Enhancing the Use of Water Quality Data to Support Reclassification Decisions

Case Study: A Statistical Defense in Favor of Reclassifying Chandler Bay, ME as Class SA

Prepared By Remote Ecologist:

Darby Pochtar, Ph.D, Jason Krumholz, Ph.D, and David Hudson, Ph.D

Executive Summary

Maine's marine water classifications (SA, SB, SC) set the legal baseline for water-quality protection and determine how antidegradation requirements are implemented under both state and federal law. Most coastal waters are SB by default, which makes the ability to accurately identify waters that meet SA standards essential to protect "higher water quality" where it exists.

This report uses statewide water quality data to build a screening model that learns the multivariate environmental patterns historically associated with SA versus SB waters. The model is then applied—without refitting or tuning—to an independent dataset from Chandler Bay collected under the Kingfish Company permit monitoring requirements. The purpose is not to replace Maine's statutory classification process, but to evaluate whether observed Chandler Bay conditions are more consistent with statewide historical SA or SB waters.

Key findings are as followed:

1. A statewide random forest model can meaningfully distinguish historical SA from SB waters based on multivariate water quality patterns and monitoring structure.

2. Chandler Bay observations fall firmly in the SA-like region of the model's probability space. **Chandler Bay's predicted probability of SA-like conditions is higher than the median of historically designated SA waters statewide**.

3. The model relies strongly on both (1) measured environmental conditions and (2) monitoring-regime structure (whether key parameters were measured). This reflects a real and policy-relevant feature of Maine's monitoring landscape: higher-quality and/or priority waters tend to be monitored more intensively for regulatory-relevant variables.

4. A sensitivity analysis removing all monitoring-indicator variables shows that Chandler Bay remains strongly SA-like based on measured chemistry alone, though the strength of the signal decreases. This indicates that the sociopolitical factors impacting monitoring structure amplify, but do not create, the observed patterns.

5. These results underscore a critical implementation problem for 38 M.R.S. § 464(4)(F)(4): without consistent statewide measurement of classification-relevant parameters, Maine cannot reliably identify where actual water quality exceeds the minimum standards of the next highest class, which poses serious reclassification challenges.

6. **Based on these findings, we request that LD 2187 be amended to include reclassification of Chandler Bay as "SA" prior to passage.**

Background

Marine Classification and Regulatory Significance

Maine's coastal waters are classified under state law (38 M.R.S. §§ 464, 465-B) in accordance with the federal Clean Water Act (33 U.S.C. § 1313). These classifications (SA, SB, and SC) establish the designated uses and water quality standards that apply to each waterbody. Under both federal and state law, discharge permits must include limitations stringent enough to meet water quality standards and protect designated uses (33 U.S.C. § 1311(b)(1)(C); 40 C.F.R. § 122.4(d); 38 M.R.S. § 414-A). Therefore, marine classification is not simply descriptive; it establishes the legal benchmark against which environmental protection and regulatory decisions are measured.

Under 38 M.R.S. § 465-B(1), Class SA is Maine's highest marine classification and applies to waters that are outstanding natural resources and should be preserved because of their ecological, social, scenic, economic, or recreational importance. Class SB is the second-highest classification and applies to waters that must support recreation, fishing, aquaculture, navigation, and unimpaired marine habitat (38 M.R.S. § 465-B(2)).

Antidegradation and the Protection of Higher Water Quality

Maine's antidegradation policy (38 M.R.S. § 464(4)(F)) requires protection of existing uses and the water quality needed to maintain them. Critically, § 464(4)(F)(4) provides that when the "actual quality of a classified water exceeds the minimum standards of the next highest classification, that higher water quality must be maintained and protected", and "the board shall recommend to the Legislature that the water be reclassified" accordingly. Given that most coastal waters are classified as SB by default under 38 M.R.S. § 469, the practical effectiveness of antidegradation depends on Maine's ability to (1) detect waters whose measured conditions exceed SB and align with SA standards, and (2) maintain and protect those conditions through classification accuracy and permitting decisions. However, achieving this requires consistent and comparable monitoring of the water quality parameters that differentiate marine classifications.

Purpose and Research Questions

Marine classification reflects both environmental conditions and regulatory considerations. Accordingly, the modeling framework developed in this report is intended to assess consistency between observed water quality conditions and historically classified waters, rather than to replace or supersede Maine's statutory classification process.

To evaluate how water quality conditions relate to marine classification at a statewide scale, we developed a data-driven modeling framework using a random forest approach. This framework was designed to learn how combinations of water quality measurements historically align with class SA and class SB designations across coastal Maine and to place individual waterbodies within that broader environmental context. By focusing on multivariate patterns rather than single parameters, the approach reflects the reality that marine classification is influenced by overall environmental conditions rather than any one measurement in isolation.

Within this framework, Chandler Bay is treated as a case study. Water quality observations collected throughout Chandler Bay under the Kingfish permit monitoring program (permit #ME0037559) [1] were evaluated as an independent dataset and compared against statewide patterns learned from historically designated SA and SB waters. The central research question addressed in this report is, “Given water quality conditions measured throughout Chandler Bay, Maine, are those conditions more consistent with historically designated SA waters than with historically designated SB waters?”

The objective of this analysis is not to assign or modify regulatory classifications, but to generate quantitative evidence regarding whether Chandler Bay’s observed environmental conditions align more closely with the characteristics of SA waters than SB waters. In doing so, the analysis provides a scientifically grounded basis for evaluating whether Chandler Bay may warrant closer review under Maine’s antidegradation framework, including the provisions of 38 M.R.S. § 464(4)(F)(4) concerning waters whose actual quality exceeds the minimum standards of their assigned classification.

Methods

All data curation and analyses were conducted in R Studio [2]. Reproducible code and outputs are provided in Appendix 1.

Statewide Historical Dataset

Water-quality observations were compiled from multiple monitoring programs and harmonized into a single analytical dataset. The two data sources that had the best coverage of regulatory-relevant water quality parameters were from the National Water Quality Monitoring Council (NWQMC) [3] and Maine Department of Environmental Resources Environmental and Geographic Analysis Database (EGAD) [4]. Each observation included location, sampling date, and available physical, chemical, and microbial measurements, including dissolved oxygen, salinity, temperature, pH, nitrogen concentrations (i.e., ammonia), and indicator bacteria (i.e., most probable number (MPN) of enterococci per 100mL). Observations that contained only temperature and salinity measurements and no additional water quality variables were excluded from the analysis to ensure that model inputs reflected parameters relevant to marine

classification decisions. While temperature and salinity provide important environmental context, they are not used independently to determine marine classification and do not, on their own, distinguish among SA, SB, or SC waters. Requiring at least one additional water-quality parameter per observation ensured that model training was informed by variables directly tied to regulatory standards and water quality conditions. Variables that were infrequently measured or inconsistently defined across monitoring programs (e.g., chlorophyll, light availability, nitrate and nitrite, total nitrogen, and total phosphorus) were excluded to maintain interpretability across the combined dataset.

The response variable was marine classification, categorized as SA, SB, or SC. Marine classification was assigned by overlaying sampling observations on the official Maine DEP statutory water classification geographic information system (GIS) map [5]. Observations that did not fall within a classified marine area were excluded from analysis. Since the purpose of this report was to isolate and evaluate the environmental gradient separating SA and SB waters, SC observations were excluded from our analyses. Removing SC observations allowed the models to focus on the more subtle distinctions between adjacent regulatory classes most relevant to Chandler Bay's current designation. In addition, SC observations represented a very small proportion of the available dataset (3.6% of all observations) and following the removal of observations lacking all water-quality measurements except temperature and salinity, SC observations accounted for only 1.8% of the remaining data. Given their limited representation and distinct regulatory status, inclusion of SC observations would have added noise without improving the model's ability to resolve the SA–SB decision boundary. All included observations were then mapped to observe overall sampling coverage (Figure 1).

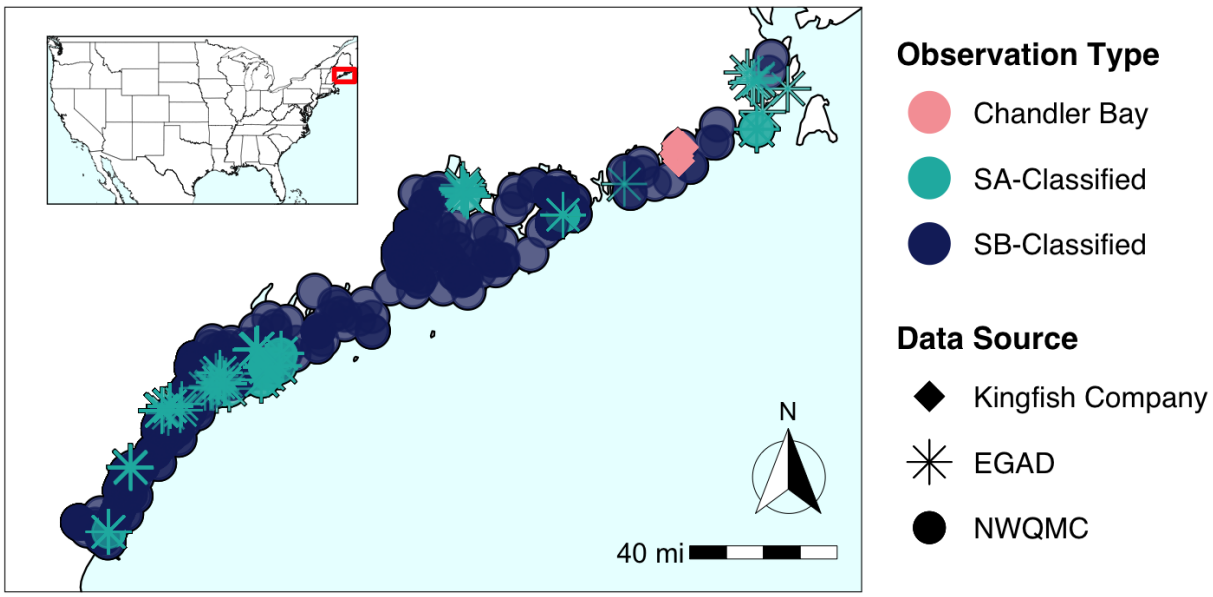


Figure 1. Map of the water quality monitoring locations across coastal Maine used in our analyses, categorized by marine classification and data source. Points represent individual sampling locations, with darker shading indicating a greater number of observations at a given location (i.e., higher sampling intensity).

Predictor Variables and Missing Data Handling

Predictor variables were selected to represent key aspects of coastal water quality, including physical conditions (temperature and salinity), dissolved oxygen, pH, microbial indicators, and nutrient-related parameters where available. Since data was compiled from multiple monitoring programs with differing objectives and sampling frequencies, many predictors were not measured for every observation. These missing values primarily reflect differences in monitoring design rather than data collection errors.

Instead of removing incomplete observations or replacing missing measurements with estimated values, missing data were left as collected and handled directly within the modeling framework. For each predictor, the model was given two types of information: (1) the measured value when available and (2) a simple indicator noting whether that variable was measured for a given observation (yes/no). This allowed the model to use both the environmental measurements and the monitoring patterns themselves, while retaining all available observations in the analysis and avoiding artificial value substitution.

Modeling Framework

Prior to model fitting, the dataset was partitioned into training and testing subsets using an 80/20 stratified split. Stratification was performed on the response variable (SA vs SB) to preserve class proportions in both subsets and ensure unbiased evaluation of predictive performance [6]. The resulting training set was used to fit the model, while the held-out test set was reserved for out-of-sample performance evaluation.

To evaluate how multivariate water quality conditions relate to marine classification, we implemented random forest models [7]. The random forest model works by building many simple decision trees and combining their results. Each tree examines the data in a slightly different way, and the final prediction reflects the overall agreement across all trees. By averaging across many trees, the model produces stable and reliable results rather than relying on a single decision rule. The model was configured to generate probabilities rather than only categorical classifications. For each observation, the model returned a value between 0 and 1 representing how strongly the measured water quality conditions resemble those historically associated with SA waters. Values closer to 1 indicate strong similarity to typical SA conditions, while values closer to 0 indicate stronger similarity to SB conditions. This probability-based approach allows assessment of degree of environmental similarity rather than forcing a strict binary classification.

Model performance was first evaluated on the held-out test data using a default probability threshold of 0.5 to establish baseline discrimination between SA and SB waters. To support screening use, model performance was then evaluated across a range of probability thresholds (a “threshold sweep”) to quantify tradeoffs between (1) recall (sensitivity) for SA: how many true SA observations are flagged as SA; (2) precision for SA: how often SA predictions are correct; and (3) flagged rate: how aggressively the model label observations SA. As an initial benchmark, the threshold that maximized the F1 score (balanced recall and precision) was selected using historical statewide data only and fixed prior to application to Chandler Bay.

To identify which environmental factors most strongly distinguish class SA from class SB waters, permutation-based variable importance was used. This approach evaluates how prediction accuracy changes when individual variables are randomly shuffled, with larger reductions in accuracy indicating greater importance. This process provided an objective ranking of which parameters most strongly contribute to classification differences. Since multiple water quality parameters were evaluated together, the model captured how combinations of parameters jointly characterize regulatory classes, reflecting the multivariate nature of marine classification.

Finally, to ensure that results were not driven primarily by differences in monitoring practices across agencies, a sensitivity analysis was conducted in which all indicator variables reflecting

whether a parameter was measured were removed. This chemistry-only model tested whether classification patterns persisted when predictions were based solely on measured environmental values. Comparison of results across model specifications allowed separation of environmental signals from potential effects of monitoring structure.

Intended Use and Limitations

This modeling framework is intended as a screening and decision-support tool to help identify locations where observed water quality conditions may warrant closer examination. Model results do not replace regulatory criteria or formal classification procedures and should be interpreted alongside site-specific context, monitoring history, and regulatory standards. Because marine classification reflects factors beyond measured water quality alone, including management considerations and historical designations, perfect replication of classifications using environmental variables alone is neither expected nor necessary for the intended application.

Chandler Bay Water Quality Data

Chandler Bay water quality data were collected from monitoring efforts conducted under the Kingfish Company discharge permit requirements (permit #ME0037559) [1]. These monitoring data were collected independently of this analysis and used as an external application dataset, not for training, variable selection, or threshold tuning. The Chandler Bay dataset includes sampling location and date information and the available water quality variables that correspond to those used in the statewide model. However, since the Kingfish company was not required to measure any microbial indicators, the Chandler Bay dataset was missing all enterococcus measurements.

Summary of Results

A total of 26,810 observations were initially assigned a marine classification based on spatial overlay with Maine DEP statutory GIS layers. Of these, 81.8% were classified as SB, 13.5% as SA, and 4.7% as SC. After removing observations that contained only temperature and salinity measurements (i.e., lacked any additional water quality parameter), 2,480 observations remained. Within this subset, SA comprised 48.1%, SB comprised 47.7%, and SC comprised 4.2%. Since the primary research question concerns the environmental gradient between SA and SB waters, the remaining 105 SC observations were excluded. The final SA-SB modeling dataset consisted of 2,375 observations (n = 1,193 SA; 1,182 SB). This near-balanced distribution reduced potential bias in model training and evaluation.

Model Results

Using the statewide reference dataset, the random forest model was able to tell apart water quality patterns typical of historically designated SA waters versus historically designated SB waters. In the held-out test set (i.e., data not used to train the model), the model correctly classified the majority of observations at the standard 0.5 threshold. When the model predicted that an observation was SA, it was correct about 85% of the time, meaning the model's SA predictions were usually reliable. However, since the purpose of this work is to support regulatory screening, and not just to maximize overall accuracy, we then examined performance across many probability cutoffs to understand the tradeoff between (1) identifying more waters that may be SA-like and (2) maintaining confidence when the model flags a water as SA-like. The threshold that provided the best overall balance was 0.38, which maximized a combined performance measure that rewards both "finding SA-like waters" and "being correct when calling SA-like." At this cutoff, the model identified about 82.8% of true SA observations while remaining correct about 77.6% of the time when it labeled an observation as SA-like (overall balance score = 0.80). Together, these results show the model learned meaningful, repeatable differences in environmental conditions, along with differences in monitoring patterns, that have historically aligned with SA versus SB classifications in the statewide data. Importantly, the 0.38 cutoff was selected using only historical statewide data and was fixed before applying the model to the independent Chandler Bay Kingfish dataset, so the Chandler Bay evaluation reflects a fair comparison rather than a threshold tuned to the case study.

To understand what the model relied on most when distinguishing historically designated SA waters from SB waters, we calculated "relative importance" for each input by measuring how much model predictions worsened when that input was randomly shuffled. For ease of comparison, we scaled these values so the most influential input equals 100, using the formula 100 × importance(variable) / max(importance) (Figure 2). The results show that water-quality measurements themselves did matter. However, the strongest drivers of the SA vs SB distinction were not the measured values alone, but whether certain parameters were measured at all (for example, whether dissolved oxygen or enterococcus were included in a given record). In the importance rankings, the "measured/not measured" indicators (especially DO measured and enterococcus measured) were more influential than any single water-quality value (Figure 2), meaning the model partly learned differences in monitoring patterns between the datasets historically associated with SA and SB. Because these measurement-pattern variables can reflect how monitoring is conducted, not just environmental conditions, we treated this as an important caution for interpretation and ran a separate "chemistry-only" sensitivity check to confirm that Chandler Bay's conclusions were not dependent on monitoring-pattern signals.

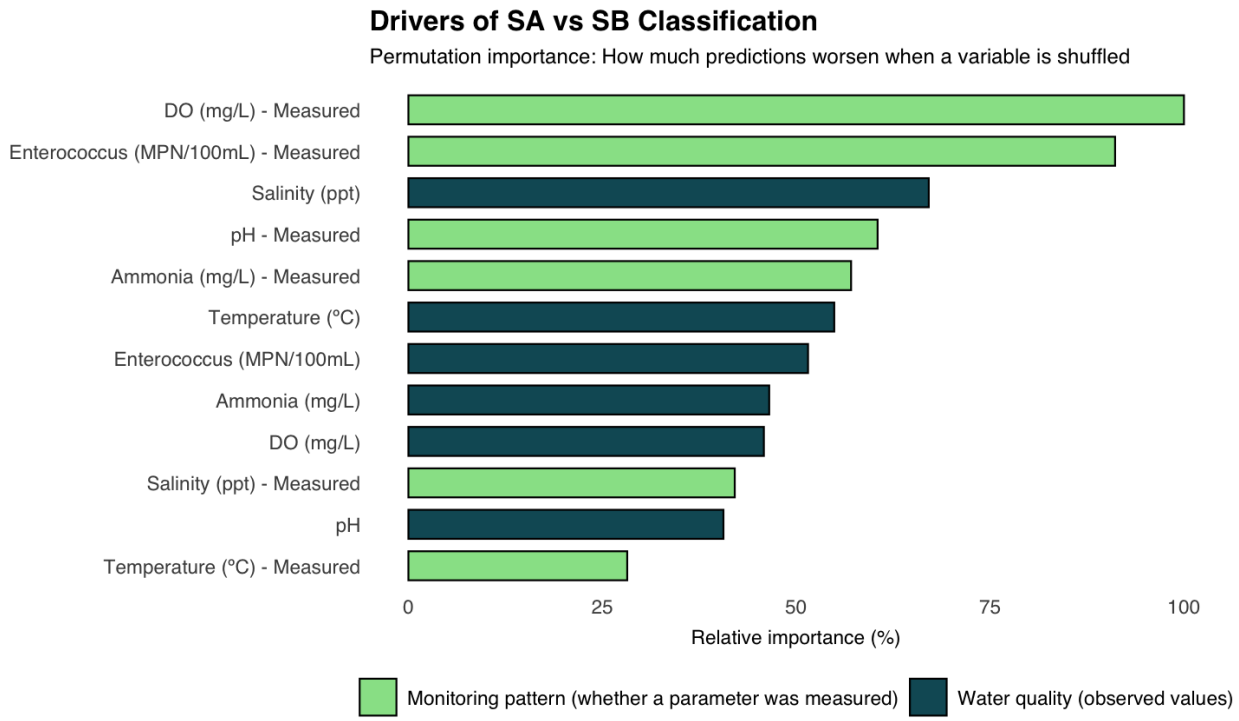


Figure 2. Relative importance of model inputs used to distinguish historically designated SA versus SB waters. Bays show permutation importance, defined as how much the model's predictions worsened when each input was randomly shuffled. Importance values are scaled so the most influential input equals 100% (relative importance = 100 × importance(variable) / max(importance)). Dark green bars represent measured water-quality values, while light green bars represent monitoring-pattern indicators (whether a parameter was measured or not).

To ensure that model results were not driven primarily by differences in monitoring practices across agencies, an additional random forest model was fit using only measured environmental values and excluding all indicator variables that reflected whether a parameter was recorded. This sensitivity analysis tested whether classification performance depended on patterns of data availability rather than underlying environmental conditions. Model performance under this restricted specification remained strong and yielded a similar separation between SA and SB waters. Variable importance rankings continued to emphasize dissolved oxygen, microbial indicators, salinity, pH, and nitrogen-related parameters.

Applying Models to Chandler Bay Data

When the trained model was applied to the Chandler Bay monitoring dataset, Chandler Bay observations consistently aligned more closely with the statewide pattern typical of historically designated SA waters than SB waters (Figure 3, Figure 4). Using the pre-selected cutoff of 0.38, 155 out of 156 Chandler Bay observations were classified as SA-like. The predicted SA-likeness scores for Chandler Bay were generally high (median 0.83; mean 0.80). In contrast, the model scores for statewide historical waters showed lower predicted SA-likeness values for SB and

intermediate-to-high values for SA (historical SA test-set median 0.71; historical SB test-set median 0.28), with Chandler Bay's median score exceeding both and sitting well above the typical SA level. Consistent with this finding, about 80% of Chandler Bay observations had scores higher than the median score of historical SA observations, indicating that most Chandler Bay samples were not only SA-like, but often more SA-like than the "typical" SA observation in the statewide comparison set.

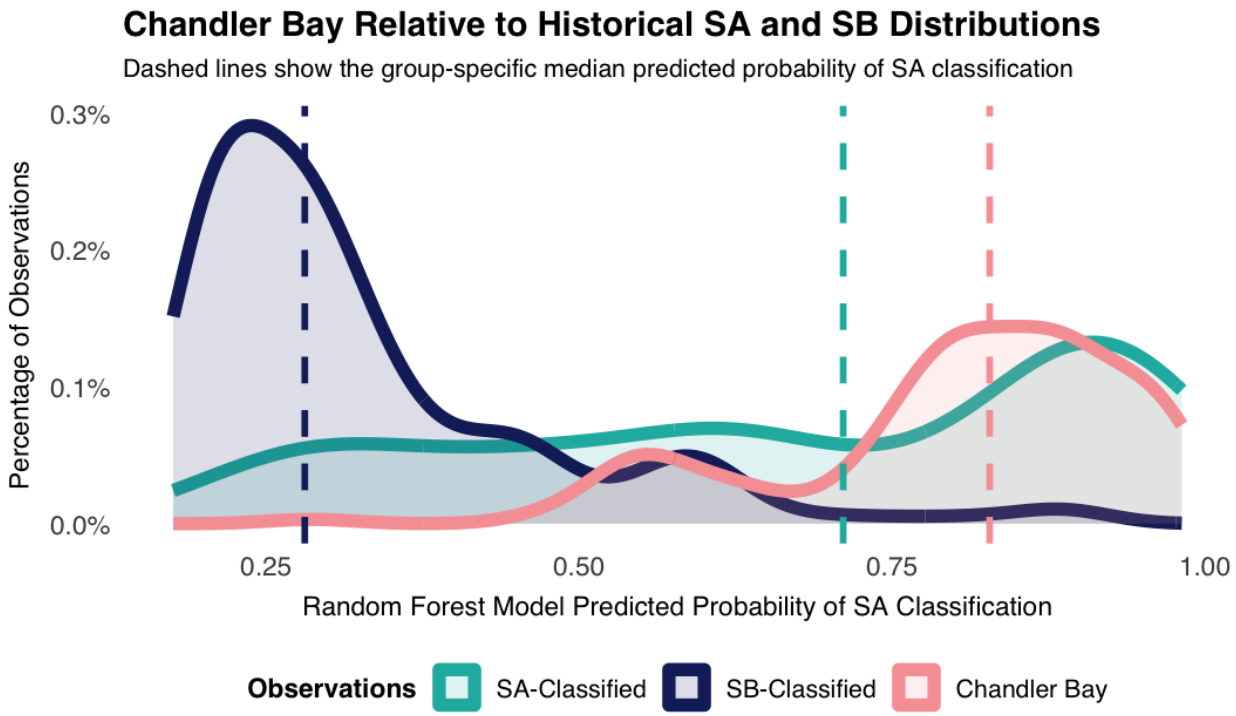


Figure 3. Frequency of model scores for waters designated SB, SA, and the observed data from Chandler Bay. Although there was a large range in model scores for waters classified SB, the median model score was 0.28. For waters designated SA, the median model score was 0.71, indicating that the model demonstrates skill in determining between SA and SB waters based on the provided parameters. The median model score for the Chandler Bay data is 0.83, which is above the SA average.

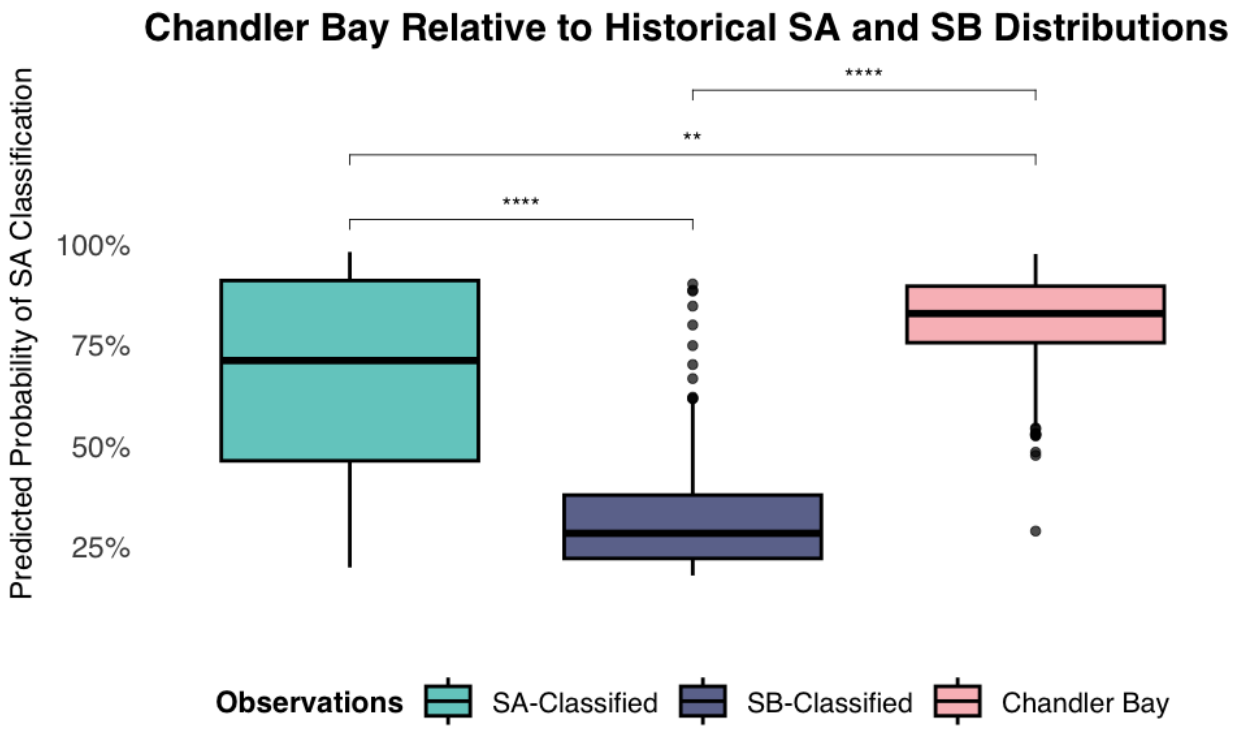


Figure 4. Distribution of the random forest's predicted probability that an observation is SA-like for three groups: historically SA-classified waters, historically SB-classified waters, and Chandler Bay. Higher values indicate conditions more similar to historically designated SA waters. The center of each box represents the group-specific median predicted probability of SA classification. Brackets indicate pairwise statistical comparisons of the three groups using a Dunn's test with Bonferroni adjustment for multiple comparisons. Asterisks represent the adjusted significance level.

Because statewide datasets can differ in how often parameters are measured, we also ran a sensitivity check that removed the "measured/not measured" indicator variables and relied only on the actual water-quality values (i.e., chemistry-only model; Figure 5). Under this chemistry-only approach, Chandler Bay still aligned more closely with historical SA conditions (median 0.80; mean 0.79). However, removing the monitoring indicators reduced the strength of the original signal: the proportion of Chandler Bay observations exceeding the historical SA median declined from roughly 80% to roughly 60%. This change does not mean Chandler Bay appears SB-like, and it does not mean the model result was artificial. Instead, it indicates that the original 80% finding reflected two components—measured environmental conditions and differences in monitoring structure—and that monitoring patterns modestly amplified the separation between SA and SB learned from the statewide data. Even after removing monitoring structure, the majority of Chandler Bay observations still fell above the typical historical SA level, indicating that Chandler Bay's SA-like result is primarily driven by measured environmental conditions, with monitoring patterns providing a secondary reinforcement rather than creating the conclusion.

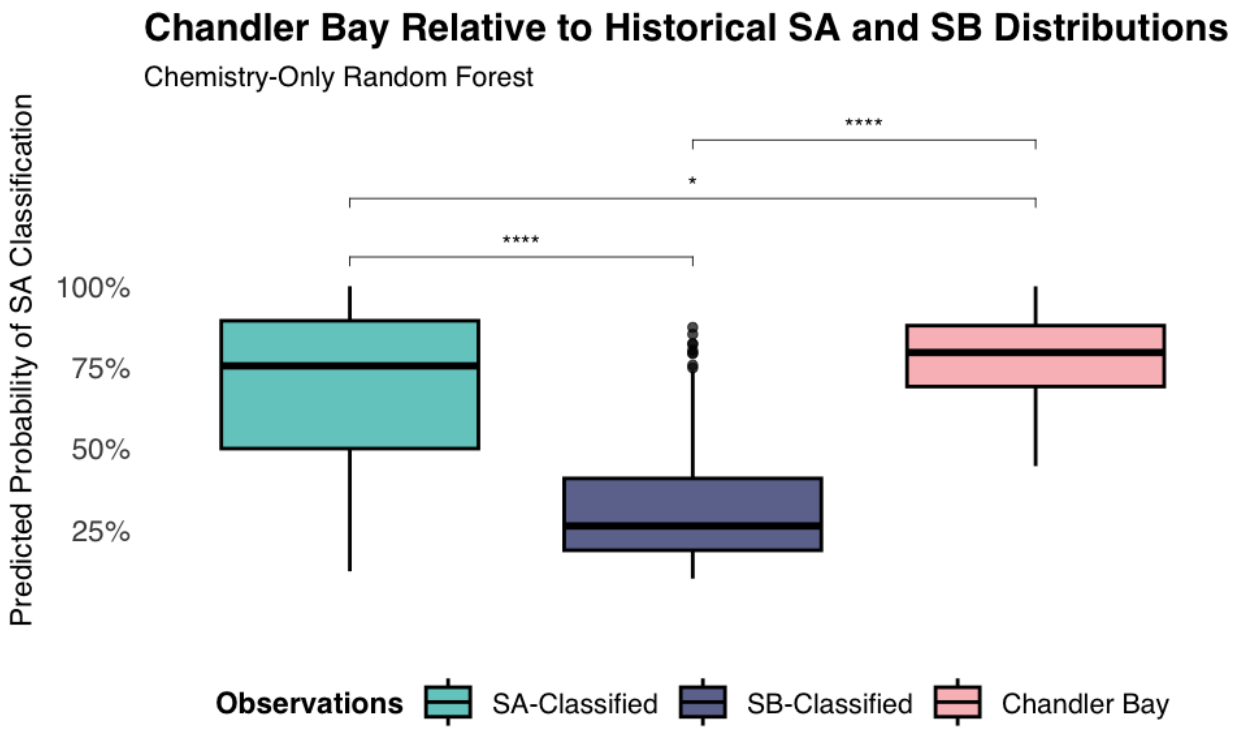


Figure 5. Boxplots show the **chemistry-only random forest model's predicted probability** that an observation is SA-like when the model is trained and applied using only measured water-quality values (i.e., excluding variables that indicate whether a parameter was measured). Higher values indicate conditions more similar to historically designated SA waters. The center of each box represents the group-specific median predicted probability of SA classification. Brackets indicate pairwise statistical comparisons of the three groups using a Dunn's test with Bonferroni adjustment for multiple comparisons. Asterisks represent the adjusted significance level.

Management Implications

The modeling framework presented in this report is designed to support, not replace, Maine's statutory marine classification process. Marine classifications are legal designations set by statute (38 M.R.S. §§ 464, 465-B), and they reflect a combination of environmental conditions, historical decisions, socioeconomic considerations, management priorities, and policy objectives. Because of this, it is neither expected nor appropriate for a statistical model based on water quality alone to exactly reproduce official classifications. What water quality data *can* do, however, is help answer a practical and policy-relevant question, "do conditions we are measuring in the water actually match what the given classification is meant to represent?" By being able to place individual waterbodies within a statewide environmental context, this analysis provides an objective way to evaluate whether observed conditions are more consistent with higher-quality (class SA) waters or with class SB waters.

Importantly, this analysis does *not* change or assign classifications. The random forest model does not “upgrade” or “downgrade” any waterbody and does not make regulatory decisions. Instead, it evaluates how similar a set of measured environmental conditions is to those typically found in SA waters versus SB waters. When model results differ from official classifications, that difference reflects patterns in the data, not a change in legal status. From a management perspective, this distinction matters. Maine’s antidegradation policy (38 M.R.S. § 464(4)(F)) requires the State to identify and protect waters whose actual quality exceeds the minimum standards of their assigned classification. That responsibility depends on being able to recognize when higher-quality conditions are present in the first place. This study shows that, when water quality data are evaluated in a statewide context, it is possible to identify consistent and repeatable environmental patterns that distinguish higher-quality waters from those under greater stress.

The analysis also identifies which water quality measurements are most useful for distinguishing higher-quality waters from those under greater stress. Dissolved oxygen and microbial indicators, parameters already central to regulatory decision-making, were among the strongest predictors in the model. In addition, pH and ammonia consistently emerged as important indicators of overall environmental condition. Although ammonia is not commonly highlighted in marine classification discussions, its influence in the model likely reflects its role as a signal of nutrient inputs, organic enrichment, and biological stress. Considered together, pH and ammonia capture broader water quality conditions that help differentiate higher-quality systems from those experiencing greater anthropogenic influence. While these variables are not currently emphasized to the same extent as dissolved oxygen or bacterial indicators in marine classification determinations (38 M.R.S. § 456-B), their strong and consistent statistical signal suggests that they should be considered important supporting variables when evaluating whether observed conditions align with assigned classifications.

A key finding of this study is that monitoring practices themselves influence how clearly higher-quality waters can be identified. The initial statewide dataset included more than 26,000 observations, but many of those records (90.7%) contained only temperature and salinity measurements. Because temperature and salinity alone do not determine marine classification, those observations could not meaningfully inform the SA–SB distinction and were excluded from the final analysis. As a result, fewer than 2,500 observations ultimately contained at least one additional classification-relevant water quality parameter. This highlights a practical limitation of current monitoring programs: when key parameters are not measured consistently, the ability to detect higher-quality conditions is reduced.

The chemistry-only analysis further showed that differences in monitoring intensity can amplify, but do not create, signals of higher water quality. Even when all monitoring-structure information was removed, Chandler Bay still aligned more closely with historically SA waters based on

measured chemistry alone. This finding underscores that environmental signals are real, while also emphasizing that inconsistent monitoring makes those signals harder to detect.

From a management standpoint, these results point to a clear opportunity for improvement. More consistent documentation and standardized labeling of water quality measurements across agencies would greatly expand the usefulness of existing data. Improved consistency would facilitate the inclusion of additional water quality variables (such as chlorophyll, phosphorus, light availability, and total nitrogen) into future versions of the model. These variables are known to be ecologically important but are currently difficult to integrate due to inconsistencies in how agencies label, record, and document their water quality measurements. Standardizing how data are collected, labeled, and archived would also allow additional datasets and monitoring programs to be incorporated into the model. Expanding the range of comparable data sources would increase sample size, improve model accuracy, and strengthen the State's ability to detect and protect higher-quality waters. In this way, better data consistency directly translates into better management tools.

Overall, the modeling framework presented here is best viewed as a screening and prioritization tool. It can help identify locations where observed conditions appear inconsistent with assigned classifications and where additional monitoring, review, or regulatory attention may be warranted. Used alongside professional judgment and statutory procedures, this approach can strengthen the factual foundation of marine classification review and improve implementation of Maine's antidegradation policy as coastal uses and environmental pressures continue to evolve.

Data Limitations

The datasets used in this study did not include complete or consistent measurements across all coastal Maine waters. Monitoring coverage varied substantially across agencies, locations, and time periods. Salinity (missing from 2.7% of observations) and temperature (missing from 3.4%) were measured consistently across nearly all records, whereas dissolved oxygen was missing from about 96.5% of observations, enterococcus was missing from 94.2%, pH was missing from 97.8%, and ammonia was missing from 98.5%. In fact, no observation ever recorded all five of our core variables. After removing observations that only recorded salinity and/or temperature and all SC classified waters, approximately 61.8%, 37.4%, 75.7%, and 84.1% of observations were missing DO, enterococcus, pH, and ammonia readings, respectively. This uneven data availability is particularly important because dissolved oxygen and enterococcus are currently the top predictors associated with regulatory decisions.

Rather than filling missing values with artificial estimates, this analysis retained data gaps transparently and explicitly accounted for whether variables were measured. This approach avoids introducing assumptions into the dataset and ensures that model outputs reflect real-world monitoring constraints. However, it also means that uneven sampling directly influences

model performance. As monitoring programs expand and more uniform measurements become available, our random forest model is expected to become increasingly effective tools for identifying trends, flagging potential inconsistencies, and evaluating longer-term dynamics relevant to marine classification review.

Future Considerations

Maine's statutory framework already provides mechanisms to maintain and protect high-quality waters. The effectiveness of that framework depends in part on whether marine classifications accurately reflect current environmental conditions. As coastal uses expand (e.g., expansion of the aquaculture industry in Maine (e.g, [8]) and monitoring technologies improve, integrating modern analytical tools into classification review can help ensure that Maine's coastal waters receive the full level of protection provided under Maine's statutory framework.

We also want to emphasize that this report does not oppose aquaculture development or other coastal economic activity. Rather, it emphasizes that accurate, data-informed classification supports regulatory clarity, environmental stewardship, and long-term sustainability. By aligning marine classifications with observed environmental conditions, Maine can strengthen implementation of its existing statutory protections and preserve the ecological and economic value of its coastal waters for future generations.

References

[1] The University of Maine, “2025 Ambient Water Quality Monitoring Report for Kingfish Maine, Land Based Aquaculture Project Jonesport, Washington County, Maine, USA,” Jonesport, Maine, Nov. 2025.

[2] Posit team, *RStudio: Integrated Development Environment for R*. (2024). Posit Software, PBC, Boston, MA. [Online]. Available: http://www.posit.co/

[3] National Water Quality Monitoring Council, “Maine Biological Water Quality Metadata.” Water Quality Portal. doi: https://doi.org/10.5066/P9QRKUVJ.

[4] Maine Department of Environmental Protection, “Environmental and Geographic Analysis Database (EGAD) Maine SA Waters Water Quality.” Accessed: Feb. 04, 2026. [Online]. Available: https://www.maine.gov/dep/maps-data/egad

[5] B. Schaffner, “Current Maine Statutory Water Classification,” Maine Department of Environmental Protection, State of Maine, Apr. 28, 2025. Accessed: Feb. 01, 2026. [Online]. Available: https://maine.hub.arcgis.com/maps/maine::current-maine-statutory-water-classification-1/explore?location=44.963167%2C-68.548530%2C4&path=

[6] Hannah Frick, Fanny Chow, Max Kuhn, Michael Mahoney, Julia Silge, and Hadley Wickham, *rsample: General Resampling Infrastructure*. (2026). [Online]. Available: https://CRAN.R-project.org/package=rsample

[7] Marvin N. Wright and Andreas Ziegler, “ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R.,” *Journal of Statistical Software*, vol. 77, no. 1, pp. 1–17, 2017, doi: 10.18637/jss.v077.i01.

[8] Frenchman Bay United, “DEP Triennial Review of Maine’s Water Quality Standards- 2024 - 2026 Written Testimony, Henry Sharpe, President, Frenchman Bay United,” 2024. [Online]. Available: https://www.maine.gov/dep/water/wqs/ProposalDocs/2024-2026/2024-TR_FBU_FinfishAquaculture.pdf

Appendix 1

R Code to Create Figures and Random Forest Model

Including Additional Analyses and Figures not Included in Final Report

NOTE - in this file I am going to remove all SC observations since my main research question is “Does Chandler Bay environmentally resemble historically SA-designated waters more than SB-designate waters?”

By removing SC observations this helps the decision boundary between SA and SB to become more subtle in the binary models, forcing the model to learn the environmental gradient between adjacent regulatory classes.

Model	Question it Answers	When Appropriate
SA vs (SB + SC)	Is this water SA-like vs everything else?	General SA screening
SA vs SB only	Is this water more like SA or SB?	Reclassification argument
Multiclass SA/SB/SC	Can chemistry reproduce regulation?	Structural/diagnostic analysis

How to explain in report what/why we did this: Because Chandler Bay is currently designated SB, and the central research question concerns whether its environmental conditions more closely resemble SA or SB waters, we restricted the binary classification analysis to SA and SB observations only. SC waters were excluded from this model because they represent a distinct regulatory category not directly relevant to the SA–SB decision boundary. This approach focuses the analysis on the environmental gradient between adjacent classifications.

Figure 1 - Map Creation

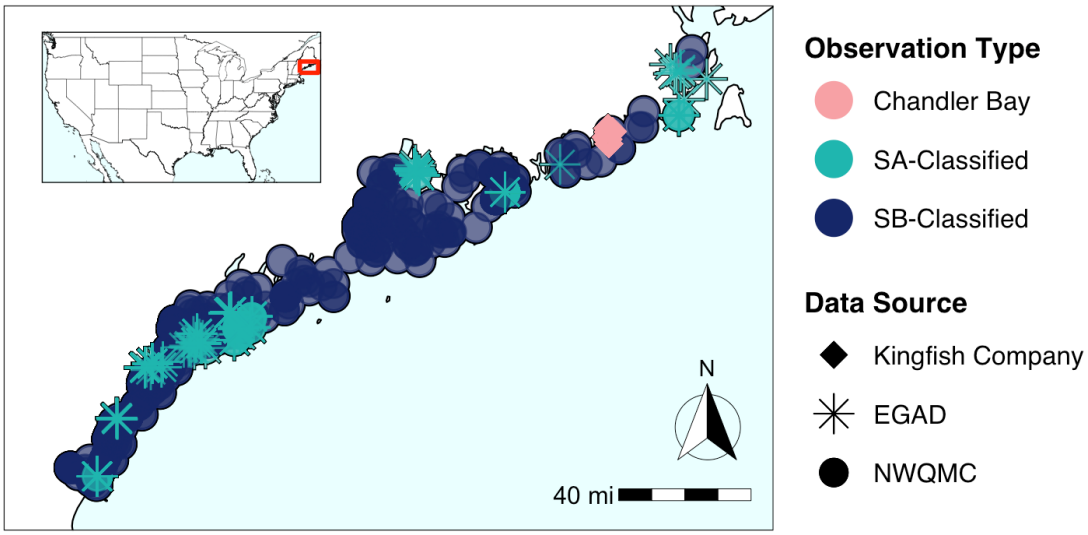
```
rm(list=ls())

NWQMC = fread("Data Files/Cleaned Data/Manually Edited/NWQMC water quality download_m
anually edited_created 2.18.26.csv")

EGAD = fread("Data Files/Cleaned Data/Manually Edited/EGAD Manually Edited Data creat
ed 2.9.26.csv")

combined <- bind_rows(source_NWQMC = NWQMC,
                      source_EGAD = EGAD,
                      .id = "source")

## New names:
## • `Total Phosphorus_mixed forms_mgL` -> `Total Phosphorus_mixed forms_mgL...36`
## • `Nitrate + Nitrite_mgL` -> `Nitrate + Nitrite_mgL...43`
```

```
## • `Nitrate + Nitrite_mgL` -> `Nitrate + Nitrite_mgL...54`
## • `Total Phosphorus_mixed forms_mgL` -> `Total Phosphorus_mixed forms_mgL...55`

df <- combined |>
  mutate(
    Marine_Class = factor(`Marine Class`, levels = c("SA", "SB", "SC")),
    Date = as.Date(Date, format = "%m/%d/%y"))

core_vars <- c("Temperature_degC",
               "Salinity")

h20_vars <- c("DO_mgL",
              "pH")

nutrient_vars <- c("Ammonia_mgL")

bio_vars <- c("Enterococcus_MPNper100mL")

all_vars <- c(core_vars, h20_vars, nutrient_vars, bio_vars)

non_core_vars <- setdiff(all_vars, core_vars)

historical_df <- df |>
  dplyr::filter(
    dplyr::if_any(
      dplyr::any_of(non_core_vars),
      \(x) !is.na(x))) |>
  dplyr::filter(`Marine Class` %in% c("SA", "SB")) |>
  dplyr::mutate(
    class_sasb = factor(`Marine Class`, levels = c("SA", "SB"))
  ) |>
  dplyr::select(-`Marine Class`) |>
  droplevels() |>
  st_as_sf(coords = c("Longitude", "Latitude"), crs = 4326, remove = FALSE)

chandler_df <- fread("Data Files/Cleaned Data/Manually Edited/Kingfish Only Chandler
Bay WQData Manually edited 2.13.26.csv") |>
  st_as_sf(coords = c("Longitude", "Latitude"), crs = 4326, remove = FALSE) |>
  mutate(Date = as.Date(Date, format = "%m/%d/%y"))

historical_plot <- st_transform(historical_df, 4326)
chandler_plot <- st_transform(chandler_df, 4326)


# Get land polygons (countries; fine for Maine coast context)
land <- ne_download(scale = 10, type = "land", category = "physical", returnclass = "
sf") |>
  st_transform(4326)

## Reading layer `ne_10m_land' from data source
##   `/private/var/folders/yl/72hfylyj5g138hjk8btzbzfh0000gn/T/RtmpFs82oT/ne_10m_land
.shp'
##   using driver `ESRI Shapefile'
## Simple feature collection with 11 features and 3 fields
## Geometry type: MULTIPOLYGON
```

```
## Dimension:     XY
## Bounding box:  xmin: -180 ymin: -90 xmax: 180 ymax: 83.6341
## Geodetic CRS:  WGS 84

# Optional: zoom to your data extent with a little padding
bb <- st_bbox(historical_plot)
x_pad <- (bb$xmax - bb$xmin) * 0.05
y_pad <- (bb$ymax - bb$ymin) * 0.05

historical_plot$source = as.factor(historical_plot$source)

chandler_plot$Marine_Class <- "Chandler"
chandler_plot$source <- "Kingfish"

combined_plot <- dplyr::bind_rows(historical_plot, chandler_plot)

## New names:
## • `Nitrate + Nitrite_mgL` -> `Nitrate + Nitrite_mgL...65`
## • `Total Phosphorus_mixed forms_mgL` -> `Total Phosphorus_mixed forms_mgL...66`

box <- data.frame(long = c(bb$xmin,bb$xmin, bb$xmax, bb$xmax, bb$xmin),
                  lat = c(bb$ymin, bb$ymax, bb$ymax, bb$ymin, bb$ymin))
states <- st_as_sf(map("state", plot = FALSE, fill = TRUE)) #package 'sf'
world = ne_countries(scale = "large", returnclass = "sf") #ne_countries{rnaturalearth}

g2 <- ggplotGrob(
   ggplot(data = world) +
     geom_sf(fill= "white") +
     geom_sf(data = states, fill = NA) +
     theme_void() +
     theme(panel.grid.major = element_line(color = gray(.5),
                                           linetype = 0,
                                           linewidth = 0.5),
           panel.background = element_rect(fill = "#EAFFFFFF")) +
     coord_sf(xlim = c(-125.2154, -66.3087),
              ylim = c(24.6016,49.9417), expand = FALSE) +
     geom_path(data = box, aes(x = long, y = lat),
               linewidth = 2, color = "red") +
     theme(
       plot.title = element_blank(),
       axis.title.x = element_blank(),
       axis.title.y = element_blank(),
       legend.title = element_blank() ,
       axis.text = element_blank(),
       axis.ticks = element_blank()) +
     theme(panel.border = element_rect(colour = "black", fill=NA, linewidth=1)))

ggplot() +
  annotate("rect", xmin = -Inf, xmax = Inf, ymin = -Inf, ymax = Inf,
           fill = "#EAFFFFFF", color = NA) + # water background
  geom_sf(data = land, fill = "white", color = "black", linewidth = 0.8) + # land layer
  ggspatial::annotation_scale(location = "br",
                              width_hint = 0.2,
```

```
                              unit_category = "imperial",
                              text_cex = 2,
                              pad_x = unit(0.3, "in"), pad_y = unit(0.4, "in"),
                              height = unit(0.4, "cm")) +
  ggspatial::annotation_north_arrow(location = "br", which_north = "true",
                            height = unit(3.7, "cm"),
                            width = unit(4, "cm"),
                            pad_x = unit(0.1, "in"), pad_y = unit(0.8, "in"),
                            style = ggspatial::north_arrow_fancy_orienteering(text_size
= 22)) +
  geom_sf(data = combined_plot, #[combined_plot$source == "source_EGAD",],
        aes(shape = source),
        color = "black",
        stroke = 2.5,
        size = 12,
        alpha = 1) +
  geom_sf(data = combined_plot, #[combined_plot$source == "source_EGAD",],
        aes(shape = source),
        color = "white",
        stroke = 1,
        size = 12,
        alpha = 1) +
  geom_sf(data = combined_plot, #[combined_plot$source == "source_NWQMC",],
        aes(color = Marine_Class, shape = source),
        #color = "black",
        stroke = 1.5,
        size = 12,
        alpha = 0.7) +
  coord_sf(xlim = c(bb$xmin - x_pad, bb$xmax + x_pad),
           ylim = c(bb$ymin - y_pad, bb$ymax + y_pad),
           datum = NA) +
  scale_shape_manual(values =c(source_EGAD = 8,
                               source_NWQMC = 19,
                               Kingfish = 18),
                     labels =c(source_EGAD ="EGAD",
                               source_NWQMC ="NWQMC",
                               Kingfish ="Kingfish Company"),
                     name = "Data Source") +
  scale_color_manual(values =c(SA = "#1BB6AFFF",
                               SB = "#172869FF",
                               Chandler = "#F6A1A5FF"),
                     labels =c(SA ="SA-Classified",
                               SB ="SB-Classified",
                               Chandler ="Chandler Bay"),
                     name = "Observation Type") +
  hrbrthemes::theme_ipsum(base_family = "Helvetica", base_size = 20, grid = "N",
                            axis_title_size = 22, axis_title_just = "cc", axis_title_fa
ce = "plain",
                            plot_margin = margin(0,0,0,0)) +
  guides(color = guide_legend(order = 1,
                              override.aes = list(shape = 19, #color = "black",
                                                  stroke = 4,  alpha = 1)),
         shape = guide_legend(order = 2,
                              override.aes = list(fill = "white", color = "black",
                                                  stroke = 1,  alpha = 1))) +
```

```
  annotation_custom(grob = g2,
                    xmin = -71 , xmax = -69.3,
                    ymin = 44.3, ymax = 45.1) +
  theme(panel.grid = element_blank(),
        legend.position = "right",
        legend.justification.inside = c(0, 1),
        legend.key.size = unit(2, "cm"),
        legend.text = element_text(size = 25, color = "black"),
        legend.title = element_text(face = "bold",size = 27),
        panel.border = element_rect(color = "black", fill = NA, linewidth = 0.8),
        plot.margin = unit(c(.5,.5,.5,.5), "cm"))
```



.

Random Forests

```
rm(list=ls())


NWQMC = fread("Data Files/Cleaned Data/Manually Edited/NWQMC water quality download_m
anually edited_created 2.18.26.csv")


EGAD = fread("Data Files/Cleaned Data/Manually Edited/EGAD Manually Edited Data creat
ed 2.9.26.csv")

combined <- bind_rows(source_NWQMC = NWQMC,
                      source_EGAD = EGAD,
                      .id = "source")
```

```
## New names:
## • `Total Phosphorus_mixed forms_mgL` -> `Total Phosphorus_mixed forms_mgL...36`
## • `Nitrate + Nitrite_mgL` -> `Nitrate + Nitrite_mgL...43`
## • `Nitrate + Nitrite_mgL` -> `Nitrate + Nitrite_mgL...54`
## • `Total Phosphorus_mixed forms_mgL` -> `Total Phosphorus_mixed forms_mgL...55`
```

Standardize column names and types

```
df <- combined |>
  mutate(
    Marine_Class = factor(`Marine Class`, levels = c("SA", "SB", "SC")),
    Date = as.Date(Date, format = "%m/%d/%y")
  )
```

Data Engineering: Build a Model-Ready Table

Define variable groups:

```
core_vars <- c("Temperature_degC",
               "Salinity")

h20_vars <- c("DO_mgL",
              "pH")

nutrient_vars <- c("Ammonia_mgL") # removed "Nitrate + Nitrite_mgL" since missing a l
ot of data

bio_vars <- c("Enterococcus_MPNper100mL")

all_vars <- c(core_vars, h20_vars, nutrient_vars, bio_vars)
```

NOTE - here, I did remove chlorophyll a, total nitrogen, and total phosphorus from the list of predictor variables since they were rarely ever recorded and honestly had a hard time determining how each data set measured these variables. They generally would have multiple columns for these three variables labeled as something slightly different - I tried to make sure I was only combining columns that were measuring the same thing but not sure if that was all that correct for these variables.

Only keep observations where there is at least one non-core parameter present (i.e., remove all observations that only had salinity/temperature data available - make sure there is at least one other parameter available for that observation):

```
non_core_vars <- setdiff(all_vars, core_vars)

df_clean2 <- df |>
  dplyr::filter(
    dplyr::if_any(
      dplyr::any_of(non_core_vars),
      \(x) !is.na(x)
    )
  )
```

Create missingness flags:

```
df_ml <- df_clean2 |>
  mutate(Marine_Class = factor(`Marine Class`, levels = c("SA", "SB", "SC"))) |>
  select(Marine_Class, all_of(all_vars)) |>
  mutate(
    across(
      all_of(all_vars),
      list(measured = ~ !is.na(.x)),
      .names = "{.col}_measured"
    )
  )
```

Now every variable has (1) parameter and (2) parameter_measured (T/F). This allows the model to learn (using DO as an example): 1. When DO is low AND measured –> higher SC risk; 2. When DO is NEVER measured –> different regime).

```
df %>%
  count(Marine_Class) %>%
  mutate(percent = n / sum(n) * 100)

##    Marine_Class     n   percent
##          <fctr> <int>     <num>
## 1:           SA  3638 13.569564
## 2:           SB 21922 81.767997
## 3:           SC  1250  4.662439

df_ml %>%
  count(Marine_Class) %>%
  mutate(percent = n / sum(n) * 100)

##    Marine_Class     n   percent
##          <fctr> <int>     <num>
## 1:           SA  1193 48.104839
## 2:           SB  1182 47.661290
## 3:           SC   105  4.233871

df_ml_clean <- df_ml |>
  janitor::clean_names()
```

- This standardizes column names to snake_case and removes spaces/special characters in column names.
- This makes formulas and column references more reliable.

Filter to SA and SB observations only

Remove SC observations

```
df_sasb <- df_ml_clean |>
  dplyr::filter(marine_class %in% c("SA", "SB")) |>
  dplyr::mutate(
    class_sasb = factor(marine_class, levels = c("SA", "SB"))
  ) |>
```

```
  dplyr::select(-marine_class) |>
  droplevels()
```

WHY: Chandler Bay is currently SB and we want evidence that it looks more like SA than SB. SC is not part of that boundary.

```
df_sasb %>%
  count(class_sasb) %>%
  mutate(percent = n / sum(n) * 100)

##    class_sasb     n  percent
##        <fctr> <int>    <num>
## 1:         SA  1193 50.23158
## 2:         SB  1182 49.76842

df |>
  summarise(across(
    c(Temperature_degC, Salinity, DO_mgL, pH, Ammonia_mgL,  Enterococcus_MPNper100mL)
,
    ~ mean(is.na(.x))
  )) |>
  tidyr::pivot_longer(everything(), names_to = "var", values_to = "prop_na") |>
  arrange(desc(prop_na))

## # A tibble: 6 × 2
##   var                       prop_na
##   <chr>                       <dbl>
## 1 Ammonia_mgL               0.985
## 2 pH                        0.978
## 3 DO_mgL                    0.965
## 4 Enterococcus_MPNper100mL  0.942
## 5 Salinity                  0.0337
## 6 Temperature_degC          0.0271

df_ml |>
  summarise(across(
    c(Temperature_degC, Salinity, DO_mgL, pH, Ammonia_mgL,  Enterococcus_MPNper100mL)
,
    ~ mean(is.na(.x))
  )) |>
  tidyr::pivot_longer(everything(), names_to = "var", values_to = "prop_na") |>
  arrange(desc(prop_na))

## # A tibble: 6 × 2
##   var                       prop_na
##   <chr>                       <dbl>
## 1 Ammonia_mgL               0.841
## 2 pH                        0.757
## 3 DO_mgL                    0.618
## 4 Enterococcus_MPNper100mL  0.374
## 5 Salinity                  0.0605
## 6 Temperature_degC          0.0379
```

```
df_sasb |>
  summarise(across(
    c(temperature_deg_c, salinity, do_mg_l, p_h, ammonia_mg_l,  enterococcus_mp_nper1
00m_l),
    ~ mean(is.na(.x))
  )) |>
  tidyr::pivot_longer(everything(), names_to = "var", values_to = "prop_na") |>
  arrange(desc(prop_na))
```

```
## # A tibble: 6 × 2
##   var                        prop_na
##   <chr>                        <dbl>
## 1 ammonia_mg_l                0.838
## 2 p_h                         0.772
## 3 do_mg_l                     0.628
## 4 enterococcus_mp_nper100m_l  0.363
## 5 salinity                    0.0627
## 6 temperature_deg_c           0.016
```

Random Forest - Binary

Stratified Split

You need an honest performance estimate on unseen data. This avoids “training-set optimism”. Instead of randomly splitting the data, we make sure to perform a stratified split. Here we used initial_split() to make an 80/20 split. Then ‘strata - class_sasb’ forces the split to preserve the SA/SB proportions. Then training() and testing() extract the two data frames we just created

```
set.seed(123)
split_obj <- rsample::initial_split(df_sasb, prop = 0.8, strata = class_sasb)
train_sasb <- rsample::training(split_obj)
test_sasb  <- rsample::testing(split_obj)
```

Train Random Forest (probabilities and importance)

```
set.seed(123)

rf_sasb <- ranger::ranger(
  class_sasb ~ .,
  data = train_sasb,
  probability = TRUE, #forces it to return P(SA) rather than just a hard label
  importance = "permutation", #computes variable importance robustly
  num.trees = 1000,
  min.node.size = 10
)
```

- P(SA) probability allows you to do screening and thresholding
- Permutation importance gives interpretable “which variables matter most for SA vs SB?”

Predict probabilities on the test set and evaluate basic performance

```
prob_test <- predict(rf_sasb, data = test_sasb)$predictions
p_sa_test <- prob_test[, "SA"]

pred_test <- factor(ifelse(p_sa_test >= 0.5, "SA", "SB"), levels = c("SA", "SB"))

eval_test <- tibble(
  truth = test_sasb$class_sasb,
  estimate = pred_test
)

conf_mat(eval_test, truth = truth, estimate = estimate)

##           Truth
## Prediction  SA  SB
##         SA 170  29
##         SB  69 208

sens(eval_test, truth = truth, estimate = estimate, event_level = "first")

## # A tibble: 1 × 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 sens    binary         0.711

precision(eval_test, truth = truth, estimate = estimate, event_level = "first")

## # A tibble: 1 × 3
##   .metric   .estimator .estimate
##   <chr>     <chr>          <dbl>
## 1 precision binary         0.854
```

- First 2 lines of code - What this does: gets predicted probabilities and extracts the SA column
    - WHY: we want a continuous “SA-likeness” score
- WHY: baseline evaluation at 0.5 before threshold tuning

Threshold sweep

- Goal: choose a probability cutoff so that:
    - you catch more SA waters (high recall), but
    - you don’t label everything as SA

Create an evalution table:

```
eval_df <- tibble(
  truth = test_sasb$class_sasb,
  prob  = p_sa_test
)
```

This creates truth labels + predicted probabilities

Sweep thresholds and record recall/precision/flagged-rate

```
threshold_grid <- seq(0.01, 0.98, by = 0.01)

threshold_results <- lapply(threshold_grid, function(t) {

  pred <- factor(ifelse(eval_df$prob >= t, "SA", "SB"),
                 levels = c("SA", "SB"))

  tibble(
    threshold = t,
    recall_sa = yardstick::sens_vec(eval_df$truth, pred, event_level = "first"),
    precision_sa = yardstick::precision_vec(eval_df$truth, pred, event_level = "first
"),
    flagged_rate = mean(pred == "SA")
  )
}) |>
  bind_rows()
```

- Line-by-line meaning:
  - t is a candidate cutoff.
  - If P(SA) ≥ t, label as SA; otherwise SB.
  - recall_sa = % of true SA correctly flagged SA.
  - precision_sa = among predicted SA, how many are truly SA.
  - flagged_rate = what % of all samples are being labeled SA.
- Why we do this:
  - There is no “one correct” threshold.
  - Screening requires choosing an operating point explicitly.

Choose a threshold

- Recall (sensitivity for SA) answers the question “of all the true SA waters, how many did we correctly flag as SA?”
  - If recall = 0.9 –> you found 90% of the real SA waters
  - If recall = 0.3 –> you missed most SA waters
  - When recall matters: when you don’t want to miss SA-like waters
- Precision answers the question “of all waters we flagged as SA, how many actually were SA?”
  - If precision = 0.9 –> most flagged waters truly are SA
  - If precision = 0.2 –> most flagged waters are actually SB
- Flagged rate answers the question “what percentage of all waters are we calling SA?”
  - If flagged_rate = 1 –> everything is SA
  - If flagged_rate = 0.1 –> only 10% flagged SA
  - It controls how aggressive you are
- What the Threshold Actually does:
  - When you choose a threshold (e.g., 0.42), you are saying: any observation with P(SA) ≥ 0.42 will be labeled SA
  - It simply draws a horizontal line on the probability scale

- The model creates a scale of SA-likeness from 0 to 1. The thresold just defines where you draw the line between:
  - Looks more like SB
  - Looks more like SA

Choosing the threshold that maximizes F1 score (balanced precision + recall):

```
threshold_results <- threshold_results |>
  mutate(
    f1 = 2 * (recall_sa * precision_sa) / (recall_sa + precision_sa)
  )

best_f1 <- threshold_results |>
  arrange(desc(f1)) |>
  slice(1)

best_f1
```

```
## # A tibble: 1 × 5
##   threshold recall_sa precision_sa flagged_rate    f1
##       <dbl>     <dbl>        <dbl>        <dbl> <dbl>
## 1      0.38     0.828        0.776        0.536 0.802
```

```
t_star <- best_f1$threshold
```

In methods say something along the lines of: The classification threshold was selected using performance on withheld historical SA/SB observations and was fixed prior to application to the Chandler Bay data set.

Variable Importance (SA vs SB)

```
imp <- sort(rf_sasb$variable.importance, decreasing = TRUE)
head(imp, 20)
```

```
##                    do_mg_l_measured enterococcus_mp_nper100m_l_measured
##                          0.08902748                           0.08112385
##                            salinity                       p_h_measured
##                          0.05973280                           0.05386181
##                ammonia_mg_l_measured                  temperature_deg_c
##                          0.05081747                           0.04888480
##           enterococcus_mp_nper100m_l                       ammonia_mg_l
##                          0.04587557                           0.04141847
##                             do_mg_l                 salinity_measured
##                          0.04080582                           0.03746184
##                                 p_h         temperature_deg_c_measured
##                          0.03616440                           0.02511932
```

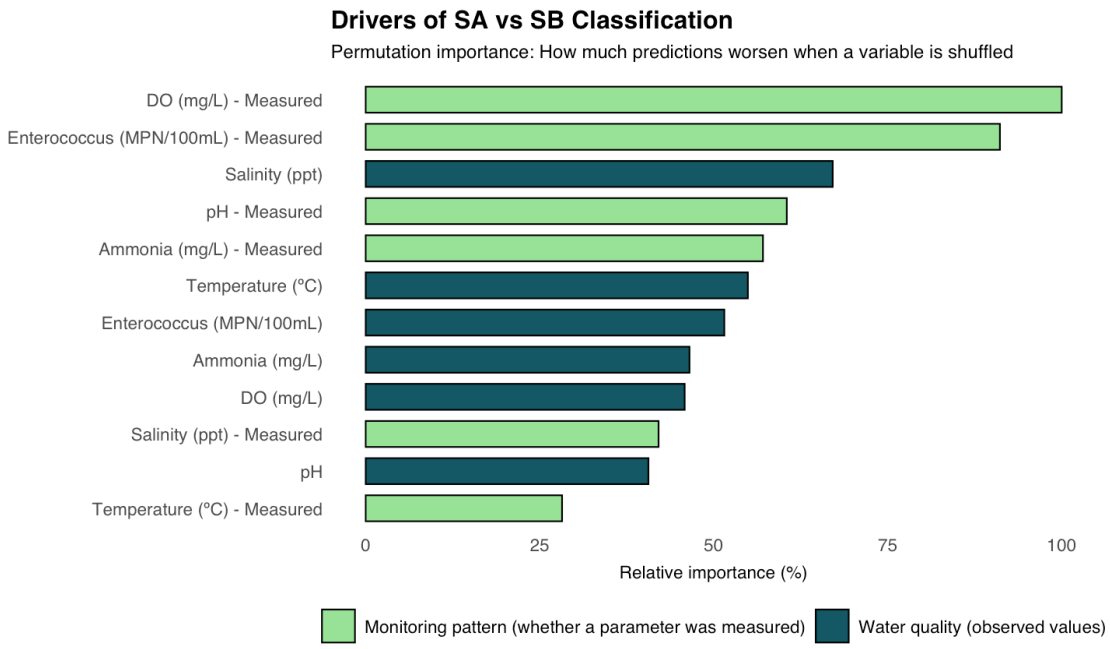
This shows you permutation importance from ranger(). For each variable, the model asks “If I randomly scramble this variable, how much worse do my predictions get?” Values are *relative*, not absolute. Thus, this ranking is telling you “which variables the forest actually relies on to predict Marine Class.

- Bigger number = more important
- 0 ≈ irrelevant

Visualize Variable Importance

```
imp_df <- tibble(
  variable = names(rf_sasb$variable.importance),
  importance = rf_sasb$variable.importance
) |>
  mutate(
    rel_importance = 100 * importance / max(importance),
    type = ifelse(grepl("_measured$", variable),
                  "Measurement pattern",
                  "Water quality")
  ) |>
  arrange(desc(rel_importance)) |>
  slice_head(n = 20)

ggplot(imp_df,
       aes(x = reorder(variable, rel_importance),
           y = rel_importance,
           fill = type)) +
  geom_col(width = 0.7, color = "black") +
  #stat_summary(geom = "bar", fun = mean, color = "black") +
  #stat_summary(geom = "errorbar", fun.data = mean_se, width = 0.2) +
  coord_flip() +
  scale_fill_manual(values = c("Measurement pattern" = "#97E196FF",
                               "Water quality" = "#105965FF"),
                    labels = c("Measurement pattern" = "Monitoring pattern (whether a
parameter was measured)",
                               "Water quality" = "Water quality (observed values)"),
                    name = NULL) +
  labs(x = NULL,
       y = "Relative importance (%)",
       title = "Drivers of SA vs SB Classification",
       subtitle = "Permutation importance: How much predictions worsen when a variabl
e is shuffled") +
  hrbrthemes::theme_ipsum(base_family = "Helvetica", base_size = 17, grid = "N",
                          axis_title_size = 17, axis_title_just = "cc", axis_title_fa
ce = "plain",
                          plot_title_family = "Helvetica", plot_title_size = 24,
                          subtitle_size = 18, subtitle_family = "Helvetica",
                          plot_margin = margin(15,15,15,15)) +
  theme(panel.grid = element_blank(),
        legend.position = "bottom",
        legend.key.size = unit(1.2, "cm"),
        legend.text = element_text(size = 17, color = "black"),
        legend.title = element_blank(),
        axis.title.x = element_text(margin = margin(t = 10, r = 0, b = 0, l = 0)),
        axis.title.y = element_text(margin = margin(t = 0, r = 10, b = 0, l = 0))) +
  scale_x_discrete(labels=c("salinity" = expression('Salinity (ppt)'),
                            "temperature_deg_c" = expression('Temperature (ºC)'),
                            "enterococcus_mp_nper100m_l" = expression('Enterococcus (
MPN/100mL)'),
                            "ammonia_mg_l" = expression('Ammonia (mg/L)'),
                            "do_mg_l" = expression('DO (mg/L)'),
                            "p_h" = expression('pH'),
                            "salinity_measured" = expression('Salinity (ppt) - Measur
```

```
ed'),
                              "temperature_deg_c_measured" = expression('Temperature (º
C) - Measured'),
                              "enterococcus_mp_nper100m_l_measured" = expression('Enter
ococcus (MPN/100mL) - Measured'),
                              "ammonia_mg_l_measured" = expression('Ammonia (mg/L) - Me
asured'),
                              "do_mg_l_measured" = expression('DO (mg/L) - Measured'),
                              "p_h_measured" = expression('pH - Measured')))
```



Option 2 visualization

```
imp_tbl <- tibble(
  variable = names(rf_sasb$variable.importance),
  importance = unname(rf_sasb$variable.importance)
) |>
  mutate(
    rel = 100 * importance / max(importance),
    group = if_else(str_detect(variable, "_measured$"),
                    "Monitoring pattern (whether a parameter was measured)",
                    "Water quality (observed values)")
  )


plot_imp_panel <- function(df, title) {
  df |>
    mutate(label = variable) |>
    ggplot(aes(x = reorder(label, rel), y = rel)) +
    geom_col(width = 0.7, fill = "#2C7FB8") +
    coord_flip() +
    labs(x = NULL, y = "Relative importance (top = 100)",
         title = title,
```

```
         subtitle = "Permutation importance: how much predictions worsen when a varia
ble is shuffled")
}

top_wq <- imp_tbl |>
  filter(group == "Water quality (observed values)") |>
  slice_max(rel, n = 7)
top_wq
```

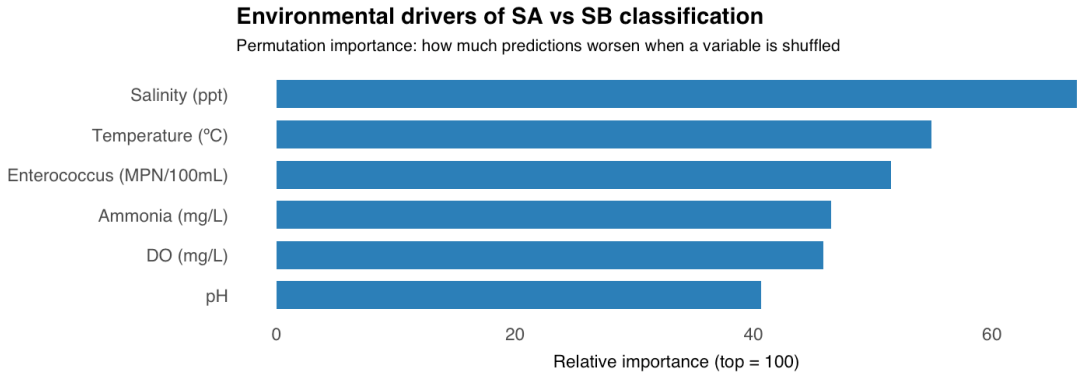
```
## # A tibble: 6 × 4
##   variable                  importance   rel group
##   <chr>                          <dbl> <dbl> <chr>
## 1 salinity                      0.0597  67.1 Water quality (observed values)
## 2 temperature_deg_c             0.0489  54.9 Water quality (observed values)
## 3 enterococcus_mp_nper100m_l    0.0459  51.5 Water quality (observed values)
## 4 ammonia_mg_l                  0.0414  46.5 Water quality (observed values)
## 5 do_mg_l                       0.0408  45.8 Water quality (observed values)
## 6 p_h                           0.0362  40.6 Water quality (observed values)
```

```
top_meas <- imp_tbl |>
  filter(group == "Monitoring pattern (whether a parameter was measured)") |>
  slice_max(rel, n = 7)
top_meas
```

```
## # A tibble: 6 × 4
##   variable                            importance   rel group
##   <chr>                                    <dbl> <dbl> <chr>
## 1 do_mg_l_measured                        0.0890 100   Monitoring pattern (whet…
## 2 enterococcus_mp_nper100m_l_measured     0.0811  91.1 Monitoring pattern (whet…
## 3 p_h_measured                            0.0539  60.5 Monitoring pattern (whet…
## 4 ammonia_mg_l_measured                   0.0508  57.1 Monitoring pattern (whet…
## 5 salinity_measured                       0.0375  42.1 Monitoring pattern (whet…
## 6 temperature_deg_c_measured              0.0251  28.2 Monitoring pattern (whet…
```

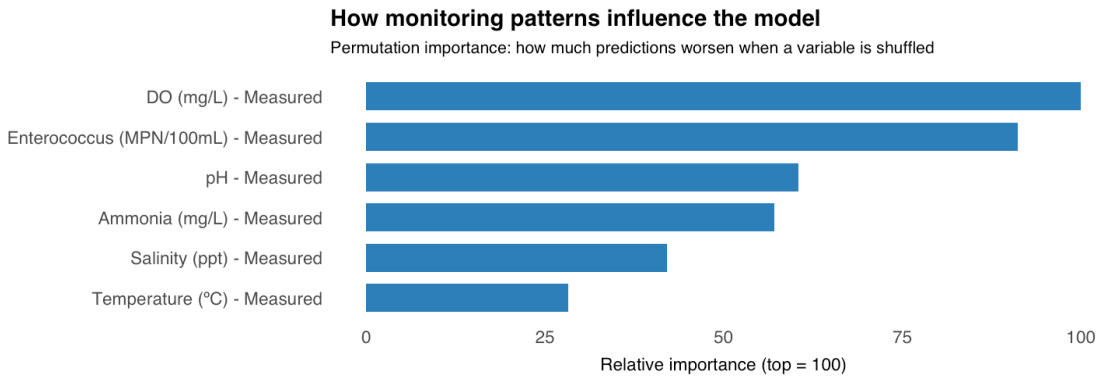
```
p1 <- plot_imp_panel(top_wq, "Environmental drivers of SA vs SB classification")
p2 <- plot_imp_panel(top_meas, "How monitoring patterns influence the model")

p1 +
  hrbrthemes::theme_ipsum(base_family = "Helvetica", base_size = 17, grid = "N",
                          axis_title_size = 17, axis_title_just = "cc", axis_title_fa
ce = "plain",
                          plot_title_family = "Helvetica", plot_title_size = 22,
                          subtitle_size = 16, subtitle_family = "Helvetica",
                          plot_margin = margin(5,5,5,5)) +
  theme(panel.grid = element_blank(),
        legend.position = "bottom",
        legend.key.size = unit(1.2, "cm"),
        legend.text = element_text(size = 16, color = "black"),
        legend.title = element_text(face = "bold",size = 17),
        axis.title.x = element_text(margin = margin(t = 10, r = 0, b = 0, l = 0)),
        axis.title.y = element_text(margin = margin(t = 0, r = 10, b = 0, l = 0))) +
  scale_x_discrete(labels=c("salinity" = expression('Salinity (ppt)'),
                            "temperature_deg_c" = expression('Temperature (ºC)'),
                            "enterococcus_mp_nper100m_l" = expression('Enterococcus (
MPN/100mL)'),
```

```
                              "ammonia_mg_l" = expression('Ammonia (mg/L)'),
                              "do_mg_l" = expression('DO (mg/L)'),
                              "p_h" = expression('pH')))
```



The environmental conditions that the model relied on

```
p2 +
  hrbrthemes::theme_ipsum(base_family = "Helvetica", base_size = 17, grid = "N",
                          axis_title_size = 17, axis_title_just = "cc", axis_title_fa
ce = "plain",
                          plot_title_family = "Helvetica", plot_title_size = 22,
                          subtitle_size = 16, subtitle_family = "Helvetica",
                          plot_margin = margin(5,5,5,5)) +
  theme(panel.grid = element_blank(),
        legend.position = "bottom",
        legend.key.size = unit(1.2, "cm"),
        legend.text = element_text(size = 16, color = "black"),
        legend.title = element_text(face = "bold",size = 17),
        axis.title.x = element_text(margin = margin(t = 10, r = 0, b = 0, l = 0)),
        axis.title.y = element_text(margin = margin(t = 0, r = 10, b = 0, l = 0))) +
  scale_x_discrete(labels=c("salinity_measured" = expression('Salinity (ppt) - Measur
ed'),
                            "temperature_deg_c_measured" = expression('Temperature (º
C) - Measured'),
                            "enterococcus_mp_nper100m_l_measured" = expression('Enter
ococcus (MPN/100mL) - Measured'),
                            "ammonia_mg_l_measured" = expression('Ammonia (mg/L) - Me
asured'),
                            "do_mg_l_measured" = expression('DO (mg/L) - Measured'),
                            "p_h_measured" = expression('pH - Measured')))
```

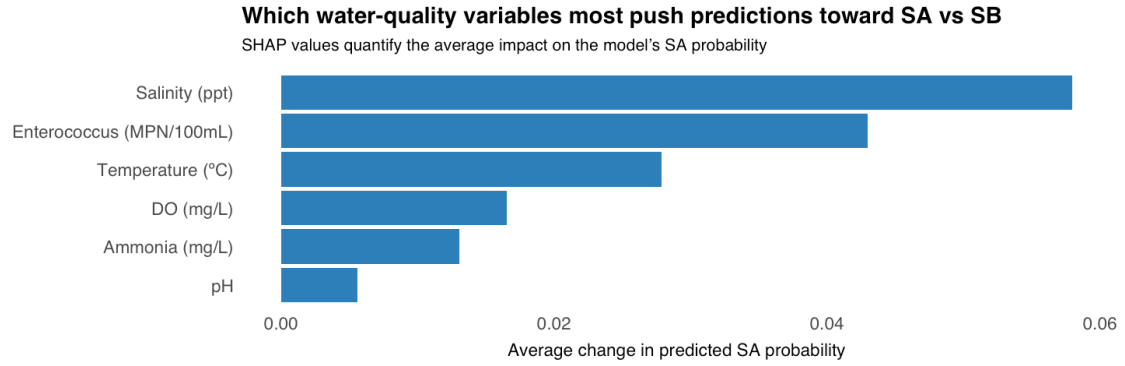


Separately, these are the monitoring patterns that also help prediction, which we treat as a caution flag about differences in sampling regimes

Option 3 - SHAP

```
x_train <- train_sasb |> select(-class_sasb)

# prediction wrapper: return P(SA)
pred_sa <- function(object, newdata) {
  predict(object, data = newdata)$predictions[, "SA"]
}

set.seed(123)
sh <- fastshap::explain(
  object = rf_sasb,
  X = x_train,
  pred_wrapper = pred_sa,
  nsim = 100
)

sv <- shapviz::shapviz(sh, X = x_train)



# Policy version: bar of mean absolute SHAP, restricted to non-_measured
mean_abs <- tibble(
  variable = colnames(sh),
  mean_abs_shap = colMeans(abs(sh))
) |>
  filter(!str_detect(variable, "_measured$")) |>
  slice_max(mean_abs_shap, n = 6) |>
  mutate(label = variable )

ggplot(mean_abs, aes(x = reorder(label, mean_abs_shap), y = mean_abs_shap)) +
  geom_col(fill = "#2C7FB8") +
  coord_flip() +
  labs(
    x = NULL,
    y = "Average change in predicted SA probability",
    title = "Which water-quality variables most push predictions toward SA vs SB",
    subtitle = "SHAP values quantify the average impact on the model's SA probability
```

```
") +
  hrbrthemes::theme_ipsum(base_family = "Helvetica", base_size = 17, grid = "N",
                          axis_title_size = 17, axis_title_just = "cc", axis_title_fa
ce = "plain",
                          plot_title_family = "Helvetica", plot_title_size = 22,
                          subtitle_size = 16, subtitle_family = "Helvetica",
                          plot_margin = margin(10,10,10,10)) +
  theme(panel.grid = element_blank(),
        legend.position = "bottom",
        legend.key.size = unit(1.2, "cm"),
        legend.text = element_text(size = 16, color = "black"),
        legend.title = element_text(face = "bold",size = 17),
        axis.title.x = element_text(margin = margin(t = 10, r = 0, b = 0, l = 0)),
        axis.title.y = element_text(margin = margin(t = 0, r = 10, b = 0, l = 0))) +
  scale_x_discrete(labels=c("salinity" = expression('Salinity (ppt)'),
                            "temperature_deg_c" = expression('Temperature (ºC)'),
                            "enterococcus_mp_nper100m_l" = expression('Enterococcus (
MPN/100mL)'),
                            "ammonia_mg_l" = expression('Ammonia (mg/L)'),
                            "do_mg_l" = expression('DO (mg/L)'),
                            "p_h" = expression('pH')))
```



Here, "Importance" becomes "influence on SA-likeness," which matches the story and the P(SA) framing.

- Each bar is: Average absolute SHAP value for that variable
- That means:
    - It measures how much that variable changes the model's predicted probability of SA
    - It averages over all observations
    - It ignores direction (only magnitude)
- So this is "average influence strength", not "positive vs negative effect."
- The x-axis: "Average influence on P(SA)"
    - means: If Salinity has ~0.06, then on average, salinity shifts predicted P(SA) by about 6 percentage points.
- You can say: "Salinity is the single most influential variable in determining whether waters look more SA-like or SB-like in our model. On average, it shifts the probability of SA classification more than any other variable."
    - Then: Enterococcus is the second strongest driver.
        - Temperature is third.
        - DO and Ammonia have moderate influence.

        - pH has relatively little influence in this classification decision.
- Important clarification:
    - It does not mean higher salinity increases SA.
    - It does not mean enterococcus makes waters worse or better.
    - It does not show direction.
- It only shows: “When this variable changes, the model’s prediction changes a lot.”
- Your plot says:
    - The model is mostly using salinity and bacteria to separate SA from SB.
    - Temperature matters.
    - Oxygen and ammonia matter somewhat.
    - pH barely influences classification.

Applying RF Model to Chandler Bay (external data)

```
# Example loading step — adjust filename to yours
chandler_clean <- fread("Data Files/Cleaned Data/Manually Edited/Kingfish Only Chandl
er Bay WQData Manually edited 2.13.26.csv") |>
  clean_names()
```

Create Chandler Bay features to match training predictors

```
make_rf_features_from_chandler <- function(ch) {
  out <- ch |>
    transmute(
      temperature_deg_c = temperature_deg_c,
      salinity = salinity,
      do_mg_l = do_mg_l,
      p_h = p_h,
      ammonia_mg_l = ammonia_mg_l,
     # nitrate_nitrite_mg_l = nitrate_nitrite_mg_l,
      enterococcus_mp_nper100m_l = NA_real_  # if Chandler doesn't have it
    )

  pred_cols <- names(out)

  out |>
    mutate(
      across(
        all_of(pred_cols),
        list(measured = ~ !is.na(.x)),
        .names = "{.col}_measured"
      )
    )
}
```

WHY: Chandler Bay data doesn’t have the same exact columns as the training data. This creates the same predictor schema the RF model expects.

Align Chandler columns to the RF training columns

RF expects the same set of columns as the training data

```
align_to_train <- function(new_df, train_cols) {

  # Add missing columns as NA
  missing_cols <- setdiff(train_cols, names(new_df))
  for (nm in missing_cols) new_df[[nm]] <- NA

  # Drop extra columns
  extra_cols <- setdiff(names(new_df), train_cols)
  if (length(extra_cols) > 0) {
    new_df <- new_df[, setdiff(names(new_df), extra_cols), drop = FALSE]
  }

  # IMPORTANT: data.table needs ..train_cols to select by a variable
  if (data.table::is.data.table(new_df)) {
    new_df <- new_df[, ..train_cols]
  } else {
    new_df <- new_df[, train_cols, drop = FALSE]
  }

  new_df
}
```

Predict Chandler P(SA) with the RF Model

```
train_cols <- setdiff(names(train_sasb), "class_sasb")

ch_feat <- make_rf_features_from_chandler(chandler_clean)
ch_feat_aligned <- align_to_train(ch_feat, train_cols)

p_sa_ch <- predict(rf_sasb, data = ch_feat_aligned)$predictions[, "SA"]
```

Apply the chosen threshold to Chandler Bay

```
chandler_scored <- chandler_clean |>
  mutate(
    p_sa_rf = p_sa_ch,
    sa_like = ifelse(p_sa_rf >= t_star, "SA-like", "SB-like")
  )

table(chandler_scored$sa_like)

##
## SA-like SB-like
##     155       1

summary(chandler_scored$p_sa_rf)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2862  0.7555  0.8280  0.8018  0.8962  0.9764
```

reprint our threshold:

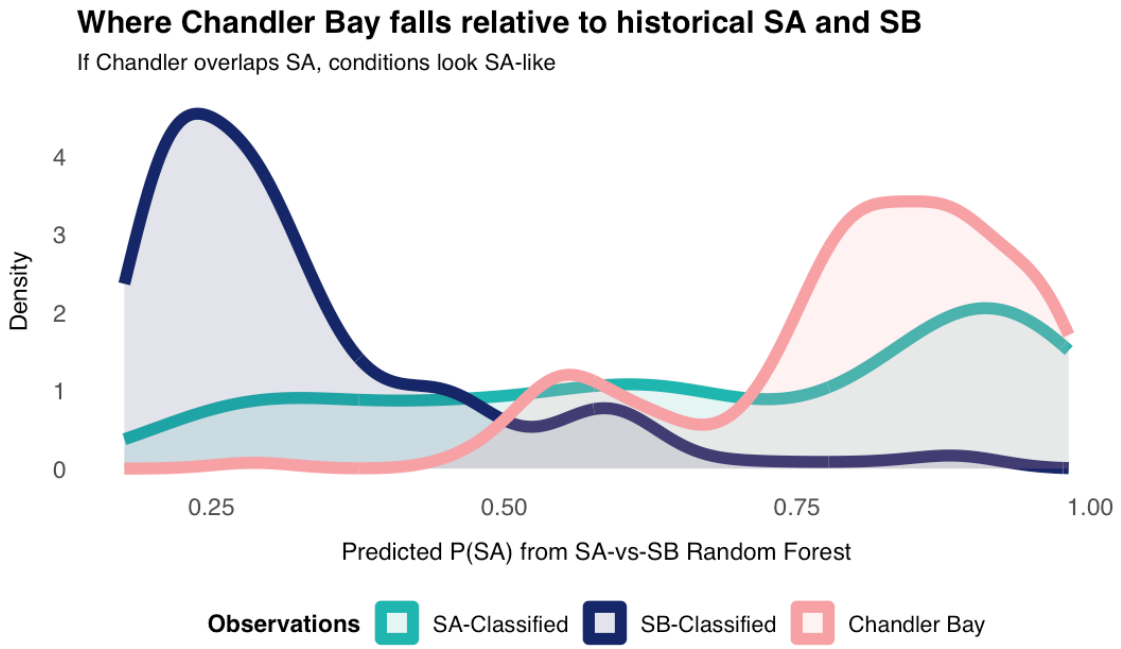
```
best_f1

## # A tibble: 1 × 5
##   threshold recall_sa precision_sa flagged_rate    f1
##       <dbl>     <dbl>        <dbl>        <dbl> <dbl>
## 1      0.38     0.828        0.776        0.536 0.802
```

Visually Compare Chandler Bay data to historical SA and SB

- Instead of hard-thresholding Chandler immediately:
  - Plot distribution of P(SA) for historical SA
  - Plot distribution for historical SB
  - Overlay Chandler distribution
- If Chandler overlaps heavily with historical SA distribution, that's powerful.
- You can do it by generating P(SA) for three groups and then plotting their distributions on the same x-axis:
  - historical SA (from your held-out test set)
  - historical SB (from your held-out test set)
  - Chandler (external)

```
hist_df <- tibble(
  group = test_sasb$class_sasb,  # SA or SB
  #do_measured = test_sasb$do_mg_l_measured,
  p_sa = p_sa_test)

ch_df <- tibble(
  group = factor("Chandler", levels = c("SA", "SB", "Chandler")),
  p_sa = p_sa_ch)

plot_df <- bind_rows(
  hist_df |> mutate(group = factor(as.character(group), levels = c("SA","SB","Chandle
r"))),
  ch_df)

ggplot(plot_df, aes(x = p_sa, fill = group, color = group)) +
  geom_density(alpha = 0.15, linewidth = 4) +
  scale_fill_manual(values = c(SA = "#1BB6AFFF",
                               SB = "#172869FF",
                               Chandler = "#F6A1A5FF"),
                    labels = c(SA = "SA-Classified",
                               SB = "SB-Classified",
                               Chandler = "Chandler Bay"),
                    name = "Observations") +
  scale_color_manual(values = c(SA = "#1BB6AFFF",
                               SB = "#172869FF",
                               Chandler = "#F6A1A5FF"),
```

```
                    labels = c(SA = "SA-Classified",
                               SB = "SB-Classified",
                               Chandler = "Chandler Bay"),
                    name = "Observations") +
  labs(x = "Predicted P(SA) from SA-vs-SB Random Forest",
       y = "Density",
       title = "Where Chandler Bay falls relative to historical SA and SB",
       subtitle = "If Chandler overlaps SA, conditions look SA-like") +
  hrbrthemes::theme_ipsum(base_family = "Helvetica", base_size = 17, grid = "N",
                          axis_title_size = 17, axis_title_just = "cc", axis_title_fa
ce = "plain",
                          plot_title_family = "Helvetica", plot_title_size = 22,
                          subtitle_size = 16, subtitle_family = "Helvetica",
                          plot_margin = margin(5,5,5,5)) +
  theme(panel.grid = element_blank(),
        legend.position = "bottom",
        legend.key.size = unit(1.2, "cm"),
        legend.text = element_text(size = 16, color = "black"),
        legend.title = element_text(face = "bold",size = 17),
        axis.title.x = element_text(margin = unit(c(t = 1, r = 0, b = 0, l = 0), "lin
e")),
        axis.title.y = element_text(margin = unit(c(t = 0, r = 1, b = 0, l = 0), "lin
e")))
```



Interpreting the Figure

- x-axis:
    - 0 = very SB-like
    - 1 = very SA-like
- y-axis = density (how common each probability is)

- SB - the model assigns low P(SA) to historical SB waters
    - mostly clustered around 0.2 - 0.4
    - very few values above 0.6
- SA - the model assigns higher P(SA) to historical SA waters
    - spread across 0.3 - 1.0
    - much heavier mass at high probabilities (0.8 - 1.0)
- **NOTE** - there is overlap between SA and SB in the middle (around 0.4 - 0.6)
    - that overlap is NOT a modeling error, it is ecological reality

TAKEAWAY - According to the SA-vs-SB model, Chandler observations fall in the range typically associated with historical SA waters — not SB waters. The Chandler Bay curve overlaps heavily with SA and barely at all with SB.
i.e., The plot is saying "when trained to distinguish SA from SB statewide, the random forest places Chandler Bay firmly in the SA-like region of environmental space."

What does this mean scientifically?

- the model has learned
    - what historical SA conditions look like
    - what historical SB conditions look like
- Research question: where does Chandler Bay fall on the learned SA-SB gradient?
    - Answer: Chandler Bay falls in the SA region of probability space

Why is this figure stronger than just choosing a threshold?

- If we had just reported 90% of Chandler Bay was classified SA at threshold 0.5, someone could argue that we schose that threshold
- This plot shows something that is threshold-independent: Chandler's entire probability distribution lies in the SA region
- Even if we raise the threshold to 0.8, most of Chandler Bay samples would still be classified as SA
    - Even at 0.9, many would still be classified as SA

How to write this up in the results section: The distribution of predicted SA probabilities for Chandler Bay was strongly shifted toward values characteristic of historical SA waters and well separated from the distribution for historical SB waters. Most Chandler observations received predicted P(SA) values above 0.8, overlapping substantially with the historical SA distribution and showing minimal overlap with historical SB. This indicates that, based solely on measured environmental conditions, Chandler Bay more closely resembles historically SA-designated waters than SB-designated waters.

Adding a simple numeric comparision to accompany the density figure

```
summary_table <- bind_rows(
  tibble(group = as.character(test_sasb$class_sasb), p_sa = p_sa_test),
  tibble(group = "Chandler", p_sa = p_sa_ch)
) |>
  mutate(group = factor(group, levels = c("SA","SB","Chandler"))) |>
```

```r
  group_by(group) |>
  summarise(
    n = n(),
    median = median(p_sa, na.rm = TRUE),
    mean = mean(p_sa, na.rm = TRUE),
    p10 = quantile(p_sa, 0.10, na.rm = TRUE),
    p90 = quantile(p_sa, 0.90, na.rm = TRUE)
  )

summary_table
```

```
## # A tibble: 3 × 6
##   group        n median  mean   p10   p90
##   <fct>    <int>  <dbl> <dbl> <dbl> <dbl>
## 1 SA         239  0.711 0.676 0.293 0.945
## 2 SB         237  0.281 0.327 0.202 0.561
## 3 Chandler   156  0.828 0.802 0.567 0.964
```

- **Interpretation** - Chandler median P(SA) = 0.83; historical SA median = 0.71; historical SB median = 0.28.
    - The model assigns Chandler Bay a *higher SA-likeness score* than typical historical SA waters
    - Chandler Bay is *extremely far from the SB distribution*
    - Chandler Bay is NOT borderline SA
    - Chandler Bay is NOT in the overlap region
    - Chandler Bay observations sits firmly in what the model has learned as SA environmental space.
    - It is NOT a weak signal - It is a VERY STRONG signal

The random forest learned: Given statewide data, these are the environmental patterns that characterize SA vs. SB waters.

Then we asked: Where does Chandler Bay sit relative to that learned SA-SB boundary?

The answer: Chandler Bay looks more like SA than typical historical SA waters do.

- What It DOES Mean
    - Chandler Bay's measured environmental conditions are consistent with waters historically designated SA.
    - Chandler does not resemble SB waters in the statewide data set.
    - If classification were based solely on measured water quality patterns in this dataset, Chandler would fall within the SA range.
    - There is no evidence (in this dataset) that Chandler exhibits SB-like environmental conditions.
- It does NOT mean:
    - The model overrides statutory classification.
    - Socioeconomic or policy considerations are irrelevant.
    - Regulatory criteria have been legally met.

- How these results supports reclassification arguments:
    - The current SB designation of Chandler Bay does not appear consistent with observed environmental patterns relative to statewide SA and SB waters.
    - The median predicted probability of SA for Chandler Bay (0.83) exceeded the median probability observed for historically designated SA waters (0.71) and was substantially higher than that for historically designated SB waters (0.28). This indicates that, based solely on measured environmental conditions included in the model, Chandler Bay aligns more closely with statewide SA-designated waters than SB-designated waters. While regulatory classification incorporates statutory and policy considerations beyond instantaneous water-quality measurements, these results suggest that Chandler Bay's environmental conditions are consistent with those historically associated with SA waters.

Additional comparisons

Historical SA median:

```
median_p_sa_hist_SA <- median(
  p_sa_test[test_sasb$class_sasb == "SA"],
  na.rm = TRUE
)

median_p_sa_hist_SA
```

```
## [1] 0.7109489
```

Historical SA median of 0.71 means that half of all historical SA waters have P(SA) below 0.71 and half have P(SA) above 0.71

Historical SB median:

```
median_p_sa_hist_SB <- median(
  p_sa_test[test_sasb$class_sasb == "SB"],
  na.rm = TRUE
)

median_p_sa_hist_SB
```

```
## [1] 0.2807209
```

Proportion of Chandler Bay observations exceeding the historical SA median:

```
prop_ch_exceeds_sa_median <- mean(
  p_sa_ch > median_p_sa_hist_SA,
  na.rm = TRUE
)

prop_ch_exceeds_sa_median
```

```
## [1] 0.8012821
```

RESULT - About 80% of Chandler Bay observations exceed the typical SA median. This means that if you compare each Chandler Bay observation to the "typical SA water," most Chandler Bay samples score

higher than what is typical for SA waters statewide. AKA Chandler Bay doesn't just "barely resemble SA waters," Chandler Bay looks like upper-tier SA conditions relative to the statewide dataset.

- IF Chandler Bay was truly SB-like, you would expect
    - Most of its scores to be near SB median (0.28)
    - Very few exceeding the SA median
- BUT instead
    - The SB median is 0.28
    - Chandler Bay median is 0.83
    - 80% of Chandler Bay observations exceed the SA median
    - meaning **Chandler Bay sits far away from SB and deep inside SA territory**

Big Takeaway - The majority of Chandler Bay observations are more environmentally similar to typical SA waters than to SB waters, indicating that Chandler aligns strongly with SA-designated conditions based on measured water quality patterns.

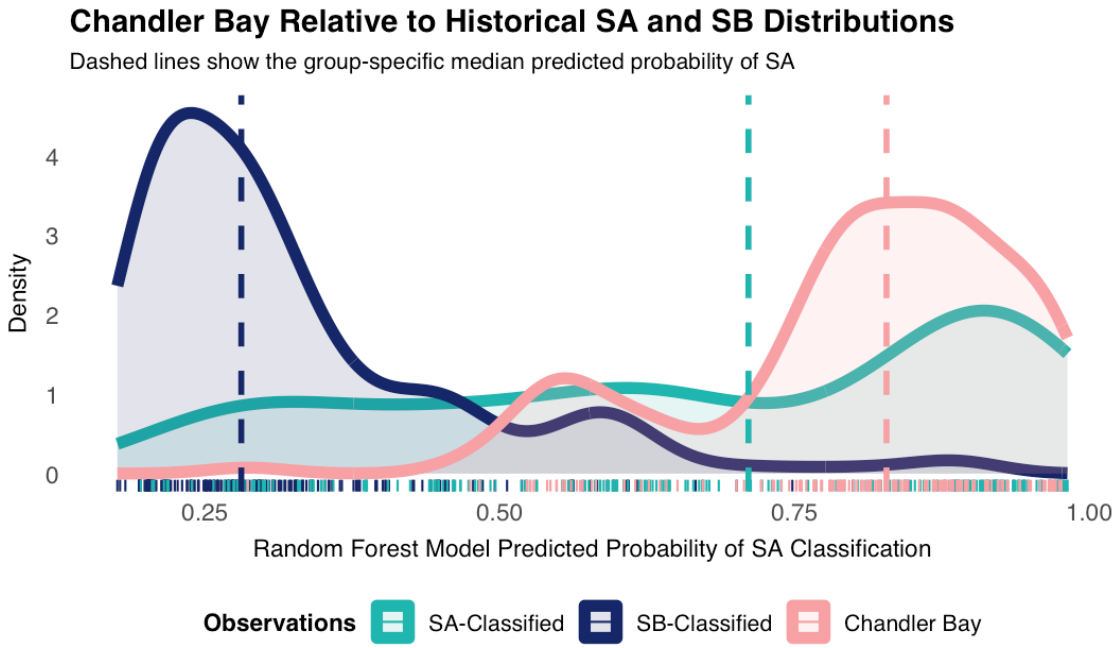
Additional Figure Options

```
sa_median <- median(
  p_sa_test[test_sasb$class_sasb == "SA"],
  na.rm = TRUE)

sb_median <- median(
  p_sa_test[test_sasb$class_sasb == "SB"],
  na.rm = TRUE)


cb_median <- median(p_sa_ch,
  na.rm = TRUE)
```

Option 2

```
ggplot(plot_df, aes(x = p_sa, fill = group, color = group)) +
  geom_density(alpha = 0.15, linewidth = 4) +
  geom_rug(alpha = 1,
           sides = "b") +
  geom_vline(xintercept = sa_median, color = "#1BB6AFFF",
             linetype = "dashed",
             linewidth = 2) +
  geom_vline(xintercept = sb_median, color = "#172869FF",
             linetype = "dashed",
             linewidth = 2) +
  geom_vline(xintercept = cb_median, color = "#F6A1A5FF",
             linetype = "dashed",
             linewidth = 2) +
  scale_fill_manual(values = c(SA = "#1BB6AFFF",
                               SB = "#172869FF",
                               Chandler = "#F6A1A5FF"),
                    labels = c(SA = "SA-Classified",
                               SB = "SB-Classified",
                               Chandler = "Chandler Bay"),
```

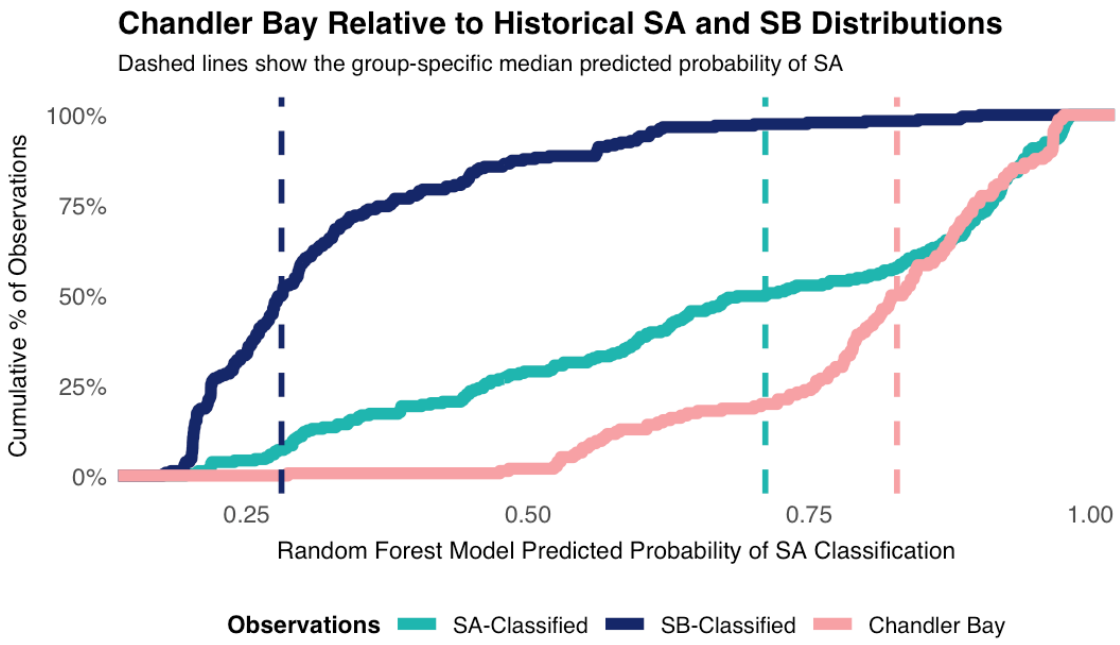
```
                    name = "Observations") +
  scale_color_manual(values = c(SA = "#1BB6AFFF",
                                SB = "#172869FF",
                                Chandler = "#F6A1A5FF"),
                     labels = c(SA = "SA-Classified",
                                SB = "SB-Classified",
                                Chandler = "Chandler Bay"),
                     name = "Observations") +
  labs(x = "Random Forest Model Predicted Probability of SA Classification",
       y = "Density",
       title = "Chandler Bay Relative to Historical SA and SB Distributions",
       subtitle = "Dashed lines show the group-specific median predicted probability
of SA") +
  hrbrthemes::theme_ipsum(base_family = "Helvetica", base_size = 17, grid = "N",
                          axis_title_size = 17, axis_title_just = "cc", axis_title_fa
ce = "plain",
                          plot_title_family = "Helvetica", plot_title_size = 22,
                          subtitle_size = 16, subtitle_family = "Helvetica",
                          plot_margin = margin(5,5,5,5)) +
  theme(panel.grid = element_blank(),
        legend.position = "bottom",
        legend.key.size = unit(1.2, "cm"),
        legend.text = element_text(size = 16, color = "black"),
        legend.title = element_text(face = "bold",size = 17),
        axis.title.x = element_text(margin = margin(t = 10, r = 0, b = 0, l = 0)),
        axis.title.y = element_text(margin = margin(t = 0, r = 10, b = 0, l = 0)))
```



Interpretation: The predicted SA probability distribution for Chandler Bay overlaps almost entirely with historically designated SA waters and is clearly separated from SB waters, indicating strong environmental similarity to SA-designated conditions.

Option 3

```
ggplot(plot_df, aes(x = p_sa, fill = group, color = group)) +
  #geom_density(alpha = 0.15, linewidth = 4) +
  stat_ecdf(linewidth = 4) +
  scale_y_continuous(labels = scales::percent_format()) +
  geom_vline(xintercept = sa_median, color = "#1BB6AFFF",
             linetype = "dashed",
             linewidth = 2) +
  geom_vline(xintercept = sb_median, color = "#172869FF",
             linetype = "dashed",
             linewidth = 2) +
  geom_vline(xintercept = cb_median, color = "#F6A1A5FF",
             linetype = "dashed",
             linewidth = 2) +
  scale_fill_manual(values = c(SA = "#1BB6AFFF",
                               SB = "#172869FF",
                               Chandler = "#F6A1A5FF"),
                    labels = c(SA = "SA-Classified",
                               SB = "SB-Classified",
                               Chandler = "Chandler Bay"),
                    name = "Observations") +
  scale_color_manual(values = c(SA = "#1BB6AFFF",
                               SB = "#172869FF",
                               Chandler = "#F6A1A5FF"),
                    labels = c(SA = "SA-Classified",
                               SB = "SB-Classified",
                               Chandler = "Chandler Bay"),
                    name = "Observations") +
  labs(x = "Random Forest Model Predicted Probability of SA Classification",
       y = "Cumulative % of Observations",
       title = "Chandler Bay Relative to Historical SA and SB Distributions",
       subtitle = "Dashed lines show the group-specific median predicted probability
of SA") +
  hrbrthemes::theme_ipsum(base_family = "Helvetica", base_size = 17, grid = "N",
                          axis_title_size = 17, axis_title_just = "cc", axis_title_fa
ce = "plain",
                          plot_title_family = "Helvetica", plot_title_size = 22,
                          subtitle_size = 16, subtitle_family = "Helvetica",
                          plot_margin = margin(5,5,5,5)) +
  theme(panel.grid = element_blank(),
        legend.position = "bottom",
        legend.key.size = unit(1.2, "cm"),
        legend.text = element_text(size = 16, color = "black"),
        legend.title = element_text(face = "bold",size = 17),
        axis.title.x = element_text(margin = margin(t = 10, r = 0, b = 0, l = 0)),
        axis.title.y = element_text(margin = margin(t = 0, r = 10, b = 0, l = 0)))
```

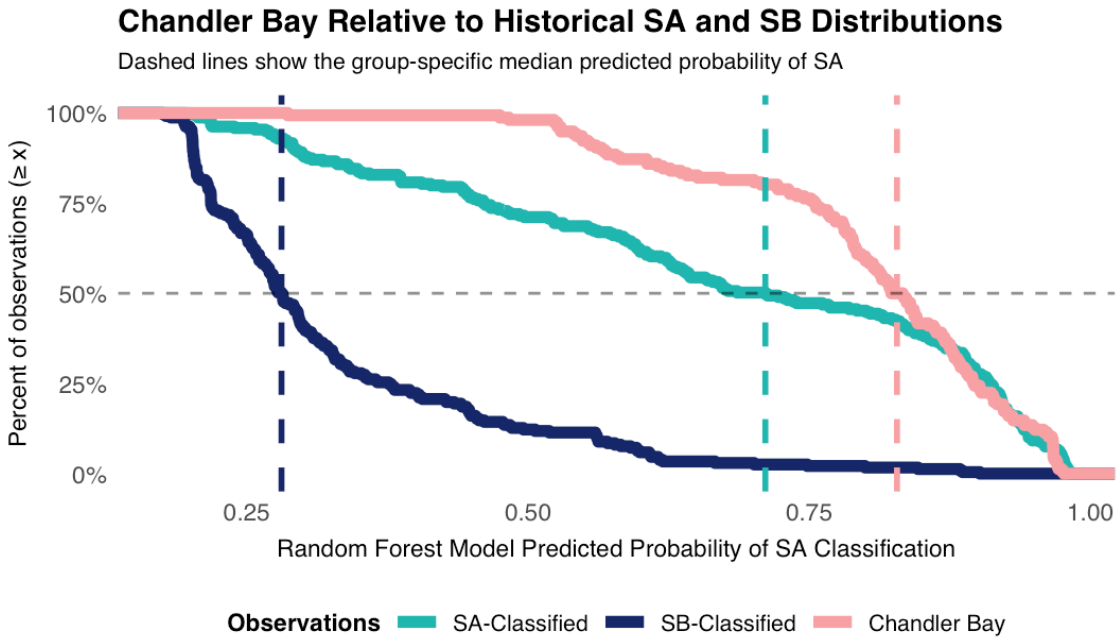


###

Option 3.1

```
ggplot(plot_df, aes(x = p_sa, fill = group, color = group)) +
  stat_ecdf(aes(y = 1 - after_stat(y)),linewidth = 4) +
  scale_y_continuous(labels = scales::percent_format()) +
  geom_vline(xintercept = sa_median, color = "#1BB6AFFF",
             linetype = "dashed",
             linewidth = 2) +
  geom_vline(xintercept = sb_median, color = "#172869FF",
             linetype = "dashed",
             linewidth = 2) +
  geom_vline(xintercept = cb_median, color = "#F6A1A5FF",
             linetype = "dashed",
             linewidth = 2) +
  geom_hline(yintercept = .5, color = "black",
             alpha = 0.5,
             linetype = "dashed",
             linewidth = 1) +
  scale_fill_manual(values = c(SA = "#1BB6AFFF",
                               SB = "#172869FF",
                               Chandler = "#F6A1A5FF"),
                    labels = c(SA = "SA-Classified",
                               SB = "SB-Classified",
                               Chandler = "Chandler Bay"),
                    name = "Observations") +
  scale_color_manual(values = c(SA = "#1BB6AFFF",
                               SB = "#172869FF",
                               Chandler = "#F6A1A5FF"),
                    labels = c(SA = "SA-Classified",
                               SB = "SB-Classified",
                               Chandler = "Chandler Bay"),
                    name = "Observations") +
  labs(x = "Random Forest Model Predicted Probability of SA Classification",
```

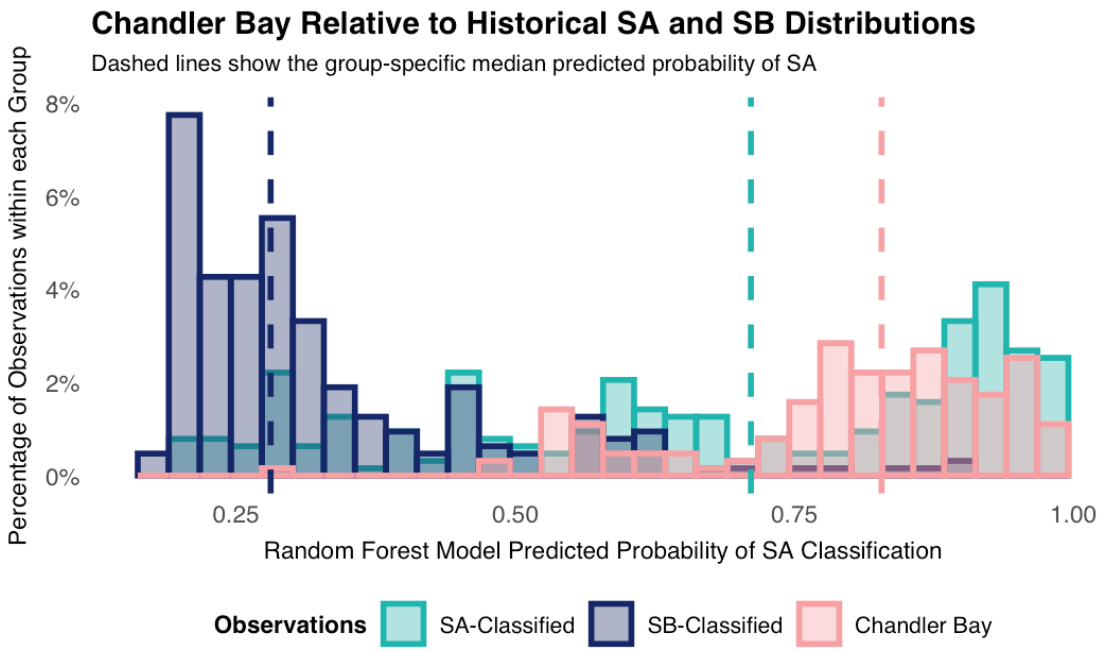
```
       y = "Percent of observations (≥ x)",
       title = "Chandler Bay Relative to Historical SA and SB Distributions",
       subtitle = "Dashed lines show the group-specific median predicted probability
of SA") +
  hrbrthemes::theme_ipsum(base_family = "Helvetica", base_size = 17, grid = "N",
                          axis_title_size = 17, axis_title_just = "cc", axis_title_fa
ce = "plain",
                          plot_title_family = "Helvetica", plot_title_size = 22,
                          subtitle_size = 16, subtitle_family = "Helvetica",
                          plot_margin = margin(5,5,5,5)) +
  theme(panel.grid = element_blank(),
        legend.position = "bottom",
        legend.key.size = unit(1.2, "cm"),
        legend.text = element_text(size = 16, color = "black"),
        legend.title = element_text(face = "bold",size = 17),
        axis.title.x = element_text(margin = margin(t = 10, r = 0, b = 0, l = 0)),
        axis.title.y = element_text(margin = margin(t = 0, r = 10, b = 0, l = 0)))
```



Option 4

```
ggplot(plot_df, aes(x = p_sa, fill = group, color = group)) +
  geom_histogram(
    aes(y = after_stat(count / sum(count))),
    position = "identity",
    alpha = 0.4, linewidth = 2,
    bins = 30) +
  scale_y_continuous(labels = scales::percent_format()) +
  #geom_density(alpha = 0.15, linewidth = 4) +
  geom_vline(xintercept = sa_median, color = "#1BB6AFFF",
             linetype = "dashed",
             linewidth = 2) +
  geom_vline(xintercept = sb_median, color = "#172869FF",
```

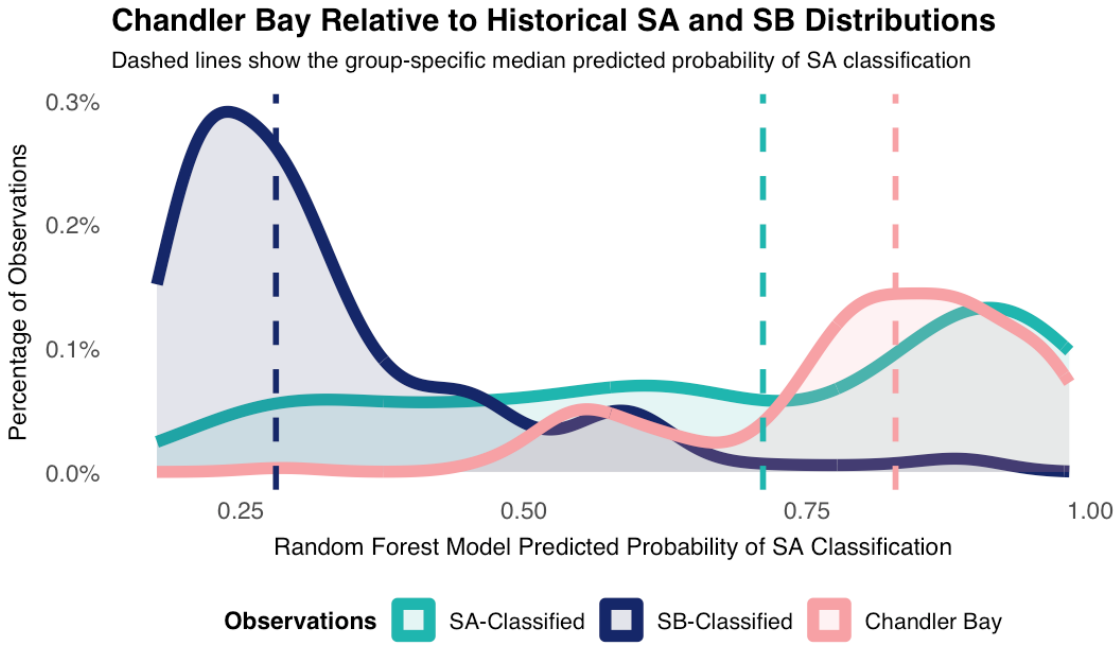
```
            linetype = "dashed",
            linewidth = 2) +
  geom_vline(xintercept = cb_median, color = "#F6A1A5FF",
            linetype = "dashed",
            linewidth = 2) +
  scale_fill_manual(values = c(SA = "#1BB6AFFF",
                               SB = "#172869FF",
                               Chandler = "#F6A1A5FF"),
                    labels = c(SA = "SA-Classified",
                               SB = "SB-Classified",
                               Chandler = "Chandler Bay"),
                    name = "Observations") +
  scale_color_manual(values = c(SA = "#1BB6AFFF",
                               SB = "#172869FF",
                               Chandler = "#F6A1A5FF"),
                    labels = c(SA = "SA-Classified",
                               SB = "SB-Classified",
                               Chandler = "Chandler Bay"),
                    name = "Observations") +
  labs(x = "Random Forest Model Predicted Probability of SA Classification",
       y = "Percentage of Observations within each Group",
       title = "Chandler Bay Relative to Historical SA and SB Distributions",
       subtitle = "Dashed lines show the group-specific median predicted probability
of SA") +
  hrbrthemes::theme_ipsum(base_family = "Helvetica", base_size = 17, grid = "N",
                          axis_title_size = 17, axis_title_just = "cc", axis_title_fa
ce = "plain",
                          plot_title_family = "Helvetica", plot_title_size = 22,
                          subtitle_size = 16, subtitle_family = "Helvetica",
                          plot_margin = margin(5,5,5,5)) +
  theme(panel.grid = element_blank(),
        legend.position = "bottom",
        legend.key.size = unit(1.2, "cm"),
        legend.text = element_text(size = 16, color = "black"),
        legend.title = element_text(face = "bold",size = 17),
        axis.title.x = element_text(margin = margin(t = 10, r = 0, b = 0, l = 0)),
        axis.title.y = element_text(margin = margin(t = 0, r = 10, b = 0, l = 0)))
```



Option 5 - use as Figure 3

```
ggplot(plot_df, aes(x = p_sa, fill = group, color = group)) +
  geom_density(aes(y = after_stat(count / sum(count))), alpha = 0.15, linewidth = 4)
+
  scale_y_continuous(labels = scales::percent_format()) +
  geom_vline(xintercept = sa_median, color = "#1BB6AFFF",
             linetype = "dashed",
             linewidth = 2) +
  geom_vline(xintercept = sb_median, color = "#172869FF",
             linetype = "dashed",
             linewidth = 2) +
  geom_vline(xintercept = cb_median, color = "#F6A1A5FF",
             linetype = "dashed",
             linewidth = 2) +
  scale_fill_manual(values = c(SA = "#1BB6AFFF",
                               SB = "#172869FF",
                               Chandler = "#F6A1A5FF"),
                    labels = c(SA = "SA-Classified",
                               SB = "SB-Classified",
                               Chandler = "Chandler Bay"),
                    name = "Observations") +
  scale_color_manual(values = c(SA = "#1BB6AFFF",
                               SB = "#172869FF",
                               Chandler = "#F6A1A5FF"),
                    labels = c(SA = "SA-Classified",
                               SB = "SB-Classified",
                               Chandler = "Chandler Bay"),
                    name = "Observations") +
  labs(x = "Random Forest Model Predicted Probability of SA Classification",
       y = "Percentage of Observations",
       title = "Chandler Bay Relative to Historical SA and SB Distributions",
```

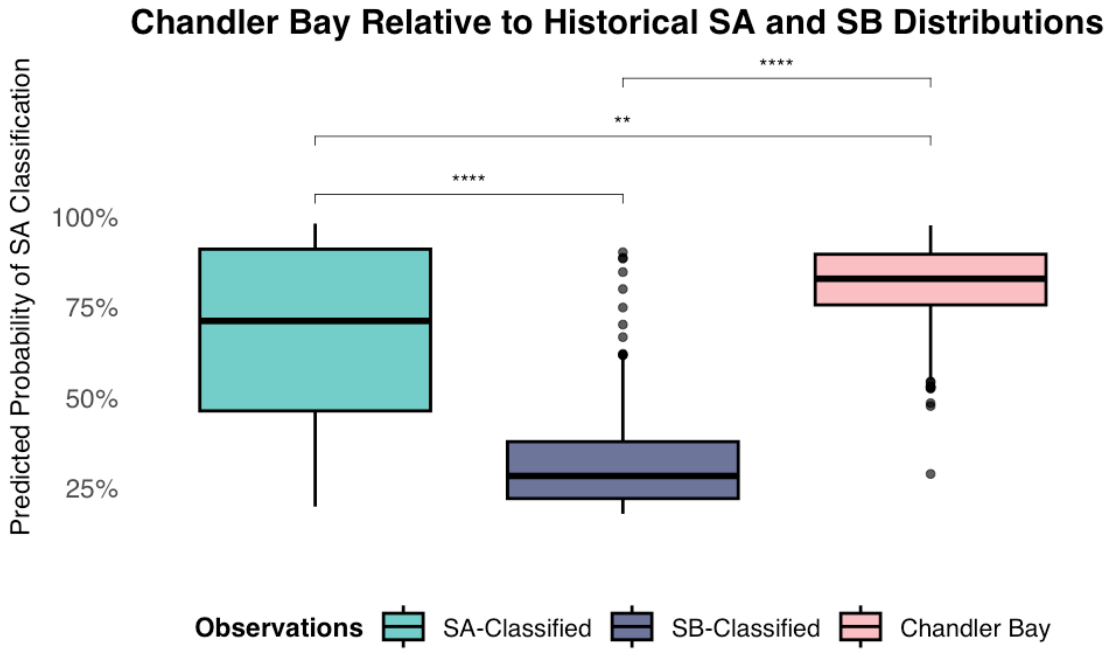
```
       subtitle = "Dashed lines show the group-specific median predicted probability
of SA classification") +
  hrbrthemes::theme_ipsum(base_family = "Helvetica", base_size = 17, grid = "N",
                          axis_title_size = 17, axis_title_just = "cc", axis_title_fa
ce = "plain",
                          plot_title_family = "Helvetica", plot_title_size = 22,
                          subtitle_size = 16, subtitle_family = "Helvetica",
                          plot_margin = margin(5,5,5,5)) +
  theme(panel.grid = element_blank(),
        legend.position = "bottom",
        legend.key.size = unit(1.2, "cm"),
        legend.text = element_text(size = 16, color = "black"),
        legend.title = element_text(face = "bold",size = 17),
        axis.title.x = element_text(margin = margin(t = 10, r = 0, b = 0, l = 0)),
        axis.title.y = element_text(margin = margin(t = 0, r = 10, b = 0, l = 0)))
```



###Option 6 – USE THIS ONE

```
ggplot(plot_df, aes(y = p_sa, fill = group,  x = group)) +
  geom_boxplot(alpha = .7, linewidth = 1,
               color = "black") +
  scale_y_continuous(breaks = c(0, 0.25, 0.5, 0.75, 1),
                     labels = scales::percent_format()) +
  scale_fill_manual(values = c(SA = "#1BB6AFFF",
                               SB = "#172869FF",
                               Chandler = "#F6A1A5FF"),
                    labels = c(SA = "SA-Classified",
                               SB = "SB-Classified",
                               Chandler = "Chandler Bay"),
                    name = "Observations") +
  scale_color_manual(values = c(SA = "#1BB6AFFF",
                               SB = "#172869FF",
                               Chandler = "#F6A1A5FF"),
```

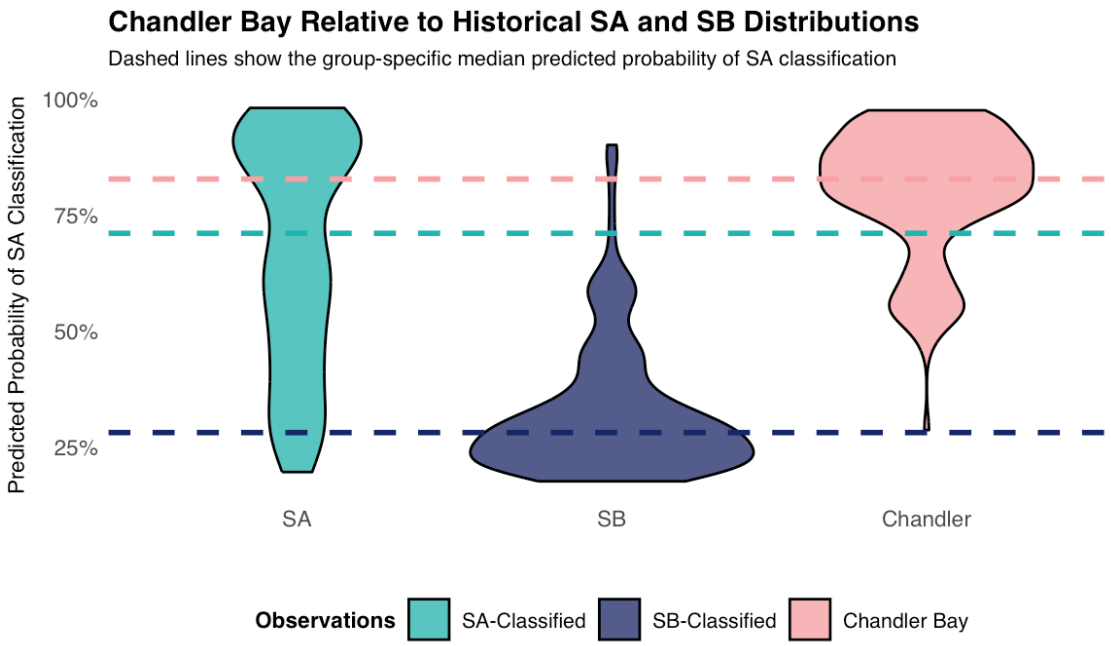
```
                    labels = c(SA = "SA-Classified",
                               SB = "SB-Classified",
                               Chandler = "Chandler Bay"),
                    name = "Observations") +
  labs(x = "",
       y = "Predicted Probability of SA Classification",
       title = "Chandler Bay Relative to Historical SA and SB Distributions") +
  hrbrthemes::theme_ipsum(base_family = "Helvetica", base_size = 17, grid = "N",
                          axis_title_size = 17, axis_title_just = "cc", axis_title_fa
ce = "plain",
                          plot_title_family = "Helvetica", plot_title_size = 22,
                          subtitle_size = 16, subtitle_family = "Helvetica",
                          plot_margin = margin(5,5,5,5)) +
  ggpubr::geom_pwc(label = "{p.adj.signif}", method = "dunn_test",
           bracket.nudge.y = 0.1, p.adjust.method = "bonferroni",
           hide.ns = TRUE, step.increase = 0.2,
           label.size = 5, family = "Helvetica") +
  theme(panel.grid = element_blank(),
        legend.position = "bottom",
        legend.key.size = unit(1.2, "cm"),
        legend.text = element_text(size = 16, color = "black"),
        legend.title = element_text(face = "bold",size = 17),
        axis.text.x = element_blank(),
        axis.title.x = element_text(margin = margin(t = 0, r = 0, b = 0, l = 0)),
        axis.title.y = element_text(margin = margin(t = 0, r = 10, b = 0, l = 0)))
```



Option 7

```
ggplot(plot_df, aes(y = p_sa, fill = group,  x = group)) +
  geom_violin(alpha = .8, linewidth = 1,
              color = "black") +
  scale_y_continuous(labels = scales::percent_format()) +
```

```
  geom_hline(yintercept = sa_median, color = "#1BB6AFFF",
             linetype = "dashed",
             linewidth = 2) +
  geom_hline(yintercept = sb_median, color = "#172869FF",
             linetype = "dashed",
             linewidth = 2) +
  geom_hline(yintercept = cb_median, color = "#F6A1A5FF",
             linetype = "dashed",
             linewidth = 2) +
  scale_fill_manual(values = c(SA = "#1BB6AFFF",
                               SB = "#172869FF",
                               Chandler = "#F6A1A5FF"),
                    labels = c(SA = "SA-Classified",
                               SB = "SB-Classified",
                               Chandler = "Chandler Bay"),
                    name = "Observations") +
  scale_color_manual(values = c(SA = "#1BB6AFFF",
                               SB = "#172869FF",
                               Chandler = "#F6A1A5FF"),
                    labels = c(SA = "SA-Classified",
                               SB = "SB-Classified",
                               Chandler = "Chandler Bay"),
                    name = "Observations") +
  labs(x = "",
       y = "Predicted Probability of SA Classification",
       title = "Chandler Bay Relative to Historical SA and SB Distributions",
       subtitle = "Dashed lines show the group-specific median predicted probability
of SA classification") +
  hrbrthemes::theme_ipsum(base_family = "Helvetica", base_size = 17, grid = "N",
                          axis_title_size = 17, axis_title_just = "cc", axis_title_fa
ce = "plain",
                          plot_title_family = "Helvetica", plot_title_size = 22,
                          subtitle_size = 16, subtitle_family = "Helvetica",
                          plot_margin = margin(5,5,5,5)) +
  #ggpubr::geom_pwc(label = "{p.adj.format}", method = "dunn_test",
      #     bracket.nudge.y = 0.1, p.adjust.method = "bonferroni",
      #     hide.ns = TRUE, step.increase = 0.2,
      #     label.size = 5, family = "Helvetica") +
  theme(panel.grid = element_blank(),
        legend.position = "bottom",
        legend.key.size = unit(1.2, "cm"),
        legend.text = element_text(size = 16, color = "black"),
        legend.title = element_text(face = "bold",size = 17),
        axis.title.x = element_text(margin = margin(t = 10, r = 0, b = 0, l = 0)),
        axis.title.y = element_text(margin = margin(t = 0, r = 10, b = 0, l = 0)))
```



.

Removing ‘measured’ variables from the RF models

```
df_sasb_chem <- df_sasb |>
  dplyr::select(
    -dplyr::matches("_measured$")
  )

set.seed(123)

split_obj_chem <- rsample::initial_split(df_sasb_chem,
                                         prop = 0.8,
                                         strata = class_sasb)

train_chem <- rsample::training(split_obj_chem)
test_chem  <- rsample::testing(split_obj_chem)
```

Fit chemistry-only random forest

```
set.seed(123)

rf_chem <- ranger::ranger(
  class_sasb ~ .,
  data = train_chem,
  probability = TRUE,
  importance = "permutation",
  num.trees = 1000,
```

```
  min.node.size = 10
)
```

Evaluate Performance

```
prob_test_chem <- predict(rf_chem, data = test_chem)$predictions
p_sa_test_chem <- prob_test_chem[, "SA"]

pred_test_chem <- factor(
  ifelse(p_sa_test_chem >= 0.5, "SA", "SB"),
  levels = c("SA", "SB")
)

eval_chem <- tibble(
  truth = test_chem$class_sasb,
  estimate = pred_test_chem
)

yardstick::conf_mat(eval_chem, truth = truth, estimate = estimate)

##           Truth
## Prediction  SA  SB
##         SA 179  43
##         SB  60 194

sens(eval_chem, truth = truth, estimate = estimate, event_level = "first")

## # A tibble: 1 × 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 sens    binary         0.749

precision(eval_chem, truth = truth, estimate = estimate, event_level = "first")

## # A tibble: 1 × 3
##   .metric   .estimator .estimate
##   <chr>     <chr>          <dbl>
## 1 precision binary         0.806
```

Applying Chemistry-Only Model to Chandler Bay

```
make_rf_features_chandler_chem <- function(ch) {
  ch |>
    dplyr::transmute(
      temperature_deg_c = temperature_deg_c,
      salinity = salinity,
      do_mg_l = do_mg_l,
      p_h = p_h,
      ammonia_mg_l = ammonia_mg_l,
      #nitrate_nitrite_mg_l = nitrate_nitrite_mg_l,
      enterococcus_mp_nper100m_l = NA_real_
    )
}
```

```r
train_cols_chem <- setdiff(names(train_chem), "class_sasb")

ch_feat_chem <- make_rf_features_chandler_chem(chandler_clean)

ch_feat_chem <- ch_feat_chem[, ..train_cols_chem]

p_sa_ch_chem <- predict(
  rf_chem,
  data = ch_feat_chem
)$predictions[, "SA"]

summary(p_sa_ch_chem)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4453  0.6901  0.7948  0.7856  0.8775  0.9985



median(p_sa_ch_chem)

## [1] 0.7948262



mean(p_sa_ch_chem)

## [1] 0.7856425
```

RESULTS - The SA-like classification of Chandler Bay is NOT being driven by monitoring regime variables. It is being driven by measured environmental chemistry alone. Even after removing all monitoring pattern information, Chandler Bay remains strongly SA-like.

```r
median_sa_chem <- median(
  p_sa_test_chem[test_chem$class_sasb == "SA"],
  na.rm = TRUE
)

mean(p_sa_ch_chem > median_sa_chem)

## [1] 0.6025641
```

This tells you: What proportion of Chandler observations exceed the typical SA water under chemistry-only modeling?

- This means: Monitoring regime variables were contributing positively to Chandler's SA-likeness score.
    - IT DOES NOT MEAN CHANDLER BAY IS SB-LIKE
    - It means:
        - Some of the strength of the original signal came from measurement patterns.

        - When you remove those, the signal weakens slightly.
        - But the majority (60%) still exceed typical SA.
- Before: Chandler looked strongly SA-like both environmentally and structurally.
- After removing monitoring information: Chandler still looks more SA-like than SB — but not as overwhelmingly.
- So the environmental signal is real, but part of the earlier strength reflected monitoring structure.

What this means Scientifically

- The original 80% result reflected two components:
    - Environmental chemistry
    - Monitoring structure
- When you removed monitoring structure: Only environmental chemistry remained.
- The drop from 80% → 60% suggests:
    - Monitoring patterns amplified the signal.
    - But they did not create the signal.
- It does NOT mean:
    - The model is invalid.
    - The chemistry signal disappears.
    - Chandler is SB-like.
    - The result was artificial.
- It means: The strength of evidence becomes moderate-to-strong rather than extremely strong.

Could say something in the report along the lines of: When monitoring-indicator variables were removed, the proportion of Chandler observations exceeding the historical SA median decreased from approximately 80% to 60%. This indicates that while sampling structure contributed to the strength of the original signal, the majority of Chandler observations remained more environmentally similar to typical SA waters than SB waters. Thus, the SA-like classification is primarily driven by measured environmental conditions, although monitoring patterns modestly reinforced this signal.

Variable importance

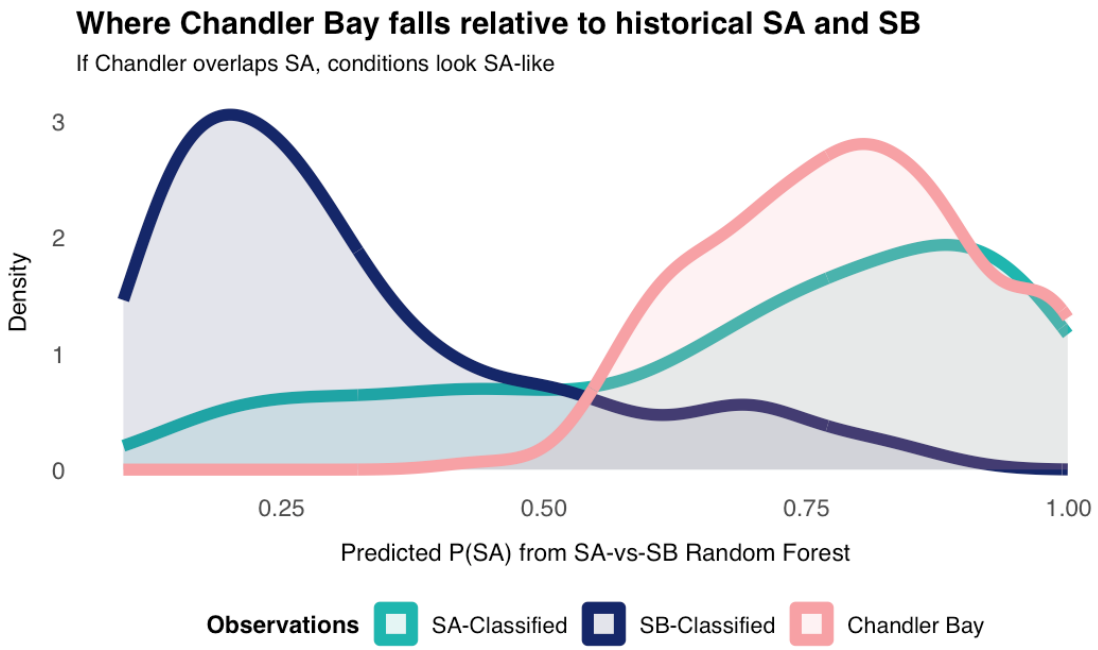
```
imp_chem <- sort(rf_chem$variable.importance, decreasing = TRUE)
head(imp_chem, 20)
```

```
## enterococcus_mp_nper100m_l                   salinity
##                  0.1471179                  0.1327646
##          temperature_deg_c                    do_mg_l
##                  0.1312691                  0.1164823
##                        p_h                ammonia_mg_l
##                  0.1115071                  0.1102456
```

Other visualization of the parameter only model

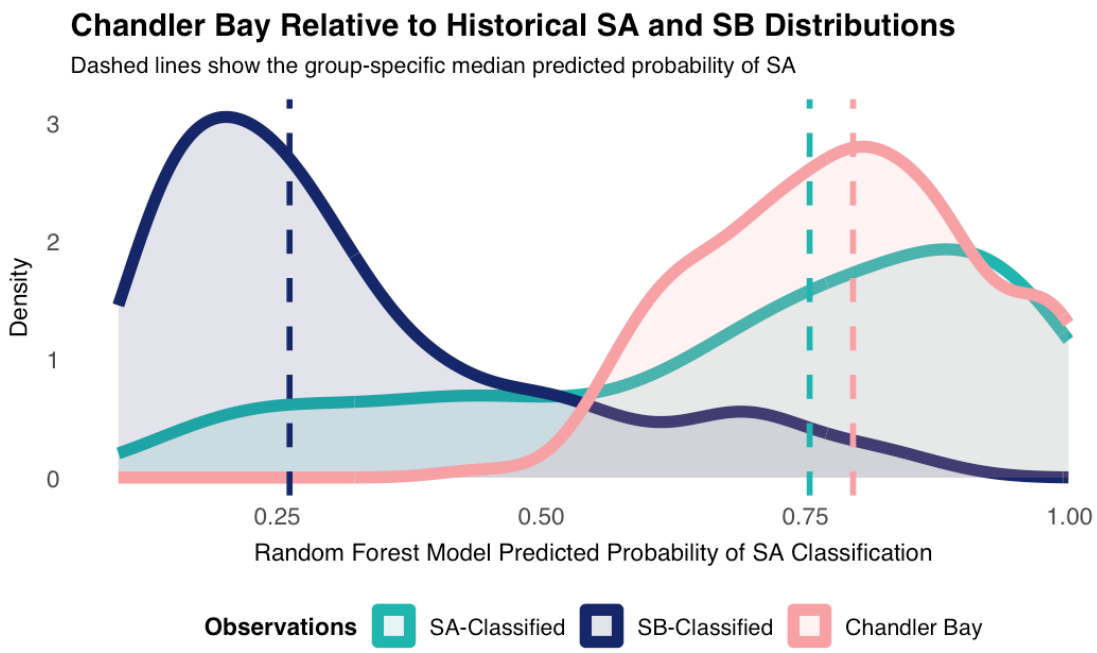
```
hist_df <- tibble(
  group = test_chem$class_sasb,  # SA or SB
  p_sa = p_sa_test_chem)
```

```r
ch_df <- tibble(
  group = factor("Chandler", levels = c("SA", "SB", "Chandler")),
  p_sa = p_sa_ch_chem)

plot_df <- bind_rows(
  hist_df |> mutate(group = factor(as.character(group), levels = c("SA","SB","Chandle
r"))),
  ch_df)

ggplot(plot_df, aes(x = p_sa, fill = group, color = group)) +
  geom_density(alpha = 0.15, linewidth = 4) +
  scale_fill_manual(values = c(SA = "#1BB6AFFF",
                               SB = "#172869FF",
                               Chandler = "#F6A1A5FF"),
                    labels = c(SA = "SA-Classified",
                               SB = "SB-Classified",
                               Chandler = "Chandler Bay"),
                    name = "Observations") +
  scale_color_manual(values = c(SA = "#1BB6AFFF",
                               SB = "#172869FF",
                               Chandler = "#F6A1A5FF"),
                    labels = c(SA = "SA-Classified",
                               SB = "SB-Classified",
                               Chandler = "Chandler Bay"),
                    name = "Observations") +
  labs(x = "Predicted P(SA) from SA-vs-SB Random Forest",
       y = "Density",
       title = "Where Chandler Bay falls relative to historical SA and SB",
       subtitle = "If Chandler overlaps SA, conditions look SA-like") +
  hrbrthemes::theme_ipsum(base_family = "Helvetica", base_size = 17, grid = "N",
                          axis_title_size = 17, axis_title_just = "cc", axis_title_fa
ce = "plain",
                          plot_title_family = "Helvetica", plot_title_size = 22,
                          subtitle_size = 16, subtitle_family = "Helvetica",
                          plot_margin = margin(5,5,5,5)) +
  theme(panel.grid = element_blank(),
        legend.position = "bottom",
        legend.key.size = unit(1.2, "cm"),
        legend.text = element_text(size = 16, color = "black"),
        legend.title = element_text(face = "bold",size = 17),
        axis.title.x = element_text(margin = unit(c(t = 1, r = 0, b = 0, l = 0), "lin
e")),
        axis.title.y = element_text(margin = unit(c(t = 0, r = 1, b = 0, l = 0), "lin
e")))
```



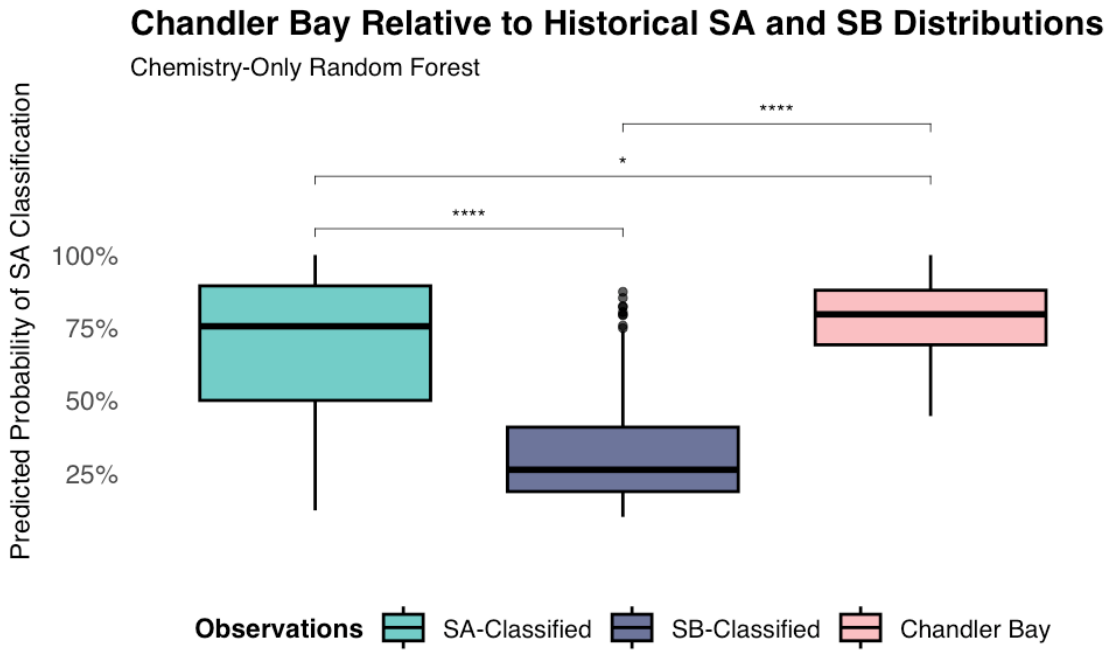

```r
sa_median <- median(
  p_sa_test_chem[test_chem$class_sasb == "SA"],
  na.rm = TRUE)

sb_median <- median(
  p_sa_test_chem[test_chem$class_sasb == "SB"],
  na.rm = TRUE)


cb_median <- median(p_sa_ch_chem,
  na.rm = TRUE)

ggplot(plot_df, aes(x = p_sa, fill = group, color = group)) +
  geom_density(alpha = 0.15, linewidth = 4) +
  geom_vline(xintercept = sa_median, color = "#1BB6AFFF",
             linetype = "dashed",
             linewidth = 2) +
  geom_vline(xintercept = sb_median, color = "#172869FF",
             linetype = "dashed",
             linewidth = 2) +
  geom_vline(xintercept = cb_median, color = "#F6A1A5FF",
             linetype = "dashed",
             linewidth = 2) +
  scale_fill_manual(values = c(SA = "#1BB6AFFF",
                               SB = "#172869FF",
                               Chandler = "#F6A1A5FF"),
                    labels = c(SA = "SA-Classified",
                               SB = "SB-Classified",
                               Chandler = "Chandler Bay"),
                    name = "Observations") +
  scale_color_manual(values = c(SA = "#1BB6AFFF",
                               SB = "#172869FF",
                               Chandler = "#F6A1A5FF"),
```

```
                    labels = c(SA = "SA-Classified",
                               SB = "SB-Classified",
                               Chandler = "Chandler Bay"),
                    name = "Observations") +
  labs(x = "Random Forest Model Predicted Probability of SA Classification",
       y = "Density",
       title = "Chandler Bay Relative to Historical SA and SB Distributions",
       subtitle = "Dashed lines show the group-specific median predicted probability
of SA") +
  hrbrthemes::theme_ipsum(base_family = "Helvetica", base_size = 17, grid = "N",
                          axis_title_size = 17, axis_title_just = "cc", axis_title_fa
ce = "plain",
                          plot_title_family = "Helvetica", plot_title_size = 22,
                          subtitle_size = 16, subtitle_family = "Helvetica",
                          plot_margin = margin(5,5,5,5)) +
  theme(panel.grid = element_blank(),
        legend.position = "bottom",
        legend.key.size = unit(1.2, "cm"),
        legend.text = element_text(size = 16, color = "black"),
        legend.title = element_text(face = "bold",size = 17),
        axis.title.x = element_text(margin = margin(t = 10, r = 0, b = 0, l = 0)),
        axis.title.y = element_text(margin = margin(t = 0, r = 10, b = 0, l = 0)))
```




```
ggplot(plot_df, aes(y = p_sa, fill = group,  x = group)) +
  geom_boxplot(alpha = .7, linewidth = 1,
               color = "black") +
  scale_y_continuous(breaks = c(0, 0.25, 0.5, 0.75, 1),
                     labels = scales::percent_format()) +
  scale_fill_manual(values = c(SA = "#1BB6AFFF",
                               SB = "#172869FF",
                               Chandler = "#F6A1A5FF"),
                    labels = c(SA = "SA-Classified",
                               SB = "SB-Classified",
```

```
                                Chandler = "Chandler Bay"),
                     name = "Observations") +
  scale_color_manual(values = c(SA = "#1BB6AFFF",
                                SB = "#172869FF",
                                Chandler = "#F6A1A5FF"),
                     labels = c(SA = "SA-Classified",
                                SB = "SB-Classified",
                                Chandler = "Chandler Bay"),
                     name = "Observations") +
  labs(x = "",
       y = "Predicted Probability of SA Classification",
       title = "Chandler Bay Relative to Historical SA and SB Distributions",
       subtitle = "Chemistry-Only Random Forest") +
  hrbrthemes::theme_ipsum(base_family = "Helvetica", base_size = 17, grid = "N",
                          axis_title_size = 17, axis_title_just = "cc", axis_title_fa
ce = "plain",
                          plot_title_family = "Helvetica", plot_title_size = 22,
                          subtitle_size = 16, subtitle_family = "Helvetica",
                          plot_margin = margin(5,5,5,5)) +
  ggpubr::geom_pwc(label = "{p.adj.signif}", method = "dunn_test",
           bracket.nudge.y = 0.1, p.adjust.method = "bonferroni",
           hide.ns = TRUE, step.increase = 0.2,
           label.size = 5, family = "Helvetica") +
  theme(panel.grid = element_blank(),
        legend.position = "bottom",
        legend.key.size = unit(1.2, "cm"),
        legend.text = element_text(size = 16, color = "black"),
        legend.title = element_text(face = "bold",size = 17),
        axis.text.x = element_blank(),
        axis.title.x = element_text(margin = margin(t = 0, r = 0, b = 0, l = 0)),
        axis.title.y = element_text(margin = margin(t = 0, r = 10, b = 0, l = 0)))
```

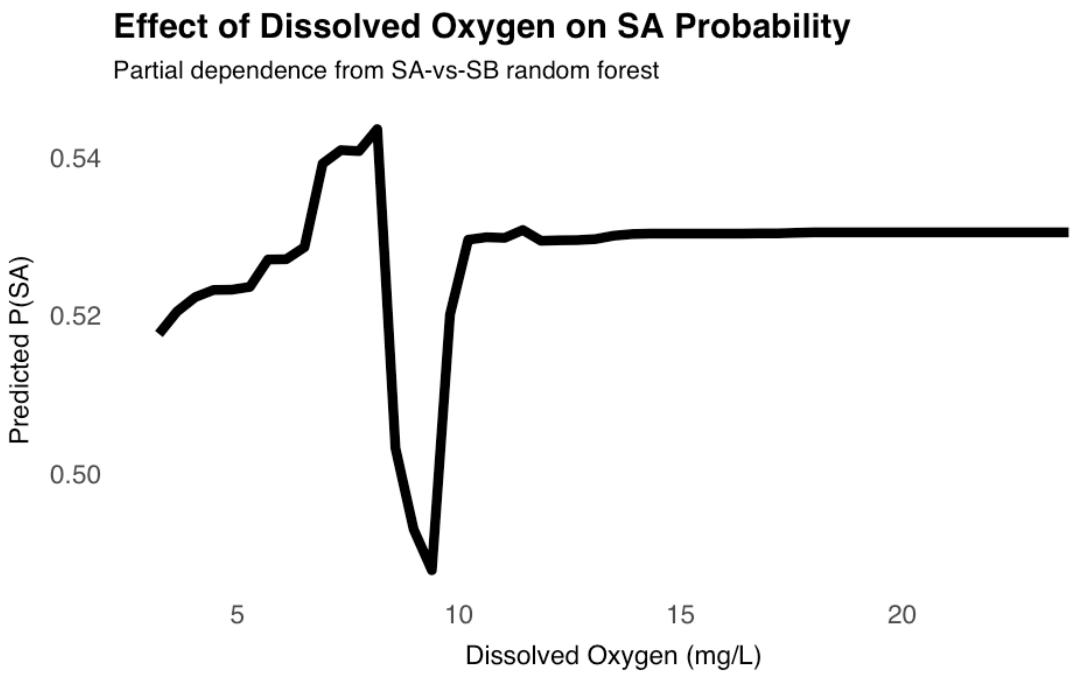


.

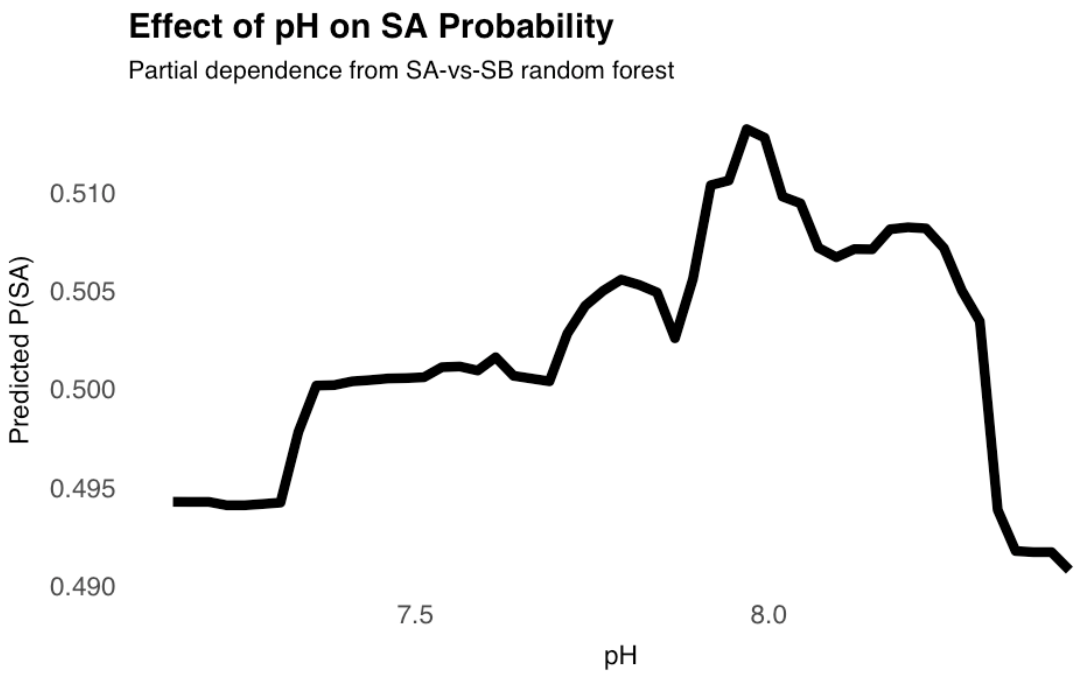
Additional Analyses and Figures that are NOT included in Final Report:

Visualizing Parameters

```
pdp_do <- partial(
  rf_sasb,
  pred.var = "do_mg_l",
  train = train_sasb,
  prob = TRUE,
  which.class = "SA")

ggplot(pdp_do, aes(x = do_mg_l, y = yhat)) +
  geom_line(linewidth = 3) +
  labs(x = "Dissolved Oxygen (mg/L)",
       y = "Predicted P(SA)",
       title = "Effect of Dissolved Oxygen on SA Probability",
       subtitle = "Partial dependence from SA-vs-SB random forest") +
  hrbrthemes::theme_ipsum(base_family = "Helvetica", base_size = 17, grid = "N",
                          axis_title_size = 17, axis_title_just = "cc", axis_title_fa
ce = "plain",
                          plot_title_family = "Helvetica", plot_title_size = 22,
                          subtitle_size = 16, subtitle_family = "Helvetica",
                          plot_margin = margin(5,5,5,5)) +
  theme(panel.grid = element_blank(),
        legend.position = "bottom",
        legend.key.size = unit(1.2, "cm"),
        legend.text = element_text(size = 16, color = "black"),
        legend.title = element_text(face = "bold",size = 17),
        axis.title.x = element_text(margin = margin(t = 10, r = 0, b = 0, l = 0)),
        axis.title.y = element_text(margin = margin(t = 0, r = 10, b = 0, l = 0)))
```




```
pdp_ph <- partial(
  rf_sasb,
  pred.var = "p_h",
  train = train_sasb,
  prob = TRUE,
  which.class = "SA")

ggplot(pdp_ph, aes(x = p_h, y = yhat)) +
  geom_line(linewidth = 3) +
  labs(x = "pH",
       y = "Predicted P(SA)",
       title = "Effect of pH on SA Probability",
       subtitle = "Partial dependence from SA-vs-SB random forest") +
  hrbrthemes::theme_ipsum(base_family = "Helvetica", base_size = 17, grid = "N",
                          axis_title_size = 17, axis_title_just = "cc", axis_title_fa
ce = "plain",
                          plot_title_family = "Helvetica", plot_title_size = 22,
                          subtitle_size = 16, subtitle_family = "Helvetica",
                          plot_margin = margin(5,5,5,5)) +
  theme(panel.grid = element_blank(),
        legend.position = "bottom",
        legend.key.size = unit(1.2, "cm"),
        legend.text = element_text(size = 16, color = "black"),
        legend.title = element_text(face = "bold",size = 17),
        axis.title.x = element_text(margin = margin(t = 10, r = 0, b = 0, l = 0)),
        axis.title.y = element_text(margin = margin(t = 0, r = 10, b = 0, l = 0)))
```



.

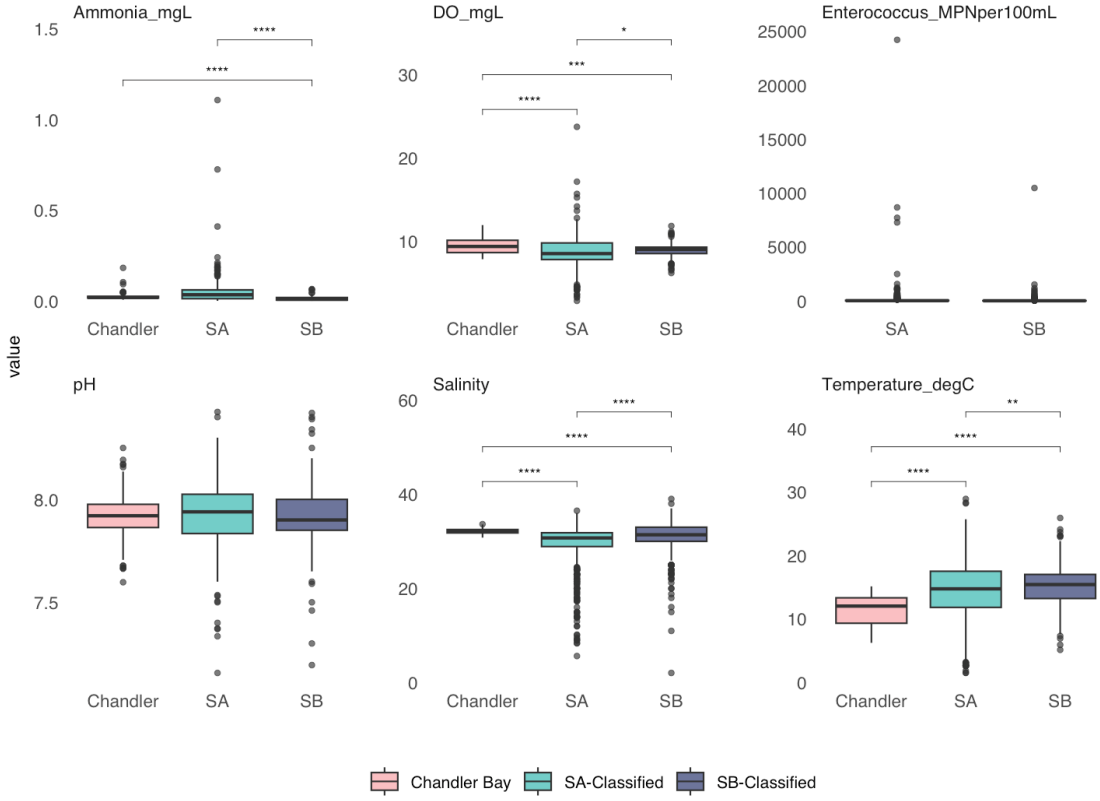
Visualizing the raw data

```
rm(list=ls())

#NWQMC = fread("Data Files/Cleaned Data/Manually Edited/NWQMC Manually Edited Data cr
eated 2.9.26.csv")
#NWQMC = fread("Data Files/Cleaned Data/Manually Edited/NWQMC_biological query_Maine
Water Quality created 2.18.26.csv")

NWQMC = fread("Data Files/Cleaned Data/Manually Edited/NWQMC water quality download_m
anually edited_created 2.18.26.csv")

EGAD = fread("Data Files/Cleaned Data/Manually Edited/EGAD Manually Edited Data creat
ed 2.9.26.csv")

combined <- bind_rows(source_NWQMC = NWQMC,
                      source_EGAD = EGAD,
                      .id = "source")

## New names:
## • `Total Phosphorus_mixed forms_mgL` -> `Total Phosphorus_mixed forms_mgL...36`
## • `Nitrate + Nitrite_mgL` -> `Nitrate + Nitrite_mgL...43`
## • `Nitrate + Nitrite_mgL` -> `Nitrate + Nitrite_mgL...54`
## • `Total Phosphorus_mixed forms_mgL` -> `Total Phosphorus_mixed forms_mgL...55`

df <- combined |>
  mutate(
```

```
    Marine_Class = factor(`Marine Class`, levels = c("SA", "SB", "SC")),
    Date = as.Date(Date, format = "%m/%d/%y"))

core_vars <- c("Temperature_degC",
               "Salinity")

h20_vars <- c("DO_mgL",
              "pH")

nutrient_vars <- c("Ammonia_mgL")

bio_vars <- c("Enterococcus_MPNper100mL")

all_vars <- c(core_vars, h20_vars, nutrient_vars, bio_vars)

non_core_vars <- setdiff(all_vars, core_vars)

historical_df <- df |>
  dplyr::filter(
    dplyr::if_any(
      dplyr::any_of(non_core_vars),
      \(x) !is.na(x))) |>
  dplyr::filter(`Marine Class` %in% c("SA", "SB")) |>
  dplyr::mutate(
    class_sasb = factor(`Marine Class`, levels = c("SA", "SB"))) |>
  dplyr::select(-`Marine Class`) |>
  droplevels() |>
  st_as_sf(coords = c("Longitude", "Latitude"), crs = 4326, remove = FALSE)

chandler_df <- fread("Data Files/Cleaned Data/Manually Edited/Kingfish Only Chandler
Bay WQData Manually edited 2.13.26.csv") |>
  st_as_sf(coords = c("Longitude", "Latitude"), crs = 4326, remove = FALSE) |>
  mutate(Date = as.Date(Date, format = "%m/%d/%y"))

historical_df$source = as.factor(historical_df$source)

chandler_df$Marine_Class <- "Chandler"
chandler_df$source <- "Kingfish"

combined_plot <- dplyr::bind_rows(historical_df, chandler_df)

## New names:
## • `Nitrate + Nitrite_mgL` -> `Nitrate + Nitrite_mgL...65`
## • `Total Phosphorus_mixed forms_mgL` -> `Total Phosphorus_mixed forms_mgL...66`

analysis_df <- combined_plot |>
  st_drop_geometry() |>
  select(
    source,
    class_sasb,
    Temperature_degC,
    Salinity,
    DO_mgL,
    pH,
    Ammonia_mgL,
```

```
    Enterococcus_MPNper100mL
  )

analysis_df <- analysis_df |>
  mutate(
    WaterType = case_when(
      source == "Kingfish" ~ "Chandler",
      class_sasb == "SA" ~ "SA",
      class_sasb == "SB" ~ "SB"
    )
  )

colMeans(is.na(analysis_df))

##                   source               class_sasb         Temperature_degC
##               0.00000000               0.06163572               0.01501383
##                 Salinity                    DO_mgL                       pH
##               0.05887001               0.58909522               0.72421968
##              Ammonia_mgL Enterococcus_MPNper100mL                WaterType
##               0.82971158               0.40181746               0.00000000

long_df <- analysis_df |>
  pivot_longer(
    cols = -c(source, class_sasb, WaterType),
    names_to = "variable",
    values_to = "value"
  ) |>
  filter(!is.na(value))

ggplot(long_df, aes(y = value, x = WaterType,
                    fill = WaterType)) +
  geom_boxplot(alpha = 0.7) +
  facet_wrap(~ variable, scales = "free") +
  scale_fill_manual(values = c(SA = "#1BB6AFFF",
                               SB = "#172869FF",
                               Chandler = "#F6A1A5FF"),
                    labels = c(SA = "SA-Classified",
                               SB = "SB-Classified",
                               Chandler = "Chandler Bay"),
                    name = NULL) +
  hrbrthemes::theme_ipsum(base_family = "Helvetica", base_size = 17, grid = "N",
                          axis_title_size = 17, axis_title_just = "cc", axis_title_fa
ce = "plain",
                          strip_text_size = 17,
                          plot_title_family = "Helvetica", plot_title_size = 22,
                          subtitle_size = 16, subtitle_family = "Helvetica",
                          plot_margin = margin(5,5,5,5)) +
  xlab("") +
  ggpubr::geom_pwc(label = "{p.adj.signif}", method = "dunn_test",
           bracket.nudge.y = 0.1, p.adjust.method = "bonferroni",
           hide.ns = TRUE, step.increase = 0.2,
           label.size = 5, family = "Helvetica") +
  theme(panel.grid = element_blank(),
        legend.position = "bottom",
        legend.key.size = unit(1.2, "cm"),
```

```
        legend.text = element_text(size = 16, color = "black"),
        legend.title = element_text(face = "bold",size = 17),
        axis.title.x = element_text(margin = margin(t = 10, r = 0, b = 0, l = 0)),
        axis.title.y = element_text(margin = margin(t = 0, r = 10, b = 0, l = 0)))
```

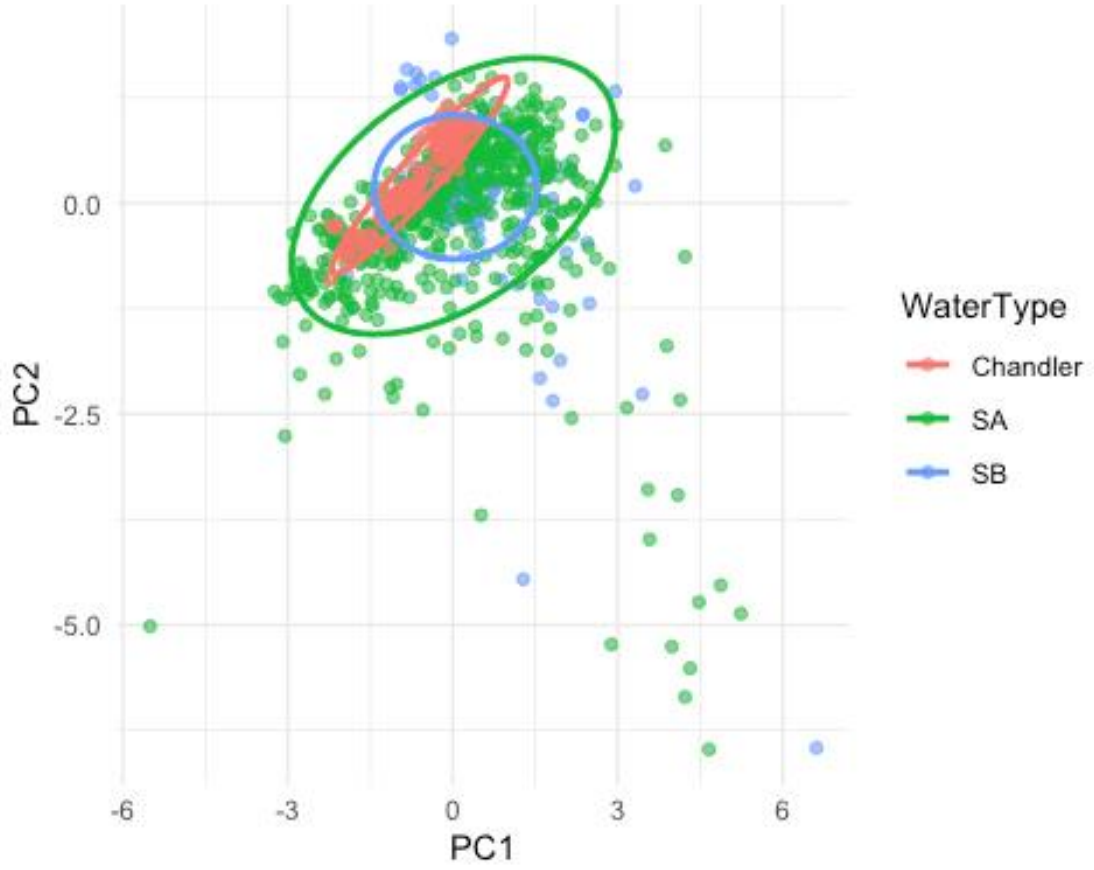



```
analysis_df |>
  summarise(across(
    c(Temperature_degC, Salinity, DO_mgL, pH, Ammonia_mgL,  Enterococcus_MPNper100mL)
,
    ~ mean(is.na(.x))
  )) |>
  tidyr::pivot_longer(everything(), names_to = "var", values_to = "prop_na") |>
  arrange(desc(prop_na))
```

```
## # A tibble: 6 × 2
##   var                       prop_na
##   <chr>                       <dbl>
## 1 Ammonia_mgL               0.830
## 2 pH                        0.724
## 3 DO_mgL                    0.589
## 4 Enterococcus_MPNper100mL  0.402
## 5 Salinity                  0.0589
## 6 Temperature_degC          0.0150
```

```
vars_3 <- c("Temperature_degC", "Salinity", "DO_mgL")

pca_df <- analysis_df |>
  drop_na(all_of(vars_3))
```

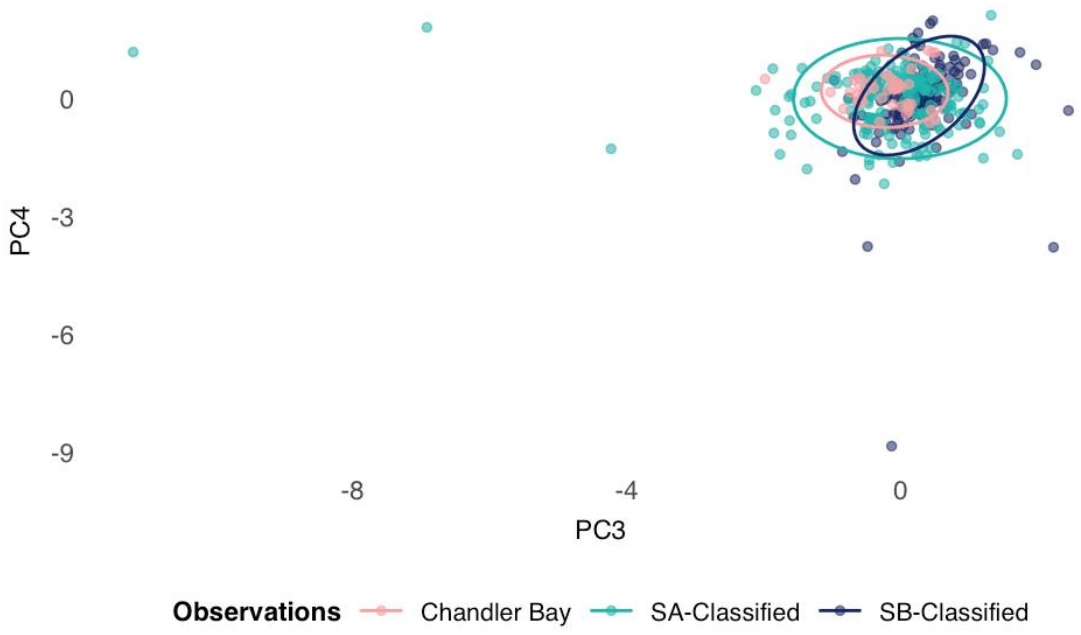
```r
pca_scaled <- scale(pca_df |> select(all_of(vars_3)))

pca_fit <- prcomp(pca_scaled)

scores <- as.data.frame(pca_fit$x[, 1:2]) |>
  dplyr::bind_cols(WaterType = pca_df$WaterType)

ggplot(scores, aes(PC1, PC2, color = WaterType)) +
  geom_point(alpha = 0.6) +
  stat_ellipse(level = 0.95, linewidth = 1) +
  theme_minimal()
```




```r
sa_mat <- pca_scaled[pca_df$WaterType == "SA", ]
sb_mat <- pca_scaled[pca_df$WaterType == "SB", ]
ch_mat <- pca_scaled[pca_df$WaterType == "Chandler", ]

mean(mahalanobis(ch_mat, colMeans(sa_mat), cov(sa_mat)))

## [1] 1.066639

mean(mahalanobis(ch_mat, colMeans(sb_mat), cov(sb_mat)))

## [1] 3.056191
```

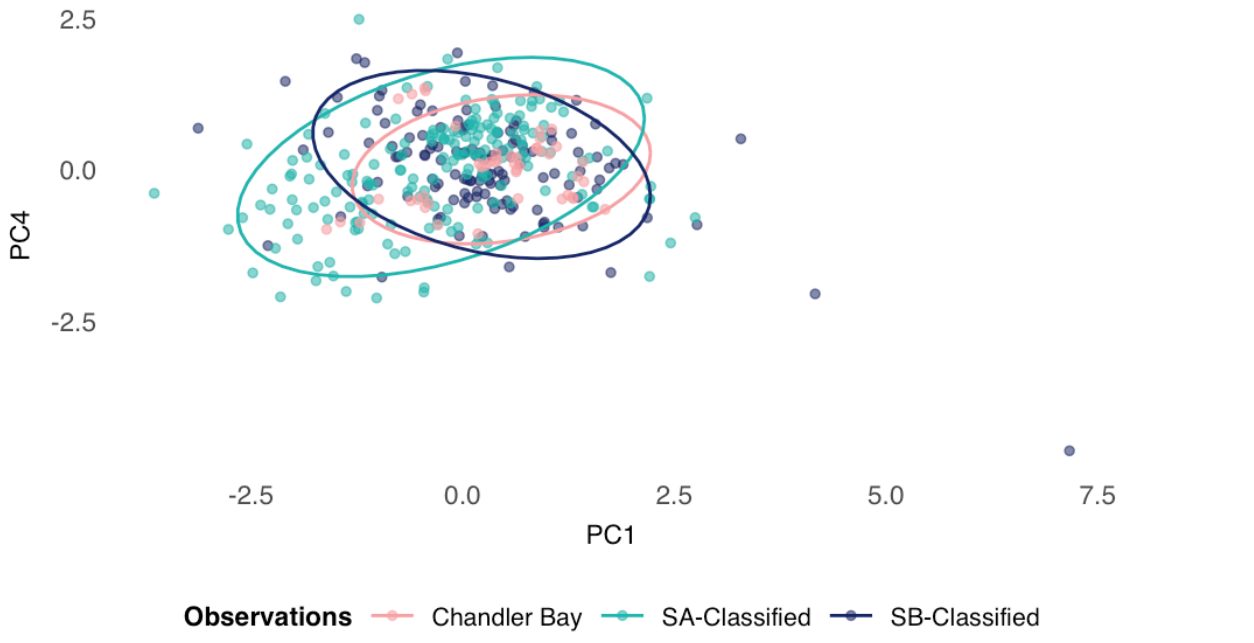
Interpretation: Whichever distance is smaller → closer multivariate similarity (here we found, based on salinity, temperature, and DO that Chandler Bay multivariate structure is closer to SA than SB.)

```r
vars_3 <- c("Temperature_degC", "Salinity", "DO_mgL", "pH", "Ammonia_mgL")

pca_df <- analysis_df |>
  drop_na(all_of(vars_3))

pca_scaled <- scale(pca_df |> select(all_of(vars_3)))

pca_fit <- prcomp(pca_scaled)

scores <- as.data.frame(pca_fit$x[, 1:5]) |>
  dplyr::bind_cols(WaterType = pca_df$WaterType)

ggplot(scores, aes(PC3, PC4, color = WaterType)) +
  geom_point(alpha = 0.6) +
  stat_ellipse(level = 0.95, linewidth = 1) +
  scale_color_manual(values = c(SA = "#1BB6AFFF",
                                SB = "#172869FF",
                                Chandler = "#F6A1A5FF"),
                     labels = c(SA = "SA-Classified",
                                SB = "SB-Classified",
                                Chandler = "Chandler Bay"),
                     name = "Observations") +
  hrbrthemes::theme_ipsum(base_family = "Helvetica", base_size = 17, grid = "N",
                          axis_title_size = 17, axis_title_just = "cc", axis_title_fa
ce = "plain",
                          plot_title_family = "Helvetica", plot_title_size = 22,
                          subtitle_size = 16, subtitle_family = "Helvetica",
                          plot_margin = margin(5,5,5,5)) +
  theme(panel.grid = element_blank(),
        legend.position = "bottom",
        legend.key.size = unit(1.2, "cm"),
        legend.text = element_text(size = 16, color = "black"),
        legend.title = element_text(face = "bold",size = 17),
        axis.title.x = element_text(margin = margin(t = 10, r = 0, b = 0, l = 0)),
        axis.title.y = element_text(margin = margin(t = 0, r = 10, b = 0, l = 0)))
```




```
sa_mat <- pca_scaled[pca_df$WaterType == "SA", ]
sb_mat <- pca_scaled[pca_df$WaterType == "SB", ]
ch_mat <- pca_scaled[pca_df$WaterType == "Chandler", ]

mean(mahalanobis(ch_mat, colMeans(sa_mat), cov(sa_mat)))

## [1] 3.221322

mean(mahalanobis(ch_mat, colMeans(sb_mat), cov(sb_mat)))

## [1] 12.86867

summary(pca_fit)

## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5
## Standard deviation     1.3855 1.0696 0.9774 0.8873 0.44027
## Proportion of Variance 0.3839 0.2288 0.1910 0.1575 0.03877
## Cumulative Proportion  0.3839 0.6127 0.8038 0.9612 1.00000

pca_fit$rotation

##                          PC1        PC2        PC3         PC4        PC5
## Temperature_degC  0.64872446  0.1806380  0.1114430  0.25769079 -0.6838877
## Salinity         -0.36592218 -0.4927463 -0.1546431  0.74135537 -0.2231138
## DO_mgL           -0.62855325  0.1912776  0.1543883 -0.34690844 -0.6512699
## pH               -0.21327465  0.6774216  0.4285888  0.50571581  0.2370171
## Ammonia_mgL      -0.06850015  0.4786271 -0.8695642  0.08884667 -0.0467787

vars_3 <- c("Salinity", "DO_mgL", "pH", "Ammonia_mgL")

pca_df <- analysis_df |>
  drop_na(all_of(vars_3))
```

```r
pca_scaled <- scale(pca_df |> select(all_of(vars_3)))

pca_fit <- prcomp(pca_scaled)

scores <- as.data.frame(pca_fit$x[, c(1,4)]) |>
  dplyr::bind_cols(WaterType = pca_df$WaterType)

ggplot(scores, aes(PC1, PC4, color = WaterType)) +
  geom_point(alpha = 0.6) +
  stat_ellipse(level = 0.95, linewidth = 1) +
  scale_color_manual(values = c(SA = "#1BB6AFFF",
                                SB = "#172869FF",
                                Chandler = "#F6A1A5FF"),
                     labels = c(SA = "SA-Classified",
                                SB = "SB-Classified",
                                Chandler = "Chandler Bay"),
                     name = "Observations") +
  hrbrthemes::theme_ipsum(base_family = "Helvetica", base_size = 17, grid = "N",
                          axis_title_size = 17, axis_title_just = "cc", axis_title_fa
ce = "plain",
                          plot_title_family = "Helvetica", plot_title_size = 22,
                          subtitle_size = 16, subtitle_family = "Helvetica",
                          plot_margin = margin(5,5,5,5)) +
  theme(panel.grid = element_blank(),
        legend.position = "bottom",
        legend.key.size = unit(1.2, "cm"),
        legend.text = element_text(size = 16, color = "black"),
        legend.title = element_text(face = "bold",size = 17),
        axis.title.x = element_text(margin = margin(t = 10, r = 0, b = 0, l = 0)),
        axis.title.y = element_text(margin = margin(t = 0, r = 10, b = 0, l = 0)))
```



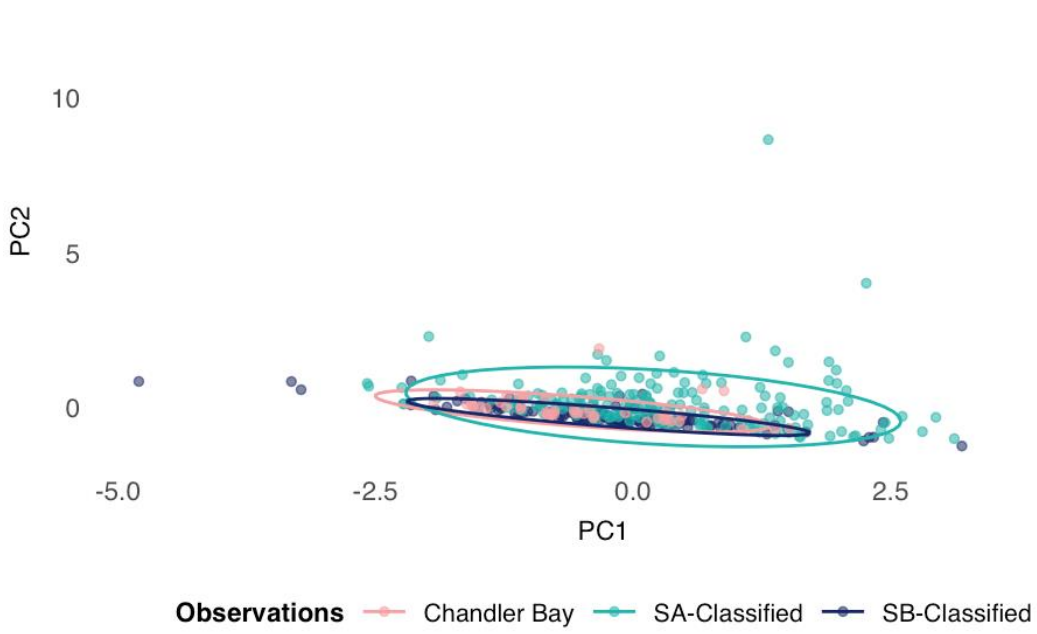

```r
sa_mat <- pca_scaled[pca_df$WaterType == "SA", ]
sb_mat <- pca_scaled[pca_df$WaterType == "SB", ]
```

```r
ch_mat <- pca_scaled[pca_df$WaterType == "Chandler", ]

mean(mahalanobis(ch_mat, colMeans(sa_mat), cov(sa_mat)))

## [1] 2.424875

mean(mahalanobis(ch_mat, colMeans(sb_mat), cov(sb_mat)))

## [1] 11.68563

summary(pca_fit)

## Importance of components:
##                           PC1    PC2    PC3    PC4
## Standard deviation     1.1577 1.0310 0.9650 0.8158
## Proportion of Variance 0.3351 0.2657 0.2328 0.1664
## Cumulative Proportion  0.3351 0.6008 0.8336 1.0000

pca_fit$rotation

##                    PC1        PC2         PC3         PC4
## Salinity    -0.2349991 -0.7320455  0.56300057  0.30317518
## DO_mgL      -0.6933998 -0.1394682 -0.09343323 -0.70072508
## pH          -0.6506415  0.2177581 -0.33592757  0.64529038
## Ammonia_mgL -0.2015881  0.6302694  0.74930184 -0.02587473
```

-

```r
vars_3 <- c("DO_mgL", "pH", "Ammonia_mgL")

pca_df <- analysis_df |>
  drop_na(all_of(vars_3))

pca_scaled <- scale(pca_df |> select(all_of(vars_3)))

pca_fit <- prcomp(pca_scaled)

scores <- as.data.frame(pca_fit$x[, 1:2]) |>
  dplyr::bind_cols(WaterType = pca_df$WaterType)

ggplot(scores, aes(PC1, PC2, color = WaterType)) +
  geom_point(alpha = 0.6) +
  stat_ellipse(level = 0.95, linewidth = 1) +
  scale_color_manual(values = c(SA = "#1BB6AFFF",
                                SB = "#172869FF",
                                Chandler = "#F6A1A5FF"),
                     labels = c(SA = "SA-Classified",
                                SB = "SB-Classified",
                                Chandler = "Chandler Bay"),
                     name = "Observations") +
  hrbrthemes::theme_ipsum(base_family = "Helvetica", base_size = 17, grid = "N",
                          axis_title_size = 17, axis_title_just = "cc", axis_title_fa
ce = "plain",
                          plot_title_family = "Helvetica", plot_title_size = 22,
```

```
                        subtitle_size = 16, subtitle_family = "Helvetica",
                        plot_margin = margin(5,5,5,5)) +
  theme(panel.grid = element_blank(),
        legend.position = "bottom",
        legend.key.size = unit(1.2, "cm"),
        legend.text = element_text(size = 16, color = "black"),
        legend.title = element_text(face = "bold",size = 17),
        axis.title.x = element_text(margin = margin(t = 10, r = 0, b = 0, l = 0)),
        axis.title.y = element_text(margin = margin(t = 0, r = 10, b = 0, l = 0)))
```



- What the figure shows:
  - SA forms a broad, elongated cloud
  - SB is more compact and shifted left along PC1
  - Chandler sits almost entirely inside the SA ellipse
  - Very little Chandler overlap with SB ellipse
- So visually: Chandler's water chemistry structure overlaps strongly with SA and not SB.
  - This supports the RF result!!

```
sa_mat <- pca_scaled[pca_df$WaterType == "SA", ]
sb_mat <- pca_scaled[pca_df$WaterType == "SB", ]
ch_mat <- pca_scaled[pca_df$WaterType == "Chandler", ]

mean(mahalanobis(ch_mat, colMeans(sa_mat), cov(sa_mat)))

## [1] 1.738752

mean(mahalanobis(ch_mat, colMeans(sb_mat), cov(sb_mat)))

## [1] 10.52233
```

- Here, we computed the mean Mahalanobis distance of Chandler points to:
    - SA centroid
    - SB centroid
- Mahalanobis distance measures: How far a point is from a group center, accounting for covariance structure.
    - So it considers:
        - Scale
        - Correlation between variables
        - Shape of the SA/SB cloud
- Chandler observations are:
    - Very close to SA centroid
    - Extremely far from SB centroid
    - And not just slightly — 10.5 vs 1.7 is a dramatic difference.
    - That's about 10.5/1.7 = 6x farther from SB than SA
        - That is a strong multivariate separation

Interpretation: Whichever distance is smaller → closer multivariate similarity (here we found, based on DO, pH, and Ammonia that Chandler Bay multivariate structure is closer to SA than SB.)

What This Means Scientifically

- Based on DO, pH, and Ammonia:
    - Chandler water chemistry structure aligns strongly with SA regime
    - It is not intermediate
    - It is not SB-like
    - It is well within SA multivariate space
    - This is independent confirmation of your RF probability plot.

```
summary(pca_fit)

## Importance of components:
##                           PC1    PC2    PC3
## Standard deviation     1.1492 0.9909 0.8352
## Proportion of Variance 0.4402 0.3273 0.2325
## Cumulative Proportion  0.4402 0.7675 1.0000

pca_fit$rotation

##                    PC1        PC2         PC3
## DO_mgL      0.6789301 -0.2315301  0.69674081
## pH          0.6924849 -0.1133956 -0.71246483
## Ammonia_mgL 0.2439644  0.9661963  0.08334337
```

- Variance Explained:
    - PC1 = 44.0%
    - PC2 = 32.7%
    - PC3 = 23.3%

- So:
    - PC1 + PC2 = 76.8% of total variance
    - Your 2D plot captures most of the chemistry structure
    - Interpretation based on PC1/PC2 is valid
- Interpreting the Loadings (code: pca_fit$rotation)(the more important part):
- PC1:
    - PC1 is driven primarily by:
        - DO (strong positive)
        - pH (strong positive)
        - Minor contribution from ammonia
    - PC1 ≈ Oxygen–pH gradient
    - High PC1 values =
        - High DO
        - High pH
        - Slightly higher ammonia
    - Low PC1 values =
        - Lower oxygen
        - Lower pH
    - So PC1 represents overall “oxygenated water quality”.
- PC2 is almost entirely Ammonia
    - PC2 ≈ Ammonia gradient
    - High PC2 values = high ammonia
    - Low PC2 values = low ammonia
    - DO and pH barely matter here
- Since PC1 = oxygen/pH axis and PC2 = ammonia axis, the PCA plot is essentially:
    - Horizontal axis = oxygenated vs less oxygenated
    - Vertical axis = ammonia concentration
- What Your Plot + Mahalanobis Now Mean
    - Given the loadings, Chandler matches SA in terms of
        - oxygen levels
        - pH structure
        - ammonia concentrations
    - And differs strongly from SB across that same structure
- **Ecological Interpretation**: Chandler is not just “close to SA” in some abstract PCA space.
- Chandler Bay and historical SA sites are specifically similar in:
    - Oxygenation regime
    - Acid–base chemistry
    - Nutrient (ammonia) profile
- Because PC2 = ammonia (loading = 0.966), if Chandler overlapped SA strongly along PC2, it means that ammonia behavior in Chandler Bay is much more SA-like than SB-like
    - This could be due:
        - circulation-driven
        - flushing rate-driven

        - biological processing differences
- Notice that the SA ellipse was larger in our plot. This means:
    - SA chemistry is more variable
    - Chandler fits within that natural SA variability
    - So Chandler is not “special SA” - it is inside normal SA structure
- Based on DO, pH, and ammonia: Chandler Bay is chemically SA-like and not intermediate and the structure driving that similarity is oxygen regime, acid-base chemistry, and ammonia levels

MANOVA

```
perm_df <- analysis_df |>
  drop_na(all_of(vars_3)) |>
  select(WaterType, all_of(vars_3))

perm_df <- perm_df |>
  mutate(
    Ammonia_mgL = log1p(Ammonia_mgL)
  )

manova_fit <- manova(
  cbind(DO_mgL, pH, Ammonia_mgL) ~ WaterType,
  data = perm_df
)

summary(manova_fit, test = "Pillai")

##            Df  Pillai approx F num Df den Df    Pr(>F)
## WaterType   2 0.17533   11.402      6    712 3.402e-12 ***
## Residuals 357
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary.aov(manova_fit)

##  Response DO_mgL :
##              Df Sum Sq Mean Sq F value   Pr(>F)
## WaterType     2  12.82  6.4125  5.9029 0.003005 **
## Residuals   357 387.82  1.0863
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response pH :
##              Df Sum Sq  Mean Sq F value    Pr(>F)
## WaterType     2 0.4346 0.217281  8.9885 0.0001554 ***
## Residuals   357 8.6299 0.024173
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Ammonia_mgL :
##              Df  Sum Sq  Mean Sq F value    Pr(>F)
## WaterType     2 0.15416 0.077078  23.444 2.714e-10 ***
## Residuals   357 1.17371 0.003288
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hotelling's T² (Direct Pairwise Multivariate Test)

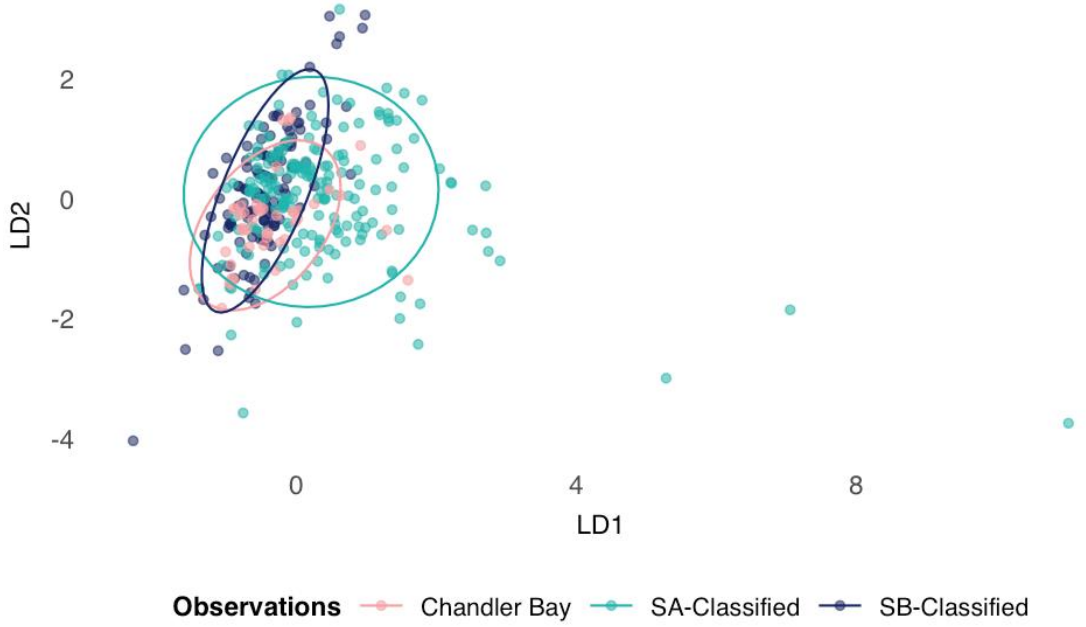
```
library(Hotelling)

sa_ch_df <- perm_df |>
  filter(WaterType %in% c("SA", "Chandler"))

res <- hotelling.test(
  sa_ch_df |> filter(WaterType == "SA") |> select(DO_mgL, pH, Ammonia_mgL),
  sa_ch_df |> filter(WaterType == "Chandler") |> select(DO_mgL, pH, Ammonia_mgL))

res

## Test stat:  26.545
## Numerator df:  3
## Denominator df:  223
## P-value:  1.601e-05
```

This directly answers: Are the multivariate means different?

Linear Discriminant Analysis

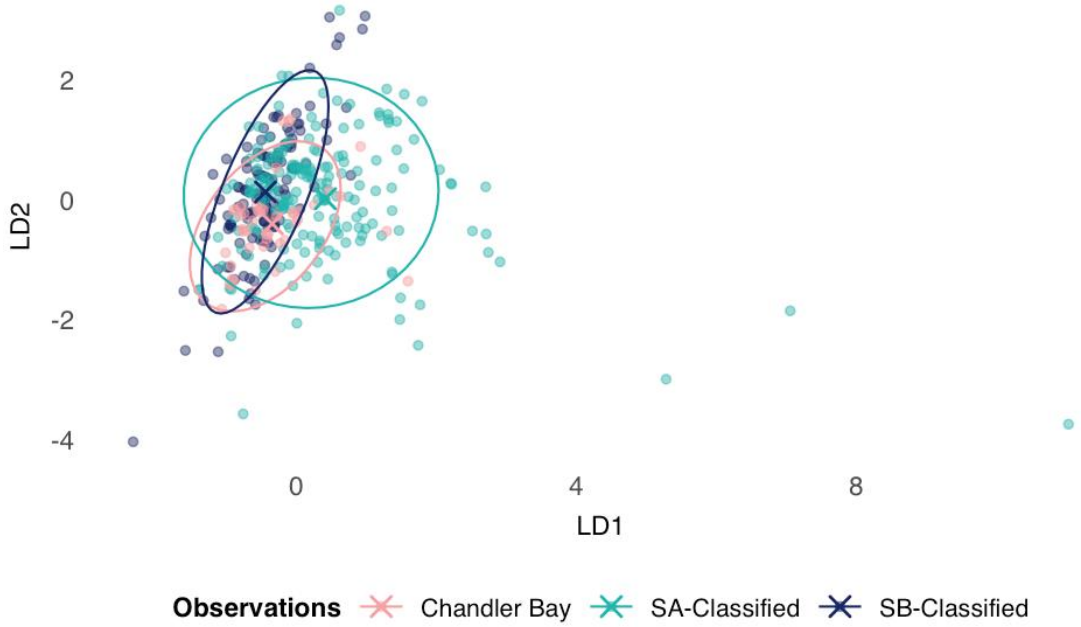
```
library(MASS)

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

lda_fit <- lda(
  WaterType ~ DO_mgL + pH + Ammonia_mgL,
  data = perm_df
)

lda_pred <- predict(lda_fit)

lda_df <- as.data.frame(lda_pred$x) |>
  dplyr::bind_cols(WaterType = perm_df$WaterType)

ggplot(lda_df, aes(LD1, LD2, color = WaterType)) +
  geom_point(alpha = 0.6) +
  stat_ellipse(level = 0.95) +
  scale_color_manual(values = c(SA = "#1BB6AFFF",
                                SB = "#172869FF",
                                Chandler = "#F6A1A5FF"),
                     labels = c(SA = "SA-Classified",
                                SB = "SB-Classified",
                                Chandler = "Chandler Bay"),
                     name = "Observations") +
```

```
  hrbrthemes::theme_ipsum(base_family = "Helvetica", base_size = 17, grid = "N",
                          axis_title_size = 17, axis_title_just = "cc", axis_title_fa
ce = "plain",
                          plot_title_family = "Helvetica", plot_title_size = 22,
                          subtitle_size = 16, subtitle_family = "Helvetica",
                          plot_margin = margin(5,5,5,5)) +
  theme(panel.grid = element_blank(),
        legend.position = "bottom",
        legend.key.size = unit(1.2, "cm"),
        legend.text = element_text(size = 16, color = "black"),
        legend.title = element_text(face = "bold",size = 17),
        axis.title.x = element_text(margin = margin(t = 10, r = 0, b = 0, l = 0)),
        axis.title.y = element_text(margin = margin(t = 0, r = 10, b = 0, l = 0)))
```



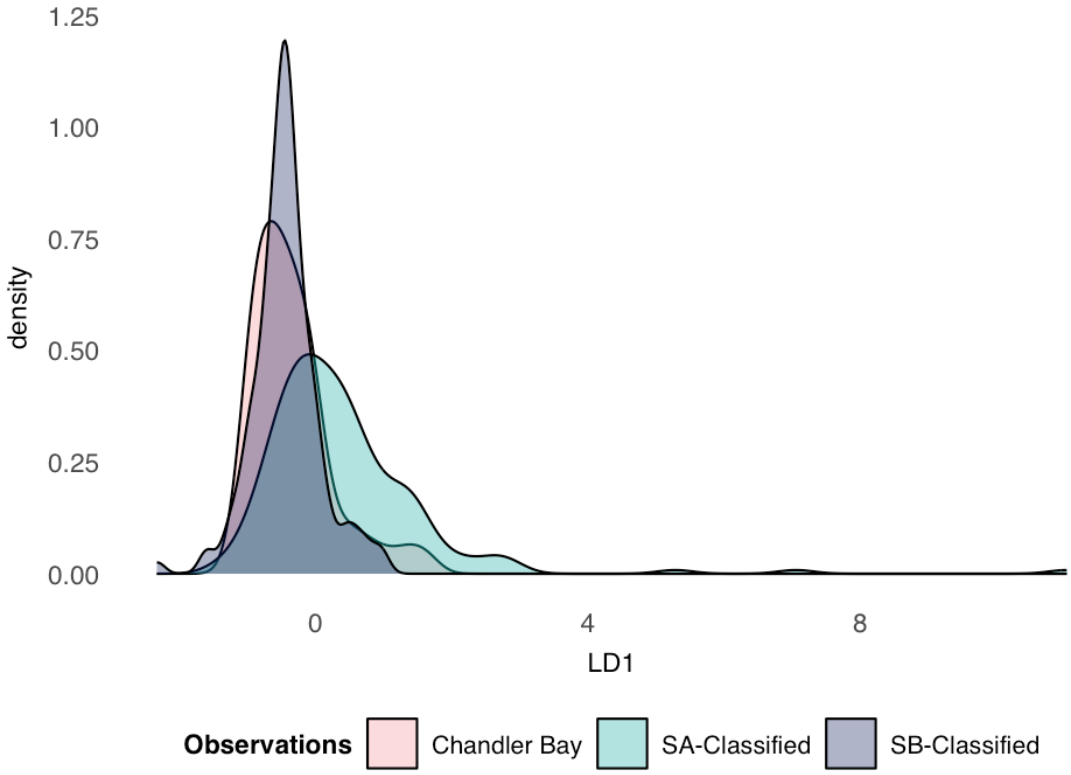

```
library(dplyr)

centroids <- lda_df |>
  dplyr::group_by(WaterType) |>
  dplyr::summarise(
    dplyr::across(
      .cols = dplyr::starts_with("LD"),
      .fns  = mean
    )
  )

ggplot(lda_df, aes(LD1, LD2, color = WaterType)) +
  geom_point(alpha = 0.5) +
  stat_ellipse(level = 0.95) +
  geom_point(data = centroids, size = 5, shape = 4, stroke = 1.5) +
  scale_color_manual(values = c(SA = "#1BB6AFFF",
                                SB = "#172869FF",
                                Chandler = "#F6A1A5FF"),
                     labels = c(SA = "SA-Classified",
```

```
                              SB = "SB-Classified",
                              Chandler = "Chandler Bay"),
                    name = "Observations") +
  hrbrthemes::theme_ipsum(base_family = "Helvetica", base_size = 17, grid = "N",
                          axis_title_size = 17, axis_title_just = "cc", axis_title_fa
ce = "plain",
                          plot_title_family = "Helvetica", plot_title_size = 22,
                          subtitle_size = 16, subtitle_family = "Helvetica",
                          plot_margin = margin(5,5,5,5)) +
  theme(panel.grid = element_blank(),
        legend.position = "bottom",
        legend.key.size = unit(1.2, "cm"),
        legend.text = element_text(size = 16, color = "black"),
        legend.title = element_text(face = "bold",size = 17),
        axis.title.x = element_text(margin = margin(t = 10, r = 0, b = 0, l = 0)),
        axis.title.y = element_text(margin = margin(t = 0, r = 10, b = 0, l = 0)))
```




```
ggplot(lda_df, aes(LD1, fill = WaterType)) +
  geom_density(alpha = 0.4) +
  scale_fill_manual(values = c(SA = "#1BB6AFFF",
                              SB = "#172869FF",
                              Chandler = "#F6A1A5FF"),
                    labels = c(SA = "SA-Classified",
                              SB = "SB-Classified",
                              Chandler = "Chandler Bay"),
                    name = "Observations") +
  hrbrthemes::theme_ipsum(base_family = "Helvetica", base_size = 17, grid = "N",
                          axis_title_size = 17, axis_title_just = "cc", axis_title_fa
ce = "plain",
                          plot_title_family = "Helvetica", plot_title_size = 22,
                          subtitle_size = 16, subtitle_family = "Helvetica",
                          plot_margin = margin(5,5,5,5)) +
  theme(panel.grid = element_blank(),
```

```
        legend.position = "bottom",
        legend.key.size = unit(1.2, "cm"),
        legend.text = element_text(size = 16, color = "black"),
        legend.title = element_text(face = "bold",size = 17),
        axis.title.x = element_text(margin = margin(t = 10, r = 0, b = 0, l = 0)),
        axis.title.y = element_text(margin = margin(t = 0, r = 10, b = 0, l = 0)))
```



Jason Krumholz
Remote Ecologist
LD 2187

I will be briefly summarizing the results of the attached modeling effort, which uses statewide, historical water quality data from waters currently classified "SA" and "SB" to train a model to distinguish between "SA" and "SB" water quality.

Our results show that this approach is appropriate, and capable of discerning between "SA" and "SB" waters in the majority of cases. Although water quality is not the only factor governing "SA" classification, statewide "SA" waters DO have measurably better water quality.

We then used this model to test whether water quality in Chandler Bay is more representative of "SA" or "SB" waters statewide. The model finds that water quality in Chandler Bay strongly resembles the "SA" portion of the dataset, with the median model score for Chandler bay HIGHER than the median model score for currently designated "SA" waters statewide.

Although we recognize that water quality is not the only factor governing the decision to reclassify waters as "SA", we feel that this is strong evidence in favor of Chandler Bay's waters as "of outstanding ecological significance" and worthy of protection.

Therefore, we request that LD2187 be amended to include reclassification of Chandler Bay to "SA" prior to passage.