UNIVERSITY OF CALIFORNIA
SANTA CRUZ

FORECASTING MARINE HEATWAVES IN THE NORTHEAST PACIFIC OCEAN: A COMPARATIVE ANALYSIS OF MACHINE LEARNING APPROACHES

A thesis submitted in partial satisfaction of the
requirements for the degree of

MASTER OF SCIENCE

in

EARTH SCIENCES

by

Courtney A. Stratton

June 2024

The Thesis of Courtney A. Stratton
is approved:

Professor Claudie Beaulieu, Chair

Professor Mathis Hain

Professor Christopher Edwards

Peter Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

Abstract

Forecasting Marine Heatwaves in the Northeast Pacific Ocean: A Comparative Analysis of Machine Learning Approaches

by

Courtney A. Stratton

Marine heatwaves (MHWs) are periods of abnormally high sea surface temperatures (SSTs) that persist for periods of time, causing adverse impacts of marine ecosystems and coastal communities. With projections that MHWs will become more frequent and severe, it is increasingly important to be able to predict MHW events to mitigate risks for ecosystems and communities. Here, we investigate the predictability of MHWs in the northeast Pacific Ocean by employing a suite of models, including logistic regression, naive Bayes, gradient boosting, random forest, and feedforward neural network, all of which are trained on selected oceanic and atmospheric variables. We find that the random forest model performs best at predicting the presence or absence of a MHW event at the 90$^{\text{th}}$ percentile MHW threshold using cluster centroid balanced data. The model is able to predict with accuracy ranging from 0.98 to 0.97 for leads spanning from 1 day to 2 weeks. While all models encounter difficulties in accurately categorizing MHWs, predicting their presence or absence remains a valuable metric for informing managers and industries about impending MHW events. Short-term forecasts can be especially advantageous in alerting industries and communities to these events, empowering them to implement adaptive measures against the detrimental impacts of MHWs.

Dedicated to my late father, Greg Stratton, and late grandmother, Mary Ann Semler. I am forever grateful for the endless love and support they gave me. I would not be who I am today without them.

Acknowledgments

I am humbled and very grateful for the the endless support from my family, friends and committee members. Without the support I received, none of this would have been possible.

First off, I am very grateful for my advisor, Claudie Beaulieu. Thank you Claudie for sparking my interest in data analysis all the way back in ESCI 160 during my time as an undergraduate at UCSC. Your class is what inspired me to continue my curiosity in data analysis and ultimately what lead me to persue a Master's at UCSC. Thank you for your continued encouragement, feedback, guidance and support throughout the years, during my time as both an undergraduate and graduate. I cannot thank you enough for all that you have done for me during my time at UCSC. I am also grateful to my committee members, Mathis Hain and Chris Edwards, for their availability, encouragement, feedback and support. My time at UCSC has been truly a joy in large part to my committee members enthusiasm and support.

Thank you to my Mom, Danna, and brother, Conner. With all of life's bumps and roadblocks, your love and support has never wavered. I am so grateful for the endless encouragement and to be a part of such a loving family. Thank you to all my friends, both near and far, for their unwavering support and endless laughter. Whether it's through cracking silly jokes or simply spending time together, I can always count on them to bring a smile to my face. Last but not least, thank you to my fiancé, Wilson, for your constant love and support, and for never loosing faith in me. I feel lucky everyday to be sharing life with you.

Lastly, I acknowledge the funding sources, both during my undergraduate and graduate studies at UCSC, that made this research possible. I acknowledge the Undergraduate Research in Science and Technology Grant, Richard Libby Scholarship, SURGES Award and partial funding from NSF award AGS-2143550.

Chapter 1

Introduction

Marine heatwaves (MHWs) are prolonged periods of abnormally warm sea surface temperatures (SSTs), often lasting from days to months [21]. Marine heatwave events have been linked to numerous ecosystem, biological, and economic disruptions in many regions across the globe [2, 14, 42, 41, 8]. Unusually high SSTs have been linked to irregular weather patterns, suppressed nutrient transports leading to low chlorophyll events and species range shift, as seen in the record high MHW event named the “Blob” in the northeast Pacific Ocean from 2013-2015 [2, 14]. This prolonged warm water anomaly in the northeast Pacific Ocean provided valuable insights into the repercussions of extreme weather events on temperature-sensitive marine ecosystems. During this event, the largest recorded outbreak of neurotoxins and domoic acid occurred along the North American west coast, impacting fisheries in the region resulting in extensive closures [34, 42]. The “Blob” was associated with heightened vertical stratification, inhibiting the vertical transport of nutrients and consequently triggering food web disruption [8]. Phytoplankton, crucial as the basis of the food web, experienced a decline, adversely affecting a multitude of marine species, including those at higher trophic levels like sea lions and baleen whales

[8]. Experiments have indicated that increased coastal nutrient runoff from human activities, combined with warmer ocean temperatures, could exacerbate the severity of toxic events in oceans [34]. Not only can harmful algal blooms impact fisheries, but increased SSTs can result in species range shift and decrease in fish biomass, with projections that fish populations of pacific cod, California anchovy and sockeye salmon will drastically decrease in the coming decades in the northeast Pacific Ocean [9].

While our focus lies on the northeast Pacific Ocean, it is essential to acknowledge the broader implications of MHWs across various oceanic regions and ecosystems. For example, the MHW that occurred off the coast of Western Australia from 2010-2011 caused numerous ecosystem, social and economic impacts including a decline in keystone species, such as seagrass and kelps, and loss of habitat, impacting the toursim industry and resulting in extensive fishery closures [36]. Extreme SSTs have also been linked to coral bleaching in many regions, which disrupts the ecosystem by decreasing ecosystem function, loss of habitat and a decline in biodiversity [2, 36]. With projections of more severe and frequent MHWs under anthropogenic climate change, coral reefs are expected to suffer a tough future as bleaching intensity and coral mortality continue to climb [14]. Among localized impacts, MHWs can influence global weather patterns. Sea surface temperatures play an important role in global weather patterns, “with phenomena such as El Niño-Southern Oscillation (ENSO) regarded as a major source of interannual climate variability at the global scale” [45]. El Niño-Southern Oscillation is known to regulate global SST “variability and the MHW frequency, duration, and intensity”, with global impacts [45].

Several oceanic and atmospheric phenomena are known to influence extreme SSTs on both a regional and global scale. Research utilizing an operational coupled ocean-atmosphere prediction system (ACCESS-S1) focused on forecasting MHWs on the Great Barrier Reef found that shortwave radiation, low cloud cover, and latent heat flux anomalies influenced MHW patterns [4]. In the northeast Pacific Ocean, it has been suggested that strong positive anomalies in sea level pressure and low net heat flux, partially driven by wind speeds, led to the warming event that occurred during the 2014-2015 winter [5]. In the Kuroshio-Oyashio Extension Region, a study identified key factors behind intense summer MHWs occurring in 1999, 2008, 2012, and 2016 [13]. These events were primarily influenced by air-sea heat flux anomalies and reduced cloud cover, while regional factors such as the strengthened North Pacific High system and the Philippine-Japan teleconnection also played roles [13]. Another study explored the potential causes of anomalous events in the Indian Ocean Dipole (IOD), which can stem from teleconnections with the equatorial Pacific, including evolving El Niño events, as well as cooling phenomena along the Australian coast [15]. These findings underscore the complex interplay of oceanic and atmospheric dynamics in shaping extreme SST events.

Recent studies have advanced our understanding of MHW forecasting, revealing that numerical multimodel ensemble and machine learning approaches offer promising accuracy in long-range predictions spanning from 1 month to 1 year. However, these approaches face limitations in short-term forecasting below 1 month. A study forecasting MHWs using a multimodel ensemble has shown that MHWs can be forecasted accurately using monthly data with leads spanning from 1 month to 1 year [23]. The accuracy of the model is highly dependent on the region, season and the regime of large-scale climate modes and cannot forecast at time scales smaller than 1 month [23]. Prior forecasting based on machine learning, using a deep learning time series prediction model (Unet-LSTM), has proven skillful in long-range forecasting for up to 18 months [45]. Both cases of MHW forecasting based on numerical models or machine learning discussed above focus on long-term forecasting using monthly resolution data at a global scale, leaving a gap for forecasting MHWs on short-term timescales with daily resolution data [23, 45].

Studies focusing on short-term forecasts at both the global and regional level have shown promising results in predicting MHWs. Research using a deep learning convolutional neural network (CNN) model demonstrated good accuracy on the global scale with a 2 week lead time, with forecast improvements of 10% when combined with a physical forecast model [43]. Additional research using the ACCESS-S1 operational coupled ocean-atmosphere prediction system at the Great Barrier Reef found the model can accurately predict MHW spatial extent, but struggled with longer-term forecasting due to unaccounted sub-seasonal weather variability in the region [4]. Other research used previous daily SSTs and forecast atmospheric temperature to predict SST extremes in Chesapeake Bay, USA, using a 35-day probabilistic forecast and found the model is skillful at predicting SST extremes with lead times ranging from 1-2 weeks using two predictors (SST and air temperature) as precursors [40]. Another study demonstrated the efficacy of machine learning models in predicting SSTs with high accuracy at a fraction of the computational cost of physics-based models on a global scale [50]. Employing various machine learning algorithms, including random forest, generalized additive models, and extreme gradient boosting, the study investigated their effectiveness in predicting MHWs across different regions, highlighting the variability in model performance and the absence of a one-size-fits-all solution [50]. Similarly, a localized study in the Mediterranean Sea utilized machine learning models such as random forests, long short-term memory, and convolutional neural networks to predict MHWs [6]. By incorporating lagged SSTs and selected atmospheric variables as predictors, all models successfully forecasted MHWs with at least 50% confidence at a lead time of 1 week [6]. Further studies have proven promising with the use of a random forest model predicting the presence or absence and MHW using spatial, temporal and climate variables known to impact SSTs with an accuracy of 76% at weekly time leads [19]. This study further assessed predictive power using the random forest model to predict the category of MHW as either no event, moderate, strong or severe/extreme, but forecasting accuracy dropped to 38% at weekly time leads [19]. Such studies give insight into the ability to predict sub-seasonal MHWs using various techniques across several oceanic regions.

Among previous research addressing predictability of MHWs, there is limited research that uses oceanic and atmospheric variables to predict MHWs on short timescales using daily resolution data in the northeast Pacific Ocean. Lagged sea surface temperature remains the most widely documented and adopted precursor to MHW prediction, leading to a deficit in studies that address other atmospheric and oceanic variables as sole predictors to MHWs on short timescales [20]. This highlights the significance of evaluating daily to weekly forecasts of ocean temperature extremes to develop operational forecast products for early warning systems. With climate models predicting longer and more severe MHWs, short-term predictions of MHWs are increasingly important to mitigate risks on marine ecosystems and communities dependent on these ecosystems to take adaptive measures to alleviate impacts [39].

In this thesis, we aim to enhance our comprehension of the predictive capacities of five distinct models; logistic regression, naive Bayes, gradient boosting, random forest, and feedforward neural network utilizing daily resolution data to forecast MHW events. The main goal is to assess and quantify the ability of these five models in predicting MHW events based on the selected predictor variables (wind speed, net heat flux, surface air temperature and sea level pressure), that are known to drive MHW events, as indicators to accurately predict these extreme events [41]. The final model seeks to enhance our comprehension of the precursors leading to MHWs and to establish an early warning mechanism for MHW events on a sub-seasonal timescale. Utilizing sub-seasonal forecasts can be valuable for industries such as shipping, fisheries, and coastal water management, especially in regions like the northeast Pacific Ocean where variability can significantly impact operational planning and decision-making [12]. With anthropogenic climate change projected to drive more frequent and extreme MHW events, it is increasingly important to be able to predict MHWs to help mitigate risks [30].

Chapter 2

Data and Methods

2.1 Data

2.1.1 Data Pre-Processing

The National Oceanic and Atmospheric Administration (NOAA) Daily Optimum Interpolation Sea Surface Temperature (DOISST) version 2.1 dataset spanning from 1981 to 2019 is employed to examine SST extremes and extrapolate MHW data [3]. The DOISST dataset is a comprehensive global dataset measuring SST anomalies on a daily timescale at .25° x .25° grid cell locations. Marine heatwaves are defined as "prolonged discrete anomalously warm water event that can be described by its duration, intensity, rate of evolution, and spatial extent" [21]. Similar to Hobday's definition, MHW events are defined here as at least 5 consecutive days of extreme SST anomalies exceeding at least the 90th percentile [21]. Throughout data processing, four unique MHW event dataset splits were generated: those exceeding the 90^{th} SST percentile, those exceeding the 95^{th} SST percentile, as well as two-class and four-class MHW event datasets. In the two-class datasets, class 0 represents non-MHW with SST $< 90^{\text{th}}$ percentile, class 1 represents MHWs with 90^{th}

< SST < 95th percentile and class 2 represents MHW with SST > 95th percentile. In the four-class datasets, class 0 represents non-MHW with SST < 90th percentile, class 1 represents MHW with 90th < SST < 92.5th percentile, class 2 represents MHW with 92.5th < SST < 95th percentile, class 3 represents MHW with 95th < SST < 97.5th percentile, class 4 represents MHW with SST > 97.5th percentile.

Climate variables including net heat flux, sea level pressure, surface air temperature and wind speed on daily timescales are used to predict MHWs. The explanatory variables were obtained from the National Centers for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) Reanalysis 1 [26]. The high resolution SST dataset is regridded to 2.5° x 2.5° resolution to align with both the spatial and temporal parameters of the explanatory variables [26]. The final dataset consists of daily observations from 1992 to 2019, covering variables recorded at each grid cell within the northeast Pacific Ocean, equating to 3,668,621 data points prior to data balancing. This area extends from 10°N to 65°N and from 100°W to 175°W, an ecologically significant region renowned for its rich biodiversity and vital contribution to the habitats and migration routes of numerous marine species. Codebase used for regridding can be found at `https://github.com/KatGiamalaki/Random_Forest_for_MHW`.

2.1.2 Data Balancing

Since MHW events are infrequent occurrences, they represent a minority within the dataset, with the majority of data consisting of days without MHWs. This results in class imbalance, necessitating data balancing. Failure to balance the data means models would be predominantly trained on non-MHW instances, leading the models to disproportionately favor predicting non-MHW occurrences.

To address the class imbalance issue, we use two dataset balancing approaches: random balancing and cluster centroid balancing. These techniques are used to undersample the majority class in the dataset in order to achieve a balanced class distribution for both binary and multi-class classification tasks. Random sampling entails the selection of samples from the dataset such that each majority class sample has an equal chance of being selected, without replacement and any discernible pattern, ensuring equitable representation. When applied to an imbalanced dataset, random sampling the majority class means selecting a subset of the majority class randomly to balance the class distribution relative to the minority class. Whereas cluster centroid sampling organizes similar samples into clusters and subsequently selects representative majority samples, without replacement, from the centroids of these clusters. To determine the optimal number of clusters, an elbow curve was constructed using binary MHW data. The elbow curve is a visual tool used in clustering analysis to find the optimal number of clusters in a dataset. It plots the number of clusters against a measure of clustering quality, such as within-cluster sum of squares (WCSS). The "elbow" point on the curve, where the rate of decrease in WCSS slows down significantly, indicates the optimal number of clusters. Here, seven clusters were chosen as sufficient (Figure 2.1).

Both balancing methods were geared towards addressing class imbalances by concentrating on the majority class and selecting exemplar samples to help make more robust models. Each MHW data split (90$^{\text{th}}$, 95$^{\text{th}}$, 2-Class and 4-Class MHW split) was further refined into three variations, ensuring a comprehensive exploration of data balancing techniques: unbalanced, randomly balanced, and cluster centroid balanced datasets. This approach resulted in the creation of a total of 12 distinct datasets to provide detailed insights and support thorough analysis of MHW occur-

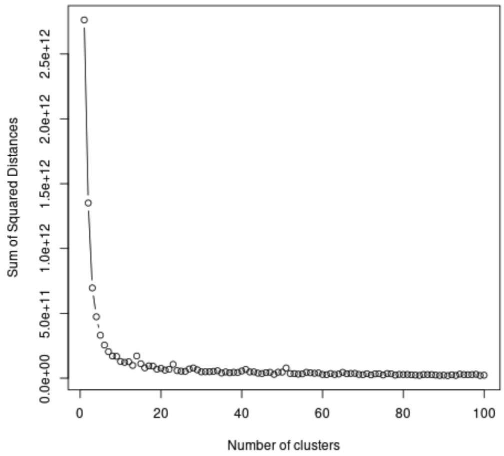


Figure 2.1: Within-cluster sum of squared distances as a function of the number of clusters. The “elbow” point on the curve, where the rate of decrease in within-cluster sum of squares slows down significantly, indicates the optimal number of clusters. Seven clusters seem sufficient.

rences and their patterns.

Additionally, each dataset was split into two subsets: a training dataset used to train the models and a testing dataset used to assess model performance. In total, there are 309,976 MHW data points exceeding the 90^{th} percentile of SSTs. The dataset was split such that 75% of the data was used for training (observations from 1992 $\sim$ 2015) while the remaining 25% was reserved for testing (observations from 2015 $\sim$ 2019). The bulk of the dataset was allocated for training to optimize model performance, while ensuring a sufficient amount of data remained available for thorough performance testing. To evaluate model performance, leads of 1, 3, 5, 7, and 14 days were incorporated into the predictor variable, which denoted the presence, absence or classification of MHWs. This step aimed to assess the predictive capability of the model across various temporal scales, exploring how effectively previous day values of predictor variables could anticipate future occurrences of MHW events. Additionally, the models were evaluated as a "nowcast" (referred here as lag 0) to gauge their ability to predict outcomes for the current day. To preserve the integrity of variable relationships, all lagging procedures were executed prior to any dataset processing and cleansing tasks, such as MHW grouping and data balancing. All analysis was conducted in R programming using version R/4.2.0 and Python programming version 3.9.14 [46, 16]. The codebase for this report can be found at `https://git.ucsc.edu/castratt/masters-project`.

2.2 Models

2.2.1 Logistic Regression Model

Logistic regression is a statistical method primarily used for binary classification tasks and can be extended to multi-classification tasks, where the goal is to predict the probability of an event occurring based on one or more predictor variables. In the context of predicting MHWs, logistic regression can be utilized to forecast the probability of an event based on various predictor variables. Logistic regression is defined as:

$$P(y = 1 \,|\, x) = \frac{1}{1 + e^{-(0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p)}} \tag{2.1}$$

where $P(y = 1 \,|\, x)$ is the probability that the target variable y equals 1 given the input features x, β_0 is the intercept term, $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients (weights) associated with each input feature $x_1, x_2, \ldots, x_p$. Here, logistic regression serves as the baseline model due to its simplicity and ease of implementation. The `glmnet` R package was used to implement logistic regression models [17, 44].

2.2.2 Naive Bayes Model

Naive Bayes is a statistical classification algorithm based on Bayes' theorem. Naive Bayes calculates the probability of a given instance belonging to each class based on the features or attributes associated with that instance [24]. It assumes that the presence of a particular feature in a class is independent of the presence of other features. Bayes' theorem is expressed as:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{2.2}$$

where $P(B|A)$ is the posterior probability of class B given predictor A, $P(A|B)$ is the likelihood, the probability of predictor A given class B, $P(B)$ is the prior probability of class B and $P(A)$ is the marginal likelihood, the probability of predictor A.

Naive Bayes is known to perform well in practice, especially with large datasets and when the independence assumption amongst predictors holds approximately true [24]. To ensure the independence assumption is respected, a simple correlation test was conducted between predictor variables, finding that there is an insignificant amount of correlation among predictors and thus the independence assumption is sustained. The algorithm works by first estimating the probabilities of each class and the conditional probabilities of each feature given the class. These probabilities are typically estimated from the training data using maximum likelihood estimation. When given a new instance with features, Naive Bayes calculates the probability of the instance belonging to each class based on these probabilities. The class with the highest probability is then predicted.

Naive Bayes is computationally efficient and simple, especially with high-dimensional data, and is commonly used in text classification [22]. It requires minimal training data to estimate the parameters, and the classification process is

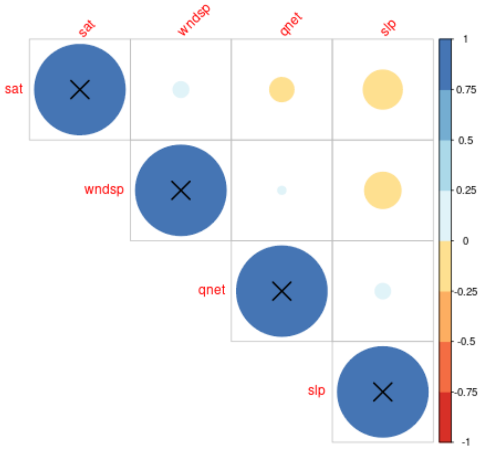


Figure 2.2: Correlation between predictor variables: wind speed (wndsp), net heat flux (qnet), sea level pressure (slp), surface air temperature (sat).

straightforward, fast and typically provides good performance [22, 52]. This makes it particularly attractive for tasks where computational resources are limited or where rapid deployment is necessary. As such, no additional model parameters were implemented. R packages that were used to achieve outlined tasks were; `naivebayes` and `e1071` [33, 35].

2.2.3 Gradient Boosting Model

Gradient boosting classification is a machine learning technique used to predict categorical outcomes, such as whether a MHW will occur based on various input variables. The model works by combining multiple weak learners, typically decision trees, to form a strong predictive ensemble [18]. The model iteratively fits new trees to the residuals of the previous trees, with each tree aiming to correct the errors made by the previous ones [18]. This process continues until a predefined number of trees are built or until no further improvements can be made.

The model starts with a single decision tree, often referred to as a weak learner. This decision tree is trained on the predictor variables and their corresponding MHW labels. However, this initial tree may not accurately capture all the patterns in the data. To improve prediction accuracy, subsequent decision trees are added to the ensemble in an iterative manner. Each new tree is trained on the residuals (the differences between the actual MHW labels and the predictions of the existing ensemble). The goal of each new tree is to correct the errors made by the previous trees, focusing on the instances where the model's predictions were inaccurate. During each iteration, the model places more emphasis on the instances where it previously made mistakes, effectively learning from its errors. This process continues until a predefined number of decision trees are added to the ensemble, or until no further improvements can be made.

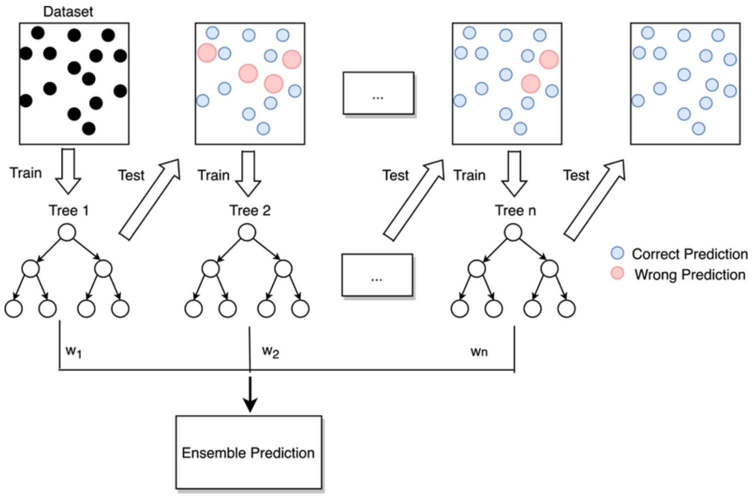


Figure 2.3: Figure from [53] demonstrates the basic mechanics of the gradient boosting ensemble prediction.

The ensemble of decision trees works together to make predictions on new data points. Each tree in the ensemble independently predicts whether a given observation corresponds to a MHW event based on the input variables. The final prediction is then determined by aggregating the individual predictions of all the trees in the ensemble. R packages that were used to achieve outlined tasks were; `‘gbm’` [17].

2.2.4 Random Forest Model

Random Forest is a non-parametric machine learning method in which the combined results of an ensemble of decision trees are used to arrive at a single result. Each decision tree in the ensemble is trained on a random subset of the data and predictor variables, reducing overfitting and capturing diverse patterns in the data [7]. The model aggregates predictions through bagging, where each tree's prediction is voted for classification [7]. A decision tree is made of a root node at the start of a tree, internal nodes that govern features of the branches and leaf nodes where a branch terminates and a decision is returned. Each tree splits into smaller groups of data, creating branches, and when the relationship between the independent variables and the dependent variable is assessed and understood by the decision trees, the model can be used to predict either binary or categorical MHW events given a set of independent predictor variables provided by the testing dataset.

The model, a supervised learning algorithm, has gained popularity for its robustness and accuracy in predictive modeling tasks where parameters are chosen for optimal model performance. Although default hyperparameters have proven to be sufficient, it is suggested that tuning specific hyperparameters further improves model performance; number of trees grown, number of variables randomly sampled each split (mtry) and minimal number of data points in node required to split ($\min_n$) [38].

The number of trees determines the ensemble's size, balancing model com-

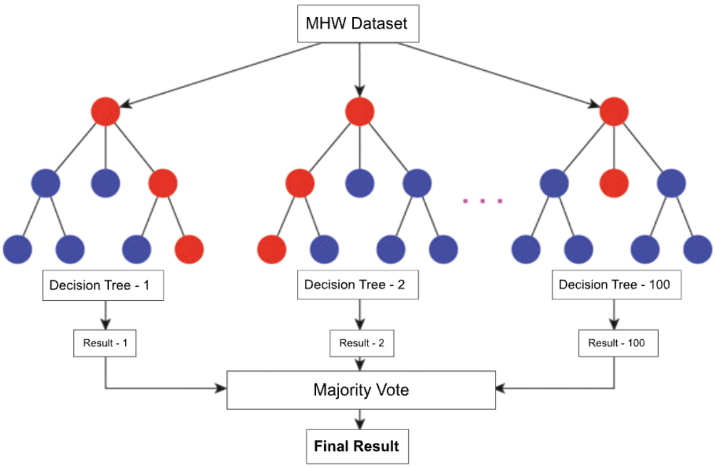


Figure 2.4: Figure adapted from [27]. The schematic illustrates the tree growth of a random forest model. One hundred trees are grown for each model configuration, from which the majority vote is used to arrive at a final result. The red indicates a potential path that the data may follow and the resulting terminal node.

plexity with computational efficiency. A trees versus error plot, shown in Figure 2.5, is used to visually determine the optimal number of trees to grow in the forest through comparing error rate for each additional tree added to the forest [32].

It was determined that 100 trees were sufficient as the gain in accuracy is negligible beyond 100 trees and any additional trees do not significantly increase performance and only increase runtime (Figure 2.5). This approach ensures a balance between predictive power and computational cost. A 10-fold cross validation grid search was then performed to determine the optimal mtry and $\min_n$ parameters based on root mean square error (RMSE) values for various combinations of these selected hyperparameters [51]. Once the optimal final model parameters were determined, the final model was fitted. R packages that were used to achieve outlined tasks were; `randomForest`, `ranger`, `tidyverse`, `tidymodels` and `caret`

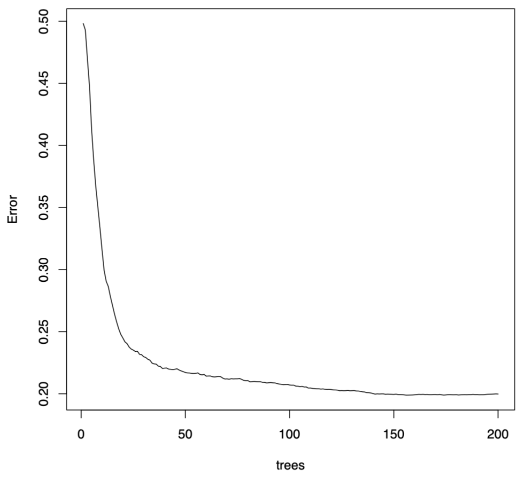


Figure 2.5: Out-of-bag (OOB) error rate as additional trees are added in the random forest model. The OOB measures the predictive accuracy of the random forest model using samples not included in each tree's bootstrap sample. The OOB error rate decreases in a logarithmic manner, until it reaches a plateau at around 100 trees. One hundred trees were chosen for the final model, as the gain in accuracy beyond 100 trees is negligible and is outweighed by the increase in additional computation time. A 0.2 error rate signifies a 20% error rate in the random forest model before additional metrics are optimized.

[28, 32, 48, 49, 51].

2.2.5 Feedforward Neural Network Model

Feedforward neural networks are a widely adopted simple neural network from which many more complex neural networks, such as recurrent neural networks (RNNs), are derived from [31]. Feedforward neural networks consist of interconnected layers of neurons, each layer passing its output as input to the next layer without any feedback loops. This structure enables the network to map input data to output predictions through a series of transformations [31]. Feedforward neural network comprises an input layer, one or more hidden layers, and an output layer. Each neuron in the network is associated with a set of learnable parameters, including weights and biases, which are adjusted during the training process to minimize the discrepancy between predicted and actual outputs [31]. Through a process known as forward propagation, input data is passed through the network, and successive layers apply nonlinear activation functions to generate increasingly complex representations of the input [47].

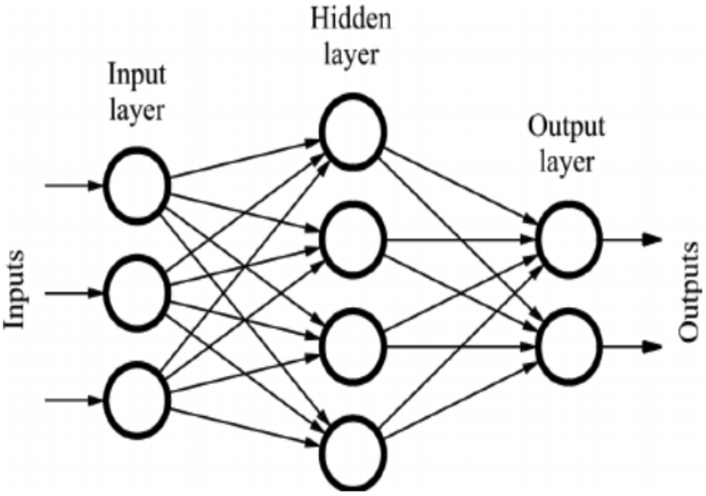


Figure 2.6: Figure from [37] demonstrates the mapping of a feedforward neural network model.

The Rectified Linear Unit (ReLU) is a nonlinear activation function and is used for both the binary and multi-class classification models and implemented in the hidden layer. The Rectified Linear Unit (ReLU) activation function is defined as:

$$f(x) = \max(0, x) \tag{2.3}$$

where x is the weighted sum of the input layers from the previous layer. The function returns the input value if it is positive, and zero otherwise. This simple yet effective activation function introduces non-linearity to the neural network, enabling it to learn complex relationships in the data. Its nonlinear nature allows the network to capture and represent non-linear relationships between input features and target outputs. This is essential for learning complex patterns in the data, making ReLU a popular choice in neural network architectures. Whether the task is binary classification or multiclass classification, ReLU can be utilized effectively in the hidden layers of the FNN.

The sigmoid function, also known as the logistic function, is a nonlinear activation function and is used for the binary classification models and implemented in the output layer. It takes an input (usually the weighted sum of inputs plus a bias term) and compresses it to a value between 0 and 1, making it suitable for modeling binary classification problems and producing probability-like outputs. The sigmoid function is defined as:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2.4}$$

where x is the weighted sum of inputs plus a bias term. As the input x increases, the sigmoid function asymptotically approaches 1, indicating a high probability of the positive class. Conversely, as x decreases towards negative infinity, the function approaches 0, signaling a high probability of the negative class. In situations where x is close to 0, the sigmoid function yields a probability close to 0.5, denoting uncertainty between classes.

The softmax function is implemented for multiclass classification tasks in the output layer. It transforms the raw output scores from the neural network's final layer into probabilities, ensuring they sum to 1. The softmax function is defined as:

$$\text{softmax}(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_i}} \tag{2.5}$$

where z_i is the raw score (logit) for the i$^{\text{th}}$ class and N is the total number of classes. This function exponentiates each score and normalizes it by dividing it by the sum of all exponentiated scores, resulting in a probability distribution across all classes. The output probabilities facilitate classification by indicating the likelihood of each class, allowing the model to predict the class with the highest probability as the final output. R packages that were used to achieve outlined tasks were; `tensorflow` and `keras` [1, 25].

2.3 Model Evaluation and Performance Metrics

When evaluating performance across all models, Receiver Operating Characteristic curve (ROC) and Precision Recall curve (PR) are used to provide a graphical representation of model evaluation. The ROC curve is a graphical representation of the true positive rate (Sensitivity) against the false positive rate (1-Specificity) across different decision thresholds. The ROC curve provides the trade-off between the true positive rate (TPR) and the false positive rate (FPR) across different threshold values. It aids in visualizing the performance of binary classification models, illustrating the model's ability to discriminate between positive and negative instances. A perfect classifier would exhibit an ROC curve that reaches the upper left corner of the plot (Sensitivity=1, Specificity=1), whereas a random classifier would resemble the diagonal line (AUC=0.5), represented in the following plots as the light grey line. The Area Under the ROC Curve (AUC) quantifies the overall performance of the model across all possible thresholds. A higher AUC value indicates better discriminatory power, with a maximum value of 1 indicating perfect classification. True positive rate (TPR) is defined as:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2.6}$$

where true positive (TP) is the instance where the model correctly predicted a MHW (i.e. predicted class 1 MHW event and it indeed occurred in reality). False negative (FN) is defined as the instances where the model incorrectly flagged non-MHW instances as MHW events (i.e. predicted a MHW event (class 1), but it does not occur in reality (the actual class is 0)). False positive rate (FPR) is defined as:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{2.7}$$

where false positive (FP) is the instance where the model failed to detect MHW events (i.e. predicted the absence of a MHW (class 0), but it occurs in reality (the actual class is 1)). True negative (TN) is defined as the number of instances where the model correctly identified non-MHW (i.e. (class 0), and it indeed does not occur in reality). AUC_{ROC} is defined as:

$$AUC_{ROC} = \int_0^1 \mathrm{TPR}\, d(FPR) \tag{2.8}$$

where the integral represents the area under the ROC curve and provides a measure of the overall discriminatory power of a classification model across all possible thresholds. An AUC_{ROC} of 1 indicates a perfect classifier and 0.5 indicates a random classifier. The AUC_{ROC} is a helpful metric because it offers a comprehensive evaluation of model performance across all classification thresholds, effectively capturing the model's ability to discriminate between positive and negative instances regardless of class distribution.

The Precision-Recall curve is another evaluation metric commonly employed in binary classification tasks. Unlike the ROC curve, which focuses on the true positive rate against the false positive rate, the PR curve depicts the trade-off between precision (positive predictive value) and recall (sensitivity) across various decision thresholds. In situations where the positive class is rare or of particular interest, the PR curve can provide a more informative assessment of the model's performance compared to the ROC curve. The area under the Precision-Recall curve (AUC_{PR}) quantifies the model's ability to balance precision and recall, with higher values indicating superior performance in capturing relevant instances while minimizing false positives. Precision (P) is defined as:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{2.9}$$

and recall (R) is defined as:

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2.10}$$

where TP, FP and FN are the same as defined above. AUC_{PR} is defined as:

$$AUC_{PR} = \int_0^1 \mathrm{Precision}\, d(Recall) \tag{2.11}$$

where the integral represents the area under the PR curve, where 0 indicates a poor model and 1 indicates perfect model. The AUC_{PR} metric provides a balanced assessment of model performance by focusing on the precision-recall trade-off, specifically evaluating the classifier's ability to rank positive instances higher than negative ones.

Two performance metrics are used to evaluate and compare the performance across all models: accuracy and hit-rate. All models are tested using a holdout dataset to assess predictive power as a nowcast and leads of 1, 3, 5, 7 and 14 days.

Accuracy measures the proportion of correctly classified instances among the total instances. Here, mean accuracy is calculated by averaging accuracy across all correctly predicted MHW classes. This provides a robust assessment of model performance under conditions of class balance, where each class is evenly represented within the dataset, but may not be sufficient when dealing with imbalanced datasets, where one class dominates the mean accuracy score. In the case of testing accuracy of MHWs using unbalanced data, where one class disproportionately outweighs the others, accuracy is heavily influenced by the dominant class (i.e. no MHW instance) and therefore falsely inflates the accuracy metric. To address these issues of class imbalance and rare event detection, hit-rate is used as complementary metrics to enhance the comprehensive evaluation of model performance. Accuracy (A) is defined as:

$$A = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2.12}$$

and mean accuracy (MA) is defined as:

$$MA = \frac{A_0 + A_1 + \ldots + A_n}{\text{number of MHW classes}} \tag{2.13}$$

where A_n represents MHW class accuracy n. Accuracy and mean accuracy quantify the proportion of all correctly predicted instances by the model.

Hit-rate evaluates the model's ability to predict rare events accurately. It focuses specifically on the model's performance concerning positive instances and is especially valuable as a complimentary metric to assess predictive power of MHW events. Hit-rate is defined as:

$$\text{Hit Rate} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.14}$$

where TP represents true positives and FP represents false positives. This metric helps quantify the proportion of correctly predicted positive instances out of all instances predicted as positive by the model.

For evaluating spatial accuracy and hit rate, spatial plots were employed to depict the effectiveness of each model in predicting the spatial distribution of a MHW occurrence on a chosen day that showcased a significant number of MHW events. To broaden the spatial visualization, the hit rate was computed for each grid cell over the period of the testing data to provide a comprehensive understanding of the model's predictive performance across the temporal and spatial domain.

In the final section of the results (3.3), we apply the best-performing model identified in the preceding sections; the random forest model. Additionally, we augment the model with additional predictor variables to evaluate its predictive efficacy. The presentation of the final model results mirrors that of the previous sections, featuring similar figures and spatial plots.

Chapter 3

Results

In this section, we present the results of our model performance in two different model configurations: 90th percentile binary MHW prediction (Section 3.1) and 2-Class MHW prediction (class 0 = non-MHW with SST < 90th percentile, class 1 = MHW with 90th < SST < 95th percentile, class 2 = MHW with SST > 95th percentile (Section 3.2)). The remaining model configurations, 95th percentile binary MHW prediction (Section A.2) and 4-class MHW (class 0 = non-MHW with SST < 90th percentile, class 1 = MHW with 90th < SST < 92.5th percentile, class 2 = MHW with 92.5th < SST < 95th percentile, class 3 = MHW with 95th < SST < 97.5th percentile, class 4 = MHW with SST > 97.5th percentile (Section A.4)), are touched on in this section, but the main results are presented in the appendix as they exhibit similar behavior to that of the selected model configurations presented in this section. We present the findings of the two main selected configurations, 90th percentile binary MHW and 2-class MHW, via various methods to provide a comprehensive evaluation of each model's predictive capabilities.

3.1 90th Percentile MHW

Figure 3.1 represents the ROC curve and associated AUC for each model trained on random balanced data for a binary MHW outcome at the 90th threshold. The diagonal light grey line indicates random chance (associated with an AUC of 0.50), demonstrating all models perform better than random chance. The random forest model, performs significantly better than other models, with an AUC score of 0.89 for class 0 and class 1 (Figure 3.1D). While the logistic regression model performed the worst (AUC = 0.56 for class 0 and class 1, Figure 3.1A), the remaining models shown on Figure 3.1 B, C and E do not perform much better, all with AUC scores near 0.61 for class 0 and class 1. The ROC curves for the cluster centroid balanced data, depicted in Figure 3.2, demonstrated slight enhancements across all models in terms of AUC scores. Notably, the dominance of the random forest model (AUC = 0.93 for both class 0 and class 1, Figure 3.2D), the underperformance of logistic regression (AUC = 0.58 for both class 0 and class 1, Figure 3.2A), and the intermediate to poor performance of the remaining models (AUC ranging from 0.63 to 0.64 for both class 0 and class 1, Figure 3.2 B, C, and E) were consistent.

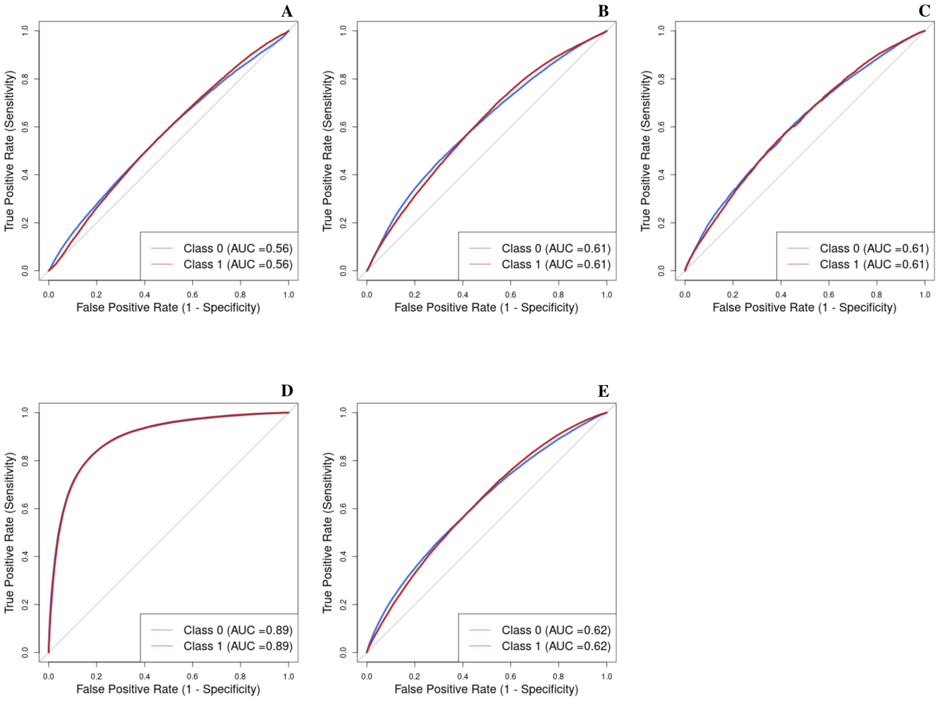


Figure 3.1: ROC curve for the 90th percentile random balanced binary MHW prediction across all models: A) Logistic Regression, B) Naive Bayes, C) Gradient Boosting, D) Random Forest, E) Feedforward Neural Network.

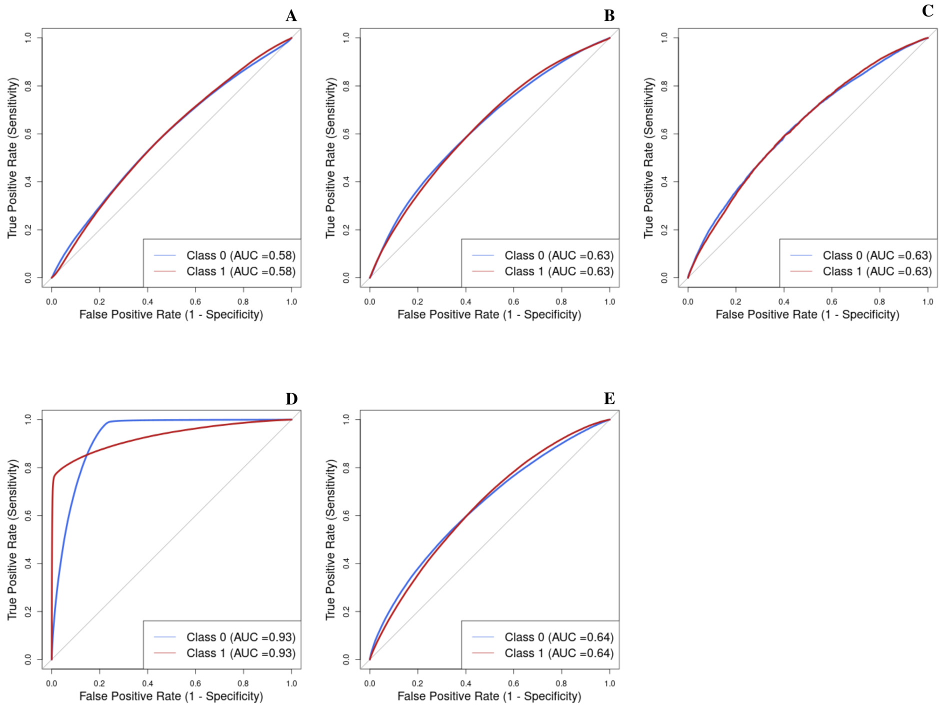


Figure 3.2: ROC curve for the 90th percentile cluster centroid balanced binary MHW prediction across all models: A) Logistic Regression, B) Naive Bayes, C) Gradient Boosting, D) Random Forest, E) Feedforward Neural Network.

Figure 3.3 displays the Precision-Recall (PR) curve and corresponding AUC values for all models using random balanced data, providing an additional metric for model evaluation. Notably, the feedforward neural network model (AUC = 0.22 for class 0 and AUC = 0.65 for class 1, Figure 3.3A) exhibits the poorest performance, while the random forest model (AUC = 0.78 for class 0 and AUC = 0.95 for class 1, Figure 3.3D) demonstrates the highest performance. It is worth mentioning that across all models, the AUC score for class 0 is consistently lower than that for class 1, suggesting that in the balanced dataset, the classifiers tend to achieve higher precision but lower recall. Figure 3.4 presents the PR curves for cluster centroid balanced data and exhibit similar performance to random balanced data PR performance, with slight AUC improvements across both classes and all models. Notable across all PR curves is the poor performance of the feedforward neural network, rather than the baseline logistic regression model.

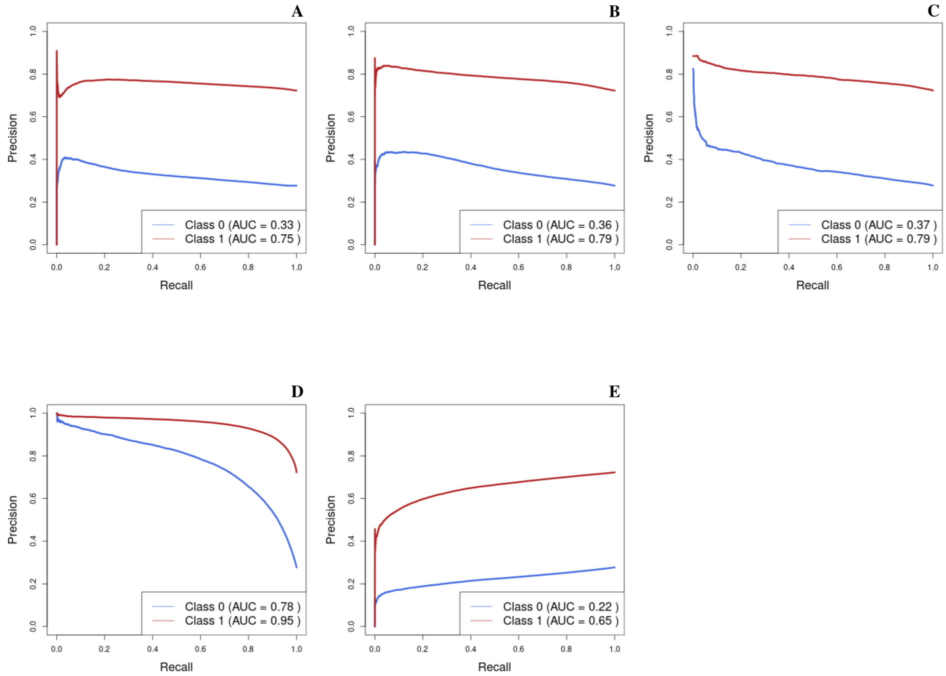


Figure 3.3: PR curve for the 90$^{\text{th}}$ percentile random balanced binary MHW prediction across all models: A) Logistic Regression, B) Naive Bayes, C) Gradient Boosting, D) Random Forest, E) Feedforward Neural Network.

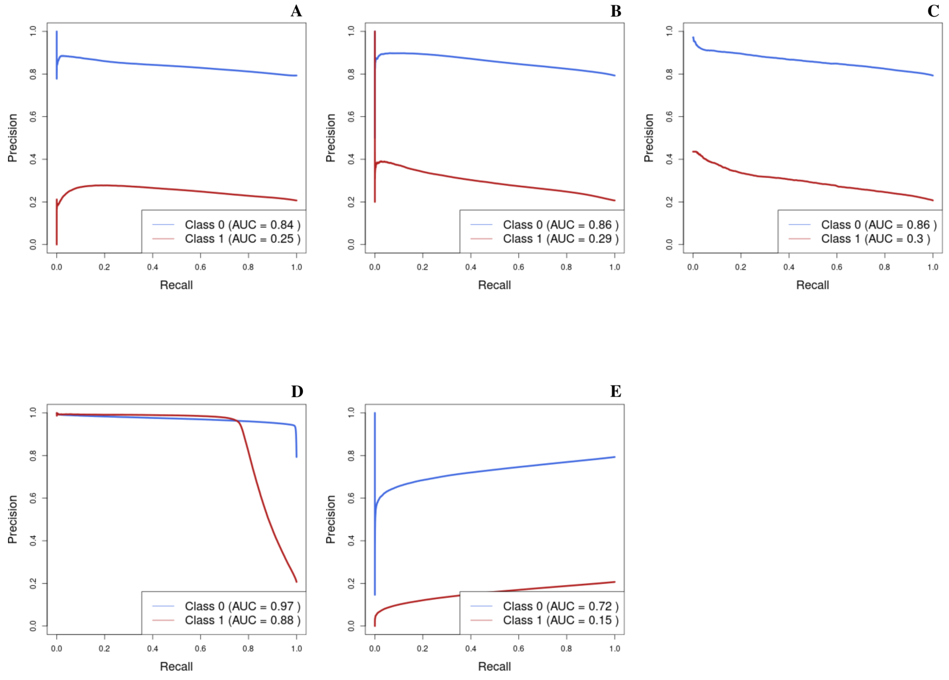
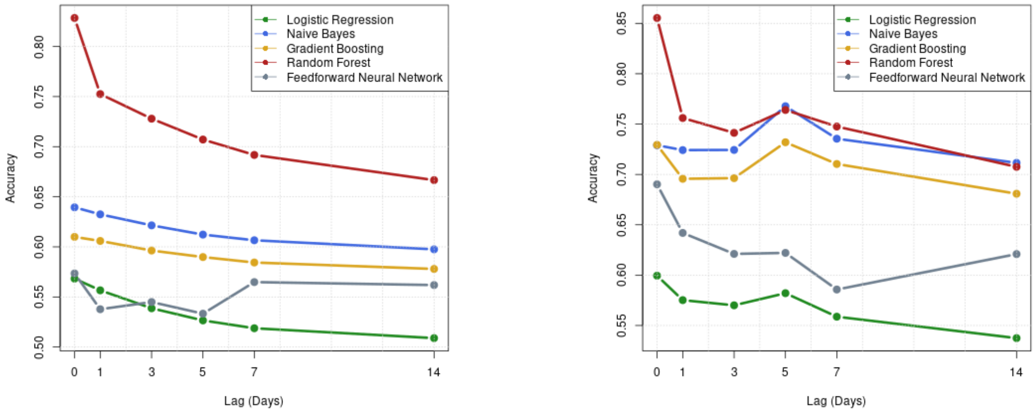


Figure 3.4: PR curve for the 90th percentile cluster centroid balanced binary MHW prediction across all models: A) Logistic Regression, B) Naive Bayes, C) Gradient Boosting, D) Random Forest, E) Feedforward Neural Network.

To track model performance over time, we evaluated accuracy and hit rate against lagged predictor variables. Figures 3.5a and 3.5b illustrate the accuracy of each model across different lag periods, ranging from 1 day to 2 weeks, using random balanced and cluster centroid balanced data, respectively. In both cases, lag 0 consistently exhibited the highest accuracy across all models, while increasing lag resulting in reduced accuracy. There is marginal improvement in accuracy observed with the cluster centroid balanced data, indicating its slightly superior performance. Although the random forest model remains dominant in both scenarios, the implementation of cluster centroid balanced data notably enhanced the performance of the naive Bayes model, rendering it comparable to the random forest.

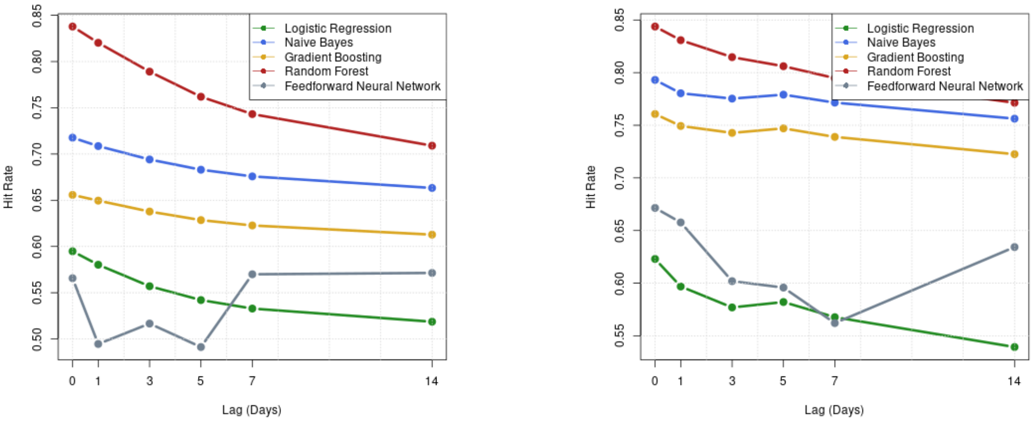


(a) Random Balanced Data

(b) Cluster Centroid Balanced Data

Figure 3.5: Accuracy versus lagged predictor variables using balanced binary 90th percentile MHW data across all models: Logistic Regression, Naive Bayes, Gradient Boosting, Random Forest and Feedforward Neural Network.

To narrow down the focus of model evaluation and look specifically at instances where models accurately predicted a MHW event, we constructed a lag versus hit rate plot. This plot encompasses lag periods ranging from 1 day to 2 weeks for all models using random balanced and cluster centroid balanced data, as shown in Figure 3.6a and Figure 3.6b, respectively. Similar to the accuracy versus lag plot, the lag versus hit rate plot reveals a decrease in hit rate with increased lag time. Across both balanced datasets, the random forest model consistently performed the best, followed by naive Bayes, gradient boosting, feedforward neural network and logistic regression. There was a slight improvement observed when using the cluster centroid balanced data. Across all models when using cluster centroid balanced data, the decline in hit rate is slower than that of the random balanced data, with notable enhancements observed in the naive Bayes and gradient boosting models. Overall scoring of the hit rate versus lag plot across all models was better than the accuracy versus lag plot, indicating that all models are better at predicting the presence of a MHW rather than the absence of a MHW.



(a) Random Balanced Data

(b) Cluster Centroid Balanced Data

Figure 3.6: Hit rate versus lagged predictor variables using balanced binary 90th percentile MHW data across all models : Logistic Regression, Naive Bayes, Gradient Boosting, Random Forest and Feedforward Neural Network.

To illustrate the models' ability to forecast the presence or absence of a MHW event, spatial plots were generated for each model. These plots depict the observed MHW occurrences on a specific day alongside the model predictions utilizing lagged data at intervals of 7 and 14 days, where the presence of a MHW is indicated by a black dot and the data is superimposed onto observed SST anomalies for the specified day. This analysis encompassed both data balancing techniques, providing insights into the predictive capacity of each model across different temporal contexts and balancing strategies. To highlight the strongest model's predictive capability, Figure 3.7 provides insights into the predictive capabilities of the random forest model using cluster centroid balanced data regarding the presence or absence of a MHW on August 18, 2015 — a day characterized by a notable frequency of MHW occurrences. Figures A.1 through A.10, found in the appendix, provide further insights into the predictive capabilities of each model regarding the presence or absence of a MHW on the same day, August 18, 2015.

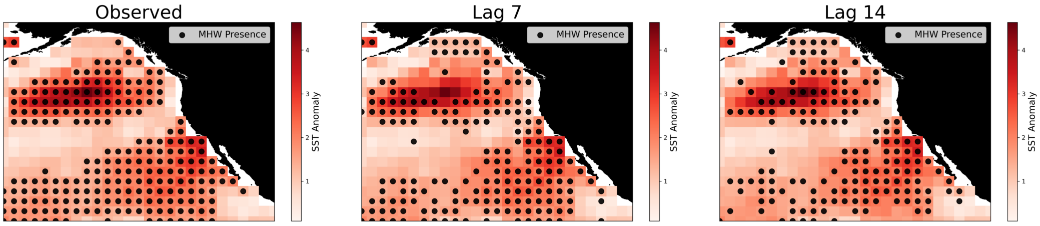


Figure 3.7: Random forest model forecasts for the 90th percentile MHW using cluster centroid balanced data for August 18, 2015, with no lag, 7 days lag and 14 days lag. The color bar represents the observed sea surface temperature anomaly and dots indicate where a MHW was correctly predicted.

Notable trends represented in the spatial plots is the inability of the logistic regression model and feedforward neural network to capture spatial patters on the given day, though there is slight improvement in prediction in the feedforward neural network when using cluster centroid balanced data. While the naive Bayes and gradient boosting models demonstrate moderate predictive power at 1 week and 2 week lead times, the random forest model demonstrates the strongest ability to predict MHW occurrence. On the selected day, the random forest model (Figure 3.7) is able capture the spatial extent of the MHW with good accuracy, whereas the remaining models (Figures A.1 to A.10) struggled to capture the MHW spatial shape and extent.

To further illustrate each model's ability to correctly predict MHW events, a spatial hit rate plot was compiled for each model for both data balancing techniques. The hit rate is averaged across the entire testing dataset for each grid cell, resulting in a comprehensive spatial plot that illustrates the model's effectiveness in predicting MHWs where a higher hit rate (indicated by dark red here) represents a higher model performance. Figure 3.8 represents the random forest model with cluster centroid balanced data, demonstrating the model's strong ability to correctly predict MHWs in the region. Figures A.11 through A.19, found in the appendix, demonstrate the remaining model's spatial hit rate plots. Again, we are seeing similar patters that have been observed with supporting data: poor performance with logistic regression, moderate performance for naive Bayes and gradient boosting, moderate to poor performance for feedforward neural network and high performance for random forest. Across all models, performance increases when using cluster centroid balanced data.

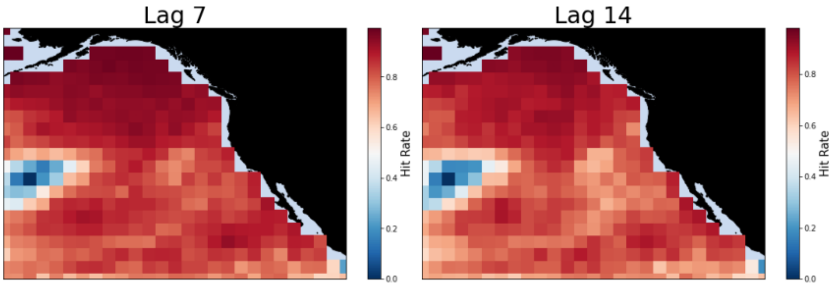


Figure 3.8: Random Forest hit rate spatial plot for binary 90th percentile MHW outcome with cluster centroid balanced data.

When the binary MHW threshold was extended to 95th percentile SST as a MHW events, the results were very similar to that of the 90th percentile MHW exhibited above. The ROC and PR curves for the 95th percentile MHW exhibited similar AUC values to the 90th percentile MHW, with the random forest performing the best (Figures A.22 through A.25). The accuracy versus lag and hit rate versus lag plots exhibited almost the same behavior as the 90$^{\text{th}}$ percentile MHW, with only small decreases in accuracy and hit rate across many of the models while the model ranking and behavior with increased lags remained the same as the 90$^{\text{th}}$ percentile (Figures A.26 through A.29). Figures A.30 through A.39 representing the spatial accuracy for each model, with notable differences from the 90$^{\text{th}}$ percentile being poor performance across all models at the 2 week lag time, demonstrating all models struggle to predict a higher threshold binary MHW event at longer lag times. Further emphasizing the same patterns of model performance using the 95$^{\text{th}}$ percentile, Figures A.40 through A.48 represent the spatial hit rate plots across all models and both balancing methods. Overall, the 95$^{\text{th}}$ percentile MHW threshold exhibited the same behaviors as the 90$^{\text{th}}$ percentile MHW threshold: random forest using cluster centroid data had the best overall performance. Seeing the similar patterns among both binary MHW thresholds suggests that despite raising the threshold for identifying a MHW, the models did not demonstrate significant adverse impacts.

3.2 2-Class MHW

In this section, the primary emphasis will be on utilizing cluster centroid balanced data, which has been demonstrated as the most effective balancing method in the preceding section (Section 3.1). Figure 3.9 depicts the Receiver Operating Characteristic (ROC) curve and associated Area Under the Curve (AUC) for each model trained on cluster centroid balanced data for a 2-class MHW outcome. All models demonstrate poor performance in distinguishing between MWH classes. Notably, the logistic regression, gradient boosting, and feedforward neural network exhibit particularly inadequate performance (Figure 3.9 A, C, and E, respectively).

Both the naive Bayes and random forest perform relatively better than the other models, but even so these models do not perform well, with the highest AUC of 0.63. Four of five models perform worse for class 1, suggesting that the models struggle to effectively differentiate between MHW classes, often misclassifying MHW events as either non-MHW (class 0) or the most extreme MHW class (class 2).

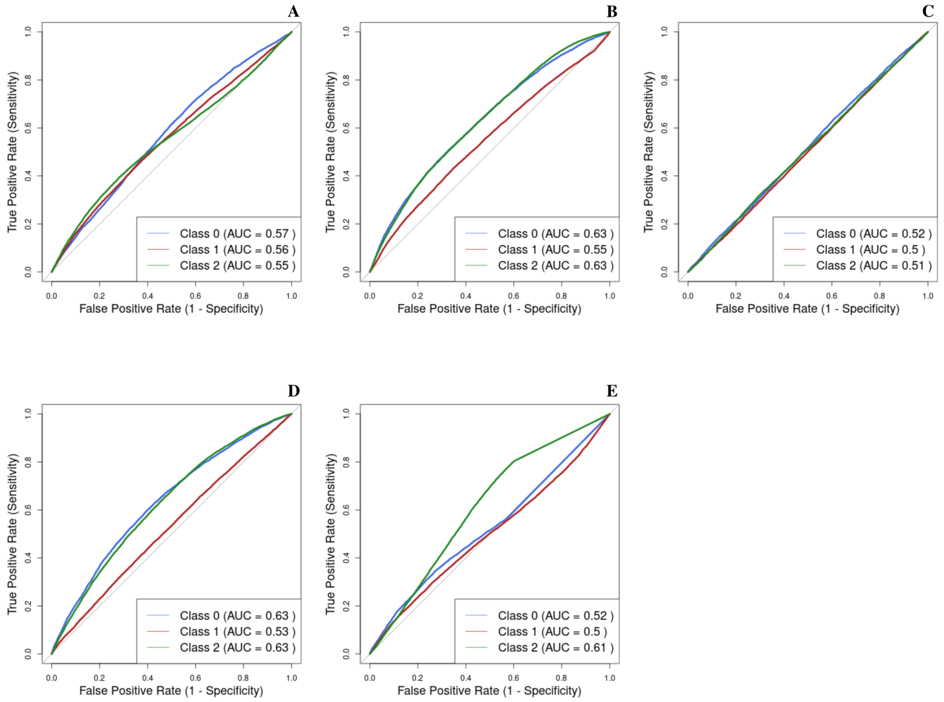


Figure 3.9: ROC curve for cluster centroid balanced 2-class MHW prediction across all models: A) Logistic Regression, B) Naive Bayes, C) Gradient Boosting, D) Random Forest, E) Feedforward Neural Network.

Figure 3.10 illustrates the Precision-Recall (PR) curves and corresponding Area Under the Curve (AUC) values for all models trained on cluster centroid balanced data, providing further insights into the predictive performance of 2-class MHWs. Despite utilizing cluster centroid balanced data, the feedforward neural network demonstrates the poorest performance overall. It exhibits an AUC of 0.18

for class 0, 0.48 for class 1, and 0.43 for class 2 (Figure 3.10E). Conversely, the best-performing models, the random forest and naive Bayes, do not show significant improvements in AUC across all classes (Figure 3.10 B and D).

These findings underscore the challenges faced by all models in distinguishing and accurately classifying MHW events across different classes. Even when cluster centroid balanced data is employed, all models continue to struggle, as evidenced by similarly poor AUC scores to those observed with random balanced data (Figure A.52 and A.53).

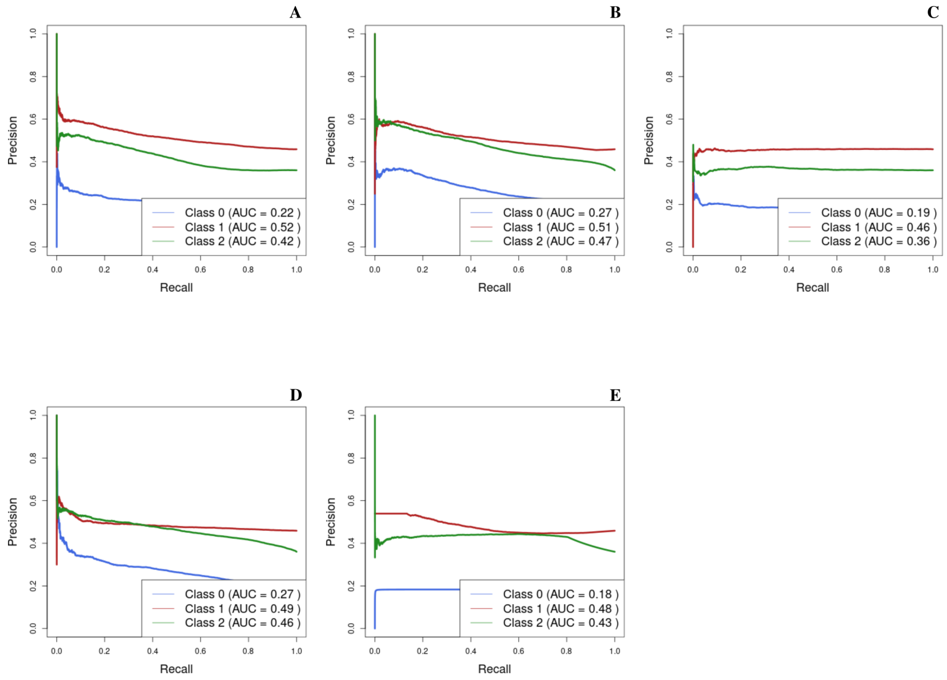
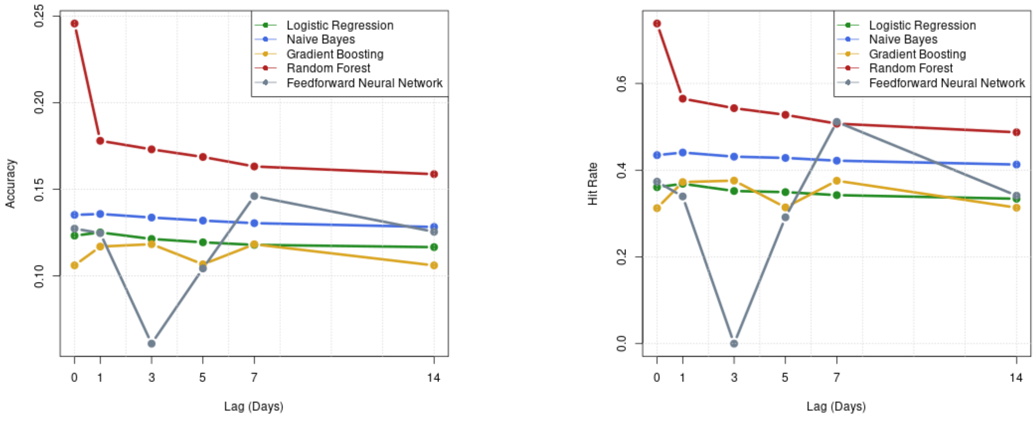


Figure 3.10: PR curve for cluster centroid balanced 2-class MHW prediction across all models: A) Logistic Regression, B) Naive Bayes, C) Gradient Boosting, D) Random Forest, E) Feedforward Neural Network.

To track the evolution of model performance, we evaluated accuracy and hit rate against lagged predictor variables. In Figures 3.11a and 3.11b, we present the accuracy and hit rate of each model across different lag periods, ranging from 1 day to 2 weeks, using cluster centroid balanced data. Across both scenarios, lag 0 consistently showcased the highest accuracy across all models. All model accuracies and hit rates are averaged between all classes and thus are likely dragged down by intermediate classes that perform very poorly (in this case, class 1 predictions). Random forest remained the dominant model for both accuracy and hit rate, followed by naive Bayes, logistic regression, gradient boosting and feedforward neural network. Note here the poor performance of gradient boosting, even worse than the logistic regression baseline model. Performance among all models is very poor, indicating all models struggle to predict MHW classes.



(a) Accuracy versus lagged predictor variables using cluster centroid balanced 2-class MHW data across all models

(b) Hit rate versus lagged predictor variables using cluster centroid balanced 2-class MHW data across all models

Figure 3.11: Lag Accuracy and hit rate for 2-Class MHW with cluster centroid balanced data across all models : Logistic Regression, Naive Bayes, Gradient Boosting, Random Forest and Feedforward Neural Network.

To illustrate the models overall hit rate, spatial plots were generated for each model demonstrating each model's average hit rate for lag 7 and lag 14 across both random and cluster centroid balanced data. Here, we show only random forest with cluster centroid data to demonstrate the relatively best 2-class MHW predictions, demonstrating that even the best of the multi-class configuration struggles to correctly classify MHWs into categories (Figure 3.12). Figures A.56 through A.64, found in the appendix, demonstrate all models struggled to correctly predict MHW classes, with minor improvements when using cluster centroid data rather than random balanced data.

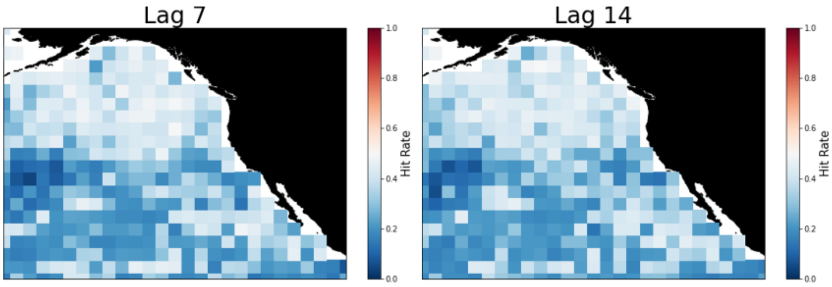


Figure 3.12: Random forest hit rate spatial plot for 2-class MHW outcome with cluster centroid balanced data.

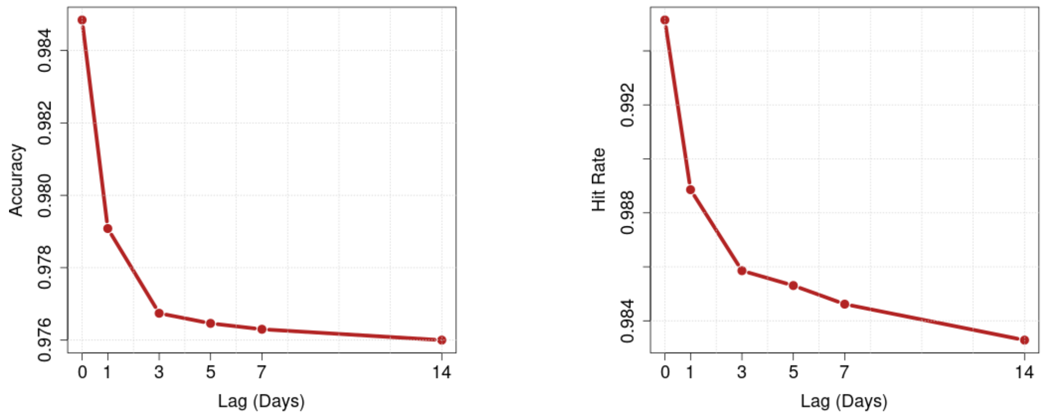
Figures A.67 and A.68 present ROC curves and associated AUC values for models trained on randomly balanced and cluster centroid balanced data for 4-class MHW predictions. Despite varying approaches, all models demonstrate poor performance in distinguishing between MHW classes. Notably, logistic regression and gradient boosting models exhibit particularly weak predictive power. However, random forest and naive Bayes models show relatively better performance, especially in distinguishing extreme MHW classes (class 0 and class 4). Similarly, Precision-Recall (PR) analysis, depicted in Figures A.69 and A.70, reveals consistent poor predictive power across all models. This underscores the challenges in discerning relative performance nuances from PR curves due to minimal differences.

Evaluation of accuracy and hit rate over various lag periods further reinforces the models' limitations in correctly predicting MHW classes. Figures A.71 and A.72 illustrate poor accuracy across different lag periods for both randomly balanced and cluster centroid balanced data. Similarly, Figures A.73 and A.74 highlight poor hit rate performance across lags. Spatial hit rate analysis (Figures A.75 through A.84) demonstrates marginal improvement compared to 2-class MHW predictions in average hit rate for 4-class MHW, particularly with cluster centroid data. However, overall performance remains poor across all models and data balancing techniques.

Across the model configurations in this study (90th percentile MHW, 95th percentile MHW, 2-class MHW and 4-class MHW) the performance of the feedforward neural network in predicting MHWs has been marked by inconsistency and poor performance. The model struggles to provide accurate predictions, yielding probabilities for events that closely resemble random chance. Consequently, its reliability and consistency are compromised. It is important to acknowledge that while the results of the feedforward neural network are presented here, its unreliability necessitates limited discussion.

3.3 Final Model

From the results presented above, we implemented the best model with the best performing specifications; the random forest model with the 90th percentile MHW binary outcome using cluster centroid data. To further evaluate and enhance the final model, sea surface temperature, temporal (day of year) and spatial location



(a) Accuracy versus Lags

(b) Hit Rate versus Lags

Figure 3.13: Accuracy and hit rate for final random forest model with cluster centroid balanced data.

were added as a predictor to the previously outlined predictor variables. The results show a substantial improvement in overall model accuracy and hit rate. Accuracy for the previous random forest model ranged from 0.85 to 0.71 for lag 0 through lag 14, while the new random forest model has accuracy ranging from 0.98 to 0.97 for lag 0 through 14 (Figures 3.13a and 3.5b). Hit rate exhibited similar improvements, with the previous random forest model having hit rates ranging from 0.85 to 0.77 while the new fitted model hit rates range from 0.99 to 0.98 (Figures 3.6b and 3.13b).

The final fitted random forest model exhibits the best predictive behavior, with the spatial plot demonstrating the model's ability to correctly predict almost all locations of a MHW event on an arbitrarily selected day, August 15, 2018 (Figure 3.14). As shown in the figure, the model is able to capture almost all of the spatial extent and shape of the MHW event on the given day for both a lag 7 and lag 14. Although, it is important to note that unlike the original random forest model, the new fitted model over predicted in this region on this given day, especially for lag 7

(Figure 3.14). The average hit rate plot provides additional evidence of the model's consistently high hit rate across all grid cells, with each cell's hit rate averaged over the entire testing dataset (Figure 3.15).

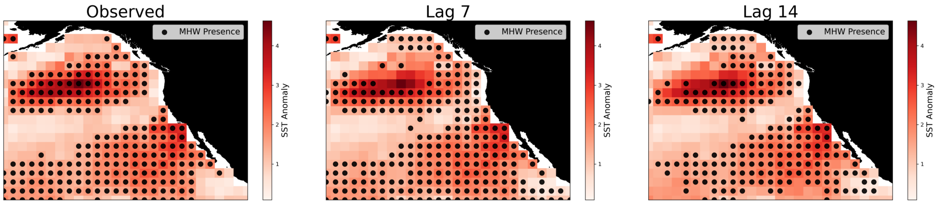


Figure 3.14: Final random forest model spatial plot with cluster centroid balanced data.

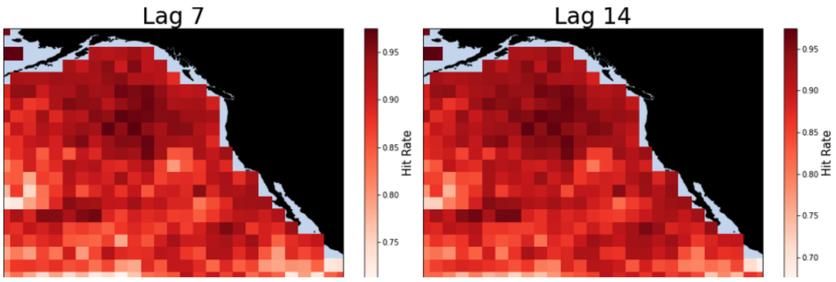


Figure 3.15: Final random forest model spatial hit rate plot with cluster centroid balanced data.

Sea surface temperature is by far the most important predictor variable in the final model (Figure 3.16). Although the remaining predictor variables in the final model exhibit comparatively less significance, date, latitude, and longitude emerge as the subsequent most important predictors following SST. For comparison, we also show the predictor variable importance for the baseline random forest model that was implemented with just net heat flux, sea level pressure, surface air temperature and wind speed to demonstrate these predictor variables show relatively similar importance, with surface air temperature showing the most variable importance

(Figure 3.17).

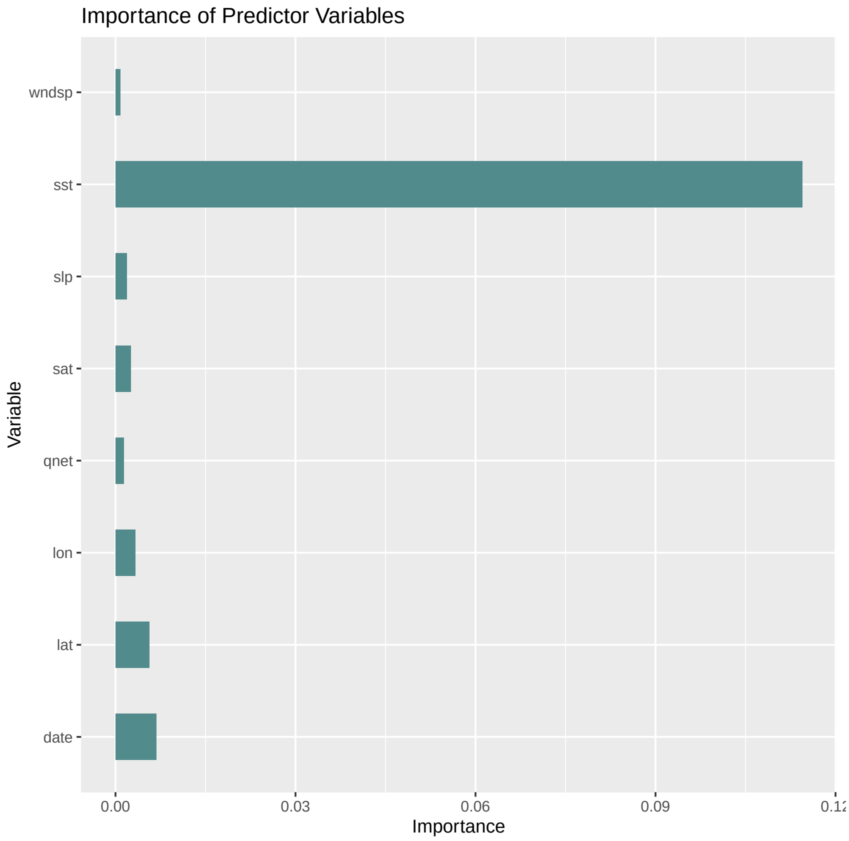


Figure 3.16: Final random forest model variable importance plot with cluster centroid balanced data. The final model includes sea surface temperature, date, latitude, longitude, wind speed, net heat flux, sea level pressure and surface air temperature as predictors to MHWs.

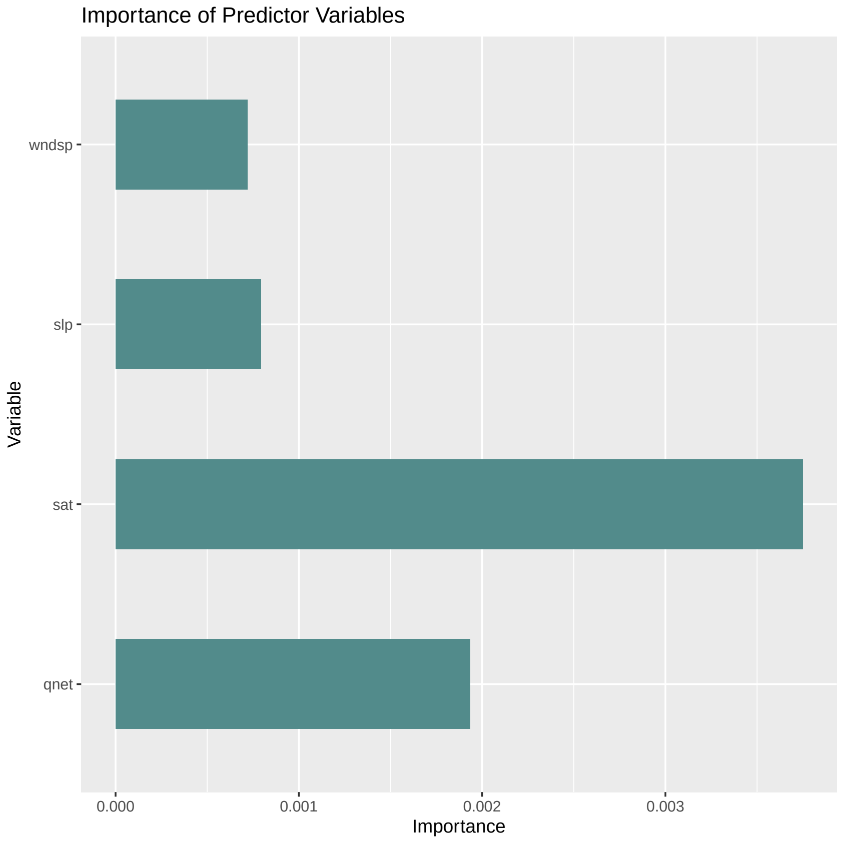


Figure 3.17: Random forest model variable importance plot with cluster centroid balanced data, excluding SST, temporal and spatial predictor variables.

Chapter 4

Discussion and Conclusion

4.1 Binary MHW Classification

In the first part of the study, we evaluate the predictive capabilities of five models in forecasting MHWs within the northeast Pacific Ocean; logistic regression, naive Bayes, gradient boosting, random forest, and feedforward neural network. We utilize selected atmospheric and oceanic variables to gauge their efficacy in predicting binary MHW outcomes, defined by either the 90$^{\text{th}}$ or 95$^{\text{th}}$ of SST anomalies in the dataset. Based on the selected predictor variables, we find that random forest is the overall best performed model in predicting the presence of absence or a MHW at both the 90$^{\text{th}}$ and 95$^{\text{th}}$ MHW thresholds.

When assessing the accuracy and hit rate for the 90$^{\text{th}}$ and the 95$^{\text{th}}$ percentile binary MHW lagged data across all models and balancing techniques, the general performance ranking holds true: random forest, naive Bayes, gradient boosting, feedforward neural network and logistic regression. The best performing model, the random forest model, for the 90$^{\text{th}}$ percentile MHW has accuracy for lag 0 using random and cluster centroid balanced data of 0.83 and 0.85 with the accuracy of that same model configuration slowly decreasing to 0.67 and 0.71 for lag 14, respectively (Figures 3.5a and 3.6b), while the accuracy of the random forest model at the 95th percentile threshold is similar to that of the 90th (Figures A.26 and A.27).

Across all models, there is a slight increase in accuracy and hit rate when using cluster centroid data compared to random balanced data, due to cluster centroid's ability to select more representative samples when undersampling the majority class. Notably, the naive Bayes model shows significant improvement with cluster centroid balanced data, achieving performance comparable to the random forest model. The hit rate plots across all models and data balancing techniques exhibit similar behavior to the lag accuracy plots, with hit rate being slightly higher than accuracy, indicating that all models are slightly better able to predict the presence of a MHW rather than the absence (Figures 3.6a and 3.6b).

At both the 90th and 95th MHW thresholds, the random forest model performs the best in terms of accuracy and hit rate when tested using lag 0 predictor data for both the random balanced data and the cluster centroid balanced data, which is no surprise as we expect the accuracy to be high at lag 0 and decrease with lagged data (Figures 3.5 through 3.6 and A.26 through A.29).

Figures A.1 through A.10 exemplify the predictive capabilities of each model in discerning the occurrence or absence of MHWs at the 90th percentile MHW threshold on a randomly selected day. These figures provide a snapshot into the performance of each model in predicting MHWs with temporal lags of 1 and 2 weeks, highlighting their effectiveness in capturing both the spatial accuracy and extent of a given MHW event. Gradient boosting and random forest appear to be able to perform the best in capturing the MHW event, while the remaining models lack in capturing spatial extent.

To extend the analysis further, figures A.30 through A.39 provide insight into the same MHW event, but using the 95th percentile MHW threshold. As the MHW extent on the same day is now smaller due to the higher threshold, there are less MHW locations to predict. Unlike the 90th percentile MHW threshold, the 95th percentile MHW predictions showed large improvement when using cluster centroid data across all models.

These findings are further supported by the ROC and PR curves, where random forest exhibits consistently better AUC values across both MHW percentile thresholds and data balancing techniques (Figures 3.1 through 3.4 for 90th percentile MHW and A.22 through A.25 for 95th percentile MHW). After considering all model evaluation techniques, the random forest model demonstrates superior performance in predicting the presence or absence of MWHs at both the 90th and 95th percentiles.

4.2 Multi-Class MHW Classification

In the second part of the study, we evaluate the predictive capabilities of the five models in forecasting MHW classes within the northeast Pacific Ocean. Using the same predictor variables as the binary MHW classification (wind speed, net heat flux, sea level pressure, surface air temperature) we find that all models have difficulty correctly categorizing MHWs for both the 2-class and 4-class MHWs.

Figures A.54 and 3.11a demonstrate the accuracy for various lag times for each model using 2-class random balanced and cluster centroid balanced data, respectively. The highest performing model, the random forest model, has the highest average accuracy of just under 0.25 for lag 0 (Figure 3.11a). All other models perform similarly poor, with average accuracies ranging from 0.25 to less than 0.10.

Similar to the binary MHW predictions, there are improvements with each model when just assessing hit rate through various lags. Figures A.55 and 3.11b show the hit rate across each model through time, with again the highest score occurring with the random forest model at lag 0 using cluster centroid balanced data (hit rate = 0.70). Overall hit rate is relatively low with random balanced data, with hit rates ranging from 0.42 to 0.25 for all models except feedforward neural network, which exhibits sporadic and unreliable beha for the multi-class MHW predictions (Figure A.55). While average hit rate is still relatively low, we do see an improvement in overall hit rate when using cluster centroid balanced data, with hit rate ranging from 0.70 to 0.32 (Figure 3.11b). Hit rate scores are substantially higher than the average accuracy indicating that the models perform better at correctly predicting MHW classes than no MHW. The associated ROC and PR curves further conclude the inability of the models to accurately predict MHW classes, with even the best model, random forest, demonstrating poor ROC and PR AUC scores (Figures A.52, 3.9, A.53, and 3.10).

Extending the study to forecasting 4-class MHWs results in similarly poor model performance as in the 2-class MHW results. Figures A.71 and A.72 display the accuracy for lags ranging from 1 to 2 weeks. Again, we see a poor average accuracy among all models and both data balancing techniques. Four-class MHW average accuracy is notably lower than 2-class, with the highest accuracy occurring at lag 0 with the random forest model using cluster centroid balanced data with an accuracy of 0.18 (Figure A.72). Although the highest average accuracy is 0.18, the majority of the accuracies range from 0.08 to 0.04, further emphasizing the inability of the models to correctly predict the category of MHW (Figures A.71 and A.72).

Similar to the 2-class lag results, we again see an improvement in average hit rate for 4-class MHW compared to the average accuracy. Figures A.73 and A.74 represent the hit rate over lagged times for random and cluster centroid balanced data, respectively (Figures A.73 and A.74). While the general model performance structure holds true (random forest performing the best and logistic regression performing the poorest), we see an notable improvement in the hit rate when using the cluster centroid balanced data specifically for the random forest model over the lagged time. Although the accuracy and hit rate across all models is relatively poor, again, we see the random forest emerging as the leader in both the accuracy and hit rate performance metrics.

We represented the average hit rate over time across all models at each grid cell to provide a visual into the spatial hit rate of each model (Figures A.56 through A.64 and 3.12 for 2-class, Figures A.75 through A.84 for 4-class). Across all models, we see an improvement in spatial hit rate when using cluster centroid data, with random forest and gradient boosting classifier exhibiting the most cohesive spatial extent hit rate with 4-class MHW data.

4.3 Final Model

This study explored the forecasting potential of selected atmospheric variables, including wind speed, net heat flux, sea level pressure, and surface air temperature, in predicting MHWs and sought to identify the most effective models for this purpose. While it has been well documented and studied that lagged SSTs are the most popular input variable for forecasting SSTs, our investigation aimed to assess the predictive performance of other atmospheric variables, excluding SSTs [20]. Thus, the majority of the study lacks arguably one of the most important predictor variables: lagged SSTs. To address this issue, the best performing model, the random forest model, was selected to include more predictor variables as it is a non-parametric model that can handle correlation among predictors.

After evaluating previous model configurations, as stated previously, the random forest emerged as the top-performing model. Subsequently, a final model was tailored specifically using configurations that were showcased as resulting in the best performance: the 90^{th} percentile binary MHW event using cluster centroid balanced data. To enhance final model's performance, we incorporated additional predictor variables including sea surface temperature, along with temporal and spatial data. Consequently, the final model comprises a comprehensive set of seven variables encompassing spatial, temporal, atmospheric, and oceanic dimensions: location, day of year, sea surface temperature, wind speed, net heat flux, sea level pressure, and surface air temperature. The final model is able to predict the presence or absence of a MHW event with high accuracy and hit rates at lead times of 1 and 2 weeks in the region (Figures 3.13 and 3.15). The final model heavily relies on lagged SST as a predictor, while the remaining predictors are comparatively less important (Figure 3.16).

Random forests are a popular machine learning method as they are fast to implement, handle imbalanced datasets and complex relationship, generally work well “out of the box”, and are robust against overfitting [11]. The superiority of random forest over logistic regression, naive Bayes, gradient boosting, and feedforward neural networks for predicting marine heatwaves with the selected predictor variables may be attributed to several key factors. Random forest are known to perform well and capture non-linear relationships and complex interactions among predictors, which here is particularly advantageous in the context of MHW prediction, where the relationships between atmospheric and oceanic variables predicting a MHW are often complex and non-linear [11]. Random forests perform feature selection by evaluating the importance of each predictor variable, allowing it to identify the most relevant features for MHW prediction more effectively than the other models evaluated in this study [11]. Further, random forest's ensemble learning approach, which combines multiple decision trees, offers increased complexity and flexibility allowing it to capture complex patterns in the data more effectively [7].

While the final model demonstrates superior performance, it is important to note that the 90th percentile MHW random forest model achieved commendable accuracy in predicting MHW presence or absence using only wind speed, net heat flux, sea level pressure, and surface air temperature as predictors. This suggests that these selected oceanic and atmospheric variables serve as robust indicators of upcoming MHWs (Figures 3.5, 3.6, and 3.17). While sea surface temperature is the strongest predictor variable, wind speed, sea level pressure, net heat flux and surface air temperature together are able to predict with good accuracy (Figures 3.16 and 3.17).

4.4 Implications and Future Research

In this study, we conducted a thorough evaluation of binary and multiclass MHW predictions. Consistent with our previous discussions, a clear pattern emerges in the performance of the classification models. The random forest consistently achieves the highest overall performance, followed by the naive Bayes, gradient boosting, feedforward neural network, and logistic regression, ranked in descending order of performance. Notably, employing cluster centroid balanced data yields superior performance. This enhancement can be attributed to the balancing technique's ability to select more representative data samples, consequently ensuring that the models are trained on more comprehensive and meaningful sampled data. The ability of the random forest model, and by close proximity the naive Bayes model, in predicting MHWs given the selected predictor variables signifies that MHWs can be predicted with moderate accuracy using wind speed, net heat flux, sea level pressure and surface air temperature and very good accuracy when additional predictor variables are included in the final random forest model.

Previous research using monthly averages of SSTs have proven skillful at predicting MHWs in the global ocean 1 to 12 months in advance using a large multi-model ensemble of global climate forecasts [23]. In addition, other research has developed a deep learning time series prediction model (Unet-LSTM) that uses SST and air temperature to predict SST variability at various lead times using monthly mean values to forecast up to 2 years [45]. It has been noted that monthly forecasts are more useful to capture seasonal variability of SST in the global ocean, whereas daily resolution forecasting is better suited for short-term forecasting [45, 23]. Downscaling global forecasts to regional forecasts "may provide enhanced skill for specific areas" to supplement the global forecasting model [23].

Other research has addressed forecasting SST extremes on a localized region and a smaller timescale. Such research used previous SSTs and forecast atmospheric temperature to predict SST extremes in Chesapeake Bay, USA, using a 35-day probabilistic forecast [40]. The Chesapeake Bay study found that the model is skillful at predicting SST extremes with lead times ranging from 1-2 weeks using two predictors (SST and air temperature) as precursors [40]. Further studies have proven promising with the use of a random forest model predicting the presence or absence and MHW using spatial, temporal and climate variables known to impact SSTs with an accuracy of 76% at weekly time leads [19]. This study further assessed predictive power using the random forest model to predict the category of MHW as either no event, moderate, strong or severe/extreme, but forecasting accuracy dropped to 38% at weekly time leads [19]. An additional localized case study in the Mediterranean Sea implemented machine learning models including random forests, long short-term memory and convolution neural network, using lagged SSTs and selected atmospheric variables as predictors, found that all models were able to predict MHWs at a lead of 1 week with at least 50% confidence [6]. Another study focused on data-driven modeling of SSTs with in-situ observations and demonstrated that machine learning models can predict SSTs with good accuracy at a fraction of the computational cost as physics-based models on a global scale [50]. In contrast to the present study, this study implemented various machine learning algorithms to predict MHWs (including random forest, generalized additive models, extreme gradient boosting) and found that some models work better in particular regions, demonstrating that no one model fits all regions [50].

Additional studies have investigated the drivers and influences of MHWs on a regional basis. A study in Kuroshio-Oyashio Extension Region found that the driving factors of intense summer MHWs that occurred during 1999, 2008, 2012, and 2016 were primarily driven by by air-sea heat flux anomalies and reduced cloud cover, but were also influenced by region factors including the strengthened North Pacific High system and the Philippine-Japan teleconnection [13]. Another regional study employed a convolutional neural network and found that the SST anomalies in the Indian Ocean Dipole (IOD) region could be predicted up to 6 months in advance [15]. Notably, the study investigates the potential causes of anomalous events in the

IOD, which can arise from teleconnections with the equatorial Pacific, including evolving El Niño events, as well as cooling phenomena along the Australian coast [15].

The research outlined in this paper extends previous research by providing a comprehensive evaluation of several models in their ability to predict MHWs using selected atmospheric and oceanic variables in the northeast Pacific Ocean. The final model, the random forest model with additional predictor variables including SST, temporal and spatial variables, is able to predict the presence and absence of MHWs with high accuracy for lag times ranging from 1 to 2 weeks, adding to the research in short term MHW forecasting (Figures 3.13a and 3.13b).

The ability to predict MHWs on sub-seasonal timescales is becoming increasingly crucial for mitigating associated risks. Sea surface temperature extremes can influence coastal flooding and disrupt fisheries, highlighting the critical need for accurate MHW predictions [12]. Effective MHW forecasting "can help minimize disruptions to everyday public and commercial activity, keep coastal and maritime workers safe, and aid marine conservation efforts" [12]. Additionally, forecasting allows for timely adjustments in commercial fisheries management plans and strategies, alleviating the impact on businesses and resources [12]. These forecasts are essential for proactive planning and response, ultimately safeguarding both economic and environmental interests.

Future work should encompass several avenues for enhancement and refinement. There is a potential to incorporate neighboring effects into MHW predictions [50]. Additionally, the integration of more predictor variables such as mixed layer depth and water column stability could be explored, with an acknowledgment of potential regional variations in the relevance of these variables. Another avenue for research could involve assessing the ability of the random forest model to predict MHWs on a global scale, considering that prediction accuracy may vary across different regions. There is also potential to explore the utilization of higher resolution data to further improve prediction accuracy and capture finer-scale phenomena. These potential future directions could contribute to advancing the understanding and predictive capabilities in the field of MHW modeling.

Bibliography

[1] JJ Allaire. R Interface to 'TensorFlow', 2024.

[2] Andrew C Baker, Peter W Glynn, and Bernhard Riegl. Climate change and coral reef bleaching: An ecological assessment of long-term impacts, recovery trends and future outlook. *Estuar Coast Shelf Sci*, 80(4):435–471, Dec 2008.

[3] Viva Banzon, Thomas M. Smith, Michael Steele, Boyin Huang, and Huai-Min Zhang. Improved Estimation of Proxy Sea Surface Temperature in the Arctic. *Journal of Atmospheric and Oceanic Technology*, 37(2):341–349, February 2020.

[4] Jessica Benthuysen, Grant Smith, Claire Spillman, and Craig Steinberg. Subseasonal prediction of the 2020 great barrier reef and coral sea marine heatwave. *Environmental Research Letters*, 16, 12 2021.

[5] Nicholas A. Bond, Meghan F. Cronin, Howard Freeland, and Nathan Mantua. Causes and impacts of the 2014 warm anomaly in the ne pacific. *Geophysical Research Letters*, 42(9):3414–3420, 2015.

[6] G. Bonino, G. Galimberti, S. Masina, R. McAdam, and E. Clementi. Machine learning methods to predict sea surface temperature and marine heatwave occurrence: a case study of the mediterranean sea. *EGUsphere*, 2023:1–22, 2023.

[7] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[8] Letícia Cavole, Alyssa Demko, Rachel Diner, Ashlyn Giddings, Irina Koester, Camille Pagniello, May-Linn Paulsen, Arturo Ramírez-Valdez, Sarah Schwenck, Nicole Yen, Michelle Zill, and Peter Franks. Biological impacts of the 2013–2015 warm-water anomaly in the northeast pacific: Winners, losers, and the future. *Oceanography (Washington D.C.)*, 29, 06 2016.

[9] William WL Cheung and Thomas L Frölicher. Marine heatwaves exacerbate climate change impacts for fisheries in the northeast pacific. *Sci Rep*, 10(1):6678, Apr 2020.

[10] CSIRO, Alistair Hobday, Eric Oliver, Alex Sen Gupta, Jessica Benthuysen, Michael Burrows, Markus Donat, Neil Holbrook, Pippa Moore, Mads Thomsen, Thomas Wernberg, and Dan Smale. Categorizing and Naming Marine Heatwaves. *Oceanography*, 31(2), June 2018.

[11] A. Cutler, D.R. Cutler, and J.R. Stevens. Random forests. In C. Zhang and Y. Ma, editors, *Ensemble Machine Learning*. Springer, New York, NY, 2012.

[12] C. DeMott, Á. G. Muñoz, C. D. Roberts, C. M. Spillman, and F. Vitart. The benefits of better ocean weather forecasting. *Eos*, 102, 2021. Published on 12 November 2021.

[13] Yanzhen Du, Ming Feng, Zhenhua Xu, Baoshu Yin, and Alistair Hobday. Summer marine heatwaves in the kuroshio-oyashio extension region. *Remote Sensing*, 14:2980, 06 2022.

[14] C Mark Eakin, Jessica A Morgan, Scott F Heron, Tyler B Smith, Gang Liu, Lorenzo Alvarez-Filip, et al. Caribbean corals in crisis: Record thermal stress, bleaching, and mortality in 2005. *PLoS ONE*, 5(11):e13969, 2010.

[15] Ming Feng, Fabio Boschetti, Fenghua Ling, Xuebin Zhang, Jason Hartog, Mahmood Akhtar, Li Shi, Brint Gardner, Jing-Jia Luo, and Alistair Hobday. Predictability of sea surface temperature anomalies at the eastern pole of the indian ocean dipole—using a convolutional neural network model. *Frontiers in Climate*, 4, 08 2022.

[16] Python Software Foundation. Python programming language, 2022.

[17] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1), 2010.

[18] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

[19] Klea Giamalaki, Claudie Beaulieu, and Jason X Prochaska. Assessing predictability of marine heatwaves with random forests. *Geophys Res Lett*, 49(23), Dec 2022.

[20] M. Haghbin, A. Sharafati, and D. et al. Motta. Applications of soft computing models for predicting sea surface temperature: a comprehensive review and assessment. *Prog Earth Planet Sci*, 8(4), 2021.

[21] Alistair J. Hobday, Lisa V. Alexander, Sarah E. Perkins, Dan A. Smale, Sandra C. Straub, Eric C.J. Oliver, Jessica A. Benthuysen, Michael T. Burrows, Markus G. Donat, Ming Feng, Neil J. Holbrook, Pippa J. Moore, Hillary A. Scannell, Alex Sen Gupta, and Thomas Wernberg. A hierarchical approach to defining marine heatwaves. *Progress in Oceanography*, 141:227–238, February 2016.

[22] Yuguang Huang and Lei Li. Naive bayes classification algorithm based on small sample set. In *2011 IEEE International Conference on Cloud Computing and Intelligence Systems*, pages 34–39, 2011.

[23] Michael G Jacox, Michael A Alexander, Diana Amaya, Emily Becker, Steven J Bograd, Stephanie Brodie, et al. Global seasonal forecasts of marine heatwaves. *Nature*, 604(7906):486–490, Apr 2022.

[24] George H. John and Pat Langley. Estimating continuous distributions in bayesian classifiers. In Philippe Besnard and Steve Hanks, editors, *UAI*, pages 338–345. Morgan Kaufmann, 1995.

[25] Tomasz Kalinowski. R Interface to 'Keras', August 2023.

[26] E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, A. Leetmaa, R. Reynolds, Roy Jenne, and Dennis Joseph. The NCEP/NCAR 40-year reanalysis project.

[27] Muhammad Yaseen Khan, Abdul Qayoom, Muhammad Suffian Nizami, Muhammad Shoaib Siddiqui, Shaukat Wasi, and Syed Muhammad Khaliq-ur-Rahman Raazi. Automated Prediction of Good Dictionary EXamples (GDEX): A Comprehensive Experiment with Distant Supervision, Machine Learning, and Word Embedding-Based Deep Learning Techniques. *Complexity*, 2021:1–18, September 2021.

[28] Max Kuhn. Building Predictive Models in R Using the caret Package, 2008.

[29] Vipin Kumar, editor. *Introduction to parallel computing: design and analysis of algorithms*. Benjamin/Cummings Pub. Co, Redwood City, Calif, 1994.

[30] Charlotte Laufkötter, Jakob Zscheischler, Thomas L. Frölicher, et al. High-impact marine heatwaves attributable to human-induced global warming. *Science*, 369:1621–1625, 2020.

[31] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[32] Wiener M Liaw A. Classification and Regression by randomForest, 2002.

[33] Michal Majka. High Performance Implementation of the Naive Bayes Algorithm, March 2024.

[34] Ryan M. McCabe, Barbara M. Hickey, Raphael M. Kudela, Kathi A. Lefebvre, Nicolaus G. Adams, Brian D. Bill, Frances M. D. Gulland, Richard E. Thomson, William P. Cochlan, and Vera L. Trainer. An unprecedented coastwide toxic algal bloom linked to anomalous ocean conditions. *Geophysical Research Letters*, 43(19):10,366–10,376, 2016.

[35] David Meyer. Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien, December 2023.

[36] Brett Molony, Damian Thomson, and Ming Feng. What can we learn from the 2010/11 western australian marine heatwave to better understand risks from the one forecast in 2020/21? *Frontiers in Marine Science*, 8:645383, 02 2021.

[37] Arnaud Nguembang Fadja, Evelina Lamma, and Fabrizio Riguzzi. Vision inspection with neural networks. 12 2018.

[38] Philipp Probst, Marvin Wright, and Anne-Laure Boulesteix. Hyperparameters and Tuning Strategies for Random Forest. *WIREs Data Mining and Knowledge Discovery*, 9(3), May 2019. arXiv:1804.03515 [cs, stat].

[39] Zijian Qiu, Fangli Qiao, Chan Joo Jang, Lujun Zhang, and Zhenya Song. Evaluation and projection of global marine heatwaves based on CMIP6 models. *Deep Sea Research Part II: Topical Studies in Oceanography*, 194:104998, 2021.

[40] Alanna C Ross and Charles A Stock. Probabilistic extreme sst and marine heatwave forecasts in chesapeake bay: A forecast model, skill assessment, and potential value. *Front Mar Sci*, 9:896961, Oct 2022.

[41] Alex Sen Gupta, Mads Thomsen, Jessica A Benthuysen, et al. Drivers and impacts of the most extreme marine heatwave events. *Sci Rep*, 10(1):19359, Nov 2020.

[42] Dan A Smale, Thomas Wernberg, Eric CJ Oliver, Mads Thomsen, Ben P Harvey, Susanne C Straub, et al. Marine heatwaves threaten global biodiversity and the provision of ecosystem services. *Nat Clim Change*, 9(4):306–312, Apr 2019.

[43] Di Sun, Zhao Jing, and Hailong Liu. Deep learning improves sub-seasonal marine heatwave forecast. *Environmental Research Letters*, 2024.

[44] J. Kenneth Tay, Balasubramanian Narasimhan, and Trevor Hastie. Elastic Net Regularization Paths for All Generalized Linear Models. *Journal of Statistical Software*, 106(1), 2023.

[45] James Taylor and Ming Feng. A deep learning model for forecasting global

monthly mean sea surface temperature anomalies. *Front Clim*, 4:932932, Sep 2022.

[46] R Core Team. R: A language and environment for statistical computing., 2022.

[47] Jun Wang and B. Malakooti. A feedforward neural network for multiple criteria decision making. *Computers & Operations Research*, 19(2):151–167, 1992.

[48] Hadley Wickham, Mara Averick, Jennifer Bryan, Lucy D'Agostino McGowan, Romain François, Alex Hayes, Lionel Henry, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, Winston Chang, Garrett Grolemund, Jim Hester, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse, 2019.

[49] Hadley Wickham and Max Kuhn. Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles., 2020.

[50] Stefan Wolff, Fearghal O'Donncha, and Bei Chen. Statistical and machine learning ensemble modelling to forecast sea surface temperature. *Journal of Marine Systems*, 208:103347, 05 2020.

[51] Ziegler A Wright MN. Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R, 2017.

[52] Harry Zhang. The optimality of naive bayes. *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2004*, 2, 01 2004.

[53] Tao Zhang, Wuyin Lin, Andrew M. Vogelmann, Minghua Zhang, Shaocheng

Xie, Yi Qin, and Jean-Christophe Golaz. Improving Convection Trigger Functions in Deep Convective Parameterization Schemes Using Machine Learning. *Journal of Advances in Modeling Earth Systems*, 13(5):e2020MS002365, May 2021.

Appendix A

Appendix

A.1 90$^{\text{th}}$ Percentile MHW

A.1.1 Balanced Data

To illustrate the models' ability to forecast the presence or absence of a MHW event, spatial plots were generated for each model. These plots depict the observed MHW occurrences on a specific day alongside the model predictions utilizing lagged data at intervals of 7 and 14 days, where the presence of a MHW is indicated by a black dot and the data is superimposed onto observed SST anomalies for the specified day. This analysis encompassed both data balancing techniques, providing insights into the predictive capacity of each model across different temporal contexts and balancing strategies. Figures A.1 through A.10 provide insights into the predictive capabilities of each model regarding the presence or absence of a MHW on August 18, 2015 — a day characterized by a notable frequency of MHW occurrences.

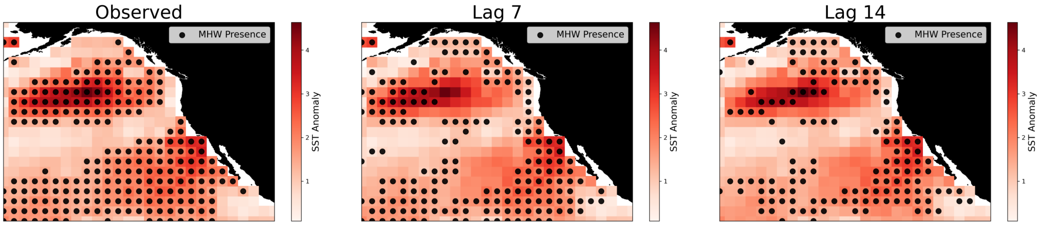


Figure A.1: Logistic regression model forecasts for the 90th percentile MHW using random balanced data for August 18, 2015, with no lag, 7 days lag and 14 days lag. The color bar represents the observed sea surface temperature anomaly and dots indicate where a MHW was correctly predicted.

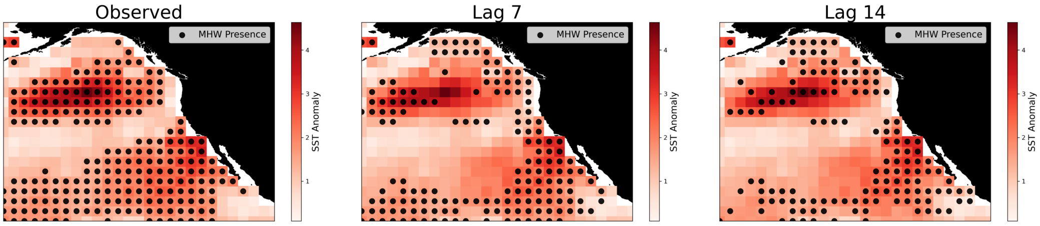


Figure A.2: Logistic regression model forecasts for the 90th percentile MHW using cluster centroid balanced data for August 18, 2015, with no lag, 7 days lag and 14 days lag. The color bar represents the observed sea surface temperature anomaly and dots indicate where a MHW was correctly predicted.

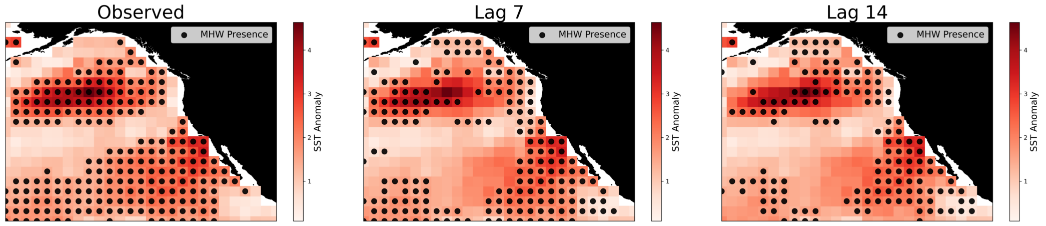


Figure A.3: Naive Bayes model forecasts for the 90th percentile MHW using random balanced data for August 18, 2015, with no lag, 7 days lag and 14 days lag. The color bar represents the observed sea surface temperature anomaly and dots indicate where a MHW was correctly predicted.

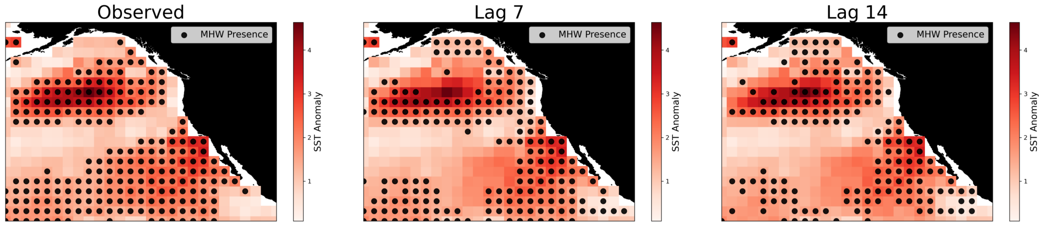


Figure A.4: Naive Bayes model forecasts for the 90th percentile MHW using cluster centroid balanced data for August 18, 2015, with no lag, 7 days lag and 14 days lag. The color bar represents the observed sea surface temperature anomaly and dots indicate where a MHW was correctly predicted.

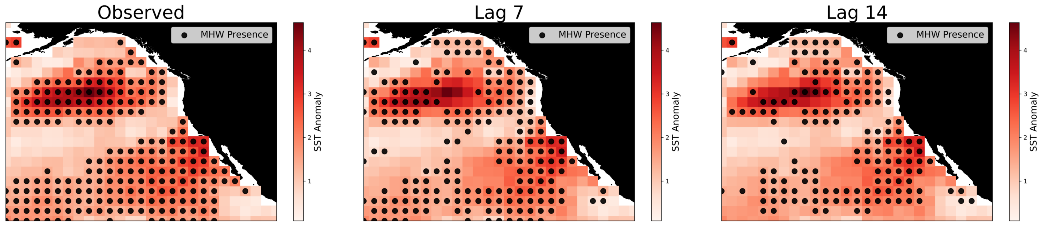


Figure A.5: Gradient boosting model forecasts for the 90th percentile MHW using random balanced data for August 18, 2015, with no lag, 7 days lag and 14 days lag. The color bar represents the observed sea surface temperature anomaly and dots indicate where a MHW was correctly predicted.

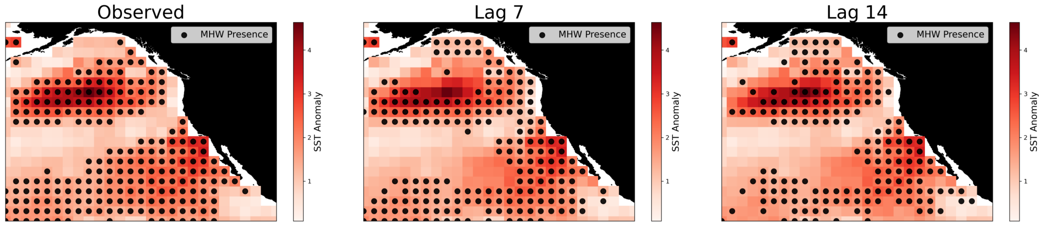


Figure A.6: Gradient boosting model forecasts for the 90th percentile MHW using cluster centroid balanced data for August 18, 2015, with no lag, 7 days lag and 14 days lag. The color bar represents the observed sea surface temperature anomaly and dots indicate where a MHW was correctly predicted.

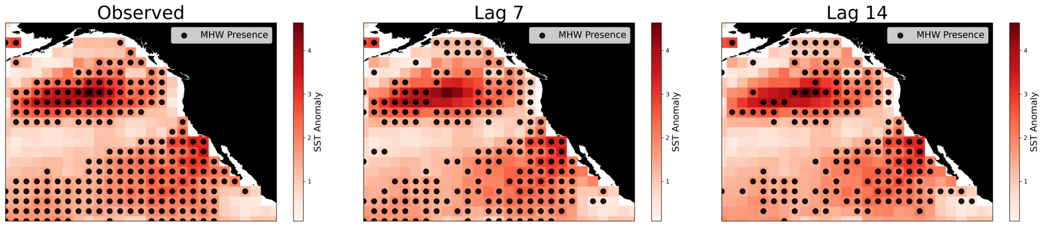


Figure A.7: Random Forest model forecasts for the 90$^{\text{th}}$ percentile MHW using random balanced data for August 18, 2015, with no lag, 7 days lag and 14 days lag. The colorbar represents the observed sea surface temperature anomaly and dots indicate where a MHW was correctly predicted.

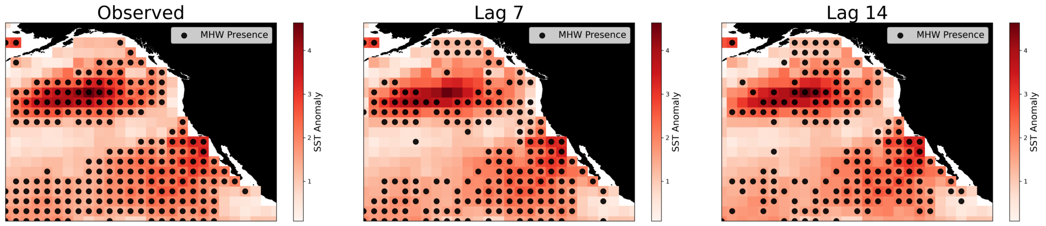


Figure A.8: Random forest model forecasts for the 90$^{\text{th}}$ percentile MHW using cluster centroid balanced data for August 18, 2015, with no lag, 7 days lag and 14 days lag. The color bar represents the observed sea surface temperature anomaly and dots indicate where a MHW was correctly predicted.

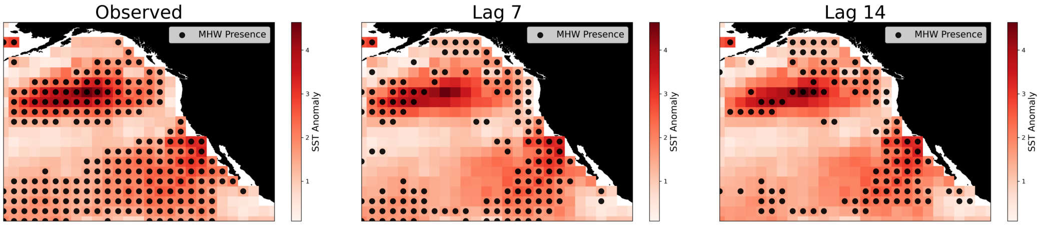


Figure A.9: Feedforward neural network model forecasts for the 90th percentile MHW using random balanced data for August 18, 2015, with no lag, 7 days lag and 14 days lag. The color bar represents the observed sea surface temperature anomaly and dots indicate where a MHW was correctly predicted.

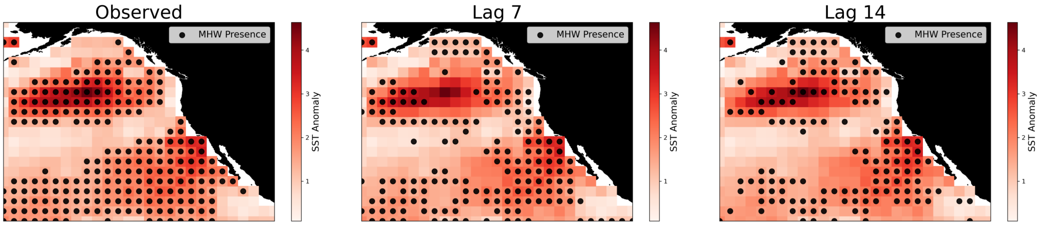


Figure A.10: Feedforward neural network model forecasts for the 90th percentile MHW using cluster centroid balanced data for August 18, 2015, with no lag, 7 days lag and 14 days lag. The color bar represents the observed sea surface temperature anomaly and dots indicate where a MHW was correctly predicted.

Notable trends represented in the spatial plots is the inability of the logistic regression model and feedforward neural network to capture spatial patters on the given day, though there is slight improvement in prediction in the feedforward neural network when using cluster centroid balanced data. While the naive bayes an gradient boosting models demonstrate moderate predictive power at 1 week and 2 week lead times, the random forest model demonstrates the strongest ability to predict a MHW occurrence.

To further illustrate each model's ability to correctly predict MHW events, a spatial hit rate plot was generated for each model at both the 90th percentile MHW for both data balancing techniques. The hit rate is averaged across the entire testing dataset for each grid cell, resulting in a comprehensive spatial plot that illustrates the models' effectiveness in predicting MHWs where a higher hit rate (indicated by dark red here) represents a higher model performance. Figures A.11 through A.18 represent the spatial hit rate plots across all models and both balancing methods.

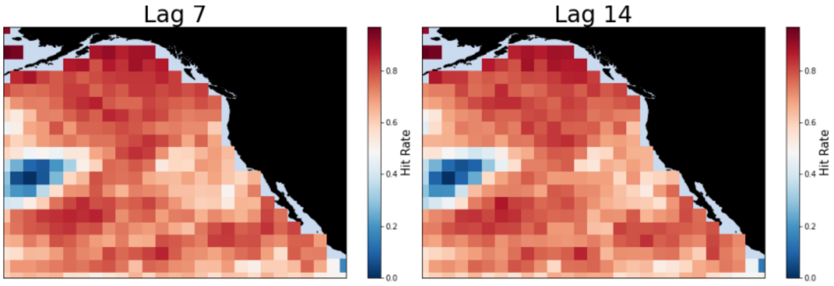


Figure A.11: Logistic regression hit rate spatial plot for binary 90th percentile MHW outcome with random balanced data.

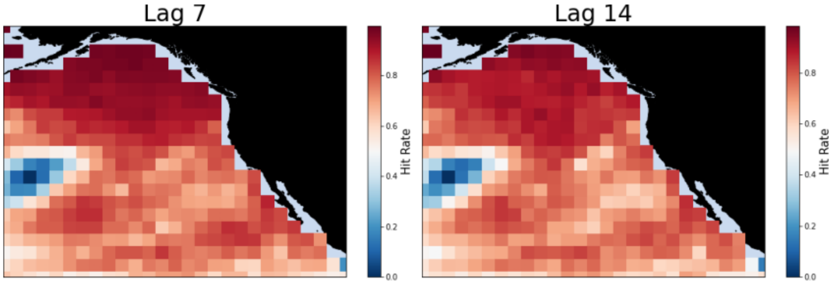


Figure A.12: Logistic regression hit rate spatial plot for binary 90th percentile MHW outcome with cluster centroid balanced data.

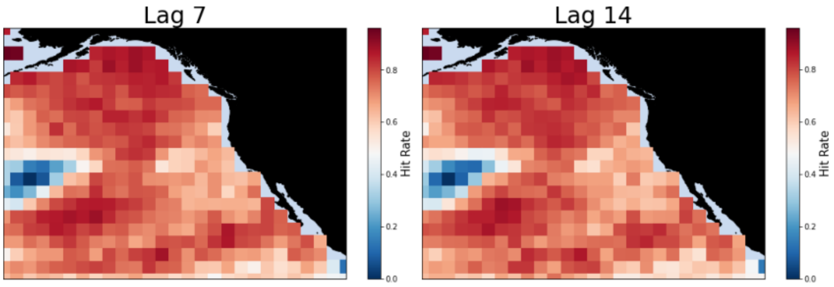


Figure A.13: Naive Bayes hit rate spatial plot for binary 90$^{\text{th}}$ percentile MHW outcome with random balanced data.

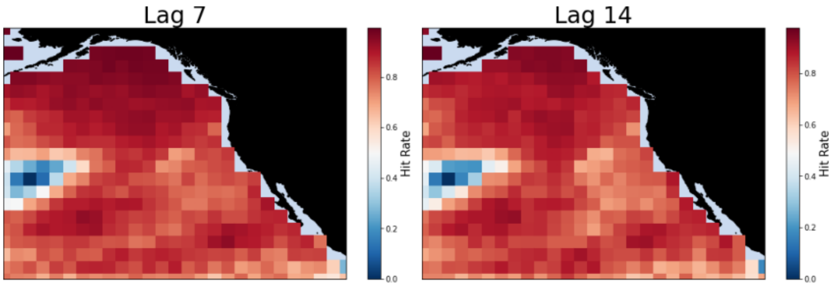


Figure A.14: Naive Bayes hit rate spatial plot for binary 90$^{\text{th}}$ percentile MHW outcome with cluster centroid balanced data.

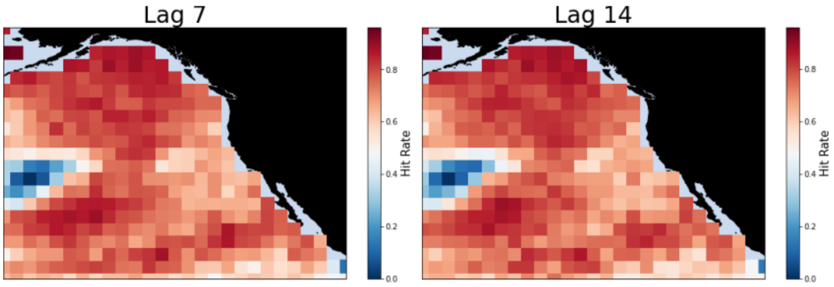


Figure A.15: Gradient boosting hit rate spatial plot for binary 90$^{\text{th}}$ percentile MHW outcome with random balanced data.

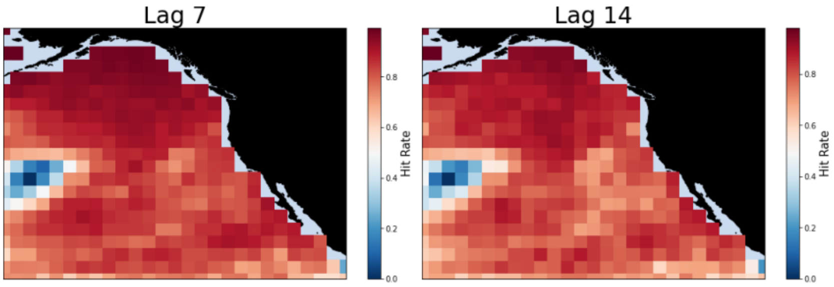


Figure A.16: Gradient boosting hit rate spatial plot for binary 90th percentile MHW outcome with cluster centroid balanced data.

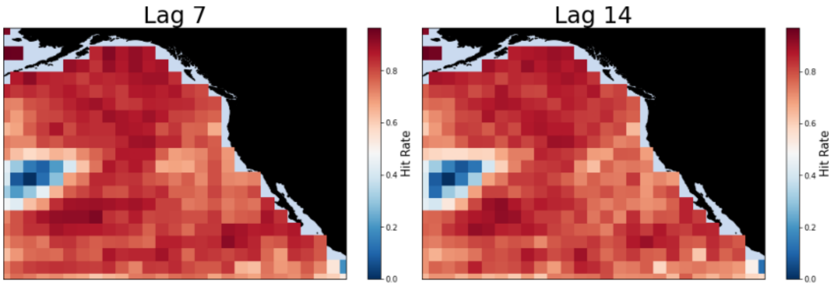


Figure A.17: Random Forest hit rate spatial plot for binary 90th percentile MHW outcome with random balanced data.

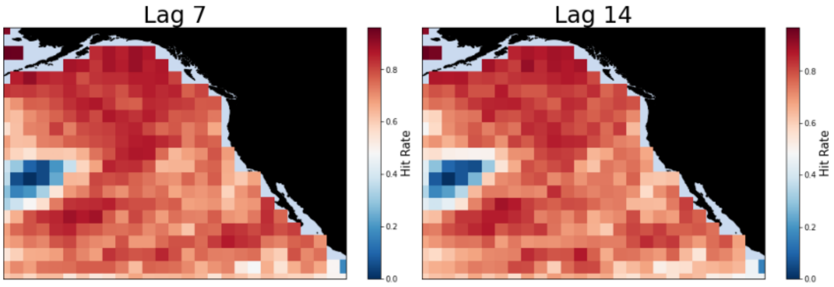


Figure A.18: Feedforward neural network hit rate spatial plot for binary 90th percentile MHW outcome with random balanced data.

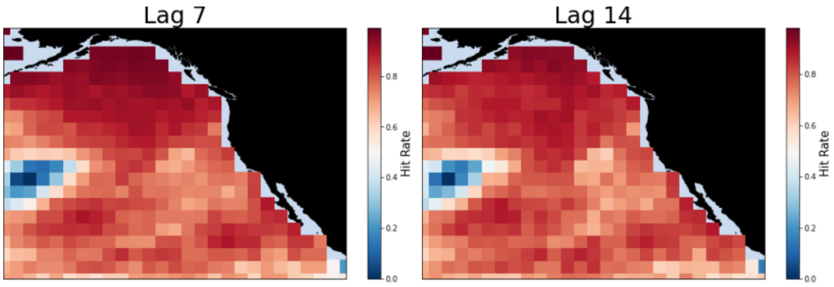


Figure A.19: Feedforward neural network hit rate spatial plot for binary 90th percentile MHW outcome with cluster centroid balanced data.

The hit rate spatial plots provide insight into the model's overall performance. Again, we are seeing similar patters that have been observed with supporting data: poor performance with logistic regression, moderate performance for naive bayes and gradient boosting, moderate to poor performance for feedforward neural network and high performance for random forest. Across all models, performance increases when using cluster centroid balanced data.

A.1.2 Unbalanced Data

The ROC curve (A.20) and PR curve (A.21) illustrate the 90th percentile MHW's model performance, revealing its inadequate performance. Class 0 dominating due to its majority, resulting in poor predictive performance for class 1, which is outweighed by the prevalence of class 0.

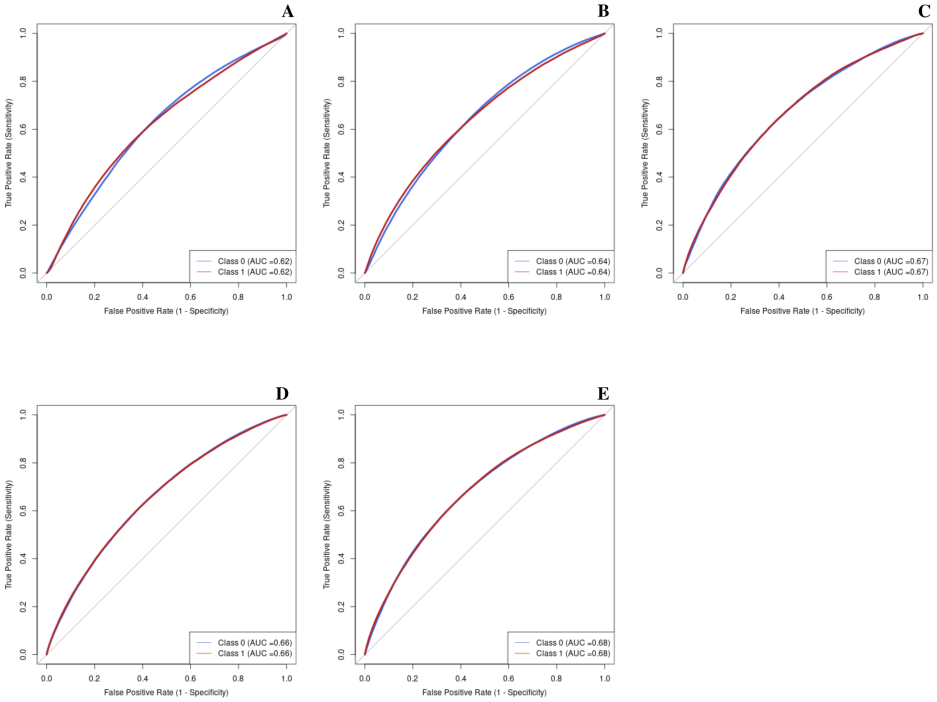


Figure A.20: ROC Curve for 90th Percentile MHW with Unbalanced Data.

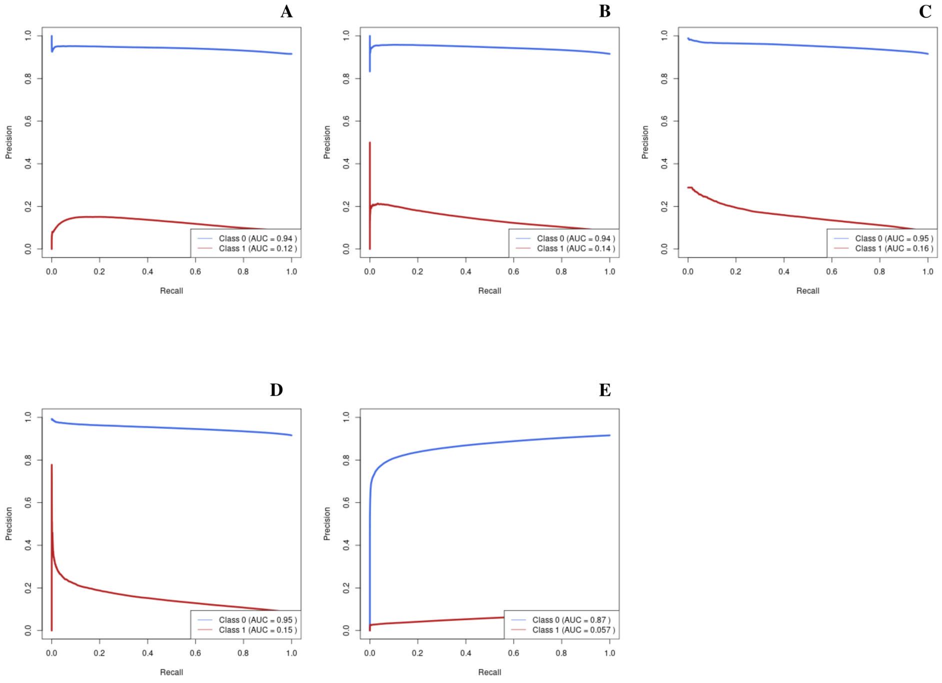


Figure A.21: PR Curve for 90th Percentile MHW with Unbalanced Data.

A.2 95th Percentile MHW

A.2.1 Balanced Data

Figure A.22 represents the ROC curve and associated AUC for each model trained on random balanced data for a binary MHW outcome at the 95th threshold. While all models perform better than random chance, Figure A.22D, representing the random forest model, performs significantly better than other models indicated by the ROC curve closely hugging the upper left corner, with an AUC score of 0.90 for class 0 and class 1. The logistic regression performed the worse (AUC = 0.57 for class 0 and class 1, Figure A.22A) while the remaining models Figure A.22 B, C and E had poor to moderate performance, all with AUC scores around 0.65 for class 0 and class 1.

The cluster centroid balanced data ROC curves, shown in Figure A.23, showed slight improvements compared to the random balanced data, with AUC score improvements across all models. The ROC curves for the cluster centroid balanced data, depicted in Figure A.23, demonstrated slight enhancements across all models in terms of AUC scores. Notably, the dominance of the random forest model (AUC = 0.94 for both class 0 and class 1, Figure A.23D), the under performance of logistic regression (AUC = 0.60 for both class 0 and class 1, Figure A.23A), and the intermediate performance of the remaining models (AUC ranging from 0.70 to 0.78 for both class 0 and class 1, Figure A.23 B, C, and E) were consistent.

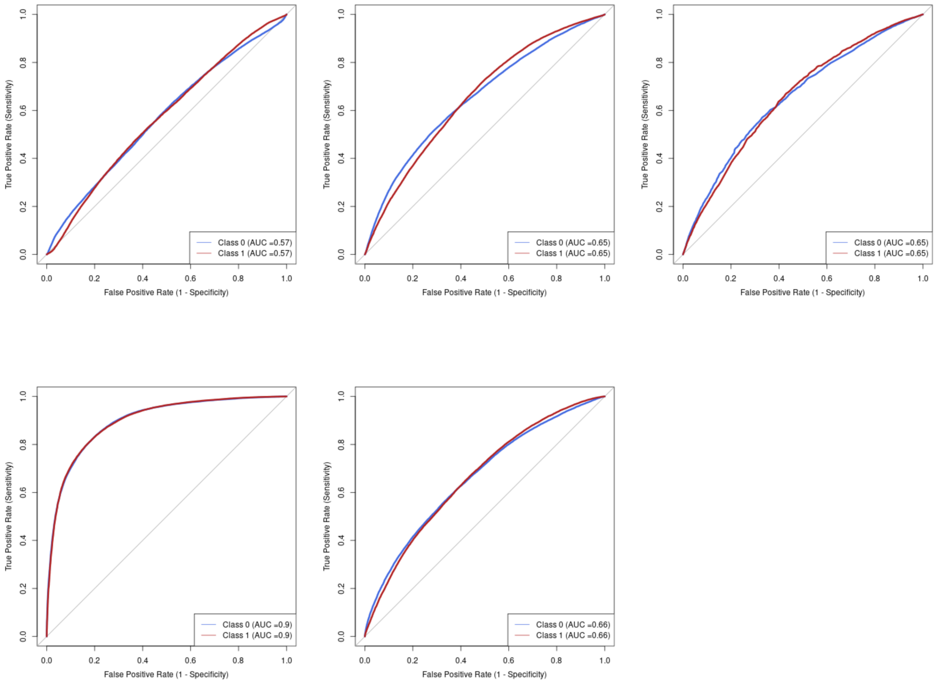


Figure A.22: ROC curve for the 95th percentile random balanced binary MHW prediction across all models: A) Logistic Regression, B) Naive Bayes, C) Gradient Boosting, D) Random Forest, E) Feedforward Neural Network.

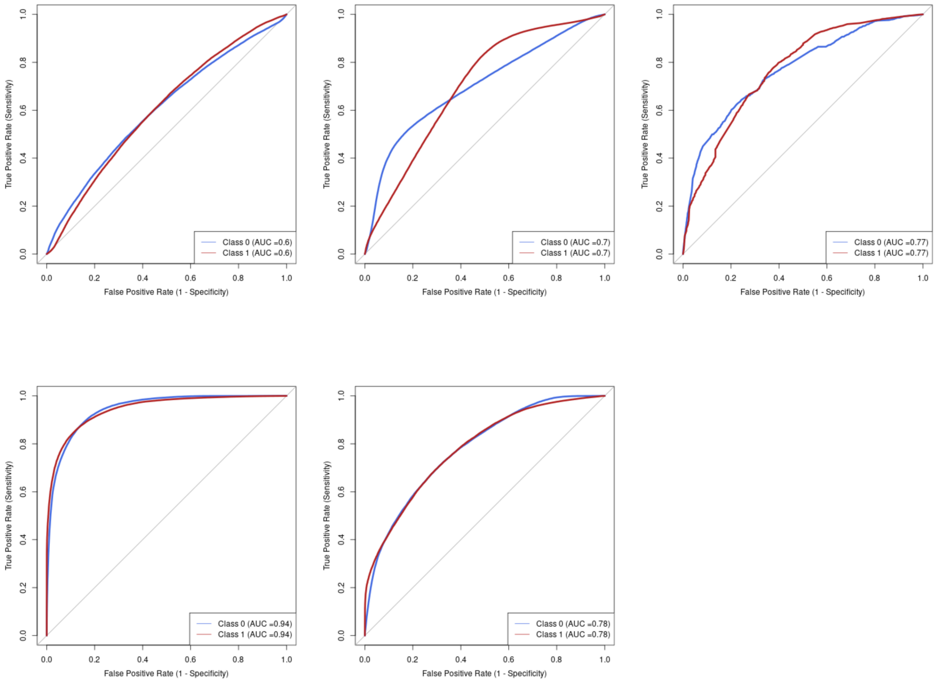


Figure A.23: ROC curve for the 95th percentile cluster centroid balanced binary MHW prediction across all models: A) Logistic Regression, B) Naive Bayes, C) Gradient Boosting, D) Random Forest, E) Feedforward Neural Network.

Figure A.24 displays the Precision-Recall (PR) curve and corresponding AUC values for all models for random balanced data, further contributing to model evaluation. Notably, the feedforward neural network model (AUC = 0.22 for class 0 and AUC = 0.60 for class 1, Figure A.24A) exhibits the poorest performance, while the random forest model (AUC = 0.80 for class 0 and AUC = 0.95 for class 1, Figure A.24D) demonstrates the highest performance. It's worth mentioning that across all models, the AUC score for class 0 is consistently lower than that for class 1, suggesting that in the balanced dataset, the classifiers tend to achieve higher precision but lower recall. Figure A.25 presents the PR curves for cluster centroid balanced data and exhibit similar performance to random balanced data

PR performance, with slight AUC improvements across both classes and all models. Notable across all PR curves is the poor performance of the feedforward neural network, rather than the baseline logistic regression model. It's worth noting that the ROC and PR curve results at the 95th percentile displayed analogous patterns to those at the 90th percentile. This suggests that despite raising the threshold for identifying a MHW, the models did not demonstrate significant adverse effects, as observed from the ROC and PR curves.

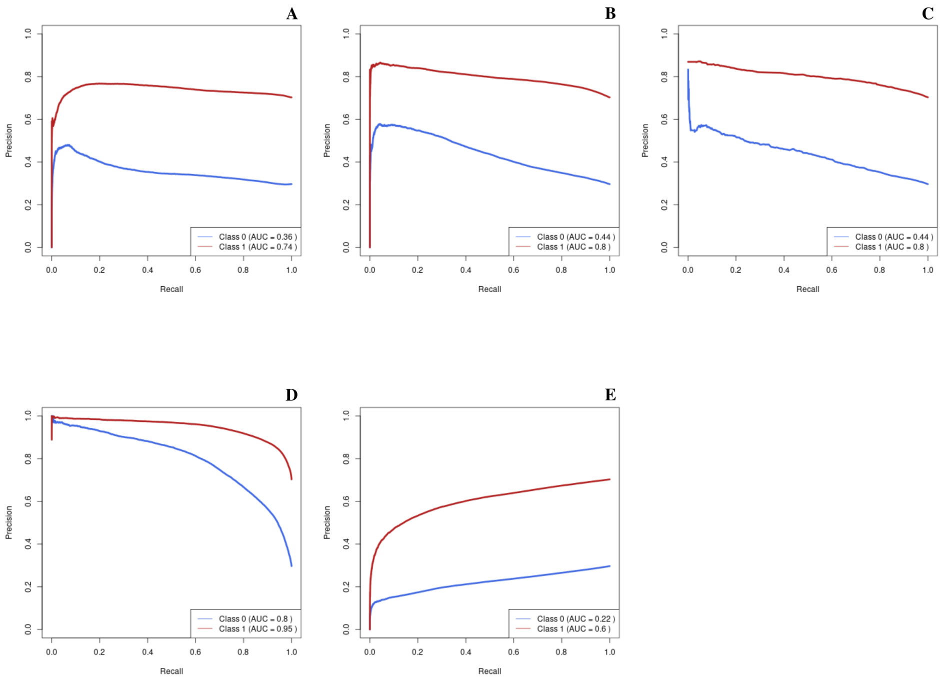


Figure A.24: PR curve for the 95th percentile random balanced binary MHW prediction across all models: A) Logistic Regression, B) Naive Bayes, C) Gradient Boosting, D) Random Forest, E) Feedforward Neural Network.

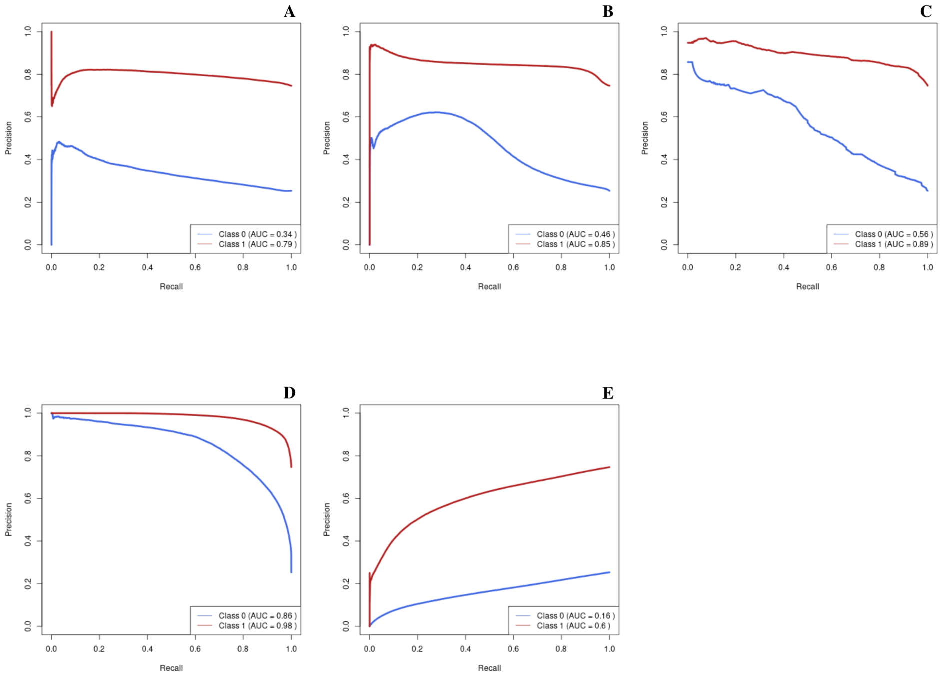


Figure A.25: PR curve for the 95th percentile cluster centroid balanced binary MHW prediction across all models: A) Logistic Regression, B) Naive Bayes, C) Gradient Boosting, D) Random Forest, E) Feedforward Neural Network.

To monitor model performance over time, we assessed accuracy and hit rate against lagged predictor variables. Figures A.26 and A.27 depict the accuracy of each model across various lag periods, ranging from 1 day to 2 weeks, using random balanced and cluster centroid balanced data, respectively. In both scenarios, lag 0 consistently demonstrated the highest accuracy across all models, with accuracy decreasing as the lag increased. While there was a slight improvement in accuracy observed with the cluster centroid balanced data, indicating its slightly superior performance, the improvement was not significant. Although the random forest model remained dominant in both scenarios, the use of cluster centroid balanced data notably enhanced the performance of the naive Bayes model, mirroring the pattern observed with 90th percentile MHW and making it comparable to the random forest model.

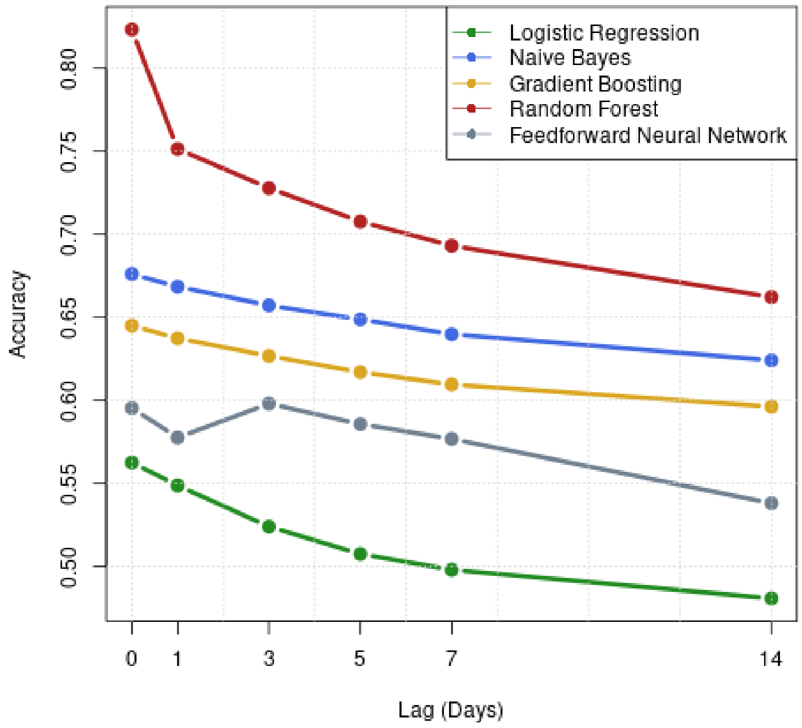


Figure A.26: Accuracy versus lagged predictor variables using random balanced binary 95th percentile MHW data across all models: Logistic Regression, Naive Bayes, Gradient Boosting, Random Forest and Feedforward Neural Network.

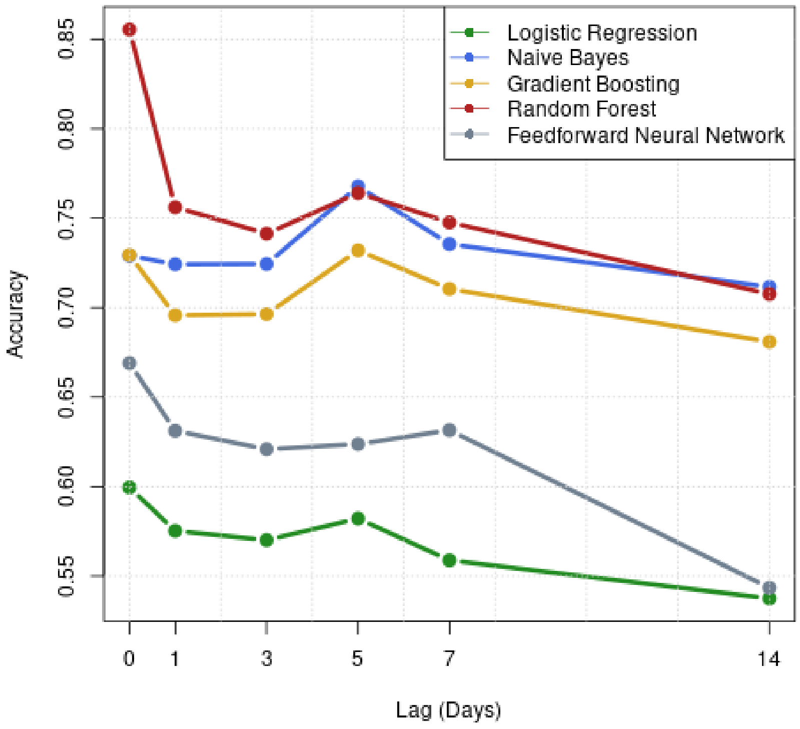


Figure A.27: Accuracy versus lagged predictor variables using cluster centroid balanced binary 95$^{\text{th}}$ percentile MHW data across all models: Logistic Regression, Naive Bayes, Gradient Boosting, Random Forest and Feedforward Neural Network.

To refine our model evaluation and concentrate on instances where models successfully predicted a MHW event, we constructed a plot illustrating the lag versus hit rate for the 95$^{\text{th}}$ percentile MHW. This plot encompasses lag periods ranging from 1 to 14 days for all models, utilizing both random balanced and cluster centroid balanced data. Refer to Figure A.28 for results with random balanced data and Figure A.29 for results with cluster centroid balanced data.

Similar to the accuracy versus lag plot, the lag versus hit rate plot reveals a decrease in hit rate with increased lag time. Across both balanced datasets, the random forest model consistently performed the best, followed by naive Bayes, gradient boosting, feedforward neural network and logistic regression. There was a slight improvement observed when using the cluster centroid balanced data. Across all models, the decline in hit rate for cluster centroid balanced data (Figure A.29) is slower than that of the random balanced data (Figure A.28), with notable enhancements observed in the naive Bayes and gradient boosting models. Similar to that of the results observed with the 90th, the overall scoring of the hit rate versus lag plot across all models was better than the accuracy versus lag plot, indicating that all models are better at predicting the presence of a MHW rather than the absence of a MHW with the balanced datasets.

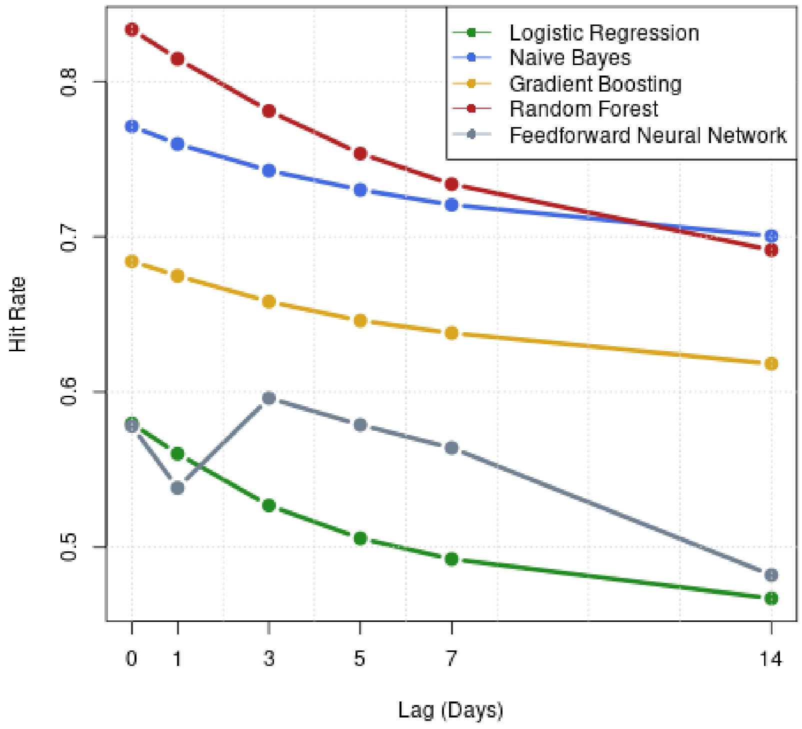


Figure A.28: Hit rate versus lagged predictor variables using random balanced binary 95th percentile MHW data across all models : Logistic Regression, Naive Bayes, Gradient Boosting, Random Forest and Feedforward Neural Network.

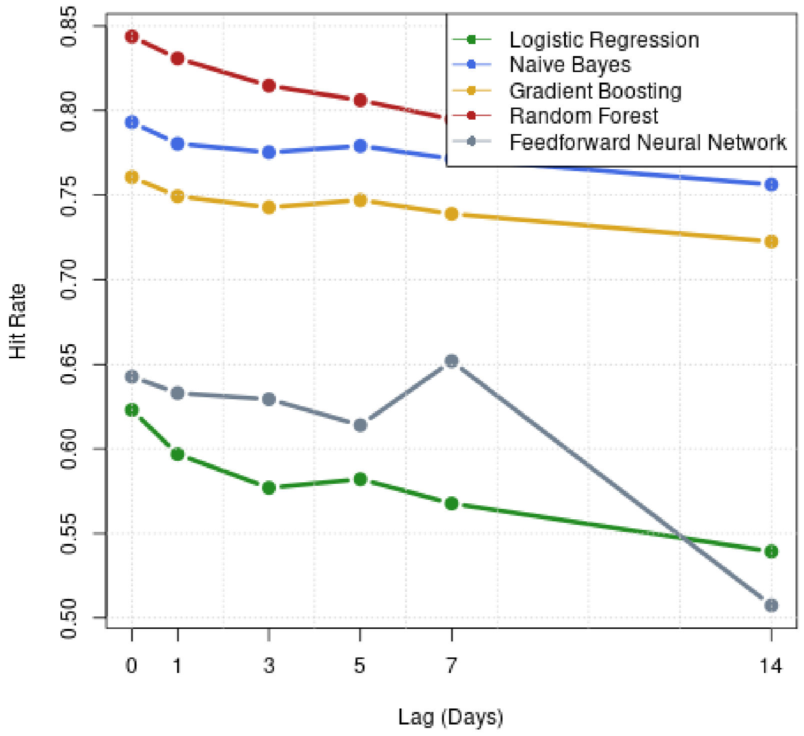


Figure A.29: Hit rate versus lagged predictor variables using cluster centroid balanced binary 95$^{\text{th}}$ percentile MHW data across all models : Logistic Regression, Naive Bayes, Gradient Boosting, Random Forest and Feedforward Neural Network.

Figures A.30 through A.39 represent the spatial accuracy in predicting the presence or absence of MHW across each model for both data balancing techniques on an arbitrarily selected day, August 18, 2015. Each figure represents the observed MHW data and the subsequent MHW predictions for each model using 1 week and 2 week lags.

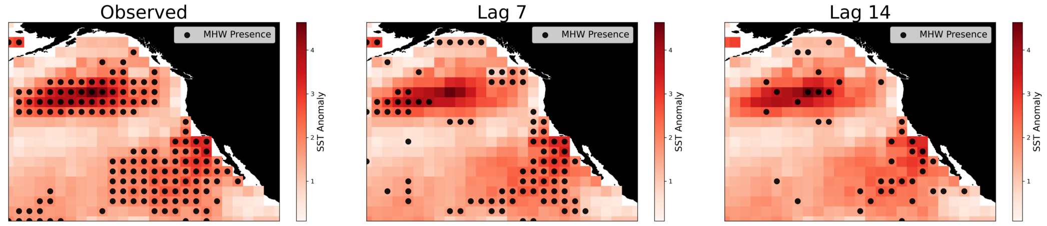


Figure A.30: Logistic regression model forecasts for the 95th percentile MHW using random balanced data for August 18, 2015, with no lag, 7 days lag and 14 days lag. The color bar represents the observed sea surface temperature anomaly and dots indicate where a MHW was correctly predicted

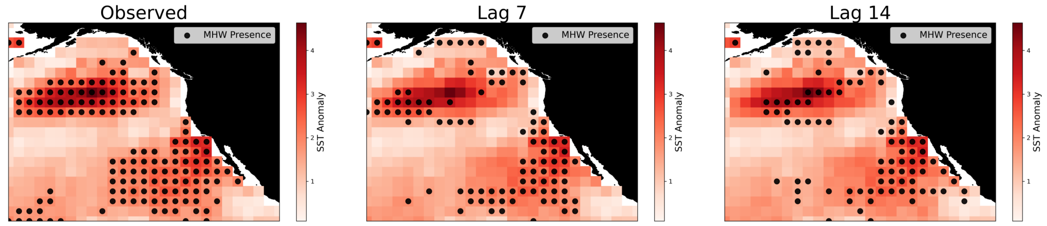


Figure A.31: Logistic regression model forecasts for the 95th percentile MHW using cluster centroid balanced data for August 18, 2015, with no lag, 7 days lag and 14 days lag. The color bar represents the observed sea surface temperature anomaly and dots indicate where a MHW was correctly predicted

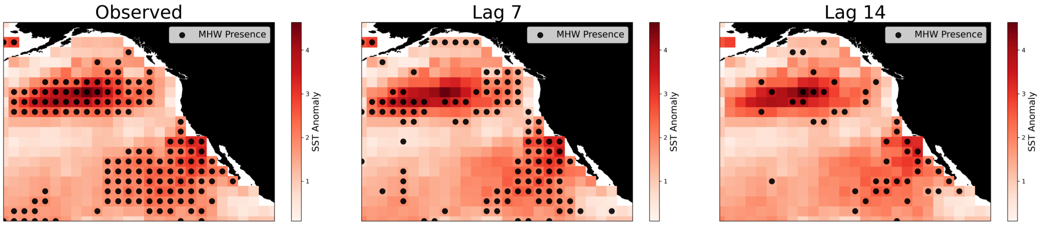


Figure A.32: Naive Bayes model forecasts for the 95th percentile MHW using random balanced data for August 18, 2015, with no lag, 7 days lag and 14 days lag. The color bar represents the observed sea surface temperature anomaly and dots indicate where a MHW was correctly predicted.

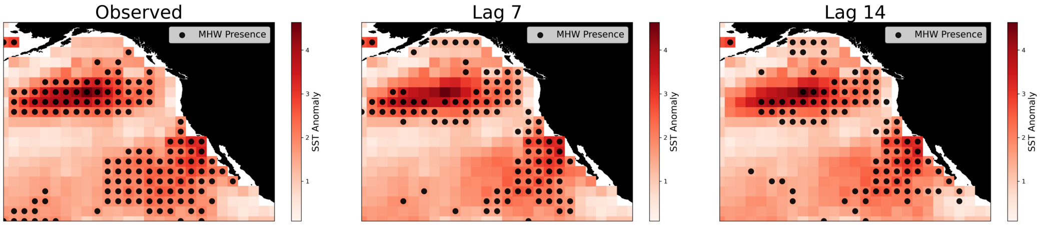


Figure A.33: Naive Bayes model forecasts for the 95th percentile MHW using cluster centroid balanced data for August 18, 2015, with no lag, 7 days lag and 14 days lag. The color bar represents the observed sea surface temperature anomaly and dots indicate where a MHW was correctly predicted.

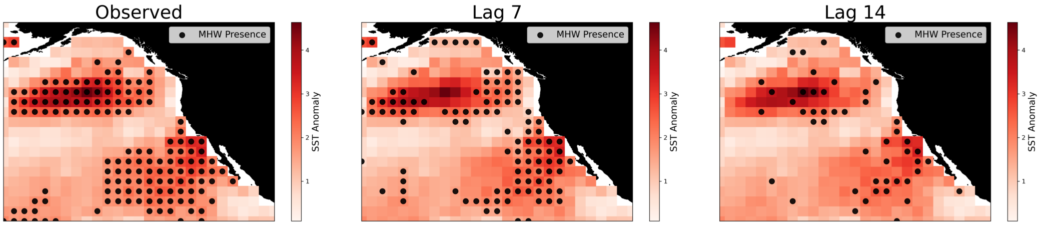


Figure A.34: Gradient boosting model forecasts for the 95$^{\text{th}}$ percentile MHW using random balanced data for August 18, 2015, with no lag, 7 days lag and 14 days lag. The color bar represents the observed sea surface temperature anomaly and dots indicate where a MHW was correctly predicted.

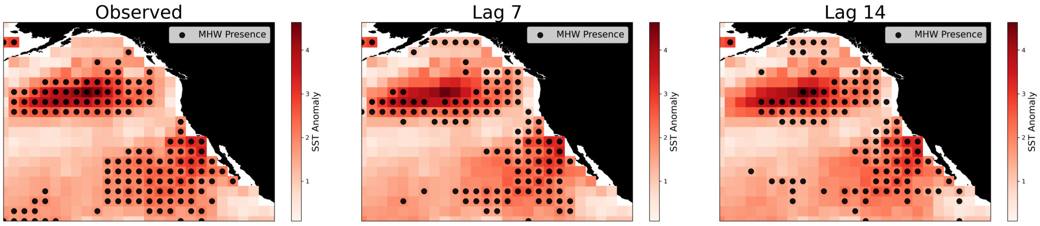


Figure A.35: Gradient boosting model forecasts for the 95$^{\text{th}}$ percentile MHW using cluster centroid balanced data for August 18, 2015, with no lag, 7 days lag and 14 days lag. The color bar represents the observed sea surface temperature anomaly and dots indicate where a MHW was correctly predicted.

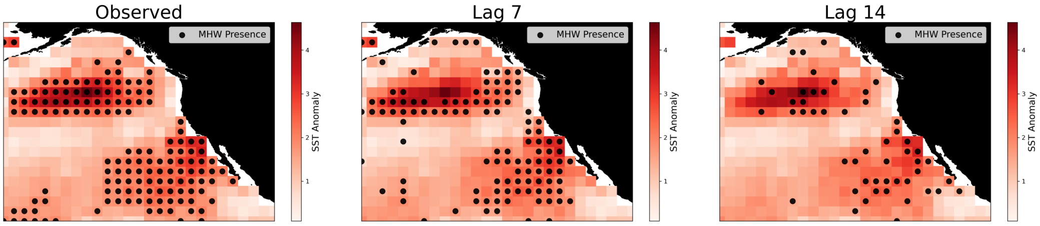


Figure A.36: Random Forest model forecasts for the 95th percentile MHW using random balanced data for August 18, 2015, with no lag, 7 days lag and 14 days lag. The colorbar represents the observed sea surface temperature anomaly and dots indicate where a MHW was correctly predicted.

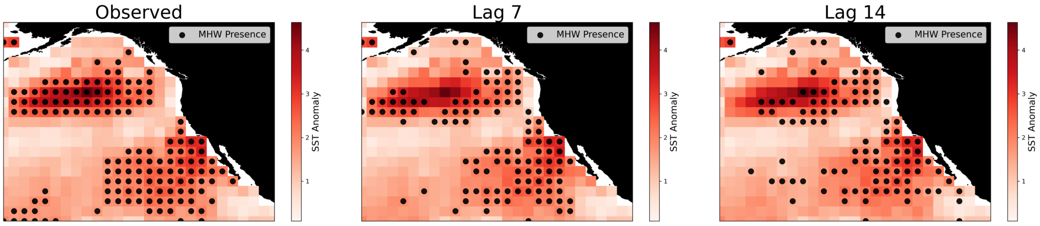


Figure A.37: Random forest model forecasts for the 95th percentile MHW using cluster centroid balanced data for August 18, 2015, with no lag, 7 days lag and 14 days lag. The color bar represents the observed sea surface temperature anomaly and dots indicate where a MHW was correctly predicted.

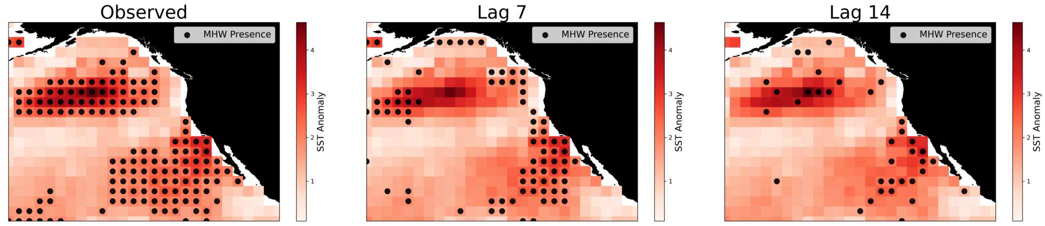


Figure A.38: Feedforward neural network model forecasts for the 95th percentile MHW using random balanced data for August 18, 2015, with no lag, 7 days lag and 14 days lag. The color bar represents the observed sea surface temperature anomaly and dots indicate where a MHW was correctly predicted.

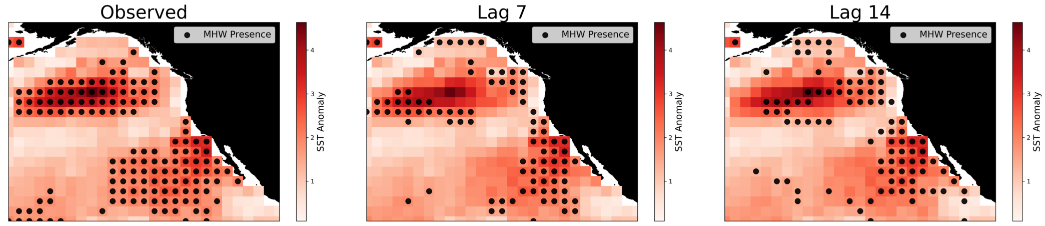


Figure A.39: Feedforward neural network model forecasts for the 95th percentile MHW using cluster centroid balanced data for August 18, 2015, with no lag, 7 days lag and 14 days lag. The color bar represents the observed sea surface temperature anomaly and dots indicate where a MHW was correctly predicted.

Notable trends represented in the spatial plots are the weak forecasting capability for 2 week lag across all models with the random balanced data. Across all models, there was significant improvement in MHW prediction in both extent and spatial accuracy with cluster centroid data. Similar to that of the 90th MHW threshold, the random forest is able to best predict the MHW event represented in these figures.

To further illustrate each model's ability to correctly predict MHW events, a spatial hit rate plot was generated for each model using both data balancing techniques. Again, the hit rate is averaged across the entire testing dataset for each grid cell, resulting in a comprehensive spatial plot that illustrates the models' effectiveness in predicting MHWs where a higher hit rate (indicated by dark red here) represents a higher model performance. Figures A.40 through A.48 represent the spatial hit rate plots across all models and both balancing methods.

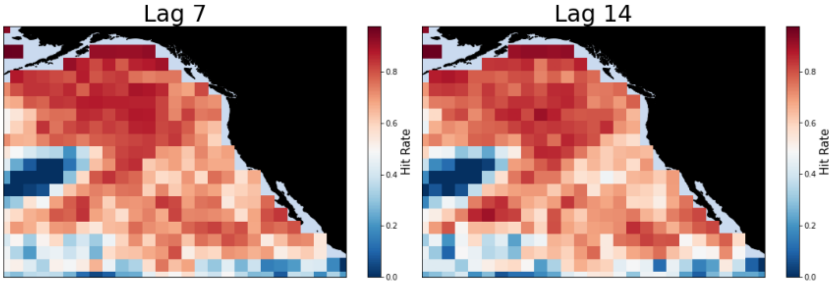


Figure A.40: Logistic regression hit rate spatial plot for binary 95th percentile MHW outcome with random balanced data.

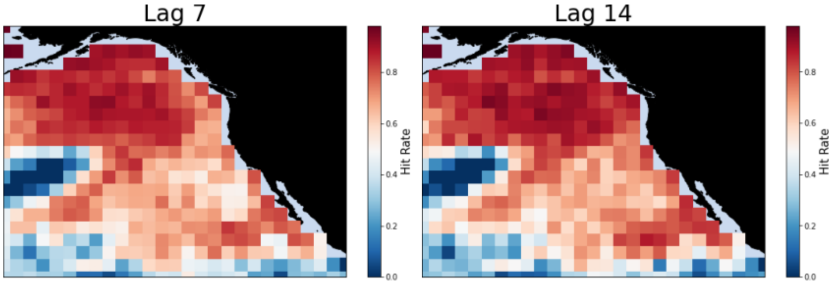


Figure A.41: Logistic regression hit rate spatial plot for binary 95th percentile MHW outcome with cluster centroid balanced data.

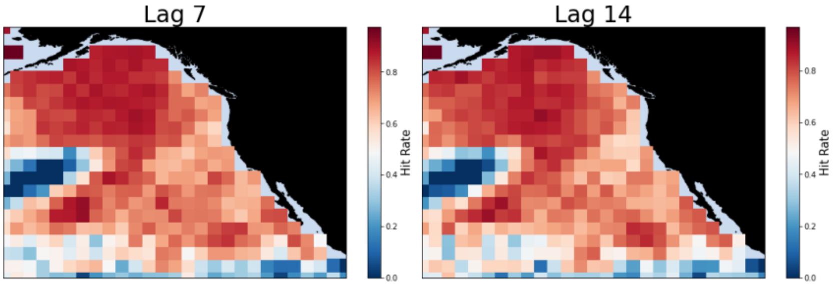


Figure A.42: Naive Bayes hit rate spatial plot for binary 95th percentile MHW outcome with random balanced data.

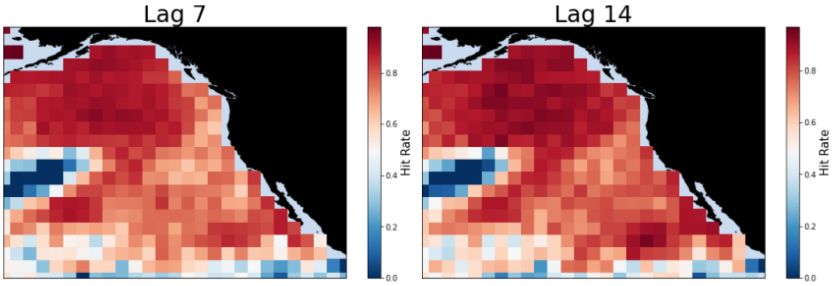


Figure A.43: Naive Bayes hit rate spatial plot for binary 95th percentile MHW outcome with cluster centroid balanced data.

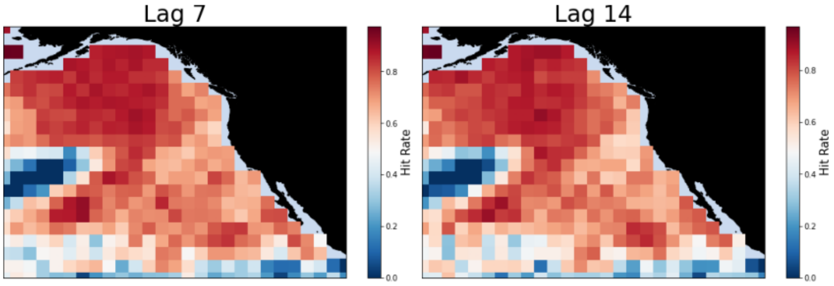


Figure A.44: Gradient boosting hit rate spatial plot for binary 95th percentile MHW outcome with random balanced data.

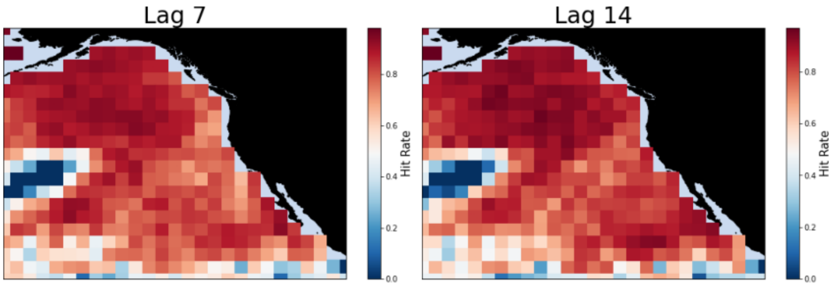


Figure A.45: Gradient boosting hit rate spatial plot for binary 95th percentile MHW outcome with cluster centroid balanced data.

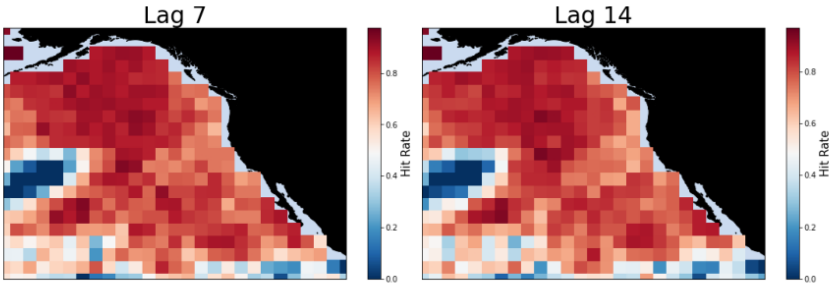


Figure A.46: Random forest hit rate spatial plot for binary 95th percentile MHW outcome with random balanced data.

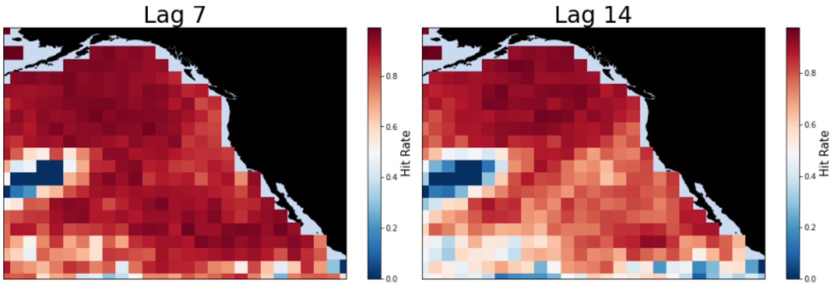


Figure A.47: Random forest hit rate spatial plot for binary 95th percentile MHW outcome with cluster centroid balanced data.

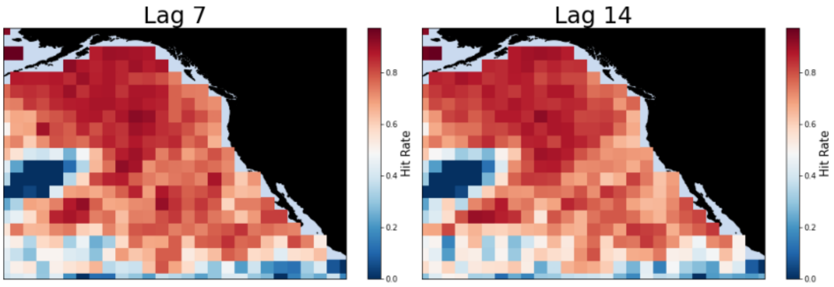


Figure A.48: Feedforward neural network hit rate spatial plot for binary 95th percentile MHW outcome with random balanced data.

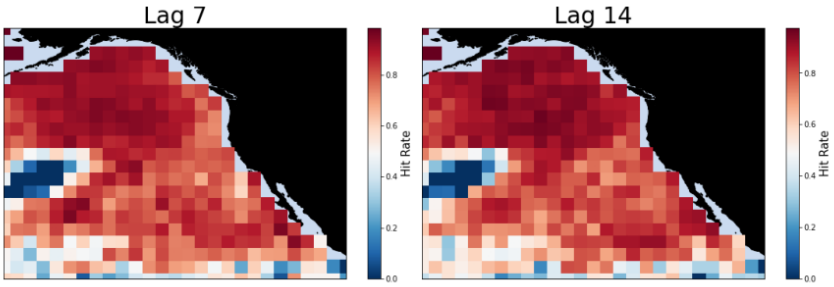


Figure A.49: Feedforward neural network hit rate spatial plot for binary 95th percentile MHW outcome with cluster centroid balanced data.

The hit rate spatial plots provide insight into the model's overall performance. Again, we are seeing similar patters that have been observed with the 90th percentile MHW and supporting data: poor performance with logistic regression, moderate performance for naive Bayes and gradient boosting, moderate to poor performance for feedforward neural network and high performance for random forest. Again, we see performance increases when using cluster centroid balanced data.

A.2.2 Unbalanced Data

The ROC curve (A.50) and PR curve (A.51) illustrate the 95th percentile MHW's model performance, revealing all models poor performance with unbalanced data. The 95th percentile MHW's model performance exhibits similar behavior to that of the 90th percentile MHW's model performance, with class 0 dominating due to its majority, resulting in poor predictive performance for class 1, which is outweighed by the prevalence of class 0.

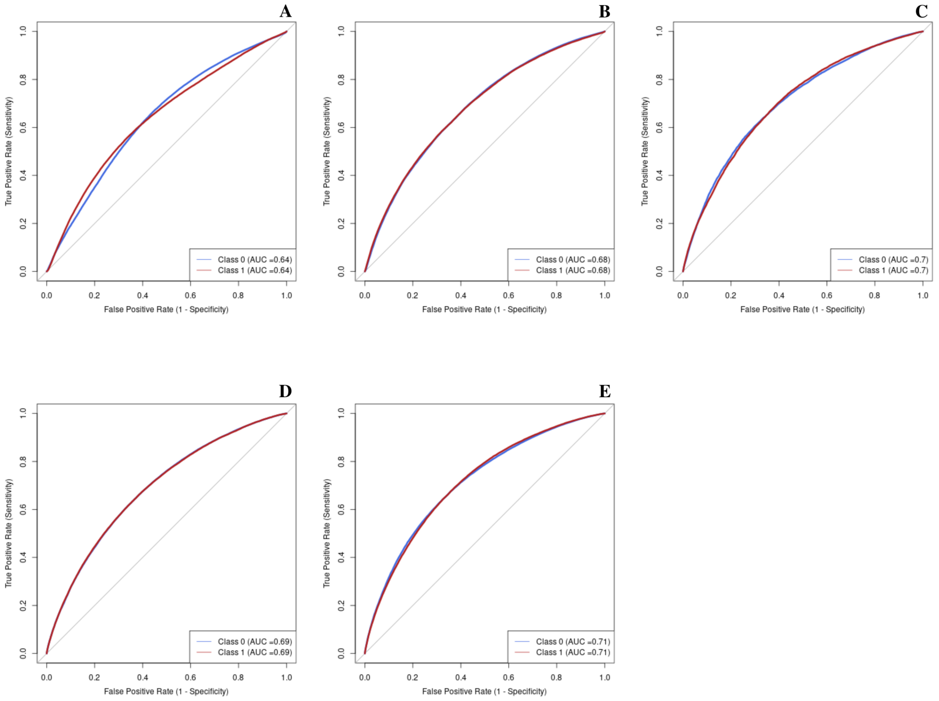


Figure A.50: ROC Curve for 95th Percentile MHW with unbalanced Data

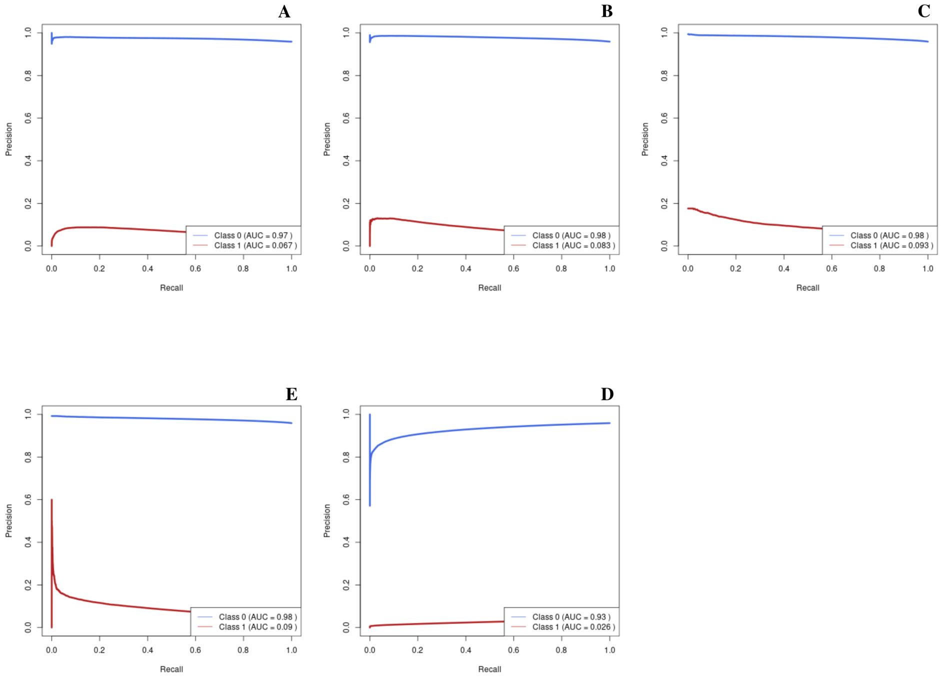


Figure A.51: PR Curve for 95$^{\text{th}}$ Percentile MHW with unbalanced Data

A.3 2-Class MHW

A.3.1 Balanced Data

Figure A.52 depicts the Receiver Operating Characteristic (ROC) curve and associated Area Under the Curve (AUC) for each model trained on randomly balanced data for a 2-class MHW outcome. Despite training on random balanced data, all models demonstrate poor performance in distinguishing between MHW classes. Notably, the logistic regression, gradient boosting, and feedforward neural network exhibit particularly inadequate performance (Figure A.52 A, C, and E, respectively), with the feedforward neural network even performing worse than random chance for class 2.

Both the random forest and naive Bayes models show relatively better performance for classes 0 and 2 with random balanced data, with an AUC of 0.63 for both models and classes. However, they perform worse for class 1, with an AUC of 0.53 for both models. This suggests that the models struggle to effectively differentiate between MHW classes, often misclassifying MHW events as either non-MHW (class 0) or the most extreme MHW class (class 2).

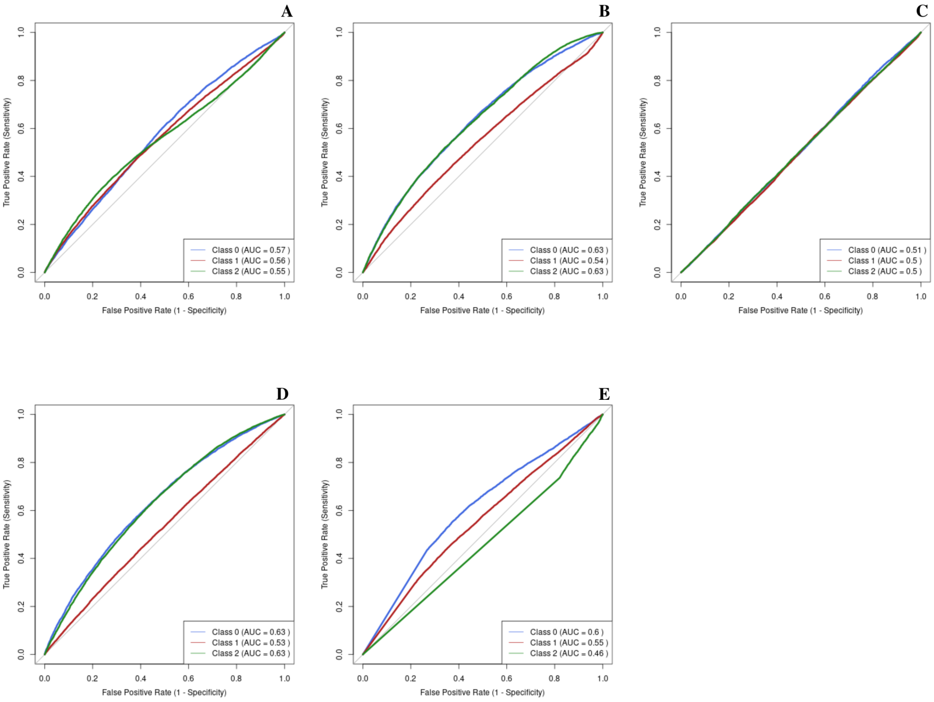


Figure A.52: ROC curve for random balanced 2-class MHW prediction across all models: A) Logistic Regression, B) Naive Bayes, C) Gradient Boosting, D) Random Forest, E) Feedforward Neural Network.

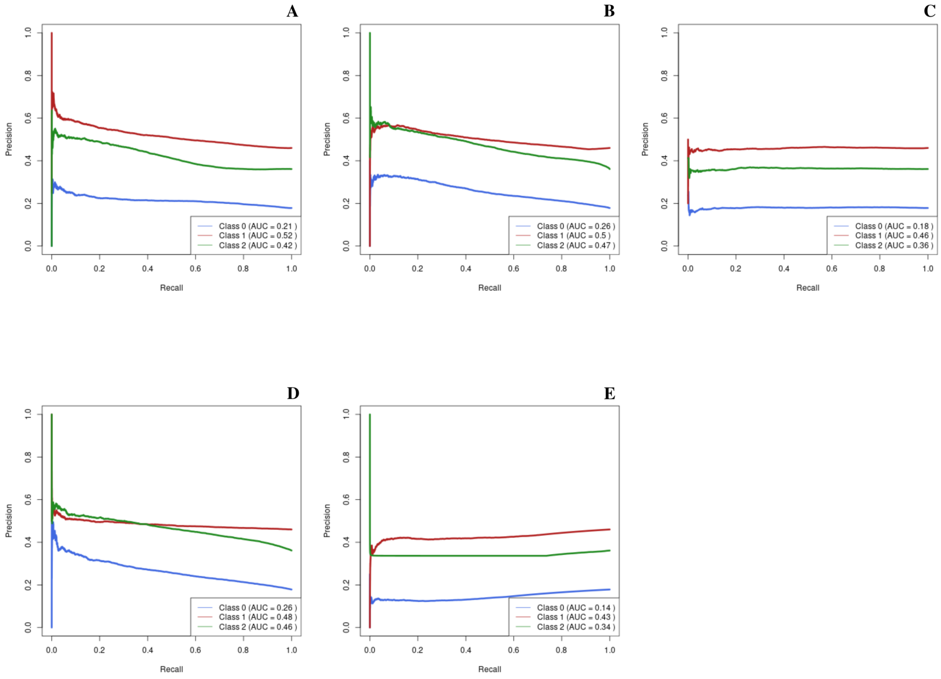


Figure A.53: PR curve for random balanced 2-class MHW prediction across all models: A) Logistic Regression, B) Naive Bayes, C) Gradient Boosting, D) Random Forest, E) Feedforward Neural Network.

Figure A.53 illustrates the Precision-Recall (PR) curves and corresponding Area Under the Curve (AUC) values for all models trained on random balanced data, providing further insights into the predictive performance of 2-class MHWs. Despite utilizing random balanced data, the feedforward neural network demonstrates the poorest performance overall. It exhibits an AUC of 0.14 for class 0, 0.43 for class 1, and 0.34 for class 2 (Figure A.53E). Conversely, the best-performing models, the random forest and naive Bayes, do not show significant improvements in AUC across all classes (Figure A.53 B and D).

These findings underscore the challenges faced by all models in distinguishing and accurately classifying MHW events across different classes. Even when cluster centroid balanced data is employed, all models continue to struggle, as evidenced by similarly poor AUC scores to those observed with random balanced data (Figure 3.10).

In order to refine the evaluation of our models, comprehensive plots illustrating lag versus accuracy and lag versus hit rate were constructed (Figures A.54 and A.55). All models exhibited poor average accuracy and hit rate through all lags.

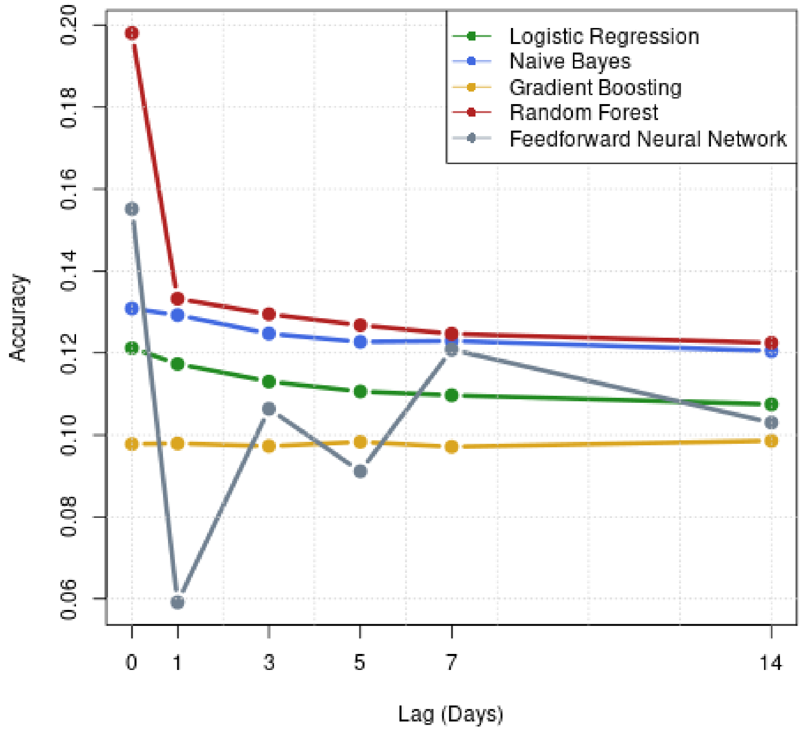


Figure A.54: Accuracy versus lagged predictor variables using random balanced 2-class MHW data across all models: Logistic Regression, Naive Bayes, Gradient Boosting, Random Forest and Feedforward Neural Network.

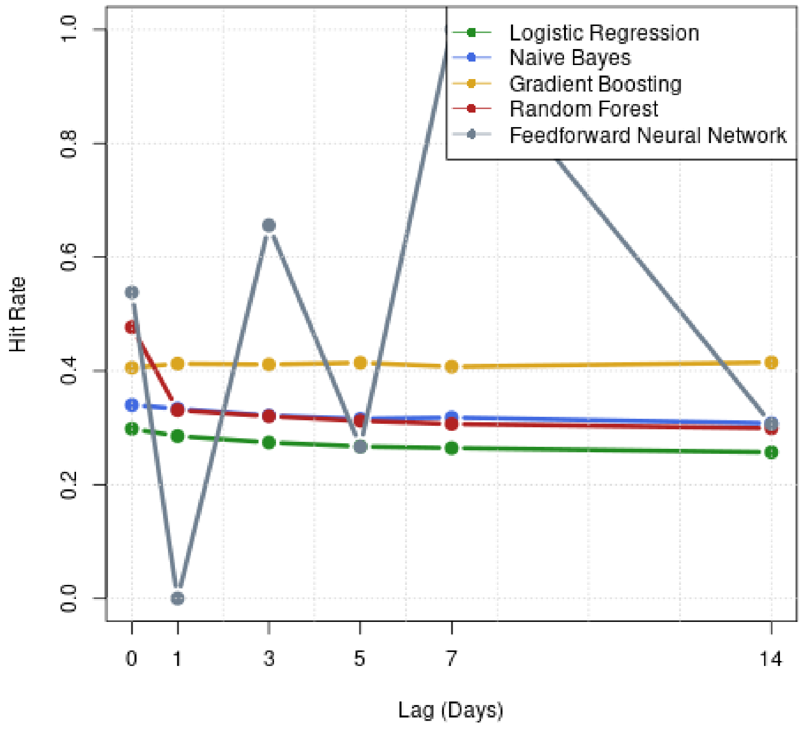


Figure A.55: Hit rate versus lagged predictor variables using random balanced 2-class MHW data across all models: Logistic Regression, Naive Bayes, Gradient Boosting, Random Forest and Feedforward Neural Network.

Figures A.56 through A.64 demonstrate each model's averaged hit rate. All models struggled to correctly predict MHW classes, with minor improvements when using cluster centroid data rather than random balanced data.

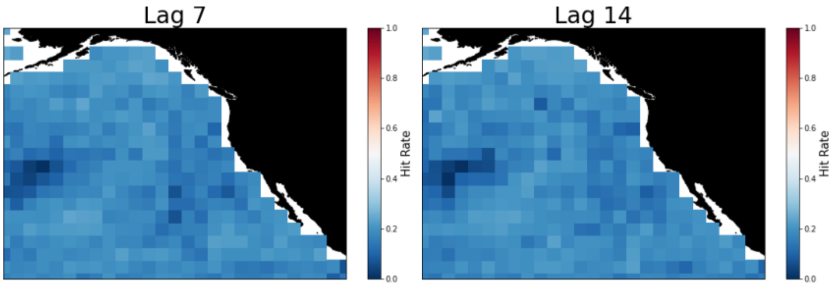


Figure A.56: Logistic regression hit rate spatial plot for 2-class MHW outcome with random balanced data.

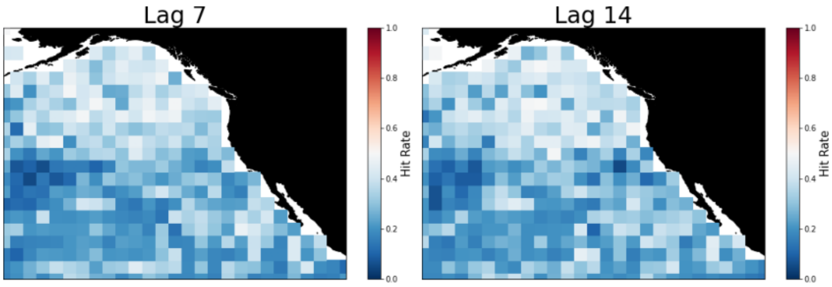


Figure A.57: Logistic regression hit rate spatial plot for 2-class MHW outcome with cluster centroid balanced data.

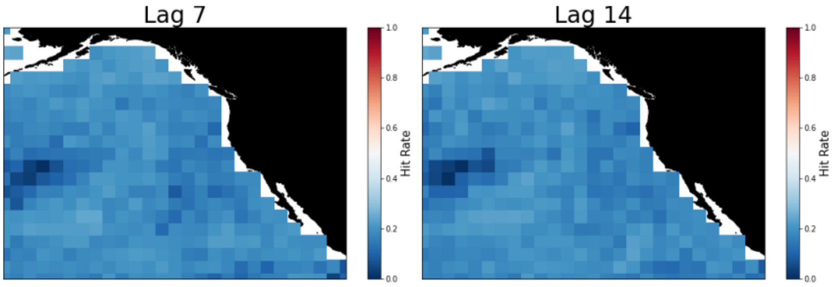


Figure A.58: Naive Bayes hit rate spatial plot for 2-class MHW outcome with random balanced data.

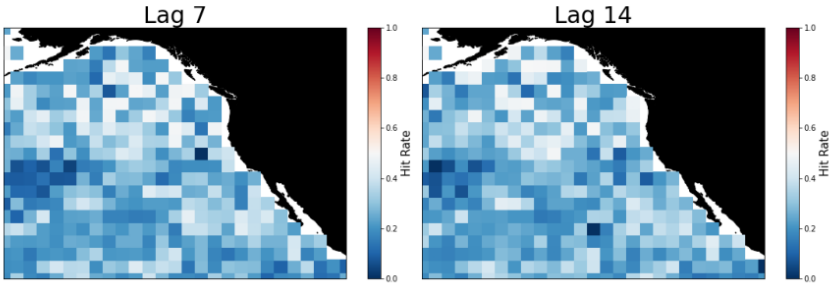


Figure A.59: Naive Bayes hit rate spatial plot for 2-class MHW outcome with cluster centroid balanced data.

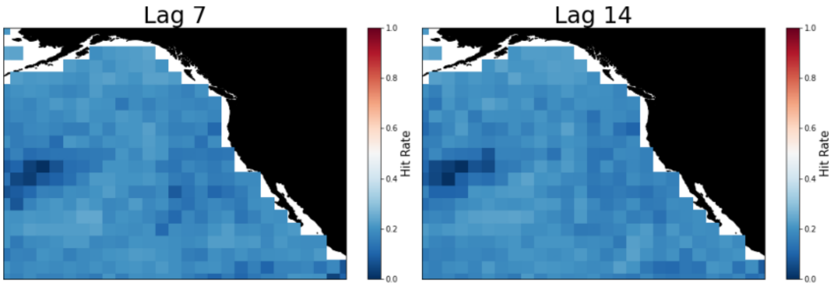


Figure A.60: Gradient boosting hit rate spatial plot for 2-class MHW outcome with random balanced data.

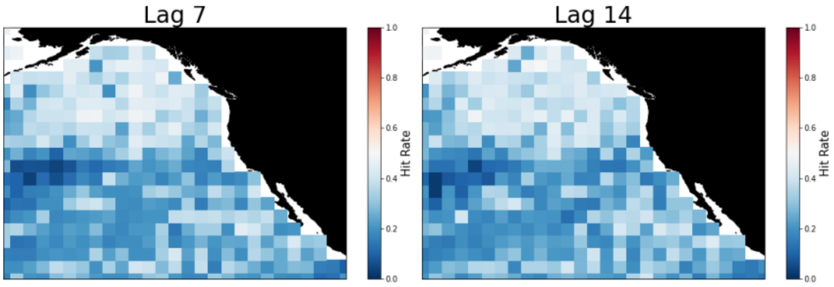


Figure A.61: Gradient boosting hit rate spatial plot for 2-class MHW outcome with cluster centroid balanced data.

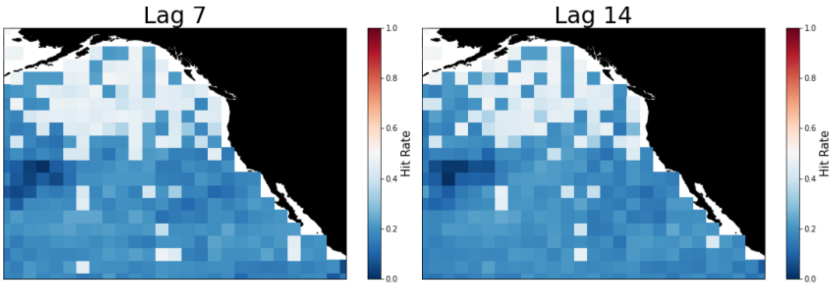


Figure A.62: Random forest hit rate spatial plot for 2-class MHW outcome with random balanced data.

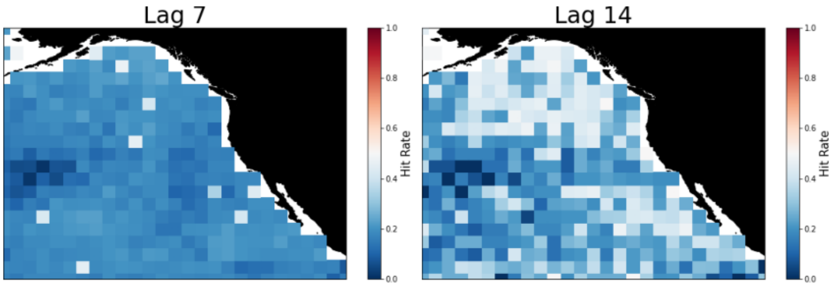


Figure A.63: Feedforward neural network hit rate spatial plot for 2-class MHW outcome with random balanced data.

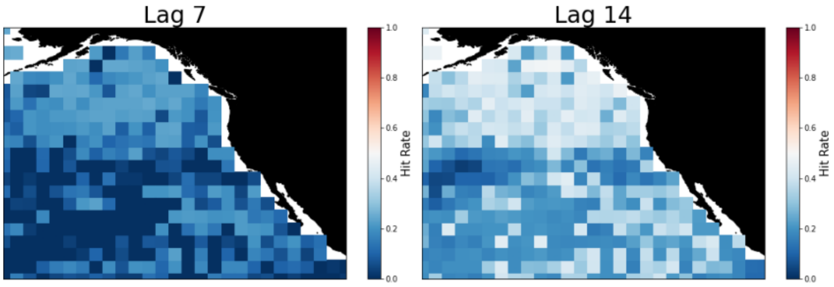


Figure A.64: Feedforward neural network hit rate spatial plot for 2-class MHW outcome with cluster centroid balanced data.

A.3.2 Unbalanced Data

The ROC curve (A.65) and PR curve (A.66) illustrate the 2-class MHW's model performance, revealing all models perform poorly.

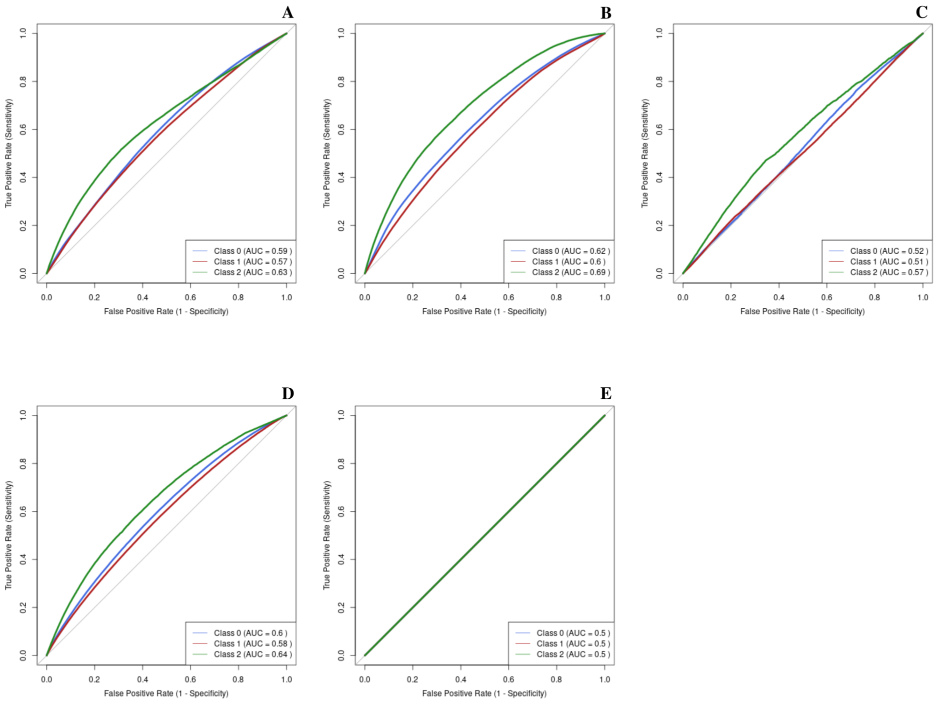


Figure A.65: ROC Curve for 2-Class MHW with Unbalanced Data

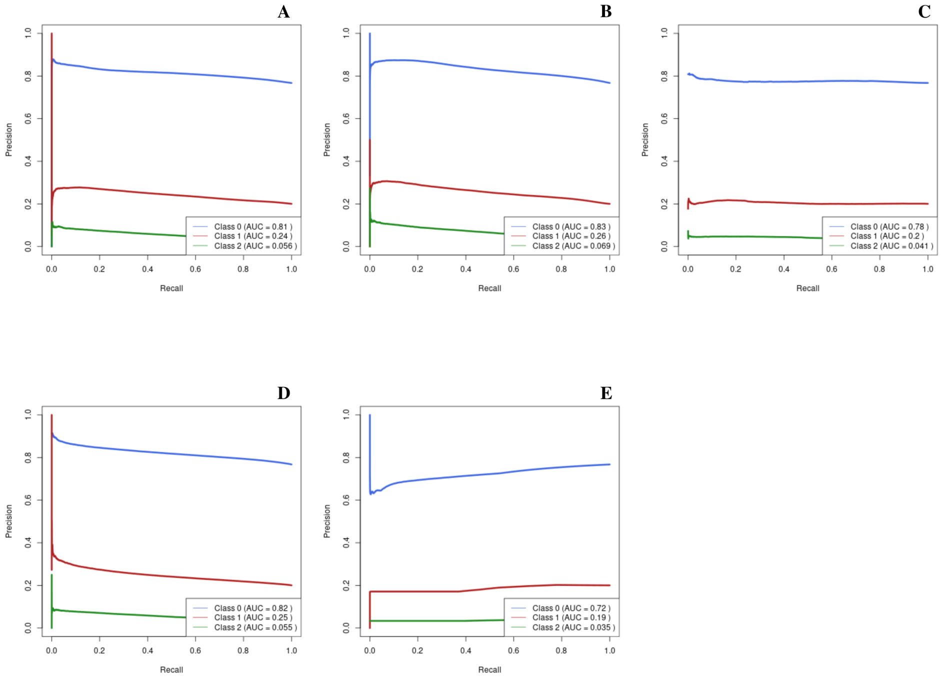


Figure A.66: PR Curve for 2-Class MHW with Unbalanced Balanced Data

A.4 4-Class MHW

A.4.1 Balanced Data

Figure A.67 depicts the Receiver Operating Characteristic (ROC) curve and associated Area Under the Curve (AUC) for each model trained on randomly balanced data for a 4-class MHW outcome while figure A.68 represents the ROC curve and associated AUC using cluster centroid balanced data. All models demonstrate poor performance at distinguishing between MHW classes for both the random and cluster centroid balanced data, notably poor prediction is the logistic regression and gradient boosting. Across both data balancing techniques, random forest and naive Bayes exhibit relatively good performance, with class 0 and class 4 having

AUC ranging from 0.66 to 0.55 (Figures A.67 and A.68, B and D). The relatively highest AUC scores among the most extreme classes (class 0 and class 4), similar to 2-class results, demonstrate that even the best of the models struggle to distinguish between MHW classes. The remaining models (logistic regression, gradient boosting, feedforward neural network) for both datasets sparsely demonstrate predictive performance better than random chance (Figures A.67 and A.68 A, C and E).

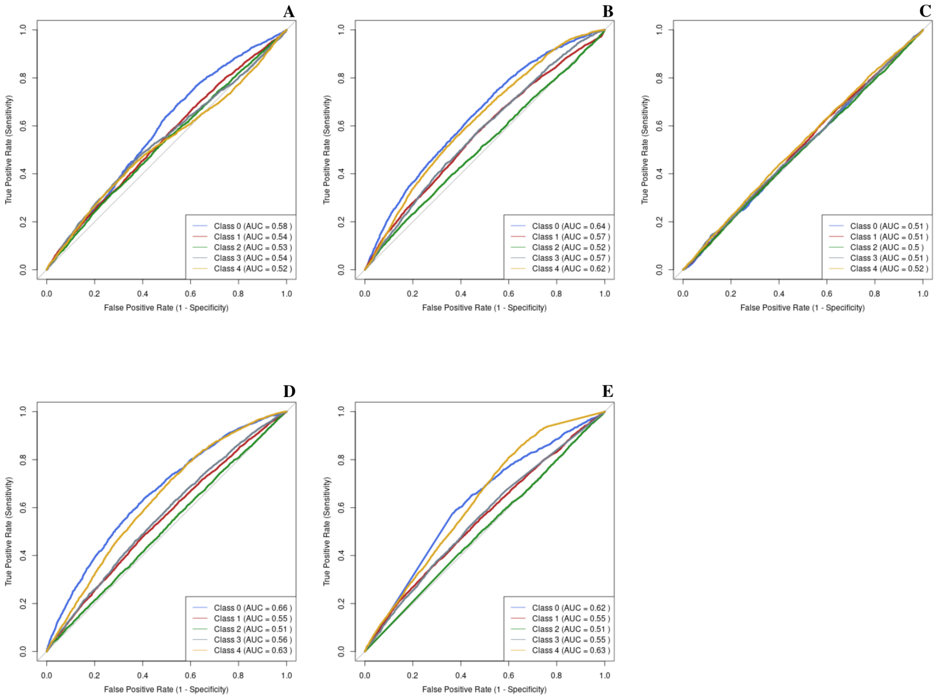


Figure A.67: ROC curve for random balanced 4-class MHW prediction across all models: A) Logistic Regression, B) Naive Bayes, C) Gradient Boosting, D) Random Forest, E) Feedforward Neural Network.

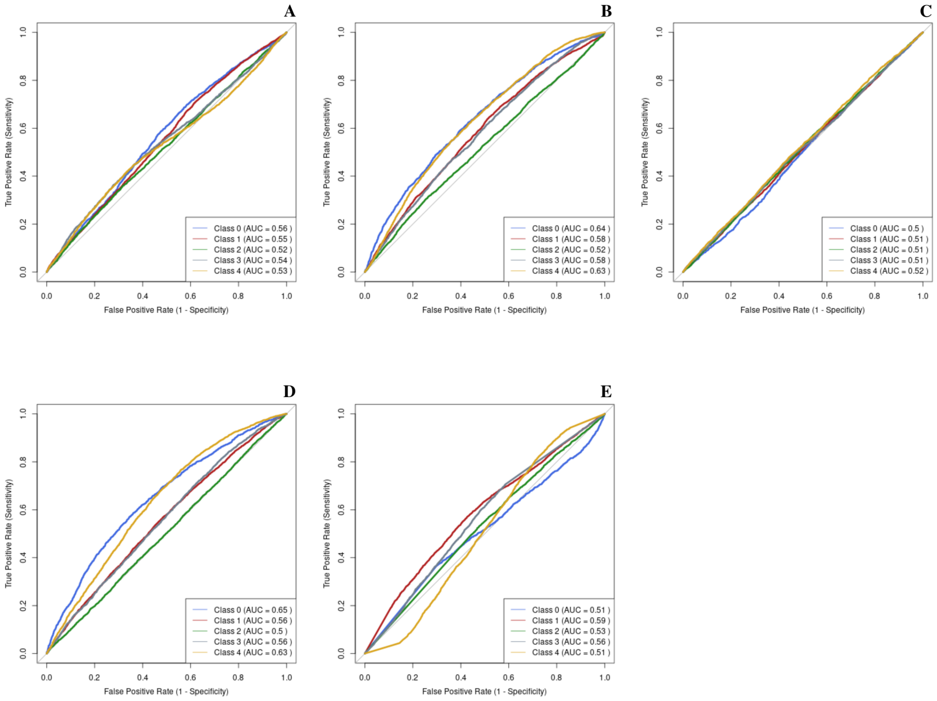


Figure A.68: ROC curve for cluster centroid balanced 4-class MHW prediction across all models: A) Logistic Regression, B) Naive Bayes, C) Gradient Boosting, D) Random Forest, E) Feedforward Neural Network.

Figures A.69 and A.70 depict the Precision-Recall (PR) curves and corresponding Area Under the Curve (AUC) values for all models trained on randomly balanced and cluster centroid balanced data. These visualizations offer deeper insights into the predictive efficacy of 4-class MHW models. Consistent with the ROC curves, the PR curves indicate generally poor predictive power across all models. However, discerning relative performance nuances from the PR curves is challenging due to minimal differences. Therefore, in this context, ROC curves provide a clearer representation of relative model performance.

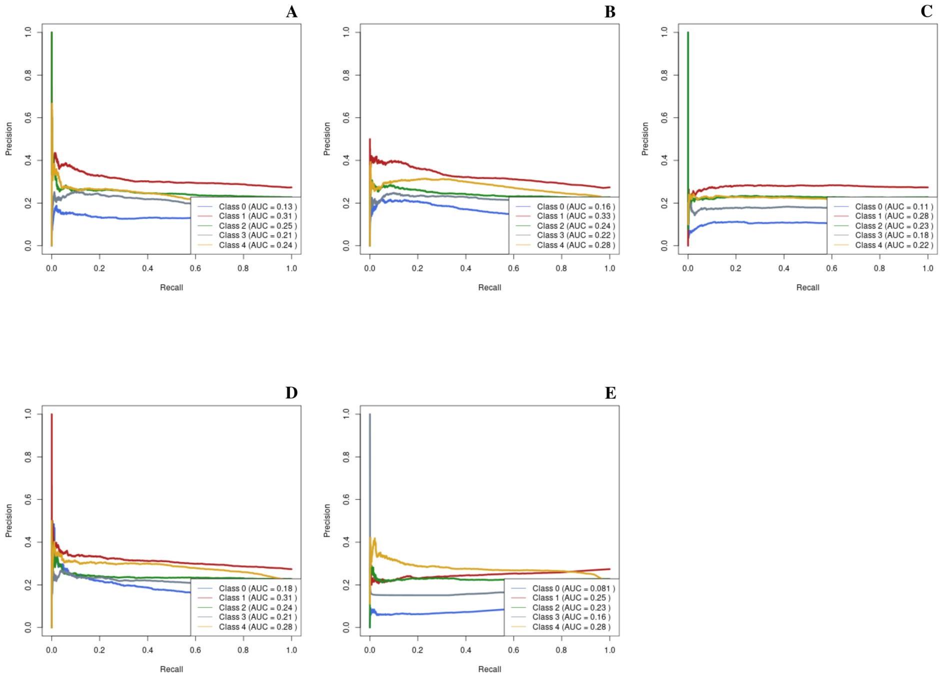


Figure A.69: PR curve for random balanced 4-class MHW prediction across all models: A) Logistic Regression, B) Naive Bayes, C) Gradient Boosting, D) Random Forest, E) Feedforward Neural Network.

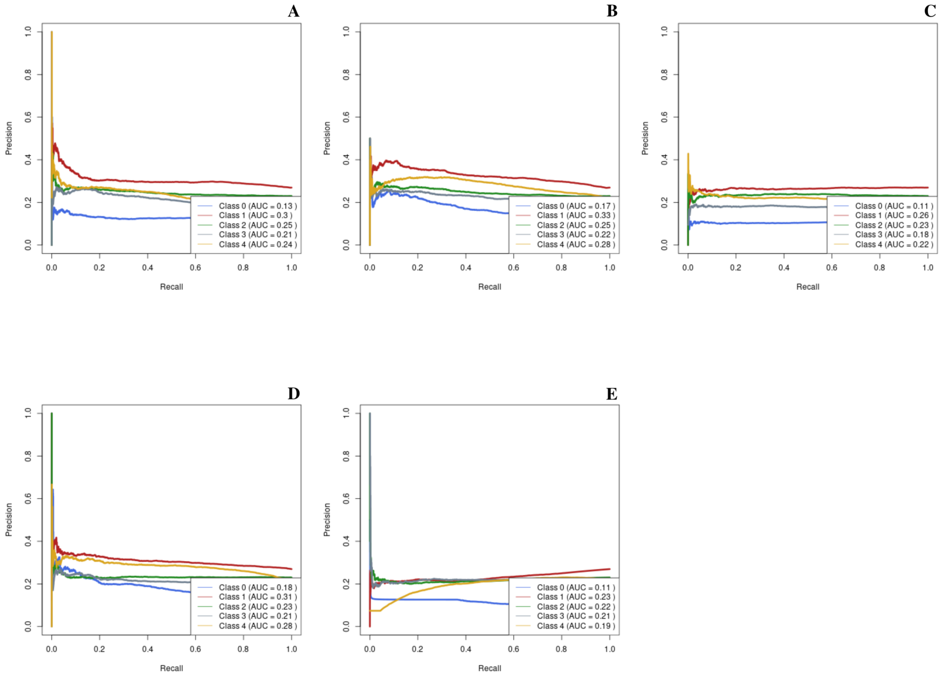


Figure A.70: PR curve for cluster centroid balanced 4-class MHW prediction across all models: A) Logistic Regression, B) Naive Bayes, C) Gradient Boosting, D) Random Forest, E) Feedforward Neural Network.

To track the evolution of model performance, we evaluated accuracy and hit rate against lagged predictor variables. In Figures A.71 and A.72, we present the accuracy of each model across different lag periods, ranging from 1 day to 2 weeks, using random balanced and cluster centroid balanced data, respectively. Across both data balancing scenarios, all models performed very poor. As the accuracy represented here is averaged among all classes, the average accuracy is likely weighed down heavily by classes that performed very poorly (the intermediate MHW classes, class 1, class 2 and class 3).

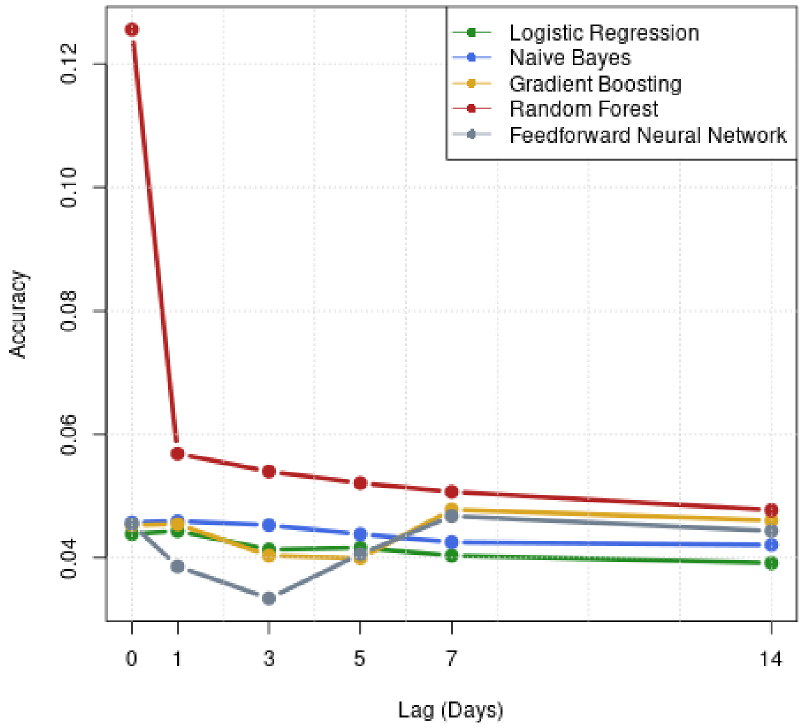


Figure A.71: Accuracy versus lagged predictor variables using random balanced 4-class MHW data across all models: Logistic Regression, Naive Bayes, Gradient Boosting, Random Forest and Feedforward Neural Network.

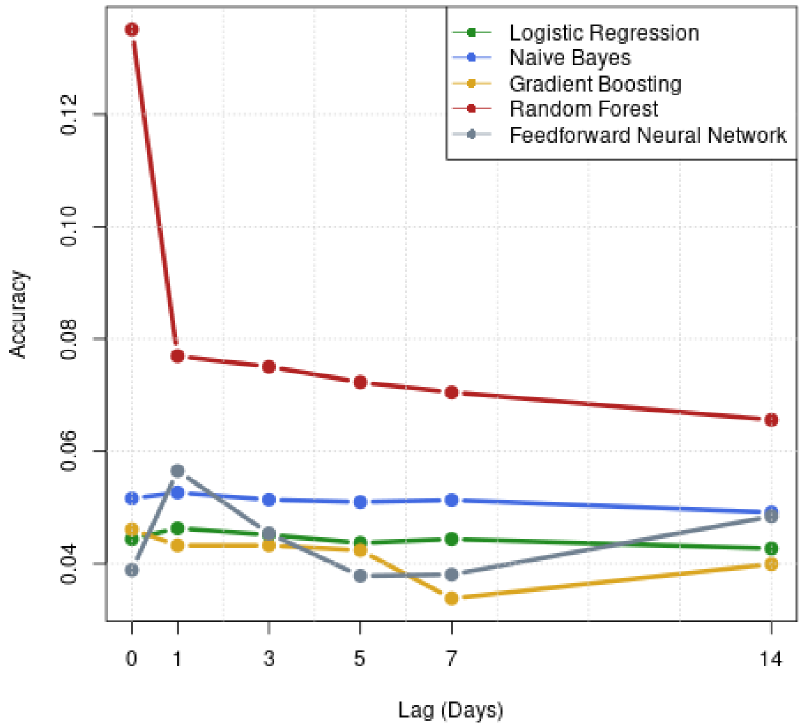


Figure A.72: Accuracy versus lagged predictor variables using cluster centroid balanced balanced 4-class MHW data across all models: Logistic Regression, Naive Bayes, Gradient Boosting, Random Forest and Feedforward Neural Network.

Similar to the accuracy versus lag plot, the hit rate versus lag plots assess each model's hit rate over lags ranging from 1 day to 2 weeks. Figures A.73 and A.74 display each model's hit rate over various lags. Much like the accuracy versus lag plot, each model for both the random balanced and cluster centroid balanced data exhibit poor hit rate performance, further emphasizing all model's inability to correctly classify MHWs.

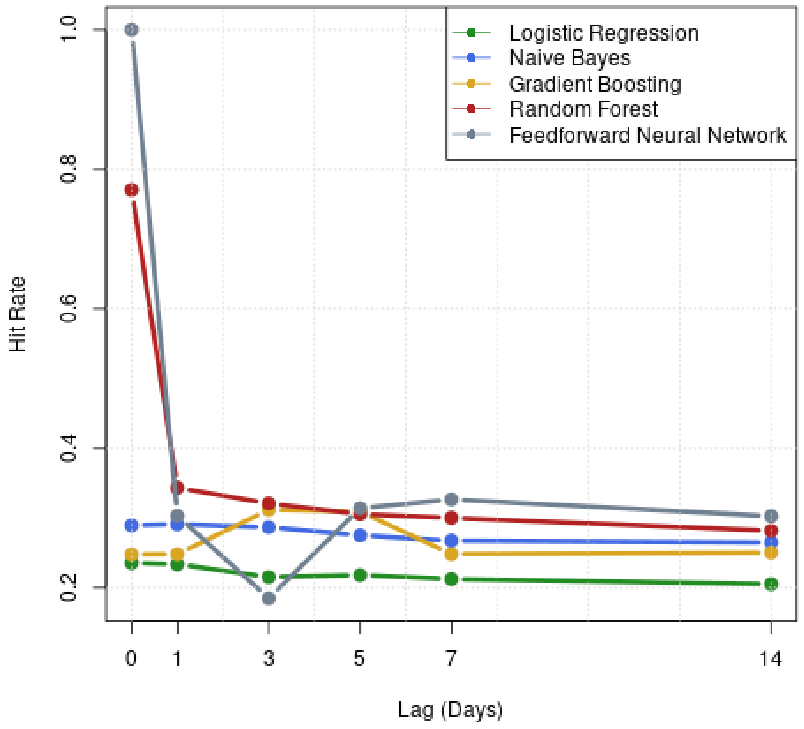


Figure A.73: Hit rate versus lagged predictor variables using random balanced 4-class MHW data across all models: Logistic Regression, Naive Bayes, Gradient Boosting, Random Forest and Feedforward Neural Network.

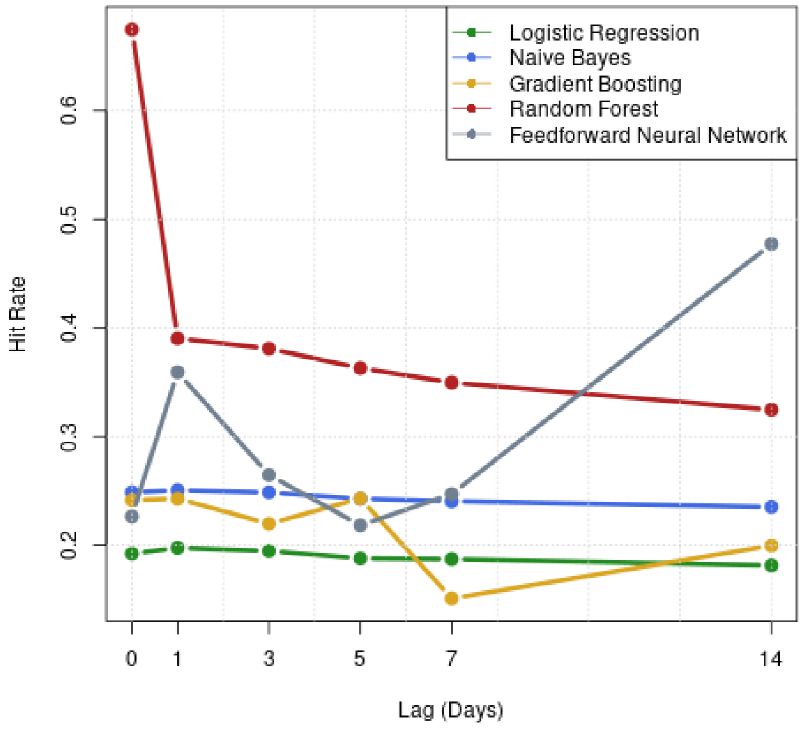


Figure A.74: Hit rate versus lagged predictor variables using cluster centroid balanced 4-class MHW data across all models: Logistic Regression, Naive Bayes, Gradient Boosting, Random Forest and Feedforward Neural Network.

Figures A.75 through A.84 represent the spatial hit rate across each models and both balancing techniques. Similar to 2-class MHW predictions, 4-class performs poor. Although, with 4-class there is some improvement in the average hit rate, especially with cluster centroid data.

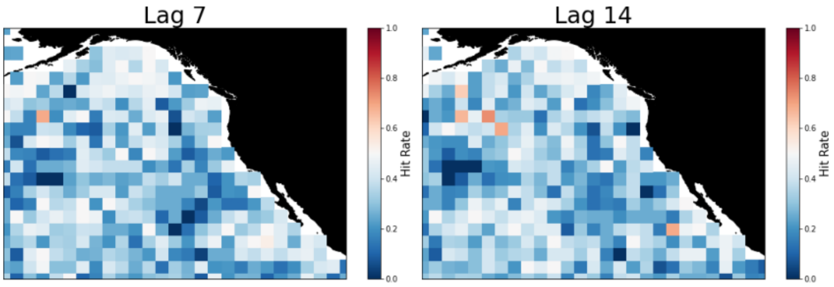


Figure A.75: Logistic regression hit rate spatial plot for 4-class MHW outcome with random balanced data.

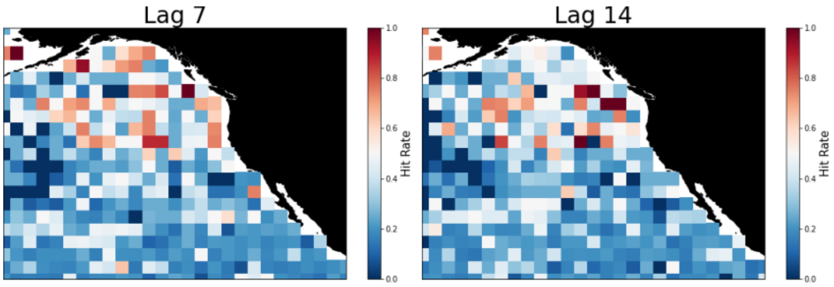


Figure A.76: Logistic regression hit rate spatial plot for 4-class MHW outcome with cluster centroid balanced data.

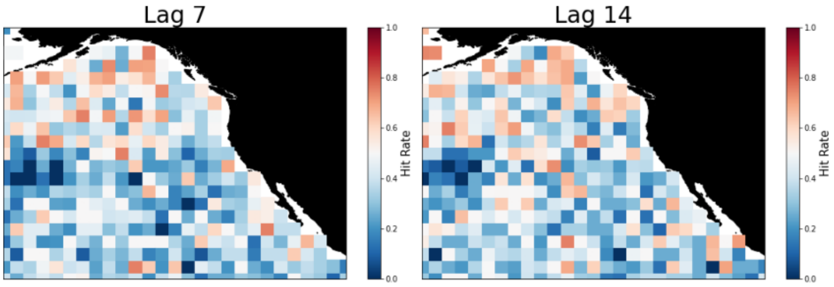


Figure A.77: Naive Bayes hit rate spatial plot for 4-class MHW outcome with random balanced data.

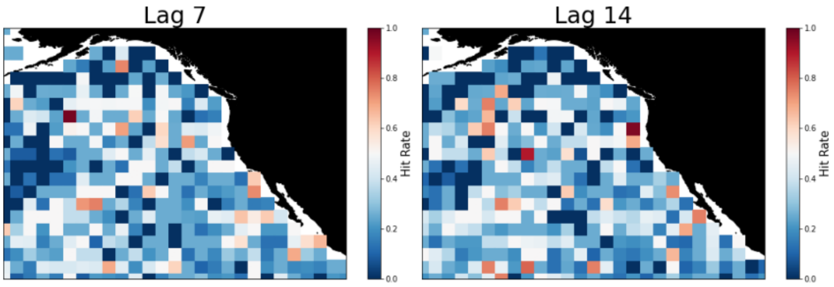


Figure A.78: Naive Bayes hit rate spatial plot for 4-class MHW outcome with cluster centroid balanced data.

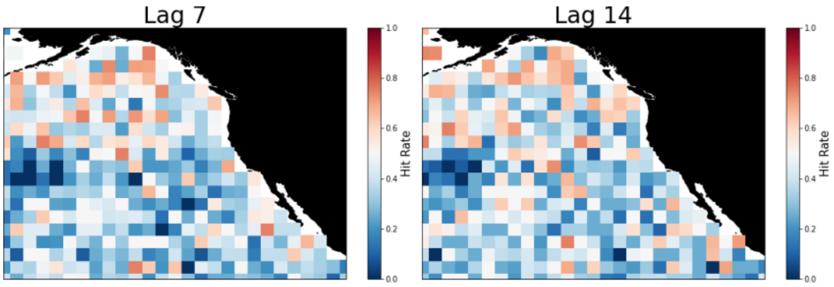


Figure A.79: Gradient boosting hit rate spatial plot for 4-class MHW outcome with random balanced data.

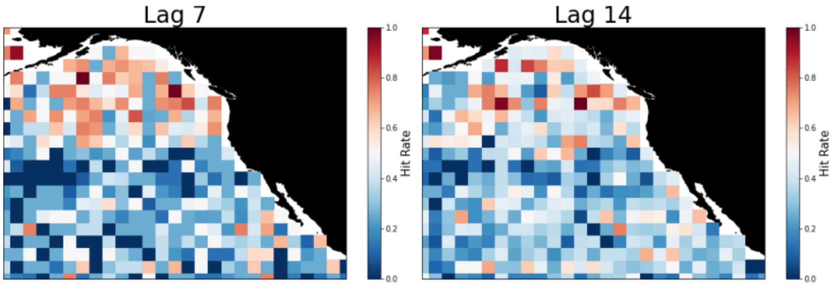


Figure A.80: Gradient boosting hit rate spatial plot for 4-class MHW outcome with cluster centroid balanced data.

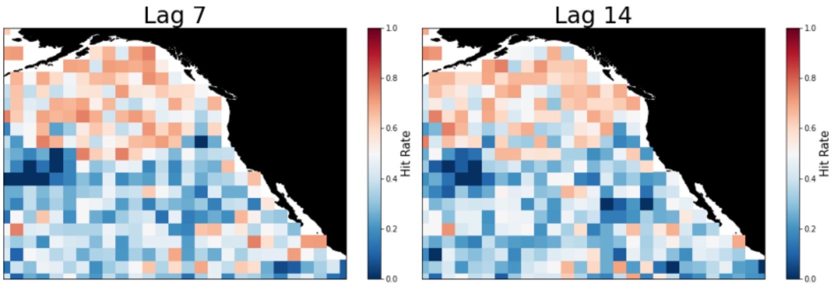


Figure A.81: Random forest hit rate spatial plot for 4-class MHW outcome with random balanced data.

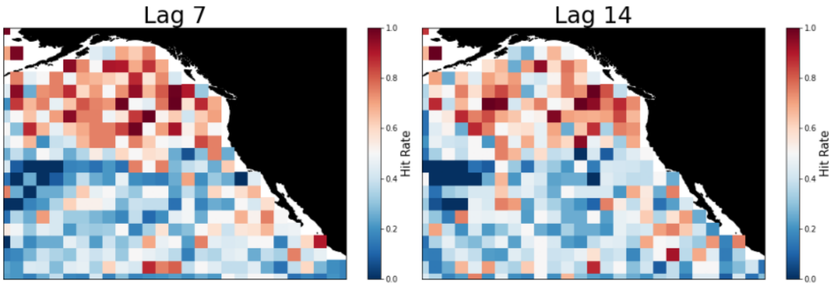


Figure A.82: Random forest hit rate spatial plot for 4-class MHW outcome with cluster centroid balanced data.

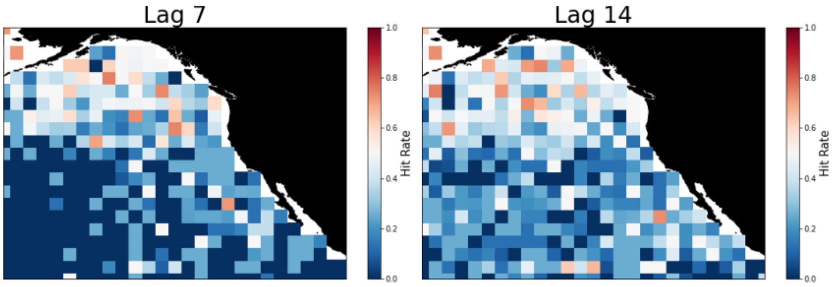


Figure A.83: Feedforward neural network hit rate spatial plot for 4-class MHW outcome with random balanced data.

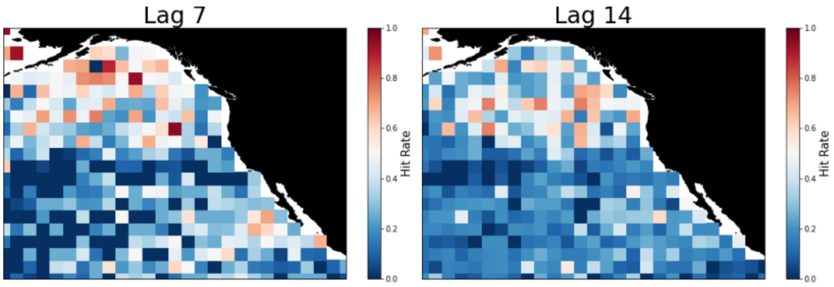


Figure A.84: Feedforward neural network hit rate spatial plot for 4-class MHW outcome with cluster centroid balanced data.

A.4.2 Unbalanced Data

The ROC curve (A.85) and PR curve (A.86) illustrate the 2-class MHW's model performance, revealing all models perform poorly.

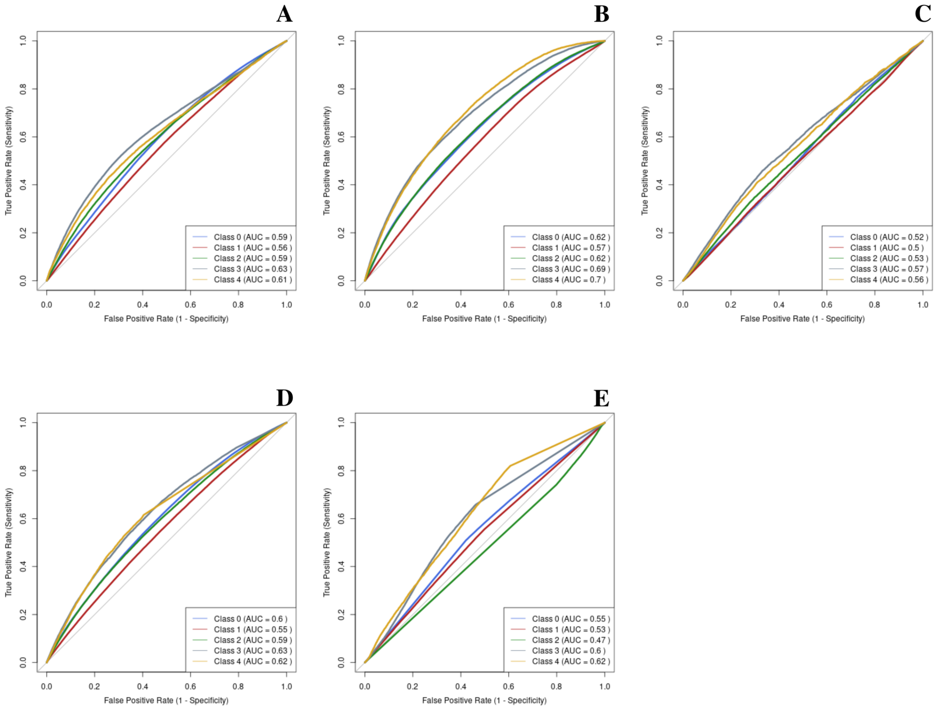


Figure A.85: ROC Curve for 4-Class MHW with unbalanced Data

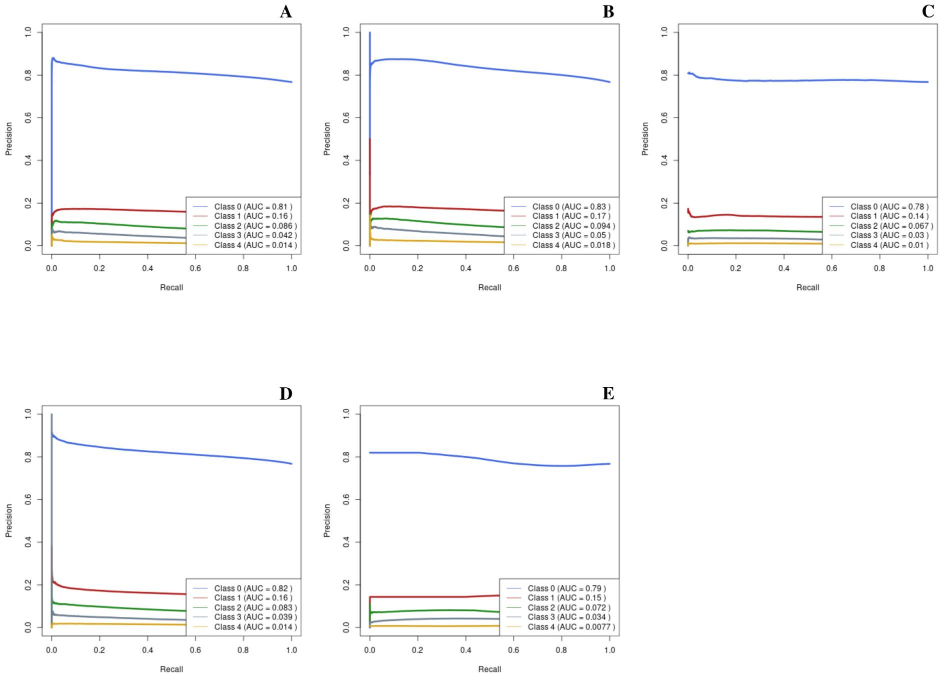


Figure A.86: PR Curve for 4-Class MHW with unbalanced Balanced Data