Classification of Phytoplankton Communities

Goal: investigate the presence/absence and robustness of a relationship between phytoplankton taxa and various environmental predictors

Main conclusions:

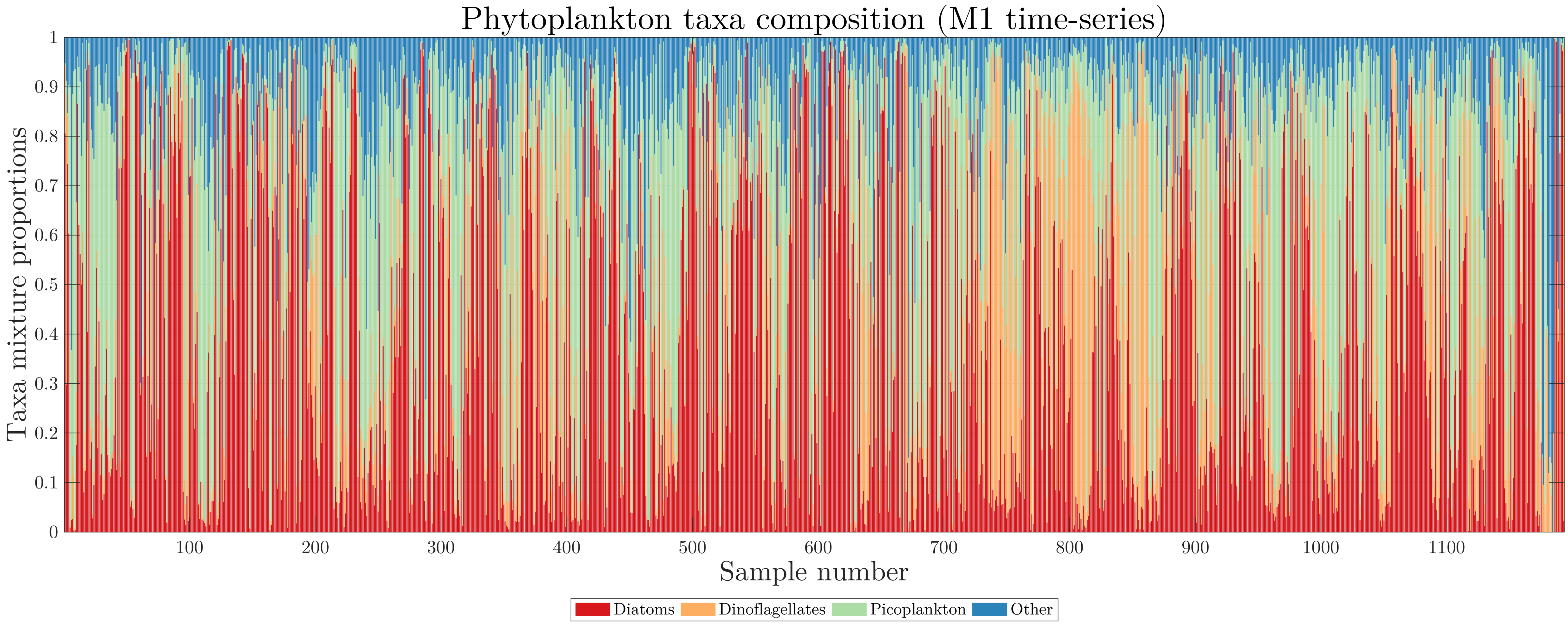
1. In the CTD dataset, phytoplankton taxa classification performance is limited by feature discriminatory power, not by sample size.
2. Explored features are limited in their ability to separate communities dominated by Dinoflagellates.
3. Consistent high-bias model performance across all feature combinations may indicate the presence of noise in the response variable, i.e., the count data from which the community taxa composition is derived.
4. The predictive power of features that do not measure biological properties directly (e.g., time of year, water temp, distance from coast, etc.) suggests that a hierarchical approach that leverages these variables as priors (e.g., in a Bayesian network) may be more appropriate.

Overview of the approach

Classification of phytoplankton communities

1. Use *k*-means clustering on phytoplankton taxa composition data from M1 time-series to index community "types" - these clusters will serve as our classes.
2. Each community is a mixture of Diatoms, Dinoflagellates, Picoplankton and Other.
3. Train a classification model to try and separate the communities: Support Vector Machine (SVM).
4. Perform exhaustive search over our feature set to identify optimal predictors for each class - consider all possible permutations (524,287 models).
5. Compute performance metrics and “learning curve” for best models.

Deriving phytoplankton community types from in-situ count data

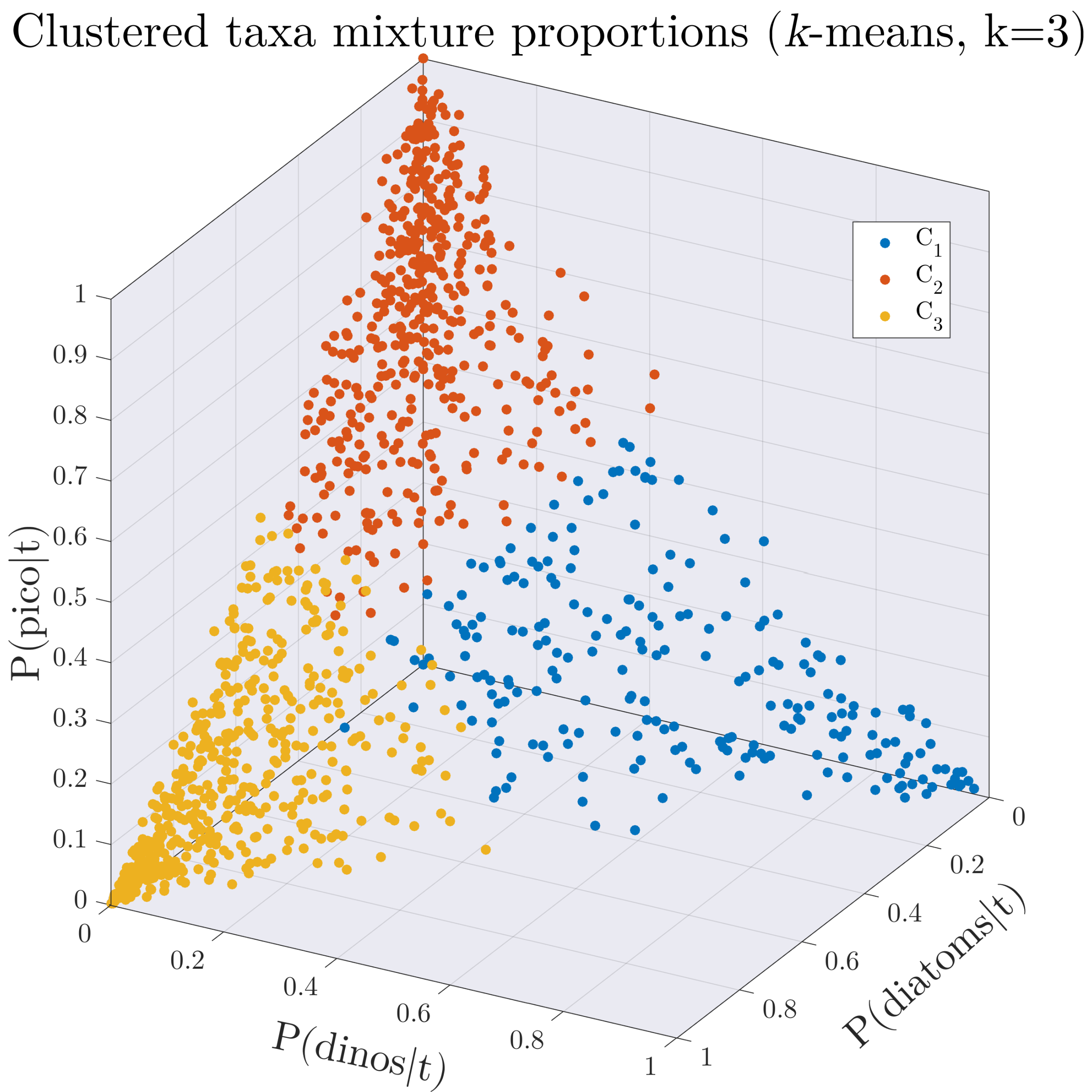


Community types using k-means clustering

k-means clustering:

- Partition n observations into k clusters.
- Each observation belongs to the cluster with the nearest mean (centroid)
- k is user specified.

- Tried k=3-12.

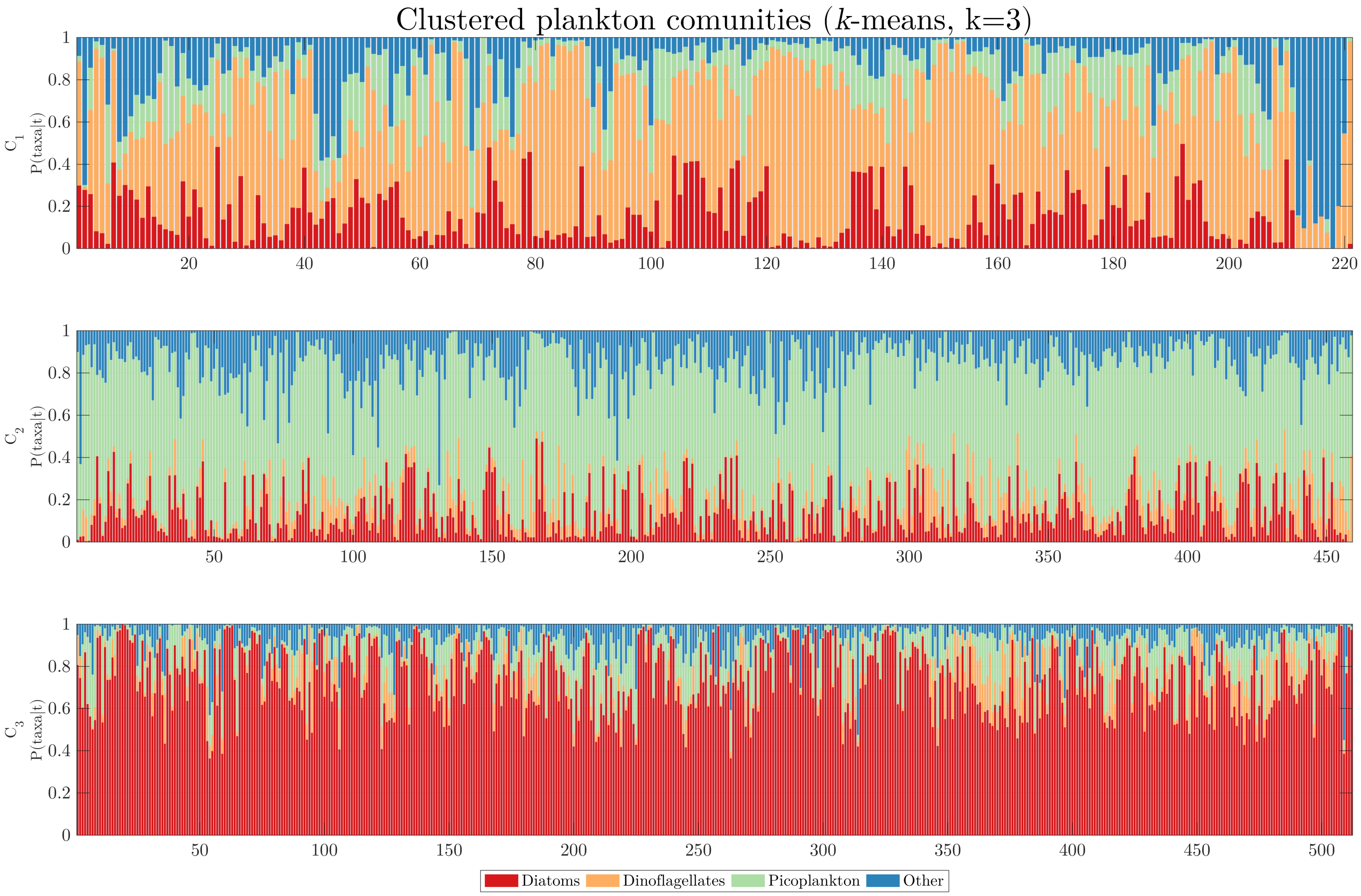


Community types using k-means clustering

Dinoflagellate dominated

Picoplankton dominated

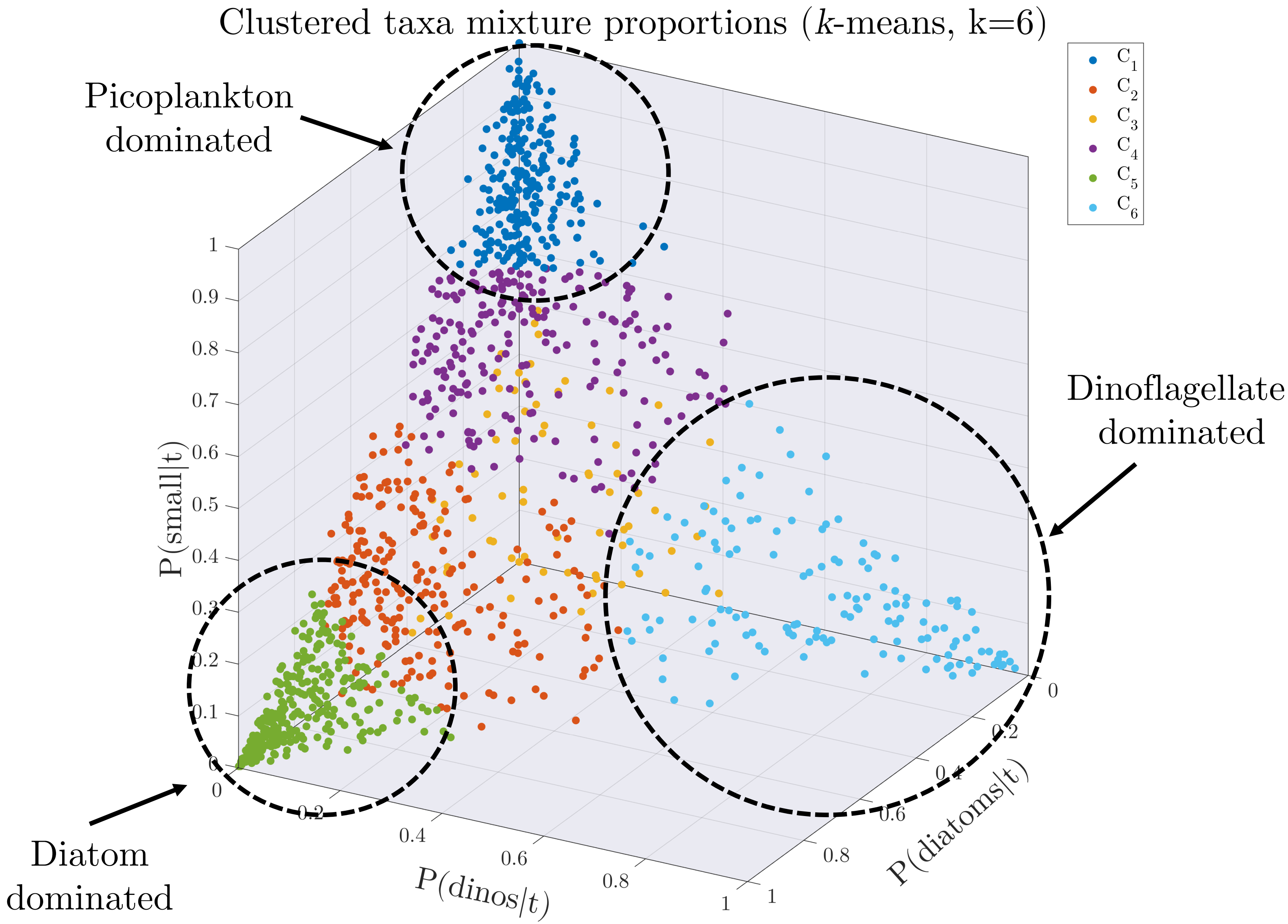
Diatom dominated



Community "types" using k-means clustering

***k*-means clustering:**

- Use larger number of clusters (e.g., k=6) to extract communities that are dominated by a single taxon
- Process only "uni-taxa" communities in classification models
- Dinoflagellate dominated community (C_6) incorporates observations that are more mixed to increase the number of samples in the group

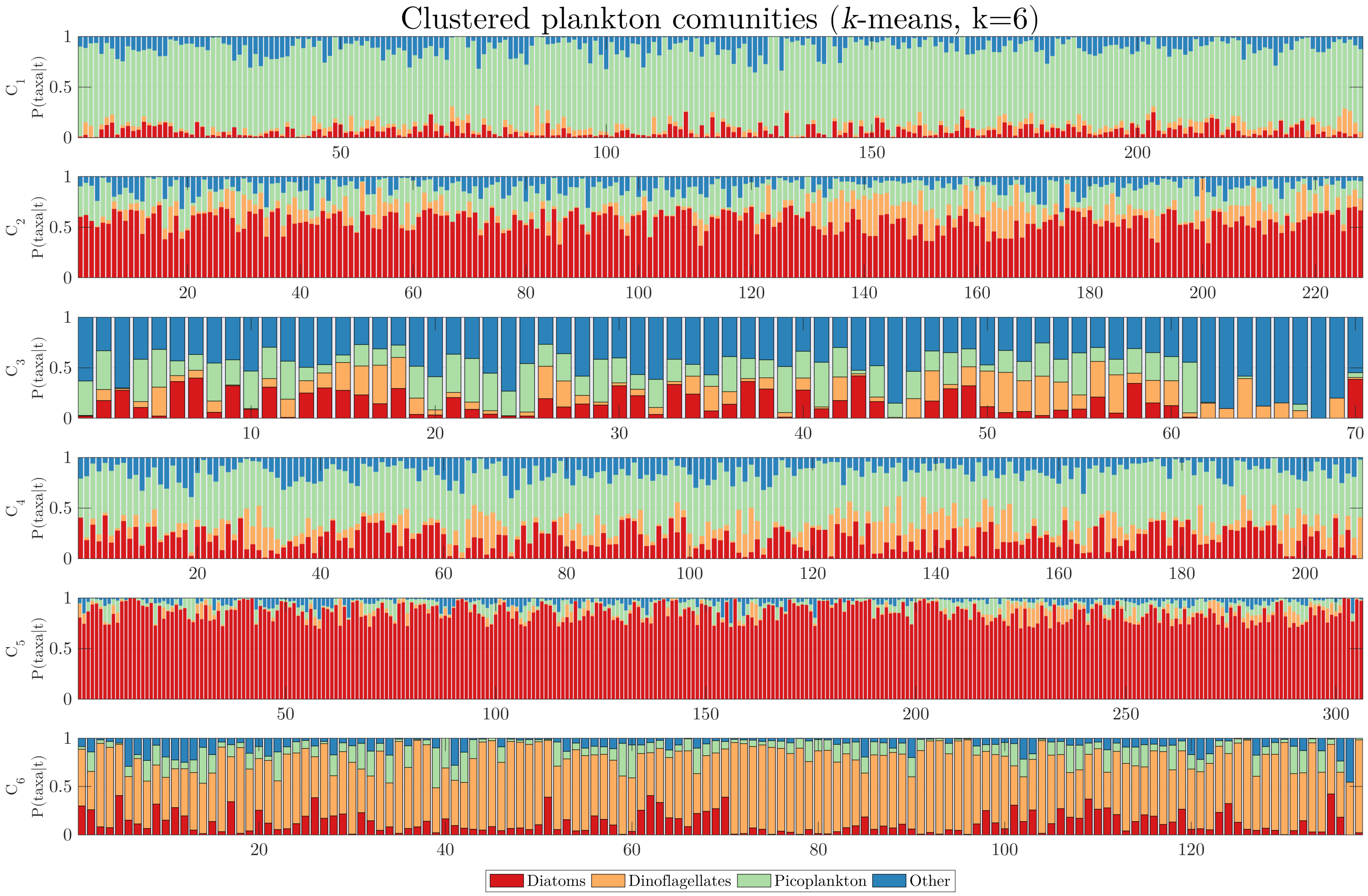


Community "types" using k-means clustering

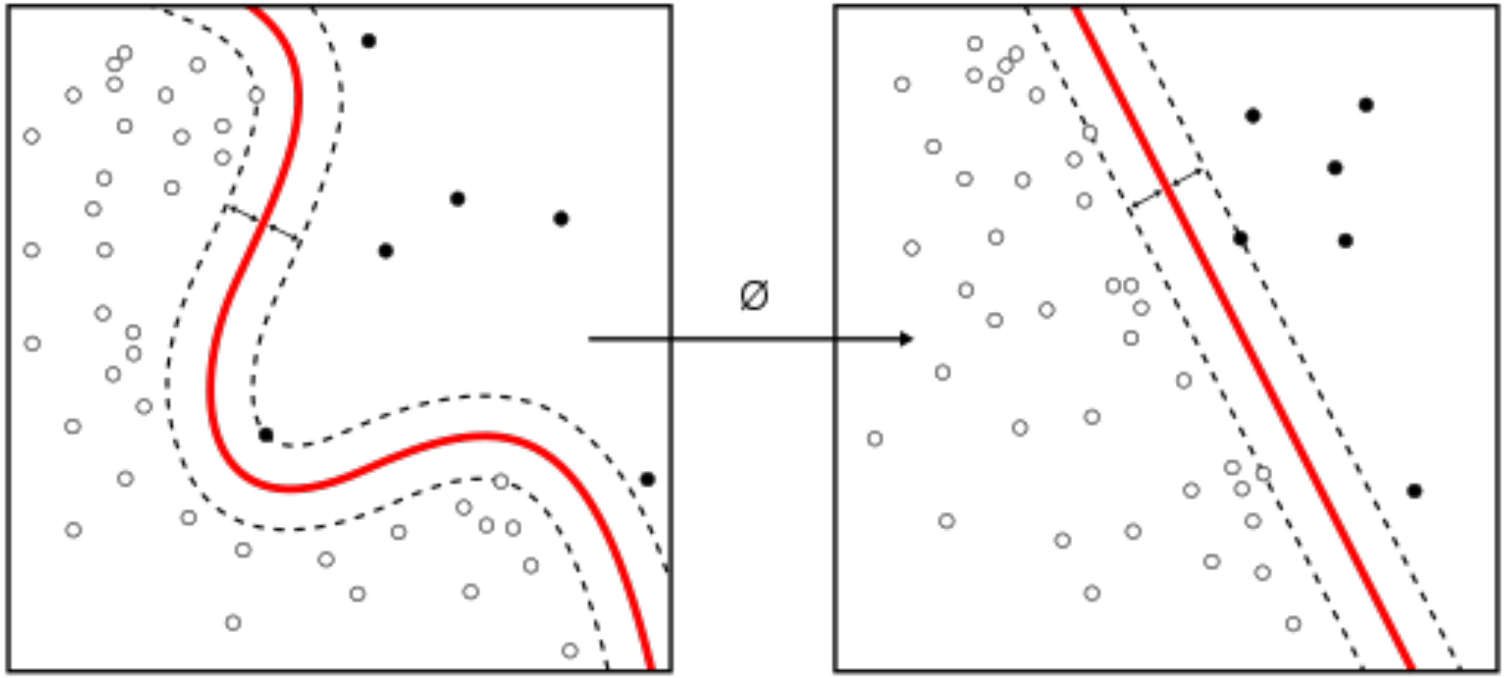
Picoplankton dominated

Diatom dominated

Dinoflagellate dominated



Classification model: Support Vector Machine (SVM)



SVM:

- Binary classifier.
- Find widest separating margin between two classes.
- Find complex margins using transformation functions (kernels)
- "One-vs-all" for multi-class case

- Partition data: K-fold cross-validation (we used k=5)

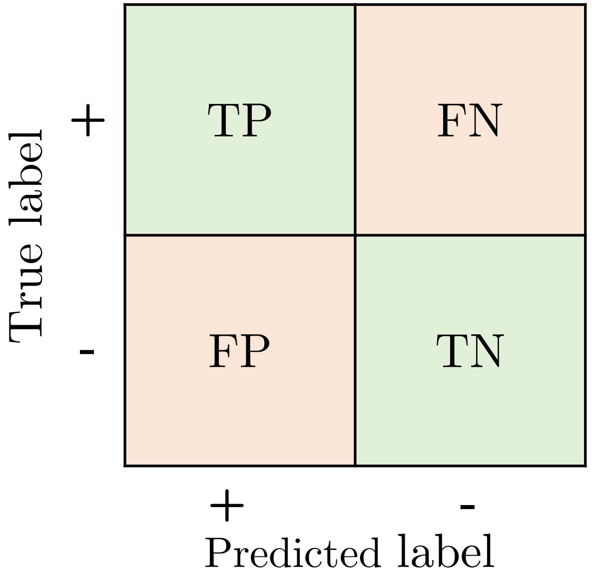
RBF Kernel:

- Allows fitting smooth functions to datasets that aren't linearly separable
- RBF kernel works well in practice, and it is relatively easy to tune
- γ variance hyperparameter – I performed a coarse grid search: $\gamma = 0.05, 0.1, 0.5, 1, 5$
- I used a fixed regularization penalty factor (*"box constraint"*) of $C = 1$. C controls the maximum penalty imposed on margin-violating observations and aids in preventing overfitting (regularization).

RBF (Gaussian) Kernel
$$K(x, x') = \exp(-\gamma\|x - x'\|^2)$$
where, $\gamma = 1/2\sigma$

Evaluation metrics

Confusion matrix



TP: true positive
TN: true negative
FP: false positive
FN: false negative

Model evaluation metrics

- **True positive rate (TPR) or Recall** (Sensitivity, probability of detection) — the number of positive (+) predictions divided by the number of positive class values in the test data. A low recall score may indicate many False Negatives.

$$\mathbf{TPR = TP\ /\ (TP + FN)}$$

- **False positive rate (FPR) or fall-out** (probability of false alarm) — the ratio between the number of negative (-) events wrongly categorized as positive (false positives) and the total number of actual negative events.

$$\mathbf{FPR = FP\ /\ (FP + TN)}$$

- **Precision or Positive predictive value (PPV)** — the ratio of correctly predicted positive observations to the total predicted positive observations. Precision can be thought of as a measure of a classifiers exactness. A low precision score may indicate a large number of False Positives.

$$\mathbf{Precision = TP\ /\ (TP + FP)}$$

Evaluation metrics

Confusion matrix

True label \ Predicted label	+	-
+	TP	FN
-	FP	TN

TP: true positive
TN: true negative
FP: false positive
FN: false negative

Model evaluation metrics

- **Accuracy** (ACC) — the ratio of correctly predicted observation to the total observations. Accuracy is a good measure of model performance, but only for symmetric datasets where values of false positive and false negatives are almost same.

$$\mathbf{ACC = TP{+}TN/TP{+}FP{+}FN{+}TN}$$

- **F1-score** — the harmonic mean of Precision and Recall. This score takes both false positives and false negatives into account. F1 is usually more useful than accuracy, especially if you have an uneven class distribution.

$$\textbf{F1 Score} = \textbf{2*(Recall * Precision) / (Recall + Precision)}$$

- **Matthews Correlation Coefficient** (MCC) — takes into account true and false positives and negatives and is regarded as a balanced measure, which can be used even if the classes are of very different sizes. MCC returns a value between -1 and $+1$. A coefficient of $+1$ represents a perfect prediction, 0 no better than random prediction and -1 indicates total disagreement between prediction and observation.

$$\mathbf{MCC} = \frac{\mathbf{TP{\times}TN - FP{\times}FN}}{\sqrt{\mathbf{(TP{+}FP)(TP{+}FN)(TN{+}FP)(TN{+}FN)}}}$$

CTD dataset, $k$means=6

Picoplankton	= 242 samples
Diatoms	= 306 samples
Dinoflagellates	= 138 samples

Predictors

Name	Description
distcoast	Distance from coast
bathy	Bathymetry (water depth)
hour	Hour of day
timeofyear	Year day
chl	Chlorophyll (bottle data)
no3	Nitrate (bottle data)
allhet	(taxa)
ratiohet	Ratio of heterotrophs to autotrophs biomass
par	Photosynthetically Active Radiation (CTD)
temp	Temperature (CTD)
salt	Salinity (CTD)
strat	Stratification (0-100m average)
strat2	Stratification (5-100m density difference)
MLD	Mixed layer depth (deltaT = 0.2°C relative to 5m)
beamc	Beam c (CTD)
fluo	Fluorescence (CTD)
slope	
maxfluo	Maximum fluorescence
beamc_maxfluo	Beam-c at the maximum fluorescence

1. Consider all possible feature permutations: 524,287 models

2. Compute performance metrics for each model on a per class basis (e.g., what combination of features best predicts diatoms, pico, dino?)

3. Sort list by per class classification performance (MCC)

SVM model selection ($k$means=6)

SVM model selection

Top 10 models sorted by MCC score (averaged over all 3 classes):

Model name		MCC	F1-score	ACC	TPR	FPR	Precision	Recall	gamma
1	timeofyear+chl+allhet+ratiohet+temp+maxfluo	0.6622	0.7613	0.8620	0.7466	0.1137	0.8013	0.7466	1
2	timeofyear+chl+allhet+temp+maxfluo	0.6561	0.7578	0.8601	0.7452	0.1141	0.7918	0.7452	1
3	distcoast+timeofyear+chl+allhet+ratiohet+maxfluo	0.6449	0.7481	0.8542	0.7334	0.1205	0.7956	0.7334	1
4	timeofyear+chl+no3+allhet+temp+maxfluo	0.6378	0.7444	0.8513	0.7280	0.1239	0.7902	0.7280	1
5	bathy+timeofyear+chl+allhet+maxfluo	0.6377	0.7465	0.8523	0.7343	0.1205	0.7786	0.7343	1
6	timeofyear+chl+allhet+beamc+maxfluo	0.6348	0.7346	0.8513	0.7183	0.1249	0.7980	0.7183	1
7	timeofyear+chl+allhet+maxfluo	0.6347	0.7424	0.8513	0.7298	0.1218	0.7801	0.7298	1
8	bathy+timeofyear+chl+allhet+ratiohet+temp+maxfluo	0.6344	0.7452	0.8494	0.7312	0.1236	0.7818	0.7312	1
9	distcoast+timeofyear+chl+allhet+beamc+maxfluo	0.6307	0.7352	0.8484	0.7193	0.1260	0.7928	0.7193	1
10	timeofyear+chl+allhet+ratiohet+maxfluo	0.6307	0.7382	0.8494	0.7253	0.1238	0.7815	0.7253	1

Dominating features are: timeofyear, chl, and allhet

SVM model selection

Top 10 models sorted by MCC score (averaged over all 3 classes):

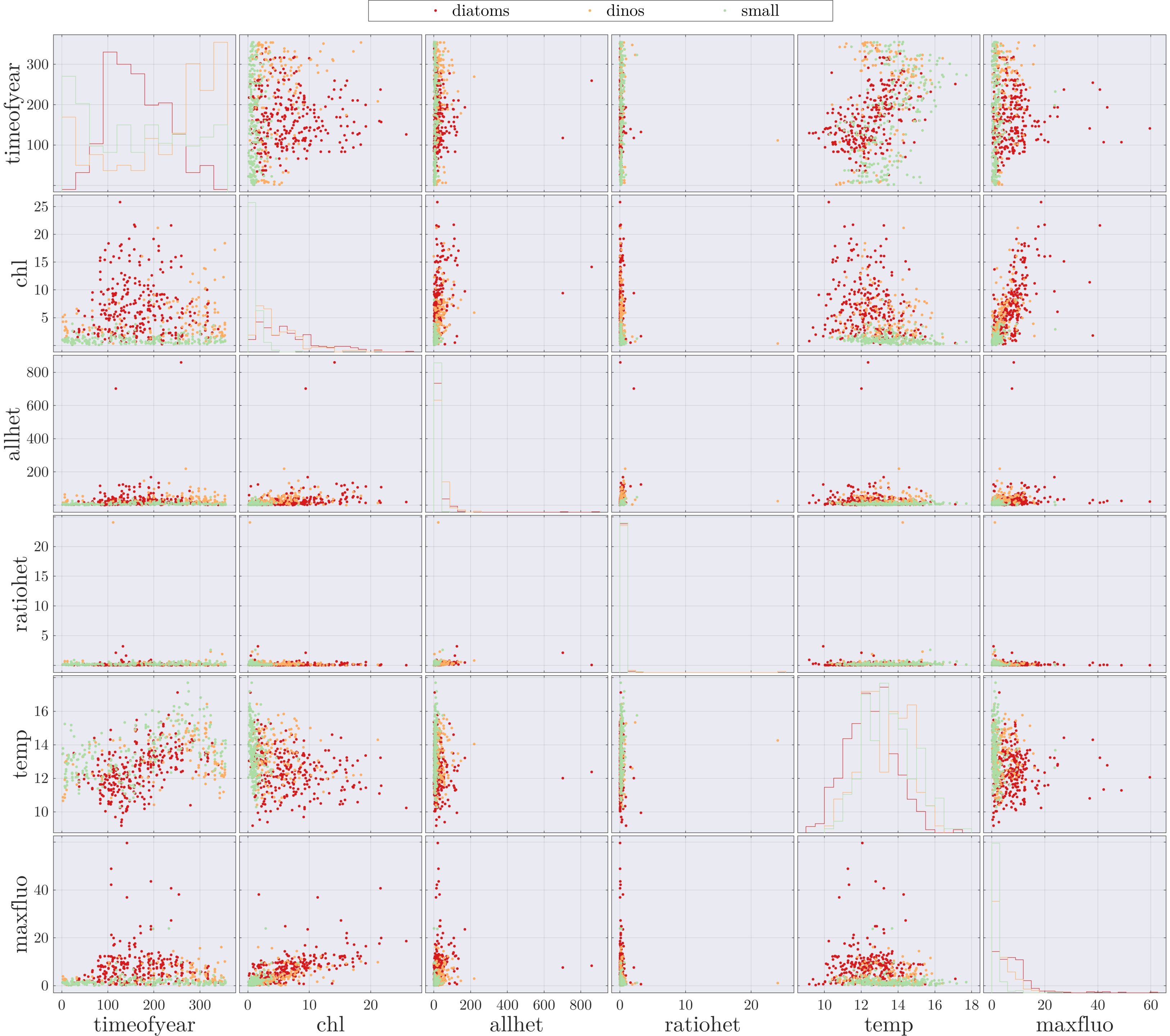
Model name		MCC	F1-score	ACC	TPR	FPR	Precision	Recall	gamma
1	bathy+timeofyear+temp+salt+beamc+beamc_maxfluo	0.5840	0.7061	0.8309	0.6933	0.1389	0.7480	0.6933	1
2	distcoast+timeofyear+temp+beamc+fluo+maxfluo+beamc_maxfluo	0.5815	0.6997	0.8319	0.6896	0.1377	0.7465	0.6896	1
3	bathy+timeofyear+temp+beamc+beamc_maxfluo	0.5810	0.7009	0.8309	0.6894	0.1387	0.7468	0.6894	1
4	timeofyear+temp+beamc+beamc_maxfluo	0.5735	0.6968	0.8280	0.6858	0.1408	0.7389	0.6858	1
5	distcoast+timeofyear+temp+salt+strat+beamc+beamc_maxfluo	0.5677	0.6918	0.8231	0.6773	0.1474	0.7503	0.6773	1
6	distcoast+timeofyear+temp+salt+beamc+maxfluo+beamc_maxfluo	0.5671	0.6815	0.8241	0.6692	0.1476	0.7627	0.6692	1
7	timeofyear+temp+salt+beamc+beamc_maxfluo	0.5652	0.6860	0.8270	0.6765	0.1425	0.7339	0.6765	1
8	distcoast+timeofyear+temp+beamc+maxfluo+beamc_maxfluo	0.5647	0.6816	0.8251	0.6726	0.1445	0.7473	0.6726	1
9	distcoast+timeofyear+temp+beamc+fluo+beamc_maxfluo	0.5643	0.6835	0.8251	0.6736	0.1442	0.7425	0.6736	1
10	distcoast+timeofyear+temp+beamc+beamc_maxfluo	0.5640	0.6853	0.8251	0.6760	0.1437	0.7380	0.6760	1

Features from autonomous sensors

Analysis of best SVM model (optimized hyperparameters)

Model inputs
timeofyear + chl + allhet + ratiohet + temp + maxfluo

Feature space



SVM performance

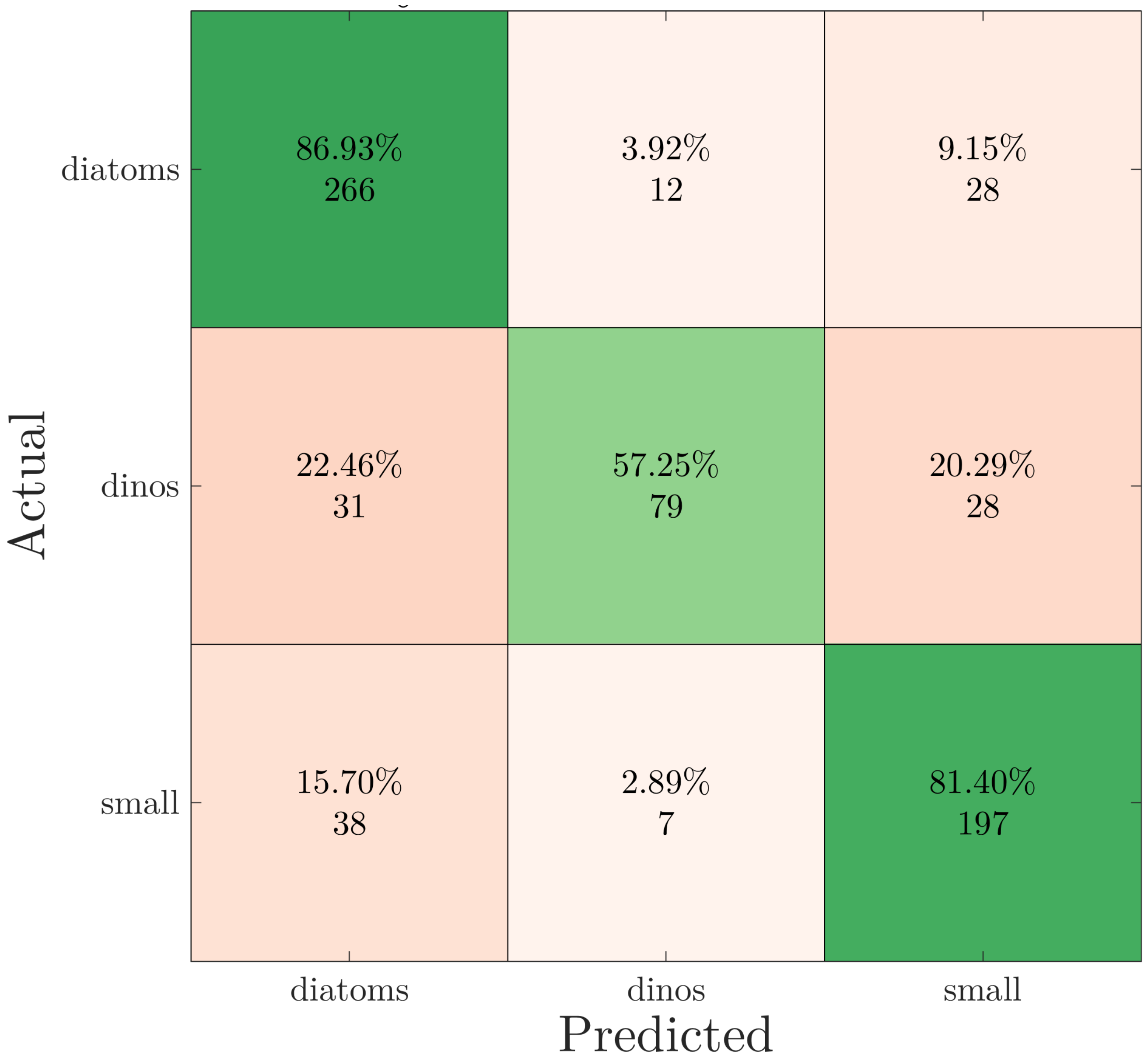
Model inputs
timeofyear + chl + allhet + ratiohet + temp + maxfluo

Results after hyperparameter optimization:

Best observed feasible point:

Coding	BoxConstraint	KernelScale
onevsone	**2.66**	**1.7602**

Overall loss = 0. 19825
Evaluation time = 0.23475 sec

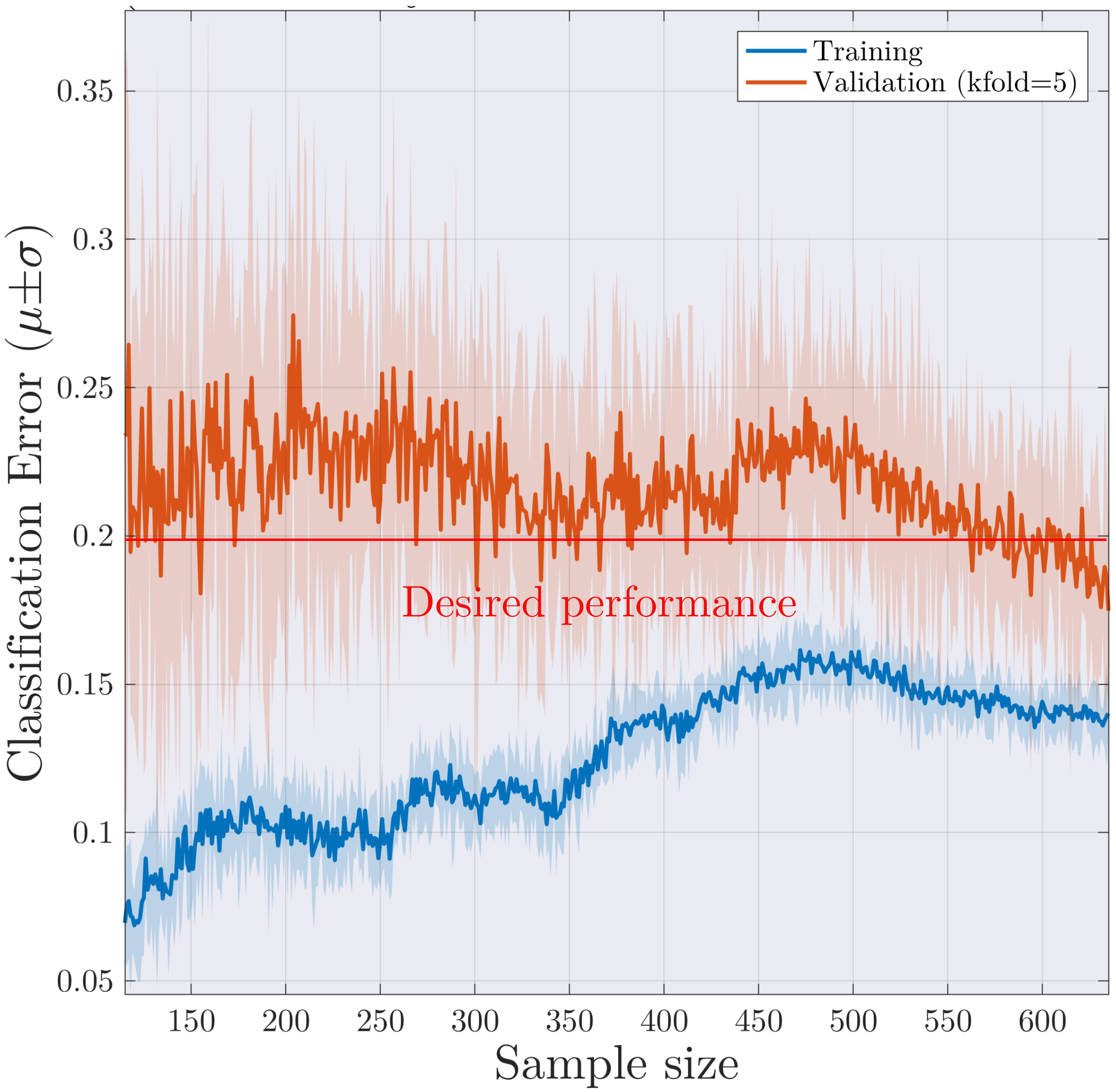


SVM model performance: learning curve

Learning curve:
(avg. classification error vs. sample size)

- High bias learning curve:
  - Training error is still relatively high (~15%).
  - A small gap (<5%) between error in training and validation sets.

- These results suggest:
  - Limited predictive power using current features, i.e., the classes aren’t well separated in the feature space.
  - Increasing the size of the training set will likely have minor impact on classification performance.
  - May indicate issue with class labeling, i.e., the response variable (count dataset in our case)



Analysis of 2D SVM model (optimized hyperparameters)

Model inputs
maxfluo+beamc_maxfluo

SVM performance

Model inputs
maxfluo+**beamc_maxfluo**

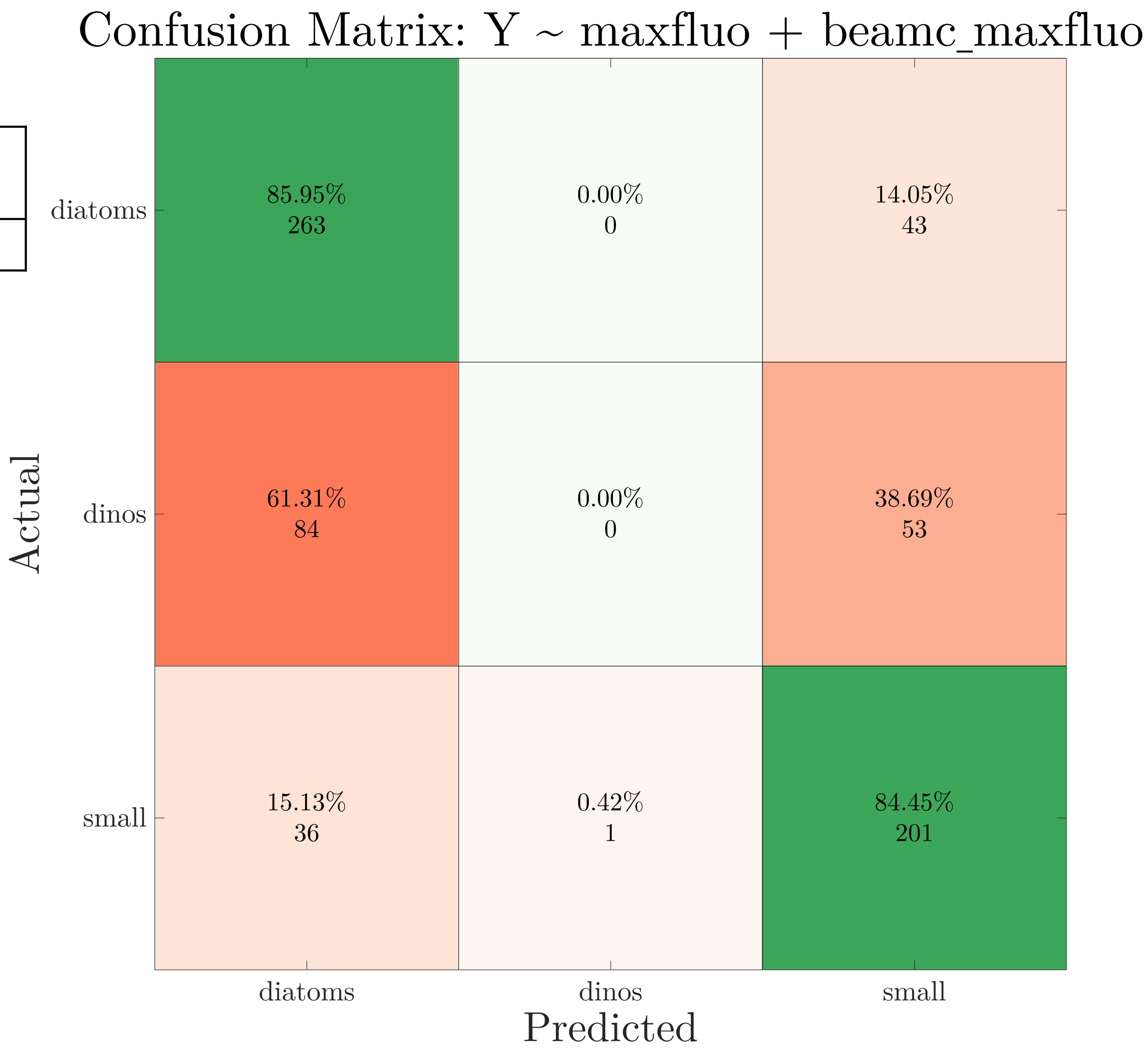
Results after hyperparameter optimization:

Best observed feasible point:

Coding	BoxConstraint	KernelScale
onevsone	**3.0918**	**0.46196**

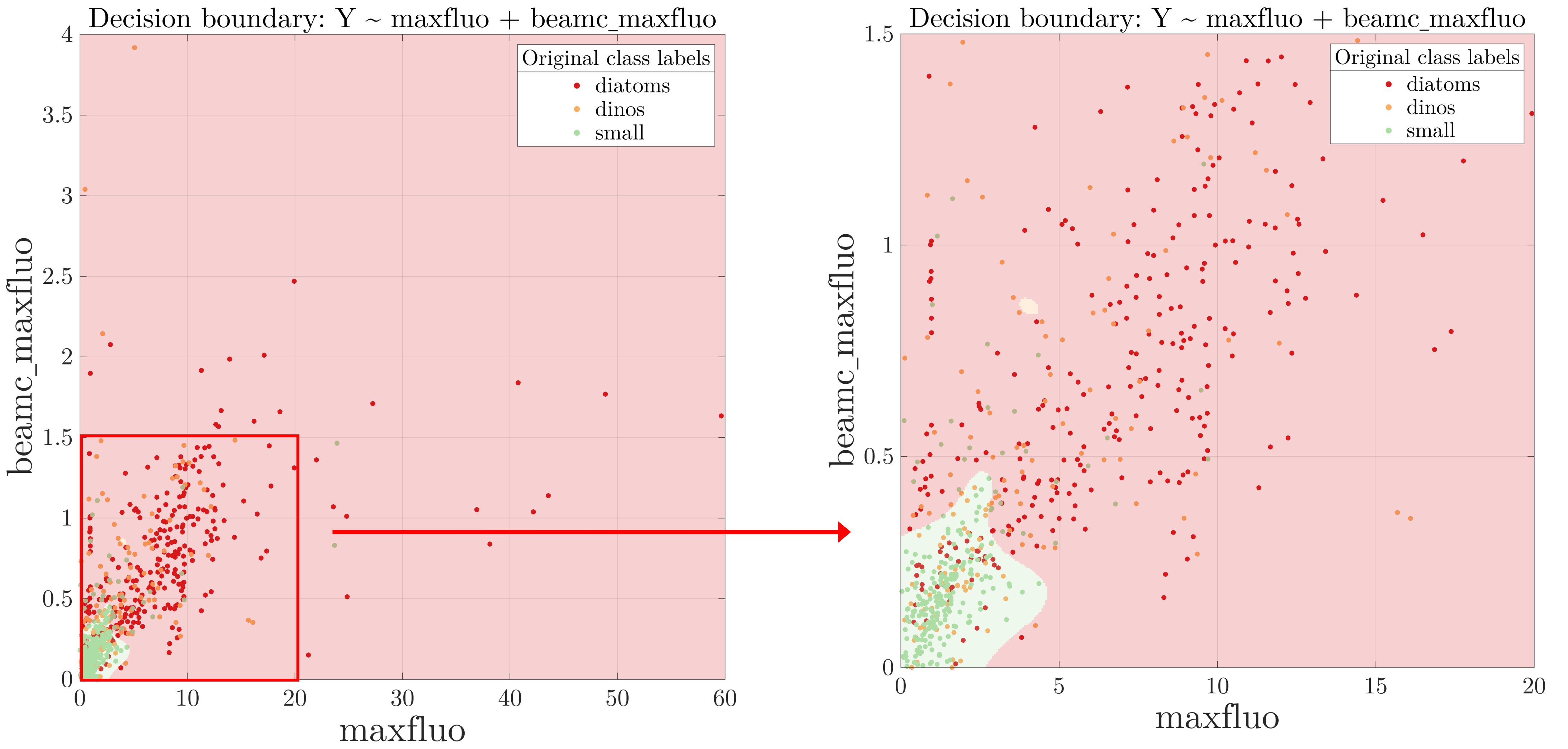
Observed loss value = 0.31865
Function evaluation time = 0.31224

Confusion matrix obtained from cross-validated model:



SVM performance:

Features provide reasonable separation between Diatoms and Pico classes, but are unable to separate Dinoflagellate class



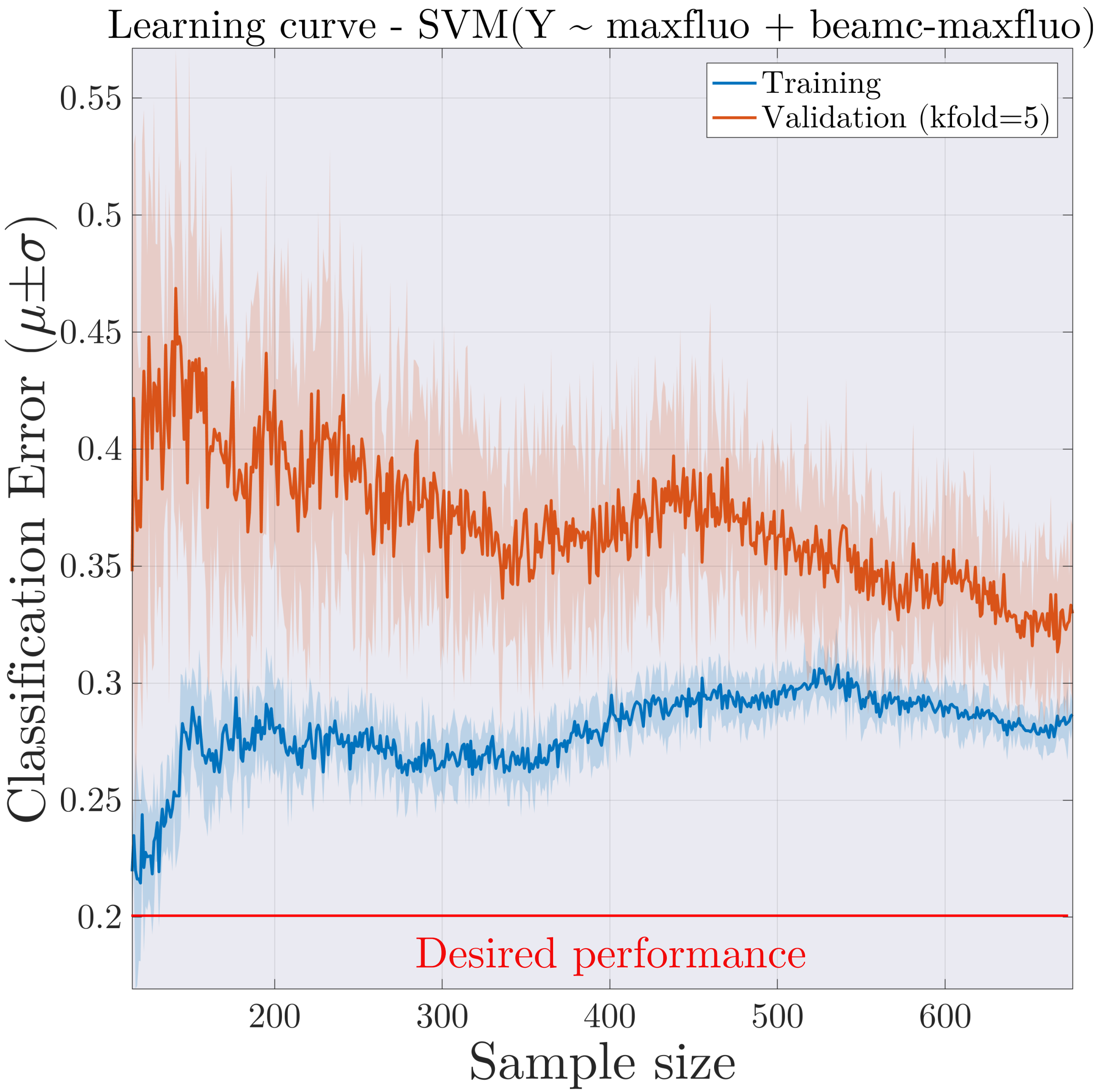
The classification regions from the SVM are shown in the background shading, with data points (actual labels) overlaid

SVM model performance: learning curve

Learning curve:

- High bias learning curve:
  - Training error is high (>20%).
  - Small gap between training and test error.

- These results suggest:
  - Increasing the size of the training set will likely have minor impact on classification performance.
  - Classes aren't well separated in the current feature space.



SVM model selection ($k$means=6)
Sorted per class

SVM model selection – Picoplankton

Top 10 models for predicting Picoplankton dominated communities sorted by MCC:

	Model name	MCC	F1-score	ACC	TPR	FPR	Precision	Recall	gamma
1	timeofyear+chl+allhet+maxfluo	0.7398	0.8327	0.8805	0.8430	0.0991	0.8226	0.8430	0.5
2	timeofyear+chl+maxfluo	0.7227	0.8226	0.8717	0.8430	0.1126	0.8031	0.8430	0.5
3	timeofyear+chl+allhet+beamc+maxfluo	0.7206	0.8176	0.8732	0.8058	0.0901	0.8298	0.8058	0.5
4	chl+allhet+ratiohet+salt+fluo+maxfluo+beamc_maxfluo	0.7192	0.8145	0.8732	0.7893	0.0811	0.8414	0.7893	0.5
5	distcoast+chl+ratiohet+maxfluo	0.7173	0.8180	0.8703	0.8264	0.1059	0.8097	0.8264	0.5
6	timeofyear+chl+ratiohet+beamc+maxfluo+beamc_maxfluo	0.7161	0.8128	0.8717	0.7893	0.0833	0.8377	0.7893	0.5
7	chl+allhet+ratiohet+beamc+maxfluo+beamc_maxfluo	0.7151	0.8150	0.8703	0.8099	0.0968	0.8201	0.8099	0.5
8	distcoast+chl+allhet+ratiohet+maxfluo	0.7151	0.8171	0.8688	0.8306	0.1104	0.8040	0.8306	0.5
9	timeofyear+chl+allhet+fluo+maxfluo+beamc_maxfluo	0.7151	0.8171	0.8688	0.8306	0.1104	0.8040	0.8306	1
10	timeofyear+chl+allhet+ratiohet+fluo+maxfluo+beamc_maxfluo	0.7151	0.8171	0.8688	0.8306	0.1104	0.8040	0.8306	1

*Reported stats are for only the Picoplankton class

Dominating features are: timeofyear, chl, maxfluo, and allhet

SVM model selection – Picoplankton

Top 10 models for predicting Picoplankton dominated communities sorted by MCC:

Model name		MCC	F1-score	ACC	TPR	FPR	Precision	Recall	gamma
1	distcoast+timeofyear+temp+beamc+fluo+maxfluo+beamc_maxfluo	0.6370	0.7701	0.8294	0.8099	0.1599	0.7341	0.8099	1
2	distcoast+timeofyear+strat+beamc+fluo+maxfluo+beamc_maxfluo	0.6335	0.7652	0.8309	0.7810	0.1419	0.7500	0.7810	1
3	bathy+timeofyear+temp+salt+strat+fluo+maxfluo+beamc_maxfluo	0.6319	0.7633	0.8309	0.7727	0.1374	0.7540	0.7727	1
4	timeofyear+temp+salt+beamc+beamc_maxfluo	0.6299	0.7627	0.8294	0.7769	0.1419	0.7490	0.7769	1
5	timeofyear+temp+salt+strat+beamc+beamc_maxfluo	0.6297	0.7640	0.8280	0.7893	0.1509	0.7403	0.7893	1
6	distcoast+bathy+strat+beamc+maxfluo+beamc_maxfluo	0.6245	0.7623	0.8236	0.8017	0.1644	0.7266	0.8017	1
7	distcoast+timeofyear+temp+beamc+maxfluo+beamc_maxfluo	0.6239	0.7626	0.8222	0.8099	0.1712	0.7206	0.8099	1
8	timeofyear+temp+strat+beamc+maxfluo+beamc_maxfluo	0.6235	0.7613	0.8236	0.7975	0.1622	0.7283	0.7975	1
9	distcoast+timeofyear+temp+beamc+beamc_maxfluo	0.6219	0.7608	0.8222	0.8017	0.1667	0.7239	0.8017	1
10	timeofyear+beamc+fluo+beamc_maxfluo	0.6203	0.7602	0.8207	0.8058	0.1712	0.7196	0.8058	1

*Reported stats are for only the Picoplankton class

Features from autonomous sensors

SVM model selection – Diatoms

Top 10 models for predicting Diatoms dominated communities sorted by MCC:

	Model name	MCC	F1-score	ACC	TPR	FPR	Precision	Recall	gamma
1	timeofyear+chl+allhet+ratiohet+temp+maxfluo	0.6830	0.8302	0.8397	0.8791	0.1921	0.7865	0.8791	1
2	timeofyear+chl+allhet+temp+maxfluo	0.6802	0.8276	0.8397	0.8627	0.1789	0.7952	0.8627	1
3	distcoast+timeofyear+allhet+temp+beamc+maxfluo	0.6786	0.8274	0.8382	0.8693	0.1868	0.7893	0.8693	1
4	timeofyear+allhet+ratiohet+temp+beamc	0.6623	0.8154	0.8324	0.8301	0.1658	0.8013	0.8301	1
5	timeofyear+allhet+ratiohet+beamc+fluo	0.6621	0.8192	0.8294	0.8660	0.2000	0.7771	0.8660	1
6	timeofyear+allhet+ratiohet+temp+beamc+beamc_maxfluo	0.6603	0.8194	0.8265	0.8824	0.2184	0.7649	0.8824	1
7	timeofyear+allhet+temp+salt+beamc	0.6577	0.8162	0.8280	0.8562	0.1947	0.7798	0.8562	1
8	timeofyear+allhet+ratiohet+beamc	0.6572	0.8156	0.8280	0.8529	0.1921	0.7814	0.8529	1
9	bathy+timeofyear+allhet+beamc+maxfluo	0.6557	0.8155	0.8265	0.8595	0.2000	0.7758	0.8595	1
10	timeofyear+allhet+beamc+fluo+maxfluo	0.6556	0.8165	0.8251	0.8725	0.2132	0.7672	0.8725	1

*Reported stats are for the target class

Dominating features are: timeofyear, chl, maxfluo, and temp

SVM model selection – Diatoms

Top 10 models for predicting Diatoms dominated communities sorted by MCC:

Model name		MCC	F1-score	ACC	TPR	FPR	Precision	Recall	gamma
1	timeofyear+salt+beamc	0.6484	0.8123	0.8222	0.8627	0.2105	0.7674	0.8627	1
2	hour+timeofyear+temp+beamc+maxfluo	0.6452	0.8105	0.8207	0.8595	0.2105	0.7668	0.8595	1
3	distcoast+timeofyear+beamc+maxfluo	0.6413	0.8063	0.8207	0.8366	0.1921	0.7781	0.8366	1
4	bathy+timeofyear+temp+beamc+fluo+maxfluo	0.6413	0.8063	0.8207	0.8366	0.1921	0.7781	0.8366	1
5	distcoast+timeofyear+temp+beamc+maxfluo	0.6407	0.8075	0.8192	0.8497	0.2053	0.7692	0.8497	1
6	bathy+timeofyear+temp+beamc+beamc_maxfluo	0.6354	0.8061	0.8149	0.8627	0.2237	0.7564	0.8627	1
7	timeofyear+temp+beamc+fluo+maxfluo	0.6334	0.8043	0.8149	0.8529	0.2158	0.7609	0.8529	1
8	timeofyear+temp+salt+beamc+maxfluo	0.6328	0.8049	0.8134	0.8627	0.2263	0.7543	0.8627	1
9	timeofyear+beamc+maxfluo+beamc_maxfluo	0.6328	0.8049	0.8134	0.8627	0.2263	0.7543	0.8627	1
10	timeofyear+temp+MLD+beamc+fluo+maxfluo	0.6315	0.8037	0.8134	0.8562	0.2211	0.7572	0.8562	1

*Reported stats are for the target class

Features from autonomous sensors

onevsone 2.7551 1.9374

SVM model selection – Dinos

Top 10 models for predicting Dinos dominated communities sorted by MCC:

Model name		MCC	F1-score	ACC	TPR	FPR	Precision	Recall	gamma
1	distcoast+timeofyear+chl+allhet+ratiohet+temp+maxfluo	0.6109	0.6455	0.8863	0.5145	0.0201	0.8659	0.5145	1
2	timeofyear+chl+allhet+ratiohet+temp+maxfluo	0.5955	0.6400	0.8819	0.5217	0.0274	0.8276	0.5217	1
3	distcoast+bathy+timeofyear+chl+allhet+ratiohet+temp+salt	0.5944	0.6335	0.8819	0.5072	0.0237	0.8434	0.5072	1
4	distcoast+timeofyear+chl+allhet+ratiohet+beamc	0.5944	0.6335	0.8819	0.5072	0.0237	0.8434	0.5072	1
5	bathy+hour+timeofyear+chl+allhet+temp+salt+beamc+maxfluo+ beamc_maxfluo	0.5939	0.6301	0.8819	0.5000	0.0219	0.8519	0.5000	5
6	hour+timeofyear+chl+temp+salt+MLD	0.5897	0.6339	0.8805	0.5145	0.0274	0.8256	0.5145	1
7	distcoast+timeofyear+allhet+ratiohet+temp+beamc_maxfluo	0.5881	0.6239	0.8805	0.4928	0.0219	0.8500	0.4928	1
8	timeofyear+allhet+ratiohet+beamc	0.5875	0.6095	0.8805	0.4638	0.0146	0.8889	0.4638	1
9	distcoast+timeofyear+chl+allhet+temp	0.5873	0.6407	0.8790	0.5362	0.0347	0.7957	0.5362	1
10	bathy+timeofyear+chl+allhet+temp+salt	0.5873	0.6407	0.8790	0.5362	0.0347	0.7957	0.5362	1

*Reported stats are for the target class

Dominating features are: distcoast, timeofyear, chl, and temp

SVM model selection – Dinos

Top 10 models for predicting Dinos dominated communities sorted by MCC:

	Model name	MCC	F1-score	ACC	TPR	FPR	Precision	Recall	gamma
1	bathy+hour+timeofyear+temp+salt	0.5179	0.5950	0.8571	0.5217	0.0584	0.6923	0.5217	1
2	bathy+timeofyear+temp+salt+beamc+beamc_maxfluo	0.5101	0.5636	0.8601	0.4493	0.0365	0.7561	0.4493	1
3	distcoast+timeofyear+temp+salt+strat+beamc+beamc_maxfluo	0.5044	0.5429	0.8601	0.4130	0.0274	0.7917	0.4130	1
4	distcoast+hour+timeofyear+temp+salt+beamc	0.5018	0.5763	0.8542	0.4928	0.0547	0.6939	0.4928	1
5	distcoast+bathy+timeofyear+temp+salt+maxfluo	0.4999	0.5726	0.8542	0.4855	0.0529	0.6979	0.4855	1
6	distcoast+bathy+hour+timeofyear+temp+salt+maxfluo	0.4914	0.5678	0.8513	0.4855	0.0566	0.6837	0.4855	1
7	distcoast+bathy+timeofyear+temp+salt+beamc_maxfluo	0.4897	0.5395	0.8557	0.4203	0.0347	0.7532	0.4203	0.5
8	bathy+timeofyear+temp+beamc+beamc_maxfluo	0.4897	0.5395	0.8557	0.4203	0.0347	0.7532	0.4203	1
9	distcoast+bathy+hour+timeofyear+temp+salt+beamc+beamc_maxfluo	0.4892	0.5495	0.8542	0.4420	0.0420	0.7262	0.4420	1
10	distcoast+bathy+hour+timeofyear+salt+strat	0.4863	0.5511	0.8528	0.4493	0.0456	0.7126	0.4493	0.5

*Reported stats are for the target class

Features from autonomous sensors