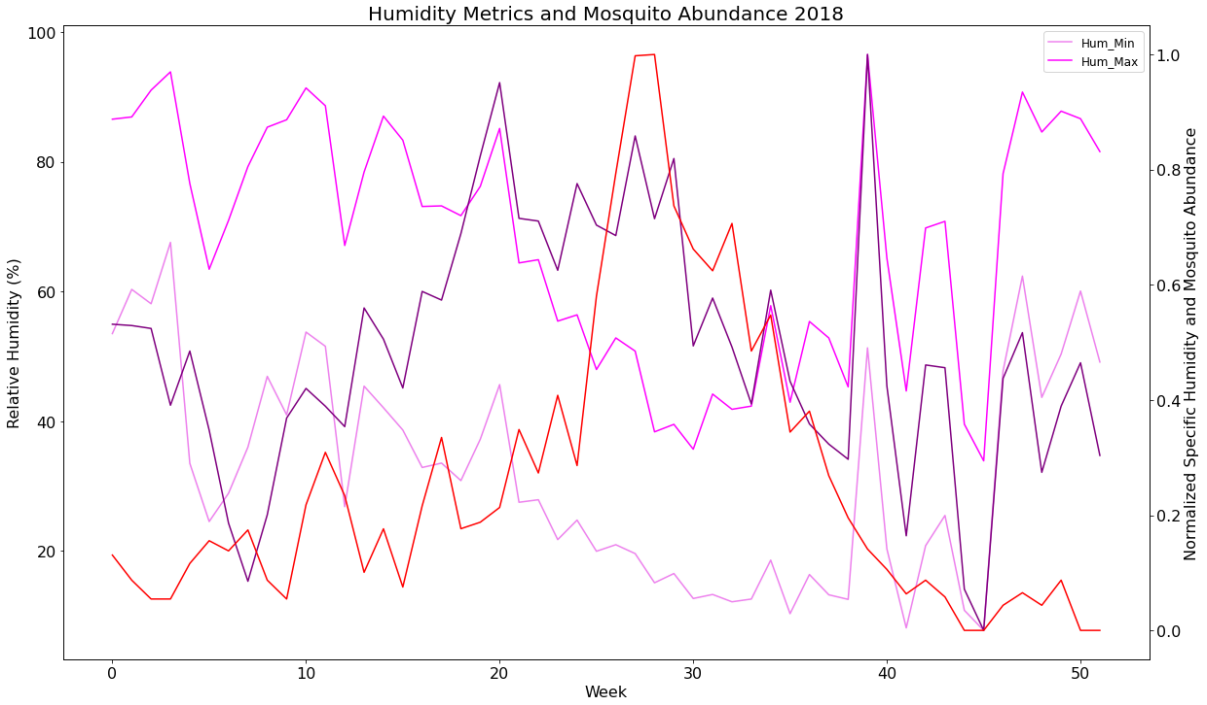
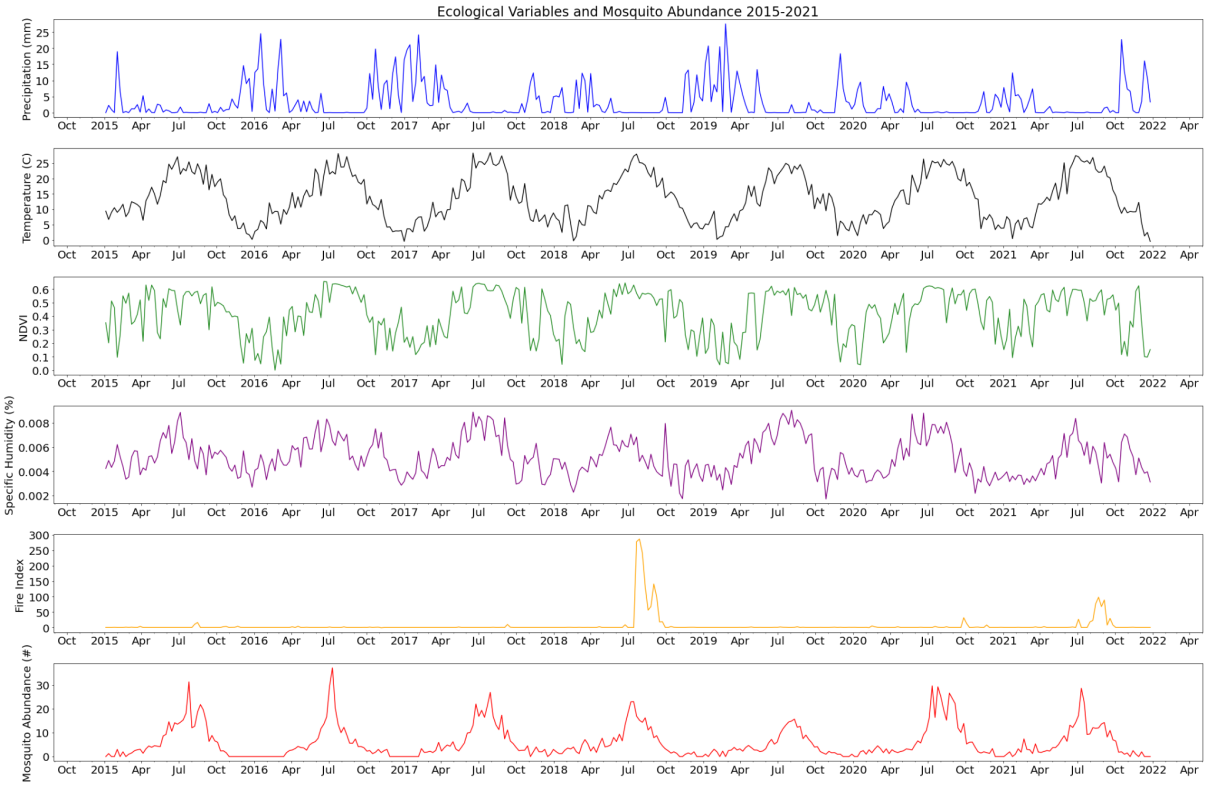


Figure 3 (b)



The previous plots have only displayed data from the year 2018, which was chosen because it had a significant fire in it. Figure 4 has each feature across 2015-2021. It shows some yearly variation, most significantly in precipitation, fire, and mosquito abundance, the same three features we found were high variability and outlier prone when looking at their statistical metrics. From this graph, it is clear that mosquito abundance is most heavily affected by temperature, displaying longer periods of high mosquito abundance when there were similarly long periods of high temperatures, as seen in 2015, 2017, and 2020. Slightly lower temperatures can also have drastic effects on mosquito abundance, as shown in 2019. This follows the prior consensus that mosquitos have a sensitive optimal temperature range where they flourish (Mordecai EA et al, 2012).

Figure 4



Finding Optimal Lag

There is significant precedence of feature lag among mosquito abundance models since the impact of an event will likely not affect adult mosquito populations till a few weeks later: Schneider et al. (2021) used 3 weeks of lag on all features. Chang et al. (2016) discussed how feature lag times will differ under different climates, so instead of using previous lag values, we took a quantative approach to finding the lag values best suited to our application. There are many ways of finding optimal lag: Lopez et al. (2014) suggested visual evaluation using graphs; Ruiz et al. (2010) and Lee et al. (2016) exhaustively tested lag for the highest correlation coefficient (Pearson's R). We decided to exhaustively test lag because it seems more robust and objective. In Figure 5, we exhaustively tested lag by plotting each feature's correlation with mosquito abundance on the y-axis, and the number of weeks lagged on the x-axis up till 26 weeks. Further than 26 weeks did not make sense because then the data began having inaccurate correlations because it is closer to last year's summer instead of correlating with this year's mosquito data. Lagging a negative number of weeks is also illogical since these variables are supposed to influence mosquito abundance.

Table 2 shows the results of Figure 5, with the best correlation coefficient achieved for each feature and how many weeks created that correlation. We decided to use the absolute correlation coefficient since negative or positive correlation does not matter to machine learning algorithms, as long as it is strong. Every feature's highest absolute correlation was with 0 weeks of lag except specific humidity, which performed best with 3 weeks of lag. Therefore, the first three weeks had to be cut from the timeframe, making it from 1/24/2010 - 7/10/2022.

These results contradict previous findings, where most features were lagged between 1 and 8 weeks (Guo et al., 2014). We hypothesize that this is because precipitation is the main feature that

logically benefits from lag, since the effect of precipitation is it creates oviposition habitats, and from egg to adult mosquito takes around 2-3 weeks depending on the species, which is the most common lag time for precipitation data (Hwang et al., 2020). As discussed earlier, our AOI has abnormal precipitation patterns, making short-term precipitation effects almost non-existent because there is nearly no precipitation during the summer months when mosquitos are active.

Figure 5 (a): Weeks of Lag vs Pearson's R for Temperature, Fire, and NDVI

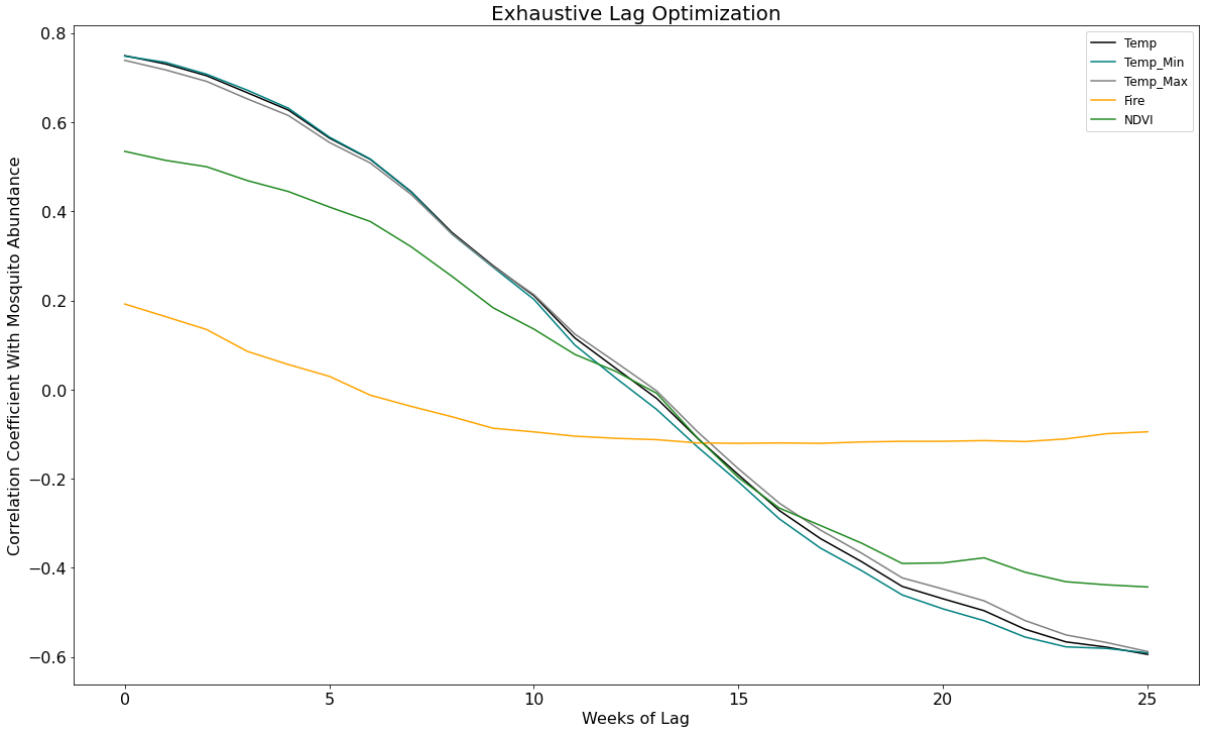


Figure 5 (b): Weeks of Lag vs Pearson's R for Precipitation, and Specific and Relative Humidity

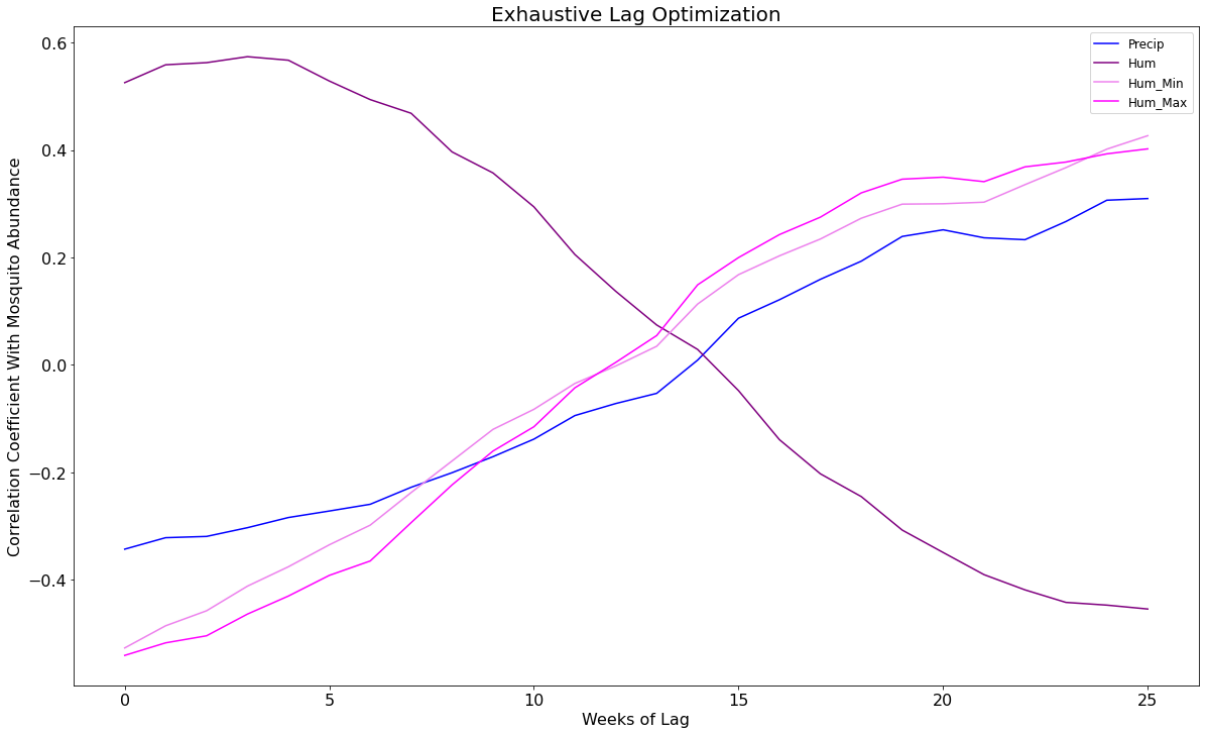


Table 2: Optimal Lag Weeks

	Precip	Temp	NDVI	Hum	Temp Min	Temp Max	Hum Min	Hum Max	Fire
Max R	0.343	0.749	0.535	0.574	0.748	0.739	0.527	0.541	0.192
Lag	0	0	0	3	0	0	0	0	0

Train Test Split

We created a train and a test dataset, following the common machine learning model validation method: train-test split. However, instead of splitting completely randomly, we used an 80/20 train test stratified random split, "StratifiedShuffleSplit", from the machine learning Python package Scikit-Learn (SKL). We chose to stratify by year to eliminate any bias in the model that may be caused by annually changing conditions like climate change, or how 2022 only has half a year of data. We checked stratification for both train and test sets, and each year's representation in the set was within 5% of its representation in the full dataset.

Hyperparameter Optimization

To train the models, we first compared three different hyper-parameter (HP) optimization techniques. The first is for control purposes: it uses the default values of SKL's Random Forest Regressor (RFR); the second is Grid Search (SKL's Grid Search Cross Validation), which brute force tries every combination of the HPs given to it, making it very inefficient but effective if given enough time; the third is Bayesian Search, which is similar to Random Search, but instead of using completely randomized HPs, it guesses what will be the best HPs based on past trials and uses that for its next trial. It does this by creating a model of how well the actual model will do based on what HPs are set. Bayesian Search is significantly faster and can achieve better results than Grid Search because it makes informed decisions and improves upon itself (Wu et al., 2019; Snoek et al., 2012).

To control for the difference in effectiveness of the three optimization methods, we fitted them on the same dataset: base dataset (Temp, Hum, Precip, NDVI), without any special variations (min or max) on any of the features. The results are shown in Table 3 below using the metrics: root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2). Results are as expected, with Bayesian Search being the best, then Grid Search, and then no optimization.

Table 3: Hyperparameter Optimization Method Comparison

	Default HPs	Grid Search	Bayesian Search
RMSE	4.1727	4.0721	4.0315
MAE	2.7601	2.6501	2.6199
R^2	0.62595	0.64377	0.65084

Table 4 shows the HPs tried and the optimal HP values found by the Grid Search method. We chose "n_estimators," "max_features," and "max_samples" because they are the most basic and

commonly optimized HPs of RFR (Snoek et al,. 2012). We then chose “min_samples_leaf” and “max_depth” because they make each tree less specific and increase overall variance, which helps prevent overfitting.

Table 4: Grid Search Tried and Optimal Values

	Tried Values	Optimal Values
n_estimators	[100, 200, 300, 400]	400
max_features	[“sqrt”, 1, 2, 3]	“sqrt”
max_samples	[0.5, 0.6, 0.7, 0.8 0.9, 1]	0.6
min_samples_leaf	[1, 2, 3, 4, 5, 8]	2
max_depth	[20, 30, 40, 50, 60, 70]	20

Table 5 shows the HPs tried and optimal HPs found by the Bayesian Search method. Square brackets denote discrete choices and parenthesis denote ranges. Again, the same five HPs were chosen to be optimized along with a sixth “max_leaf_nodes,” which operates similarly to “max_depth” and helps prevent overfitting.

Table 5: Bayesian Search Tried and Optimal Values

	Tried Values	Optimal Values
n_estimators	(100 - 600)	494
max_features	[“sqrt”, “log2”]	“sqrt”
max_samples	(0.1 - 1)	0.980
min_samples_leaf	(1-20)	5
max_depth	(10-60)	30
max_leaf_nodes	(15-100)	96

Feature Optimization

We trained RFR models using different combinations of features to select the best ones instead of using RF variable importances or trying each feature in linear models like past research has done (Belgiu & Drăguţ, 2016; Schneider et al., 2021). We did so because directly using the feature gives the best estimate as to how it will do in the final model. Additionally, training an RFR is extremely quick since we only have 651 samples.

We quickly found that having multiple varieties of the same feature caused poor performance as the repeated features only added extra training time and useless dimensionality. . Therefore, we tried each variation of each feature one at a time using the basic variation of all the other features and chose the best variations accordingly.

Table 6 shows the results of these trials with different variations of features. The Base model was fitted with the base dataset (Temp, Hum, Precip, NDVI), and for all the other models, whatever the model is named is the tested variation. So the Temp_Min model uses (Temp_Min, Hum, Precip, NDVI). The final model uses (Temp_Min, Hum, Precip, NDVI, Fire) because Temp_Min performed the best out of the three temperature variations, barely edging out the Base model. Temp_Max was significantly worse than the other temperature variations. This reflects previously mentioned research that shows that mosquitos are more sensitive to low temperatures (Arora et al., 2022). All the humidity variations proved useless, and were actually detrimental to the model. When variable importances were calculated for Hum_Min, and Hum_Max, the humidity variations received negative importance, meaning the models performed better without any humidity data than with Hum_Min and Hum_Max.

Table 6. Test Set Results For Each Model

Model Name / Tested Variation	Base	Temp_ Min	Temp_ Max	Hum_ Min	Hum_ Max	Final
RMSE	4.0315	4.0223	4.3313	4.4453	4.2525	3.9476
MAE	2.6199	2.6487	2.7397	2.7993	2.7196	2.5746
R^2	0.65084	0.65244	0.59697	0.57548	0.6115	0.66522

Table 7. Optimal HP values for each model, using the same trial HP values as Table 5.

Model Name / Tested Variation	Base	Temp_ Min	Temp_ Max	Hum_ Min	Hum_ Max	Final
n_estimators	494	415	324	500	219	100
max_features	“sqrt”	“sqrt”	“sqrt”	“log2”	“log2”	“sqrt”
max_samples	0.980	0.989	0.966	1.00	0.810	1.00
min_samples_leaf	5	3	3	1	3	1
max_depth	30	20	42	10	48	60
max_leaf_nodes	96	96	100	100	84	100

Variable Importances

While we did not use variable importances to choose features for the final model, they are still valuable for showing how much each feature actually contributes to estimating mosquito abundance.

Variable importance is more complicated than meets the eye, and there are many different ways of calculating it. SKL’s RFR comes with a default variable importance based on the mean decrease in Gini impurity. This method is cheap to compute given a fitted RFR, but it is biased, favoring features with high cardinality (continuous or very many categories) (Strobl et al., 2007; Boulesteix et al., 2012).

Permutation importance is another choice that randomly scrambles one feature in the dataset at a time, then uses a fitted model on it. This allows it to test one feature at a time, checking whether the model relied on it to create a good prediction. The output is also very handy because it outputs the raw increase in RMSE, or whichever error metric is used when the feature is shuffled vs normal (Strobl et al., 2008; Altmann et al., 2012). This method is a good middle ground between variable importance accuracy and calculation speed since it avoids retraining the whole model for each feature.

Drop-column importance is the most accurate measurement of a feature’s significance to a model. As its name suggests, it drops a feature, then fits a model without using the feature, and outputs the raw increase in error when the model is fitted without the feature, just like Permutation importance. This is the purest way of testing if a feature is truly important, and we chose this since we can afford to train many models due to our small sample count.

Results

Model Performance

Table 6 in Methodology displays the best results for our final model. Overall performance is very good, exceeding Lee's Artificial Neural Network mosquito abundance model RMSE of 14.38 by a significant margin (Lee et al., 2016). Figure 6 shows our final model's predictions on the test set. Since the test set was stratified across years, the plot has about 20% of the weeks from each year, so the weeks are not displayed continuously. Simply observing Figure 6 reveals the predictions are highly accurate, as prediction and actual mosquito abundance peaks line up almost every year. The model sometimes struggles to predict the peaks' exact magnitude; however, this is understandable given the extreme right-skewed and outlier-prone mosquito abundance data.

Figure 6

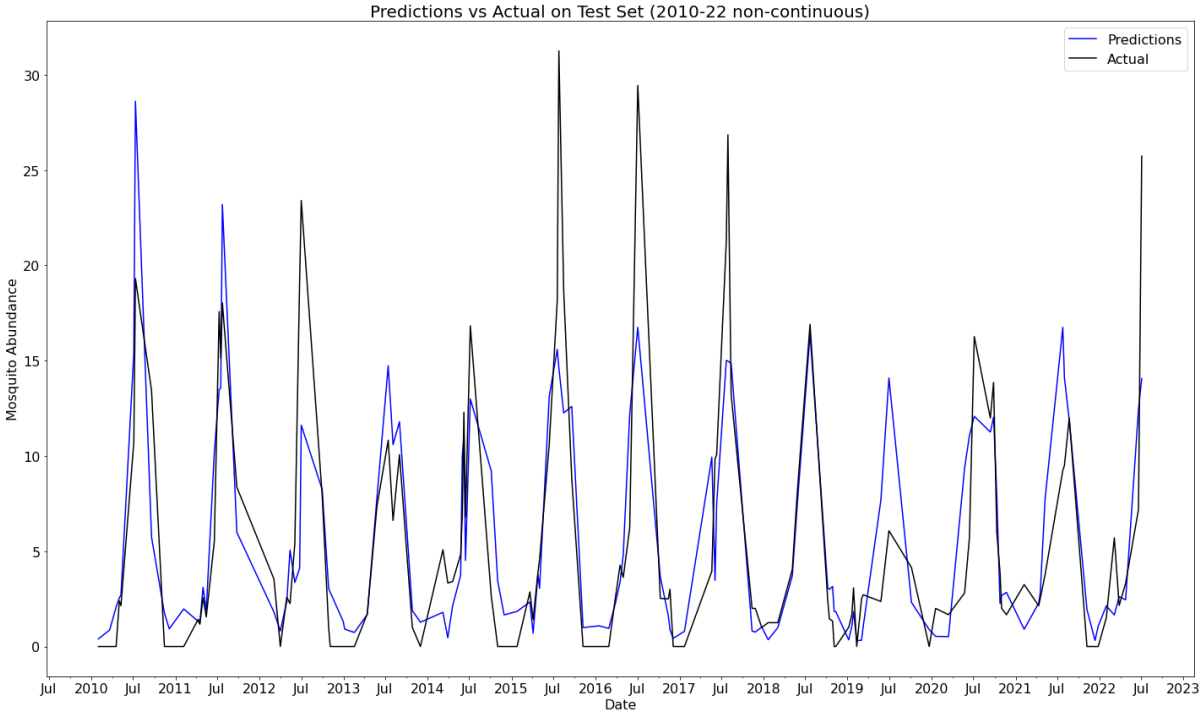
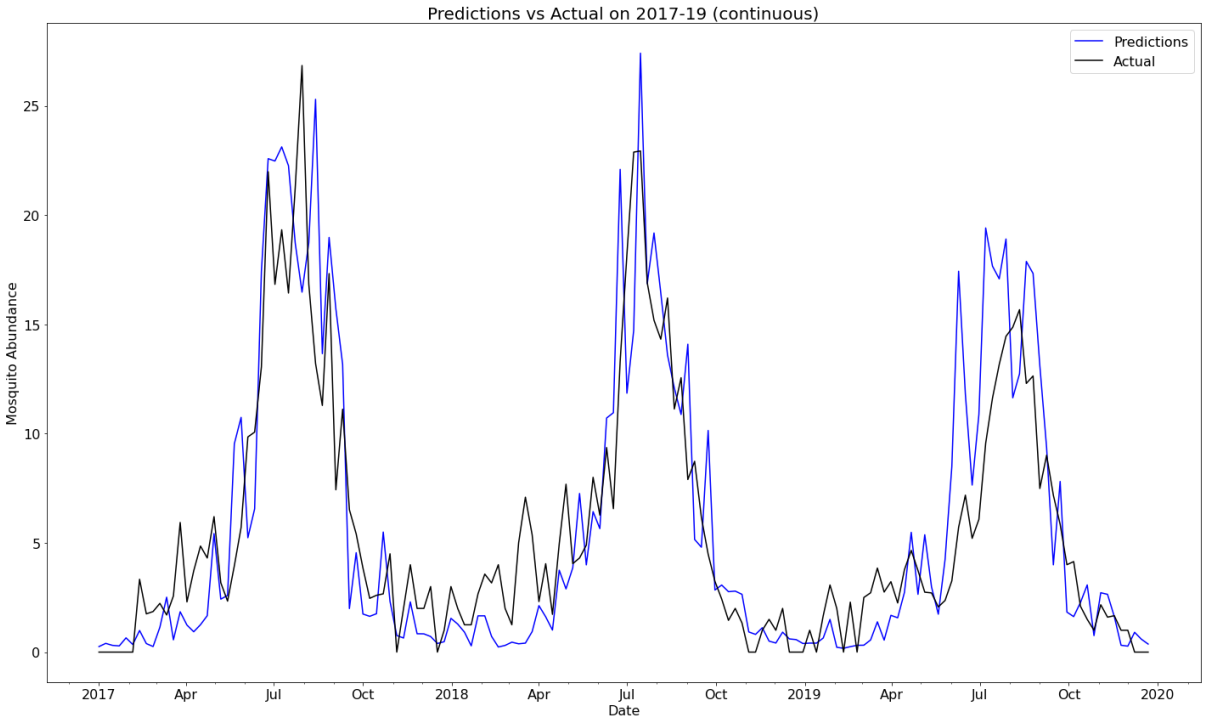


Figure 7 shows predictions for a continuous 3 years from 2017-19. In order to make it fair, we refitted our final model using all the data from 2010-16 and 2020-22, saving 2017-19 as a test set. We then used this new model to predict mosquito abundances for 2017-19. Figure 7 shows similar results as Figure 6, with the model correctly predicting peak mosquito season for all three years. Although the model overshot the height of the 2019 peak, its prediction was still significantly shorter than its prediction for 2017 and 18, showing it was able to recognize an abnormally low year.

Figure 7



Feature Evaluation

Variable Importance Observations

Apart from training an accurate model, another aspect of our research is evaluating different features, feature variations, and HP optimization techniques. We have already compared feature variations and HP optimization techniques in methodology extensively. Table 8 shows a comparison between features used in the final model using various variable importance techniques. The first two techniques are drop-column importances, the only difference being how RMSE is being measured. “Drop-Column Importance: Test Set” uses the test set to evaluate RMSE, while “Drop-Column Importance: Cross Validation” uses 10-fold cross-validation to calculate RMSE. SKL RFR’s default Gini Importance is also included here for comparison, but it should not be used to determine variable importance due to possible bias as mentioned earlier.

In both drop-column importance columns, it can be seen that Temp_Min is by far the most important feature, causing a large increase in RMSE when not used to train the model. Precipitation and NDVI are also very important. Interestingly, Humidity is judged as very important by “Drop-Column Importance: Test Set,” but more than ten times less important by “Drop-Column Importance: Cross Validation.” We do not know why this is, as there should be no significant difference in data characteristics between the test and train sets. Overall Fire is found least important,

however, it is still worthwhile to note that it does improve accuracy when used, even if not by a large amount.

Table 8: Various Variable Importance Results

Importance Method	Drop-Column Test Set	Drop-Column CV Score	Drop-Column OOB Score	Permutation Importance	Impurity Importance
Temp_Min	0.84956	1.1699	0.18070	3.7641	0.58123
Hum	0.20771	0.049327	0.01045	0.22517	0.15522
NDVI	0.10136	0.26923	0.052345	0.08498	0.18613
Precip	0.081598	0.075605	0.014726	0.11564	0.044031
Fire	0.03779	0.019902	0.001193	0.04180	0.033388

Partial Dependence Plots and Individual Condition Expectation Plots

While variable importance is useful for evaluating features, it is ultimately limited and does not say anything about how the feature contributes to the model. To find out how each feature contributes, we used Partial Dependence Plots (PDP) and Individual Condition Expectation (ICE) plots. These are model-agnostic machine learning inspection techniques, meaning they can be used on any model. We chose these techniques because Random Forest is mostly considered a “black box” model, meaning it cannot be inspected directly like a decision tree for example (Molnar, 2022). Figure 8 shows PDP and ICE plots for every feature in the final model.

Technical Explanation of PDP and ICE

ICE plots displays one line per sample, or more practically, one line per sample of a small subset of the total data points. Each line represents the value the model predicts for that sample as one feature in the sample changes. For example, if one chosen sample had the values: {precip: 5, temp_min: 12, NDVI: 0.5, hum: 0.006, fire: 0}, and we are graphing the ICE for temp_min, we would generate this sample’s line by plugging values from -5 to 20 ℃ in for temp_min (ignoring the actual value there) while keeping all the other feature values constant. This way we can control for the effect of temp_min. PDP is the averaging of all ICE lines (Molnar, 2022).

In Figure 8 PDP and ICE are plotted together, ICE lines are blue, and the PDP line is orange. On the top x-axis of each plot are the values that are plugged into the feature. So values 0 to 30 mm were plugged in for Precipitation, -5 to 20 ℃ for Temperature, and so on. On the y-axis of each plot is the prediction that the model gave for mosquito abundance based on the newly augmented sample.

Interpreting PDP and ICE

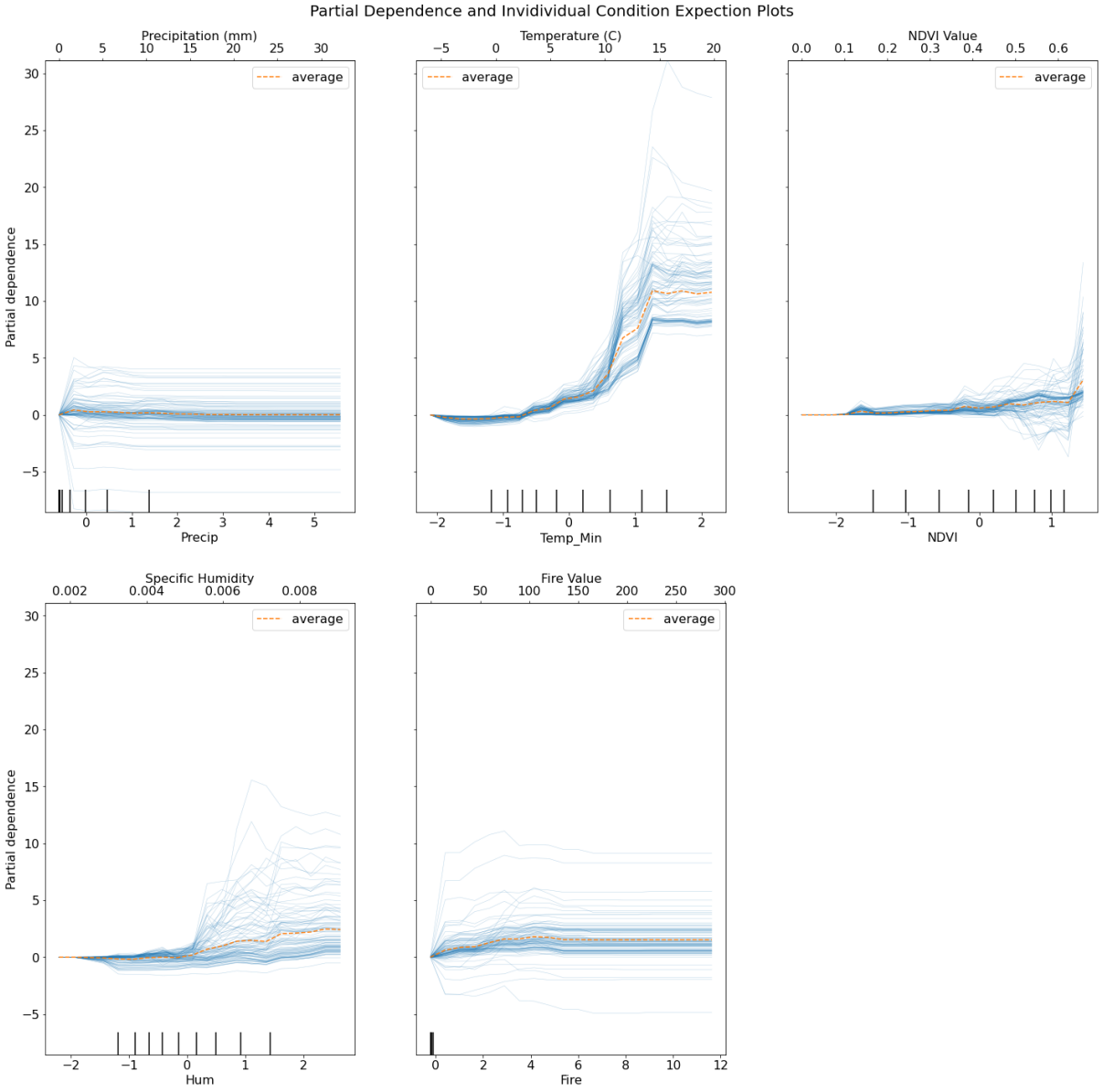
The plots show how the model reacts when a feature changes, for example, in the Figure 8 Temperature plot, mosquito abundance rises above zero at around temp_min=2 ℃, then continues rising until a maximum at temp_min=15 ℃, and plateaus afterward. This means the model generally predicts that mosquito abundance will change this way based on temp_min alone since it is controlled for by keeping all other feature values constant (Molnar, 2022).

Observations

Some conclusions can be reached by combining these plots with the variable importance values we discussed earlier because PDP can be used as a representation of variable importance (Greenwell et al., 2018). These plots still show that Temperature is clearly the most important feature, displaying a significant change in predicted mosquito abundance when temp_min is changed. NDVI and Humidity are also shown to be crucial. It should be noted that for Humidity, the ICE plot lines are very dispersed, which may indicate that how mosquito abundance reacts to humidity changes depends on other features such as temperature. Contrary to variable importance findings, Precipitation does not seem to significantly affect mosquito abundance based on the PDP. However many ICE lines jump in both directions at around 2 mm of precipitation, which may indicate that like Humidity, Precipitation's effect on mosquito abundance depends on other features. The Fire plot behaves as expected, showing some mosquito abundance variation is explained by Fire data, but not a lot. Fire's effect on mosquito abundance likely depends on other features as well, as shown by the large variation in ICE lines.

Clear numerical conclusions can be reached for Temperature and NDVI. The Temperature plot shows that mosquito abundance begins rising at around 2 ℃, but that the optimal minimum daily temperature for mosquitoes is likely around 15 ℃, which reflects mosquito literature: Ciota et al. (2014) found the highest proportions of blood-fed Culex mosquitoes laying eggs at the 16-28 ℃ range. The NDVI plot clearly shows that at very high values, 0.6 and above, mosquito abundance drastically increases. This is likely because high NDVI values represent leafy, green, lush vegetation that can provide necessary shade for mosquito oviposition.

Figure 8


Partial Dependence and Invidividual Condition Expection Plots
Precipitation (mm)
Temperature (C)
NDVI Value
Specific Humidity
Fire Value
average
Partial dependence
Precip
Temp_Min
NDVI
Hum
Fire

Discussion

In this study, we investigated the importance of wildfire data on mosquito populations by programming two distinct random forest regression models, one leveraging the wildfire data and one as a control. Previous studies mostly opt to use continuous ecological and meteorological data; for example, a study in Kern County, California used rainfall, snow depth, and water content as factors to analyze weather's impact on mosquito abundance (Wegbreit & Reisen, 2000).

However, our study provides another factor that can aid researchers in gaining more accurate statistics on mosquito abundance. In the research, it was concluded that while wildfire data aids the accuracy of the model, it was unclear how it does so. This is because our study used random forest regression models, which is a black box model – a model where the processes between the input and output are "blacked out" and not shown. Though, we did uncover that while wildfire data holds fractional importance in predicting mosquito abundance, models should continue to emphasize factors such as temperature, humidity, and NDVI in their models. Ultimately, our results do not fully align with other papers' claims that fire helps regulate or has a significant impact on mosquito populations (Agee, 1996; Scasta, 2015; Whittle et al., 1993). Though nearly 43% of Shasta's land cover is comprised of conifers, it may not set fire as often due to Shasta's wetter landscape and larger amounts of precipitation compared to the rest of California (GLOBE, 2022; Barker et al., 2010). The lower significance of wildfire data was hypothesized earlier on in the study, as wildfires are rare occurrences and have fewer data points that impact the model's result.

Additionally, the creation of the models in this study brought along key information that can aid future mosquito abundance models in gaining more accurate results. Specifically, our study found interesting findings on both lag and variable choice. Though many mosquito abundance papers used months of lag, those lag times did not improve the accuracy of our model when we tested them, confirming that lag times vary in different areas of the world (Wegbreit & Reisen, 2000; Poh et al., 2019; Chang et al., 2016). While looking for meteorological variables, many models prioritized relative humidity and precipitation for predicting mosquito abundance; however, we found that specific humidity correlates with summer mosquito abundance better, and precipitation was not nearly as important since it did not directly affect mosquito abundance as usual (Drakou et al., 2020). Past literature used different variations of temperature, but mosquitoes requiring a minimum temperature to function made the most sense for our model through testing (Arora et al., 2022; Reisen et al., 2008). We suspect this is largely due to location; for instance, California itself had vastly different precipitation patterns in the Northern and Southern parts (Barker et al., 2010). Thus, future mosquito abundance models should focus on testing for their own lag times and variable importance since they are not universally similar.

We also took a different approach to trap types. Many papers used New Jersey Light traps (NJLT), which use light to attract mosquitoes and kill them with poison, for mosquito abundance counts; however, we chose gravid and CO2 traps, which mimic natural conditions, to create a parallel model to nature (Barker et al., 2010; Jian et al., 2014). We encourage future papers to do so as well, especially if the focus is on a natural disaster, such as wildfires.

While improvements to the study such as expanding the area of interest or not using a black box model to see the inner workings of the model can be made, this research provides key information for researchers regarding the importance of wildfire data, the utilization of regression models, and mosquito trapping to accurately predict mosquito populations.

Conclusion

In summary, our final model predicts mosquito abundance in Shasta County, CA with remarkable accuracy. However, unlike previous papers, our model cannot be used for prediction since most features do not have any lag (Schneider et al., 2021). Therefore it has no immediate public health applications. Because of this, our contribution to machine learning and mosquito research is more through our methods than our final model. We show that feature variations like Specific Humidity and Minimum Temperature perform the best in climates like Shasta, which has wet winters and dry summers. Going forward we suggest research into feature variations for every mosquito abundance model because optimal feature variation changes based on AOI. We also evaluated optimization techniques like Grid Search and Bayesian Search and found that in a practical application, Bayesian Search reflects its theoretical effectiveness and was the best HP optimization technique. We believe the results are sound enough that Bayesian Search should be suggested as the HP optimization technique for any future mosquito abundance model research. We also found Bayesian Search to be simple to implement because the available libraries integrated right into SKL's ecosystem. If research is using the popular statistical machine learning language R instead of Python, there are also many ready-made packages in R for Bayesian Search.

Apart from model building, our research also provides insight into model inspection and interpretation. We compared different variable importance techniques and going forward we suggest using drop-column importance if feasible, and permutation importance otherwise. On top of that, we combined variable importance results with PDP and ICE plots to show a full picture of how our model is using each feature. These inspection techniques reveal that our model found real relationships between features and mosquito abundance, most notably for temperature. The fact that our model discovered very similar optimal mosquito temperatures as field and experimental research is astounding. Further research should be done in this direction because our research proves machine learning can be a supplement to observational and experimental findings.

There are of course many venues for improvement. One that should be heavily considered is using daily data instead of consolidating daily data into weekly data. Daily data will be far more continuous, less volatile, and will result in a more accurate model. Another area that can be looked into is other fire measurements. We created our own fire index because it was convenient for the data we could find on Google Earth Engine. Other fire measurements such as area burned and radiative power should be evaluated as well. For further public health research, using our methods on West Nile Virus transmission rates or another disease-related metric may be useful. One idea could be to use PDP and ICE plots to analyze temperature's effect on disease metrics instead of abundance. Studies like Kilpatrick et al. (2008) have experimentally shown that temperatures exceeding 28℃ cause a dramatic increase in Culex mosquito disease transmission. Machine learning analysis could confirm these findings using real-world data in areas that are vulnerable to mosquito-borne diseases.

Acknowledgements

IVSS Badges

I am

References

About the area. Forest Service National Website. (n.d.). https://www.fs.usda.gov/main/stnf/about-forest/about-area

Arora, A. K., Sim, C., Severson, D. W., & Kang, D. S. (1AD, January 1). *Random Forest analysis of impact of abiotic factors on Culex pipiens and Culex quinquefasciatus occurrence*. Frontiers. Retrieved July 29, 2022, from https://doi.org/10.3389/fevo.2021.773360

Atlas of the biodiversity of California. (2003). In E. Kauffman, M. Parisi, D. Stermer, & California (Eds.), *Berkeley Law*. California Department of Fish and Game. https://lawcat.berkeley.edu/record/426640

Barker, C. M., Eldridge, B. F., & Reisen, W. K. (2010). Seasonal Abundance of *Culex tarsalis* and *Culex pipiens* Complex Mosquitoes (Diptera: Culicidae) in California. *Journal of Medical Entomology*, *47*(5), 759–768. https://doi.org/10.1603/me09139

Barker, C. M., Eldridge, B. F., Reisen, W. K., Park, B. K., Johnson, W. O., & Melton, F. (2010). Temporal Connections between Culex tarsalis Abundance and Transmission of Western Equine Encephalomyelitis Virus in California. *The American Journal of Tropical Medicine and Hygiene*, *82*(6), 1185–1193. https://doi.org/10.4269/ajtmh.2010.09-0324

Belgiu, M., & Drăguţ, L. (2016). Random Forest in remote sensing: A review of applications and Future Directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Boulesteix, A.-L., Bender, A., Lorenzo Bermejo, J., & Strobl, C. (2011). Random Forest Gini Importance favours snps with large minor allele frequency: Impact, sources and recommendations. Briefings in Bioinformatics, 13(3), 292–304. https://doi.org/10.1093/bib/bbr053

California Department of Forestry and Fire Protection. (2022). Incidents Overview. California. https://www.fire.ca.gov/incident

Carbajo, A. E., Curto, S. I., & Schweigmann, N. J. (2006). Spatial distribution pattern of oviposition in the Mosquito aedes aegypti in relation to urbanization in Buenos Aires: Southern Fringe Bionomics of an introduced vector. *Medical and Veterinary Entomology*, *20*(2), 209–218. https://doi.org/10.1111/j.1365-2915.2006.00625.x

Chang, K., Chen, C.-D., Shih, C.-M., Lee, T.-C., Wu, M.-T., Wu, D.-C., Chen, Y.-H., Hung, C.-H., Wu, M.-C., Huang, C.-C., Lee, C.-H., & Ho, C.-K. (2016). Time-Lagging Interplay Effect and Excess Risk of Meteorological/Mosquito Parameters and Petrochemical Gas Explosion on Dengue Incidence. *Scientific Reports*, *6*(1). https://doi.org/10.1038/srep35028

Ciota, A. T., Matacchiero, A. C., Kilpatrick, A. M., & Kramer, L. D. (2014). The Effect of Temperature on Life History Traits of *Culex* Mosquitoes. *Journal of Medical Entomology*, *51*(1), 55–62. https://doi.org/10.1603/me13003

Cleckner, H. L., Allen, T. R., & Bellows, A. S. (2011). Remote Sensing and modeling of mosquito abundance and habitats in coastal Virginia, USA. *Remote Sensing*, *3*(12), 2663–2681. https://doi.org/10.3390/rs3122663

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, *88*(11), 2783–2792. https://doi.org/10.1890/07-0539.1

Drakou, K., Nikolaou, T., Vasquez, M., Petric, D., Michaelakis, A., Kapranas, A., Papatheodoulou, A., & Koliou, M. (2020). The Effect of Weather Variables on Mosquito Activity: A Snapshot of the Main Point of Entry of Cyprus. *International Journal of Environmental Research and Public Health*, *17*(4), 1403. https://doi.org/10.3390/ijerph17041403

Earth Engine Data Catalog (2022). MOD14A1.006: Terra Thermal Anomalies & Fire Daily Global 1km. (2022). https://developers.google.com/earth-engine/datasets/catalog/MODIS_006_MOD14A1?hl=en#citations

Fire, D., & California. (n.d.). *National Interagency Coordination Center Wildland Fire Summary and Statistics Annual Report 2021*. https://www.predictiveservices.nifc.gov/intelligence/2021_statssumm/annual_report_2021.pdf

Global Learning and Observations to Benefit the Environment (GLOBE) Program, 2022, https://www.globe.gov/globe-data

Greenwell, B.M., Boehmke, B.C., McCarthy, A.J.: A simple and effective modelbased variable importance measure. arXiv preprint arXiv:1805.04755 [stat.mL] (2018)

govinfo. (n.d.). www.govinfo.gov. Retrieved July 29, 2022, from https://www.govinfo.gov/app/details/FR-2016-02-05/2016-02269

Guo, S., Ling, F., Hou, J., Wang, J., Fu, G., & Gong, Z. (2014). Mosquito surveillance revealed lagged effects of mosquito abundance on mosquito-borne disease transmission: A retrospective study in Zhejiang, China. *PLoS ONE*, *9*(11). https://doi.org/10.1371/journal.pone.0112975

How different tree species impact the spread of wildfire - Alberta. (n.d.). Retrieved July 30, 2022, from https://www1.agric.gov.ab.ca/$department/deptdocs.nsf/all/formain15744/$FILE/tree-species-impact-wildfire-aug03-2012.pdf

Hwang, M.-J., Kim, H.-C., Klein, T. A., Chong, S.-T., Sim, K., Chung, Y., & Cheong, H.-K. (2020). Comparison of climatic factors on mosquito abundance at US Army Garrison Humphreys, republic of korea. *PLOS ONE*, *15*(10). https://doi.org/10.1371/journal.pone.0240363

Kofidou, M., de Courcy Williams, M., Nearchou, A., Veletza, S., Gemitzi, A., & Karakasiliotis, I. (2021). Applying remotely sensed environmental information to model mosquito populations. *Sustainability*, *13*(14), 7655. https://doi.org/10.3390/su13147655

Lee, K. Y., Chung, N., & Hwang, S. (2016). Application of an artificial neural network (ANN) model for predicting Mosquito abundances in urban areas. *Ecological Informatics*, *36*, 172–180. https://doi.org/10.1016/j.ecoinf.2015.08.011

Madzokere, E. T., Hallgren, W., Sahin, O., Webster, J. A., Webb, C. E., Mackey, B., & Herrero, L. J. (2020). Integrating statistical and mechanistic approaches with biotic and environmental variables improves model predictions of the impact of climate and land-use changes on future mosquito-vector abundance, diversity and distributions in Australia. *Parasites & Vectors, 13*(1). https://doi.org/10.1186/s13071-020-04360-3

Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Github.

Mordecai EA, Paaijmans KP, Johnson LR, Balzer C, Ben-Horin T, De Moor E, McNally A, Pawar S, Ryan SJ, Smith TC, Lafferty KD. Optimal temperature for malaria transmission is dramatically lower than previously predicted. *Ecology Letters*. 2012;16(1):22-30. https://doi.org/10.1111/ele.12015

NASA Earth Observatory. (n.d.). *What's behind California's surge of large fires?* NASA. Retrieved July 29, 2022, from https://earthobservatory.nasa.gov/images/148908/whats-behind-californias-surge-of-large-fires

Number of wildfires to rise by 50% by 2100 and governments are not prepared, experts warn. (2022, February 23). UN Environment. https://www.unep.org/news-and-stories/press-release/number-wildfires-rise-50-2100-and-governments-are-not-prepared

Poh, K. C., Chaves, L. F., Reyna-Nava, M., Roberts, C. M., Fredregill, C., Bueno, R., Debboun, M., & Hamer, G. L. (2019). The influence of weather and weather variability on mosquito abundance and infection with West Nile virus in Harris County, Texas, USA. *Science of The Total Environment*, *675*, 260–272. https://doi.org/10.1016/j.scitotenv.2019.04.109

PRISM Climate Group, Oregon State University, https://prism.oregonstate.edu/, data created 4 Feb 2014, accessed 20 Jul 2022.

Reisen, W.K., Cayan, D.R., Tyree, M., Barker, C.M., Eldridge, B.F., & Dettinger, M.D. (2008). Impact of climate variation on mosquito abundance in California. *Journal of vector ecology : journal of the Society for Vector Ecology*. https://doi.org/10.3376/1081-1710(2008)33[89:IOCVOM]2.0.CO;2

Ruiz, M.O., Chaves, L.F., Hamer, G.L. *et al.* Local impact of temperature and precipitation on West Nile virus infection in *Culex* species mosquitoes in northeast Illinois, USA. *Parasites Vectors* 3, 19 (2010). https://doi.org/10.1186/1756-3305-3-19

Schneider, J., Greco, A., Chang, J., Molchanova, M., & Shao, L. (2021). Predicting West Nile virus mosquito positivity rates and abundance: A comparative evaluation of machine learning methods for epidemiological applications. *Earth and Space Science Open Archive*. https://doi.org/10.1002/essoar.10509422.1

Shasta County MVCD. (2022). [Shasta County data for Culex pipiens and Culex tarsalis from CO2 and gravid traps from 2010-2022] [Unpublished raw data]. Shasta County MVCD.

Snoek, J., Larochelle, H, & Adams, R. P. (Eds.). (2012). *Proceedings of the 25th International Conference on Neural Information Processing Systems.* ACM Digital Library. https://dl.acm.org/doi/10.5555/2999325.2999464

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*(1). https://doi.org/10.1186/1471-2105-9-307

Strobl, C., Boulesteix, AL., Zeileis, A. et al. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 8, 25 (2007). https://doi.org/10.1186/1471-2105-8-25

Trees of the Shasta-Trinity. (n.d.). https://www.fs.usda.gov/Internet/FSE_DOCUMENTS/fsm9_008614.pdf

United States Department of Agriculture Forest Service. (2016). Ochoco, Umatilla, Wallowa-Whitman National Forests; Oregon and Washington; Blue Mountains Forest Resiliency Project. Federal Register.

Voiland, A. (2021, October 4). *What's Behind California's Surge of Large Fires?* Earthobservatory.nasa.gov.

https://earthobservatory.nasa.gov/images/148908/whats-behind-californias-surge-of-large-fires

Weather averages Shasta, California. Temperature - Precipitation - Sunshine - Snowfall. (n.d.). https://www.usclimatedata.com/climate/shasta/california/united-states/usca1045

Wegbreit, J., & Reisen, W. K. (2000). Relationships among weather, mosquito abundance, and encephalitis virus activity in California: Kern County 1990-98. *Journal of the American Mosquito Control Association*, *16*(1), 22–27.

World Health Organization. (n.d.). *Vector-borne diseases*. World Health Organization. Retrieved July 29, 2022, from https://www.who.int/news-room/fact-sheets/detail/vector-borne-diseases

Wu, J., Chen, X.-Y., Zhang, H., Xiong, L.-D., Lei, H., & Deng, S.-H. (2019). Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimization. *Journal of Electronic Science and Technology*, *17*(1), 26–40. https://doi.org/https://doi.org/10.11989/JEST.1674-862X.80904120

Zanaga, D., Van De Kerchove, R., De Keersmaecker, W., Souverijns, N., Brockmann, C., Quast, R., Wevers, J., Grosu, A., Paccini, A., Vergnaud, S., Cartus, O., Santoro, M., Fritz, S., Georgieva, I., Lesiv, M., Carter, S., Herold, M., Li, Linlin, Tsendbazar, N.E., Ramoino, F., Arino, O., 2021. ESA WorldCover 10 m 2020 v100. https://doi.org/10.5281/zenodo.5571936