Supporting Information for 'How predictable is plankton biogeography using statistical learning methods?'

L.R. Bardon[1], B.A. Ward[1], S. Dutkiewicz[2], and B.B. Cael[3]

[1]University of Southampton, UK

[2]Massachusetts Institute of Technology, Cambridge, MA, USA

[3]National Oceanography Centre, Southampton, UK

Contents of this file:

1. Figures S1 to S13
2. Tables S1 to S4
3. Access to Darwin Model output data
4. Access to code used to produce results

Introduction The analyses used to generate the following results are described in the Materials and Methods section of the main text.

Figures S1 to S6 is a complete set of figures, equivalent to Figure (1) in the main text, for all remaining plankton functional groups included in this study.

Figure S7 shows the true/false positives (TP, FP) and true/false negatives (TN, FN) from the GAMs predictions for all functional groups in the four different scenarios: GAMs trained on measurements-derived datapoints versus random datapoints, and spatial-only predictions (historical) versus end-of-century predictions (future). Note that the format of this figure is best understood as a bar plot visualisation of a confidence matrix, such that $TP + FP + TN + FN = 1$.

Figures S8 and S9 are the relative difference maps between Darwin Model "true" values and GAMs predictions for all functional groups, in the historical period (1987-2008) and by end-of-century (2079-2100).

Figure S10 is the partial dependence plots for all functional groups besides Coccolithophores, which are given in the main text. GAMs trained on data within 3586 Darwin surface cells, from the 1987-2008 period, and the 2079-2100 period. These demon-

strate how relationships between each predictor variable and the target variable (plankton biomass) change over time, for each functional group.

Figure S11 is equivalent to S10, but for 3586 randomly-distributed cells.

Figure S12 shows the correlations between predictors and functional group biomass within measurement-derived and randomly-distributed samples, of varying sizes, historical and future. Several methods are used for comparison: Distance Correlations, Pearson's Correlation Coefficient (ρ), calculated after data are transformed via natural log (ρ_{ln}), Spearman's Rank Correlation Coefficient (ρ_s.

Figure S13 shows the distribution of randomly-selected datapoints (the ocean observation analogue points are included in Figure 1a in the main text).

Table S1 Summary data for a range of sensitivity tests done on varied random sample sizes, from number of cells N=63 to N=11,557, and in predicting both historical and future biogeography.

Table S2 Summary of results for the predictions generated from the main 3586 cell testcases.

Table S3 Proportion of the functional group biomass measurements that were below the absence cut-off, for the 3586 cell training sets.

Table S4 Summary data for a range of sensitivity tests done on how varying presence-absence cut-off by a factory of ten in either direction affects results.

Finally, the raw Darwin Model output used for this work is available at `http://www.dataverse.harvard.edu/dataverse/`. The processed surface (top 10m) ocean ecosystem and physical data for the years 1991-2012 (which we consider equivalent to 1987-2008, for reasons explained in Methods and Materials) and 2079-2100, will also be made publicly available via Harvard Dataverse. Should the manuscript be accepted, DOIs for all associated code and data will be provided.

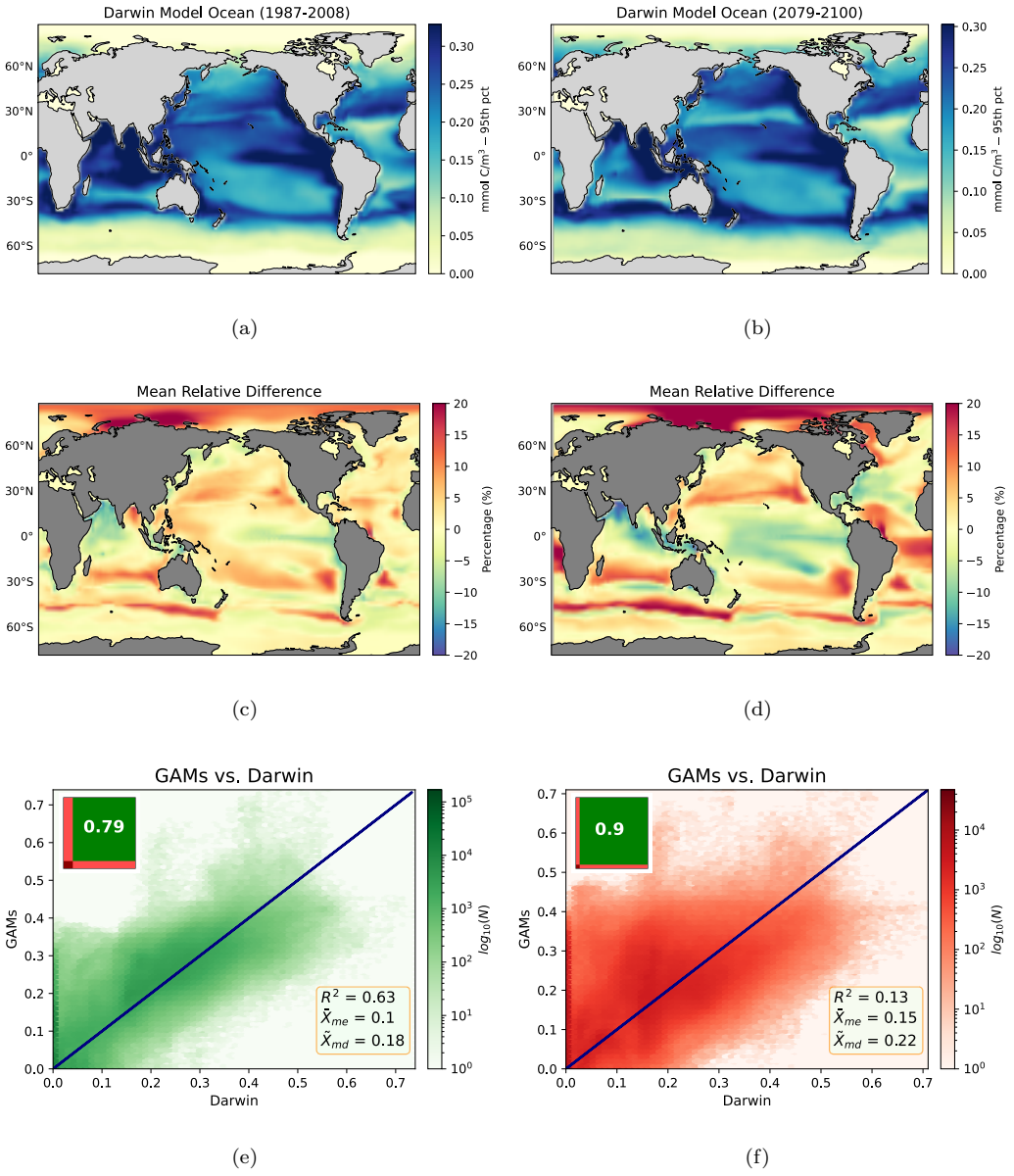


Figure S1: **(a)** Mean prokaryote surface biomass (1987 - 2008) from the Darwin model. **(b)** As per S1(a) for the years 2079 - 2100. **(c)** Relative (percent) difference between mean diatom surface biomass from the Darwin model and the GAMs (1987 - 2008) **(d)** As per S1(c) for the years 2079 - 2100. **(e)** Hexagonally binned scatterplot of 1987-2008 GAMs predictions vs 1987-2008 Darwin model. Colorbar shows log-scaled density of observations. *Top inset:* Fraction of data above the presence/absence threshold (10^{-5} mmol C/m^3)(green box), GAMs below threshold (left, light red), Darwin below threshold (bottom, light red), both below threshold (dark red). *Bottom inset:* The R^2, relative difference of the means ($\bar{X}_{me}$), and relative difference of the medians ($\tilde{X}_{md}$). **(f)** As per S1(e) but for 2079-2100.

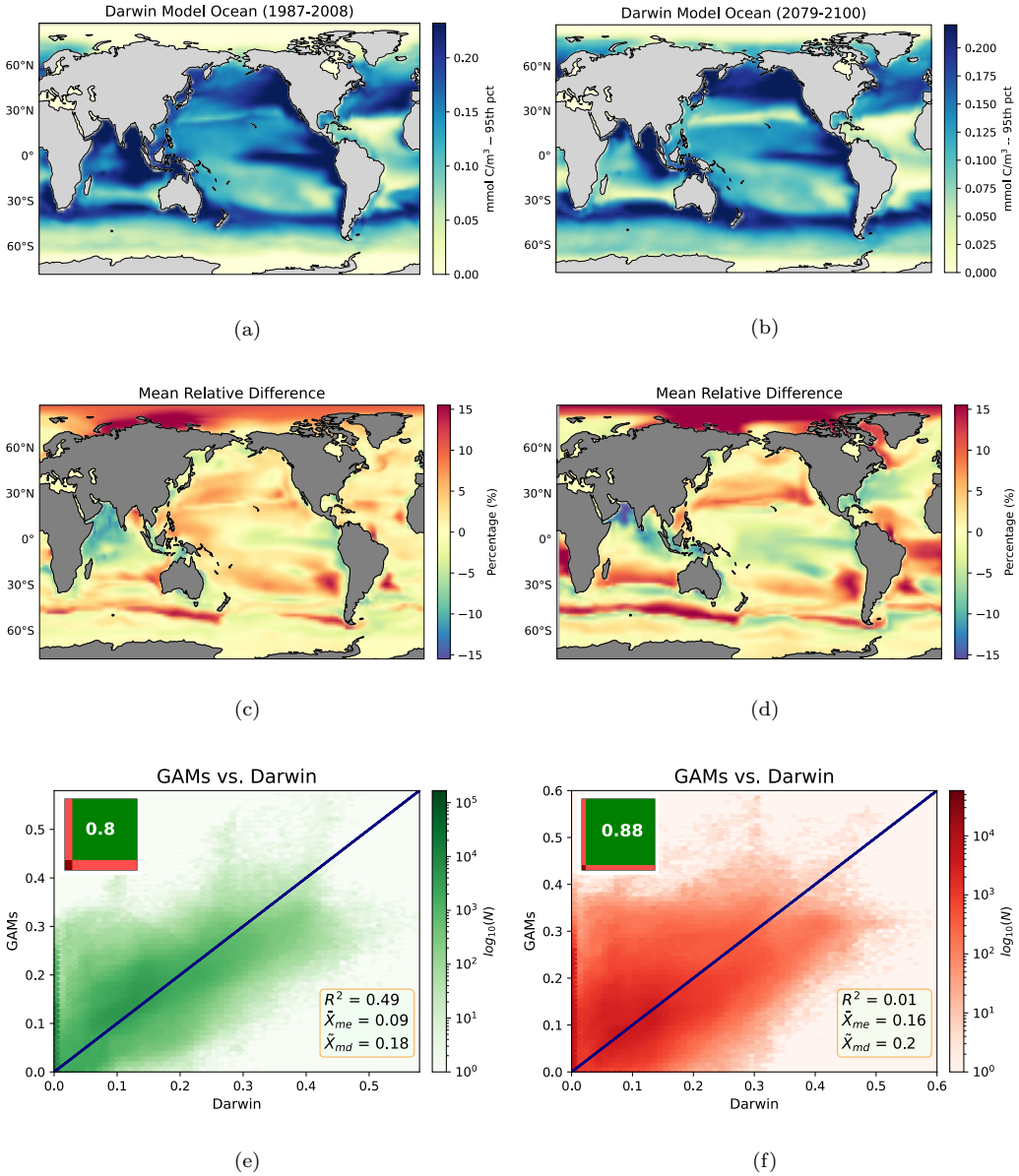


Figure S2: Picoeukaryotes, layout as per Figure S1.

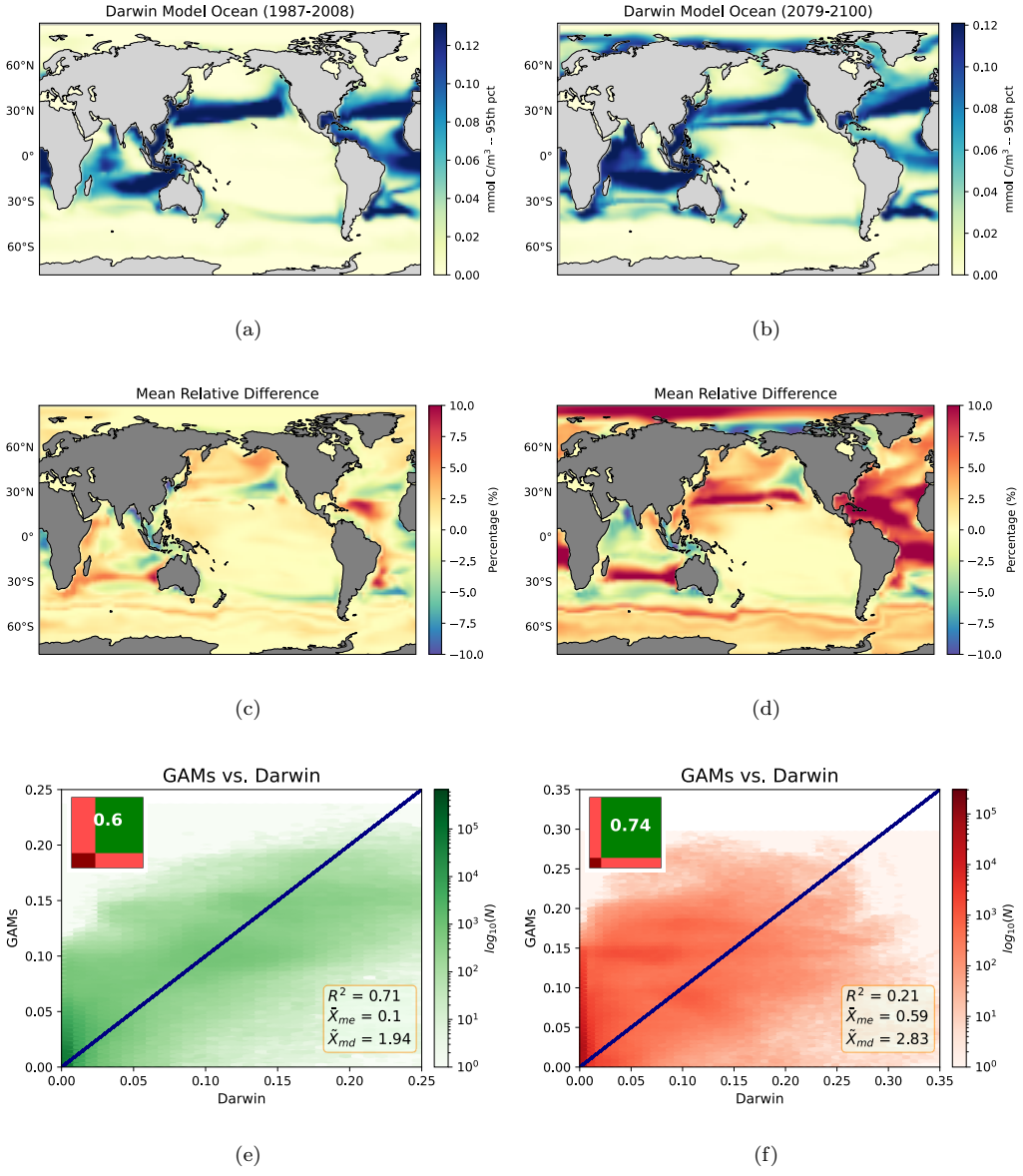


Figure S3: Diazotrophs, layout as per Figure S1.

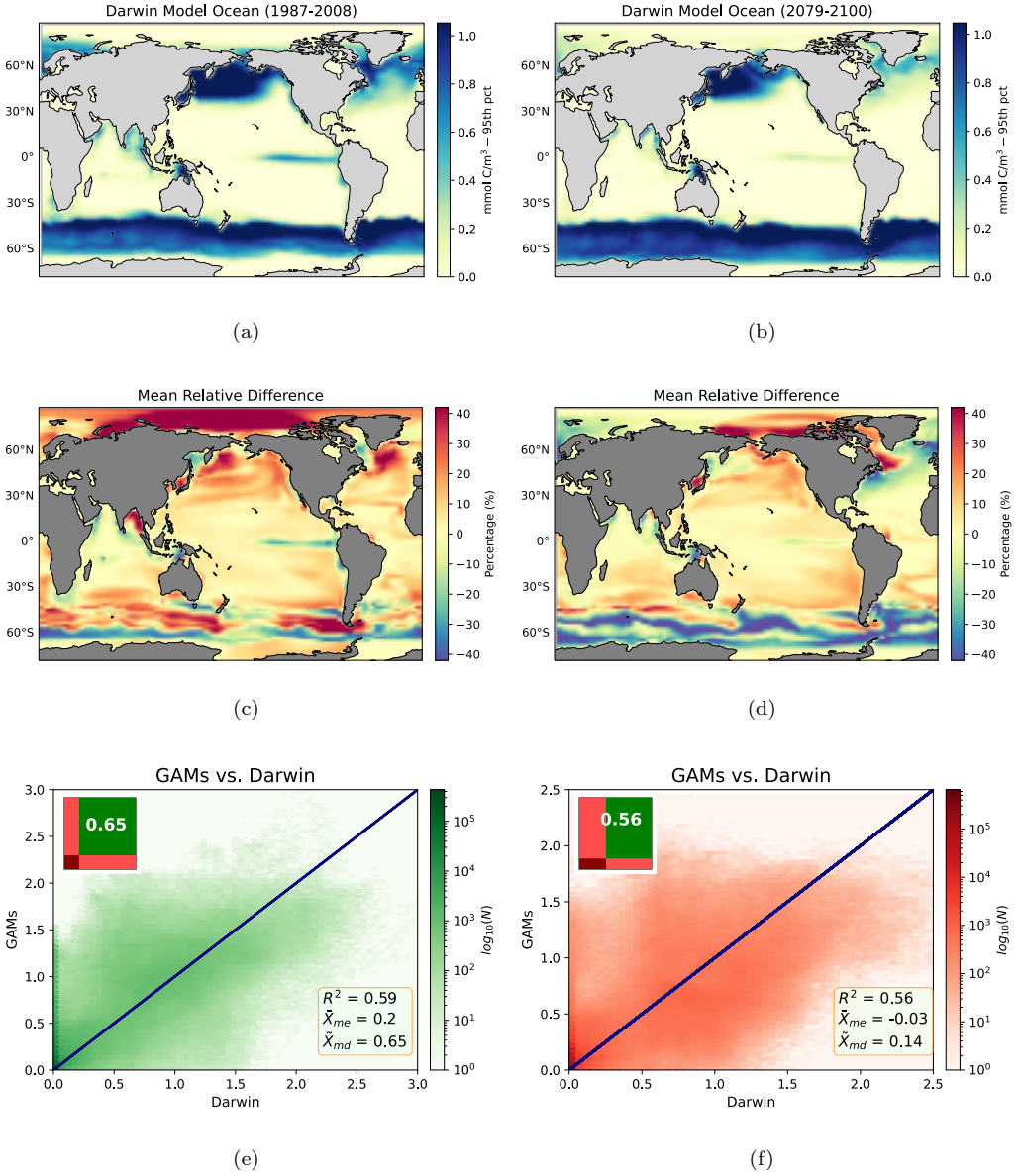


Figure S4: Diatoms, layout as per Figure S1.

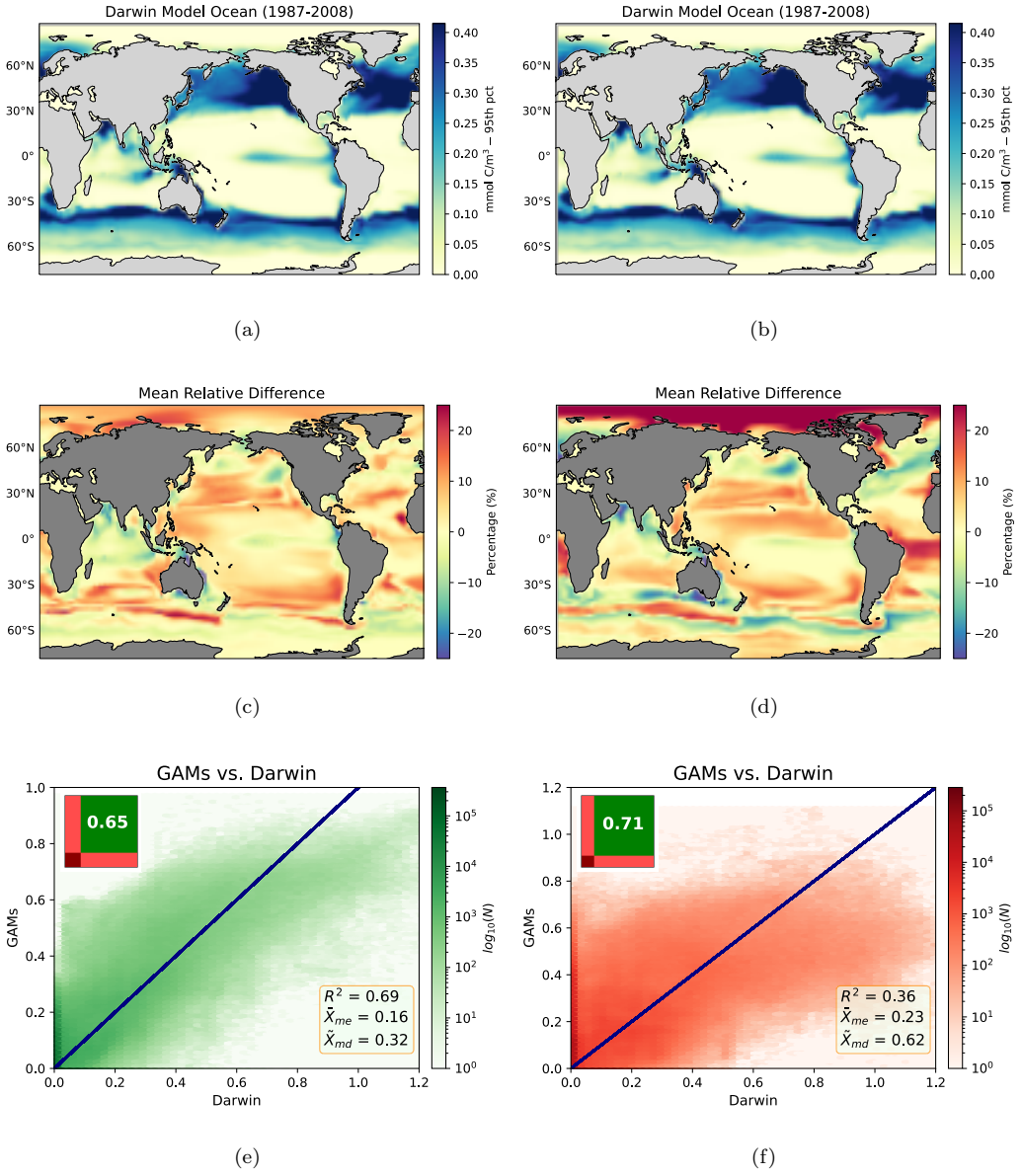


Figure S5: Mixotrophic dinoflagellates, layout as per Figure S1.

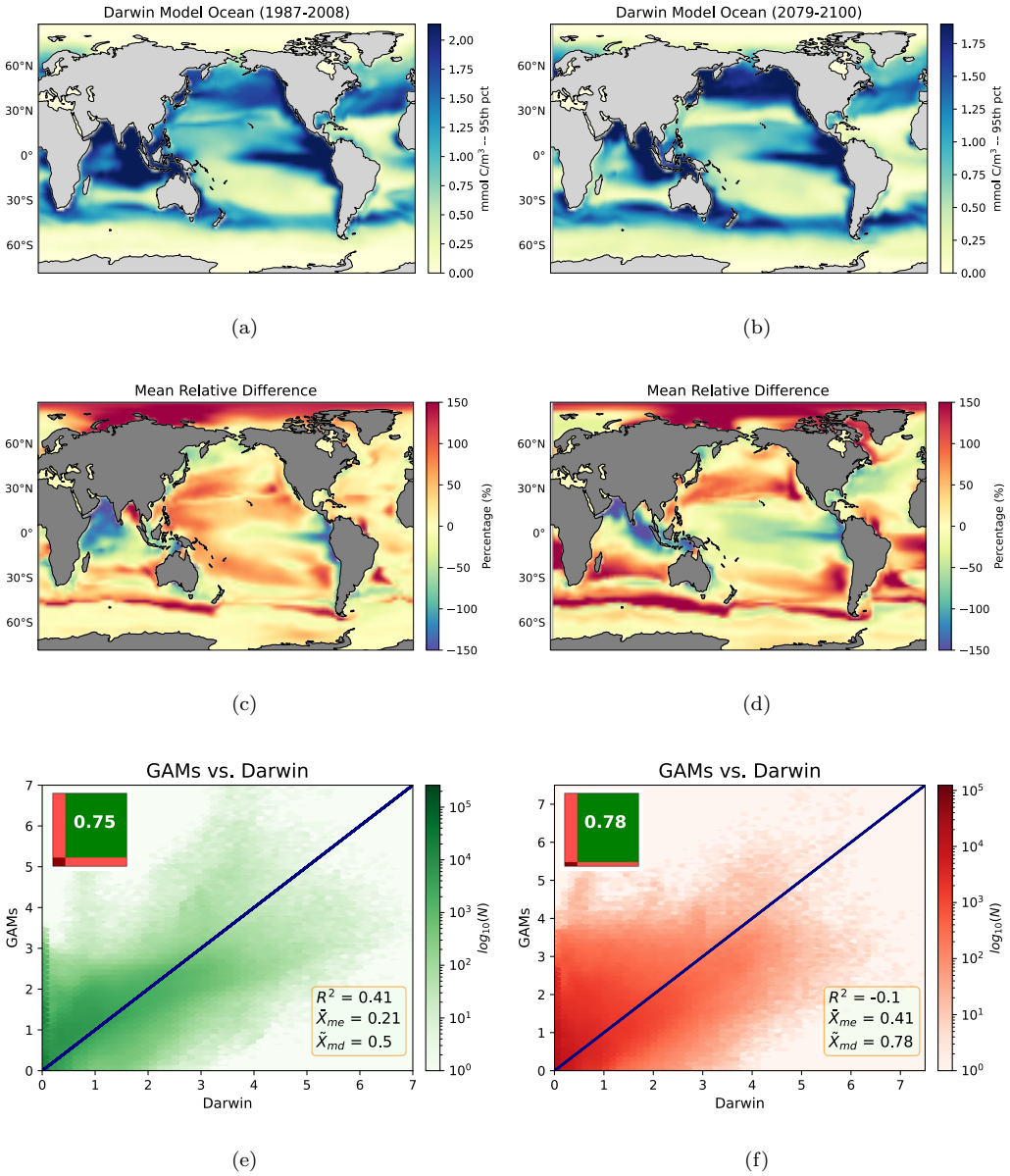


Figure S6: Zooplankton, layout as per Figure S1.

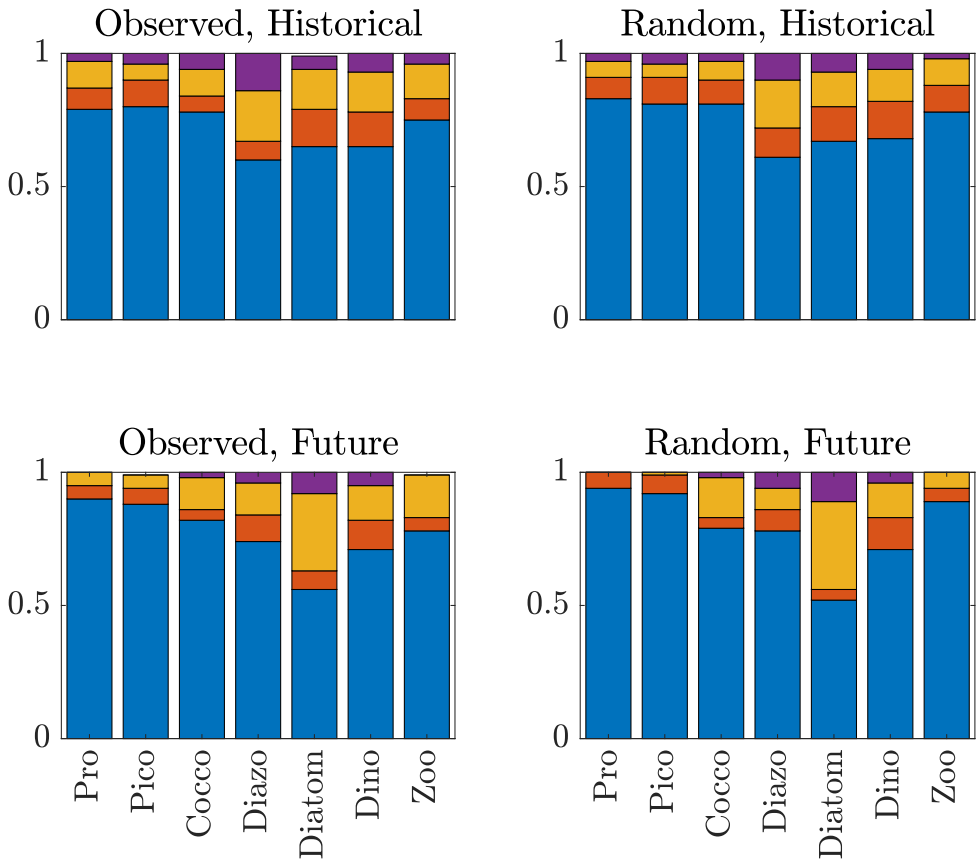
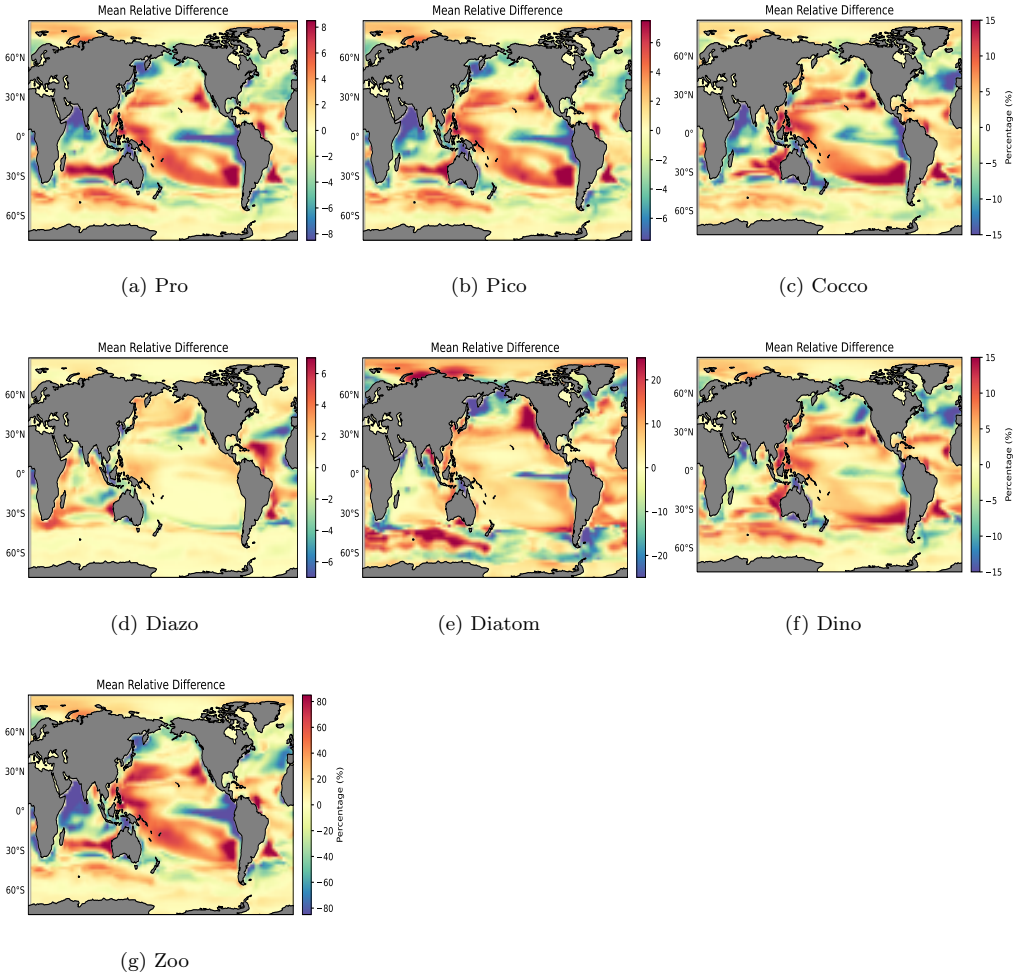


Figure S7: True positive (blue), false positive (orange), false negative (yellow), and true negative (purple), in terms of presence/absence above the cutoff biomass threshold, for each functional group for historical and future predictions, with observations-derived and random training sets. Note that the format of this figure is best understood as a bar plot visualisation of a confidence matrix, such that $TP + FP + TN + FN = 1$.

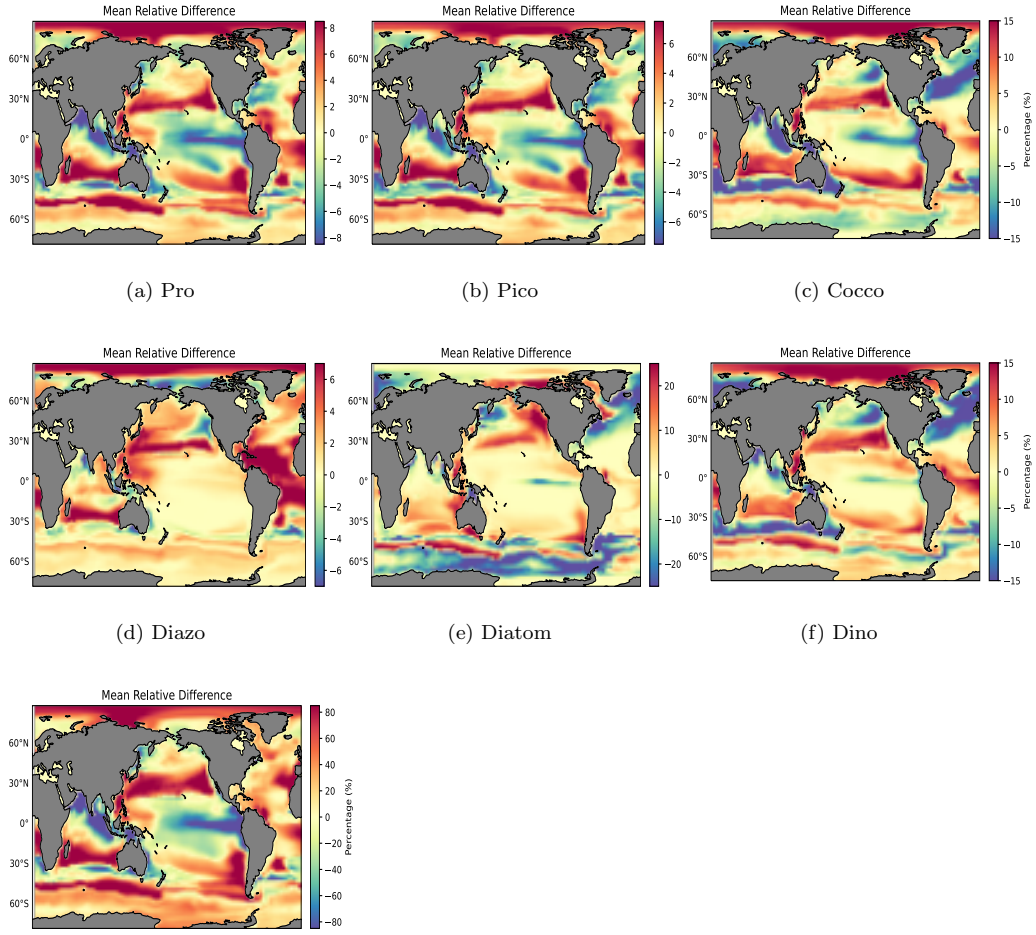


(a) Pro (b) Pico (c) Cocco

(d) Diazo (e) Diatom (f) Dino

(g) Zoo

Figure S8: Relative (%) difference in mean surface biomass (1987-2008) between the Darwin model and GAMs, where the latter has been trained on 3586 randomly-selected cells (S11(a)).



(a) Pro (b) Pico (c) Cocco

(d) Diazo (e) Diatom (f) Dino

(g) Zoo

Figure S9: Relative (%) difference in mean surface biomass (2079-2100) between the Darwin model and GAMs, where the latter has been trained on 3586 randomly-selected cells (S11(a))

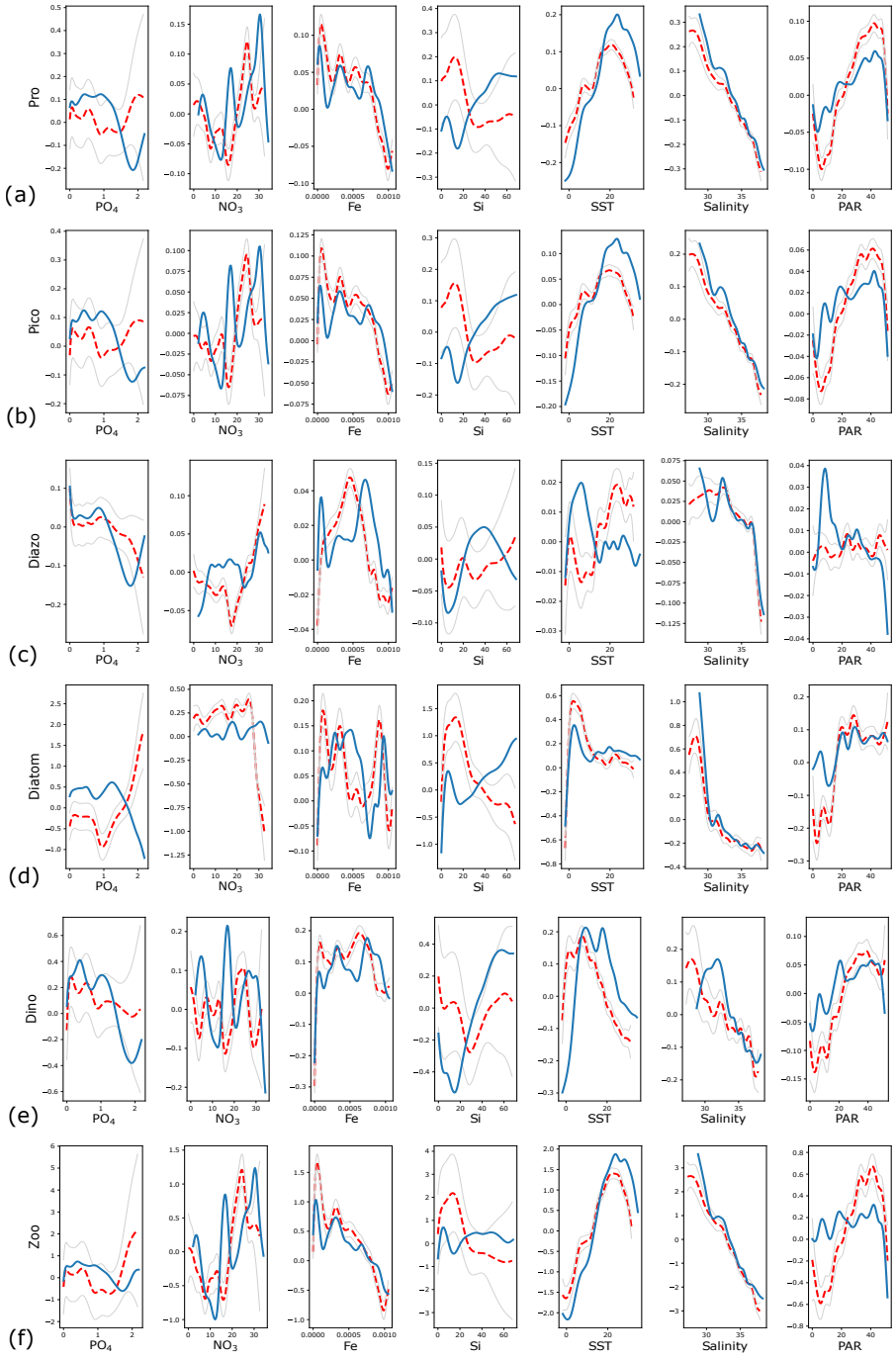


Figure S10: *Changing Relationships* (models trained at ocean measurement locations): Difference in partial dependence plots of plankton biomass for GAMs trained on data from 1987-2008 (dashed red line) and from 2079-2100 (blue line), for each predictor (PO_4, NO_3, Fe, Si in $mmol\ X/m^3$, SST in $°C$, SSS in PSU, PAR in $E/m^2/day$). From top to bottom: (a) Pro, (b) Pico, (c) Diazo, (d) Diatom, (e) Dino, (f) Zoo.

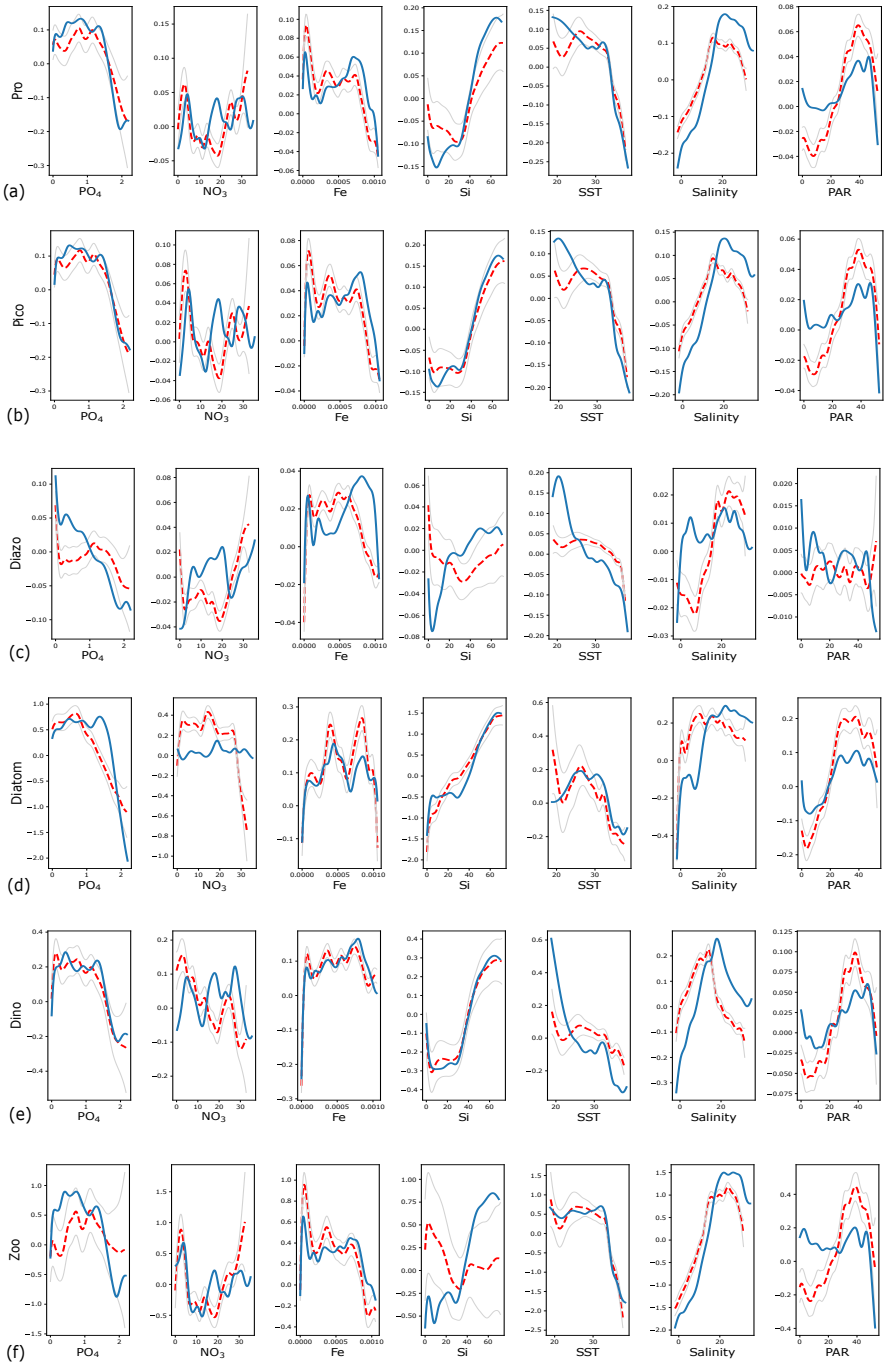


Figure S11: *Changing Relationships* (models trained at random locations): Difference in partial dependence plots of plankton biomass for GAMs trained on data from 1987-2008 (dashed red line) and from 2079-2100 (blue line), for each predictor (PO_4, NO_3, Fe, Si in $mmol\ X/m^3$, SST in $°C$, SSS in PSU, PAR in $E/m^2/day$). From top to bottom: (a) Pro, (b) Pico, (c) Diazo, (d) Diatom, (e) Dino, (f) Zoo.

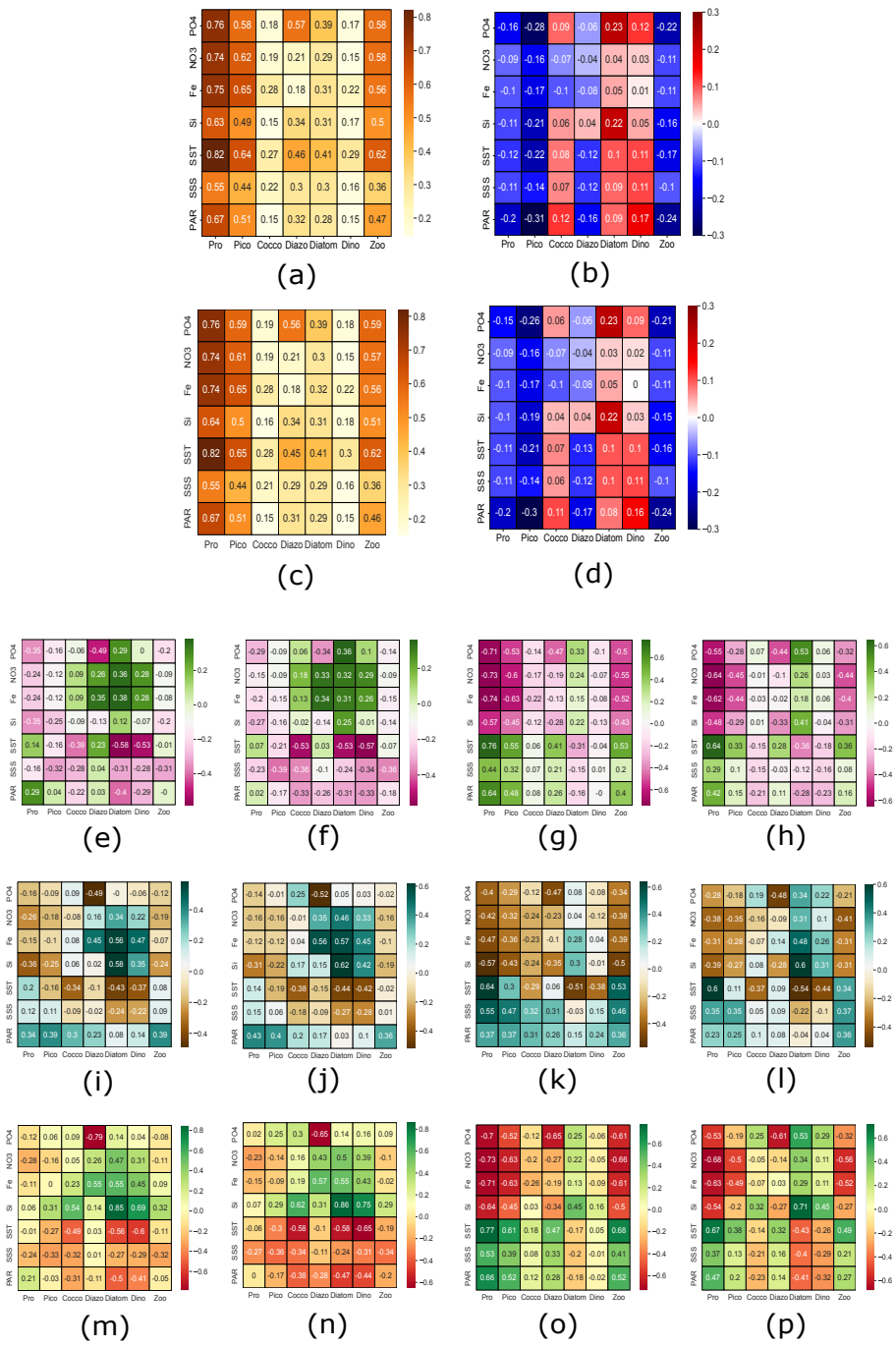
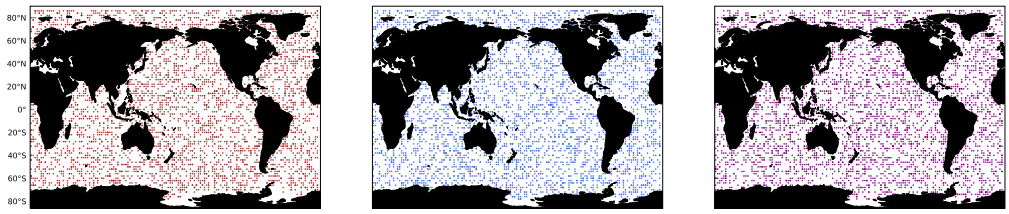


Figure S12: Correlation Matrices: (a) Distance Correlations, 12,894 randomly-sampled cells (1987-2008). (b) The difference between S12(a) and the same locations, 2079-2100. (c) As per S12(a) for 25,683 cells. (d) As per S12(b) for 25,683 cells. (e) Pearson's Correlation Coefficients at ocean measurement locations, 3586 cells, 1987-2008. (f) As per S12(e), 2079-2100. (g) As per S12(e) for random locations. (h) As per S12(g), 2079-2100. (i) Pearson's Correlation Coefficients of Log_{10} transformed data at ocean measurement locations, 1987-2100, 3586 cells. (j) As per S12(i), 2079-2100. (k) As per S12(i), for random sample. (l) As per S12(k), 2079-2100. (m) Spearman's Rank correlation, from measurements, 3586 cells, 1987-2008. (n) As per S12(m), 2079-2100. (o) As per S12(m), for random locations. (p) As per S12(o), 2079-2100.



(a) (b) (c)

Figure S13: Example of the sample distributions used for testing the effect of sample size on results. Shown are the three independent configurations of 3586 cell test cases.

	group & no. cells	GAMs absence	Darwin absence	Both presence	Sensitivity	Specificity	Balanced Acc.	Means Ratios	Medians Ratios	r-squared
1987-2008	cocco_63	78088	270362	1920351	0.90	0.59	0.74	-0.02	0.23	0.25
	cocco_130	264221	270362	1772383	0.90	0.32	0.61	-0.01	0.07	0.41
	cocco_262	242445	270362	1801160	0.91	0.37	0.64	-0.03	0.05	0.45
	cocco_387	237705	270362	1795115	0.90	0.34	0.62	-0.03	0.06	0.53
	cocco_506	235813	270362	1804463	0.91	0.37	0.64	-0.03	0.08	0.56
	cocco_642	249108	270362	1806183	0.91	0.41	0.66	-0.02	0.1	0.57
	cocco_951	214895	270362	1832190	0.91	0.44	0.68	-0.04	0.1	0.59
	cocco_1273	215849	270362	1826321	0.91	0.41	0.66	-0.04	0.1	0.6
	cocco_1914	218032	270362	1824663	0.91	0.41	0.66	-0.04	0.09	0.62
	cocco_2576	220313	270362	1827632	0.91	0.43	0.67	-0.02	0.09	0.63
	cocco_3189	237944	270362	1819557	0.92	0.44	0.68	-0.01	0.1	0.64
	cocco_3823	235662	270362	1816323	0.91	0.42	0.67	-0.02	0.1	0.65
	cocco_5105	238891	270362	1813581	0.91	0.42	0.67	-0.02	0.1	0.65
	cocco_6385	235153	270362	1814256	0.91	0.41	0.66	-0.01	0.1	0.66
	cocco_7694	235133	270362	1812618	0.91	0.40	0.66	-0.02	0.09	0.67
	cocco_8987	239231	270362	1808001	0.91	0.40	0.65	-0.01	0.1	0.68
	cocco_10278	238511	270362	1809469	0.91	0.40	0.66	-0.01	0.1	0.68
	cocco_11557	240296	270362	1811764	0.91	0.41	0.66	-0.01	0.09	0.68
2079-2100	cocco_63	24345	143357	2066494	0.94	0.46	0.70	-0.03	0.2	0.13
	cocco_130	254855	143357	1880659	0.96	0.22	0.59	-0.07	0.03	0.24
	cocco_262	409770	143357	1722453	0.95	0.13	0.54	-0.12	-0.03	0.3
	cocco_387	380840	143357	1750392	0.95	0.14	0.54	-0.05	0.07	0.38
	cocco_506	363121	143357	1770250	0.95	0.15	0.55	-0.06	0.05	0.42
	cocco_642	411339	143357	1731899	0.96	0.15	0.56	-0.08	0.03	0.43
	cocco_951	287034	143357	1847036	0.95	0.19	0.57	-0.07	0.02	0.42
	cocco_1273	288454	143357	1846416	0.95	0.19	0.57	-0.03	0.09	0.44
	cocco_1914	297245	143357	1834847	0.95	0.18	0.56	-0.03	0.1	0.45
	cocco_2576	325909	143357	1812718	0.96	0.18	0.57	-0.02	0.1	0.45
	cocco_3189	347584	143357	1792247	0.96	0.17	0.56	-0.03	0.07	0.47
	cocco_3823	362848	143357	1777020	0.96	0.17	0.56	-0.04	0.05	0.48
	cocco_5105	357788	143357	1781900	0.96	0.17	0.56	-0.06	0.02	0.49
	cocco_6385	344507	143357	1794904	0.96	0.17	0.56	-0.05	0.02	0.5
	cocco_7694	340198	143357	1799150	0.96	0.18	0.57	-0.06	0.01	0.52
	cocco_8987	337699	143357	1801904	0.96	0.18	0.57	-0.05	0.02	0.53
	cocco_10278	341563	143357	1798800	0.96	0.18	0.57	-0.05	0.01	0.53
	cocco_11557	338192	143357	1801954	0.96	0.18	0.57	-0.06	0	0.53

Table S1: Testing Sample Size: The results from a range of sensitivity tests exploring the effect of sample size on GAMs performance when trained on random sample distributions of varying cell size, as compared to the 'true' Darwin values.

		GAMs Absence	Darwin Absence	Both Presence	Sensitivity	Specificity	Bal. Acc.	Mean Ratio	Med. Ratio	r-squared
Obvs.	Pro	289643	234332	1750622	0.91	0.18	0.54	0.1	0.18	0.63
1987-2008	Pico	233020	315132	1769297	0.89	0.40	0.65	0.09	0.18	0.49
	Cocco	346786	270362	1724826	0.92	0.34	0.63	0.11	0.22	0.62
	Diazo	740089	465617	1328257	0.90	0.42	0.66	0.1	1.94	0.71
	Diatom	464219	434788	1434453	0.82	0.24	0.53	0.2	0.65	0.59
	Dino	483597	448636	1450406	0.83	0.33	0.58	0.16	0.32	0.69
	Zoo	377708	263891	1664984	0.90	0.22	0.56	0.21	0.5	0.41
Random	Pro	249868	234332	1820431	0.92	0.33	0.62	-0.03	0.09	0.75
1987-2008	Pico	239126	315132	1793730	0.90	0.52	0.71	-0.01	0.1	0.63
	Cocco	244842	270362	1807747	0.91	0.41	0.66	-0.01	0.12	0.64
	Diazo	664359	465617	1363055	0.87	0.41	0.64	0.04	1.62	0.73
	Diatom	463093	434788	1475506	0.84	0.32	0.58	0.02	0.35	0.78
	Dino	430112	448636	1505060	0.84	0.37	0.61	0.03	0.29	0.71
	Zoo	306205	263891	1737890	0.91	0.28	0.59	-0.03	0.21	0.59
Obvs.	Pro	118442	121404	2004870	0.95	0.18	0.57	0.15	0.22	0.13
2079-2100	Pico	124381	158466	1964180	0.94	0.19	0.56	0.16	0.2	0.01
	Cocco	308183	143357	1820087	0.95	0.16	0.55	0.17	0.36	0.45
	Diazo	361561	316968	1653024	0.89	0.30	0.59	0.59	2.83	0.21
	Diatom	823779	338479	1245015	0.89	0.22	0.56	-0.03	0.14	0.56
	Dino	392510	357224	1574498	0.86	0.26	0.56	0.23	0.62	0.36
	Zoo	378572	124043	1737890	0.94	0.05	0.49	0.41	0.78	-0.1
Random	Pro	36486	121404	2071146	0.95	0.16	0.56	0.05	0.15	0.57
2079-2100	Pico	41648	158466	2027800	0.93	0.12	0.52	0.03	0.11	0.42
	Cocco	344182	143357	1794021	0.95	0.17	0.56	0.01	0.16	0.5
	Diazo	284328	316968	1747900	0.90	0.44	0.67	0.39	2.19	0.43
	Diatom	936357	338479	1203113	0.94	0.27	0.60	0.02	0.39	0.74
	Dino	433499	357224	1551800	0.87	0.28	0.57	0.05	0.47	0.52
	Zoo	134244	124043	1965283	0.94	0.00	0.47	0.14	0.47	0.36

Table S2: Summary of results for the predictions generated from the main 3586 cell test cases. Note that the absence values are out of a total of 2,223,085 data points, and that 'Both presence' refers to where both GAMs and Darwin predict presence.

	Pro	Pico	Cocco	Diazo	Diatom	Dino	Zoo
Obvs.	31	43	44	368	628	544	42
Random	309	438	359	680	661	678	380

Table S3: Proportion of the functional group biomass measurements that were below the absence cut-off, for the 3586 cell training sets.

Cut -off	R^2		$\bar{X}_{me}$		$\tilde{X}_{md}$		Darwin removed		GAMs removed		
	1987 -2008	*2079 -2100*	*1987 -2008*	*2079 -2100*	*1987 -2008*	*2079 -2100*	*1987 -2008*	*2079 -2100*	*1987 -2008*	*2079 -2100*	
10^{-6}	0.64	0.13	0.16	0.17	0.22	0.23	0.15	0.03	0.13	0.05	*Pro*
	0.49	0	0.09	0.17	0.19	0.22	0.12	0.06	0.11	0.06	*Pico*
	0.62	0.45	0.11	0.19	0.23	0.41	0.1	0.04	0.16	0.14	*Cocco*
	0.72	0.21	0.11	0.63	2.13	3.59	0.16	0.1	0.33	0.16	*Diazo*
	0.59	0.57	0.22	-0.02	1.14	0.38	0.14	0.09	0.21	0.37	*Diatom*
	0.69	0.35	0.18	0.27	0.38	0.82	0.16	0.11	0.22	0.18	*Dino*
	0.39	-0.12	0.24	0.44	0.55	0.85	0.1	0.04	0.17	0.17	*Zoo*
10^{-5}	0.63	0.13	0.1	0.15	0.18	0.22	0.11	0.05	0.13	0.05	*Pro*
	0.49	0.01	0.09	0.16	0.18	0.2	0.14	0.07	0.10	0.06	*Pico*
	0.62	0.45	0.11	0.17	0.22	0.36	0.12	0.06	0.16	0.14	*Cocco*
	0.71	0.21	0.1	0.59	1.94	2.83	0.21	0.14	0.33	0.16	*Diazo*
	0.59	0.56	0.2	-0.03	0.65	0.14	0.20	0.15	0.21	0.37	*Diatom*
	0.69	0.36	0.16	0.23	0.32	0.62	0.20	0.16	0.22	0.18	*Dino*
	0.41	-0.1	0.21	0.41	0.5	0.78	0.12	0.06	0.17	0.17	*Zoo*
10^{-4}	0.62	0.13	0.09	0.14	0.16	0.21	0.15	0.05	0.13	0.05	*Pro*
	0.49	0.02	0.08	0.14	0.17	0.19	0.16	0.07	0.10	0.05	*Pico*
	0.62	0.46	0.1	0.15	0.21	0.3	0.15	0.06	0.16	0.14	*Cocco*
	0.7	0.21	0.09	0.54	1.59	2.03	0.28	0.14	0.33	0.16	*Diazo*
	0.59	0.53	0.17	-0.04	0.32	-0.1	0.26	0.15	0.21	0.37	*Diatom*
	0.68	0.36	0.15	0.18	0.26	0.42	0.25	0.16	0.22	0.18	*Dino*
	0.43	-0.09	0.19	0.38	0.46	0.72	0.14	0.06	0.17	0.17	*Zoo*

Table S4: Testing Cutoff Value Sensitivity: The results of a suite of tests designed to assess the effect of varying the absence cut-off value from by a factor of ten on either side of the $1e^{-5}$ value used for the main body of results.