Supplementary Material for Paper

Data-based assessment of environmental controls on global marine nitrogen fixation

Ya-Wei Luo (Woods Hole Oceanographic Institution, Woods Hole, MA),

Ivan D. Lima (Woods Hole Oceanographic Institution, Woods Hole, MA),

David M. Karl (University of Hawaii, Honolulu, HI) and

Scott C. Doney (Woods Hole Oceanographic Institution, Woods Hole, MA)

(1) Supplementary texts: the procedure using the Cook's Distance to remove outliers from multiple linear regression (MLR).

Cook's distance [*Cook*, 1977], which measures the effect of deleting an observation from the regression, was used to identify outliers to the MLR. The Cook's distance for the *i*th observation is defined as:

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{P \cdot \mathrm{MSE}}, \qquad (1)$$

where $\hat{y}_j$ and $\hat{y}_{j(i)}$ are the *j*th predicted value with and without the *i*th observation used in the regression, respectively; *n* is the total number of observations; MSE is the mean squared error; and *P* is the number of coefficients in the regression model. Thus an observation with a high Cook's distance indicates that the observation has great impact on the regression. Although different critical values for D_i are proposed for identifying

outliers, such as $D_i > 1$ [*Cook and Weisberg*, 1982] or $D_i > 4/n$ [*Bollen and Jackman*, 1990], we followed the suggestions of *Chatterjee et al.* [2000] to identify outliers by graphing the D_i values and examining any one or two points having a much higher D_i than the others (Figure S1). The MLR was redone after removing outliers.

Reference

Bollen, K. A., and R. W. Jackman (1990), Regression diagnostics: An expository treatment of outliers and influential cases, in *Modern Methods of Data Analysis*, edited by J. Fox and J. S. Long, pp. 257-291, Sage, Newbury Park, CA.

Chatterjee, S., A. S. Hadi, and B. Price (2000), *Regression Analysis by Example, 3rd edition*, Wiley, New York.

Cook, R. D. (1977), Detection of influential observation in linear regression, *Technometrics*, *19*(1), 15-18, doi:10.2307/1268249.

Cook, R. D., and S. Weisberg (1982), *Residuals and influence in regression*, Chapman & Hall, New York, NY.

(2) Supplemental Figure

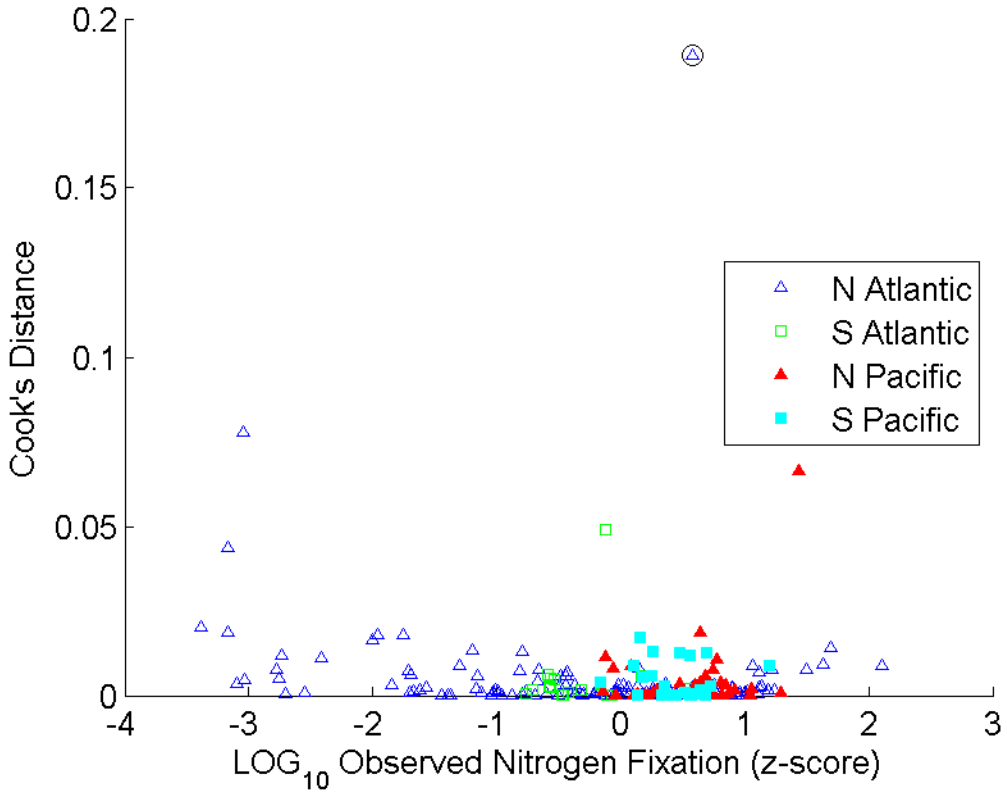

Fig. S1. Cook's distance of each data point in the multiple linear regression (MLR) with all the available data points used. The data point with black circle is the actual data point excluded from the final MLR.