A Section 508–conformant HTML version of this article is available at https://doi.org/10.1289/EHP15593.

Machine Learning Potential for Identifying and Forecasting Complex Environmental Drivers of *Vibrio vulnificus* Infections in the United States

Amy Marie Campbell,[1,2] Jordi Manuel Cabrera-Gumbau,[3] Joaquin Trinanes,[4] Craig Baker-Austin,[2] and Jaime Martinez-Urtaza[2,3]

[1]School of Ocean and Earth Science, University of Southampton, National Oceanography Centre, Southampton, UK
[2]Centre for Environment, Fisheries and Aquaculture Science (CEFAS), Weymouth, UK
[3]Department of Genetics and Microbiology, Autonomous University of Barcelona, Barcelona, Spain
[4]Department of Electronics and Computer Sciences, University of Santiago de Compostela, Santiago de Compostela, Spain

Background: Environmental change in coastal areas can drive marine bacteria and resulting infections, such as those caused by *Vibrio vulnificus*, with both foodborne and nonfoodborne exposure routes and high mortality. Although ecological drivers of *V. vulnificus* in the environment have been well-characterized, fewer models have been able to apply this to human infection risk due to limited surveillance.

Objectives: The Cholera and Other *Vibrio* Illness Surveillance (COVIS) system database has reported *V. vulnificus* infections in the United States since 1988, offering a unique opportunity to both explore the forecasting capabilities machine learning could provide and to characterize complex environmental drivers of *V. vulnificus* infections.

Methods: Machine learning models, in the form of random forest classification models, were trained and refined using the epidemiological data from 2008 to 2018, six environmental variables (sea surface temperature, salinity, chlorophyll *a* concentration, sea level, land surface temperature, and runoff rate) and categorical encoders to assess our predictive potential to forecast *V. vulnificus* infections based on environmental data.

Results: The highest-performing model, which used balanced classes, had an Area Under the Curve score of 0.984 and a sensitivity of 0.971, highlighting the potential of machine learning to anticipate areas and periods of *V. vulnificus* risk. A higher false positive rate was found when the model was applied to real-world imbalanced surveillance data, which is pertinent amid modeled underreporting and misdiagnosis ratios of *V. vulnificus* infections. Further models were also developed to explore multilevel spatial resolution, finding state-specific models can improve specificity and early warning system potential by exclusively using lagged environmental data.

Discussion: The machine learning approach was able to characterize nonlinear and interacting environmental associations driving *V. vulnificus* infections. This study accentuates the potential of machine learning and robust surveillance for forecasting environmentally associated marine infections, providing future directions for improvements, further application, and operationalization. https://doi.org/10.1289/EHP15593

Introduction

Consequences of climate change in coastal areas, a significant interface between humans and marine microbial communities, include the shifting dynamics of environmentally sensitive marine pathogens pertinent to human health. A key example of pathogens responding to climate change are *Vibrio* spp.,[1] a group of Gram-negative bacteria that reside in coastal waters, either free-living or attached to chitinous marine organisms,[2] and exhibit environmental dependencies. Increases in *Vibrio*-associated infections amid warming marine environments[3] are of growing concern to human health. Of these, *Vibrio vulnificus* (Vv) is transmitted to humans when contaminated water enters open wounds, where infections progress to necrotizing fasciitis and septicemia, or through the consumption of contaminated raw shellfish, similar to other pathogenic *Vibrio* bacteria.[1] Common exposure routes alongside gastrointestinal transmission include recreational water activities, such as swimming or fishing, or occupational exposure to contaminated water such as oyster harvesting, particularly when open wounds are present. It has an exceptionally high mortality rate ranging from 20% to 50%, depending on severity of infection,[4] and has been attributed to 95% of seafood-related deaths in the United States, making it the most fatal reported foodborne pathogen.[5]

As a constituent of the marine environment, Vv bacteria exhibit particular environmental tolerances and thresholds. Generally, Vv favors water temperatures above 16°C and moderate salinities between 5 and 20 practical salinity units (psu).[6] Plankton is a key reservoir for Vv bacteria in the marine environment, and it has been found that copepod presence, for which chlorophyll *a* concentration (as a proxy for phytoplankton density) is a secondary indicator, are associated with increased Vv presence and subsequent infection risk.[7,8] Other environmental variables can affect exposure potential, such as sea level anomalies, which increase inland intrusions of plankton and Vv bacteria,[9] or increased land temperatures, which promote human interactions with coastal waters.[1] Long-term climate trends of warming and sea level rise are therefore driving Vv dynamics, such as extending the "*Vibrio* season" into cooler months beyond the summer.[1] However, Vv can opportunistically take advantage of shorter-term extreme weather events such as heatwaves,[3] storms, or hurricanes,[7,10–12] resulting in increased prevalence of Vv infections.

It is of increasing importance to understand Vv dynamics, with increasing incidence and geographic distribution of Vv infections attributed to global climate change[4,13,14] likely to continue. Deeb et al.[6] predicted infection risk could quadruple by the mid-21st century. In the United States, climate change is predicted to increase sea surface temperatures, hurricane activity, and precipitation in the Gulf of Mexico,[15] increasing Vv suitability in a hotspot for Vv infections. Vv are native inhabitants of marine microbial communities so cannot be eradicated. Instead, to protect human health, we can monitor the spatiotemporal dynamics in the marine environment and the risk of transmission to humans, which is mediated by environmental conditions and human behavior. In particular, we need to understand the

Address correspondence to Jaime Martinez-Urtaza. Email: jaime.martinez.urtaza@uab.cat

Supplemental Material is available online (https://doi.org/10.1289/EHP15593).

The authors declare no conflicts of interest.

Conclusions and opinions are those of the individual authors and do not necessarily reflect the policies or views of EHP Publishing or the National Institute of Environmental Health Sciences.

Received 17 June 2024; Revised 6 December 2024; Accepted 30 December 2024; Published 23 January 2025.

Note to readers with disabilities: *EHP* strives to ensure that all journal content is accessible to all readers. However, some figures and Supplemental Material published in *EHP* articles may not conform to 508 standards due to the complexity of the information being presented. If you need assistance accessing journal content, please contact ehpsubmissions@niehs.nih.gov. Our staff will work with you to assess and meet your accessibility needs within 3 working days.

environmental conditions conducive to Vv survival in the environment and transmission to humans, which could proliferate infections, to facilitate development of early warning systems and deployment of mitigation measures.

Robust surveillance of reported Vv infections can provide a proxy for Vv presence in the environment and subsequent risks to human health. However, despite its global presence and high disease burden, surveillance of Vv infections is characteristically poor, with few countries systematically reporting Vv infections. The actual number of Vv infections is globally uncertain due to underreporting and a lack of epidemiological frameworks leading to differing reporting procedures.[16] These disparities have resulted in studies previously reporting Vv risk in the form of hypothetical habitat suitability,[3,17] which does not directly translate into actual Vv presence nor disease risk. However, a unique opportunity is offered by the Cholera and Other *Vibrio* Illness Surveillance (COVIS) system in the United States, where Vv infections became nationally notifiable in 2007.[18] Health officials across the United States report clinically confirmed cases of vibriosis and cholera to this centralized database with accompanying metadata on the possible exposure and sources. It is unique in providing the only national robust surveillance dataset over a sufficient time period, with high-resolution spatiotemporal metadata (in terms of where and when the infection was likely contracted), to facilitate environmental association analysis. The dataset has recently been used to assess the changing disease distribution of Vv infections in eastern United States, finding an 8-fold increase and northern expansion, and to create an ecological niche model to forecast further increases under climate scenarios.[19] In addition, COVIS data have been used for correlation analysis to identify linear relationships with environmental variables, such as temperature.[20] However, the environmental drivers of Vv can be complex in dynamic marine ecosystems, wherein the variables interact among themselves through fluxes and cycles, and bacteria respond to changes through rapid replication, exhibiting nonlinear tolerances and thresholds, which introduce various layers of complexity. The dynamic interaction of environmental conditions that drive outbreaks, and our ability to predict these, has not yet been fully explored.

This study aimed to identify complex environmental drivers of Vv infections across the United States and assess our predictive potential of infections, using the unique epidemiological surveillance database available in the United States and readily available climate datasets. Specifically, we aimed to test the ability of machine learning models to characterize nonlinear associations between Vv incidence and interacting environmental conditions. Such nonlinear relationships have been explored *in situ*, such as the quadratic relationship identified between Vv incidence and salinity from sampling in an estuary[6]; however, no model has characterized such relationships for infection incidence over a wider spatial and temporal scale. In addition, this study aimed to explore the potential of machine learning models from a retrospective forecasting perspective to quantify our ability to predict future environmentally driven Vv infections in the United States using a range of environmental data. Machine learning has previously been employed to forecast cholera outbreaks[21]—caused by a *Vibrio* species that is more frequently reported, *Vibrio cholerae*—through which an 89% outbreak sensitivity was previously achieved, suggesting they could provide value when applied to Vv, but such models have not previously been applied to this surveillance dataset. Because Vv are acquired solely through environmental exposures, in comparison with the confounding effect of human–human transmission possible for toxigenic *Vibrio cholerae*,[16] there was sufficient potential to expect to achieve a similar or higher accuracy through the novel application of similar models to Vv.

Methodology

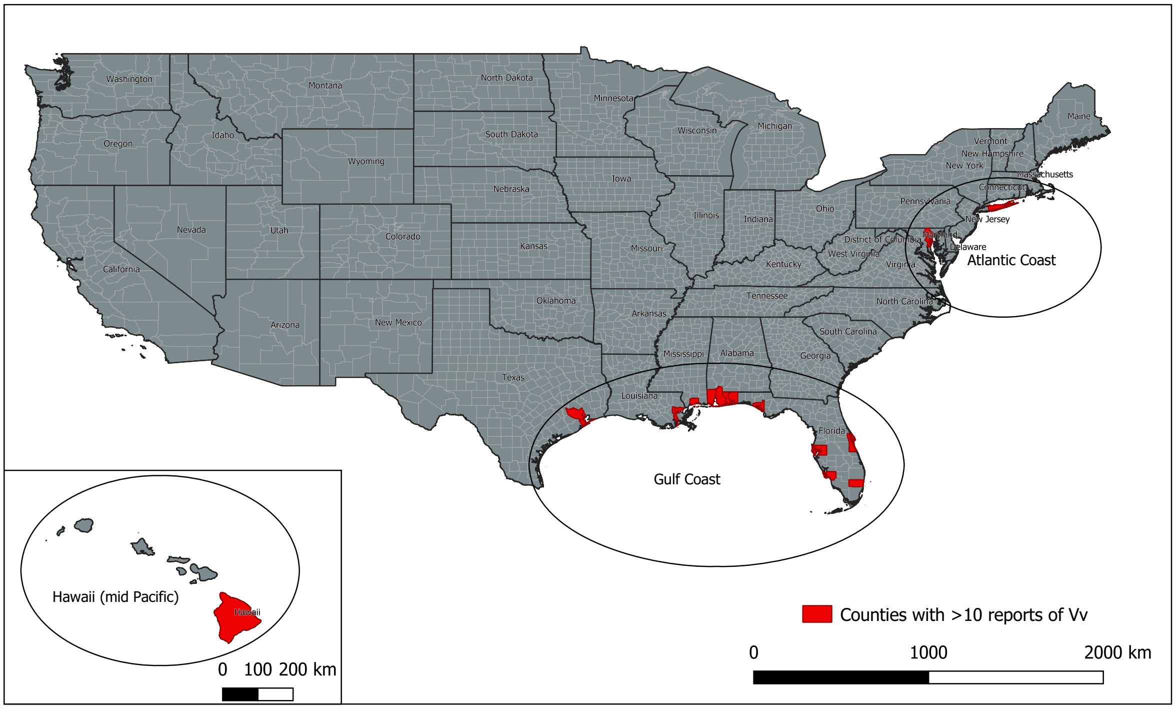
Surveillance Data

All laboratory-confirmed cases of Vv infections reported between 1988 and 2018 reported to the Cholera and Other *Vibrio* Illness Surveillance (COVIS) database were extracted (https://www.cdc.gov/vibrio/php/surveillance/) and processed into a focused subset for analysis. Data were constrained to the time period 2008–2018 to ensure standardized reporting across the study period, because *Vibrio* spp. infections only became nationally notifiable in 2007,[18] so pre-2008 records carry a temporal data reporting bias. The analysis was limited to infections contracted in the home country, as opposed to travel-associated cases, to remove spatial ambiguities. For a sufficient spatiotemporal resolution for environment-association analysis, we retained entries with a county-level spatial resolution of exposure location and monthly temporal resolution of symptom date (though these were mostly daily resolution). Sporadic reported infections, such as counties only reporting a single infection over the time period, could confound the signal by introducing noise. For this reason, we focused this study on counties that had reported 10 or more Vv infections over the time period, offering opportunity to explore sufficient epidemiological patterns. The full multiple exclusion criteria are described and quantified in Table S1. This approach resulted in 19 counties of interest in total across 8 states (Figure 1).

The Vv reports were appended into a full time series of 2008–2018 for each county. The final surveillance dataset input contained 2,641 data points in total across the 19 counties over the 11 y. Data points here refers to the rows of the dataframe that represent a particular county in a particular month (for example, January 2020 in Orleans County), for which the associated environmental data have been appended, and the presence of a Vv infection report or not recorded. A total of 289 data points represented months where Vv infections were reported, with approximately a 1:9 ratio of infection months to noninfection months, similar to that found in the surveillance dataset used to make predictive models of another *Vibrio* bacteria in Campbell et al.[21] (1:10 ratio). These infection months were encoded (1 indicating an infection reported, 0 indicating no infections reported) to provide a binary presence/absence dependence variable for the downstream machine learning models to yield a metric of infection risk presence, rather than the scale of incidence, which would be more skewed by underreporting.

Climate Data

Freely available and accessible climate data were downloaded (Table 1) for environmental variables representative of conditions that promote both Vv survival in the marine environment and the transmission potential to humans, based on previously identified associations in laboratory or environmental scenarios and the availability of open-access datasets with sufficient coverage over the study period. For the former, sea surface temperature, salinity, and chlorophyll *a* concentration time-series data were acquired. Sea surface temperature and salinity are well-reported drivers of Vv survival, with linear and threshold-based associations, respectively.[6,16] Chlorophyll *a* concentration provides a proxy for phytoplankton presence and a secondary indicator for zooplankton abundance (which, in turn, depends on phytoplankton density), because copepods provide a food source and host protection to promote *Vibrio* spp. survival, particularly in adverse weather conditions.[2,22,23] Although pH data were also acquired, they were later omitted because of their low resolution near coastal areas, prohibiting extraction of values for coastal waters corresponding to specific counties; however, the lack of spatial variability of pH in the study area[24] would likely have provided little information.



State	County
Alabama	Baldwin
Alabama	Mobile
Florida	Bay
Florida	Brevard
Florida	Broward
Florida	Escambia
Florida	Hillsborough
Florida	Lee
Florida	Pinellas
Florida	Santa Rosa
Hawaii	Hawaii
Louisiana	Jefferson
Louisiana	Orleans
Maryland	Anne Arundel
Maryland	Baltimore
Mississippi	Harrison
New York	Suffolk
Texas	Galveston
Texas	Harris

Figure 1. Selected counties of interest across the United States reporting *Vibrio vulnificus* (Vv) infections to the COVIS database. Counties were selected based on availability of Vv infection data (at least 10 reports over the period 2008–2018) and sufficient spatiotemporal metadata (monthly and county level), covering three key marine regions, with a total of 289 Vv infection months. Administrative area shapefiles provided by Global Administrative Areas database (https://gadm.org/download_country.html).

Runoff rate (representing surface and subsurface runoff of water from precipitation, melting snow, or water in soil) was used as a representative of extreme precipitation scenarios that promote salinity dilution around discharging areas and nutrient inputs promoting *Vibrio* spp. growth.[25] Variables that represented human exposure to the bacteria included sea level anomalies and land surface temperature. Specifically, sea level anomalies lead to intrusions of saline water carrying plankton and bacteria into coastal inlets where humans have greater interaction with water,[6] and land temperature affects recreational usage of water bodies.[1] The latter has previously been used as a proxy for coastal recreation in Vv studies,[19] based on a previous behavioral study.[26]

Table 1. Climate data sources and specification used in random forest classification models to predict Vv infections in the United States, 2008–2018.

Variable	Source	Specification	Available at
Sea surface temperature (SST)	Copernicus Climate Change Service[28] derived from European Space Agency (ESA) SST Climate Change Initiative (CCI) software and algorithms[29]	Daily average ocean temperature at 20 cm depth, provided at a 0.05-decimal degree resolution, in Kelvin	https://doi.org/10.24381/cds.cf608234
Sea salinity	Met Office Hadley Center Observation datasets EN.4.2.2 quality-controlled ocean data[30,31]	Monthly average subsurface salinity profiles at 5 m depth, provided at a 1-decimal degree resolution, in psu	https://www.metoffice.gov.uk/hadobs/en4/download-en4-2-2.html
Chlorophyll *a* concentration	ESA CCI Ocean Colour[32]	Chlorophyll *a* concentration, provided at a ~ 4 km resolution, in $\mathrm{mg/m^3}$	http://dx.doi.org/10.5285/1dbe7a109c0244aaad713e078fd3059a
pH	OceanSODA-ETHZ gridded data of global surface ocean carbonate system[33]	Monthly pH based on total alkalinity, provided at a 1-decimal degree resolution	https://www.ncei.noaa.gov/data/oceans/ncei/ocads/data/0220059/
Sea level anomaly	Copernicus Marine Environment Monitoring Service (CMEMS) and the Copernicus Climate Change Service (C3S) sea level gridded data from satellite observations[34]	Sea surface height (above mean sea surface computed for a 20-year mean reference period of 1993–2012), provided at a 0.25 degree resolution, in meters	https://doi.org/10.24381/cds.4c328c78
Land surface temperature	ESA CCI Land Surface Temperature[35]	Average daytime land surface temperature, provided at a 0.05-degree resolution, in Kelvin.	https://dx.doi.org/10.5285/a7e811fe11d34df5abac6f18c920bbeb
Runoff rate	ERA5-fifth generation European Center for Medium-Range Weather Forecasts (ECMWF) atmospheric reanalysis-monthly averaged data on single levels[36]	Mean rate of runoff from rainfall, melting snow or water in the soil (surface and subsurface) as if it were spread evenly over a 0.25-decimal degree grid box, in $\mathrm{kg/m^2/s}$	https://doi.org/10.24381/cds.f17050d7

Note: psu, practical salinity unit; Vv, *Vibrio vulnificus*.

Multiple variables were acquired from readily available Earth Observation sources, with satellite data previously highlighted as a key tool to track *Vibrio* spp. risk over long-term retrospective time periods and in real time.[17,27]

Each environmental time series was trimmed to the time period of interest (2008–2018) and clipped to a bounding box encompassing the entire United States (between 18.0 and 71.4° N, and between −171.8 and −67.0° E). Each environmental variable was aggregated to monthly averages, using the arithmetic mean, to match the temporal resolution being used for the epidemiological data, and interpolated to the resolution of the sea surface temperature dataset to facilitate combination into a single Network Common Data Form (NetCDF) file, an array-orientated scientific data format, containing the full environmental data time series.

Dataframe Generation

Digital vector shapefiles for each county of interest were acquired from the Global Administrative Areas database (GADM; gadm.org). Within QGIS (Open Source Geospatial Foundation Project), a 0.5-decimal degree buffer was applied to each shapefile to extend into their corresponding coastal waters, to provide a region from which the marine environmental variables could be extracted that offered sufficient data coverage of marine variables at 1-decimal degree resolution. The mean of each of the six continuous environmental variables (sea surface temperature, sea salinity, chlorophyll *a* concentration, sea level anomaly, land surface temperature, and runoff rate) for each time step within each county polygon was iteratively zonal extracted, using the Rasterio package in Python, into dataframe format. This method resulted in a spatiotemporal dataframe in which each data point represents a particular month in a particular county. The binary epidemiological feature (representing when and where infections were reported) was then appended into this climate data time series.

Several additional features were encoded. First, features were created for season (winter, spring, summer, autumn) and the three key geographic regions (Gulf Coast, Atlantic Coast, and Hawaii); full descriptions of these features can be found in Table S2. These features were generated both as categorical variables (with numbers within the same feature representing each option), and as separate binary features for each option (using One Hot Encoding). The raw environmental data were then used to generate lagged effects for 1 and 2 months previous to the infection report, due to the anticipated lag between changes in environmental conditions and the report of an infection, encompassing the full transmission pathway of population growth in the marine environment, exposure, symptoms displaying, the visiting of a clinical setting, and the confirmed laboratory report of the infection. This approach also enables the quantification of longer-term environmental drivers, including the effect of sustained optimum conditions on population growth. Previous studies have identified lagged relationships between environmental drivers in the 2 preceding months to aid prediction of *Vibrio* spp. infections.[21] Three distinct approaches were taken: first using the raw continuous monthly values from 1 and 2 months previous to the infection report, second by calculating the rate of change of that variable from 1 month and 2 months previously (continuous), and last by generating a binary feature indicating whether that environmental variable was higher or lower in the previous 1 or 2 months. Finally, any rows of the dataframe with missing data or "NAs" were removed, because machine learning methods do not support inputs with missing values.

Model Development

The spatiotemporal dataframe was randomly split into a training dataset, containing a subset of 70% of the data, a validation dataset of 15% of the data—which was used to refine the model—and a final test dataset using the remaining 15% of the data, which remained unseen during model development, to test final accuracy metrics. The number of data points these subsets for each model contained can be found in Table S3.

This training dataset was provided as an input into a random forest classification model using the scikit-learn (version 1.3.0; scikit-learn Developers) Python implementation,[37] with the binary epidemiological feature representing the presence of an infection report or not providing the target variable. Random forest classification models are an ensemble learning method using multiple decision trees that can iteratively split along gradients of environmental variables to generate a cumulative consensus.

Random forest models offer benefits over classical regression approaches such as generalized linear models (GLMs), such as used by Archer et al.,[19] by characterizing nonlinear associations. Each decision tree within the model can split the decision at any value of a predictor variable, allowing it to split at multiple points across the range of values for a predictor variable to account for nonlinear relationships. They can account for heterogeneous responses and interactions among predictor variables in different places and seasons, with different environmental contexts, without greatly increasing complexity or required informed parameterization of interactions, offering advantages as an alternative to nonlinear generalized additive models (GAMs). This characteristic provides particular value for pathogens where these relationships are not yet fully characterized and for applicable public health forecasting models that may need to be redeployed in regions with limited information. They have previously been identified as an effective, transparent model choice for complex climate-disease applications.[21,38] In addition, decision tree–based models perform well with imbalanced datasets, taking both classes into consideration at each stage when generating conditions and rules for splitting. Last, the choice of model was driven by the ability to interrogate the predictions from a random forest classification model, in comparison with black-box approaches that can provide high accuracy with complex data, at the expense of being noninterrogatable, such as neural networks.

The model setup uses bootstrap aggregating, in which the trees ($n = 100$ iterations) were trained on different bootstrapped samples of the data, a benefit of random forest classification models because it facilitates capturing the data structure more robustly and reduces variance. Model refinement was based on improving the following comprehensive accuracy metrics which were calculated for each model: AUC (area under the curve) of the ROC (receiver operating characteristic) curve, which represents model capability to distinguish between classes; sensitivity, which is affected by the false negative rate; specificity, which is affected by the false positive rate; F1 score, which is a harmonic mean between specificity and sensitivity commonly used as a comparative metric in machine learning studies; and overall accuracy calculated by the percentage of data points correctly classified.

The target variable, with a ratio of 1:9 infection months in comparison with noninfection months, was imbalanced. Imbalanced data can result in challenges in machine learning studies, because most algorithms are developed on the assumption that the target variable classes are balanced; however, real-world data is rarely perfectly balanced, particularly in disease detection applications. This frequently occurring imbalance can result in the introduction of bias in which high model accuracy can be achieved purely by predicting the majority class for each data point. Two techniques were therefore explored to address this class imbalance. First, oversampling of the minority class (infection months) to a 1:1 ratio was achieved using the synthetic minority oversampling technique (SMOTE), which generates new instances of the minority class using a k–nearest neighbor approach, using distance metrics to define new points.[39] Second, undersampling of the majority class (noninfection months) to a 1:1 ratio was achieved using Cluster Centroids—a prototype generation approach in which the majority class is synthesized into centroids using a k-means clustering approach to reduce the number of samples.[40] These two methods were implemented in distinct models using the imbalanced-learn (version 0.11.0) Python toolbox[40] within scikit-learn version 1.3.0,[37] with accuracy metrics calculated on oversampled and undersampled test datasets, respectively. The trained oversampled and undersampled models were then applied to real-world imbalanced test data (with no over- or undersampling) to explore the detection accuracy on the more representative, imbalanced number of infections reported.

The final models had 100 estimator trees in total, with the maximum number of features to identify optimum splits set as the square root of the total number of features. During the model refinement period, feature importance was calculated using two methods. First, the feature importances attribute provided within the random forest implementation in scikit-learn (version 1.3.0)[37] was extracted. This attribute computes importance based on Gini importance—the impurity (randomness) decrease across decision tree nodes when that specific feature is used to split the data. Second, permutation importance was calculated, which individually randomly shuffles each feature and refits the model to estimate the accuracy decrease when this feature is randomized. In addition, collinearity between the six environmental variables was estimated by calculating the Spearman's rank correlation coefficient between each variable (Figure S1). This, alongside estimates of Gini importance, informed the removal and retainment of particular features to reduce overfitting and provide an objective feature selection step. The full list of environmental features tested, with the retained features in each model highlighted, can be found in Table S2. The role of each environmental variable in the model predictions was further interrogated using partial dependence plots, which average over the values of all other input features to explore the expected prediction as a function of a specific environmental feature—and additionally as a function of two interacting features.

Further models were also generated throughout the development process to explore different hypotheses. This step included both a regional and state-level model for the Gulf Coast and the state of Florida (the region and county reporting the most Vv infections) to test which spatial resolution is most appropriate for model development amid distinct environmental ranges across the United States (Figure S2). The model implementation for these regional models was entirely the same as the national model, for an accurate comparison, but trained on data from the Gulf Coast and Florida only. In addition, a model was developed on a different range of features to explore the feasibility of early warning systems; the 1- and 2-month lagged environmental features were exclusively used to test predictive accuracy for the current month, using data collected from the previous months. These model parameters were set to the same as in the original model; however, fewer features were used for training and testing, omitting environmental data for the present month, ensuring the number of features was consistent between training and test data.

Results

Of the final Vv infections dataset provided to the model, containing 289 infection months, the most infections per state were reported in Florida (104 total), per month in July (76), and per year in 2015 (151), with full descriptive statistics found in Table S4. At a county level, the maximum number of infections reported in a county per month was 5 (mean of 0.14) and per year was 14 (mean of 1.7), both reported in Orleans county in Louisiana in July 2014 and 2017 and for the year in 2015, respectively.

Model Accuracy

The balanced random forest classifier models were able to retrospectively predict Vv infection months when trained on environmental data and applied to an unseen test subset of the dataframe. The accuracy of both the oversampled and undersampled approaches to create balanced models reveal particular strengths and weaknesses (Table 2). Both balanced models had an ROC AUC score above 0.98, which can be considered an outstanding discrimination[41] between months in which we would expect Vv infections or not. The sensitivity of the oversampled and undersampled models was 0.971 and 0.938, respectively, correctly predicting 333 of 343 infections and 45 of 48 infections, respectively, with notably different quantities of infections present in the test data using the two different techniques. Although both these metrics were higher for the oversampled model, specificity was higher for the undersampled model (0.957 vs. 0.888) which had a lower percentage of false positives, resulting in a higher F1 score (0.947 vs. 0.928) and overall accuracy (0.94 vs. 0.922).

However, when the models trained on balanced data were applied to the realistic imbalanced test dataset, with relatively lower number of infection reports, performance declined. Although reported sensitivity was still high (0.857–0.918), the largest differences were apparent in terms of specificity (0.158–0.417), in which a higher false positive rate was reported. Applying the model trained on undersampled data to the real imbalanced test data had a particularly poor specificity (0.158), despite a high sensitivity. The model trained on oversampled data was still able to produce a strong ROC AUC score of 0.94 when applied to the real imbalanced data (Figure 2), a 0.286 improvement in comparison with the model trained on undersampled data. There was a large

Table 2. Model development accuracy metrics for random forest classification models predicting Vv infections in the United States, 2008–2018. Comparison of accuracy metrics achieved for oversampling and undersampling models, tested on oversampled and undersampled test data, respectively, and then tested on real imbalanced test data.

Model	Metric	Tested on oversampled data (343 Vv reports total)	Tested on undersampled data (48 Vv reports total)	Tested on real imbalanced data
Trained on oversampled data	ROC AUC	0.984	—	0.94
	Sensitivity	0.971	—	0.857
	Specificity	0.888	—	0.417
	F1 Score	0.928	—	0.561
	Accuracy	0.922	—	0.869
Trained on undersampled data	ROC AUC	—	0.98	0.654
	Sensitivity	—	0.938	0.918
	Specificity	—	0.957	0.158
	F1 Score	—	0.947	0.27
	Accuracy	—	0.94	0.354

Note: —, no data; AUC, area under the curve; ROC, receiver operating characteristic; Vv, *Vibrio vulnificus*.

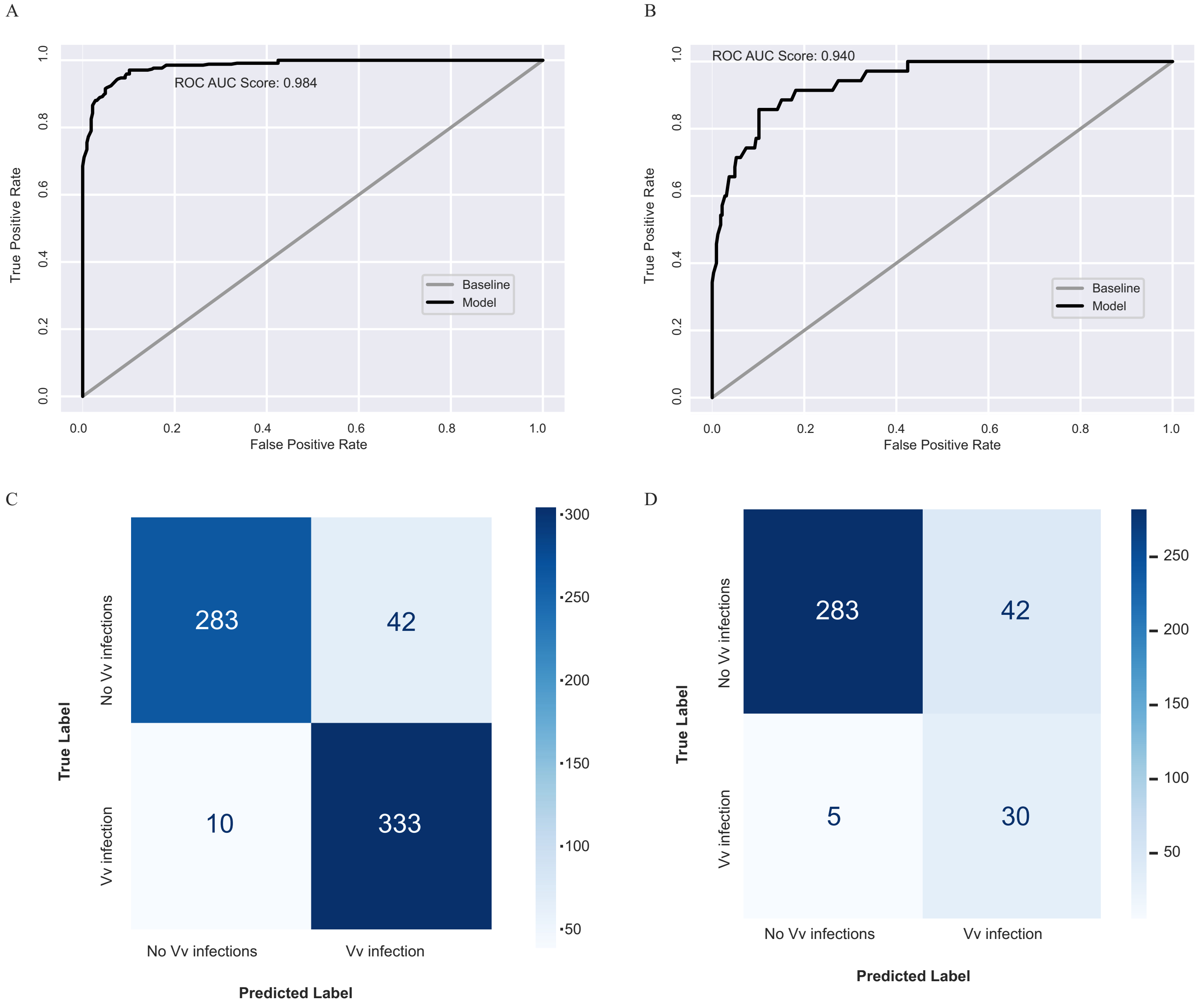


Figure 2. Comparison of results of selected random forest classification models predicting *Vibrio vulnificus* (Vv) infections in the United States, 2008–2018 (trained and validated on oversampled data). The plots on the left represent when the model was tested on oversampled test dataset containing 343 Vv infection months, and the plots on the right represent when the model was tested on the unprocessed version of that test dataset, representing imbalanced real-world data and containing 35 Vv infection months, via receiver operating characteristic curves (A,B) and confusion matrices comparing observed and predicted Vv infection numbers (C,D).

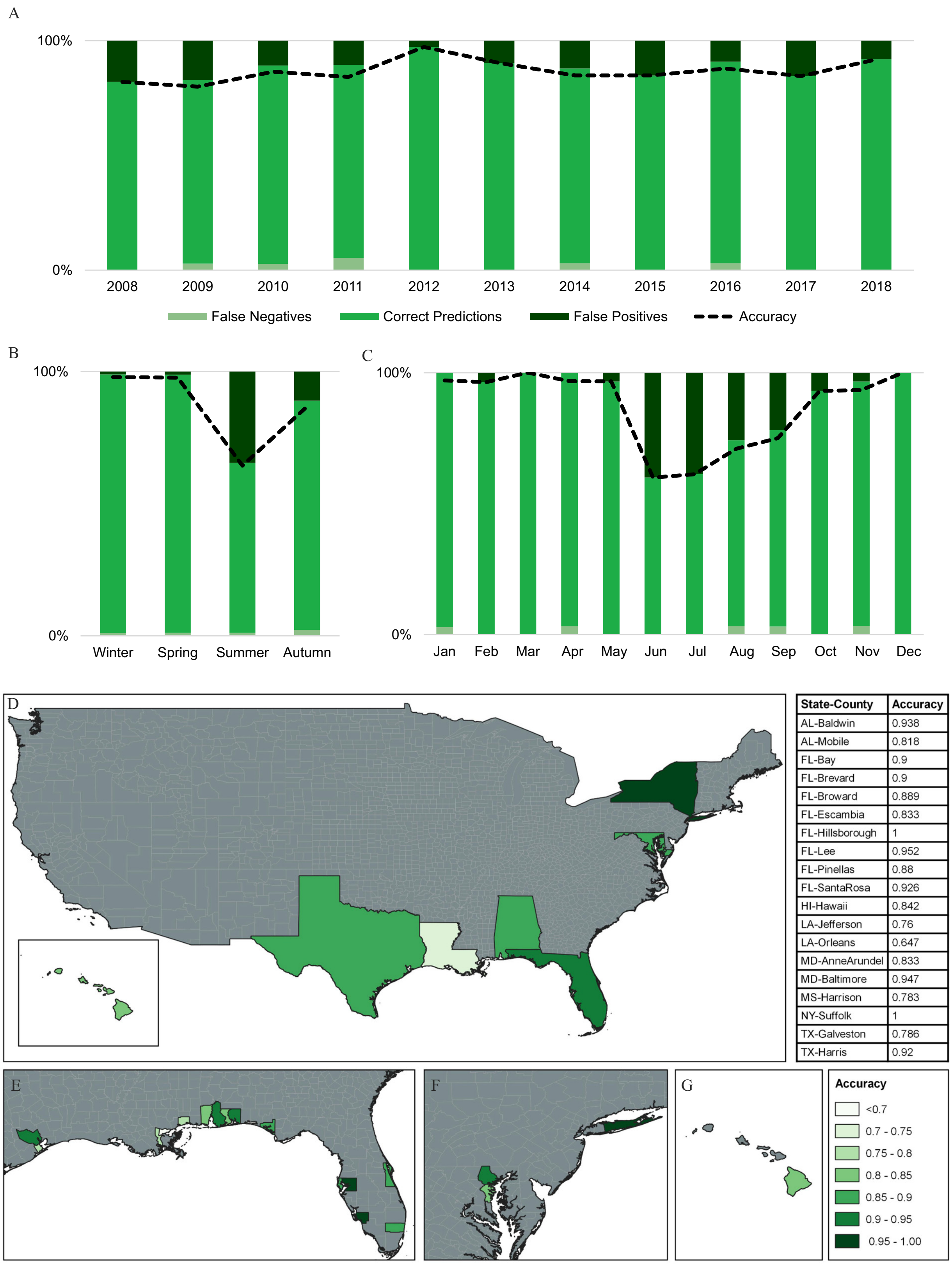
difference in overall accuracy, with the oversampled model reporting 0.869 but the undersampled model dropping to 0.354. This disparity, along with the generally more balanced result between sensitivity and specificity, drove the choice to limit the analysis to the oversampled model (and its application to both oversampled test data and real imbalanced data) hereafter.

Error Analysis

To understand the performance of the balanced model applied to the real imbalanced data, the spatial and temporal trends of errors were analyzed (Figure 3). Temporally, there were no clear interannual trends of model errors; however, a more seasonal pattern emerged in which the summer season, and particularly the months of June to September, were associated with greater numbers of false positives. Spatially, across counties, the accuracy ranged from 0.647 to 1, with an average of 0.871. Lower accuracy was found in Gulf Coast states, such as Louisiana. Notably, two counties had a perfect accuracy, with every data point from that county being correctly predicted: Hillsborough in Florida and Suffolk in New York. Errors were also found to be more prevalent for data points in the less-frequent upper ranges of various climate variables (Supplementary Figure S3).

Regional Approach

Such spatial differences in error distribution could be affected by the spatial resolution of the model development. The balanced model was developed for all three regions across the United States but contained a region-encoding categorical variable to allow the model to split the decision pathway between regions if necessary. Accuracy across the three regions in the full US model was not wide-ranging; between the Gulf Coast and Atlantic Coast, sensitivity values ranged from 0.976 to 0.979, F1 Score from 0.927 to 0.958, and accuracy from 0.920 to 0.957. A larger difference was found for specificity, with 0.883 and 0.940 reported for



State-County	Accuracy
AL-Baldwin	0.938
AL-Mobile	0.818
FL-Bay	0.9
FL-Brevard	0.9
FL-Broward	0.889
FL-Escambia	0.833
FL-Hillsborough	1
FL-Lee	0.952
FL-Pinellas	0.88
FL-SantaRosa	0.926
HI-Hawaii	0.842
LA-Jefferson	0.76
LA-Orleans	0.647
MD-AnneArundel	0.833
MD-Baltimore	0.947
MS-Harrison	0.783
NY-Suffolk	1
TX-Galveston	0.786
TX-Harris	0.92

Figure 3. Spatiotemporal errors of predictions of random forest classification model predicting *Vibrio vulnificus* (Vv) infections in the United States, 2008–2018, trained on oversampled data on unseen real-world test dataset, which contained 376 data points. Temporal errors are classified by (A) year, (B) season, and (C) month, and spatial errors averaged across state (D) and county for the Gulf Coast, Atlantic Coast, and Hawaii (E, F, and G, respectively). Full numeric data can be found in Supplementary Excel File 1. Administrative area shapefiles provided by Global Administrative Areas database (https://gadm.org/download_country.html).

Table 3. Comparison of spatial scale accuracy for random forest classification models predicting Vv infections in the United States, 2008–2018. Metrics for the Gulf Coast region and Florida, respectively, were extracted from the test results of the broadly applicable balanced model produced for entire United States and then compared to test metrics when region/state-specific models were trained and tested on only the Gulf Coast and Florida, respectively.

	Regional approach: Gulf Coast		State approach: Florida	
Metric	Extracted Gulf Coast results from full US model	Gulf Coast–specific model	Extracted Florida results from full US model	Florida-specific model
Sensitivity	0.976	0.967	0.727	0.935
Specificity	0.883	0.874	0.400	0.947
F1 Score	0.927	0.918	0.516	0.941
Accuracy	0.920	0.910	0.906	0.938

Note: Vv, Vibrio vulnificus.

the Gulf Coast and Atlantic Coast, respectively. Hawaii had a lower accuracy of 0.852; however, Hawaii had many fewer data points in the test dataframe (only 9 infection months).

When subsets of the test predictions for the Gulf Coast and Florida data points were compared to models developed specifically for the region or state, no improvements were found for a regional approach developed only on Gulf Coast data. However, the state approach using a model developed for Florida specifically yielded greater accuracy than the results for Florida counties in the original model (Table 3). Sensitivity was improved by 0.208, and specificity was notably improved by 0.547, largely reducing the false positive rate.

Early Warning System Potential

The model developed exclusively using lagged data, to simulate an early warning system, achieved high accuracy on the unseen test data. With an ROC AUC score of 0.968, similarly indicating outstanding discrimination,[41] the model had high accuracy (0.901) and sensitivity (0.965), with a slightly lower specificity (0.860).

Feature Importance

The model interrogations revealed the environmental features that had the largest contributions to the Vv infection predictions (Supplementary Figure S4). In terms of Gini importance, the top features were found to be the average sea surface temperature during the month of infection, the average land surface temperature of the month previous, and the categorical season encoder. Alongside sea and land temperatures, salinity and chlorophyll *a* concentration of the infection month also featured within the top ten feature importance. However, in terms of permutation importance, the binary lags (based on whether the value was higher or lower than the value in the previous month) of both runoff and salinity were ranked highest, followed by sea temperatures and chlorophyll *a* concentration of the month of infection.

For the early warning system model, using exclusively lagged data, the top features were similarly land and sea temperatures from the month previous, as well as the month variable. When permutation importance was calculated, the lagged chlorophyll *a* concentrations (for both 1 and 2 months previous to the infection) were found to be strong contributors to the predictions.

Environmental Associations

Partial dependence plots facilitated the interrogation of nonlinear associations between environmental variables and predicted Vv infections (Figure 4A), by presenting the relationships as a function of the average prediction (between binary variables of 0 and 1, this resulted in an averaged prediction range plotted for between 0.4 and 0.575, with higher values indicating higher probability of Vv infection prediction). Although some of these relationships followed generally positive linear patterns (such as for sea surface temperature, land surface temperature, and sea level anomalies), the random forest approach introduced step changes and thresholds for certain variables. For example, the prediction probability starts increasing more rapidly when land temperature exceeded 285 K ($\sim 11.9^{\text{o}}$C). Salinity exhibited a quadratic relationship, with prediction probability dropping between 34 and 35 practical salinity units (psu) and then again above 36 psu. An increase of Vv prediction probability was associated with a small increase in runoff, beyond which there was no further increase. Chlorophyll *a* concentration exhibited a steady positive relationship that stabilized between 3 and 8 mg/m^3.

In addition, we visualized how these environmental conditions interact to affect Vv infection probability (Figure 4B), which revealed certain limiting factors or cascading effects in a complex, interacting marine environment. For example, the increased prediction probability associated with increased sea surface temperature was limited when salinity exceeded 35 psu. sea level anomalies and chlorophyll *a* concentrations were found to have complimentary impacts on prediction probability, with sea level anomalies over 0.1 m and chlorophyll *a* concentrations above 10 mg/m^3 both needed to reach the highest prediction likelihood zone. Only a small runoff value was required to reach the highest predictive zone, with the land surface temperature signal being a stronger driver of infection prediction potential, which corroborated the relationships found in the individual feature partial dependence plots.

Transmission Routes

The models were developed using Vv infection data from both foodborne and nonfoodborne infections (such as wound infections), the majority of which were confirmed, but 4% of the transmission routes were labeled as "probable." There were a much larger number of nonfoodborne infection instances in the dataframe than foodborne (82% to 18% respectively). Despite being associated with distinct environmental variable ranges (Figure S5A), such as foodborne infections extending into cooler months with lower sea surface temperatures, there was no difference in prediction accuracy for the different exposures, with 85.2% of nonfoodborne infections detected correctly and 87.5% of foodborne infections detected correctly (Figure S5B). Although there seemed to be a positive association between how many foodborne infections a state reported and the number of false positives ($R = 0.763$) in the model predictions (Figure S5C), only six states reported both of these, so the significance of such a relationship is unclear.

Discussion

Machine Learning Performance

The results highlight the potential of machine learning approaches for forecasting Vv infections when there is a suitable surveillance system available for model training and validation. The machine learning approach allowed us to specifically explore complex, nonlinear, and previously uncharacterised environmental associations omitted by alternative models that require informed parameterization.[42]

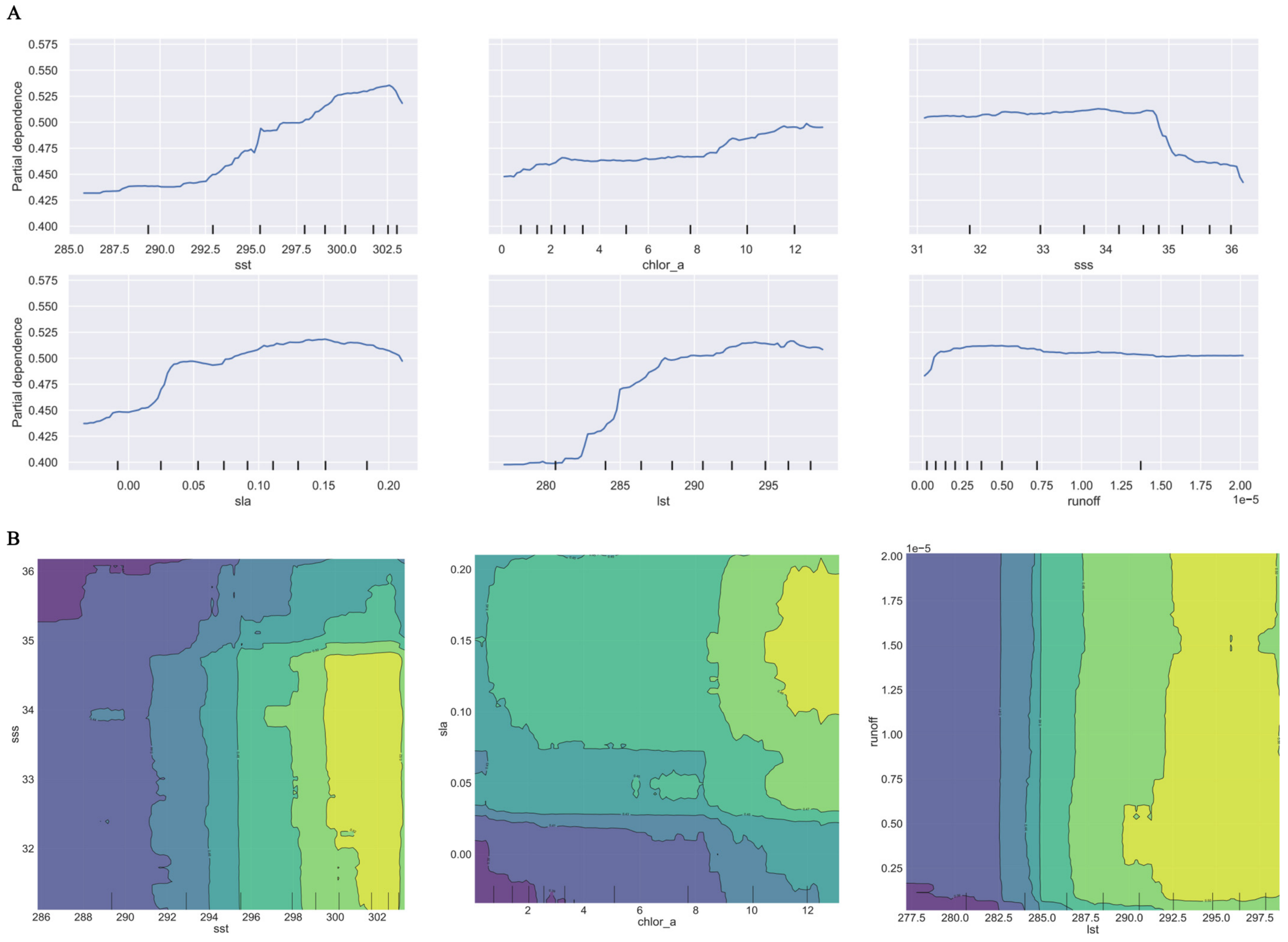


Figure 4. Role of environmental features in random forest classification model prediction of *Vibrio vulnificus* (Vv) infections in the United States, 2008–2018. (A) Partial dependence plots showing function of association between each environmental variable and the probability of an infection being predicted (on a 0–1 scale, where 1 indicates a certain infection and 0 a certain noninfection month). (B) Partial dependence plots visualizing effect of the interactions between environmental features on prediction (with lighter colors representing higher chance of an infection being predicted) for salinity and sea temperature, sea level anomalies and chlorophyll *a* concentration, and land surface temperature and runoff, respectively. Note: chlor_a, chlorophyll *a* concentration in mg/m^3; lst, land surface temperature in Kelvin; psu, practical salinity unit; sla, sea level anomaly in meters; runoff, runoff rate in $kg/m^2/s$; sss, salinity in psu; sst, sea surface temperature in Kelvin.

A specific consideration for the use of machine learning methods in this application is dealing with imbalanced datasets, which can lead to bias and overfitting. This study tested two methods to overcome such limitations, which were able to improve predictive potential. However, keeping realistic surveillance situations in mind, these balanced models were applied to imbalanced test datasets with the most notable consequence being a high false positive rate. This poor specificity is a major limitation of our model and would require improvement to become operational among the imbalanced reporting data to reduce the number of costly false alarms. The high number of false positives could be consequentially related to underreported *Vibrio* spp. infections; although the severity of Vv infections means they are less underreported than other *Vibrio* spp. infections, modeling by Scallan et al.[43] suggests the number of reported cases must be multiplied by 1.1 (to account for underreporting) and then 1.7 (to account for misdiagnosis) to calculate true Vv incidence. Applying these multipliers to the number of reported infections in the imbalanced real-world test data would suggest there could be up to 56 (rather than 30) infections actually reported in this period, which becomes closer to the 72 infections predicted by the model. However, there is no way of confirming whether any of the false positives predicted by the model are representative of missed cases, because this is our only available validation dataset, albeit with potential limitations itself. Although no clear spatiotemporal trends were identified for these false positives, we hypothesize the model may be overestimating infections due to optimum conditions existing on spatiotemporal occasions where humans are not present or interacting with the marine environment, because this part of the transmission pathway is not fully characterized in our predictor variables, despite the inclusion of some passive environmental drivers of human activity. The integration of population data and sociodemographic information, such as used in Archer et al.,[19] would specifically focus the predictions on populations at risk of infection, which could in turn reduce false positives in less-populated areas. The lack of available behavioral datasets to characterize this exposure fully is discussed below.

Although developed for the entire United States, our analysis found that state-specific models can improve specificity; however, we found little benefit in regional models. We therefore identified a trade-off between applicability and accuracy in which a lower spatial resolution model can offer greater applicability, but a higher resolution model can offer greater accuracy (however, the midrange point appears to lose the benefits of both). Our state-specific model was developed for Florida, which has two distinct coastlines with a wide range of environmental values, so models developed for states with a single, more homogenous coastline could have even higher specificity. An alternative further development option would be the inclusion of state-specific categorical encoders instead of regional encoders. State-specific models would be valuable in specific application scenarios, such as in states more heavily dependent on shellfish aquaculture, particularly in the governing structure of health departments in the United States, where each state can implement specific policies and programs.

Environmental Drivers of Vv Infections

Machine learning offers the potential to extract complex, nonlinear, and interacting relationships between environmental variables and Vv infections. The choice of a random forest classifier model, in comparison with more black-box approaches, additionally facilitated the interrogation of the contribution of each environmental variable to the model predictions of Vv infections.

All six environmental variables featured in the top feature importances, justifying the multivariate approach. Of these, some are well-characterized drivers of Vv epidemiology, such as temperature and salinity, but the less-reported use of chlorophyll *a* concentration and runoff data that reported high permutation importance should secure their inclusion in future Vv models. Specifically coastal runoff rate has previously been associated with diluting salinity and nutrient loading of coastal ecosystems, which can promote both pathogen and plankton population growth,[44] offering ecological insights into the reported association between extreme weather events and Vv infections.[7] The binary lagged values (a simple measure of whether the environmental variable was higher the month previous or not) for runoff and salinity held the highest permutation importance in the model, highlighting the importance of short-term perturbations, such as those observed during extreme weather events, as opposed to more gradual change. In addition, other environmental drivers were found to have lagged effects on Vv infection potential, particularly values of the month previous to the infection. Formal lagged time-series analysis, such as distributed nonlinear lag analysis,[45] could be used in the future as a more supervised technique to identify appropriate exposure–lag–response features as model inputs.

Alongside confirming well-established relationships, including the quadratic relationship exhibited with salinity,[6] the model predictions provided novel insights into various nonlinear threshold and step-based relationships and specifically the values at which these occur. Salinity was similarly found to be a governing factor interacting with sea temperature in Martinez-Urtaza et al.,[46] based on *Vibrio parahaemolyticus* presence in Spanish estuaries. The salinity values found here are higher than to be expected for Vv optimums (2–25 psu),[16] due to observations being measured at 5 m depth. It is difficult to measure coastal salinity from satellites due to high-resolution heterogeneity, so this approach provides a proxy for relative coastal salinity changes at the surface and nearer the coast. In addition, the cascading impact of sea level anomalies and chlorophyll *a* presence was identified, providing evidence for the impact of coastal water intrusions inland coinciding with high plankton presence as a high risk factor for Vv infections. Such relationships have been hypothesized,[9] because such mechanisms facilitate the transport of Vv bacteria in coastal waters, but this analysis quantifies this in real-world scenarios and highlights the predictive potential of exploring both variables in combination. Interrogating the environmental ranges of the predictions allows the establishment of prediction zones, ranges of environmental variables in which Vv infections are more likely (Figure S6).

Although some studies may remove variables from models based on linear collinearity, retaining these features allowed the exploration of nonlinear interactions between these. The strongest collinearity was observed between sea surface temperature and land surface temperature; however, these represent two separate sections of the transmission pathway, with the former increasing suitability in the environment and the latter related more to human exposure, with increased recreational water usage expected during warm temperatures.[1]

Applications and Future Directions

The analysis here explored the feasibility of early warning systems, using lagged environmental features from 1 and 2 months previous. If environmental data were routinely collected on a monthly basis, such collection would facilitate near real-time monitoring of Vv risk, which could inform mitigation measures such as enhanced hygiene measures for shellfish harvesting[13] or coastguard signposting to discourage recreational water usage during high-risk conditions. Longer-term lagged effects could be explored to increase the lead-time range of such warning systems. To test the applicability and validity of the predictions, the model should be implemented prospectively for a specific time period and then evaluated against the observed incidence of Vv infections over that period. Collaborations with specific health departments or coastal authorities would be required to establish the level of accuracy and specific resolution required for operationalization of such models.

To apply these models to wider climate change scenarios, forecast data of the environmental variables within the model would be necessary. Although forecast data for certain variables such as temperature (such as used in recent Vv forecasts[19]) are readily available, there is limited access to forecast data of chlorophyll *a* concentrations and runoff, for example, which proved key variables in the model here. An alternative option involves hypothetical climate modeling for particular environmental values or gradual change scenarios, where forecast data are not available, or simpler models with fewer variables to sacrifice model accuracy in the interest of forecast operationality. In addition, there is a need to explore the effects of extreme weather events on Vv, which would require a test (or real-world) dataset with a higher temporal resolution to capture shorter-term environmental variability, such as weekly or daily data. Training a model on this data would likely be infeasible, with such high temporal resolution creating an even larger imbalance of noninfection data points; however, this does not prohibit a monthly model being applied to higher temporal resolution test data. Such rare infection occurrences will likely result in a low detection rate, and trends of errors at higher ranges of variables found in this analysis will likely be exacerbated further by extreme environmental values found in daily and weekly data that were averaged over in the monthly analysis.

This model would be applicable to other *Vibrio* species and possibly other waterborne bacteria with environmental transmission routes. The lack of notable difference between accuracy of the foodborne and nonfoodborne infections in the model is promising for the representation of multiple complex transmission routes for marine bacteria like Vv. In terms of application to other locations, a similar surveillance dataset to the COVIS database would be required ideally for training, validation, and testing of models. However, if this is not possible, the model could still be hypothetically applied to a region without epidemiological data to provide hypothetical estimates of risk, but these could not be validated to assess accuracy. Such hypothetical risk models could be used as a pilot, exploratory step to encourage future surveillance of *Vibrio* spp. infections.

Some key complexities for forecasting Vv infections include the underreporting of epidemiological data and spatial inaccuracies when infections are reported in different areas from the environmental exposure, which can introduce noise and model error. Although travel-associated cases were removed from the analyses to explore local exposures, foodborne infections could be acquired from shellfish that has been imported from a different county or state. Foodborne infection data particularly might therefore need further preprocessing in the future to identify the environmental origin of Vv. In terms of limitations of environmental variables, pH data were considered for this study, due to previously reported associations[47]; however, readily available datasets covering the entire period did not have a sufficient coastal resolution for the county-level analysis.[34] In addition, coastal resolution is pertinent to several climate variables used in this study, such as the steep salinity gradients found in coastal areas, which are difficult to characterize using Earth Observation data. Coastal areas are often poorly characterized in marine datasets due to their heterogeneity, despite representing the key interface between humans and waterborne bacteria, requiring improvement to coastal observations for Vv forecasting applications.

The model assumes that all Vv populations are equally pathogenic and respond homogenously to environmental change, despite our knowledge of intraspecific heterogeneity among *Vibrio spp*. strains, due to the limited taxonomic resolution of the clinical samples that constitute the COVIS database. Similarly, the model makes major assumptions of the exposure routes, assuming optimum conditions for Vv and human water usage will lead to exposure, with no consideration for heterogeneous behaviors across seasons, regions, or human populations, or the presence of susceptible populations. The model does not contain behavioral components that can accurately represent human exposure to the pathogen, which is a critical component of the transmission pathway, thus reducing model performance. Characterizing these two aspects would negate the need for assumptions; thus two further areas of development to overcome model limitations would be to characterize human and genetic components within the model.

First, human recreational water usage data would enable the characterization of human exposure routes driving nonfoodborne infections; however, no such dataset exists for the United States. To overcome this data limitation, we recommend future collaborations with organizations, such as the US Coast Guard, who could potentially provide such data for a particular beach for a high-resolution specific study that may aid the parametrization of behavioral factors that drive the transmission process in future models. Exploring the presence of vulnerable populations, for example shellfish harvesters who both work in coastal areas and are likely to contract open wounds from sharp bivalves, would offer an alternative route to characterize potential transmission; however, such data are currently only available as a static variable per state.[48] Local scale studies specific to harvest areas could aim to characterize how this variable would affect predictions, to allow extrapolation for national studies. In terms of foodborne exposure routes, databases such as the United Nations Food and Agriculture Organization's FishStatJ provide shellfish consumption data[49] but at a maximum resolution of annually and per-country, which would be difficult to reconcile within this higher resolution model. Such data characterizing the human dimension of Vv infection routes would provide a greater characterization of *Vibrio* spp. transmission pathways to improve accuracy of predictive Vv models. However, although individual behavior can drive exposure, the environmental drivers used within the model are still critical for setting the background of risk potential on which these interactions operate and thus still prove valuable for forecasting models in public health applications, which as an intervention method can dissuade people from such interactions when the infection risk is high.

Second, a future direction would be incorporating genetic information of the reported Vv infections to explore strain-specific responses to environmental variables and identify conditions conducive to more virulent strains. Vv specifically is subdivided into three biotypes[50]; however, isolate information in the COVIS database is currently summarized at a species level rather than a strain level. Previous predictive models have been hindered by limited genomic information, such as a *Vibrio parahaemolyticus* outbreak in British Colombia in 2015, hypothesized to be attributed to the introduction of a new strain that could not be detected by models that functioned only at a species level.[51] Previous studies have identified strain-specific ecological tolerances of *Vibrio* bacteria,[52] and machine learning offers further opportunities to explore strain-specific environmental ranges by accounting for diverging pathways of environmental responses. Future developments of the COVIS database, to additionally provide genomic information associated with reported infections, would therefore open up new analysis opportunities.

Ultimately, this study has found machine learning offers potential in both forecasting Vv infections and characterizing the nonlinear and interacting environmental associations driving them. These results highlight the value of robust surveillance systems and interdisciplinary synthesis of epidemiological and environmental data for developing models and forecasts of environmentally associated waterborne diseases.

Acknowledgments

A.M.C. was supported by the Natural Environment Research Council (grant number NE/S007210/1) and Centre for Environment, Fisheries and Aquaculture Science (CEFAS) internal Seedcorn funding. J.M.U. was funded by Grants PID2021-127107NB-I00 from the Ministerio de Ciencia e Innovación (Spain), 2021 SGR 00526 from Generalitat de Catalunya and European Union's Horizon Europe research and innovation program under Grant Agreement No. 101057554 (IDAlert).

References

1. Froelich BA, Daines DA. 2020. In hot water: effects of climate change on *Vibrio*–human interactions. Environ Microbiol 22(10):4101–4111, PMID: 32114705, https://doi.org/10.1111/1462-2920.14967.
2. Vezzulli L, Pruzzo C, Huq A, Colwell RR. 2010. Environmental reservoirs of *Vibrio cholerae* and their role in cholera. Environ Microbiol Rep 2(1):27–33, PMID: 23765995, https://doi.org/10.1111/j.1758-2229.2009.00128.x.
3. Baker-Austin C, Trinanes JA, Taylor NG, Hartnell R, Siitonen A, Martinez-Urtaza J. 2013. Emerging *Vibrio* risk at high latitudes in response to ocean warming. Nat Clim Chang 3(1):73–77, https://doi.org/10.1038/nclimate1628.
4. Baker-Austin C, Oliver JD. 2018. *Vibrio vulnificus*: new insights into a deadly opportunistic pathogen. Environ Microbiol 20(2):423–430, PMID: 29027375, https://doi.org/10.1111/1462-2920.13955.
5. Coerdt KM, Khachemoune A. 2021. *Vibrio vulnificus*: review of mild to life-threatening skin infections. Cutis 107(2):E12–E17, PMID: 33891847, https://doi.org/10.12788/cutis.0183.
6. Deeb R, Tufford D, Scott GI, Moore JG, Dow K. 2018. Impact of climate change on *Vibrio vulnificus* abundance and exposure risk. Estuaries Coast 41(8):2289–2303, PMID: 31263385, https://doi.org/10.1007/s12237-018-0424-5.
7. Brumfield KD, Usmani M, Santiago S, Singh K, Gangwar M, Hasan NA, et al. 2023. Genomic diversity of *Vibrio* spp. and metagenomic analysis of pathogens in Florida Gulf coastal waters following Hurricane Ian. mBio 14(6):e01476-23, PMID: 37931127, https://doi.org/10.1128/mbio.01476-23.
8. Turner JW, Malayil L, Guadagnoli D, Cole D, Lipp EK. 2014. Detection of *Vibrio parahaemolyticus*, *Vibrio vulnificus* and *Vibrio cholerae* with respect to seasonal fluctuations in temperature and plankton abundance. Environ Microbiol 16(4):1019–1028, PMID: 24024909, https://doi.org/10.1111/1462-2920.12246.
9. Lobitz B, Beck L, Huq A, Wood B, Fuchs G, Faruque ASG, et al. 2000. Climate and infectious disease: use of remote sensing for detection of *Vibrio cholerae* by indirect measurement. Proc Natl Acad Sci USA 97(4):1438–1443, PMID: 10677480, https://doi.org/10.1073/pnas.97.4.1438.
10. Cazorla C, Guigon A, Noel M, Quilici ML, Lacassin F. 2011. Fatal *Vibrio vulnificus* infection associated with eating raw oysters, New Caledonia. Emerg Infect Dis 17(1):136–137, PMID: 21192878, https://doi.org/10.3201/eid1701.100603.

11. US CDC (US Centers for Disease Control and Prevention). 2005. *Vibrio* illnesses after Hurricane Katrina–multiple states, August–September 2005. MMWR Morb Mortal Wkly Rep 54:928–931, PMID: 16177685.
12. Morantz CA. 2005. CDC reports on illnesses in Hurricane Katrina evacuees and relief workers. Am Fam Physician 72(10):2126–2134.
13. Martinez-Urtaza J, Bowers JC, Trinanes J, DePaola A. 2010. Climate anomalies and the increasing risk of *Vibrio parahaemolyticus* and *Vibrio vulnificus* illnesses. Food Res Int 43(7):1780–1790, https://doi.org/10.1016/j.foodres.2010.04.001.
14. Paz S, Bisharat N, Paz E, Kidar O, Cohen D. 2007. Climate change and the emergence of *Vibrio vulnificus* disease in Israel. Environ Res 103(3):390–396, PMID: 16949069, https://doi.org/10.1016/j.envres.2006.07.002.
15. Biasutti M, Sobel AH, Camargo SJ, Creyts TT. 2012. Projected changes in the physical climate of the Gulf Coast and Caribbean. Climatic Change 112(3–4):819–845, https://doi.org/10.1007/s10584-011-0254-y.
16. Baker-Austin C, Oliver JD, Alam M, Ali A, Waldor MK, Qadri F, et al. 2018. *Vibrio* spp. infections. Nat Rev Dis Primers 4(1):8–19, https://doi.org/10.1038/s41572-018-0005-8.
17. Semenza JC, Trinanes J, Lohr W, Sudre B, Löfdahl M, Martinez-Urtaza J, et al. 2017. Environmental suitability of *Vibrio* infections in a warming climate: an early warning system. Environ Health Perspect 125(10):107004, PMID: 29017986, https://doi.org/10.1289/EHP2198.
18. Newton A, Kendall M, Vugia DJ, Henao OL, Mahon BE. 2012. Increasing rates of vibriosis in the United States, 1996–2010: review of surveillance data from 2 systems. Clin Infect Dis 54 (0 5):S391–S395, PMID: 22572659, https://doi.org/10.1093/cid/cis243.
19. Archer EJ, Baker-Austin C, Osborn TJ, Jones NR, Martínez-Urtaza J, Trinanes J, et al. 2023. Climate warming and increasing *Vibrio vulnificus* infections in North America. Sci Rep 13(1):3893, PMID: 36959189, https://doi.org/10.1038/s41598-023-28247-2.
20. Ayala AJ, Kabengele K, Almagro-Moreno S, Ogbunugafor CB. 2023. Meteorological associations of *Vibrio vulnificus* clinical infections in tropical settings: correlations with air pressure, wind speed, and temperature. PLoS Negl Trop Dis 17(7):e0011461, PMID: 37410780, https://doi.org/10.1371/journal.pntd.0011461.
21. Campbell AM, Racault MF, Goult S, Laurenson A. 2020. Cholera risk: a machine learning approach applied to essential climate variables. Int J Environ Res Public Health 17(24):9378, PMID: 33333823, https://doi.org/10.3390/ijerph17249378.
22. Chowdhury FR, Nur Z, Hassan N, von Seidlein L, Dunachie S. 2017. Pandemics, pathogenicity and changing molecular epidemiology of cholera in the era of global warming. Ann Clin Microbiol Antimicrob 16(1):10, PMID: 28270154, https://doi.org/10.1186/s12941-017-0185-1.
23. Racault M-F, Abdulaziz A, George G, Menon N, C J, Punathil M, et al. 2019. Environmental reservoirs of *Vibrio cholerae*: challenges and opportunities for ocean-color remote sensing. Remote Sensing 11(23):2763, https://doi.org/10.3390/rs11232763.
24. Cai W-J, Xu Y-Y, Feely RA, Wanninkhof R, Jönsson B, Alin SR, et al. 2020. Controls on surface water carbonate chemistry along North American ocean margins. Nat Commun 11(1):2691, PMID: 32483136, https://doi.org/10.1038/s41467-020-16530-z.
25. Sedas VTP. 2007. Influence of environmental factors on the presence of *Vibrio cholerae* in the marine environment: a climate link. J Infect Dev Ctries 1(3):224–241, PMID: 19734600.
26. Elliott LR, White MP, Sarran C, Grellier J, Garrett JK, Scoccimarro E, et al. 2019. The effects of meteorological conditions and daylight on nature-based recreational physical activity in England. Urban For Urban Green 42:39–50, https://doi.org/10.1016/j.ufug.2019.05.005.
27. Baker-Austin C, Trinanes J, Martinez-Urtaza J. 2020. The new tools revolutionizing vibrio science. Environ Microbiol 22(10):4096–4100, PMID: 32419260, https://doi.org/10.1111/1462-2920.15083.
28. Copernicus Climate Change Service (C3S). 2019. Sea surface temperature daily data from 1981 to present derived from satellite observations. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), https://doi.org/10.24381/cds.cf608234.
29. Merchant CJ, Embury O, Bulgin CE, Block T, Corlett GK, Fiedler E, et al. 2019. Satellite-based time-series of sea-surface temperature since 1981 for climate applications. Sci Data 6(1):223, PMID: 31641133, https://doi.org/10.1038/s41597-019-0236-x.
30. Good SA, Martin MJ, Rayner NA. 2013. EN4: quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates. JGR Oceans 118(12):6704–6716, https://doi.org/10.1002/2013JC009067.
31. Gouretski V, Cheng L. 2020. Correction for systematic errors in the global dataset of temperature profiles from mechanical bathythermographs. J Atmos Ocean Technol 37(5):841–855, https://doi.org/10.1175/JTECH-D-19-0205.1.
32. Sathyendranath S, Brewin RJW, Brockmann C, Brotas V, Calton B, Chuprin A, et al. 2019. An ocean-colour time series for use in climate studies: the experience of the ocean-colour climate change initiative (OC-CCI). Sensors (Basel) 19(19):4285, PMID: 31623312, https://doi.org/10.3390/s19194285.
33. Gregor L, Gruber N. 2020. OceanSODA-ETHZ: a global gridded data set of the surface ocean carbonate system for seasonal to decadal studies of ocean acidification. Earth Syst Sci Data 13(2):777–808, https://doi.org/10.5194/essd-13-777-2021.
34. Copernicus Climate Change Service. 2018. Sea level gridded data from satellite observations for the global ocean from 1993 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), https://doi.org/10.24381/cds.4c328c78.
35. Jimenez C, Prigent C. 2023. ESA Land Surface Temperature Climate Change Initiative (LST_cci): All-weather MicroWave Land Surface Temperature (MW-LST) Global Data Record (1996–2020), v2.33. NERC EDS Centre for Environmental Data Analysis, 26 January 2023, https://doi.org/10.5285/a7e811fe11d34df5abac6f18c920bbeb.
36. Hersbach H, Bell B, Berrisford P, Biavati G, Horányi A, Muñoz Sabater J, et al. 2023. ERA5 monthly averaged data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), https://doi.org/10.24381/cds.f17050d7.
37. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. 2011. Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830, https://doi.org/10.48550/arXiv.1201.0490.
38. Fenderson LE, Kovach AI, Llamas B. 2020. Spatiotemporal landscape genetics: investigating ecology and evolution through space and time. Mol Ecol 29(2):218–246, PMID: 31758601, https://doi.org/10.1111/mec.15315.
39. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. 2002. SMOTE: synthetic minority over-sampling technique. JAIR 16:321–357, https://doi.org/10.1613/jair.953.
40. Lemaître G, Nogueira F, Aridas CK. 2017. Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res 18(17):1–5, https://doi.org/10.48550/arXiv.1609.06570.
41. Hosmer DW Jr, Lemeshow S, Sturdivant RX. 2013. *Applied Logistic Regression*. Hoboken, NJ: John Wiley & Sons.
42. Thuiller W. 2024. Ecological niche modelling. Curr Biol 34(6):R225–R229, PMID: 38531309, https://doi.org/10.1016/j.cub.2024.02.018.
43. Scallan E, Hoekstra RM, Angulo FJ, Tauxe RV, Widdowson MA, Roy SL, et al. 2011. Foodborne illness acquired in the United States—major pathogens. Emerg Infect Dis 17(1):7–15, PMID: 21192848, https://doi.org/10.3201/eid1701.p11101.
44. Thickman JD, Gobler CJ. 2017. The ability of algal organic matter and surface runoff to promote the abundance of pathogenic and non-pathogenic strains of *Vibrio parahaemolyticus* in Long Island Sound, USA. PLoS One 12(10):e0185994, PMID: 29020074, https://doi.org/10.1371/journal.pone.0185994.
45. Gasparrini A. 2014. Modeling exposure–lag–response associations with distributed lag non-linear models. Stat Med 33(5):881–899, PMID: 24027094, https://doi.org/10.1002/sim.5963.
46. Martinez-Urtaza J, Lozano-Leon A, Varela-Pet J, Trinanes J, Pazos Y, Garcia-Martin O. 2008. Environmental determinants of the occurrence and distribution of *Vibrio parahaemolyticus* in the rias of Galicia, Spain. Appl Environ Microbiol 74(1):265–274, PMID: 17981951, https://doi.org/10.1128/AEM.01307-07.
47. Velez KC, Leighton RE, Decho AW, Pinckney JL, Norman RS. 2023. Modeling pH and temperature effects as climatic hazards in *Vibrio vulnificus* and *Vibrio parahaemolyticus* planktonic growth and biofilm formation. GeoHealth 7(4): e2022GH000769, PMID: 37091291, https://doi.org/10.1029/2022GH000769.
48. Liddel M, Yencho M. 2022. Fisheries of the United States 2020, National Marine Fisheries Service. https://repository.library.noaa.gov/view/noaa/40953 [accessed 12 May 2024].
49. Fisheries and Aquaculture Department, Food and Agriculture Organization of the United Nations. 2017. *FishStatJ: A Tool for Fishery Statistics Analysis*. Rome, Italy: FAO.
50. Baker-Austin C, Lemm E, Hartnell R, Lowther J, Onley R, Amaro C, et al. 2012. pilF polymorphism-based real-time PCR to distinguish *Vibrio vulnificus* strains of human health relevance. Food Microbiol 30(1):17–23, PMID: 22265278, https://doi.org/10.1016/j.fm.2011.09.002.
51. Galanis E, Otterstatter M, Taylor M. 2020. Measuring the impact of sea surface temperature on the human incidence of *Vibrio sp.* infection in British Columbia, Canada, 1992–2017. Environ Health 19(1):1–7, https://doi.org/10.1186/s12940-020-00605-x.
52. Chiang ML, Chen HC, Wu C, Chen MJ. 2014. Effect of acid adaptation on the environmental stress tolerance of three strains of *Vibrio parahaemolyticus*. Foodborne Pathog Dis 11(4):287–294, PMID: 24410096, https://doi.org/10.1089/fpd.2013.1641.