Supplement of Biogeosciences, 21, 3927–3958, 2024
https://doi.org/10.5194/bg-21-3927-2024-supplement

Biogeosciences Open Access EGU

Supplement of

Lipid remodeling in phytoplankton exposed to multi-environmental drivers in a mesocosm experiment

Sebastian I. Cantarero et al.

Correspondence to: Sebastian I. Cantarero (sebastian.cantarero@colorado.edu)

Supplementary

Figures

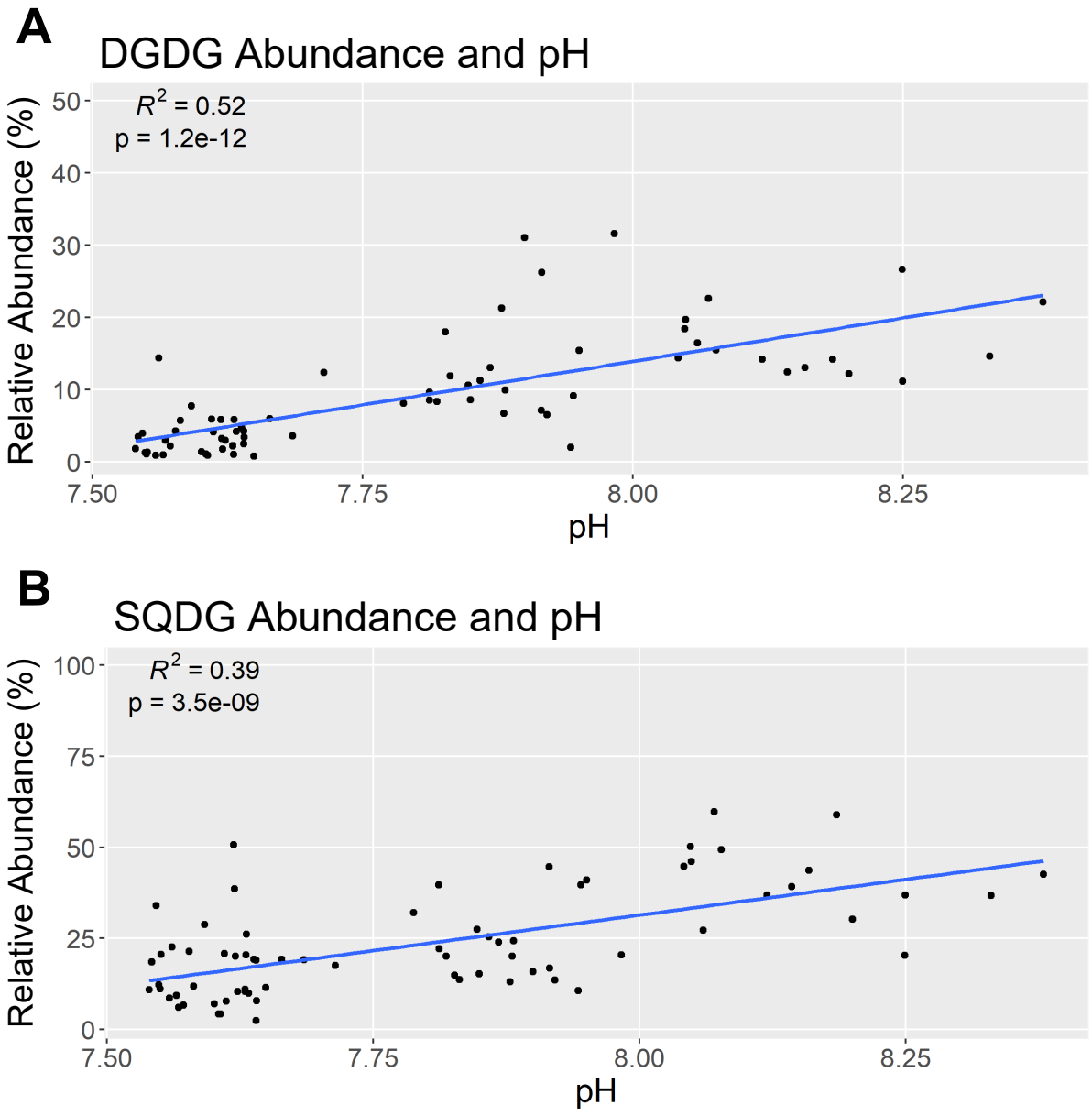

Figure S1: Linear regressions between pH and relative abundance of IPL classes DGDG (A) and SQDG (B). The statistical significance of the regression and the goodness of fit are denoted by the p value and the R^2, respectively.

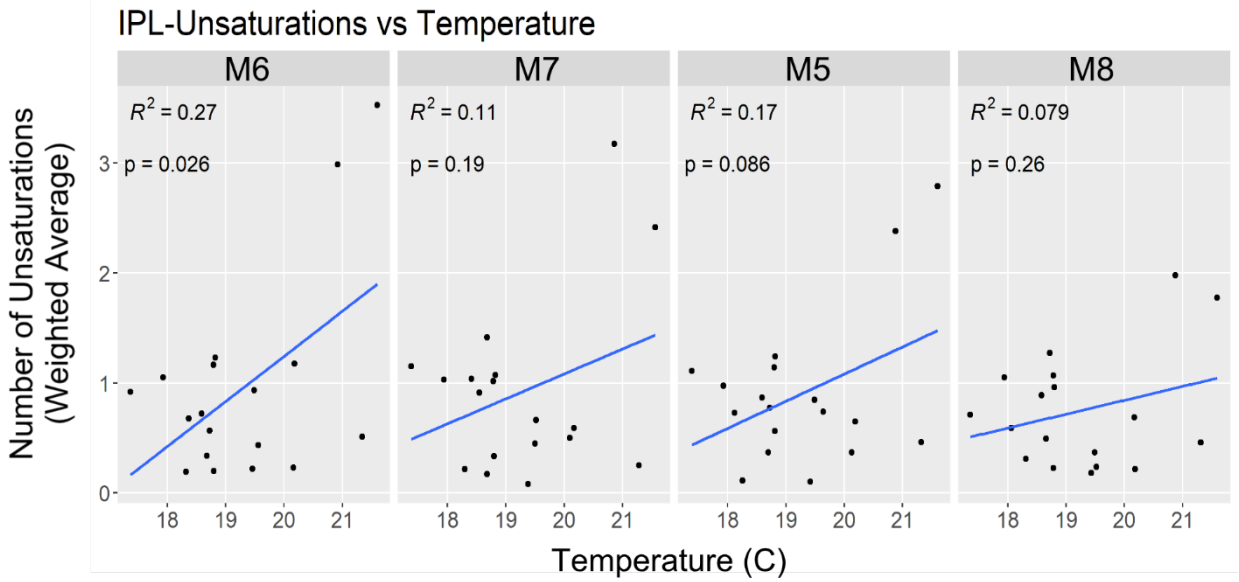

Figure S2. Linear regressions between Temperature and the number of unsaturation in all IPL classes (weighted average). The statistical significance of the regression and the goodness of the fit are denoted by the p value and the R^2, respectively.

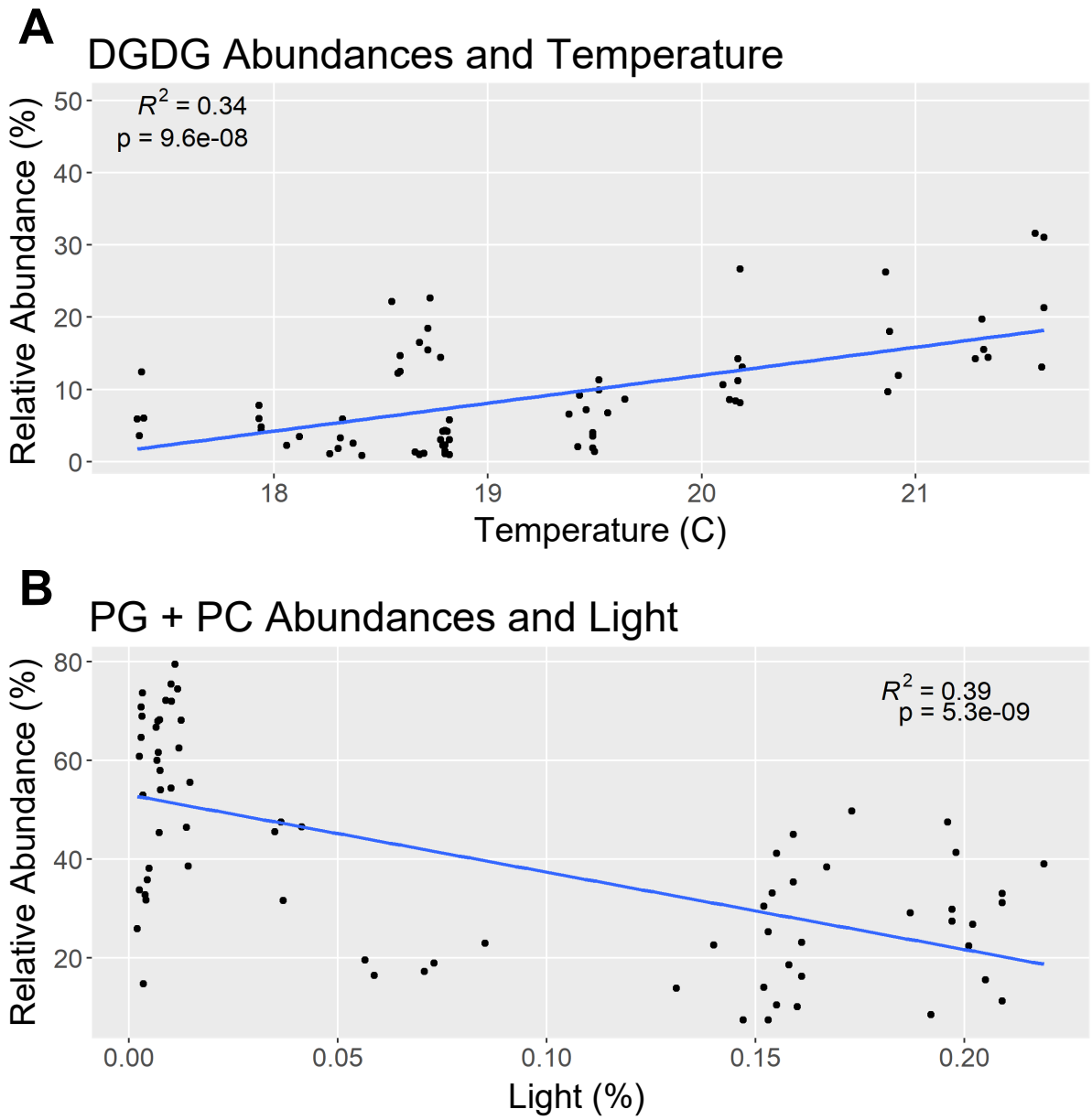

Figure S3. Linear regression between (A) Temperature and the relative abundance of DGDG lipids, and B) between the surface-normalized light level (expressed as a %) and PG+PC relative abundances. The statistical significance of the regression and the goodness of the fit are denoted by the p value and the R^2, respectively.

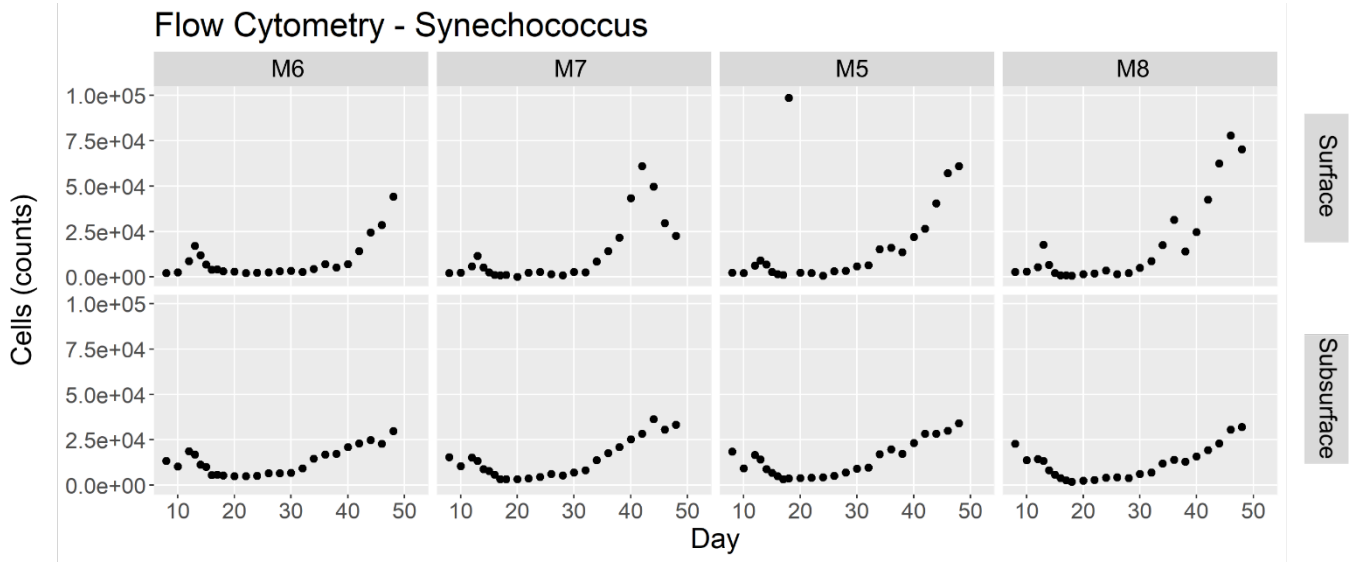

Figure S4. Cell counts of *Synechococcus* measured by flow cytometry in surface and subsurface waters of each mesocosm.

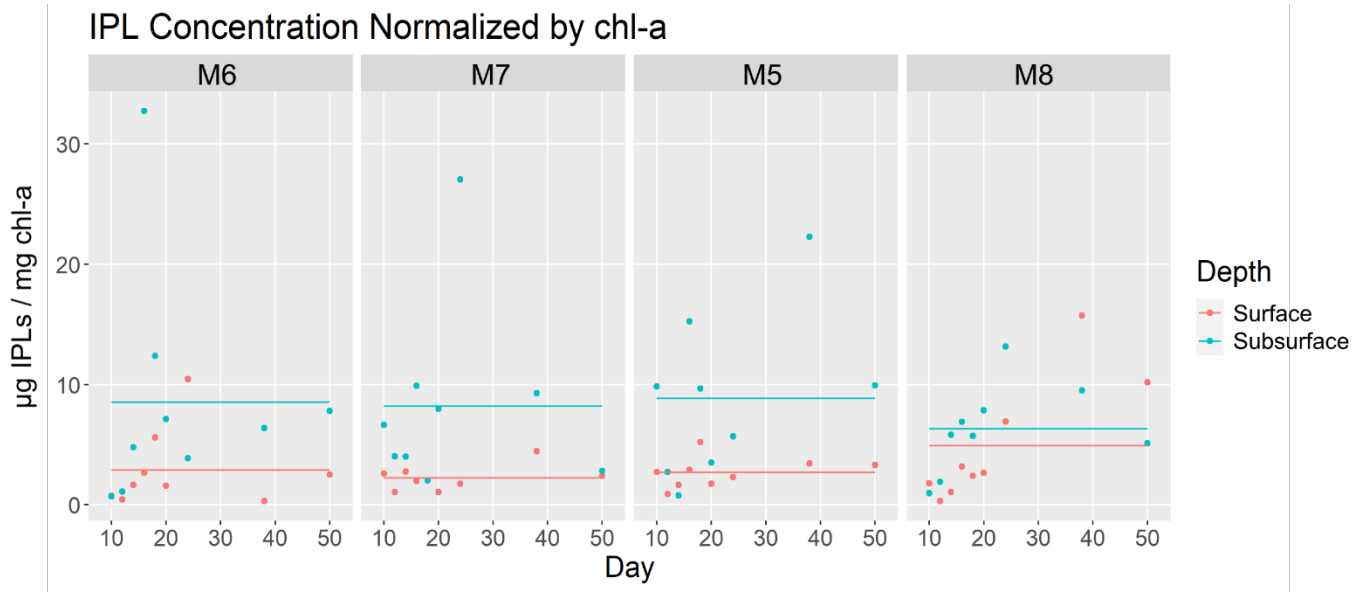

Figure S5. Concentration of total IPLs over chl-a concentration (expressed in µg/mg) in surface and subsurface waters of each mesocosm. The solid horizontal lines represent the average value for surface (red) and subsurface (blue) waters.

S1. Supplementary Notes

Classification and Regression Tree (CART) and Random Forest (Figures 4, 5, 6, and 7 in the main text).

The CART analysis is a versatile statistical method that uses decision tree structure to model complex relationships between predictor and target variables, including regression and classification problems. Essentially, it repeatedly divides data into smaller subsets, or nodes, based on specific data features to create a predictive tree structure (Breiman et al., 1984, Ho, 1995). In this study, CART, analyses employ splitting criteria determined by evaluating the sum of squared deviations in all potential splits, selecting those that yield the greatest reduction in residual error. To mitigate overfitting in the CART analysis, a pruning procedure is conducted to remove nodes that offer minimal contribution to model accuracy, as determined by cost complexity measure (Bae et al., 2010). This procedure allows us to simplify the CART results and focus our interpretations on the most significant predictors of IPL headgroups only.

A Random Forest (RF) is a machine learning algorithm that employs multiple decision trees to make more accurate predictions in classification and regression (Breiman et al., 2001). Each decision tree is constructed independently using a random subset of data and features from the training set. In the case of regression problems, such as predicting numerical values, a metric like Root Mean Square Error (RMSE) is used to evaluate the model's accuracy. The RMSE represents the square root of the mean of the squares of the errors between the model predictions and the actual values in the test dataset (Wei et al., 2010, Tyralis et al., 2019). In summary, RF utilizes multiple decision trees and metrics like RMSE to generate more accurate and robust mode. A lower RMSE indicates the model can more accurately predict the target variable's numerical values. Following the averaged cross validated accuracy estimates, we implemented a cutoff of 5% reduction in RMSE to eliminate variables that do not significantly reduce the error of the model prediction. This cutoff allows us to focus our interpretation of variables that contribute significantly to the out of bag predictor performance.

The use CART and RF algorithms helps assessing the robustness of statistical relationships and gain a more comprehensive understanding of how environmental variables influence the distribution of IPL classes, thus enhancing our understanding of lipid remodeling.

Please refer to our analyses in our R code available in this study: https://github.com/Guachan/IPLs-KOSMOS/tree/0.1.0 (https://doi.org/10.5281/zenodo.10408453, Cantarero, 2023).

References

Bae, H.K., Olson, B.H., Hsu, K.L., and Sorooshian, S. Classification and regression tree (CART) analysis for indicator bacterial concentration prediction for a Californian coastal area Water Sci. Technol., 61 (2), pp. 545-553, 2010.

Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.J.: Classification and Regression Trees (1st ed.). Routledge. https://doi.org/10.1201/9781315139470, 1984.

Breiman, L.: Random Forests, Mach. Learn., 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Ho, T. K.: Random decision forests, in Proceedings of 3rd International Conference on Document Analysis and Recognition, IEEE Comput. Soc. Press., 1, 278–282, 1995.

Tyralis, H., Papacharalampous, G., & Langousis, A. A brief review of random forests for water scientists and practitioners and their recent history in water resources. Water, 11(5), 910, https://doi.org/10.3390/w11050910, 2019.

Wei, C.-L., Rowe, G. T., Escobar-Briones, E., Boetius, A., Soltwedel, T., Caley, M. J., Soliman, Y., Huettmann, F., Qu, F., Yu, Z., Pitcher, C. R., Haedrich, R. L., Wicksten, M. K., Rex, M. A., Baguley, J. G., Sharma, J., Danovaro, R., MacDonald, I. R., Nunnally, C. C., Deming, J. W., Montagna, P., Lévesque, M., Weslawski, J. M., Wlodarska-Kowalczuk, M., Ingole, B. S., Bett, B. J., Billett, D. S. M., Yool, A., Bluhm, B. A., Iken, K., and Narayanaswamy, B. E.: Global patterns and predictions of seafloor biomass using random forests, PLoS ONE, 5, e15323, doi:10.1371/journal.pone.0015323, 2010.