

1 **Sub-monthly prediction of harmful algal blooms based on automated cell**
2 **imaging**

3 Vitul Agarwal¹, Jonathan Chávez-Casillas² and Colleen B. Mouw¹

4

5 ¹Graduate School of Oceanography, University of Rhode Island, Narragansett, USA

6 ²Department of Mathematics and Applied Mathematical Sciences, University of Rhode Island,
7 Kingston, USA

8

9 **Running title:** Timescales of HAB prediction

10 **Emails:** vitulagarwal@uri.edu; jchavezc@uri.edu; cmouw@uri.edu

11 **ORCID IDs**

12 Vitul Agarwal (0000-0002-1523-9044)

13 Jonathan Chávez-Casillas (0000-0002-8494-7538)

14 Colleen B. Mouw (0000-0003-2516-1882)

15

16 **Authors' Contributions and Conflict of Interest**

17 All authors helped prepare the manuscript and approved the final version. The authors declare no
18 competing interests.

19

20 **Data availability statement**

21 All data and code required for the analysis will be made publicly available upon publication.

22

23 **Abstract**

24 Harmful algal blooms (HABs) are an increasing threat to global fisheries and human health. The
25 mitigation of HABs requires management strategies to successfully forecast the abundance and
26 distribution of harmful algal taxa. In this study, we attempt to characterize the dynamics of 2
27 phytoplankton genera (*Pseudo-nitzschia spp.* and *Dinophysis spp.*) in Narragansett Bay, Rhode
28 Island, using empirical dynamic modeling. We utilize a high-resolution Imaging FlowCytobot
29 dataset to generate a daily-resolution time series of phytoplankton images and then characterize
30 the sub-monthly (1-30 days) timescales of univariate and multivariate prediction skill for each
31 taxon. Our results suggest that univariate predictability is low overall, different for each taxon and
32 does not significantly vary over sub-monthly timescales. For all univariate predictions, models can
33 rely on the inherent autocorrelation within each time series. When we incorporated multivariate
34 data based on quantifiable image features, we found that predictability increased for both taxa and
35 that this increase was apparent on timescales >7 days. *Pseudo-nitzschia spp.* has distinctive
36 predictive dynamics that occur on timescales of around 16 and 25 days. Similarly, *Dinophysis spp.*
37 is most predictable on timescales of 25 days. The timescales of prediction for *Pseudo-nitzschia*
38 *spp.* and *Dinophysis spp.* could be tied to environmental drivers such as tidal cycles, water
39 temperature, wind speed, community biomass, salinity, and pH in Narragansett Bay. For most
40 drivers, there were consistent effects between the environmental variables and the phytoplankton
41 taxon. Our analysis displays the potential of utilizing data from automated cell imagers to forecast
42 and monitor harmful algal blooms.

43

44 **Keywords:** Imaging FlowCytobot (IFCB); empirical dynamic modeling (EDM); Narragansett
45 Bay; phytoplankton population dynamics; ecological forecasting; *Pseudo-nitzschia*; *Dinophysis*

46 **Introduction**

47 Harmful algal blooms (HABs) are anomalous increases in phytoplankton abundance,
48 biomass, or distribution that can negatively affect marine ecosystems and public health (Fleming
49 et al. 2011; Berdalet et al. 2016; Karlson et al. 2021). The rising frequency of such events in the
50 past few decades is of increasing global concern (Xiao et al. 2019; Gobler 2020). Some estimates
51 of economic damage due to HABs exceed hundreds of millions of dollars (Anderson et al. 2000),
52 often due to fisheries closures (Brown et al. 2020; Sakamoto et al. 2021), disruption to tourism
53 (Smith et al. 2019; B  chard 2020) and damage to human health (Grattan et al. 2016; Kouakou and
54 Poder 2019). Consequently, the successful prediction and mitigation of HABs is a research priority
55 for state and national governments worldwide (Park et al. 2013; Brooks et al. 2016).

56 HAB predictions often require large amounts of data from various sources and
57 sophisticated modeling techniques (Franks 2018; Ralston and Moore 2020), as well as detailed
58 information on local and regional oceanographic features (Anderson et al. 2010; Dippner et al.
59 2011; Lapucci et al. 2022). Due to the requirement of high resolution and consistent data,
60 monitoring programs are implementing automated systems (Babin et al. 2005; Jochens et al. 2010)
61 with extensively trained algorithms (Sosik and Olson 2007; Ellen et al. 2019; Orenstein et al. 2020)
62 that can identify and alert local officials of the presence, abundance and risk of HAB development.
63 The rapid deployment of such systems has greatly expanded the ability to detect HABs; however,
64 less is known about the utility of imaging data for HAB prediction models.

65 In this study, we explored the use of phytoplankton imaging data for HAB predictions in
66 Narragansett Bay, Rhode Island (NBay). Narragansett Bay is a shallow coastal marine estuary of
67 great cultural, historical, and economic importance to local communities (Herndon and Sekatau
68 1997; Dalton et al. 2010; Nixon and Fulweiler 2012). Coastal marine estuaries are highly dynamic

69 environments that are subject to seasonality (Carstensen et al. 2015), the influence of both
70 freshwater and seawater sources (Pilson 1985), large-scale climate phenomena (Scavia et al. 2002),
71 and anthropogenic inputs of nutrients (Cundell 1973; Goldberg et al. 1977). Within the past
72 decade, toxic harmful algal blooms have led to fisheries closures in Narragansett Bay (Bates et al.
73 2018) and motivated extensive research into the potential environmental drivers and species
74 composition (Sterling et al. 2022) of the relevant bloom-causing phytoplankton genera. Of
75 particular importance in this area are *Pseudo-nitzschia spp.* and *Dinophysis spp.* due to their
76 potential toxicity and relevance for local fishery disruption.

77 Natural phytoplankton populations are variable from daily, seasonal to decadal timescales
78 (Chavez et al. 2003; Barton et al. 2016; Blauw et al. 2018). High variability in natural
79 phytoplankton populations is characteristic of non-linear and chaotic dynamics (Ascioti et al. 1993;
80 Smayda 1998). In this paper, we used empirical dynamic modeling (EDM) to predict the
81 abundance of *Pseudo-nitzschia spp.* and *Dinophysis spp.* in Narragansett Bay, Rhode Island. These
82 genera were selected for a couple of reasons: their role in local harmful algal blooms and the
83 availability of a dynamic, regular time series that would allow for the use of data-driven analyses.
84 EDM is a non-parametric framework that can avoid the pitfalls of typical statistical modeling by
85 relying on data-driven attractor reconstruction (Perretti et al. 2013; Chang et al. 2017).

86 Our goal was to characterize the sub-monthly univariate and multivariate prediction
87 timescales of *Pseudo-nitzschia spp.* and *Dinophysis spp.* utilizing a high temporal resolution
88 dataset generated with an Imaging FlowCytobot (Olson and Sosik 2007). Automated instruments
89 such as the Imaging FlowCytobot generate datasets of phytoplankton images and many associated
90 features (i.e. image texture, contrast, object size etc.). Specifically, we aim to answer (1) How
91 predictable are the harmful algal target species? (2) How does this predictability vary with time?

92 and (3) Which image features best describe the dynamics of the taxa? Once we identified the
93 important timescales, we also linked specific environmental drivers to the dynamics of the
94 phytoplankton populations. Our study did not attempt to offer detailed mechanistic explanations
95 of observed phenomena, nor develop tools that might model the growth and termination of harmful
96 algal blooms, but instead, it focused on identifying the potential of imaging data in prediction
97 models. By identifying the relevant dynamical timescales of harmful algal blooms, we also hoped
98 to provide local and regional management with a critical timeframe of action for the development
99 of environmental policy. Our underlying assumption was that the predictability of *Pseudo-*
100 *nitzschia spp.* and *Dinophysis spp.* in Narragansett Bay had distinct timescales that varied in
101 response to environmental drivers and intrinsic population dynamics.

102

103 **Materials and Methods**

104 *Automated cell imaging*

105 All the time series used in this study were collected by deploying an Imaging FlowCytobot
106 (IFCB) in Narragansett Bay, Rhode Island. The IFCB is an automated, flow-through imaging
107 system that captures images of the extant phytoplankton community in seawater. The system has
108 a maximum size limit of 150 μ m and works by drawing water at approximately 1m under the
109 surface at low tide. As our IFCB was deployed at the end of a pier (41.492°N, 71.419°W), the
110 actual sampling depth varied with the tidal cycle. Images can be observed in real-time using the
111 IFCB dashboard (<http://ifcb-dashboard.gso.uri.edu/>).

112 The IFCB samples approximately every 20 minutes depending on the number of cells
113 within a given sample. We used daily aggregated IFCB data from 14th June 2017 to 20th October

114 2021, barring gaps in the time series due to equipment malfunction or maintenance. Our data span
115 1590 days of observation with 518 days of missing data. For prediction tasks, all missing data
116 points were approximated using 30-day exponential moving averages (EMA) computed by the R
117 package “imputeTS” (Moritz and Bartz-Beielstein 2017). We used daily aggregated data, instead
118 of other shorter timescales (such as 1-hr or 12-hr) for three broad reasons: the influence of high
119 time series autocorrelation, irregular gaps in data collection, and, to strike a balance between
120 computational costs and expected analytical benefit.

121 A machine-learning approach was used to identify and classify the phytoplankton taxa
122 from a subset of annotated images (Sosik and Olson 2007). All obtained images classified as
123 *Pseudo-nitzschia spp.* and *Dinophysis spp.* were counted and reported as a concentration based on
124 the average sampling volume for each day (*images mL*). Higher concentrations of images act
125 as a proxy of higher abundance in the natural environment and lower concentrations of images
126 show that the taxon is rare/absent. To test the general ability to use image concentration as a proxy
127 for phytoplankton abundance, we visually compared our image concentration time series to a long-
128 term weekly monitoring site located approximately 12km north of our IFCB location
129 (<https://web.uri.edu/gso/research/plankton/>). Figure S1 highlights that our IFCB image
130 concentration agreed with the general pattern of *Pseudo-nitzschia spp.* abundance in Narragansett
131 Bay (as determined by microscopy counts) over the duration of our time series.

132 We evaluated the classifier’s performance for sensitivity and precision with a manually
133 annotated library of images. Table 1 reports the performance of the classifier for *Pseudo-nitzschia*
134 *spp.* and *Dinophysis spp.*

135

$$e \text{ si } i \text{ vi } y = \frac{TP}{TP + FN}$$

136
$$Precision = \frac{TP}{TP + FP}$$

137 where TP, FP and FN were the number of true positive, false positive, and false negative images.

138

139 **Table 1:** Sensitivity and precision of the automatic classifier for each of the phytoplankton classes.

	Sensitivity	Precision
<i>Pseudo-nitzschia spp.</i> (N=626)	0.85	0.94
<i>Dinophysis spp.</i> (N=412)	0.95	0.96

140

141 236 image features are automatically estimated for each IFCB image (Sosik and Olson
142 2007, <https://github.com/hsosik/ifcb-analysis/wiki>). We selected 20 features for further analysis
143 based on their relevance to phytoplankton morphology and ecology (Sonnet et al. 2022). The
144 average daily values for the image features, scaled by the average sampling volume for each day,
145 formed a multi-dimensional time series for each taxon. Table 2 lists all the features and their units.

146

147

148

149

150

151

152

153 **Table 2:** List of all image features used in this study and their units.

Feature	Units
area	<i>pixels</i> ²
biovolume	<i>pixels</i> ³
major axis	<i>pixels</i>
minor axis	<i>pixels</i>
perimeter	<i>pixels</i>
orientation	<i>degrees</i>
eccentricity	-
solidity	-
texture uniformity	-
texture smoothness	-
texture gray	-
texture entropy	-
texture contrast	-
h90	<i>pixels</i>
h180	<i>pixels</i>
hflip	<i>pixels</i>
extent	<i>pixels</i>
equivalent diameter	<i>pixels</i>
convex area	<i>pixels</i> ²
convex perimeter	<i>pixels</i>

154

155

156

157

158

159 *Environmental data*

160 We compiled data from various monitoring programs located in and around Narragansett
161 Bay. Daily averages of water temperature ($^{\circ}$), salinity (pp), chlorophyll ($\mu g L^{-1}$), and pH were
162 requested from the Narragansett Bay Fixed Site Monitoring Network (NBFSMN, personal
163 communication: Heather Stoffel). These measurements were co-located with the Imaging
164 FlowCytobot. Daily averages of wind speed ($m s^{-1}$) were drawn from the Kingston weather
165 station (41.49°N 71.54°W; U.S. Climate Reference Network;
166 <https://www1.ncdc.noaa.gov/pub/data/uscrn/products/subhourly01/>). Daily averages of tidal
167 height (Mean Sea Level; m) were calculated from measurements at the NOAA Quonset Point
168 Buoy (41° 35.2 N, 71° 24.6 W; #8454049; <https://tidesandcurrents.noaa.gov/>). Additional
169 environmental data, such as ambient nutrient concentrations, were not available at the same
170 temporal scale as the imaging data used in the study.

171

172 *Univariate predictions*

173 We used empirical dynamic modeling (EDM) to make univariate predictions for the time
174 series of each harmful algae. Every time series was normalized (i.e. subtracting the mean value of
175 the time series and dividing by the standard deviation of the time series) before the application of
176 EDM. Specifically, we relied on simplex projection (Sugihara and May 1990) with a consistent
177 embedding dimension of 4 and leave-one-out cross validation. This means that the univariate
178 attractor of a time series was embedded in a 4-dimensional space using the original times series
179 and successive lags of the same time series. Each point is described by $(x_t, x_{t-1}, x_{t-2}, x_{t-3})$ where
180 x_t is the value of x at time point t , x_{t-1} is its value at time $t-1$, and so on. The embedding

181 dimension was set at 4 to prevent overfitting and maximize the utilization of our daily-scale time
182 series data. Figure S2 shows how varying the embedding dimension affects the predictability of
183 each taxon with fixed prediction intervals of 1, 7, 14 and 28 days. An embedding dimension of 4
184 allowed for reasonable descriptions across all timescales for both taxa, with a lower risk of
185 overfitting our models to potentially noisy dynamics. The model creation and prediction sets were
186 randomly selected from the entire time series in intervals of 250 days. After bootstrapping 200
187 samples for each taxon, we generated a mean prediction and 95% confidence intervals ($1.96 \times$
188 E). By randomizing the selection of model and prediction libraries, we tried to account for the
189 effects of imputed data into the time series, as well as avoid the effects of possible non-stationarity
190 over the 1590 days of data.

191 We tested the predictability of each taxon for timescales of 1 to 30 days. Predictability was
192 described by ρ_{model} , the Pearson correlation coefficient, between the observed and the predicted
193 values after attractor reconstruction. To account for inherent autocorrelation within each time
194 series, we subtracted the absolute value of the autocorrelation coefficient at each timescale of
195 prediction. The effective value of predictability was reported as $\Delta\rho$, which is the arithmetic
196 difference of the univariate predictability ρ_{model} and the autocorrelation coefficient ρ_{auto} .
197 Therefore, $\Delta\rho$ quantifies the ability of our model to predict dynamics beyond autocorrelation
198 across a range of sub-monthly timescales. Due to the short timescales of prediction in this study
199 (<30 days), our dataset of 1590 days provided reasonable coverage of all possible sub-monthly
200 dynamics for these harmful algal taxa.

201

202

203 *Multiview Embeddings (MVE)*

204 Multiview embeddings are an effective technique for increasing predictability and drawing
205 out information from multiple related time series (Ye and Sugihara 2016). We used MVE to utilize
206 the associated dataset of image features collected by the IFCB. Once more, the embedding
207 dimension was set to 4 for all taxa and the entire time series was used for model and prediction
208 libraries. We relied on leave-one-out cross-validation instead of separate model and prediction
209 libraries.

210 Each multivariate attractor was created by randomly selecting 3 normalized time series of
211 features and the original time series of image concentration (*images mL*). Our goal was to
212 predict the proxy abundance of each taxon by leveraging information stored in the image features.
213 Predictability was evaluated for timescales of 1-30 days and reported as $\Delta\rho$ (model predictability
214 beyond autocorrelation), RMSE (root-mean-square error) and MAE (mean absolute error). We
215 considered 500 trials of image feature combinations and reported predictability as the arithmetic
216 mean with 95% confidence intervals ($1.96 \times E$).

217 For the best multivariate models (top 5% in terms of ρ_{model}), we reported the frequency of
218 appearance for each image feature as a proportion. A proportion of 0 implies that the feature did
219 not show up in the best multivariate models while a proportion of 1 implies that it was always
220 present. Based on the frequency of appearance, we could deduce the contribution of each feature
221 in improving the overall predictability of the phytoplankton species.

222

223

224

225 *Convergent Cross Mapping (CCM)*

226 Once we identified any relevant timescales of prediction, we wanted to understand whether
227 there was a link between the abundance of harmful algal taxa and relevant environmental drivers.
228 We used convergent cross mapping (CCM; Sugihara et al. 2012) to infer causation between the
229 environmental dataset and image concentration (*images mL*⁻¹). Embedding dimensions were
230 optimized (i.e. selecting the embedding dimension that provides the highest prediction skill ρ) to
231 each environmental variable (up to a maximum of 7 to prevent overfitting) and library sizes ranged
232 from 100 - 1400 in intervals of 100 days. There were 20 samples each for every library size and
233 the time to prediction ranged from 1-30 days. We tested whether we could infer causation by
234 predicting the values of past environmental variables from the abundance of the harmful algal taxa.
235 Predictability was quantified by the cross-map prediction skill (ρ), where higher values indicate
236 better predictions. Convergence was estimated using three tests – (1) Mann-Kendall trend test for
237 ρ with increasing library size, (2) a Student's t-test for the ρ distributions at the maximum and
238 minimum library size and (3) by validating that the prediction skill ρ at the maximum library size
239 was greater than the Pearson correlation coefficient between image concentration and the
240 environmental time series. Only the predictions which satisfied all conditions and were significant
241 for both the Mann-Kendall and the Student's t-test (p-value < 0.05) were deemed convergent. If
242 any of the tests failed, then the causal effect of the environmental variable on the phytoplankton
243 taxa was deemed to be unresolved at those specific timescales. Cross-map prediction skill (ρ) was
244 normalized to the embedding dimension by averaging ρ across prediction horizons (Saberski et al.
245 2021).

246

247 *Software*

248 All the analyses were conducted in R (R Core Team 2021). For plotting and data
249 visualization, we used the R packages “ggplot2” (Wickham 2016) and “cowplot” (Wilke 2020).
250 EDM was applied using pre-built functions in the R package “rEDM” (Park et al. 2022).
251 Additionally, the R package “Kendall” (McLeod 2022) was used to conduct some statistical tests.

252

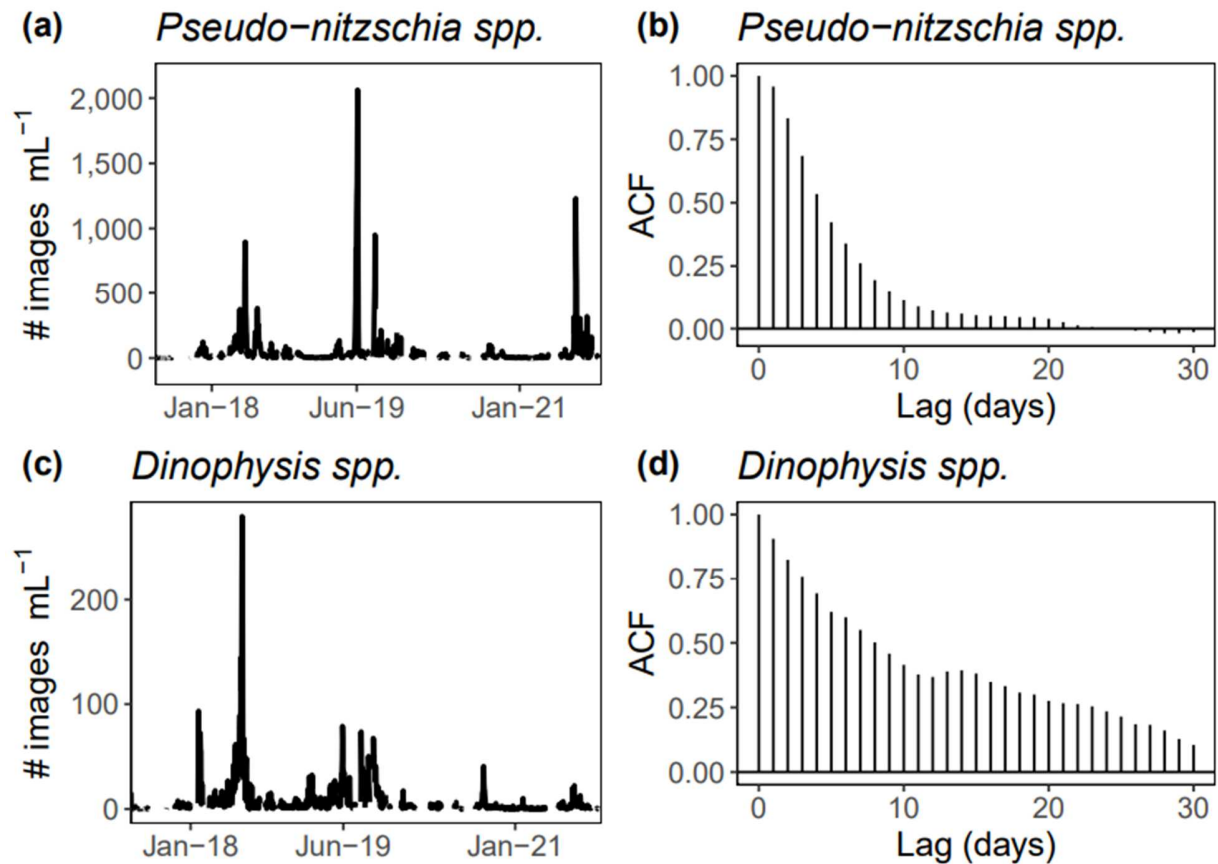
253 **Results**

254 Both *Pseudo-nitzschia spp.* and *Dinophysis spp.* in Narragansett Bay showed intermittent
255 periods of high and low abundance in Narragansett Bay. The IFCB captured such bloom dynamics
256 using the concentration of identified images of both taxa (Figure 1; left column). When evaluated
257 for the autocorrelation inherent within each time series, both *Pseudo-nitzschia spp.* and *Dinophysis*
258 *spp.* had decreasing autocorrelation with time (Figure 1; right column). The decrease was more
259 rapid for *Pseudo-nitzschia spp.* (ACF < 0.25 within 7 days), whereas *Dinophysis spp.* showed a
260 more gradual decrease over the entire 30 days.

261

262

263



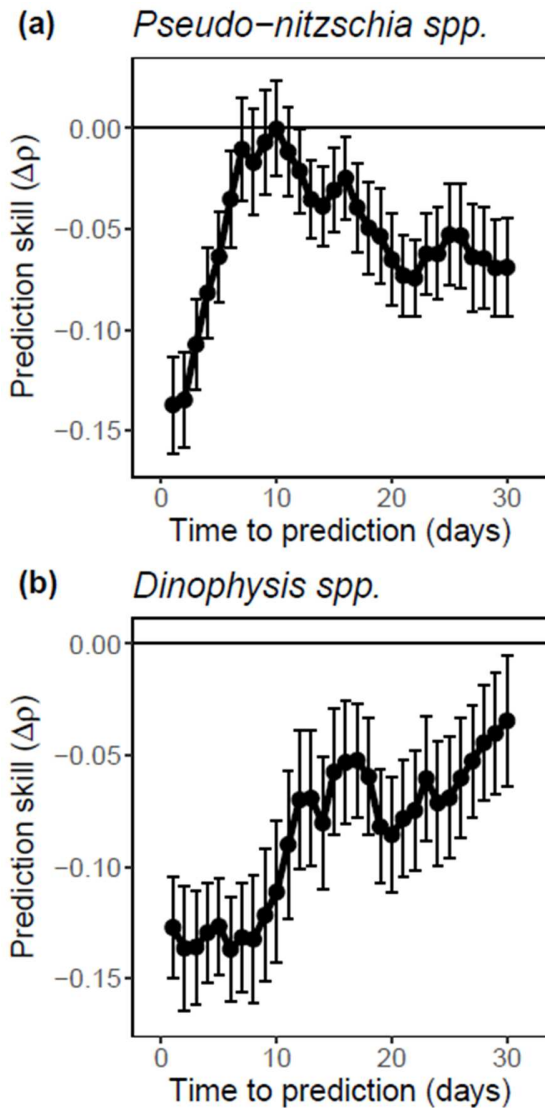
264

265 **Figure 1:** Time series of 2 harmful algal bloom-forming taxa in Narragansett Bay, Rhode Island
 266 (left column) and their associated autocorrelation functions (ACF; right column). Relative
 267 abundance is estimated from the number of unique images taken by the IFCB and classified as (a)
 268 *Pseudo-nitzschia* spp. and (c) *Dinophysis* spp. Autocorrelation decreases with time and varies
 269 depending on the dynamics of each specific taxon.

270
 271

272 The univariate predictability of both time series ($\Delta\rho$) was low overall and did not greatly
 273 change over a prediction horizon of 30 days. The univariate predictability of *Dinophysis* spp.
 274 indicated some promise of the model over autocorrelation on horizons of >28 days, whereas the
 275 results for *Pseudo-nitzschia* spp. indicated that there is little to no predictability inherent within
 276 the time series beyond autocorrelation across all sub-monthly timescales.

277



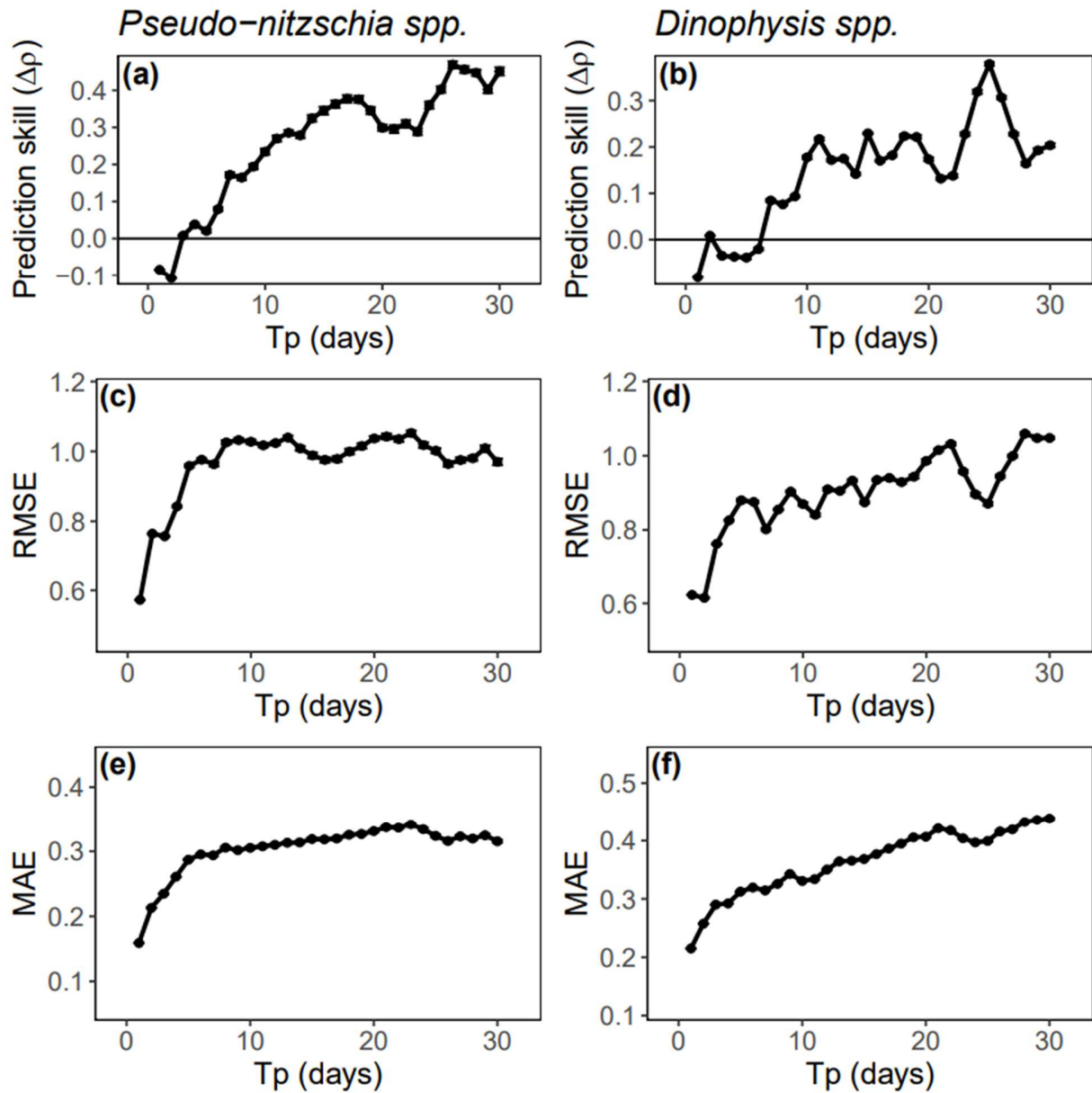
278

279 **Figure 2:** Univariate prediction skill ($\Delta\rho$) of the time series of (a) *Pseudo-nitzschia spp.* and (b)
 280 *Dinophysis spp.* over a prediction horizon of 1-30 days. Model predictions (ρ_{model}) were
 281 calculated from 200 random libraries of 250 days each and the results were reported as an
 282 arithmetic mean with 95% confidence intervals ($\pm 1.96 \times .E.$). $\Delta\rho$ was calculated by subtracting
 283 the autocorrelation coefficient at each prediction horizon.

284

285 Multivariate prediction skill ($\Delta\rho$), calculated using the time series of image abundance and
 286 3 associated image features, was much higher than the univariate prediction skill for both *Pseudo-*
 287 *nitzschia spp.* and *Dinophysis spp.* Using the original time series with only 3 image features at a

288 time (multivariate embedding dimension = 4) allowed for direct comparisons to the univariate
289 prediction skill. The predictability of *Pseudo-nitzschia spp.* had distinctive cycles with peaks every
290 16 and 25 days. An increase in model predictability over autocorrelation was most prominent after
291 a 3-day prediction horizon. The predictability of *Dinophysis spp.* was also higher than inherent
292 autocorrelation after a 6-day prediction horizon. *Dinophysis spp.* had multiple peaks in $\Delta\rho$ at
293 prediction horizons of 10-20 days, with a distinct peak at 25 days. The RMSE of the predictions
294 also showed a distinct drop around the 25-day mark.



295

296 **Figure 3:** Multivariate prediction skill of the time series of *Pseudo-nitzschia spp.* (left column)
 297 and *Dinophysis spp.* (right column) over a prediction horizon of 1-30 days. (a) and (c) report
 298 prediction skill ($\Delta\rho$) calculated by subtracting the autocorrelation coefficient at each prediction
 299 horizon, (b) and (d) report prediction error as the root-mean-squared-error (RMSE), (e) and (f)
 300 report prediction error as the mean-absolute-error (MAE). Model results were calculated from 500
 301 embeddings of phytoplankton abundance and 3 unique image features. The results were reported
 302 as an arithmetic mean with 95% confidence intervals ($\pm 1.96 \times .E.$).

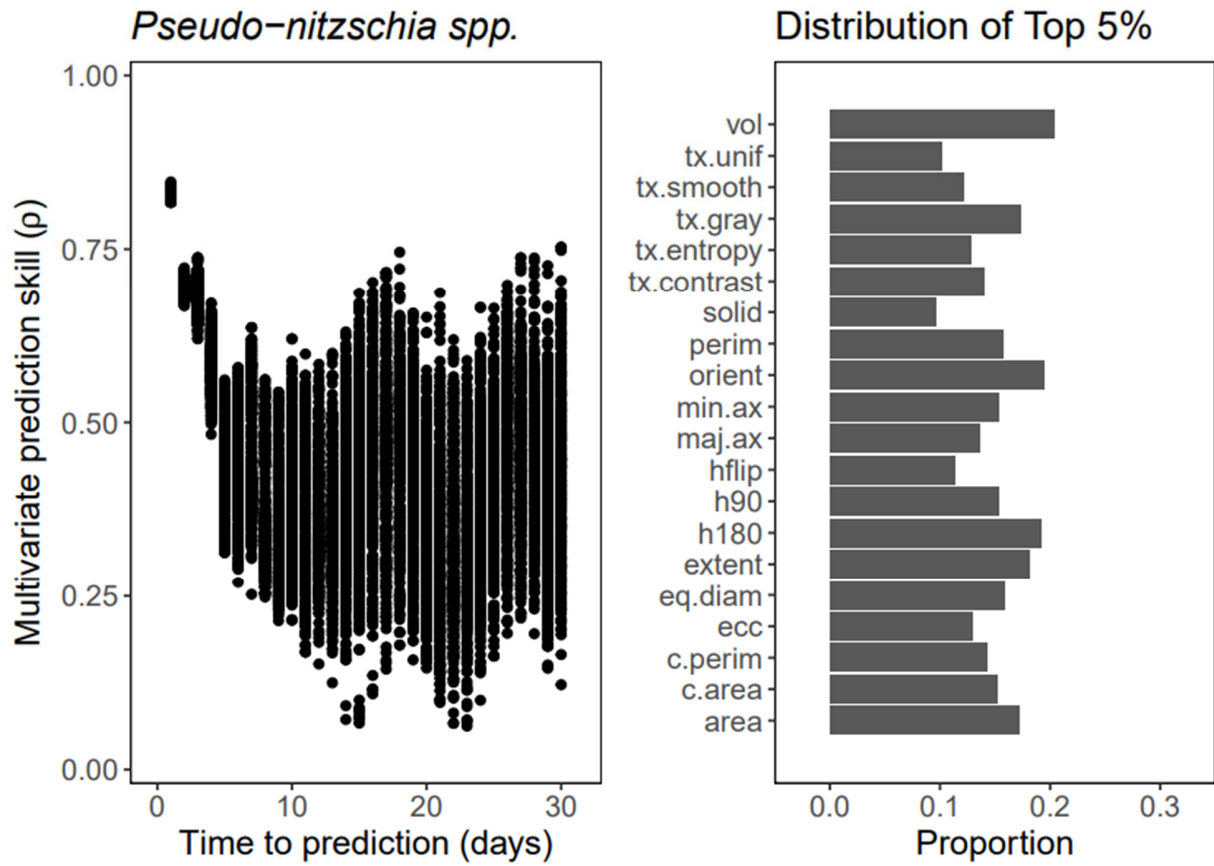
303

304 Without accounting for autocorrelation, some multivariate models for both *Pseudo-*
305 *nitzschia spp.* and *Dinophysis spp.* reached prediction skills of 0.70 and greater (Figure 4&5; left).
306 The top 5% of these multivariate models had a fairly uniform distribution of image features, with
307 some clear exceptions. The time series of biovolume and orientation prominently appeared in the
308 top multivariate models for *Pseudo-nitzschia spp.*, whereas the time series of solidity and hflip
309 were present but in a distinctly lower frequency compared to the other features. For *Dinophysis*
310 *spp.*, the time series of biovolume, texture gray and orientation were prominently present in the
311 top multivariate models.

312

313

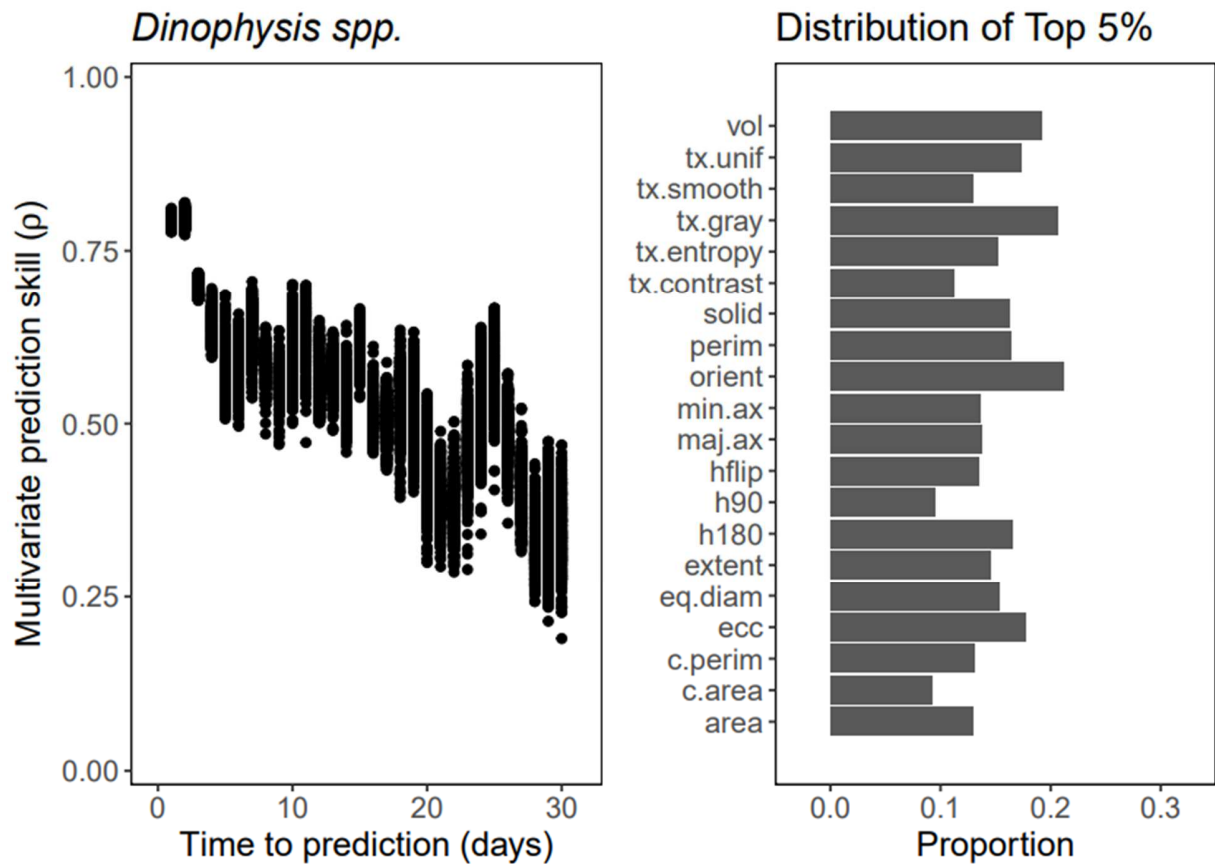
314



315

316 **Figure 4:** Multivariate prediction skill of the time series of *Pseudo-nitzschia spp.* (ρ ; left) over a
 317 prediction horizon of 1-30 days. Prediction skill (ρ) refers to the Pearson correlation coefficient
 318 between model predictions and actual observations. Model results were calculated from 1000
 319 embeddings of phytoplankton abundance and 3 unique image features. Each point is the outcome
 320 of a single model run. Frequency of image features (right) summarizes the top 5% of model
 321 outcomes and the image features included in these models.

322



323

324 **Figure 5:** Multivariate prediction skill of the time series of *Dinophysis spp.* (ρ ; left) over a
 325 prediction horizon of 1-30 days. Prediction skill (ρ) refers to the Pearson correlation coefficient
 326 between model predictions and actual observations. Model results were calculated from 1000
 327 embeddings of phytoplankton abundance and 3 unique image features. Each point is the outcome
 328 of a single model run. Frequency of image features (right) summarizes the top 5% of model
 329 outcomes and the image features included in these models.

330

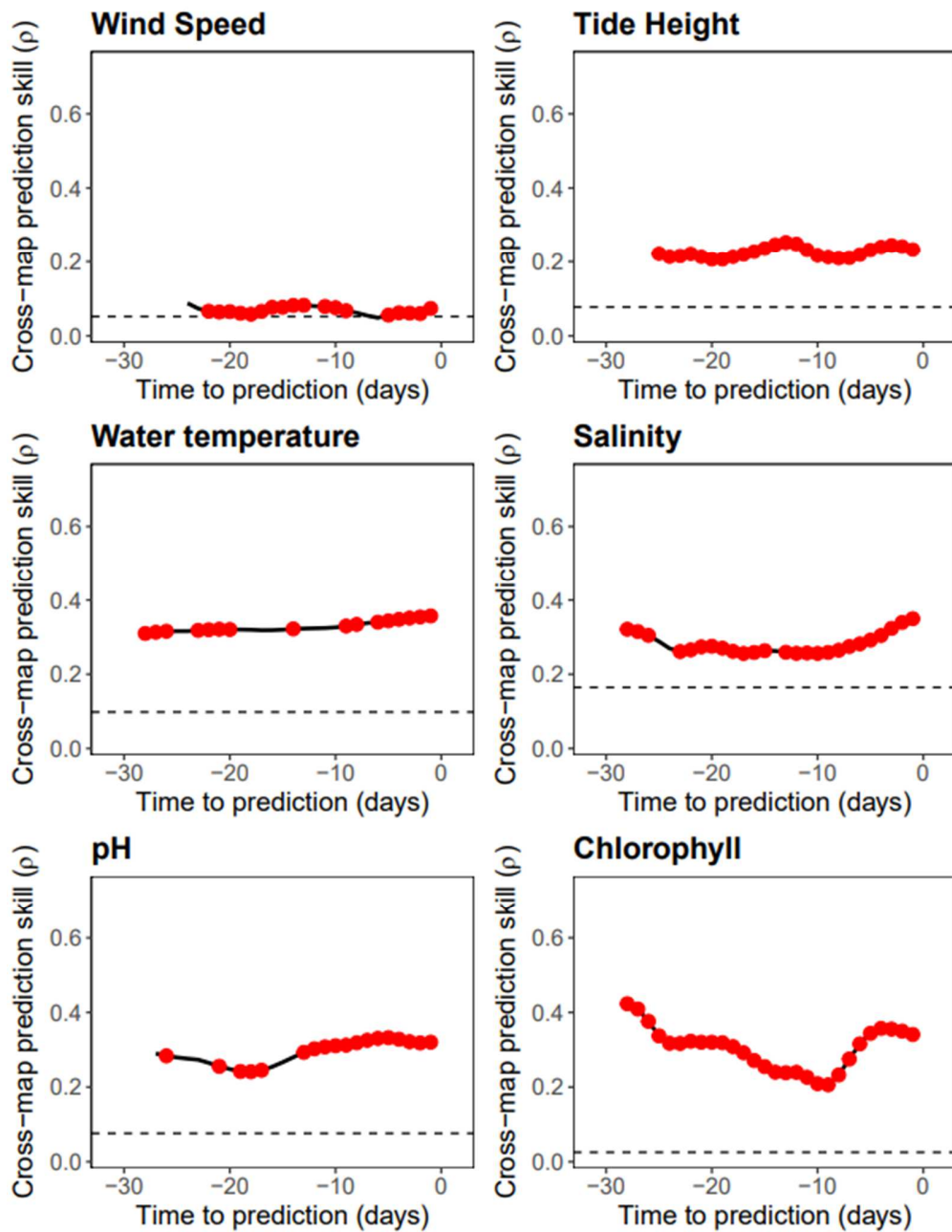
331 Environmental influence on the populations of *Pseudo-nitzschia spp.* and *Dinophysis spp.*,
 332 as measured by the cross-map prediction skill (ρ), showed variable effects across different
 333 prediction horizons. For *Pseudo-nitzschia spp.*, the prediction skill of all the environmental
 334 parameters converged with increasing library size. The strength and timescales of inferred causal
 335 influence differed across the variables. The influence of chlorophyll on *Pseudo-nitzschia spp.* had
 336 a peak at timescales around 28 days, whereas the influence of tidal height was strongest around 2

337 weeks. The time series of water temperature, pH, wind speed, and salinity showed significant and
338 consistent effects on the *Pseudo-nitzschia spp.* time series across most sub-monthly timescales.

339

340

Pseudo-nitzschia spp.

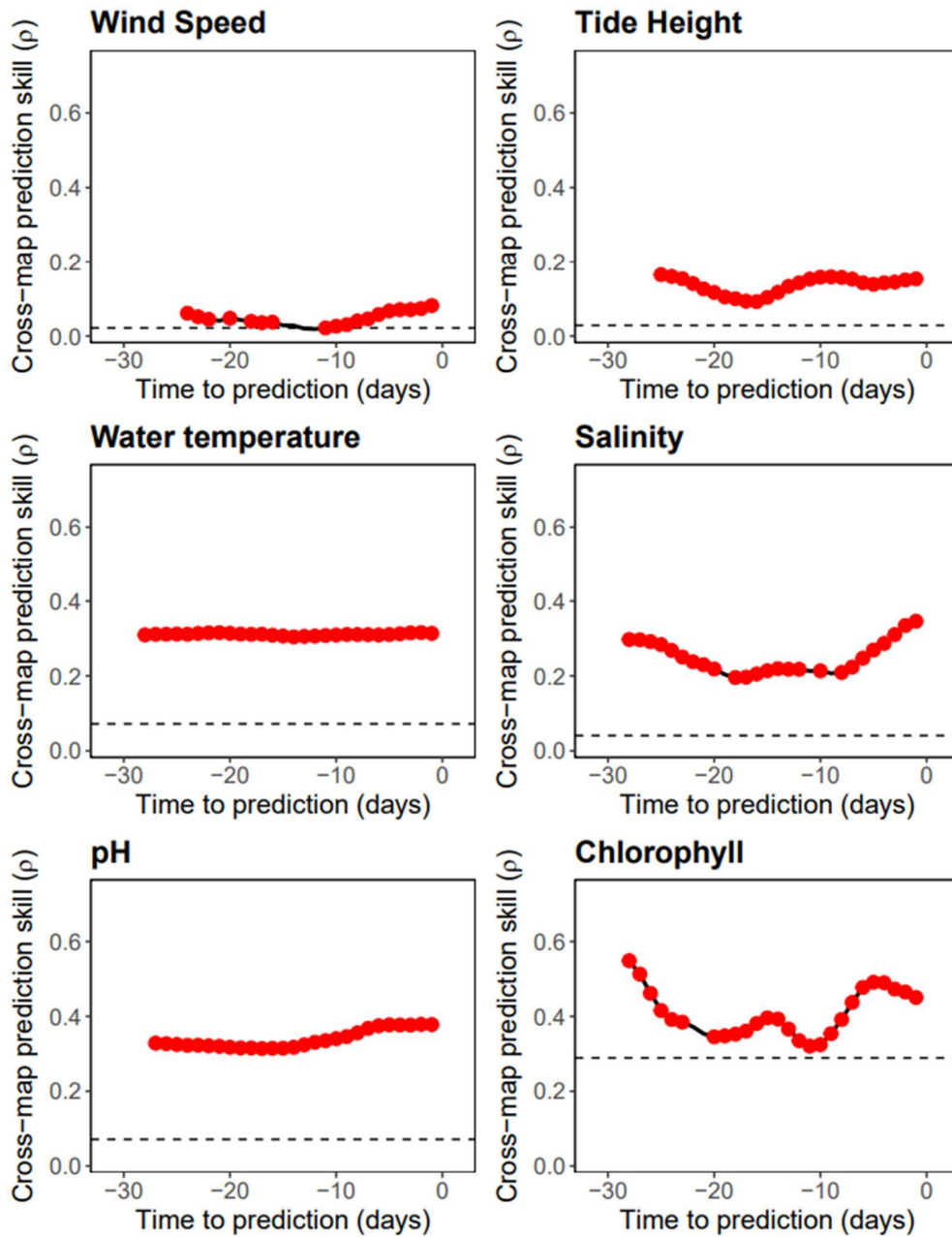


341

342 **Figure 6:** Influence of environmental drivers on *Pseudo-nitzschia spp.* in Narragansett Bay
343 quantified by the cross-map prediction skill (ρ based on convergent cross mapping; see Methods).
344 The influence was measured over a prediction horizon of 1-30 days (black line). Red points
345 indicate which models showed convergence. The dashed line refers to the Pearson correlation
346 coefficient between the time series of *Pseudo-nitzschia spp.* abundance and the environmental
347 variable.

348 For *Dinophysis spp.*, there were more models that showed convergence across all
349 prediction horizons. The time series of *Dinophysis spp.* was consistently affected by tide height,
350 water temperature, and pH across all timescales. Total biomass (chlorophyll) appeared as a
351 significant driver of *Dinophysis spp.* with peaks around 5, 14 and 27 days. The effects of salinity
352 were consistent and stronger in the short-term (1-3 days). Predictability was higher than the
353 Pearson correlation coefficient for most environmental variables and showed consistency, which
354 might suggest specific mechanisms of causal influence.

Dinophysis spp.



355

356 **Figure 7:** Influence of environmental drivers on *Dinophysis spp.* in Narragansett Bay quantified
357 by the cross-map prediction skill (ρ based on convergent cross mapping; see Methods). The
358 influence was measured over a prediction horizon of 1-30 days (black line). Red points indicate
359 which models showed convergence. The dashed line refers to the Pearson correlation coefficient
360 between the time series of *Dinophysis spp.* abundance and the environmental variable.

361

362 Discussion

363 *Predictability of Pseudo-nitzschia spp. and Dinophysis spp.*

364 Perturbations in phytoplankton population dynamics typically decorrelate within
365 timescales of a month (Kuhn et al. 2019). When we tested for the inherent autocorrelation within
366 the time series of both *Pseudo-nitzschia spp.* and *Dinophysis spp.*, we found that the
367 autocorrelation decreased significantly within the first 10 days for *Pseudo-nitzschia spp.*, but
368 *Dinophysis spp.* had higher autocorrelation for up to 30 days. After accounting for autocorrelation,
369 the univariate predictability of both *Pseudo-nitzschia spp.* and *Dinophysis spp.* was low overall;
370 however, the univariate predictability of *Pseudo-nitzschia spp.* showed some cyclical behavior.
371 Our univariate models likely picked up on repetitive population-level mechanisms that increased
372 or decreased abundance on sub-monthly timescales. Some examples of such mechanisms could
373 include regular switching between periods of growth and sexual reproduction (D'Alelio et al.
374 2009; Annunziata et al. 2022), density-dependent interactions with parasitic protists (Berdjeb et
375 al. 2018), or the tidal transport of productive populations from nearby sites (Shanks et al. 2014).
376 Part of the lack of univariate predictability could be due to the presence of measurement error and
377 stochasticity in the time series of both taxa, as well as a general lack of natural predictability for
378 larger diatoms and dinoflagellates (Agarwal et al. 2021).

379 In the multivariate case, we found the predictability of both *Pseudo-nitzschia spp.* and
380 *Dinophysis spp.* improved on timescales of greater than 1 week. Multiview embeddings have been
381 previously shown to improve the univariate predictability of short time series (Ye and Sugihara
382 2016). By leveraging information stored across multiple related image features, our approach of
383 randomly creating non-lagged embeddings could have allowed us to create better and more reliable
384 estimates of predictive dynamics (Ma et al. 2018). The cyclical predictability of *Pseudo-nitzschia*

385 *spp.* was more prominent in the multivariate models, implicating predictable behavior on 16-day
386 and 25-day timescales. *Dinophysis spp.* was most predictable on timescales of 25 days. Due to the
387 presence of distinct timescales of predictability for both taxa, our results suggest that future
388 development of HAB models would benefit by resolving dynamics on daily and weekly
389 timescales. The identification of relevant ecological and environmental drivers of population
390 dynamics on these timescales might also aid in the development of automated monitoring and
391 early-warning systems.

392

393 *Relative contribution of IFCB image features*

394 When we evaluated the relative proportions of image features among the top multivariate
395 models, the time series of biovolume was prominently present for *Pseudo-nitzschia spp.* and
396 *Dinophysis spp.* This implies that the time series of biovolume adds considerable information to
397 the future predictability of harmful algal taxa. Biovolume estimates from IFCB images (Moberg
398 and Sosik 2012) are often used as an important marker of phytoplankton community structure and
399 function (Brosnahan et al. 2015; Oliver et al. 2021). Although image-derived biovolume estimates
400 might differ from microscopy-derived estimates (Kraft et al. 2021), cell biovolume typically varies
401 linearly with other phytoplankton functional traits (Edwards et al. 2012). Our results suggest that
402 including biovolume estimates and other high-performing image descriptors into models for
403 harmful algal taxa improves predictability beyond autocorrelation.

404 Image descriptors derived from flow cytometers have found utility in studies of
405 phytoplankton morphology (Sonnet et al. 2022), as well as for the training of different image
406 classifiers (Mosleh et al. 2012; Zheng et al. 2017). In general, “features” from an IFCB image are

407 all calculated from the pixels of the image and the relationships between them (see Table 2). As
408 most features share the fundamental quantity underlying their calculations (i.e. the image itself),
409 we expect all time series to be nonlinear approximations of one another. The relatively consistent
410 proportions of most features in the top multivariate models indicate that the use of features
411 themselves, and not necessarily their “character”, increases the predictability of harmful algal taxa.
412 Unless there is a particular reason to prefer one feature for another (such as biovolume for its
413 relationship to other traits), prediction models relying on automated imaging systems would
414 benefit from using any associated image data. Detailed information on the causal relationships
415 between image features for *Pseudo-nitzschia spp.*, as well as the partial correlations between each
416 image feature and our time series of image concentration, can be found in the Supplemental
417 material.

418

419 *Potential environmental drivers*

420 To further investigate the timescales of prediction for both *Pseudo-nitzschia spp.* and
421 *Dinophysis spp.*, we evaluated any inferred causal relationships between environmental drivers
422 and the proxy abundance of each taxon. Consistent causal influence on either taxon would show
423 variable but significant, effects across sub-monthly timescales. We found that both *Pseudo-*
424 *nitzschia spp.* and *Dinophysis spp.* are affected by water temperatures, wind speed, tidal height,
425 salinity, pH, and total biomass (chlorophyll). Previous studies across various regions, have
426 hypothesized correlative relationships between harmful algal blooms and these environmental
427 drivers (Almandoz et al. 2007; Sildever et al. 2019; Zhang et al. 2020; Lima et al. 2022). In our
428 study, convergent model predictions with increasing library size and prediction skills that greatly
429 exceeded the Pearson correlation coefficients between the individual time series indicate that there

430 are causal relationships beyond simple covariance. None of the environmental drivers we tested
431 overlapped with the previously quantified multivariate timescales. This implies that the dynamics
432 of both taxa in Narragansett Bay are subject to multiple context-dependent forces that interact with
433 each other. Successful prediction models for *Pseudo-nitzschia spp.* and *Dinophysis spp.* would
434 need to incorporate the specific local conditions under which the harmful algal blooms develop.
435 An ideal prediction model would attempt to combine data from relevant image properties and
436 environmental drivers for particular timescales of prediction. Different model combinations could
437 be optimized for forecasting at certain points during the 30-day prediction horizon. Table S3
438 explores the outcomes of some illustrative combined models for both *Pseudo-nitzschia spp.* and
439 *Dinophysis spp.* with a prediction horizon of 5, 10 and 15 days. Our results also indicate that there
440 can be lags between an environmental driver and the driven harmful algal taxa. Future studies that
441 attempt to predict the dynamics of *Pseudo-nitzschia spp.* and *Dinophysis spp.* might need to
442 characterize the causal timescales of their predictors.

443 As there has been rapid deployment of automated imaging systems for the early detection
444 of harmful algal bloom events (Campbell et al. 2010, 2013), our results suggest that there is
445 potential to use such data sources in advanced prediction models. Monitoring programs that
446 concurrently deploy other environmental and biogeochemical sensors might be able characterize
447 the relevant timescales of dynamics, and consequently, predict the magnitude and spatial
448 distribution of harmful algal events across broader regions. Although this study focuses on the
449 population dynamics of the harmful algal bloom-causing taxa, our prediction models could also
450 be coupled with other broad-scale ecosystem models to potentially include the impacts on higher
451 trophic levels and human health.

452

453 *Study limitations and future directions*

454 Although we have demonstrated the potential of using automated cell imaging data in
455 prediction models, there are several considerations involved that merit further discussion. First, as
456 our sampling location is fixed, the influence of different water masses and a lack of spatial
457 information can limit real-time projections of HAB abundance across entire regions. Future studies
458 should consider the concurrent deployment of multiple different systems to accurately map and
459 forecast spatial population patterns. Second, as *Pseudo-nitzschia spp.* is a chain-forming diatom,
460 the use of image concentration is not a measure of the actual abundance of the taxon within the
461 water column – there can be a variable number of cells within an image. Instead, image
462 concentration is a measure of our ability to detect and identify the taxa. Although detection
463 numbers are high when abundance is typically high (Figure S1), future studies might need to
464 accurately quantify the relationship between the in-situ abundance of chain-forming organisms
465 and their image detection. Third, the deployment and maintenance of IFCB systems may lead to
466 some irregularities and gaps within a long-term time series. Despite multiple years of data
467 collection, a large proportion of our daily-scale time series had to be approximated from existing
468 observations. Our approach requires sufficient long-term coverage for the development of
469 prediction models and future studies could evaluate alternative methods of data processing and
470 interpolation of missing observations. Fourth, the development of harmful algal blooms likely
471 depends on a suite of unknown environmental triggers (such as the nutrient regime, ambient light
472 levels, etc.). The identification of specific causal mechanisms would depend on careful
473 experimentation in laboratory studies, where confounding factors can be controlled, and additive
474 influence can be disentangled.

475

476 **Acknowledgments**

477 This study was supported by the Rhode Island Sea Grant (Grant number: NA22OAR4170123) and
478 NASA (Grant Supplement: NNX14AB80G). We would like to thank Heather Stoffel for the
479 provision of environmental data at the sampling location and Audrey Ciochetto for managing the
480 Imaging FlowCytobot data streams, instrument maintenance, image processing and databasing.
481 We would also like to acknowledge the efforts of Jessica Carney and Virginie Sonnet for IFCB
482 image annotations. The following students have helped maintain the IFCB over the years of
483 operation: Kyle Turner, Ian Lew, Cassandra Alexander, Christopher Jenkins, Jessica Carney,
484 Virginie Sonnet, and Somang Song.

485

486

487

488

489

490

491

492

493

494

495

496 **References**

- 497 Agarwal, V., C. C. James, C. E. Widdicombe, and A. D. Barton. 2021. Intraseasonal
498 predictability of natural phytoplankton population dynamics. *Ecol. Evol.* **11**: 15720–15739.
499 doi:10.1002/ece3.8234
- 500 Almandoz, G. O., M. E. Ferrario, G. A. Ferreyra, I. R. Schloss, J. L. Esteves, and F. E.
501 Papparazzo. 2007. The genus *Pseudo-nitzschia* (Bacillariophyceae) in continental shelf
502 waters of Argentina (Southwestern Atlantic Ocean, 38-55°S). *Harmful Algae* **6**: 93–103.
503 doi:10.1016/j.hal.2006.07.003
- 504 Anderson, C. R., M. R. P. Sapiano, M. B. K. Prasad, W. Long, P. J. Tango, C. W. Brown, and R.
505 Murtugudde. 2010. Predicting potentially toxigenic *Pseudo-nitzschia* blooms in the
506 Chesapeake Bay. *J. Mar. Syst.* **83**: 127–140. doi:10.1016/j.jmarsys.2010.04.003
- 507 Anderson, D. M., P. Hoagland, Y. Kaoru, and A. W. White. 2000. Estimated annual economic
508 impacts from harmful algal blooms (HABs) in the United States.
- 509 Annunziata, R., B. H. Mele, P. Marotta, and others. 2022. Trade-off between sex and growth in
510 diatoms: Molecular mechanisms and demographic implications. *Sci. Adv.* **8**: 1–17.
511 doi:10.1126/sciadv.abj9466
- 512 Ascioti, F. A., E. Beltrami, T. O. Carroll, and C. Wirick. 1993. Is there chaos in plankton
513 dynamics? *J. Plankton Res.* **15**: 603–617. doi:10.1093/plankt/15.6.603
- 514 Babin, M., J. J. Cullen, C. S. Roesler, and others. 2005. New approaches and technologies for
515 observing harmful algal blooms. *Oceanography* **18**: 210–227. doi:10.5670/oceanog.2005.55
- 516 Barton, A. D., A. J. Irwin, Z. V. Finkel, and C. A. Stock. 2016. Anthropogenic climate change

517 drives shift and shuffle in North Atlantic phytoplankton communities. *Proc. Natl. Acad. Sci.*
518 U. S. A. **113**: 2964–2969. doi:10.1073/pnas.1519080113

519 Bates, S. S., K. A. Hubbard, N. Lundholm, M. Montresor, and C. P. Leaw. 2018. Pseudo-
520 nitzschia, *Nitzschia*, and domoic acid: New research since 2011. *Harmful Algae* **79**: 3–43.
521 doi:10.1016/j.hal.2018.06.001

522 Béchard, A. 2020. Economics losses to fishery and seafood related businesses during harmful
523 algal blooms. *Fish. Res.* **230**: 105678. doi:10.1016/j.fishres.2020.105678

524 Berdalet, E., L. E. Fleming, R. Gowen, and others. 2016. Marine harmful algal blooms, human
525 health and wellbeing: Challenges and opportunities in the 21st century. *J. Mar. Biol. Assoc.*
526 United Kingdom **96**: 61–91. doi:10.1017/S0025315415001733

527 Berdjeb, L., A. Parada, D. M. Needham, and J. A. Fuhrman. 2018. Short-term dynamics and
528 interactions of marine protist communities during the spring-summer transition. *ISME J.* **12**:
529 1907–1917. doi:10.1038/s41396-018-0097-x

530 Blauw, A. N., E. Benincà, R. W. P. M. Laane, N. Greenwood, and J. Huisman. 2018.
531 Predictability and environmental drivers of chlorophyll fluctuations vary across different
532 time scales and regions of the North Sea. *Prog. Oceanogr.* **161**: 1–18.
533 doi:10.1016/j.pocean.2018.01.005

534 Brooks, B. W., J. M. Lazorchak, M. D. A. Howard, and others. 2016. Are harmful algal blooms
535 becoming the greatest inland water quality threat to public health and aquatic ecosystems?
536 *Environ. Toxicol. Chem.* **35**: 6–13. doi:10.1002/etc.3220

537 Brosnahan, M. L., L. Velo-Suárez, D. K. Ralston, and others. 2015. Rapid growth and concerted

538 sexual transitions by a bloom of the harmful dinoflagellate *Alexandrium fundyense*
539 (Dinophyceae). *Limnol. Oceanogr.* **60**: 2059–2078. doi:10.1002/lno.10155

540 Brown, A. R., M. Lilley, J. Shutler, C. Lowe, Y. Artioli, R. Torres, E. Berdalet, and C. R. Tyler.
541 2020. Assessing risks and mitigating impacts of harmful algal blooms on mariculture and
542 marine fisheries. *Rev. Aquac.* **12**: 1663–1688. doi:10.1111/raq.12403

543 Campbell, L., D. W. Henrichs, R. J. Olson, and H. M. Sosik. 2013. Continuous automated
544 imaging-in-flow cytometry for detection and early warning of *Karenia brevis* blooms in the
545 Gulf of Mexico. *Environ. Sci. Pollut. Res.* **20**: 6896–6902. doi:10.1007/s11356-012-1437-4

546 Campbell, L., R. J. Olson, H. M. Sosik, A. Abraham, D. W. Henrichs, C. J. Hyatt, and E. J.
547 Buskey. 2010. First harmful dinophysis (dinophyceae, dinophysiales) bloom in the U.S. is
548 revealed by automated imaging flow cytometry. *J. Phycol.* **46**: 66–75. doi:10.1111/j.1529-
549 8817.2009.00791.x

550 Carstensen, J., R. Klais, and J. E. Cloern. 2015. Phytoplankton blooms in estuarine and coastal
551 waters: Seasonal patterns and key species. *Estuar. Coast. Shelf Sci.* **162**: 98–109.
552 doi:10.1016/j.ecss.2015.05.005

553 Chang, C. W., M. Ushio, and C. hao Hsieh. 2017. Empirical dynamic modeling for beginners.
554 *Ecol. Res.* **32**: 785–796. doi:10.1007/s11284-017-1469-9

555 Chavez, F. P., J. Ryan, S. E. Lluch-Cota, and C. M. Niquen. 2003. Climate: From anchovies to
556 sardines and back: Multidecadal change in the Pacific Ocean. *Science (80-.)*. **299**: 217–221.
557 doi:10.1126/science.1075880

558 Cundell, A. M. 1973. Plastic materials accumulating in Narragansett Bay. *Mar. Pollut. Bull.* **4**:

559 187–188. doi:10.1016/0025-326X(73)90226-9

560 D’Alelio, D., A. Amato, A. Luedeking, and M. Montresor. 2009. Sexual and vegetative phases in
561 the planktonic diatom *Pseudo-nitzschia multistriata*. *Harmful Algae* **8**: 225–232.
562 doi:10.1016/j.hal.2008.05.004

563 Dalton, T., R. Thompson, and D. Jin. 2010. Mapping human dimensions in marine spatial
564 planning and management: An example from Narragansett Bay, Rhode Island. *Mar. Policy*
565 **34**: 309–319. doi:10.1016/j.marpol.2009.08.001

566 Dippner, J. W., L. Nguyen-Ngoc, H. Doan-Nhu, and A. Subramaniam. 2011. A model for the
567 prediction of harmful algae blooms in the Vietnamese upwelling area. *Harmful Algae* **10**:
568 606–611. doi:10.1016/j.hal.2011.04.012

569 Edwards, K. F., M. K. Thomas, C. A. Klausmeier, and E. Litchman. 2012. Allometric scaling
570 and taxonomic variation in nutrient utilization traits and maximum growth rate of
571 phytoplankton. *Limnol. Oceanogr.* **57**: 554–566. doi:10.4319/lo.2012.57.2.0554

572 Ellen, J. S., C. A. Graff, and M. D. Ohman. 2019. Improving plankton image classification using
573 context metadata. *Limnol. Oceanogr. Methods* **17**: 439–461. doi:10.1002/lom3.10324

574 Fleming, L. E., B. Kirkpatrick, L. C. Backer, and others. 2011. Review of Florida red tide and
575 human health effects. *Harmful Algae* **10**: 224–233. doi:10.1016/j.hal.2010.08.006

576 Franks, P. J. S. 2018. Recent Advances in Modelling of Harmful Algal Blooms, p. 359–377. *In*
577 *Global Ecology and Oceanography of Harmful Algal Blooms*.

578 Gobler, C. J. 2020. Climate Change and Harmful Algal Blooms: Insights and perspective.
579 *Harmful Algae* **91**: 101731. doi:10.1016/j.hal.2019.101731

580 Goldberg, E. D., E. Gamble, J. J. Griffin, and M. Koide. 1977. Pollution history of Narragansett
581 Bay as recorded in its sediments. *Estuar.Coastal Mar.Sci.* **5**: 549–558. doi:10.1016/0302-
582 3524(77)90101-3

583 Grattan, L. M., S. Holobaugh, and J. G. Morris. 2016. Harmful algal blooms and public health.
584 *Harmful Algae* **57**: 2–8. doi:10.1016/j.hal.2016.05.003

585 Herndon, R. W., and E. W. Sekatau. 1997. The Right to a Name: The Narragansett People and
586 Rhode Island Officials in the Revolutionary Era. *Ethnohistory* **44**: 433–462.

587 Jochens, A. E., T. C. Malone, R. P. Stumpf, and others. 2010. Integrated Ocean Observing
588 System in Support of Forecasting Harmful Algal Blooms. *Mar. Technol. Soc. J.* **44**: 99–121.
589 doi:10.4031/MTSJ.44.6.16

590 Karlson, B., P. Andersen, L. Arneborg, and others. 2021. Harmful algal blooms and their effects
591 in coastal seas of Northern Europe. *Harmful Algae* **102**: 101989.
592 doi:10.1016/j.hal.2021.101989

593 Kouakou, C. R. C., and T. G. Poder. 2019. Economic impact of harmful algal blooms on human
594 health: A systematic review. *J. Water Health* **17**: 499–516. doi:10.2166/wh.2019.064

595 Kraft, K., J. Seppälä, H. Hällfors, and others. 2021. First Application of IFCB High-Frequency
596 Imaging-in-Flow Cytometry to Investigate Bloom-Forming Filamentous Cyanobacteria in
597 the Baltic Sea. *Front. Mar. Sci.* **8**: 1–17. doi:10.3389/fmars.2021.594144

598 Kuhn, A. M., S. Dutkiewicz, O. Jahn, S. Clayton, T. A. Ryneerson, M. R. Mazloff, and A. D.
599 Barton. 2019. Temporal and Spatial Scales of Correlation in Marine Phytoplankton
600 Communities. *J. Geophys. Res. Ocean.* **124**: 9417–9438. doi:10.1029/2019JC015331

601 Lapucci, C., F. Maselli, G. Chini Zittelli, and others. 2022. Towards the Prediction of Favourable
602 Conditions for the Harmful Algal Bloom Onset of *Ostreopsis ovata* in the Ligurian Sea
603 Based on Satellite and Model Data. *J. Mar. Sci. Eng.* **10**: 461. doi:10.3390/jmse10040461

604 Lima, M. J., P. Relvas, and A. B. Barbosa. 2022. Variability patterns and phenology of harmful
605 phytoplankton blooms off southern Portugal: Looking for region-specific environmental
606 drivers and predictors. *Harmful Algae* **116**: 102254. doi:10.1016/j.hal.2022.102254

607 Ma, H., S. Leng, K. Aihara, W. Lin, and L. Cheni. 2018. Randomly distributed embedding
608 making short-term high-dimensional data predictable. *Proc. Natl. Acad. Sci. U. S. A.* **115**:
609 E9994–E10002. doi:10.1073/pnas.1802987115

610 McLeod, A.I. 2022. Kendall: Kendall Rank Correlation and Mann-Kendall Trend Test. R
611 package version 2.2.1. <https://CRAN.R-project.org/package=Kendall>

612 Moberg, E. A., and H. M. Sosik. 2012. Distance maps to estimate cell volume from two-
613 dimensional plankton images. *Limnol. Oceanogr. Methods* **10**: 278–288.
614 doi:10.4319/lom.2012.10.278

615 Moritz S, Bartz-Beielstein T. 2017. “imputeTS: Time Series Missing Value Imputation in R.”
616 *The R Journal* **9**(1), 207-218. doi: 10.32614/RJ-2017-009

617 Mosleh, M. A. A., H. Manssor, S. Malek, P. Milow, and A. Salleh. 2012. A preliminary study on
618 automated freshwater algae recognition and classification system. *BMC Bioinformatics* **13**
619 **Suppl 1**. doi:10.1186/1471-2105-13-s17-s25

620 Nixon, S. W., and R. W. Fulweiler. 2012. Ecological footprints and shadows in an urban estuary,
621 Narragansett Bay, RI (USA). *Reg. Environ. Chang.* **12**: 381–394. doi:10.1007/s10113-011-

622 0221-1

623 Oliver, H., W. G. Zhang, W. O. Smith, and others. 2021. Diatom Hotspots Driven by Western
624 Boundary Current Instability. *Geophys. Res. Lett.* **48**: 1–10. doi:10.1029/2020GL091943

625 Olson, R. J., and H. M. Sosik. 2007. A submersible imaging-in-flow instrument to analyze nano-
626 and microplankton: Imaging FlowCytobot. *Limnol. Oceanogr. Methods* **5**: 195–203.
627 doi:10.4319/lom.2007.5.195

628 Orenstein, E. C., K. M. Kenitz, P. L. D. Roberts, P. J. S. Franks, J. S. Jaffe, and A. D. Barton.
629 2020. Semi- and fully supervised quantification techniques to improve population estimates
630 from machine classifiers. *Limnol. Oceanogr. Methods* **18**: 739–753.
631 doi:10.1002/lom3.10399

632 Park, T. G., W. A. Lim, Y. T. Park, C. K. Lee, and H. J. Jeong. 2013. Economic impact,
633 management and mitigation of red tides in Korea. *Harmful Algae* **30**: S131–S143.
634 doi:10.1016/j.hal.2013.10.012

635 Park, Joseph, C. Smith, G. Sugihara and E. Deyle. 2022. rEDM: Empirical Dynamic Modeling
636 ('EDM'). R package version 1.13.0. <https://CRAN.R-project.org/package=rEDM>

637 Perretti, C. T., S. B. Munch, and G. Sugihara. 2013. Model-free forecasting outperforms the
638 correct mechanistic model for simulated and experimental data. *Proc. Natl. Acad. Sci. U. S.*
639 *A.* **110**: 5253–5257. doi:10.1073/pnas.1216076110

640 Pilson, M. E. Q. 1985. On the residence time of water in Narragansett Bay. *Estuaries* **8**: 2–14.
641 doi:10.2307/1352116

642 R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for

643 Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

644 Ralston, D. K., and S. K. Moore. 2020. Modeling harmful algal blooms in a changing climate.
645 *Harmful Algae* **91**: 1–29. doi:10.1016/j.hal.2019.101729

646 Saberski, E., A. K. Bock, R. Goodridge, V. Agarwal, T. Lorimer, S. A. Rifkin, and G. Sugihara.
647 2021. Networks of causal linkage between eigenmodes characterize behavioral dynamics of
648 *caenorhabditis elegans*. *PLoS Comput. Biol.* **17**: 1–15. doi:10.1371/journal.pcbi.1009329

649 Sakamoto, S., W. A. Lim, D. Lu, X. Dai, T. Orlova, and M. Iwataki. 2021. Harmful algal blooms
650 and associated fisheries damage in East Asia: Current status and trends in China, Japan,
651 Korea and Russia. *Harmful Algae* **102**: 101787. doi:10.1016/j.hal.2020.101787

652 Scavia, D., J. C. Field, D. F. Boesch, and others. 2002. Climate change impacts on U.S. Coastal
653 and Marine Ecosystems. *Estuaries* **25**: 149–164. doi:10.1007/BF02691304

654 Shanks, A. L., S. G. Morgan, J. MacMahan, A. J. H. M. Reniers, M. Reniers, J. Brown, A.
655 Fujimura, and C. Griesemer. 2014. Onshore transport of plankton by internal tides and
656 upwelling-relaxation events. *Mar. Ecol. Prog. Ser.* **502**: 39–51. doi:10.3354/meps10717

657 Sildever, S., Y. Kawakami, N. Kanno, H. Kasai, A. Shiimoto, S. Katakura, and S. Nagai. 2019.
658 Toxic HAB species from the Sea of Okhotsk detected by a metagenetic approach,
659 seasonality and environmental drivers. *Harmful Algae* **87**: 101631.
660 doi:10.1016/j.hal.2019.101631

661 Smayda, T. J. 1998. Patterns of variability characterizing marine phytoplankton, with examples
662 from Narragansett Bay. *ICES J. Mar. Sci.* **55**: 562–573. doi:10.1006/jmsc.1998.0385

663 Smith, R. B., B. Bass, D. Sawyer, D. Depew, and S. B. Watson. 2019. Estimating the economic

664 costs of algal blooms in the Canadian Lake Erie Basin. *Harmful Algae* **87**: 101624.
665 doi:10.1016/j.hal.2019.101624

666 Sonnet, V., L. Guidi, C. B. Mouw, G. Puggioni, and S. D. Ayata. 2022. Length, width, shape
667 regularity, and chain structure: time series analysis of phytoplankton morphology from
668 imagery. *Limnol. Oceanogr.* **67**: 1850–1864. doi:10.1002/lno.12171

669 Sosik, H. M., and R. J. Olson. 2007. Automated taxonomic classification of phytoplankton
670 sampled with imaging-in-flow cytometry. *Limnol. Oceanogr. Methods* **5**: 204–216.
671 doi:10.4319/lom.2007.5.204

672 Sterling, A. R., R. D. Kirk, M. J. Bertin, and others. 2022. Emerging harmful algal blooms
673 caused by distinct seasonal assemblages of a toxic diatom. *Limnol. Oceanogr.* 0–1.
674 doi:10.1002/lno.12189

675 Sugihara, G., and R. M. May. 1990. Nonlinear forecasting as a way of distinguishing chaos from
676 measurement error in time series. *Nature* **344**: 24–26.

677 Sugihara, G., R. May, H. Ye, C. H. Hsieh, E. Deyle, M. Fogarty, and S. Munch. 2012. Detecting
678 causality in complex ecosystems. *Science* (80-.). **338**: 496–500.
679 doi:10.1126/science.1227079

680 Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

681 Wilke, Claus O. 2020. *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*.

682 Xiao, X., S. Agustí, Y. Pan, Y. Yu, K. Li, J. Wu, and C. M. Duarte. 2019. Warming Amplifies
683 the Frequency of Harmful Algal Blooms with Eutrophication in Chinese Coastal Waters.
684 *Environ. Sci. Technol.* **53**: 13031–13041. doi:10.1021/acs.est.9b03726

685 Ye, H., and G. Sugihara. 2016. Information leverage in interconnected ecosystems: Overcoming
686 the curse of dimensionality. *Science* (80-.). **353**: 922–925. doi:10.1126/science.aag0863

687 Zhang, Y., J. Z. Su, Y. P. Su, H. Lin, Y. C. Xu, B. P. Barathan, W. N. Zheng, and K. G. Schulz.
688 2020. Spatial distribution of phytoplankton community composition and their correlations
689 with environmental drivers in taiwan strait of Southeast China. *Diversity* **12**: 1–15.
690 doi:10.3390/d12110433

691 Zheng, H., R. Wang, Z. Yu, N. Wang, Z. Gu, and B. Zheng. 2017. Automatic plankton image
692 classification combining multiple view features via multiple kernel learning. *BMC*
693 *Bioinformatics* **18**: 1–18. doi:10.1186/s12859-017-1954-8

694

695

696

697

698

699

700

701