Supplementary materials for *The global distribution and climate resilience of marine heterotrophic prokaryotes*

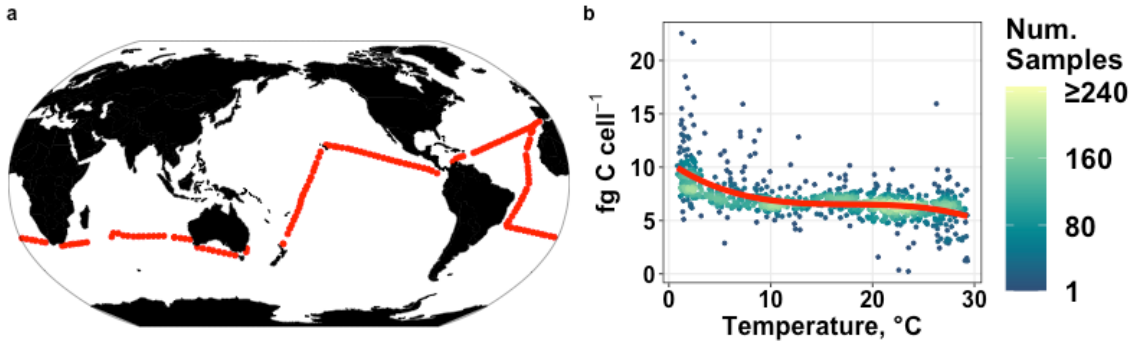

Figure S1 | Prokaryotic cell-specific carbon statistical model. a) Distribution of 1,087 *in situ* unique samples of prokaryotic cell carbon used in this study. Prokaryotic cell carbon (fg C cell^{-1}) as a function of b) temperature (°C) in the final parametric model. The red line is the fitted response, with residuals shown as dots coloured by the number of samples (Num. Samples). Source data are provided as a Source Data file.

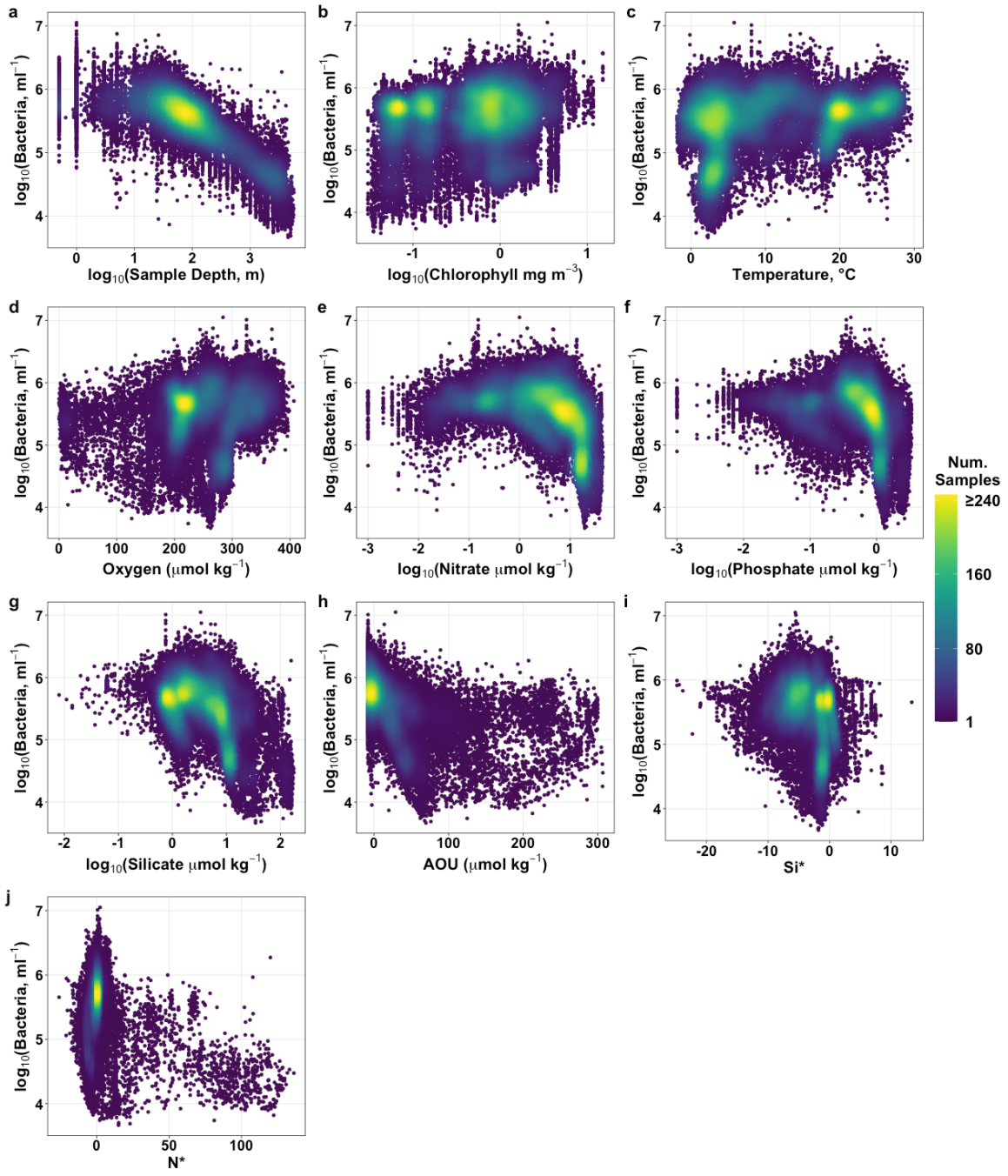

Figure S2 | Scatter plots of $\log_{10}$ prokaryotic abundance (mL^{-1}) against a) $\log_{10}$ sample depth; b) $\log_{10}$ surface chlorophyll; c) temperature; d) oxygen; e) $\log_{10}$ nitrate; f) $\log_{10}$ phosphate; g) $\log_{10}$ silicate; h) apparent oxygen utilisation **(**AOU); i) Si* and j) N*. Si* and N* are tracer variables, respectively measuring excess dissolved inorganic nitrogen relative to the Redfield ratio and the ratio of silicate to nitrate. Source data are provided as a Source Data file.

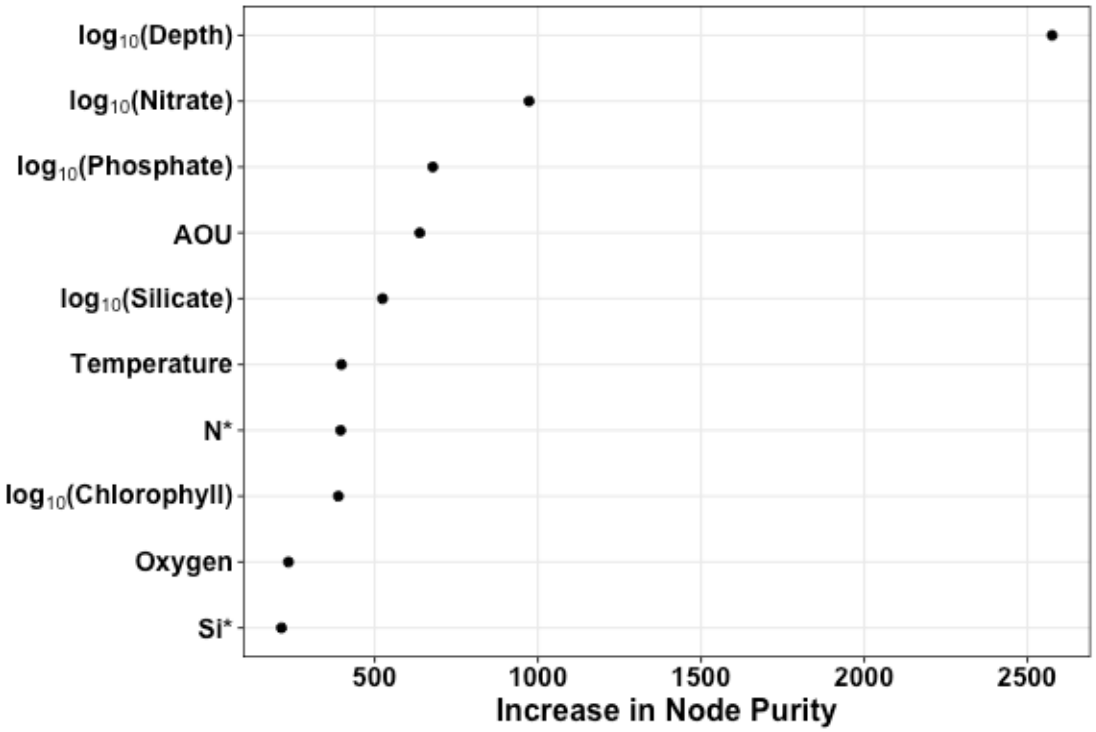

Figure S3 | Example of one of the five random forest variable importance rankings for predicting $\log_{10}$ prokaryote abundance (mL^{-1}). Increase in Node Purity measures the total decrease in error (given by the residual sum of squares) from splitting on that variable, averaged over all trees in the random forest model. AOU is apparent oxygen utilization. N* and Si* are tracer variables, respectively measuring excess dissolved inorganic nitrogen relative to the Redfield ratio and the ratio of silicate to nitrate. Source data are provided as a Source Data file.

Table S1. Comparison of different predictor variable combinations fitted with penalized regression splines in a generalised additive model for log_{10} prokaryote abundance. Root mean-square error (RMSE), mean absolute error (MAE) and deviance explained are reported for each model. Our selected candidate variable set for the final parametric equation is in bold. AOU is apparent oxygen utilization. N* and Si* are tracer variables, respectively measuring excess dissolved inorganic nitrogen relative to the Redfield ratio and the ratio of silicate to nitrate.

Predictor Variables	**RMSE**	**MAE**	**Deviance explained**
log_{10} depth	<0.001	0.187	72.2%
log_{10} nitrate	0.003	0.278	42.8%
log_{10} phosphate	0.003	0.287	38.3%
apparent oxygen utilization (AOU)	<0.001	0.285	41.4%
log_{10} silicate	0.004	0.288	37.5%
temperature	0.001	0.329	18.3%
log_{10} chlorophyll	0.002	0.364	4.7%
N*	0.002	0.339	14.6%
oxygen	<0.001	0.346	11.5%
Si*	<0.001	0.297	30.6%
log_{10} depth + log_{10} nitrate + AOU + temperature + log_{10} chlorophyll	**0.001**	**0.165**	**77.9%**
log_{10} depth + log_{10} nitrate + log_{10} phosphate + AOU + log_{10} silicate + temperature + log_{10} chlorophyll + N* + oxygen + Si*	0.001	0.160	78.9%

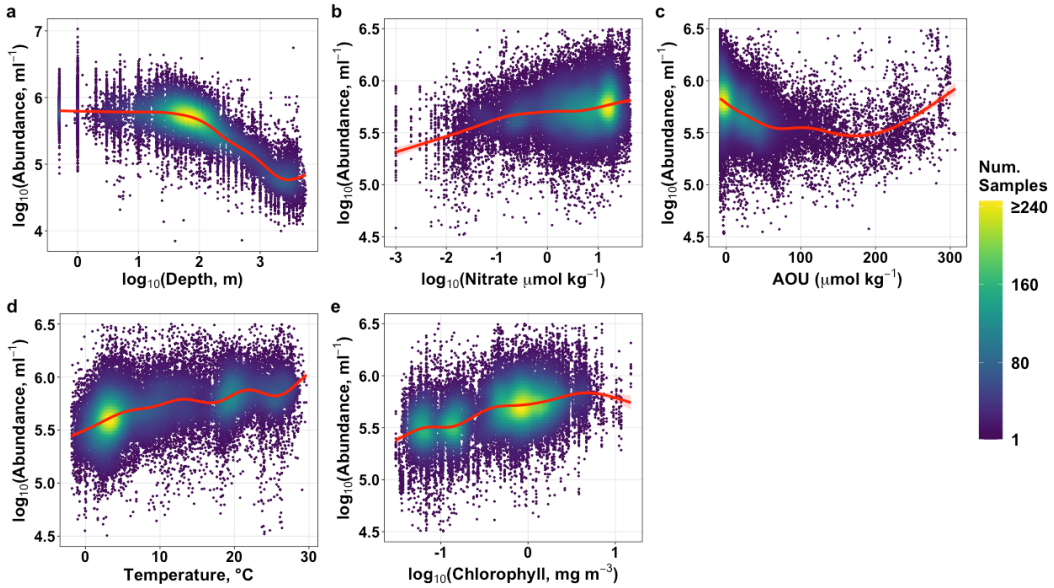

Figure S4 | Prokaryotic abundance (mL^{-1}) as a function of a) log_{10} depth, b) log_{10} nitrate, c) apparent oxygen utilization (AOU), d) temperature and e) log_{10} chlorophyll *a* in the generalised additive model. The red line and shading are the fitted response for each environmental variable, with partial residuals shown as dots coloured by the number of samples (Num. Samples). Source data are provided as a Source Data file.

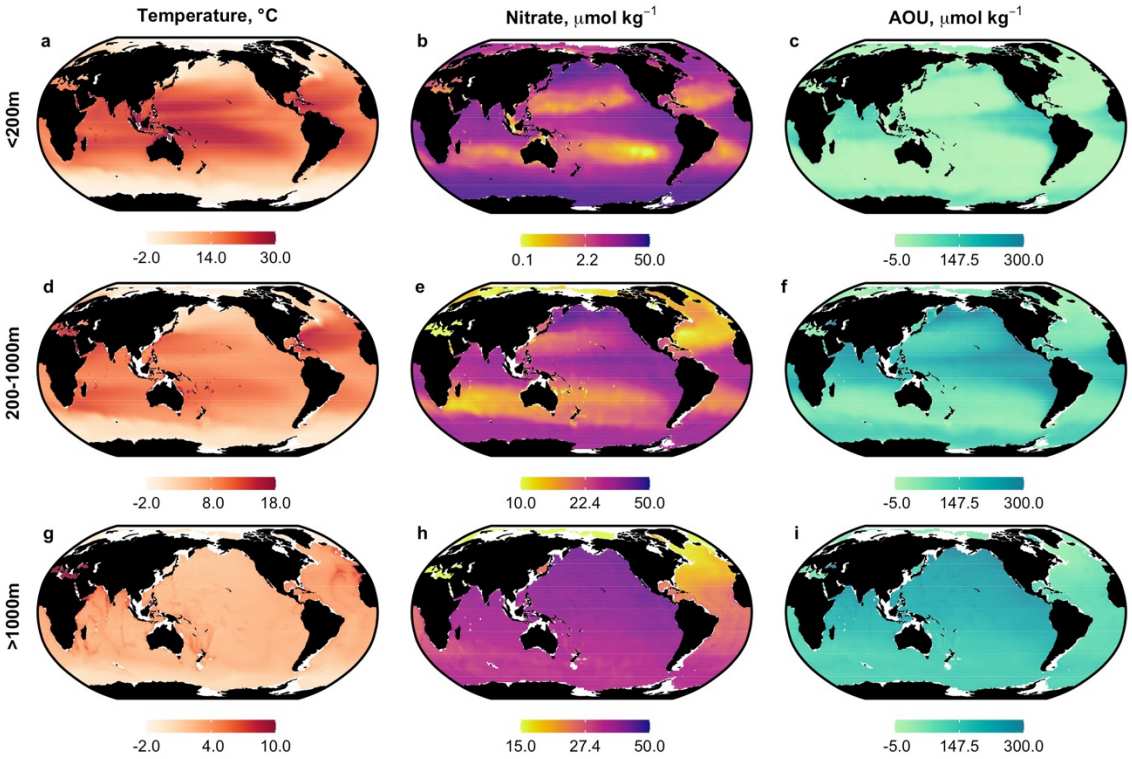

Figure S5 | Contemporary global distribution of prokaryotic environmental drivers. Temperature (a, d, g); (b, e, h) nitrate, and (c, f, i) apparent oxygen utilization (AOU) in the (a-c) top 200 m (epipelagic); (d-f) 200-1000 m (mesopelagic); and (g-i) >1000 m (bathypelagic), from the 2018 World Ocean Atlas. Note change of scale in temperature. Source data are provided as a Source Data file.

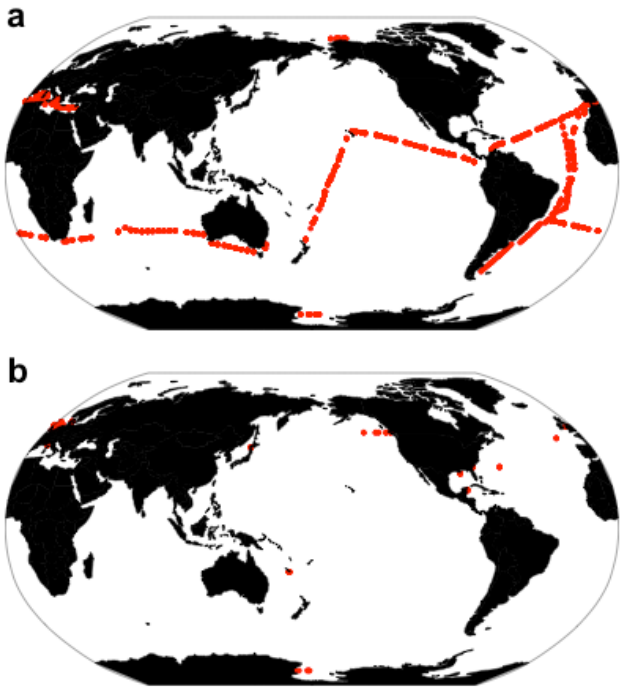

Figure S6 | Global distribution of observations to calculate prokaryotic respiration. Distribution of a) 2,092 *in situ* unique samples of heterotrophic prokaryotic specific-production rates and b) 305 *in situ* unique samples of prokaryotic growth efficiency used in this study to calculate prokaryotic respiration in the top 200 m. Source data are provided as a Source Data file.

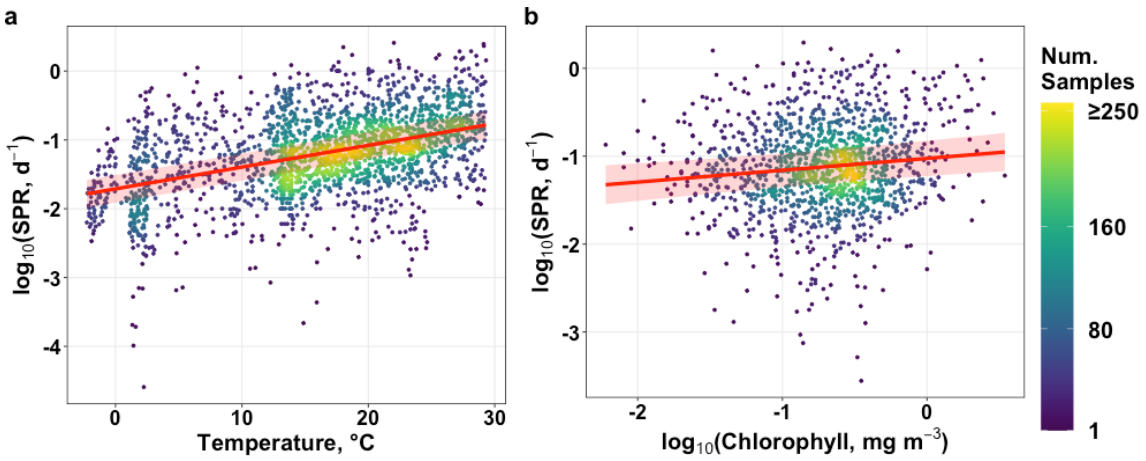

Figure S7 | Prokaryote specific-production rates (SPR; day^{-1}) statistical model. Main effects of a) temperature and b) $\log_{10}$ chlorophyll *a* in a linear mixed-effects model (which also includes data source as a random effect). The red line is the fitted response for each environmental variable, with partial residuals shown as dots coloured by the number of

observations. Red shaded regions in each figure shows the 95% confidence interval for our SPR statistical model. Source data are provided as a Source Data file.

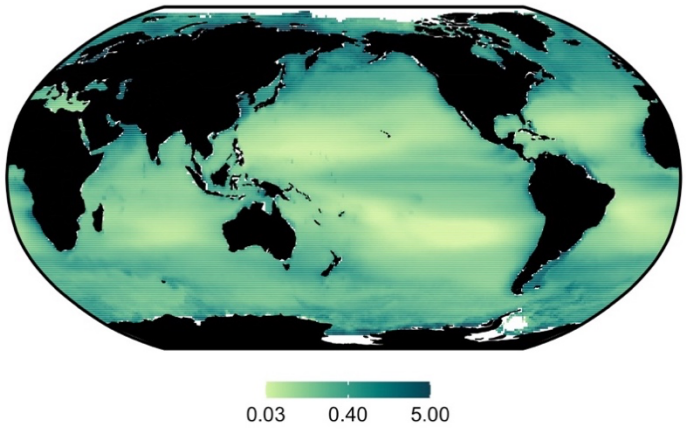

Figure S8 | Contemporary global distribution of chlorophyll *a*. The concentration of surface chlorophyll *a* (mg m^3) across the global ocean, from MODIS-Aqua averaged across 2002-2016. Source data are provided as a Source Data file.

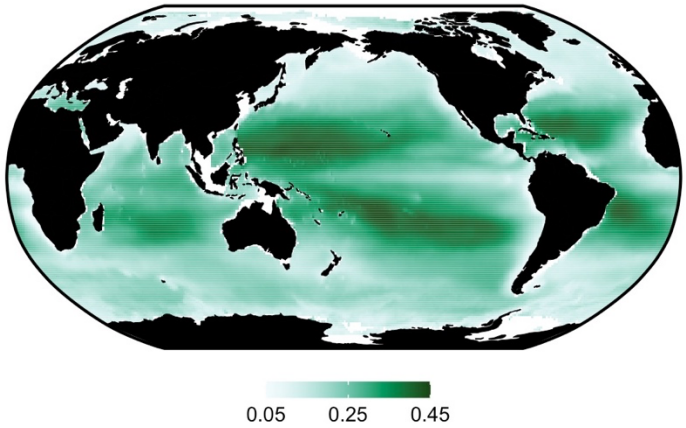

Figure S9 | Composition heterotrophic biomass composition. The ratio of prokaryotic biomass to total heterotrophic (prokaryote, zooplankton and fish) biomass in the top 200 m across the world ocean. Zooplankton and fish biomass estimates were taken from Hatton et al. (2021)[3]. Source data are provided as a Source Data file.

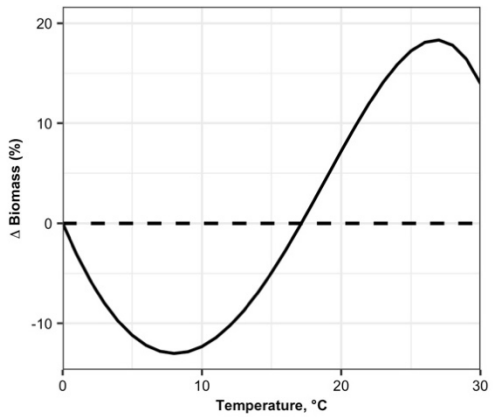

Figure S10 | Temperature impacts on prokaryotic biomass. Change (%) in prokaryotic biomass from 0 to 30°C, derived from the abundance and cell-specific models, normalised to biomass at 0°C and holding all other environmental variables constant. The horizontal dashed line indicates where there is no change in biomass, compared to biomass at 0°C. Temperature-driven declines in total biomass in <8°C and >27°C waters are caused by decreases in cell-specific carbon content (Fig. S1b) exceeding temperature-driven increases in prokaryotic abundance (Fig. 1e) in these temperature ranges. Source data are provided as a Source Data file.

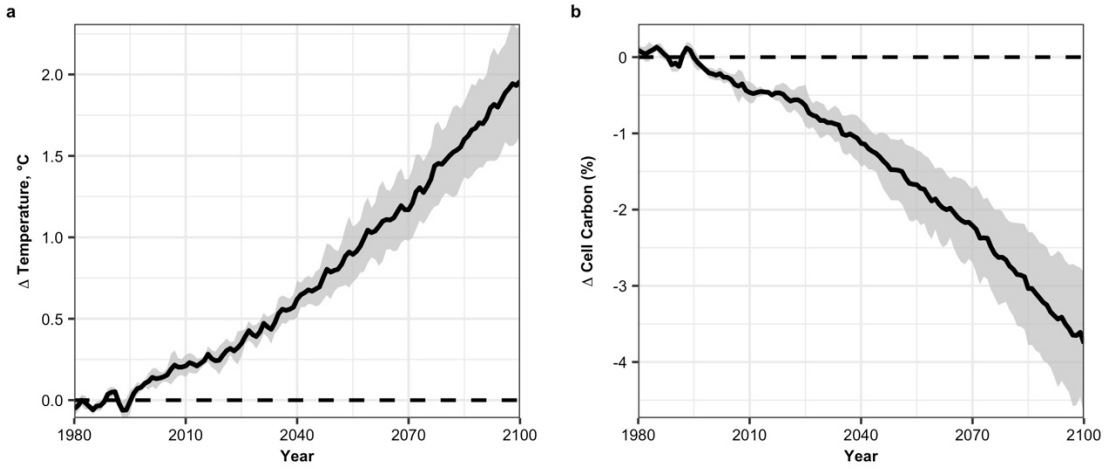

Figure S11 | Impacts of climate change on global drivers of prokaryotic biomass. Change in mean global a) temperature °C; b) prokaryotic cell-specific carbon (%) in top 200 m under emission scenario Shared Socioeconomic Pathway 3-7.0, compared to 1980-2000 levels. In both figures, solid lines represent the ensemble mean change and shared areas are the standard deviation from separate simulations, each forced by one of four earth-system models. Source data are provided as a Source Data file.

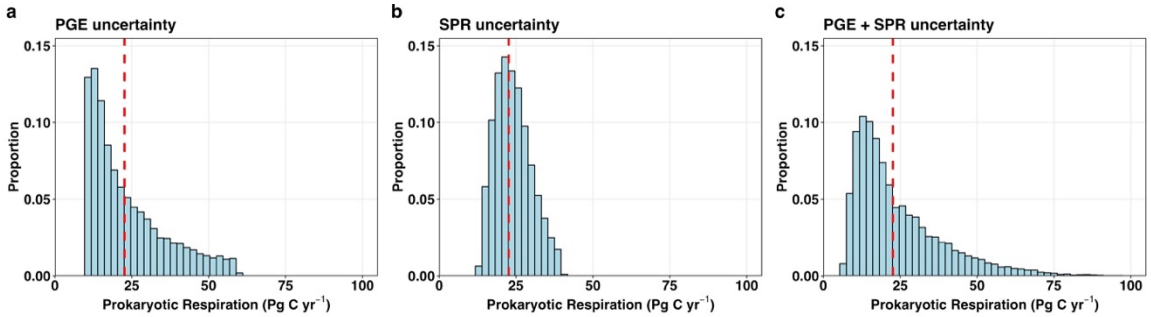

Figure S12 | Sources of uncertainty in global epipelagic prokaryotic respiration. Distribution of 10,000 estimates of global prokaryotic respiration (Pg C yr^{-1}), incorporating uncertainty from a) only prokaryotic growth efficiency (PGE); b) only specific-production rates (SPRs); and c) both PGE and SPR. The red dashed line is our reported estimate of global prokaryotic respiration (22.6 Pg C yr^{-1}). PGE uncertainty was incorporated by sampling from the interquartile range of our PGE dataset (6-27%; assuming a uniform distribution), while SPR uncertainty was resolved by drawing from the SPR statistical model's 95% confidence interval (Fig. S7). Source data are provided as a Source Data file.

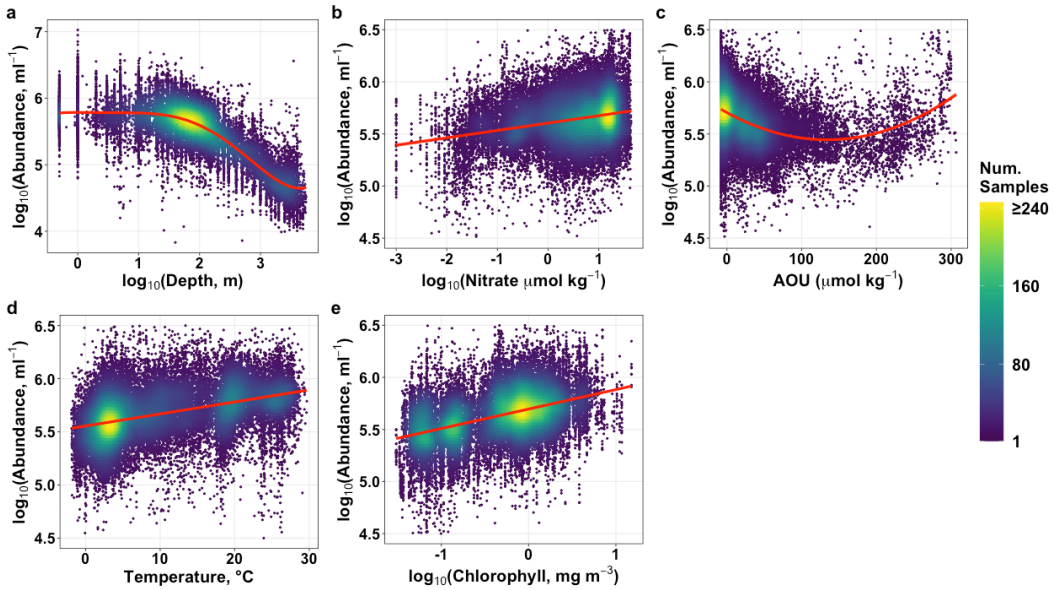

Figure S13 | Prokaryotic abundance as a function of a) $\log_{10}$ depth, b) $\log_{10}$ nitrate, c) apparent oxygen utilization (AOU), d) temperature and e) $\log_{10}$ chlorophyll *a* in the final parametric model. The red line and shading are the fitted response for each environmental variable, with partial residuals shown as dots coloured by the number of samples (Num. Samples). Source data are provided as a Source Data file.

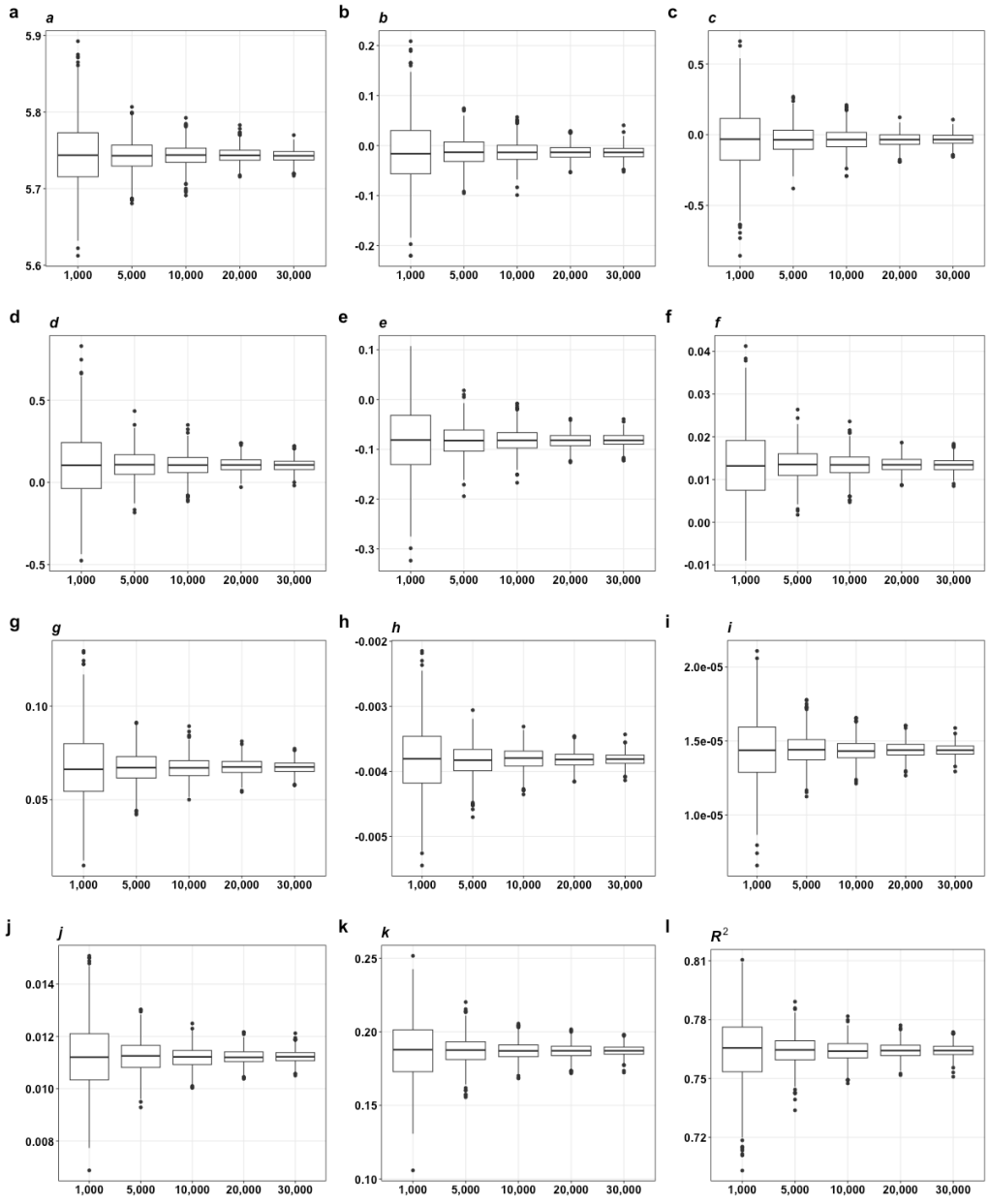

Figure S14 | Sensitivity analysis of the final parametric model. Box plots of (a-k) model parameters and l) model R^2 as a function of sample size. For each sample size, the model was fit 1,000 times using a bootstrap resampling method. Source data are provided as a Source Data file. The equation for the parametric model is:

$$\log_{10}(\text{Prokaryote Abundance}) = \text{a} + \text{b} \times \log_{10}(\text{depth}) + \text{c} \times \log_{10}(\text{depth})^2 + \text{d} \times \log_{10}(\text{depth})^3 + \text{e} \times \log_{10}(\text{depth})^4 + \text{f} \times \log_{10}(\text{depth})^5 + \text{g} \times \log_{10}(\text{nitrate}) + \text{h} \times \text{aou} + \text{i} \times \text{aou}^2 + \text{j} \times \log_{10}(\text{temperature}) + \text{k} \times \log_{10}(\text{chlorophyll}) + \epsilon.$$

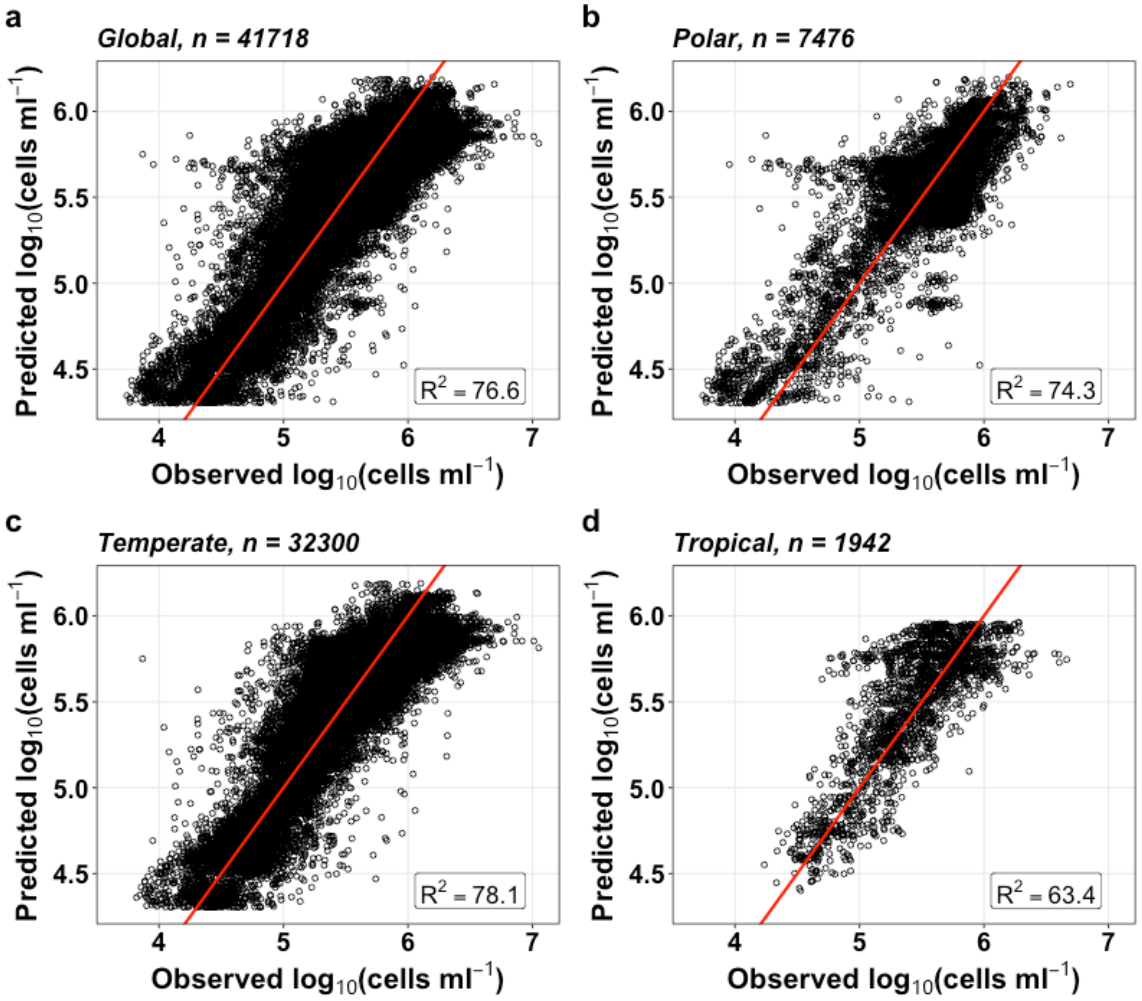

Figure S15 | Regional sensitivity of the parametric model across a) Global, b) Polar (>60°), c) Temperate (30°-60°) and d) Tropical (<30°) waters. The number of observations in each region is given in each panel title. For each sample, predicted prokaryote abundance was estimated using the parametric model and the sample's corresponding environmental variables. The solid red line is the 1:1 relationship, and R^2 values are reported for each region. Source data are provided as a Source Data file.

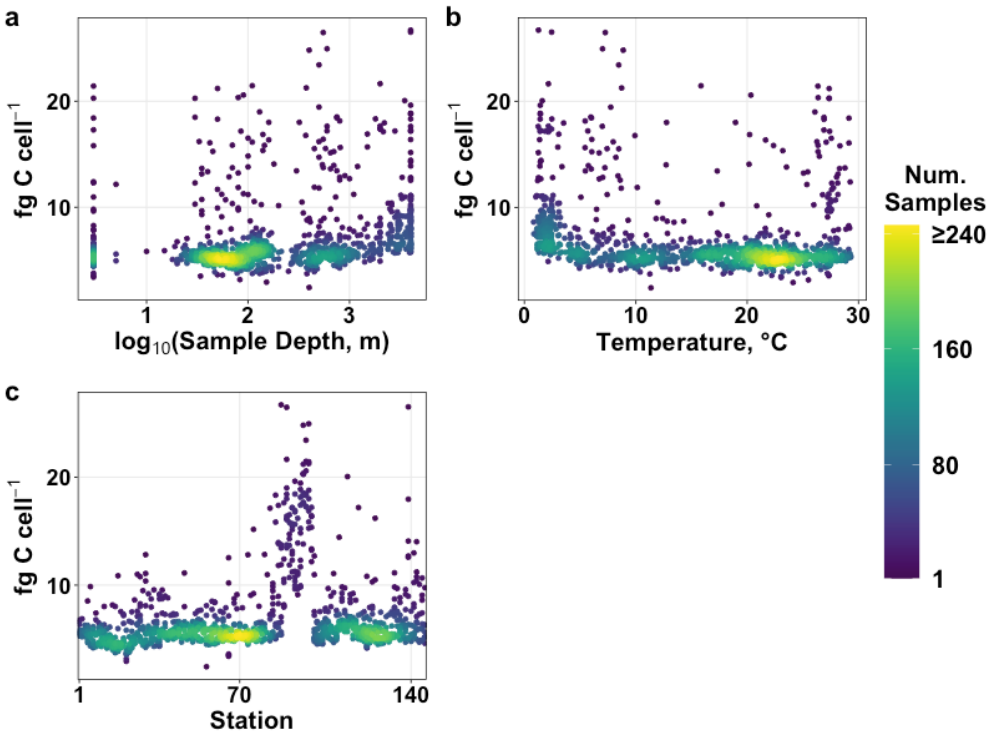

Figure S16 | Scatter plots of individual prokaryotic cell carbon (fg cell^{-1}) against a) $\log_{10}$ sample depth; b) temperature and c) station. Points are coloured by number of samples (Num. Samples). Source data are provided as a Source Data file.

Table S2. Comparison of different predictor variable combinations fitted with penalized regression splines (or factor) in a generalised additive model for individual cell-specific carbon. Root mean-square error (RMSE), mean absolute error (MAE) and deviance explained are reported for each model. Our selected candidate variable set for the final parametric equation is in bold.

Predictor Variables	**RMSE**	**MAE**	**Deviance explained**
$\log_{10}$ depth	0.21	2.25	8.6%
temperature	0.11	2.18	15.3%
station (factor)	0.16	1.56	69.8%
$\log_{10}$ depth + temperature	0.16	2.13	17.1%
temperature + station	**0.26**	**1.38**	**81.9%**
$\log_{10}$ depth + temperature + station	0.23	1.34	82.5%

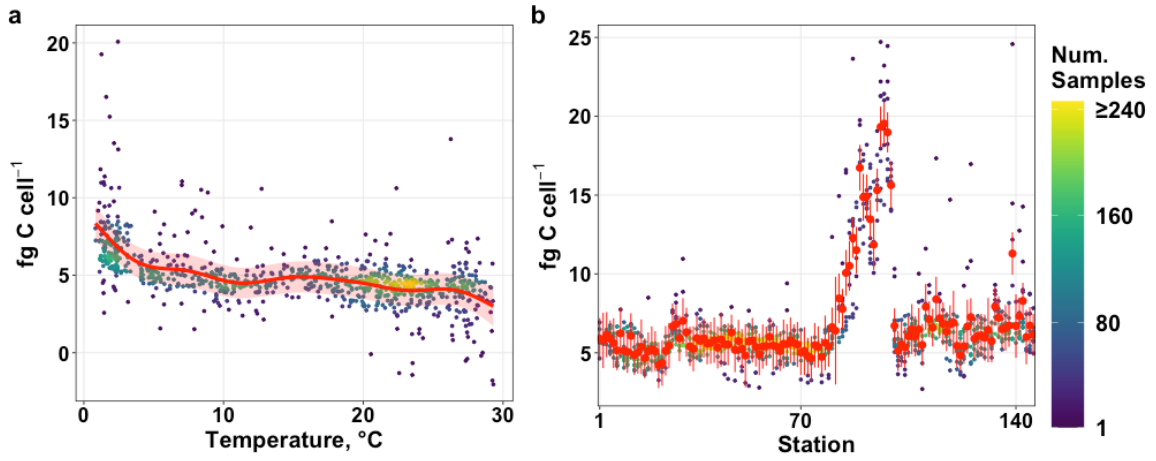

Figure S17 | Prokaryotic cell-specific carbon as a function of a) temperature and b) station in the generalised additive model. The red line and shading (or vertical line for station) is the fitted response for each environmental variable, with partial residuals shown as dots coloured by the number of samples (Num. Samples). Source data are provided as a Source Data file.

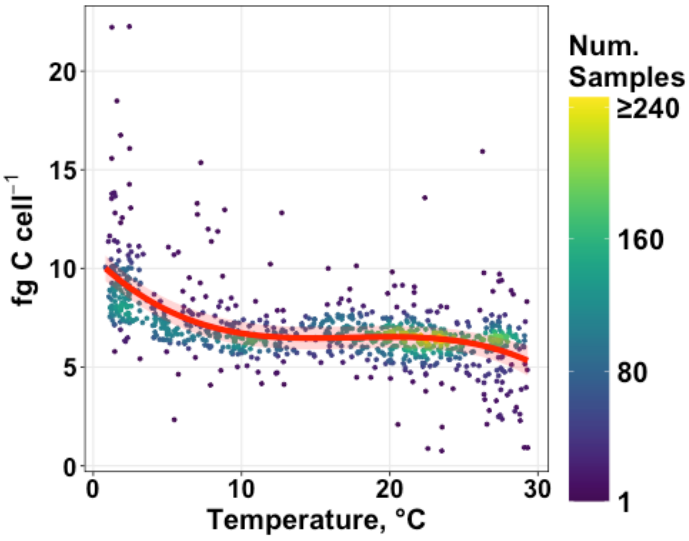

Figure S18 | Prokaryotic cell carbon as a function of temperature in the generalised linear mixed model (which also includes station as a random intercept, but this is not included in the figure above). The red line is the fitted response for each environmental variable, with partial residuals shown as dots coloured by the number of samples (Num. Samples). Source data are provided as a Source Data file.

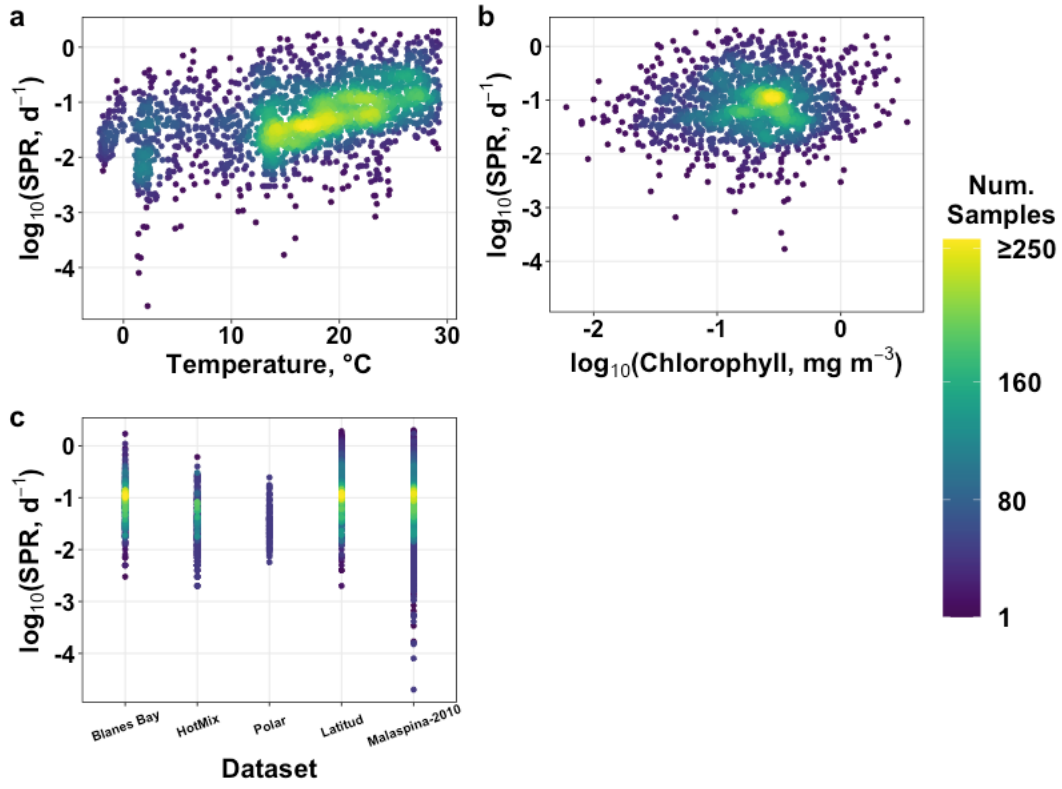

Figure S19 | Scatter plots of prokaryotic specific production rates (SPR; day^{-1}) against the full suite of predictor variables, coloured by number of samples (Num. Samples). Source data are provided as a Source Data file.

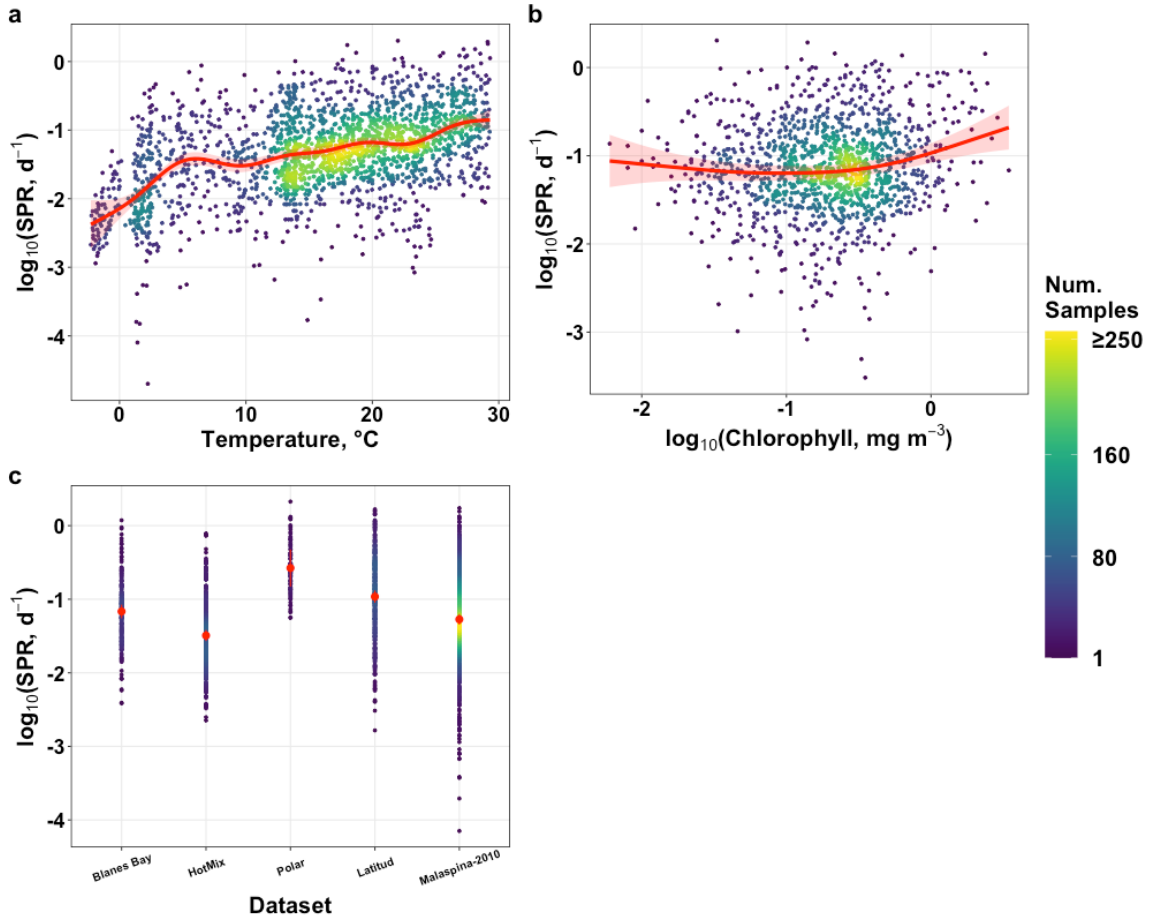

Figure S20 | Prokaryotic specific-production rates (SPR; day^{-1}) as a function of a) temperature; b) log_{10} chlorophyll and c) Dataset, in the generalised additive model. The red line and shading (or vertical line for Dataset) is the fitted response for each environmental variable, with partial residuals shown as dots coloured by the number of samples (Num. Samples). Source data are provided as a Source Data file.