[Journal of Geophysical Research: Biogeosciences]

Supporting Information for

Machine Learning Estimates of Global Marine Nitrogen Fixation

Weiyi Tang[1], Zuchuan Li[1], Nicolas Cassar[1,2,*]

[1]Division of Earth and Ocean Sciences, Nicholas School of the Environment, Duke University, Durham, NC 27708, USA

[2]Laboratoire des Sciences de l'Environnement Marin (LEMAR), UMR 6539 UBO/CNRS/IRD/IFREMER, Institut Universitaire Européen de la Mer (IUEM), Brest, France

[*]Correspondence to: Nicolas Cassar (Nicolas.Cassar@duke.edu)

Contents of this file

Additional Supporting Information (Files uploaded separately)

Text S1.

Caveats, uncertainties and future improvements

1. Sparse and uneven distribution of observations, and mismatch of N_2 fixation observations with predictors

While our training dataset includes new observations in the South Pacific, the Indian and Arctic Oceans (which were not available in Luo et al. (2014)), the majority of observations remain in the North Atlantic Ocean (Figure 1). For example, only 6 points are available in the Indian Ocean. This uneven and sparse distribution of observations may bias the statistical models by giving some regions more weight. Given that factors regulating N_2 fixation likely vary between biomes and regions (Monteiro, Dutkiewicz, & Follows, 2011; Ward, Dutkiewicz, Moore, & Follows, 2013; Weber & Deutsch, 2014), models based on data mainly collected in the warm and oligotrophic waters of the Atlantic and Pacific Ocean may not accurately represent the other regions, including the recently discovered niches of N_2 fixation in cold and/or nutrient-rich waters. In addition, spatial and temporal mismatch, and the coarse spatial and temporal resolution of our predictors and predictand may introduce noise. Some of the predictors may also work over longer timescales or larger spatial scales than the ones captured by the short-term incubations. Finally, some environmental factors like nutrient supply ratios (Ward et al., 2013) may be better predictors of the presence or absence of diazotrophs rather N_2 fixation rates. Overall, observations over broader swaths of the oceans will help further refine the biogeography and magnitude of marine N_2 fixation.

2. Difference in methods measuring N_2 fixation

Our models rely on measurements of N_2 fixation collected by three different incubation methods (i.e. ARA, $^{15}N_2$ gas addition, and dissolved $^{15}N_2$ addition). The uncertainties, assumptions and drawbacks of each method introduce biases and noise in our predictions. In line with this, training the algorithms with each individual method leads to significantly different biogeographies of N_2 fixation (Figure S2). The ARA method detects bulk N_2 fixation rates including the particulate and dissolved products (Mulholland, 2007), but suffers from variable conversion stoichiometry of acetylene reduction to N_2 fixation (Wilson, Böttjer, Church, & Karl, 2012) and other issues presented in Cassar et al. (2018). It also displays a geographical bias with most applications being conducted in the North Atlantic. The $^{15}N_2$ gas addition method is the most commonly used method so far. Unfortunately, it has been shown to underestimate N_2 fixation rates because of incomplete gas-liquid equilibration of the $^{15}N_2$ tracer (Mohr, Großkopf, Wallace, & LaRoche, 2010) and other issues (Bombar, Paerl, Anderson, & Riemann, 2018; Mulholland, 2007). While a correction may be applied for the incomplete equilibration (Böttjer et al., 2016), it comes with significant uncertainty because of variability in the degree of disequilibrium between studies. The dissolved N_2 addition method is now believed to give the best estimates of *in-situ* N_2 fixation rates. However, the measurements are too few at this time to meaningfully train our machine learning algorithms (Figure 1). Finally, varying depths of integration may also lead to significant uncertainties. Some studies report N_2 fixation rates integrated over the euphotic zone while others report rates to a specific depth (e.g. 200 m). The recent discovery of aphotic N_2 fixation (Fernandez, Farías, & Ulloa, 2011; Hamersley et al., 2011) exacerbates this issue. Although rates of N_2 fixation are low at depth, they may be significant when integrated over deep water columns. Observations should therefore be reported to a depth relevant to N_2 fixation to simplify inter-study comparisons.

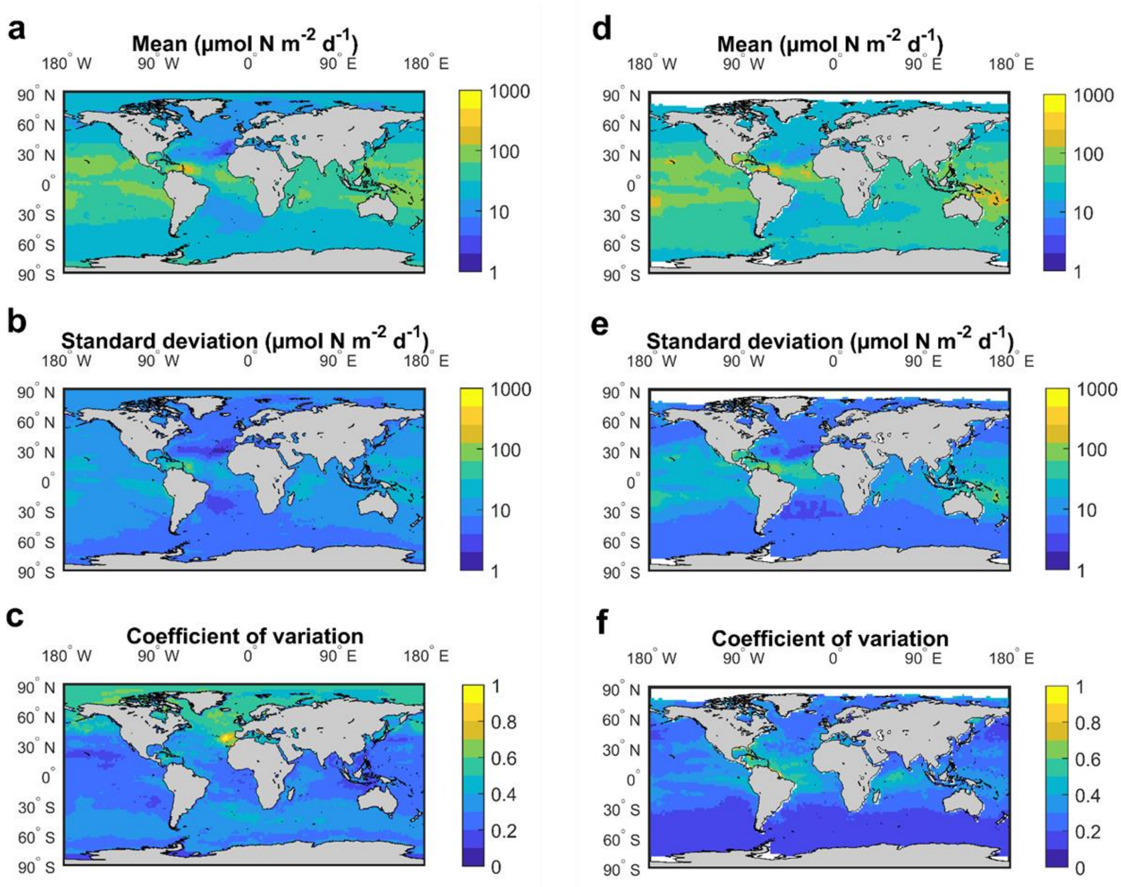

Figure S1. Mean, standard deviation and coefficient of variation of global N_2 fixation from 100 bootstrap reconstructed N_2 fixation datasets by random forest (RF, a-c), support vector regression (SVR, d-f) respectively.

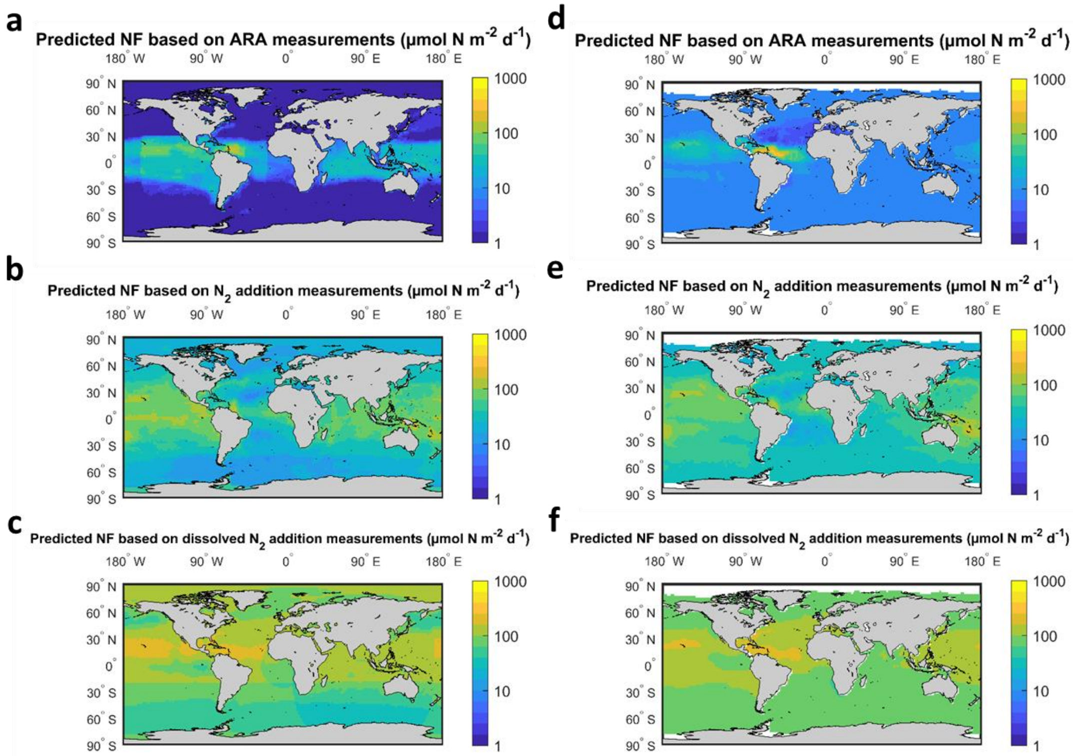

Figure S2. RF (a-c) and SVR (d-f) model predictions of N_2 fixation based on observations collected with a single method. a, d. Acetylene reduction assay (ARA). b, e. $^{15}N_2$ gas addition. c, f. Dissolved $^{15}N_2$ addition.

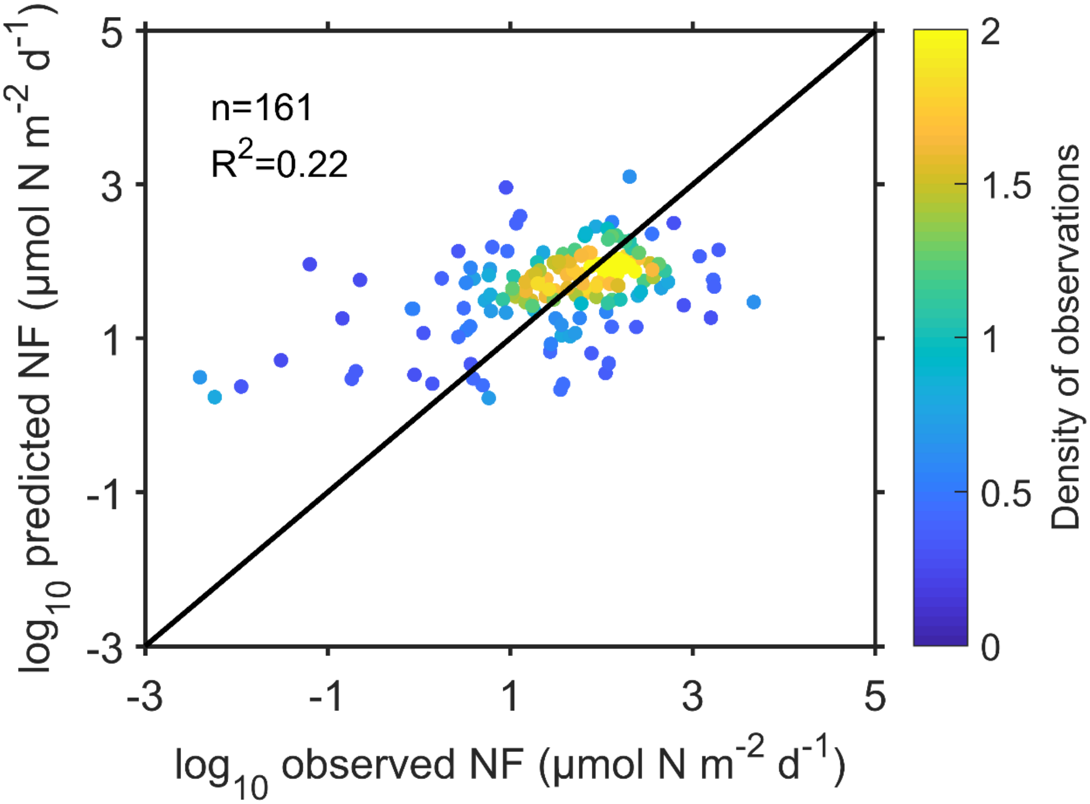

Figure S3. Observed versus simulated N_2 fixation rates by stepwise multiple linear regression.

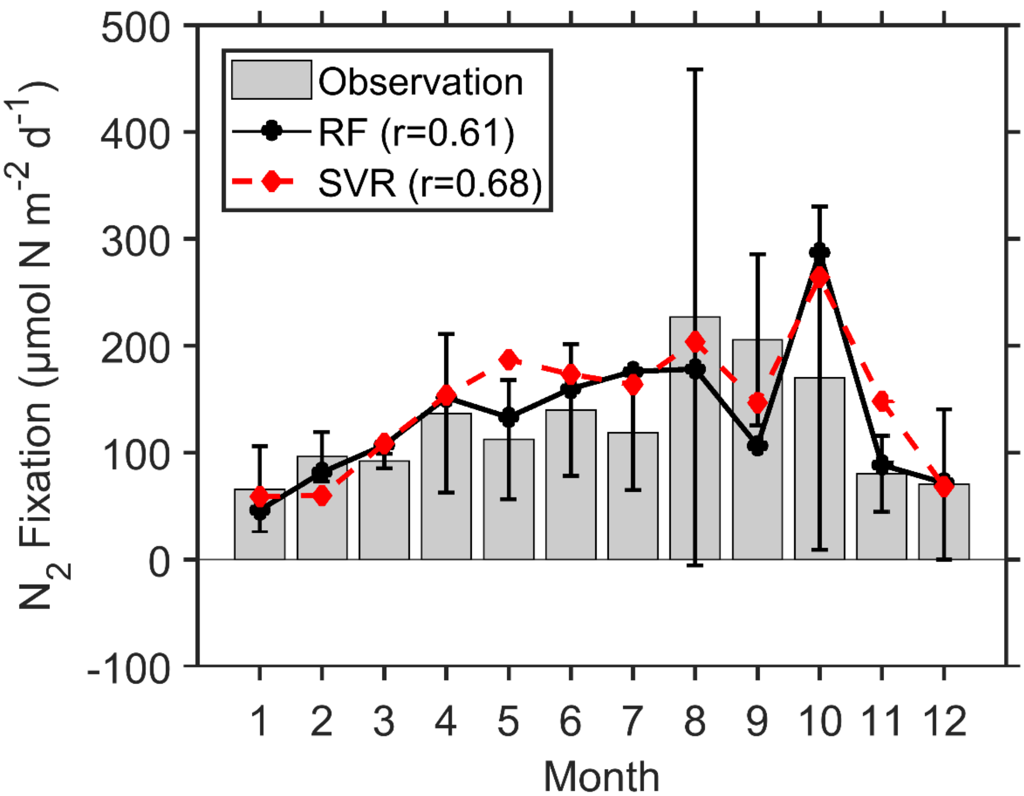

Figure S4. Comparison of observed and modeled seasonal changes in N_2 fixation rates at Hawaii Ocean Time-series (HOT). Bar plot with error bars represents the observed monthly climatology of N_2 fixation rates ± one standard deviation at HOT.

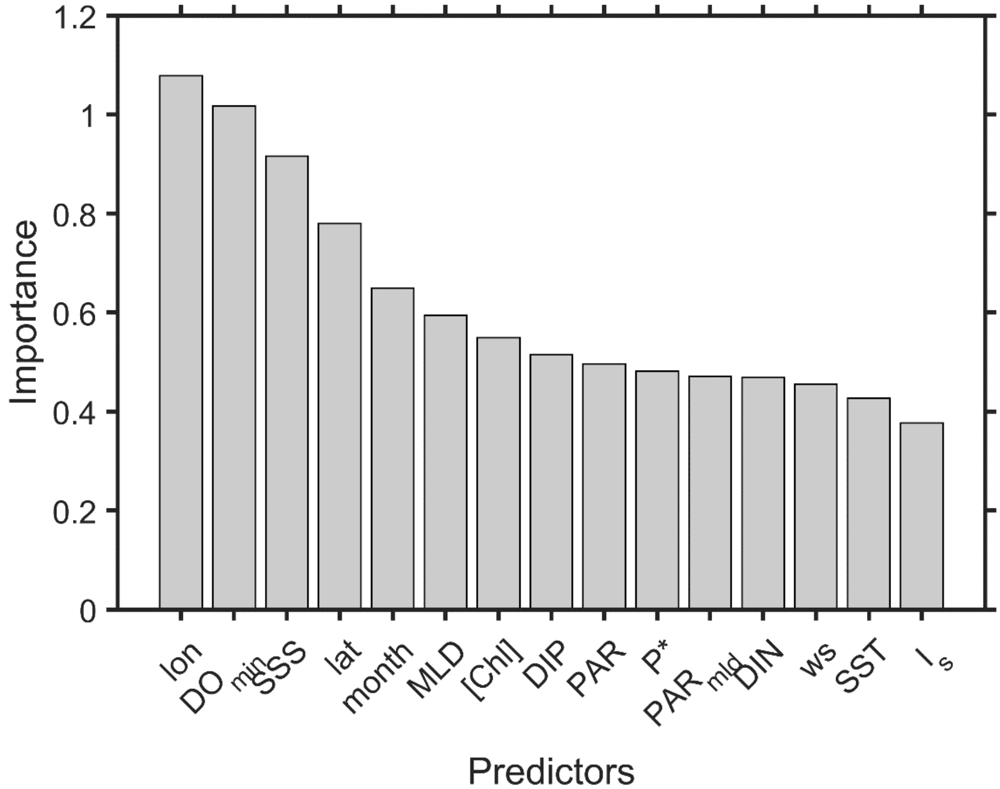

Figure S5. Predictor feature importance in random forest.

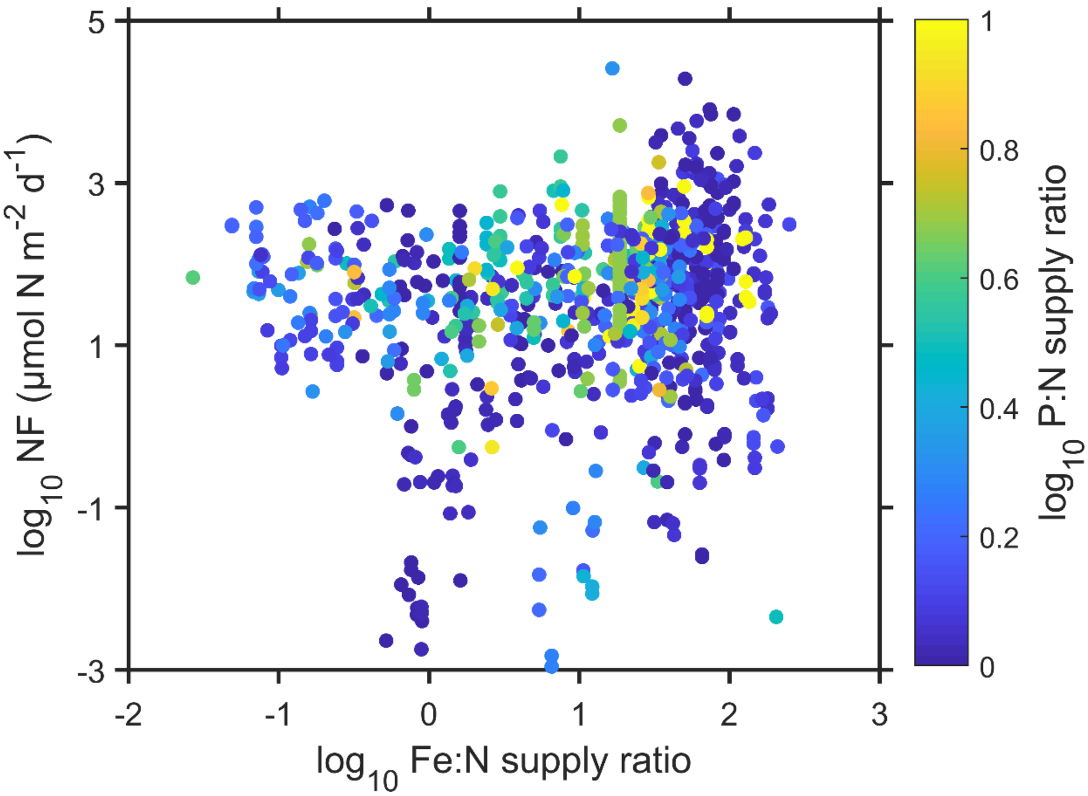

Figure S6. N_2 fixation rates versus Fe:N and P:N supply ratios. Fe:N and P:N supply ratios are from Ward et al. (2013).

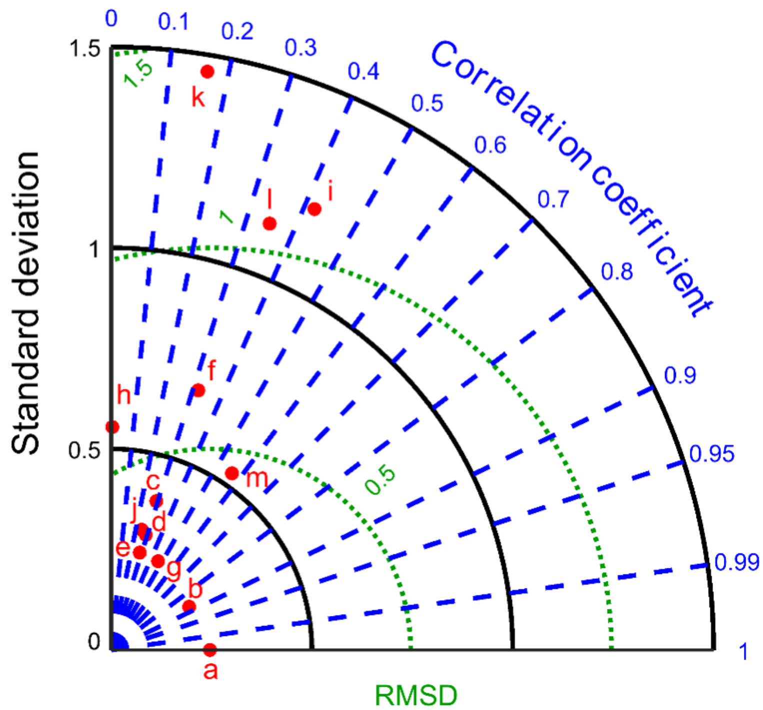

Figure S7. Taylor diagram of N_2 fixation rates (logarithmic scale) estimated by different models with the alphabetical order shown in Figure 4, with the estimate by RF (a) as the reference value. Dashed blue and dotted green lines represent the correlation and centered root-mean-square difference (RMSD) between estimates by RF and other models, respectively. Solid black lines, the radial distance from the origin, represent the standard deviation of the spatial distribution estimated by each model, with lower values indicating less spatial variability.

Movie S1. Monthly changes of predicted N_2 fixation rates by random forest over the global ocean.

Movie S2. Monthly changes of predicted N_2 fixation rates by support vector regression over the global ocean.