Using Forest Fire Data to Improve Machine Learning Modeling of Mosquito Abundance

Raymond Lin, Haley Oba, Krish Desai, Ashmit Dewan, Zarar Haider

NASA STEM Enhancement in the Earth Sciences 2022

Partner

Abstract

Mosquitoes have been a major health concern for decades, and with climate change expanding their habitat, 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 variable's 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 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 variable importance, the purest way of calculating variable importance. We found that forest fire marginally increased accuracy, which is 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 inspection and its use cases. Within mosquito research, further work can explore other novel datasets that form a more comprehensive understanding of mosquito abundance.

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 Literature Review

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). 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). 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 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 powerful random forest method can predict variable significance and has the ability to model complex relationships between variables (Cutler et al., 2007). Rainfall, specific humidity, NDVI, and temperature are critical factors in determining mosquito abundance (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.

Methodology

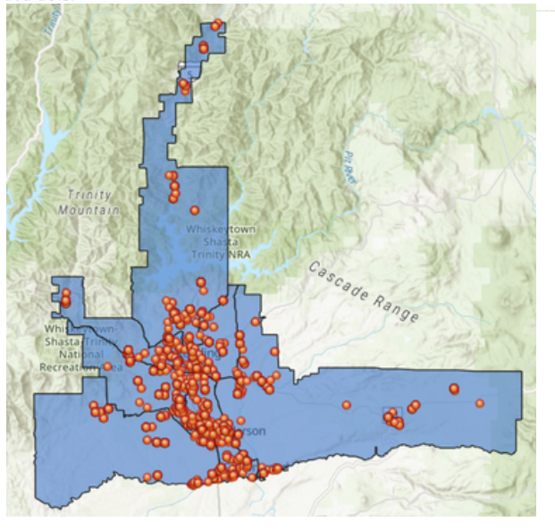


Figure 1. Shasta MVCD Surveillance Area with all mosquito observations from 2010-2022 plotted as red dots.

Shasta County, our AOI, 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). It has also suffered from a large amount of forest fires in the past 12 years (California Department of Forestry and Fire Protection, 2022).

We used Google Earth Engine and University of Idaho Gridded Surface Meteorological Dataset (GRIDMET), Daily Spatial Climate Dataset (PRISM), NDVI values from MODIS Terra Daily NDVI Dataset provided by Google, and MOD14A1.006 to acquire ecological datasets. *Culex pipiens* and *Culex tarsalis* from gravid and CO2 trap abundance data for this study were provided by Shasta MVCD. All data were acquired across our weekly time frame from 2010-01-13 to 2022-07-10 as determined by EPI. All-time GLOBE Land Cover data was also acquired for Shasta County.

Data Statistics

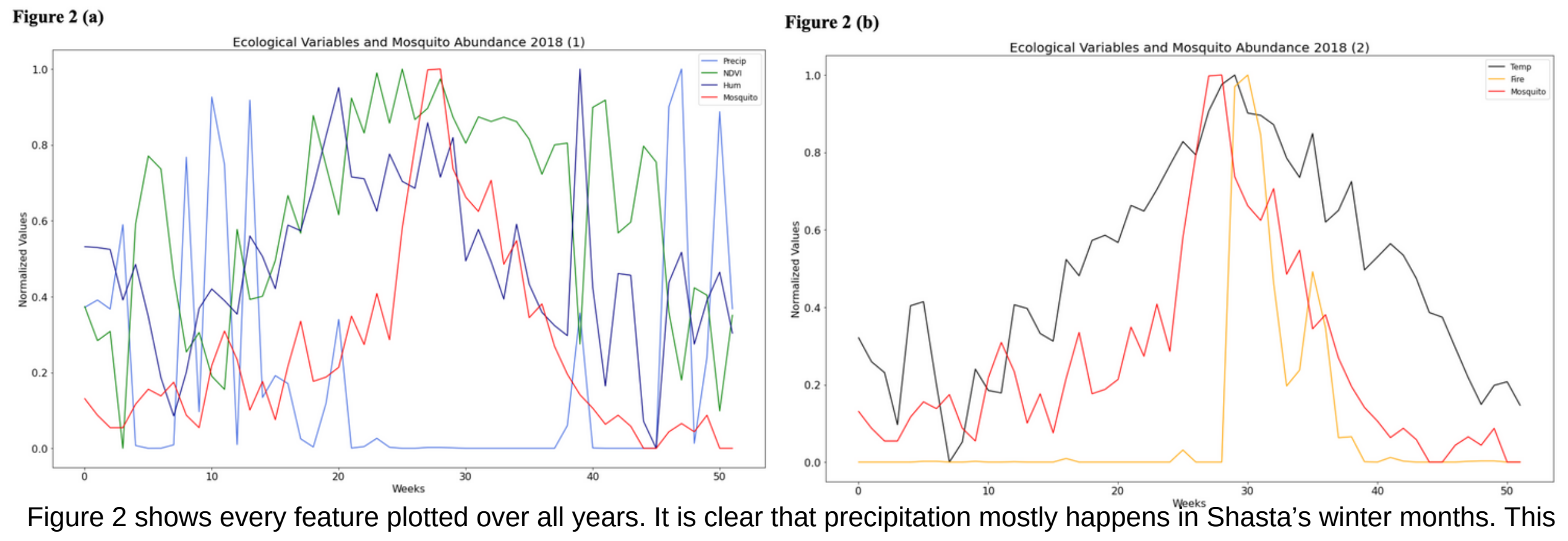


Figure 2 (a) **Figure 2 (b)**

Figure 2 shows every feature plotted over all years. It is clear 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 one has one sharp peak at around week 30 of 2018, which 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.

The features in Figure 2 are from the daily mean datasets of each feature, but previous studies have found variations like the daily minimum or maximum to be more helpful (Lee KY et al, 2017). However, we only found minimum temperature to be useful the only useful variation, and relative humidity did not work for our model.

Finding Optimal Lag

Figure 5: Weeks of Lag vs Pearson's R

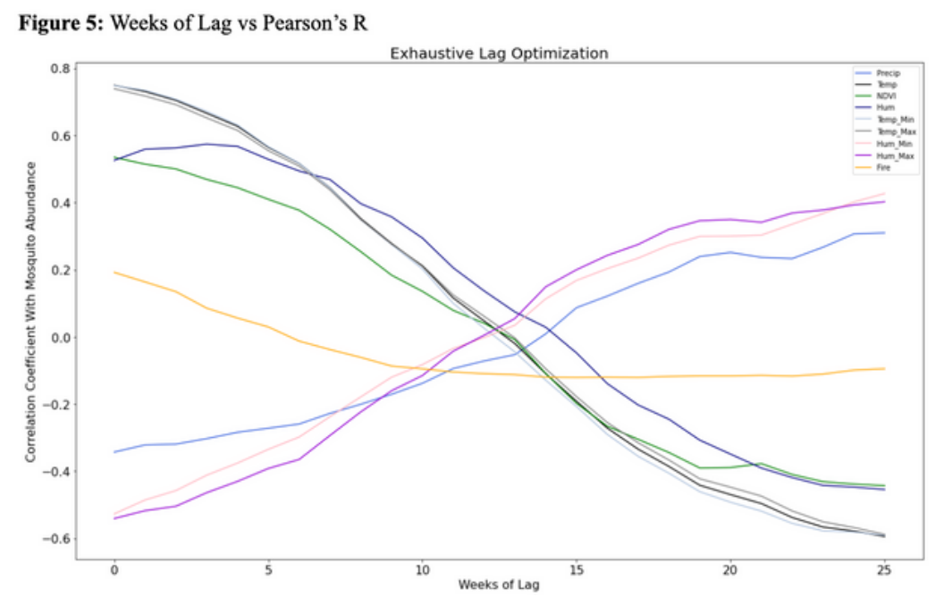


Table 2: Optimal Lag Weeks

	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

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. Chang et al. (2016) discussed how feature lag times will differ under different climates, so instead of using previous lag values, we decided to find our own. Table 2 shows the results of Figure 5, with the best correlation coefficient achieved for each feature and how many weeks created that correlation.

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. Chang et al. (2016) discussed how feature lag times will differ under different climates, so instead of using previous lag values, we decided to find our own. 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.

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 & Dragut, 2016; Schneider et al., 2021). Table 6 shows the results of these trials with different variations of features.

Table 6. Test Set Results For Each Model

Model Name / Tested Variation	Base	Temp_Min	Temp_Max	Hum_Min	Hum_Max	Final
RMSE	4.031491	4.022265	4.331339	4.445316	4.252529	3.947571
MAE	2.619923	2.648673	2.739660	2.799292	2.719619	2.574583
R^2	0.650839	0.652435	0.596969	0.575479	0.611502	0.665224

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. 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.

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/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 (R2). 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	4.172734	4.072118	4.031491
MAE	2.760101	2.650127	2.619923
R^2	0.625945	0.643766	0.650839

The tables below show hyperparameter optimization results. We chose to optimize "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," "max_depth," and "max_leaf_nodes" because they make each tree less specific and increase overall variance, which helps prevent overfitting. Square brackets denote discrete choices and parenthesis denote ranges.

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

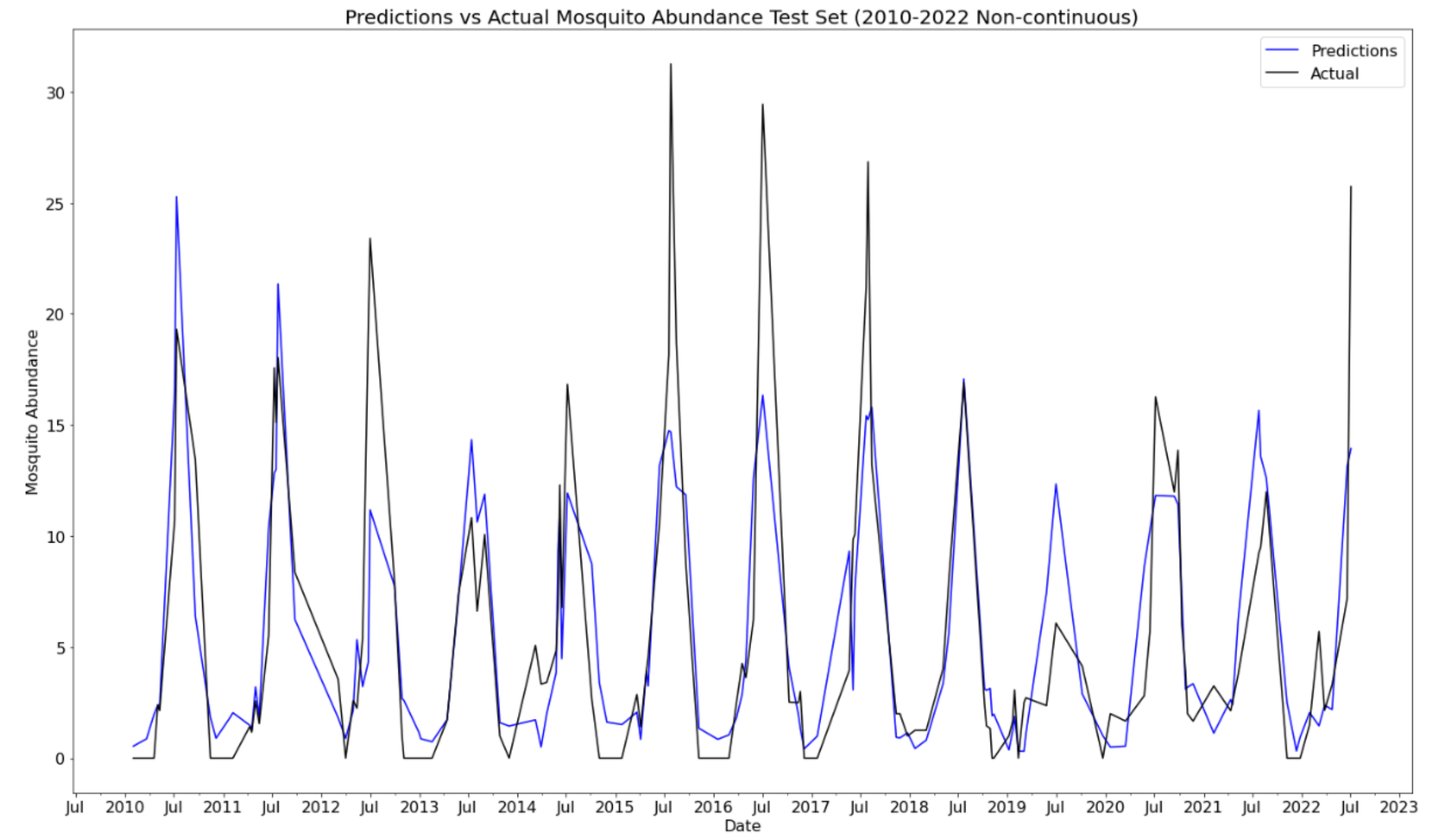
Table 5: Bayesian Search Tried and Optimal Values

	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

Results

Model Performance

Figure 6



Overall performance is very good, beating out Schneider's RFR RMSE of 7.48 by a significant margin (Schneider et al., 2021). Figure 6 shows our final model's predictions on the test set. Since the test set was stratified across years, the plot has around 20% of the weeks from each year, so it is not showing every week continuously. It can be seen that the predictions are highly accurate, with prediction and actual mosquito abundance peaks lining up almost every year. The model is less good at predicting exactly how high that peak will be, which is understandable given the extreme right-skewed and outlier-prone mosquito abundance data.

Feature Evaluation: Variable Importance, PDP, and ICE

Table 8: Variable Importance Results

Variable Importance Method	Drop Importance: Test Set
Temp_Min	1.021807
Hum	0.308829
Precip	0.277701
NDVI	0.258024
Fire	0.044106

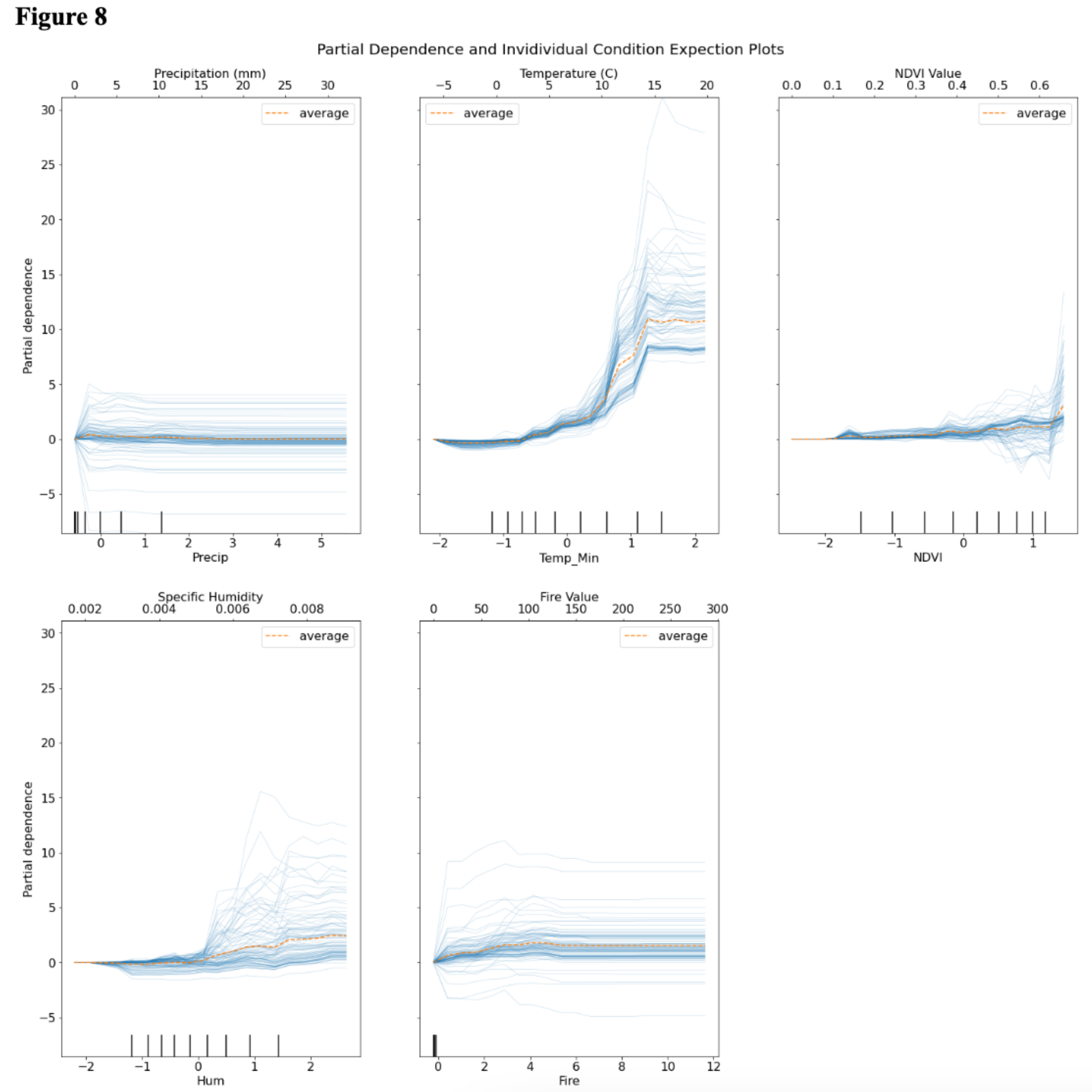
From Table 8 it can be seen that Temp_Min is by far the most important feature, causing a large increase in RMSE when not used to train the model. Precipitation and NDVI are also very important. Humidity, Precipitation, and NDVI are all judged to be fairly important. This means the model may have found a nonlinear relationship between Precipitation and Mosquito Abundance, because when looking at linear correlations, Precipitation performed poorly. 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. On top of that, Fire is a rare-occurance dataset, meaning the model will not get to use it very often, since most values are 0s.

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.

The plots show how the model reacts when a feature changes, with various values of the feature on the top x-axis, and predicted mosquito abundance on the y-axis. Each ICE blue line shows what the model would predict for that sample if all other features are kept as is, and the feature in question is changed. It is like looking into the "black box" and seeing what kinds of relationships the model was able to synthesize from the data (Molnar, 2022).

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



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. 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. 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. Wildfires hold lower significance because they are rare occurrences and have fewer data points that impact the model's result. 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). 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, something we encourage future papers to do so as well, especially if the focus is on natural disasters, such as wildfires.

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. Due to 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 proposed as the HP optimization technique for any future mosquito abundance model research.

Apart from model building, our research also provides insight into model inspection and interpretation. Further research should be done in this direction because our research proves machine learning can be a supplement to observational and experimental findings.

References

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

Belgiu, M. and Dragut, L. (2016) Random Forest in Remote Sensing: A Review of Applications and Future Directions. ISPRS Journal of Photogrammetry and Remote Sensing, 114, 24-31.

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

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 ofculexmosquitoes. Journal of Medical Entomology, 51(1), 55–62. https://doi.org/10.1603/me13003

Cleckner HL, Allen TR, Bellows AS. Remote Sensing and Modeling of Mosquito Abundance and Habitats in Coastal Virginia, USA. Remote Sensing. 2011; 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

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

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

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 KY, Chung N, Hwang S. Application of an artificial neural network (ANN) model for predicting mosquito abundances in urban areas. Ecological Informatics. 2016;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.

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

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

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.

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

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

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

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

Acknowledgements

The material contained in this poster is based upon work supported by the National Aeronautics and Space Administration (NASA) cooperative agreements NNX16AE28A to the Institute for Global Environmental Strategies (IGES) for the NASA Earth Science Education Collaborative (NESEC) and NNX16AB89A to the University of Texas Austin for the STEM Enhancement in Earth Science (SEES). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NASA.

Huge thanks to Dr. Rusanne Low, Ms. Cassie Soeffing, Mr. Peder Nelson, Dr. Erika Podest, Andrew Clark, and the rest of the SEES team for this opportunity. We would also like to give special thanks to our mentor Julianna Schneider.