Geophysical Research Letters

RESEARCH LETTER
10.1029/2022GL099069

Assessing Predictability of Marine Heatwaves With Random Forests

K. Giamalaki[1,2], C. Beaulieu[1], and J. X. Prochaska[1,3,4]

[1]University of California Santa Cruz, Santa Cruz, CA, USA, [2]Now at Moody's Analytics, London, UK, [3]Kavli IPMU (WPI), UTIAS, The University of Tokyo, Chiba, Japan, [4]Division of Sciences, National Astronomical Observatory of Japan, Mitaka, Japan

Key Points:

- Marine heatwaves in the northeast Pacific are predictable on weekly lead times using atmospheric and oceanic conditions
- Average accuracy at predicting marine heatwave presence/absence is 76%, but lowers to 38% when forecasting their severity
- Machine learning approaches should be considered to augment our ability to forecast marine heatwaves

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

C. Beaulieu,
beaulieu@ucsc.edu

Citation:

Giamalaki, K., Beaulieu, C., & Prochaska, J. X. (2022). Assessing predictability of marine heatwaves with random forests. *Geophysical Research Letters*, *49*, e2022GL099069. https://doi.org/10.1029/2022GL099069

Received 19 MAY 2022
Accepted 18 NOV 2022

Author Contributions:

Conceptualization: K. Giamalaki, C. Beaulieu, J. X. Prochaska
Formal analysis: K. Giamalaki, C. Beaulieu, J. X. Prochaska
Funding acquisition: K. Giamalaki, C. Beaulieu, J. X. Prochaska
Methodology: K. Giamalaki, C. Beaulieu, J. X. Prochaska
Project Administration: C. Beaulieu
Resources: C. Beaulieu
Supervision: C. Beaulieu
Writing – original draft: K. Giamalaki, C. Beaulieu, J. X. Prochaska
Writing – review & editing: K. Giamalaki, C. Beaulieu, J. X. Prochaska

Abstract Marine heatwaves (MHWs) have increased in frequency and duration over the last century and are expected to intensify in the future. Such events have become an increasing threat for marine ecosystems and subsequently the economies and populations that rely on them. Here we apply random forests to assess skill in forecasting MHWs onset and severity at multiple prediction lead times. Random forests models are trained on a range of atmospheric and oceanic conditions to identify precursors of MHWs. The best performing random forest model accurately captures (76%) MHW presence/absence in the northeast Pacific and is capable of forecasting realistic extreme sea surface temperature patterns at weekly lead times. However, the total accuracy drops to 38% when forecasting MHW severity. Machine learning algorithms affirm further exploration as forecasting tools and have the potential to accelerate our predictive ability and preparedness against upcoming extreme climate changes.

Plain Language Summary Marine heatwaves (MHWs) are defined as extremely high ocean temperatures that last for at least five consecutive days, threatening marine life and environment and the subsequent economies and populations that rely on them. MHWs are expected to become more intense and frequent in the future and therefore their predictability is crucial for the effective management of the marine environment and preparedness of coastal communities. Machine learning models based on algorithms are able to analyze and draw inference from patterns in data without following explicit instructions. These tools have shown promise in predicting a range of climate extreme events. Here we assess predictability of MHWs in the northeast Pacific using a Random Forest machine learning model. We train the model to recognize air-sea patterns that may act as precursors to MHWs at different time lags. Our model predicts the presence or absence of MHWs with 76% accuracy on weekly lead times. Machine learning models show promise to advance our forecasting ability of upcoming MHWs.

1. Introduction

Over the last two decades, unusual warm events have been observed in the global ocean, significantly affecting our environment and society (Frölicher & Laufkötter, 2018). A marine heatwave (MHW) is commonly defined as a "prolonged discrete anomalously warm water event that can be described by its duration, intensity, rate of evolution, and spatial extent" (Hobday et al., 2016). Thus far, understanding and monitoring MHWs have received substantial scientific and public attention (Holbrook et al., 2019, 2020; Oliver et al., 2021). Recently, the most severe MHWs were attributed to anthropogenic climate change (Laufkötter et al., 2020), highlighting the necessity for motivated efforts to limit global warming. In the North Pacific, the largest MHW ever recorded, dubbed the Warm Blob, occurred between 2013 and 2015 with maximum sea surface temperatures that reached 6°C above average in some areas along the coast of Southern California (Bond et al., 2015). Impacts on marine ecosystems include harmful algal blooms, shifts in species range, and even local extinctions (Smale et al., 2019). Consequences have also been reported in the economic sector, since aquaculture and important fisheries are vulnerable to MHW events. For example, both commercial and recreational fisheries faced major challenges and loss of millions of dollars after the 2013–2015 northeast Pacific MHW (Cavole et al., 2016). Furthermore, the largest ever-recorded harmful algal bloom in the region, caused by the extreme temperatures, produced toxins that contaminated valuable shellfish and crab fisheries (McCabe et al., 2016). Increasing sea surface temperatures may also impact regional weather by affecting storms, precipitation, air temperature and droughts, the latter posing risks for potential wildfire events (Chikamoto et al., 2017). For example, anomalously high ocean temperatures in the northeast Pacific was the key forcing for the extensive dry winters in California during 2011–2014

(Seager et al., 2015), with simultaneous multiple air heatwaves in the area (AghaKouchak et al., 2014) and some of the strongest reported North American heat waves in 2016 (NOAA, 2017). Past MHWs in the Mediterranean have also been reported as concurrent events with air heatwaves over Europe (Olita et al., 2007). Our ability to predict MHWs can ultimately inform adaptive action and help reduce the vulnerability of human communities that depend on the sea and those experiencing the effects of unforeseen extreme weather.

Triggering mechanisms that have been attributed to MHWs include changes in the ocean heat transport, persistent atmospheric patterns and ocean-atmosphere coupling mechanisms (Holbrook et al., 2019; Sen Gupta et al., 2020; Vogt et al., 2022). In the northeast Pacific, ENSO and tropical warming as well as atmospheric forcing have been suggested as the most important factors leading to MHW formation (Amaya et al., 2020; Capotondi et al., 2019). A deepening in the Aleutian Low and an intensification of the North Pacific High pressure systems have been linked to the Blob in 2013–2015 (Amaya et al., 2016) and the Blob 2.0 in 2019 (Amaya et al., 2020), respectively. The sea level pressure variability causes fluctuations in basin scale wind strength and direction, which have been previously associated with oceanic warming in the region (Sun & Okumura, 2019). Predictability efforts based on dynamical and statistical approaches are developing, with the dynamical models having the advantage of resolving the physical non-linearities of the system and demonstrated skill at the seasonal time scale (Jacox et al., 2022). Statistical models are less resource intensive and less prone to model bias errors (Jacox et al., 2020). In addition, statistical forecasting provides an alternative for predicting components for which there is a lack of knowledge needed to parameterize a dynamical model (Jacox et al., 2020). Near real-time tracking of MHWs have been developed to inform the public and the scientific community of ongoing events (Schlegel, 2020). There is therefore a multitude of methodological possibilities and needs for predicting MHWs on diverse spatial and temporal scales.

Machine learning (ML) algorithms are highly promising tools to provide warnings of approaching extreme events such as MHWs, given the ever-increasing amount of data produced by satellites and climate model outputs. Here, we focus on random forests (RF) that have shown great skill at predicting climatic variables such as precipitation and energy fluxes (Anderson & Lucas, 2018; Baez-Villanueva et al., 2020; Herman & Schumacher, 2018), and forecasting effects of climate extremes on crop yields (Beillouin et al., 2020; Vogel et al., 2019). Our objective is to assess the predictability of MHWs using a RF model. We construct a RF model to forecast MHWs using climate variables that may trigger such events in the northeast Pacific. We utilize publicly available observational and reanalysis datasets to train, test and validate a RF model to assess predictability of marine extreme events on multiple lead times.

2. Data and Methods

2.1. Data

Sea level pressure, net heat flux, surface air temperature and wind speed anomalies over northeast Pacific in addition to temporal (day of year) and spatial (longitude–latitude) information are used as predictors for MHW events using RF. Multiple lags ranging from 2 days to 1 year are tested in order to compare the RF base model performance. We focus on the northeast and eastern central Pacific regions defined by 10°N–65°N and 175°W–100°W. These boundaries encompass the entire northeast Pacific and the northern part of the eastern central Pacific major fishing regions, as per the region delimitation from the Food and Agriculture Organization of the United Nations. To identify MHW categories, we employ the high-resolution gridded daily data set National Oceanic and Atmospheric Administration Optimum Interpolation Sea Surface Temperature (NOAA OISST v2.1) with a 1/4° spatial resolution from 1981 to 2019 (Banzon et al., 2020). The data is re-gridded to 2.5° × 2.5° to match the resolution of the variables used as predictors. The pixel-wise daily SST anomalies are calculated based on the climatology using the 1983–2019 base period.

The definition of MHW categories used here is similar to the one introduced by Hobday et al. (2016). More specifically, a MHW event is defined as any interval where the SST anomalies exceeds the 90th percentile for at least five consecutive days. However, MHW events that occur within 2 days of one another (e.g., a gap of 1 day between a pair of MHWs) are collated into a single event. The MHWs are then categorized based on three levels of severity: moderate, strong, and severe/extreme, defined as multiples of the 90th percentiles of the SST anomalies (90th percentile for moderate, 92.5th for strong and above 95th percentile for the severe/extreme category) (Hobday et al., 2018). Every day is labeled as either absence of MHW or one of the three levels of MHW severity.

Daily sea level pressure, surface air temperature, net heat flux and wind speed datasets are obtained by the National Centers for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) Reanalysis 1, which uses a forecast system to perform data assimilation of past data from 1948 to present, at a $2.5° \times 2.5°$ spatial resolution (Kalnay et al., 1996). Net heat flux is calculated by summing the latent, sensible, net long wave flux, and net solar flux components. As such, negative values of net heat flux represent heat loss from the ocean to the atmosphere, and vice versa. Furthermore, the wind speed (ws) is calculated as the square root of the sum of the squares of u and v wind components $\left(ws = \sqrt{u^2 + v^2}\right)$. Anomalies for all the datasets are calculated by subtracting the daily climatology based on the period 1983–2019. To smooth out short-term fluctuations between the predictors and the MHWs categories, the model is trained using all predictors averaged over a weekly sliding window such that the value of a predictor at a given day will be its average over the current and previous 6 days. Different time lags are used to test the predictability skill of the model, with the climate variables used as predictors preceding the MHW events in multiple lead times.

The years 1981–2016 are used for the training and validation process and the last three years (2017–2019) of the available record are kept for testing the model's accuracy on unseen data for a more realistic assessment. For the 3 years (2017–2019) of testing data there are 1,094 days for each pixel on the map and 400 pixels of ocean coverage, totaling 437,600 data points, including all MHW categories. Due to the large imbalance in occurrences and absences of MHWs in the northeast Pacific (8% and 92% of total data points, respectively), we balance the presence/absence of MHW events in each location in the training data set. For each pixel with a presence of MHW event in 1 year (regardless of the severity), two absences (no MHW event) occurring in following years are removed from the data set for that specific location. This resulted in a substantial reduction of the data points used to train and validate the RF model from 13,878 days × 400 pixels (5,551,200 data points) down to 869,927 data points. In other words, the majority class (i.e., no MHW) has been down-sampled to balance out the categories. Similar techniques have been previously reported to yield higher efficiency in classifying imbalanced datasets by tree classifiers (Drummond & Holte, 2003; Kubat & Matwin, 1997) as well as random forest algorithms (Chen et al., 2004). Preliminary analyses with no balancing showed spurious extreme high accuracies due to model overfitting, that mainly identifies MHW absence throughout the whole northeast Pacific. The train-test size when considering the balanced training and unbalanced testing datasets is 67%–33%.

2.2. Random Forests

The RF algorithm is a non-parametric statistical learning method, which uses an ensemble of decision trees and can be applied to regression and classification problems (Breiman, 2001). To construct an individual decision tree, smaller sub-groups of observations of the predictors (features) are created using optimal decision rules (split rules). The groups are split such that predicted variables are maximally different between the groups and maximally homogeneous within them. In classification problems, new observations are allocated to one of the groups and decision rules return the majority value. The predicted value of the RF is the class with the most votes across the multiple trees. Here the classes correspond to MHW categories: none, moderate, strong and severe/extreme. Selecting the predicted value as the majority category predicted from individual trees effectively reduces the variance since the individual decision trees tend to over-fit.

Building a RF model involves tuning several hyper-parameters that control the structure of each individual tree as well as the forest in total, for example, its size and randomness (Probst & Boulesteix, 2018). This process may result in overfitting and reduced performance on an independent data set if done using particularly complex rules that are specific to the training data set. Using "unseen" test data and cross-validation methods assist in preventing overfitting (Probst & Boulesteix, 2018). Here, we select the optimum time-lag for forecasting MHWs with a base model of 200 trees, a sample size of at least two data points to split an internal node and at least one point is needed for a terminal node. Empirical studies have previously shown that the maximum RF model performance is often achieved by the first 100 trees, that is, a larger number of trees in a forest does not necessarily improve performance (Oshiro et al., 2012; Probst & Boulesteix, 2018). Therefore, we consider the option of 200 trees sufficient to form a base model and compare predictability amongst several time lags (Text S1; Figure S1 in Supporting Information S1).

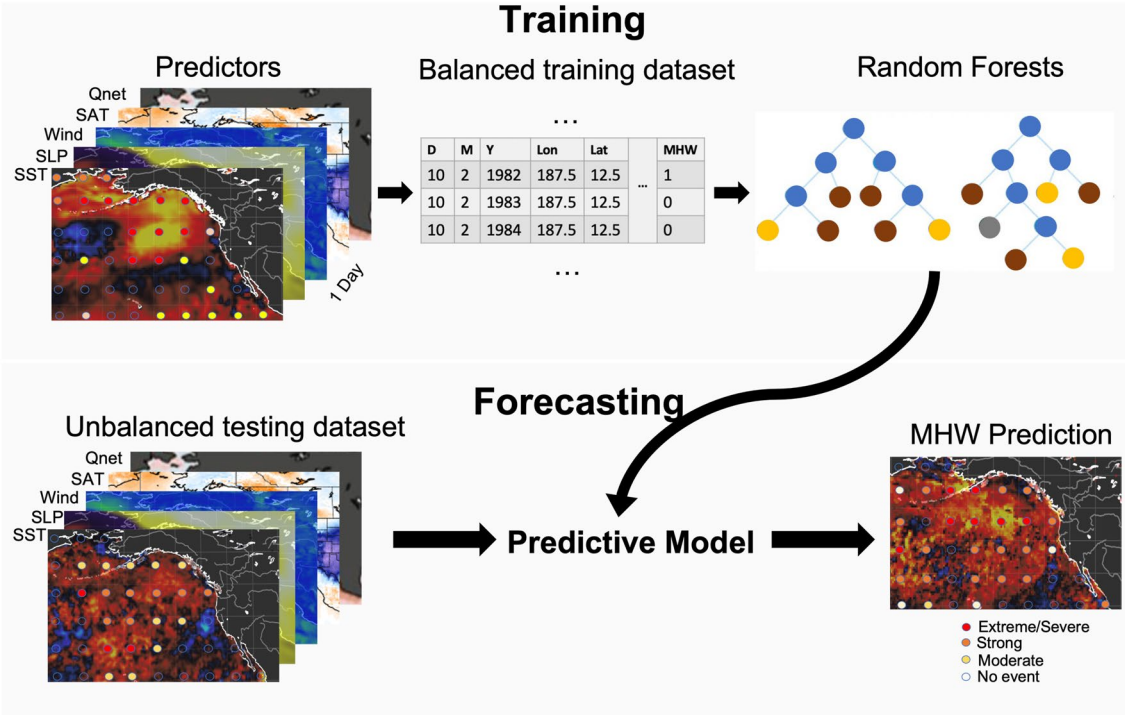

Figure 1. Schematic of the marine heatwave random forest model. The data pre-processing includes the collection of all the predictors used on a daily temporal resolution, specifically sea surface temperature (SST), sea level pressure (SLP), wind speed (Wind), surface air temperature (SAT) and net heat flux (Qnet). Each of these are gridded to 2.5° × 2.5°. The predictors together with the labels (MHW categories to predict; represented by the filled dots) are merged into a balanced training data set which is fed into the random forests. The unbalanced test data set is used to evaluate the forecasting skill of the predictive model. The model is then able to provide predictions of MHW categories and patterns a week in advance.

Once a lag is selected, we tune the RF hyper-parameters using a random grid search K-fold cross-validation. This approach separates the training data into K equal-sized randomly selected "folds," where one of the K-folds is held out in order to evaluate the model's performance on an unseen data set while the model is trained on the other folds. This results in K different models, each with an accuracy score on a different holdout set. The average of these K models' out-of-sample scores is the model's cross-validation score. Here, 3-folds are used and 100 iterations are simulated, in which a range of hyper-parameters are tested. The cross-validation showed that 700 trees are the ideal number to predict northeast Pacific MHWs and bootstrap samples should be drawn to build the optimized forest. The number of features to consider when looking for the best decision rule (split) is calculated as the square root of the number of samples in the training data set. Furthermore, five samples are required to split an internal node, and the minimum number of samples required to be at a terminal node (leaf node) is four. This allows the splitting procedure to be unrestricted of limitations relevant to node size, eventually leading to a maximum depth of a tree being 10 leaves. A schematic of the RF workflow followed here is presented in Figure 1.

2.3. Performance Metrics

The metrics used here to compare the prediction skill of different RF models are the confusion matrix, the receiver operator characteristic (ROC) curve, the area under the ROC curve (AUC), and the precision-recall (PR) curve. Furthermore, the average accuracy for all the categories for the RF is calculated on the given test data and labels.

The confusion matrix presents the proportions of actual versus predicted values for each marine heatwave category. We will number them here for simplicity: no MHW (0), moderate (1), strong (2) and severe/extreme (3). Correct predictions can be found on the diagonal of the confusion matrix.

We estimate the mean accuracy (MA) of the model, by averaging the correct predictions over the four MHW categories. That is, we average the proportions of cases that were correctly predicted as: 0, 1, 2 or 3.

We also compute the mean absence/presence accuracy (MA*) of the model, by treating the classifier as binary and aggregating the three MHW presence categories regardless of their severity:

$$MA^* = (TN^* + TP^*)/2 \tag{1}$$

Here, the rate of true negatives TN* corresponds to predicting absence of a MHW when there is none, and a true positive TP* corresponds to assigning a MHW (1, 2 or 3) given that there is presence of a MHW (1, 2 or 3).

The ROC curve presents the trade-off between the true positive rate (TPR) and false positive rate (FPR) for each category on different probability thresholds (Fawcett, 2006). Take class 1 for example, the TPR or recall (R) for this category is defined as:

$$TPR = R = \frac{TP}{TP + FN} \tag{2}$$

where TP is the number of true positives (i.e., predicting class 1 when it is true) and FN the number of false negatives (i.e., predicting class 0, 2 or 3 when class 1 is true). The false positive rate (FPR) is defined as:

$$FPR = \frac{FP}{FP + TN} \tag{3}$$

where FP is the number of false positives (i.e., predicting class 1 when class 0, 2 or 3 is true) and TN are the true negatives (i.e., not predicting class 1 when class 0, 2 or 3 is true).

The ROC curve can be summarized by the area under the ROC curve (AUC). The AUC is the probability of a classifier ranking a randomly chosen positive label higher than a randomly chosen negative one (Fawcett, 2006). The AUC is calculated as:

$$AUC(f) = \frac{\sum_{t_0 \in D^0} \sum_{t_1 \in D^1} \mathbf{1}\left[f(t_0) < f(t_1)\right]}{|D^0|.|D^1|} \tag{4}$$

where $1[f(t_0) < f(t_1)]$ is an indicator function returning 1 only if $f(t_0) < f(t_1)$ and 0 for the opposite, and D^0 and D^1 are the sets of negative and positive samples respectively (Calders & Jaroszewicz, 2007). An AUC of 1 indicates an ideal classifier that predicts categories perfectly. For a random classifier that cannot distinguish between categories, the AUC would be 0.5. One of the advantages of using AUC as a performance metric include the scale invariance, meaning that ranked predictions instead of absolute values are measured.

A preferred measure for evaluating imbalanced data is the precision-recall (PR) curve, which exposes differences between the compared classifiers that are not apparent in the ROC space (Saito & Rehmsmeier, 2015). Precision (P) represents the proportion of positive predictions that are true. For a given class, P is defined as:

$$P = \frac{TP}{TP + FP} \tag{5}$$

Overall, it has been shown that both ROC and PR curves contain important information regarding the performance of a given algorithm (Davis & Goadrich, 2006), and therefore both are used here. These quantities are also related to the $F1$ score, which is defined as the harmonic mean of precision (P) and recall (R):

$$F1 = 2\frac{P * R}{P + R} \tag{6}$$

The average precision summarizes a precision-recall curve as the weighted mean of precision achieved at each classification threshold, with the increase in recall from the previous threshold used as the weight:

$$AP = \sum_n \left(R_n - R_{n-1}\right) / P_n \tag{7}$$

where R_n and P_n are the precision and recall at the nth threshold.

We implement RF and K-fold cross validation using the scikit-learn package in Python (Pedregosa et al., 2011). Codes are available at https://github.com/KatGiamalaki/Random_Forest_for_MHW.

3. Results

3.1. Optimized Model Performance Metrics

We present results at weekly lead times, as they accurately capture the daily MHW patterns in the area. Time lags spanning from 2 to 365 days were tested here (Text S1; Figure S1a in Supporting Information S1). Based on this comparison, we focus on weekly time lags as a compromise between AUC score and time to take actions to mitigate potentially negative impacts (from a management perspective): a 1-week lag yields a high AUC score and leaves some time to mitigate the effects of a potential MHW event (compared to a 2-days lead that gives the best AUC score but a very limited window of time to implement mitigation strategies).

The performance metrics of the optimized RF model, specifically the confusion matrix, the ROC and the PR curves are presented in Figure 2. The confusion matrix presents the percentages of true and false positives and negatives for all categories. Figure 2a highlights the high accuracy of the RF model in predicting the absence of MHW events (82%), which is likely a consequence of the training data containing a majority of non MHWs. Labels with moderate MHW events are relatively well represented in the training set, and therefore correctly identified by the RF at a 57% rate. The performance of the model reduces when identifying areas where strong or severe/extreme events are recorded, the least represented categories in the training data set. The model correctly identifies strong MHWs with a rate of 13% and has zero skill in recognition of the severe/extreme class. The average accuracy (MA) of the model to predict the four MHW categories is 38%. Overall, the model tends to predict too few MHWs and to underestimate their severity when a presence of MHW is correctly predicted. It must be noted that although the RF mislabels the strong and severe MHW categories, it classifies them mostly as moderate events with a 68% and 59% accuracy for the strong and severe/extreme classes, respectively. Therefore, the model accuracy is higher when considering only presence/absence, and the optimized RF model can predict MHW occurrence with a mean absence/presence accuracy (MA*) of 76% at the selected weekly lead time.

Figure 2 presents the trade-off between the true and false positive rates (ROC curve, Figure 2b) and positive predictive value (PR curve, Figure 2c) using different probability thresholds. In the ROC curve, the peak performance is occurring in the top left corner, where TPR is 1 and FPR is 0. On the other hand, the weakest predictability area is around the diagonal dashed line, which represents random predictions. The ROC curves for the classes representing MHW absence, strong and severe/extreme events show that the model is performing well. However, for the moderate class the algorithm is under-performing with higher FPR for lower TPR of predictions. This is also shown in the confusion matrix (Figure 2a) where the moderate class is mislabeled as a no-event 35% of the times.

The PR curves for the classification are presented in Figure 2c. In the PR space the upper right corner is considered to represent ideal conditions, with high precision and high recall values. The RF model performs well for the MHW absence class, with high precision values even with increased recall. However, the precision drops substantially for the rest of the MHW categories, highlighting the fact that the performance of the RF model forecasting the severity of MHW events is weak. The difference between the ROC and PR curves is that the PR curves are only concerned with the accurate predictions of the classes, not taking into account the true negative predictions. As such, the PR curve captures the effect of negative predictions on the algorithm's performance and penalizes the incorrect labeling of points (Davis & Goadrich, 2006). Because of this difference the overall AUC is much higher than the area under the PR curve, highlighting the great effect that the class imbalance has on the performance of the RF model. As an attempt to improve performance of the model to predict strong and extreme MHWs, we have increased weights of those categories in splitting criteria (Text S2 in Supporting Information S1). Accuracy is slightly improved for the moderate and severe categories, but deteriorate results in the no MHW category (Text S2; Figure S2 in Supporting Information S1). As expected, the model performance is primarily driven by SST (Text S3; Figure S3 in Supporting Information S1). The location and time features turn out to also be important along with net heat flux (Figure S3 in Supporting Information S1).

To provide a baseline of comparison to the optimized RF model, we also fitted a simpler model based on multinomial logistic regression (Text S4; Figure S4 in Supporting Information S1). The logistic regression model achieves a lower total presence/absence accuracy of 47% (compared to 76% with the RF model), and poor skill at predicting any of the MHW categories (32% for moderate and 0% for both strong and extreme) (Figure S4 in Supporting Information S1).

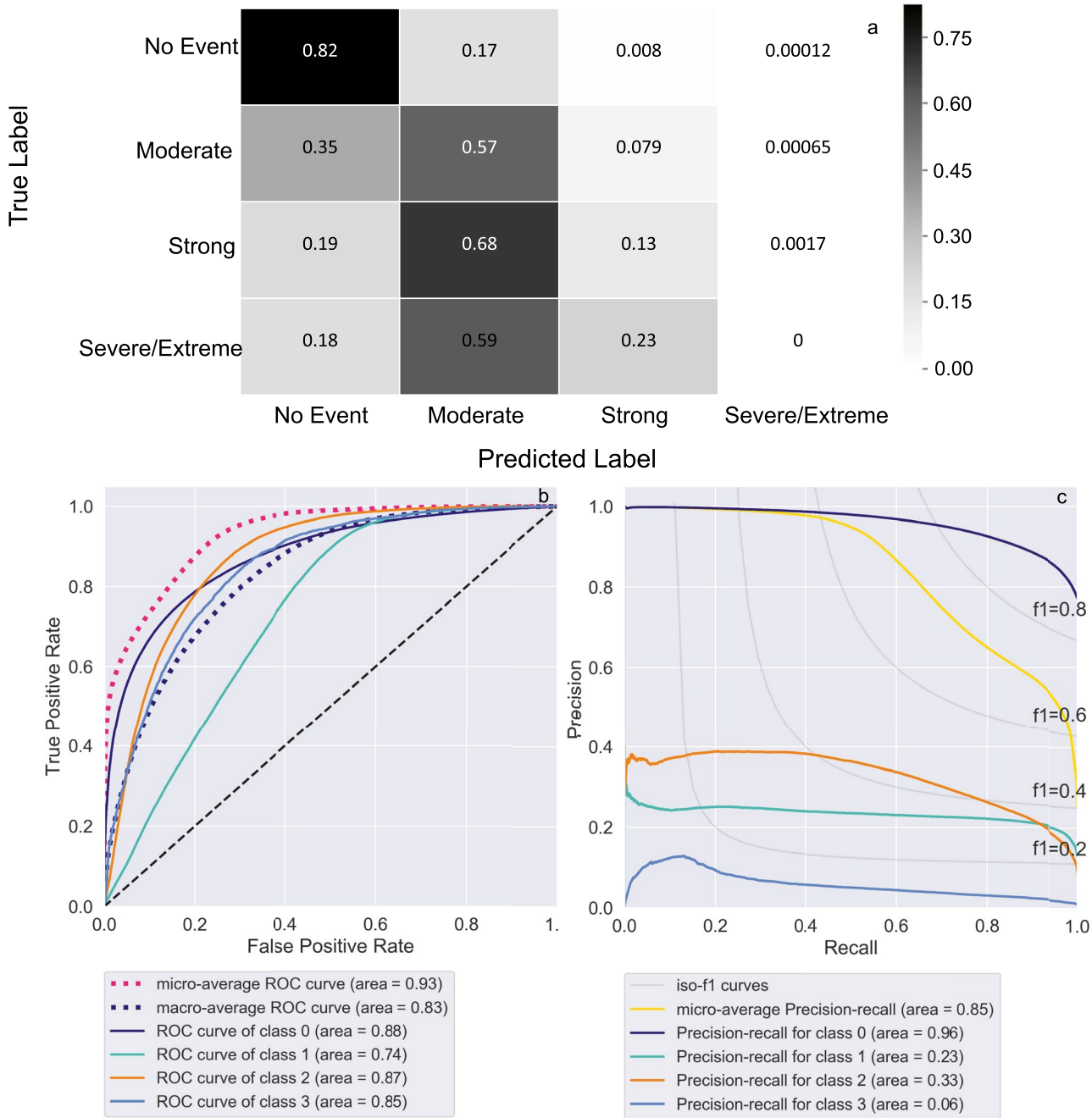

Figure 2. Performance metrics for the optimized RF model calculated on the test (unseen) data set. (a) The confusion matrix presents the rates of true and false positives and negatives. The confusion matrix shows skill in predicting the no-event and moderate MHW categories and poor skill in forecasting strong and severe/extreme MHWs. The RF model tends to predict strong and severe MHWs as moderate events. (b) Receiver Operating Characteristic (ROC) curve for different classification probability thresholds. The dashed diagonal line represents a random classifier. The micro-average provides the score of the aggregated contributions of all classes. The macro-average is the score computed independently for each class and then averaged over the number of classes, hence treating all categories equally. The micro-average is calculated by aggregating the contributions of all classes to compute the average score. (c) Precision-Recall curve for each MHW class on multiple classification probability thresholds. F1- scores represent the harmonic average of the precision and recall values. The iso-F1 curves contain all points in the precision/recall space whose F1 scores are the same. In (b and c), categories are numbered in the following order: class 0 (No Event), class 1 (Moderate), class 2 (Strong) and class 3 (Severe/Extreme).

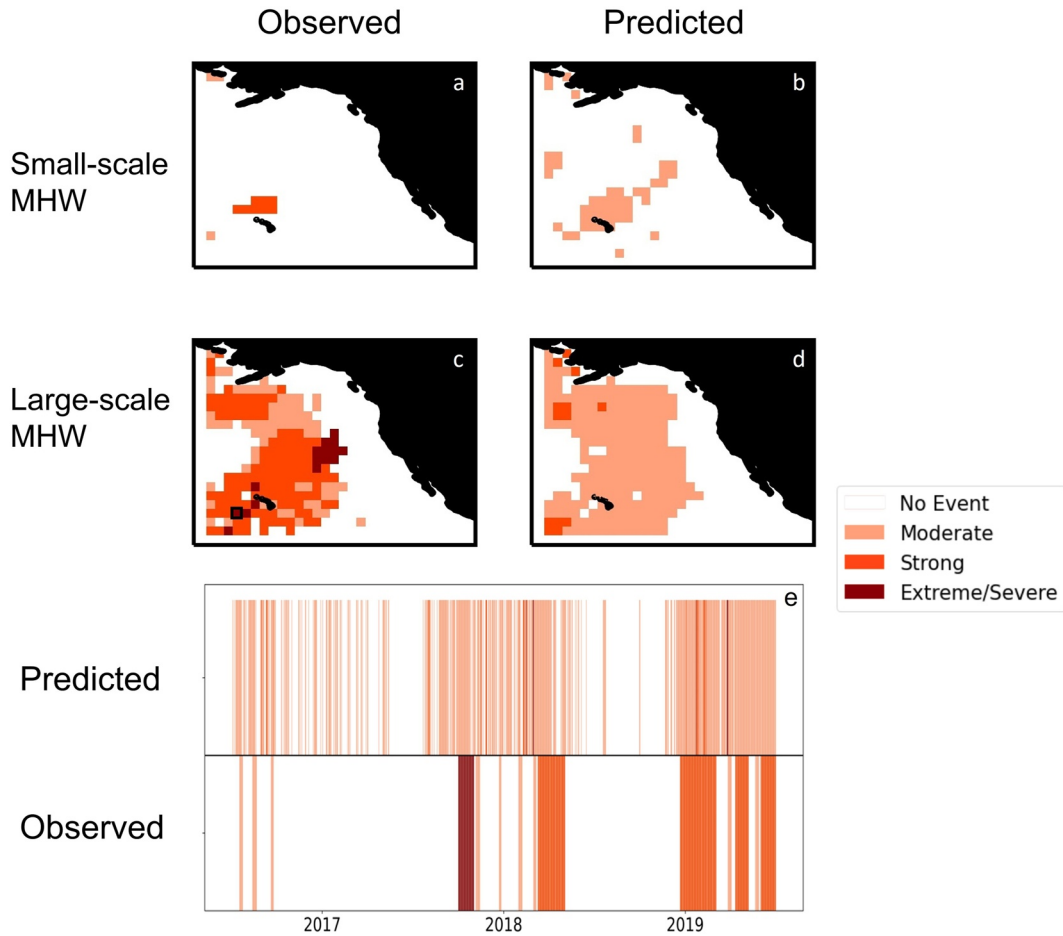

Figure 3. Examples of observed versus predicted MHWs in the testing data set using the optimized RF model at a weekly lead time. (a–d) Example observed and predicted conditions on two single days (13 October 2017 and 13 October 2019 that exhibit (a, b) small-scale MHW events and (c, d) large-scale MHWs. (e) Observed and predicted MHW time series for one location (pixel) represented by a black square in panel c) (longitude = −172.375 and latitude = 12.625) for the period 2017–2019.

3.2. Model Performance for Specific Days and Location

Two days with small-scale and large-scale MHW presence were selected as examples to compare the spatial patterns of observed and predicted MHWs (Figure 3). Overall, the general pattern of the small-scale (Figures 3a and 3b) and large-scale (Figures 3c and 3d) MHW events are well captured by the model, although some pixels with no MHW are predicted as moderate MHWs. This figure also shows that the RF model does not accurately capture the specific MHW categories. This is consistent with (Figure 2) and likely due to class imbalance. As such, the model has high predictive ability in capturing the presence or absence of extreme sea surface temperature events, without being able to identify their severity due to the limited number of strong and severe/extreme MHWs past occurrences.

The time series of observed and predicted MHW categories for one arbitrarily selected point in the Northeast Pacific are presented in Figure 3e. The comparison between the observed and predicted MHW occurrences in the figure highlights the overestimation of the moderate MHW category in the RF predictions. The RF model predicts the onset of the two strong and extreme/severe MHW events recorded in 2018 and 2019 in advance, although they are labeled as moderate events and underestimated.

Examples presented Figure 3 correspond to specific days and a specific location in the testing data set, respectively. To reveal any spatial/temporal bias in the model, metrics of model performance aggregated in space and time are presented in Text S5 in Supporting Information S1 (Figure S5 in Supporting Information S1).

4. Discussion and Conclusions

A random forest model is used here to assess the potential of forecasting marine heatwaves in northeast Pacific. We compare the skill of a base model on multiple lead times to find the optimum time lag for our predictions. Then, we validate the model using the selected week time lag and optimize the hyperparameters for the highest accuracy of the RF model. The optimized RF model presents an average 76% accuracy in predicting MHW presence/absence, however the MA drops down to 38% when predicting specific MHW categories. Our model can accurately forecast MHW patterns in the northeast Pacific a week in advance using sea surface temperature, surface air temperature, wind speed, sea level pressure and net heat fluxes. Lower skill is achieved in forecasting MHW categories with high severity. This could be due to the under-representation of these categories in the training dataset.

The RF model fitted here could be improved by including additional predictors in the model. For example, additional ocean interior variables, such as wave direction and current velocity, may improve the RF model predictions. Sea ice may also improve predictions for high latitude regions. Another improvement strategy could be to augment the amount of data used for training. This could be achieved by using a reanalysis data set with a higher spatial resolution (e.g., ERA5). In addition, using outputs from climate models to train the model has the potential to significantly augment the size of the training and testing datasets. The model could also be improved by computing uncertainties in predictions. In the present study, we do not formally integrate uncertainties in the predictor variables nor quantify prediction uncertainties. A RF is a non-parametric ensemble approach, and unlike traditional regression approaches, the variables and prediction errors are not quantified directly. Computing uncertainties will be important when implementing a real-time forecasting system and should be the focus of a future study. Research in the machine learning literature has focused on this issue (Mentch & Hooker, 2016). Similarly, reliability should be assessed before using the RF model with a different data set than the one used for training.

As the RF are not a pattern recognition algorithm, the lack of the dynamical component and physical linkages is highlighted by the fact that the physical parameters play a minor role in defining the predicted MHW categories here. Despite the simple configuration when compared to deep learning methods, the RF model provides skillful forecasts of MHW presence or absence in northeast Pacific. The accuracy of the RF in predicting patterns suggests that the application of a more sophisticated pattern recognition model should improve our ability to capture the spatial patterns, forecast the severity of MHW events and understand regional dynamics that may trigger such extreme conditions. Similarly, the high coherence of MHWs in space/time suggest that spatio-temporal statistical modeling approaches should be considered. Nevertheless, our results show promise in predicting marine temperature extremes in one of the most prominent regions in terms of climate interactions, fisheries and socio-economic importance.

Predictability of northeast Pacific MHWs on sub-seasonal, seasonal or longer time scales is desirable. Depending on the region, predictability relies on the different MHW driving processes including SST persistence, large scale atmospheric teleconnections and low-frequency climate modes (Holbrook et al., 2020). The presence of known preceding El-niño conditions have previously contributed to the increased predictability of the 2014–2016 MHW in the North Pacific using multi-model ensemble seasonal forecasts (Jacox et al., 2019). However, the model's skill weakens in the absence of the signature of the forcing mechanisms, for example, neutral El-Niño (Jacox et al., 2019, 2022). While this is beyond the goal of the present study, a model such as the one developed here could eventually complement such seasonal (and sub-seasonal) forecasts from numerical models in the northeast Pacific, and is computationally cheaper. Future work should focus on implementing a near real-time forecasting system based on a similar model architecture.

We predict MHWs on a weekly basis using a simple and interpretable model, which is less data and time expensive compared to complex pattern recognition models. In general, weekly (7–10 days) MHWs predictions are currently missing from the literature, although weekly forecasts of ocean conditions are widely available. Multiple sectors can benefit from such short-term forecasts, including fisheries and aquaculture, shipping and coastal water management (DeMott et al., 2021). Week lead warnings can assist fishery managers to avoid potential closures and prepare the industry for gear and labor modifications. Aquaculture managers can take protective measures against excessive warming or potential coral bleaching events at the aquaculture sites. Shipping companies will have the opportunity to plan accordingly for slowdowns to reduce vessel strike risks. Valuable

information can be provided to make prudent decisions in advance of potentially severe economic catastrophes. Pacific SSTs are monitored on multiple time scales (Saha et al., 2014), however predictability efforts are mainly focused on El Niño and La Niña. Although ENSO events are highly linked to the northeast Pacific MHWs (Holbrook et al., 2019), they are not always a triggering factor (Jacox et al., 2019). The northeast Pacific and California current are also closely monitored (e.g., California Current Integrated Ecosystem Assessment), due to the importance and high value of the ecological characteristics and resources of the area. Still, weekly forecasts of the intensity and spatial patterns of sea surface temperature extremes are not yet included in these efforts. Such short-term forecasts can generate opportunities to support mitigation of the consequences, providing information for short-term implementation of strategies, in order to avoid the likely ecological and economical catastrophes caused by a MHW event, even on a weekly temporal basis. Future work should focus on developing such models in other regions or the global ocean. Of course, predictors and predictability may vary depending on local driving processes.

Data Availability Statement

To label the MHW categories, we used the National Oceanic and Atmospheric Administration Optimum Interpolation Sea Surface Temperature (NOAA OISST v2.1) with a 1/4° spatial resolution from 1981 to 2019 (Banzon et al., 2020), downloaded from https://www.ncei.noaa.gov/products/optimum-interpolation-sst. The daily sea level pressure, surface air temperature, net heat flux and wind speed datasets were obtained by the National Centers for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) Reanalysis 1 (Kalnay et al., 1996), and downloaded from https://psl.noaa.gov/data/gridded/data.ncep.reanalysis.html. The Python interface of the code used to train, test and validate the random forest model, and reproduce the figures of the manuscript, is publicly available at https://doi.org/10.5281/zenodo.6533920.

Acknowledgments
We acknowledge financial support under NOAA Grant #NA18OAR4170073, California Sea Grant College Program Project # R/HCE-19PD, through NOAA'S National Sea Grant College Program, U.S. Department of Commerce. The statements, findings, conclusions and recommendations are those of the authors and do not necessarily reflect the views of California Sea Grant, NOAA or the U.S. Department of Commerce. We would like to thank National Oceanic and Atmospheric Administration and National Centers for Environmental Prediction/ National Center for Atmospheric Research for making their data publicly available. JXP acknowledges support from the Simons Foundation as a Simons Pivot Fellow.

References

AghaKouchak, A., Cheng, L., Mazdiyasni, O., & Farahmand, A. (2014). Global warming and changes in risk of concurrent climate extremes: Insights from the 2014 California drought. *Geophysical Research Letters*, *41*(24), 8847–8852. https://doi.org/10.1002/2014GL062308

Amaya, D. J., Bond, N. E., Miller, A. J., & Deflorio, M. J. (2016). The evolution and known atmospheric forcing mechanisms behind the 2013-2015 North Pacific warm anomalies. *US CLIVAR Variations*, *14*, 1–6.

Amaya, D. J., Miller, A. J., Xie, S., & Kosaka, Y. (2020). Physical drivers of the summer 2019 North Pacific marine heatwave—The Blob 2. 0. *Nature Communications*, *11*(1), 1903. https://doi.org/10.13140/RG.2.2.32308.19848/1

Anderson, G. J., & Lucas, D. D. (2018). Machine learning predictions of a multi-resolution climate model ensemble. *Geophysical Research Letters*, *45*(9), 4273–4280. https://doi.org/10.1029/2018GL077049

Baez-Villanueva, O., Zambrano-Bigiarini, M., Beck, H., McNamara, I., Ribbe, L., Nauditt, A., et al. (2020). RF-MEP: A novel random forest method for merging gridded precipitation products and ground-based measurements. *Remote Sensing of Environment*, *239*, 111606. https://doi.org/10.1016/j.rse.2019.111606

Banzon, V., Smith, T. M., Steele, M., Huang, B., & Zhang, H. (2020). Improved estimation of proxy sea surface temperature in the arctic. *Journal of Atmospheric and Oceanic Technology*, *37*(2), 341–349. https://doi.org/10.1175/JTECH-D-19-0177.1

Beillouin, D., Schauberger, B., Bastos, A., Ciais, P., & Makowski, D. (2020). Impact of extreme weather conditions on European crop production in 2018: Random forest - Yield anomalies. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1810), 20190510. https://doi.org/10.1098/rstb.2019.0510

Bond, N. A., Cronin, M. F., Freeland, H., & Mantua, N. (2015). Causes and impacts of the 2014 warm anomaly in the ne Pacific. *Geophysical Research Letters*, *42*(9), 3414–3420. https://doi.org/10.1002/2015GL063306

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Calders, T., & Jaroszewicz, S. (2007). Efficient AUC optimization for classification. Knowledge discovery in databases: PKDD 2007. *Lecture Notes in Computer Science*, 42–53. https://doi.org/10.1007/978-3-540-74976-9_8

Capotondi, A., Sardeshmukh, P. D., Di Lorenzo, E., Subramanian, A. C., & Miller, A. J. (2019). Predictability of US west coast ocean temperatures is not solely due to ENSO. *Scientific Reports*, *9*(1), 10993. https://doi.org/10.1038/s41598-019-47400-4

Cavole, L. M., Demko, A. M., Diner, R. E., Giddings, A., Koester, I., Pagniello, C., et al. (2016). Biological impacts of the 2013-2015 warm-water anomaly in the Northeast Pacific. *Oceanography*, *29*(2). https://doi.org/10.5670/oceanog.2016.32

Chen, C., Liaw, A., & Breiman, L. (2004). *Using random forest to learn imbalanced data*. University of California.

Chikamoto, Y., Timmermann, A., Widlansky, M. J., Balmaseda, M. A., & Stott, L. (2017). Multi-year predictability of climate, drought, and wildfire in southwestern North America. *Scientific Reports*, *7*(1), 6568. https://doi.org/10.1038/s41598-017-06869-7

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on machine learning* (pp. 233–240). Association for Computing Machinery. https://doi.org/10.1145/1143844.1143874

DeMott, C., Muñoz, A., Roberts, C., Spillman, C., & Vitart, F. (2021). The benefits of better ocean weather forecasting. *Eos*, *102*. https://doi.org/10.1029/2021EO210601

Drummond, C., & Holte, R. (2003). C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. In (pp. 1–8).

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Frölicher, T. L., & Laufkötter, C. (2018). Emerging risks from marine heat waves. *Nature Communications*, *9*(1), 650. https://doi.org/10.1038/s41467-018-03163-6

Herman, G. R., & Schumacher, R. S. (2018). Money doesn't grow on trees, but forecasts do: Forecasting extreme precipitation with random forests. *Monthly Weather Review*, *146*(5), 1571–1600. https://doi.org/10.1175/MWR-D-17-0250.1

Hobday, A. J., Alexander, L. V., Perkins, S. E., Smale, D. A., Straub, S. C., Oliver, E. C., et al. (2016). A hierarchical approach to defining marine heatwaves. *Progress in Oceanography*, *141*, 227–238. https://doi.org/10.1016/j.pocean.2015.12.014

Hobday, A. J., Oliver, E. C., Sen Gupta, A., Benthuysen, J. A., Burrows, M. T., Donat, M. G., et al. (2018). Categorizing and naming marine heatwaves. *Oceanography*, *31*(2), 162–173. https://doi.org/10.5670/oceanog.2018.205

Holbrook, N., Scannell, H. A., Sen Gupta, A., Benthuysen, J. A., Feng, M., Oliver, E. C., et al. (2019). A global assessment of marine heatwaves and their drivers. *Nature Communications*, *10*(1), 2624. https://doi.org/10.1038/s41467-019-10206-z

Holbrook, N., Sen Gupta, A., Oliver, E., Hobday, A., Benthuysen, J., Scannell, H., et al. (2020). Keeping pace with marine heatwaves. *Nature Reviews Earth & Environment*, *1*(9), 482–493. https://doi.org/10.1038/s43017-020-0068-4

Jacox, M. G., Alexander, M., Amaya, D., Becker, E., Bograd, S., Brodie, S., et al. (2022). Global seasonal forecasts of marine heatwaves. *Nature*, *604*(7906), 486–490. https://doi.org/10.1038/s41586-022-04573-9

Jacox, M. G., Alexander, M. A., Siedlecki, S., Chen, K., Kwon, Y., Brodie, S., et al. (2020). Seasonal-to-interannual prediction of North American coastal marine ecosystems: Forecast methods, mechanisms of predictability, and priority developments. *Progress in Oceanography*, *183*, 102307. https://doi.org/10.1016/j.pocean.2020.102307

Jacox, M. G., Tommasi, D., Alexander, M. A., Hervieux, G., & Stock, C. A. (2019). Predicting the evolution of the 2014–2016 California current system marine heatwave from an ensemble of coupled global climate forecasts. *Frontiers in Marine Science*, *6*, 497. https://doi.org/10.3389/fmars.2019.00497

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., et al. (1996). The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, *77*(3), 437–471. https://doi.org/10.1175/1520-0477(1996)077<0437:tnyrp>2.0.co;2

Kubat, M., & Matwin, S. (1997). Addressing the curse of imbalanced training sets: One-sided selection. In In *Proceedings of the fourteenth international conference on machine learning* (pp. 179–186). Morgan Kaufmann.

Laufkötter, C., Zscheischler, J., & Frölicher, T. L. (2020). High-impact marine heatwaves attributable to human-induced global warming. *Science*, *369*(6511), 1621–1625. https://doi.org/10.1126/science.aba0690

McCabe, R. M., Hickey, B. M., Kudela, R. M., Lefebvre, K. A., Adams, N. G., Bill, B. D., et al. (2016). An unprecedented coastwide toxic algal bloom linked to anomalous ocean conditions. *Geophysical Research Letters*, *43*(19), 10366–10376. https://doi.org/10.1002/2016GL070023

Mentch, L., & Hooker, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *Journal of Machine Learning Research*, *17*, 1–41. Retrieved from https://jmlr.org/papers/volume17/14-168/14-168.pdf

NOAA. (2017). State of the climate: National climate report for annual 2016 (Tech. Rep.).

Olita, A., Sorgente, R., Natale, S., Gaberšek, S., Ribotti, A., Bonanno, A., & Patti, B. (2007). Effects of the 2003 European heatwave on the central mediterranean sea: Surface fluxes and the dynamical response. *Ocean Science*, *3*(2), 273–289. https://doi.org/10.5194/os-3-273-2007

Oliver, E., Benthuysen, J. A., Darmaraki, S., Donat, M. G., Hobday, A. J., Holbrook, N. J., et al. (2021). Marine heatwaves. *Annual Review of Marine Science*, *13*(1), 313–342. https://doi.org/10.1146/annurev-marine-032720-095144

Oshiro, T., Perez, P., & Baranauskas, J. A. (2012). How many trees in a random forest? In P. Perner (Ed.), *Machine learning and data mining in pattern recognition* (pp. 154–168). Springer Berlin Heidelberg.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Probst, P., & Boulesteix, A. (2018). To tune or not to tune the number of trees in random forest. *Journal of Machine Learning Research*, *18*, 1–8.

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., et al. (2014). The ncep climate forecast system version 2. *Journal of Climate*, *27*(6), 2185–2208. https://doi.org/10.1175/JCLI-D-12-00823.1

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Schlegel, R. (2020). Marine heatwave tracker. https://doi.org/10.5281/zenodo.3787872

Seager, R., Hoerling, M., Schubert, S., Wang, H., Lyon, B., Kumar, A., et al. (2015). Causes of the 2011–2014 California drought. *Journal of Climate*, *28*(18), 6997–7024. https://doi.org/10.1175/JCLI-D-14-00860.1

Sen Gupta, A., Thomsen, M., Benthuysen, J. A., Hobday, A. J., Oliver, E., Alexander, L. V., et al. (2020). Drivers and impacts of the most extreme marine heatwaves events. *Scientific Reports*, *10*(1), 19359. https://doi.org/10.1038/s41598-020-75445-3

Smale, D. A., Wernberg, T., Oliver, E. C. J., Thomsen, M., Harvey, B. P., Straub, S. C., et al. (2019). Marine heatwaves threaten global biodiversity and the provision of ecosystem services. *Nature Climate Change*, *9*(4), 306–312. https://doi.org/10.1038/s41558-019-0412-1

Sun, T., & Okumura, Y. M. (2019). Role of stochastic atmospheric forcing from the south and North Pacific in tropical Pacific decadal variability. *Journal of Climate*, *32*(13), 4013–4038. https://doi.org/10.1175/JCLI-D-18-0536.1

Vogel, E., Donat, M. G., Alexander, L. V., Meinshausen, M., Ray, D. K., Karoly, D., et al. (2019). The effects of climate extremes on global agricultural yields. *Environmental Research Letters*, *14*(5), 054010. https://doi.org/10.1088/1748-9326/ab154b

Vogt, L., Burger, F. A., Griffies, S. M., & Frölicher, T. L. (2022). Local drivers of marine heatwaves: A global analysis with an Earth system model. In *Frontiers in climate* (Vol. *4*). https://doi.org/10.3389/fclim.2022.847995