DAtect

Forecasting Domoic Acid Levels from Harmful Algal Blooms along the Pacific Northwest Coast

Anson Chen | Tesla Stem High School | Redmond WA, USA

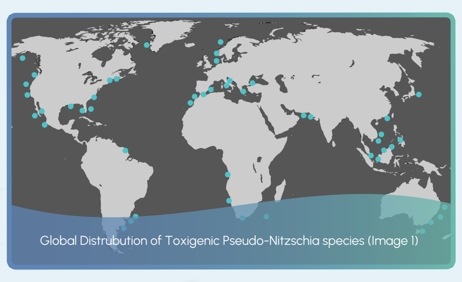
1 Background

Domoic Acid Overview

Domoic acid is a neurotoxin produced by harmful algal blooms (HABs) of marine diatoms in the genus Pseudo-Nitzschia. Consuming shellfish contaminated with Domoic Acid causes the neurological condition Amnesic Shellfish Poisoning, which is potentially fatal to not only marine birds and mammals, but also to humans.[1]

A Global Threat

- Toxigenic Pseudo-Nitzschia blooms have been found worldwide.[2]
- No visual cues - no water discoloration or visible fish kills.[3]
- Limited link between bloom size and Pseudo-Nitzschia count and Domoic Acid levels[4]
- Domoic Acid is weakly correlated with environmental metrics regionally or site-specific.[5]


Global Distrubution of Toxigenic Pseudo-Nitzschia species (Image 1)


Razor Clams along the Washington Coast (Image 2)

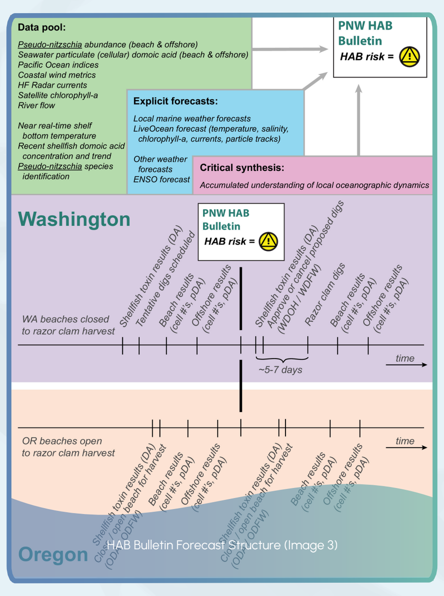
A Vulnerable Coast

- Strong upwellings and warm water from ENSO/PDO cause regular blooms in US Pacific Northwest.[6]
- Climate change expands HABs' size and duration, such as the unprecedented 2015 bloom which shut down shellfisheries for up to 5 months, resulting in $97.5 million loss for dungeness crab industry.[7]
- Warming water allow more toxigenic Pseudo-Nitzschia species (P.australis) to move into the US Pacific Northwest.[8]

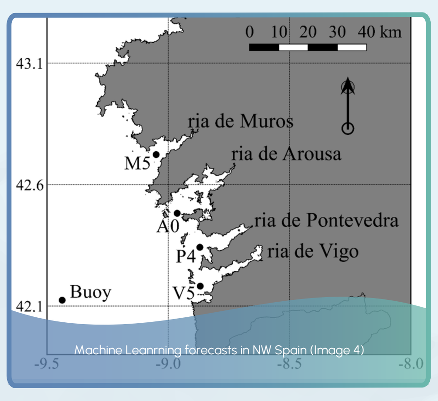
Current Forecasting

PNW HAB Bulletin[9]

- Provides a regionwide forecast for Domoic Acid risk levels since spring 2017 for the Washington State and Oregon coast.
- Uses a combination of shore-based sampling data, satellite imagery, weather forecasts, and oceanic simulations in its analysis.
- Provide biweekly to monthly easily digestible portable electronic two-page reports.
- High historic accuracy with qualitative analysis-based forecasting, but lacks any quantitative model support.



HAB Bulletin Forecast Structure (Image 3)



Machine Leanrning forecasts in NW Spain (Image 4)

ML-based Forecasting

- Prior Pseudo-Nitzschia models in NW Mediterranean, Ireland, NW Spain used limited regional geographic and time series data, and didn't predict Domoic Acid.[10][11]
- Previous Gulf of Maine Saxitoxin forecast model highly beneficial for stakeholders, but Saxitoxins are easier to predict.[12]

2 Model Objectives

"How can I quantitatively forecast Domoic Acid high-risk events in advance by using multiple moderately correlated environmental metrics and machine learning models?"

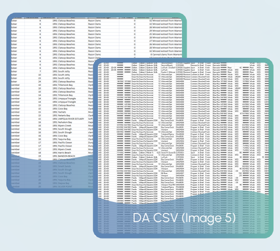
engineering goal #1 **accuracy**	DAtect model forecasts provide accurate risk level predictions, significantly outperforming traditional statistical models.
engineering goal #2 **data variety**	DAtect model utilizes a comprehensive dataset spanning at least 20 years, incorporating a wide range of abiotic and biotic data points from sites across the entire Pacific Northwest region.
engineering goal #3 **timeliness**	DAtect model accurately forecasts risk levels at least one week in advance on a regular basis in an easily accessible portable format. Similar to a weather forecast, but for high domoic acid risk.
engineering goal #4 **efficiency**	DAtect model forecasts are computed weekly, taking significantly less time than physics-based transport models or qualitative analysis. The model can efficiently retrain on new data.
engineering goal #5 **robustness**	DAtect model accounts for seasonality, lagging data, and maintains high accuracy across diverse oceanic conditions. It automatically determines the importance of different envirnomental features.
engineering goal #6 **flexibility**	DAtect model can accommodate historical data with significant clumping and gaps, as well as limited real-time data availability.

3 Methodology

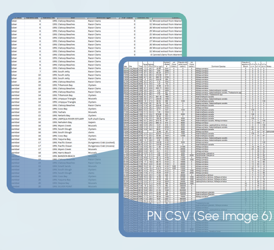
Phase 1

Data Acquisition

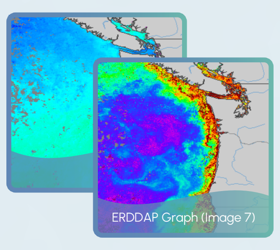
Data was collected from various sources, including records requests to the ORHAB partnership, Washington Department of Health, and Oregon Department of Fish and Wildlife. Additional data was downloaded through NOAA ERDDAP, NANOOS, and USGS public servers. The collected data was processed and consolidated, totaling more than nine gigabytes in size.


DA CSV (Image 5)

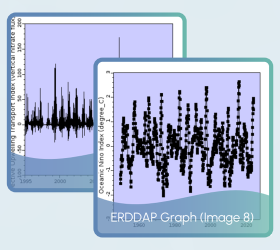
Domoic Acid (WA DoH, OR DOFW)


PN CSV (See Image 6)

Pseudo-Nitzschia (ORHAB, OR DOFW)


ERDDAP Graph (Image 7)

Satellite Imagery (NOAA ERDDAP)

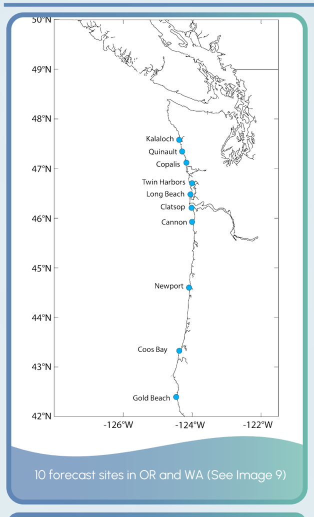

ERDDAP Graph (Image 8)

Oceanic Conditions (NOAA, USGS, NANOOS)

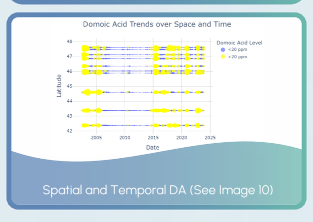
Phase 2

Data Processing

The data was first localized temporally and spatially, cleaned, gap-filled, and merged into a single CSV file with 10,660 rows across 14 features. This combined CSV file was further processed through feature engineering, including normalization, generation of time and space features, lag feature generation, and risk determination.



10 forecast sites in OR and WA (See Image 9)



Spatial and Temporal DA (See Image 10)

Preprocessing

- Convert all dates to Year-Week
- Localize data to a specific week and coastal site.
- Gap-fill missing weeks in 2002-2023 with linear interpolation and zeroes.
- Unsupervised Agglomerative Clustering for Satellite Imagery
- Merge all fourteen features to a single CSV file.

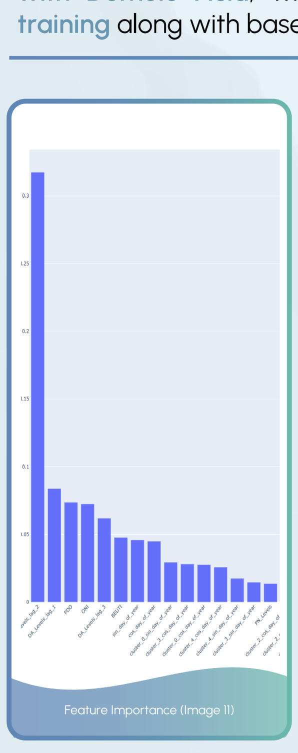
Feature Engineering

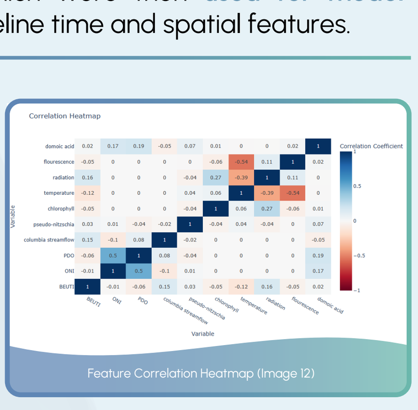
- K-Mean Clustering, Cyclical feature generation and Lag feature generation for Domoic Acid trends over space and time.
- Normalization and Multivariate interactions between various environmental features, temporal trends, and spatial trends.
- Domoic Acid risk category generation based on the 20 ppm federal limit and the HAB Bulletin.

Phase 3

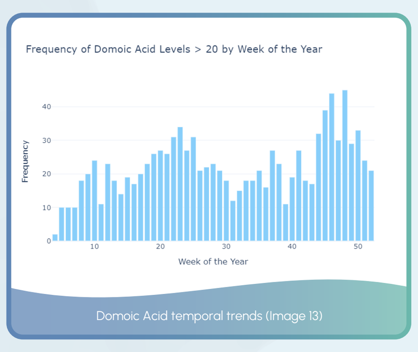
Data Selection

To optimize the dataset before model training, reduce training time, and improve the accuracy of Domoic Acid forecasts, feature selection techniques like linear correlation and tree-based analysis were used. This identified the top ten most important features in relation with Domoic Acid, which were then used for model training along with baseline time and spatial features.


Feature Importance (Image 11)



Feature Correlation Heatmap (Image 12)

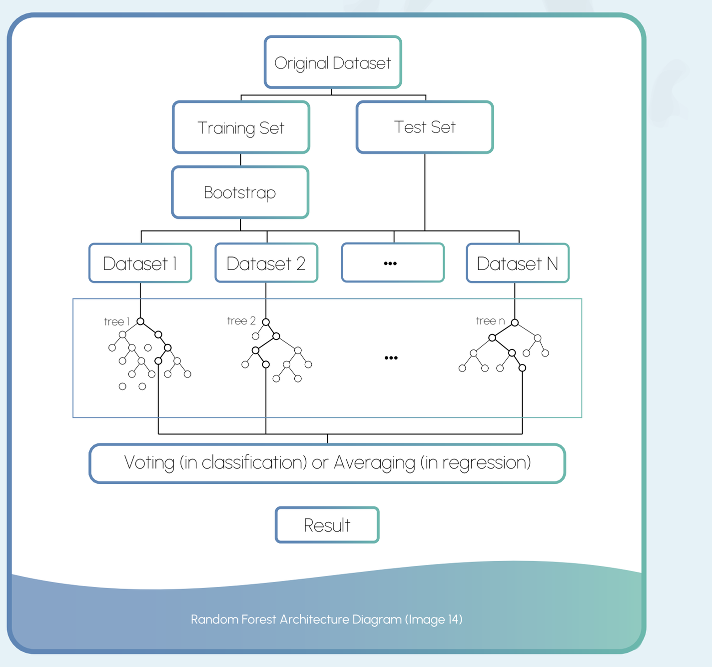


Domoic Acid temporal trends (Image 13)

Phase 4

Data Modeling

Three major categories were tested using free and open-source Python libraries. The first category were baseline statistical models. The second category were supervised Machine Learning models, which are highly efficient and flexible and can recognize non-linear patterns. The final category included Deep Learning Neural Network models, which are good at recognizing complex patterns but require specific data shapes.



Random Forest Architecture Diagram (Image 14)

4 Results & Analysis

2132 Forecasts Made

r^2 **0.59** (0.35)

- r^2 represents the proportion of the variance for the dependent variable that's explained by independent variables in a regression model. It ranges from 0 to 1, where 1 indicates a perfect fit.
- The r^2 of DAtect (Random Forest Regression) is 0.59, while the r^2 of the statistical baseline model (multiple linear regression) was 0.35.
- Formula: $1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

RMSE **10.04** (12.62)

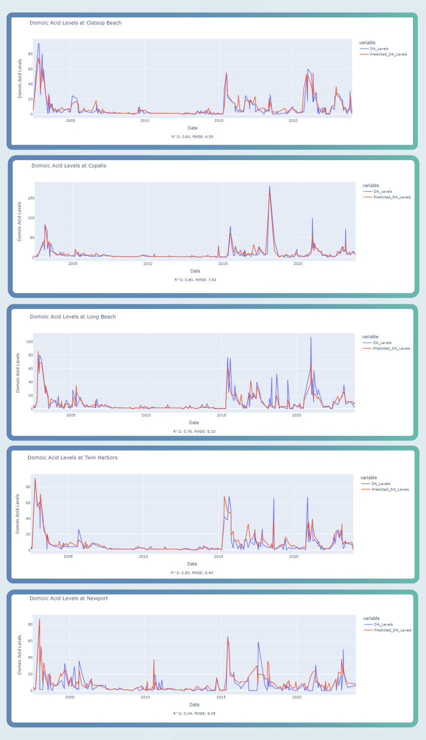
- RMSE is the differences between values predicted by a model and the actual values observed. It provides an estimate of how well the model fits the data, with lower values indicating better fit.
- The RMSE of DAtect (Random Forest Regression) is 10.04, while the RMSE of the statistical baseline model (multiple linear regression) was 12.62.
- Formula: $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$

Accuracy **82%** (69%)

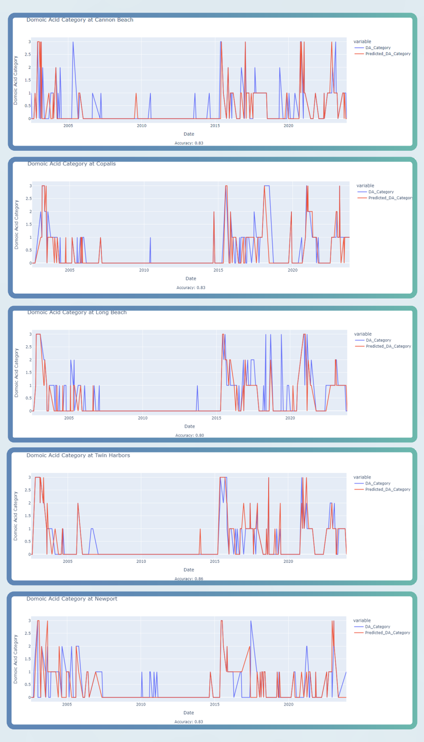
- Accuracy is the ratio of correctly predicted instances (or forecasts) to the total instances (or forecasts) done in the dataset.
- DAtect (Random Forest Classification) was 82% accurate, while the statistical baseline model (logistic regression) was 69% accurate.
- Formula: $\frac{\textit{Number of Correct Predictions}}{\textit{Total Number of Predictions}}$

Example Forecasts

Domoic Acid Level



Domoic Acid Category



DAtect Analysis

- The DAtect Model, a supervised Machine Learning model employing Random Forest Regression and Classification, outperformed traditional baseline statistical models in most of the ten sites used for forecasting particulate domoic acid concentrations. As the vast majority of coastal sites around the world at risk of high domoic acid levels have relatively similar oceanic conditions, the model has the potential for widespread application.
- While the model occasionally underestimated values, and a few forecasts at sites such as Coos Bay were not significantly better than traditional statistical models, the DAtect Model generally demonstrated accurate performance in determining whether the particulate domoic acid concentration exceeded the federal regulatory limit. Although Random Forest is a black box model and doesn't provide a direct cause and effect relationship, by forecasting domoic acid risk in advance, it can help beach managers and fisheries across the Pacific coast minimize economic harm while still protecting public health.
- The model retains its accuracy over a wide range of geographic locations and varying oceanic conditions, with the primary limiting factor being the availability of continuous Domoic Acid and Pseudo-Nitzschia data, especially in Oregon. Additionally, the reduced accuracy of the AQUA-MODIS 3.0 satellite spectrometer measurements near shore can affect the model at the coastal sites.

5 Discussion & Future Work

Significance

DAtect is the first one-week forecast model that predicts a quantitative index for the Domoic Acid risk category and Domoic Acid level, which can thus be integrated into the PNW HAB bulletin.

Accuracy
The DAtect model was significantly more accurate at forecasting Domoic Acid concentration level and threat category than baseline statistical forecasting.

Data Variety
The DAtect model uses a wide range of data including shore-based sampling, satellite imagery, and oceanic simulations.

Timeliness
The DAtect model generates forecasts for each of the ten sites on a weekly basis, using previously available data.

Efficiency
The DAtect model requires less than 10 seconds to train or retrain after optimizing the dataset.

Robustness
The DAtect model was trained on over 20 years capturing constantly changing oceanic conditions along the US Pacific Northwest coast.

Flexibility
The DAtect model automatically adjusts for gaps in the data during preprocessing, and Random Forest is optimized for clumped datasets.

01 The DAtect model's dataset can be expanded with real-time continuous shore-based and offshore Pseudo-Nitzschia and Domoic Acid data, which integrates better with satellite spectrometer data. This would allow the model to train on more edge cases, increasing accuracy given the lack of absolutes in the environment.

02 The DAtect model's dataset could be expanded with weather forecasting and particle transport oceanic simulations like the UW LiveOcean model. These are currently used in the qualitative analysis by the PNW HAB Bulletin and have potential for tracking spatial trends related to Domoic Acid and Pseudo-Nitzschia Blooms.

03 The DAtect model's accuracy could potentially improve by using a customized neural network model, as a more continuous dataset would enable the usage of neural network based models, potentially leading to more accurate and reliable forecasts.

04 To provide beach managers with easier access to forecasts and real-time Domoic Acid and satellite data, a web app could be developed, allowing them to integrate the model with the HAB Bulletin into their beach closure decisions, as well as upload shore data to update the model.

6 References

[1] Bates, S. S., Hubbard, K. A., Lundholm, N., Montresor, M., & Leaw, C. P. (2018). Pseudo-nitzschia, Nitzschia, and domoic acid: New research since 2011. Harmful Algae, 79, 3–43. https://doi.org/10.1016/j.hal.2018.06.001

[2] Trainer, V. L., Bates, S. S., Lundholm, N., Thessen, A. E., Cochlan, W. P., Adams, N. G., & Trick, C. G. (2012). Pseudo-nitzschia physiological ecology, phylogeny, toxicity, monitoring and impacts on ecosystem health. Harmful Algae, 14, 271–300. https://doi.org/10.1016/j.hal.2011.10.025

[3] Schreiber, S., Hanisak, M. D., Perricone, C. S., Fonnegra, A. C., Sullivan, J., & McFarland, M. (2023). Pseudo-nitzschia species, toxicity, and dynamics in the southern Indian River Lagoon, FL. Harmful Algae, 126, 102437. https://doi.org/10.1016/j.hal.2023.102437

[4] Pan, Y., Bates, S. S., & Cembella, A. D. (1998). Environmental stress and domoic acid production by Pseudo-nitzschia: a physiological perspective. Natural Toxins, 6(3-4), 127–135. https://doi.org/10.1002/(sici)1522-7189(199805/08)6:3/4%3C127::aid-nt9%3E3.0.co;2-2

[5] Trainer, V. L., Hickey, B. M., Lessard, E. J., Cochlan, W. P., Trick, C. G., Wells, M. L., MacFadyen, A., & Moore, S. K. (2009). Variability of Pseudo-nitzschiaand domoic acid in the Juan de Fuca eddy region and its adjacent shelves. Limnology and Oceanography, 54(1), 289–308. https://doi.org/10.4319/lo.2009.54.1.0289

[6] McKibben, S. M., Peterson, W., Wood, A. M., Trainer, V. L., Hunter, M., & White, A. E. (2017). Climatic regulation of the neurotoxin domoic acid. Proceedings of the National Academy of Sciences, 114(2), 239–244. https://doi.org/10.1073/pnas.1606798114

[7] Moore, S. K., Dreyer, S. J., Ekstrom, J. A., Moore, K., Norman, K., Klinger, T., Allison, E. H., & Jardine, S. L. (2020). Harmful algal blooms and coastal communities: Socioeconomic impacts and actions taken to cope with the 2015 U.S. West Coast domoic acid event. Harmful Algae, 96, 101799. https://doi.org/10.1016/j.hal.2020.101799

[8] Trainer, V. L., Kudela, R. M., Hunter, M. V., Adams, N. G., & McCabe, R. M. (2020). Climate Extreme Seeds a New Domoic Acid Hotspot on the US West Coast. Frontiers in Climate, 2. https://doi.org/10.3389/fclim.2020.571836

[9] McCabe, R. M., Hickey, B. M., & Trainer, V. L. (2023). The Pacific Northwest Harmful Algal Blooms Bulletin. Harmful Algae, 127, 102480. https://doi.org/10.1016/j.hal.2023.102480

[10] Aláez, F. M. B., Palenzuela, J. M. T., Spyrakos, E., & Vilas, L. G. (2021). Machine Learning Methods Applied to the Prediction of Pseudo-nitzschia spp. Blooms in the Galician Rias Baixas (NW Spain). ISPRS International Journal of Geo-Information, 10(4), 199. https://doi.org/10.3390/ijgi10040199

[11] Yu, P., Gao, R., Zhang, D., & Liu, Z.-P. (2021). Predicting coastal algal blooms with environmental factors by machine learning methods. Ecological Indicators, 123, 107334. https://doi.org/10.1016/j.ecolind.2020.107334

[12] Record, N. R., Evanilla, J., Kohl Kanwit, Burnell, C., Cartisano, C., Lewis, B. J., MacLeod, J., Tupper, B., Miller, D. W., Tracy, A. T., White, C., Moretti, M., Hamilton, B., Barner, C., & Archer, S. D. (2022). Benefits and Challenges of a Stakeholder-Driven Shellfish Toxicity Forecast in Coastal Maine. Frontiers in Marine Science, 9. https://doi.org/10.3389/fmars.2022.923738

EAEV 057