

# A coastal N<sub>2</sub> fixation hotspot at the Cape Hatteras front: Elucidating spatial heterogeneity in diazotroph activity via supervised machine learning

Corday R. Selden ,\* P. Dreux Chappell , Sophie Clayton , Alfonso Macías-Tapia, Peter W. Bernhardt, Margaret R. Mulholland 

Department of Ocean and Earth Sciences, Old Dominion University, Norfolk, Virginia

## Abstract

In the North Atlantic Ocean, dinitrogen (N<sub>2</sub>) fixation on the western continental shelf represents a significant fraction of basin-wide nitrogen (N) inputs. However, the factors regulating coastal N<sub>2</sub> fixation remain poorly understood, in part due to sharp physico-chemical gradients and dynamic water mass interactions that are difficult to constrain via traditional oceanographic approaches. This study sought to characterize the spatial heterogeneity of N<sub>2</sub> fixation on the western North Atlantic shelf, at the confluence of Mid- and South Atlantic Bight shelf waters and the Gulf Stream, in August 2016. Rates were quantified using the <sup>15</sup>N<sub>2</sub> bubble release method and used to build empirical models of regional N<sub>2</sub> fixation via a random forest machine learning approach. N<sub>2</sub> fixation rates were then predicted from high-resolution CTD and satellite data to infer the variability of its depth and surface distributions, respectively. Our findings suggest that the frontal mixing zone created conditions conducive to exceptionally high N<sub>2</sub> fixation rates (> 100 nmol N L<sup>-1</sup> d<sup>-1</sup>), which were likely driven by the haptophyte-symbiont UCYN-A. Above and below this hotspot, N<sub>2</sub> fixation rates were highest on the shelf due to the high particulate N concentrations there. Conversely, specific N<sub>2</sub> uptake rates, a biomass-independent metric for diazotroph activity, were enhanced in the oligotrophic slope waters. Broadly, these observations suggest that N<sub>2</sub> fixation is favored offshore but occurs continuously across the shelf. Nevertheless, our model results indicate that there is a niche for diazotrophs along the coastline as phytoplankton populations begin to decline, likely due to exhaustion of coastal nutrients.

## Introduction

An essential element for life, nitrogen (N) is often the nutrient limiting biological productivity in the surface ocean (Moore et al. 2013). The predominant form of N on Earth is dinitrogen (N<sub>2</sub>) gas, which is highly abundant but thermodynamically stable. Consequently, it is not readily accessible to organisms as a source of N. However, a select group of organisms, diazotrophs, can enzymatically mediate its reduction to an assimilable form of N. N<sub>2</sub> fixation may subvert N limitation and thus significantly contribute to new production by supplying biological systems with bioavailable N (e.g., Capone et al. 2005).

Marine N<sub>2</sub> fixation has historically been ascribed to the warm, well-lit surface waters of subtropical and tropical ocean basins where concentrations of dissolved inorganic N (e.g., nitrate plus nitrite [N + N] and ammonium) are at or near the limits of analytical detection (Carpenter and Capone 2008). Such conditions foster the growth of *Trichodesmium*, a well-studied and globally significant filamentous cyanobacterial diazotroph (Capone et al. 2005). *Trichodesmium* forms macroscopic colonies in the surface ocean that are easy to identify and manipulate. Consequently, *Trichodesmium* was one of the first marine diazotrophs to be established in stable culture (Prufert-Bebout et al. 1993). Bias toward *Trichodesmium* in the study of marine N<sub>2</sub> fixation and diazotroph physiology has inadvertently led to the conflation of its niche preferences with those of other pelagic diazotrophs (e.g., Knapp 2012). However, recent work (Zehr and Capone 2020 and references therein) suggests that the ecophysiology of diverse diazotrophic clades is not so generalizable.

Aided by methodological advances in molecular biology, research in the past decade has expanded the known range of pelagic N<sub>2</sub> fixers to include temperate (e.g., Moisaner

\*Correspondence: cseld001@odu.edu

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Additional Supporting Information may be found in the online version of this article.

et al. 2010) and Arctic ocean basins (e.g., Blais et al. 2012), dark and N-rich mesopelagic waters (e.g., Rahav et al. 2013), and eastern boundary upwelling systems (e.g., Selden et al. 2019). Moreover, high  $N_2$  fixation rates (NFRs) have been observed in tropical (e.g., Grosse et al. 2010), sub-tropical (Hunt et al. 2016) and temperate (e.g., Mulholland et al. 2012; Fonseca-Batista et al. 2019) coastal systems where they can drive productivity (e.g., Hunt et al. 2016). In the North Atlantic Ocean, the magnitude of NFR tends to increase with proximity to the North American continental margin (Tang et al. 2019), flouting the historical paradigm that significant NFRs are restricted to the oligotrophic ocean. Indeed, direct NFR measurements indicate that these coastal waters can contribute significantly to basin-wide fixed N inputs. Recently, Mulholland et al. (2019) estimated that the continental shelf extending from the Mid-Atlantic Bight to the Gulf of Maine alone adds  $0.28 \text{ Tg N yr}^{-1}$  to the North Atlantic Ocean. Though this region represents only 6.4% of the total North Atlantic continental shelf, this N input is equivalent to previous estimates for the entire shelf area (Nixon et al. 1996).

High NFRs ( $> 100 \text{ nmol N L}^{-1} \text{ d}^{-1}$  or  $> 1000 \text{ } \mu\text{mol N m}^{-2} \text{ d}^{-1}$ ) in cooler, temperate North Atlantic coastal waters have been associated primarily with the haptophyte-symbiont UCYN-A and diatom-diazotroph assemblages (Fonseca-Batista et al. 2019; Mulholland et al. 2019; Tang et al. 2019). Non-cyanobacterial diazotrophs may also be active in these waters, although they have not been associated with high NFRs (Fonseca-Batista et al. 2019; Mulholland et al. 2019). *Trichodesmium* has also long been observed in the Mid-Atlantic Bight and South Atlantic Bight, which are subject to inputs from the warm and

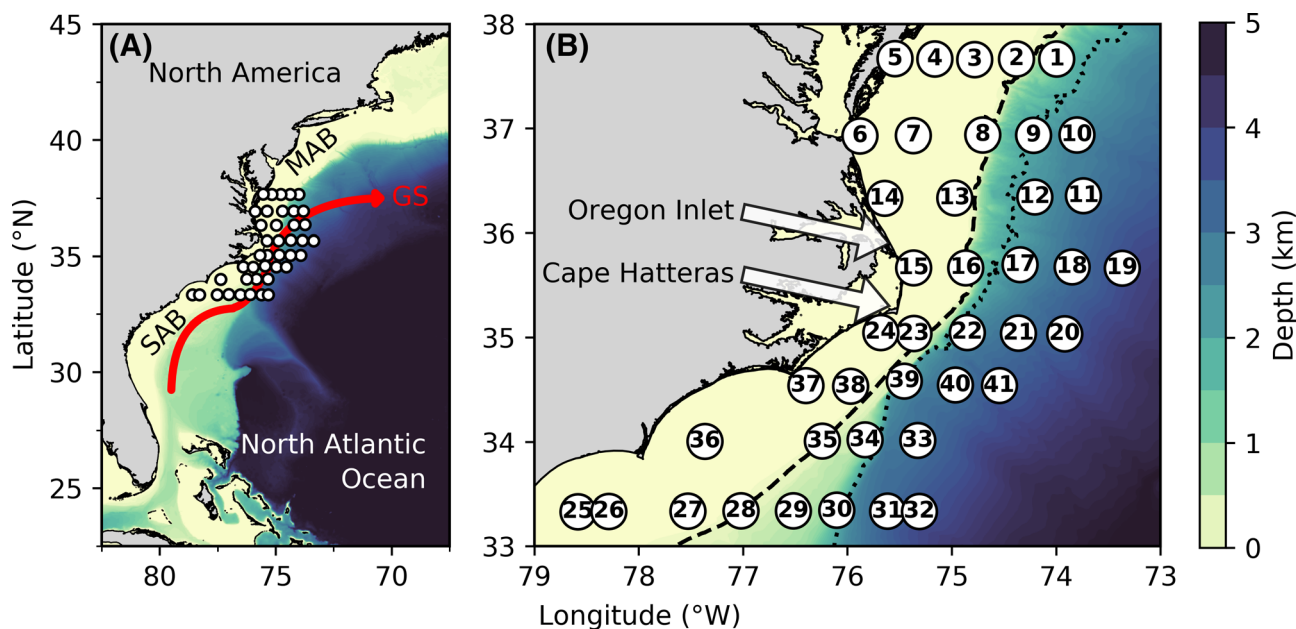
oligotrophic Gulf Stream (e.g., Mulholland et al. 2012; Tang et al. 2019; Palter et al. 2020). These distinct diazotrophic groups appear to differ in their range (Zehr and Capone 2020) and sensitivity to environmental variables (e.g., Mills et al. 2020), complicating our ability to predict NFRs through time and space (e.g., Tang et al. 2019). Moreover, understanding the factors regulating marine diazotrophy on local to regional scales will require more sophisticated characterization of the spatial and temporal heterogeneity of  $N_2$  fixation than can be gleaned from direct rate measurements alone.

This study sought to characterize the distribution of  $N_2$  fixation and diazotrophic populations across the western North Atlantic continental margin, where warm South Atlantic Bight shelf and Gulf Stream waters interact with cooler Mid-Atlantic Bight shelf waters. As NFR measurements are highly labor-intensive, a supervised machine learning algorithm was applied to augment their spatial coverage throughout the region at the time of the study. This approach facilitates examination of  $N_2$  fixation spatial heterogeneity and offers insight into the physical and chemical factors affecting its distribution and magnitude in a coastal frontal regime.

## Methods

### Sample collection

Samples were collected aboard the research vessel (R/V) *Hugh R. Sharp* in August, 2016, along eight transects (Fig. 1) extending across the northwestern mid-Atlantic continental shelf (33.3–37.7°N). The study area crossed the continental shelf and slope to the north (Mid-Atlantic Bight) and south (South Atlantic Bight) of Cape Hatteras, NC (Fig. 1B), where



**Fig 1.** Study region (A) and labeled station map (B) overlying ETOPO1 1 Arc-Minute Model bathymetry (Amante and Eakins 2009). In panel A, the solid red line indicates the approximate path of the Gulf Stream (GS). The Mid- and South Atlantic Bights are denoted as “MAB” and “SAB.” White dots indicate stations. In B, the dashed and dotted lines indicate the 200 and 2000 m isobaths, representing the edges of the shelf and slope, respectively.

the Gulf Stream separates from the shelf break. Water samples were collected from Niskin bottles mounted to a rosette, which was equipped with a SeaBird Electronics 911plus conductivity, temperature and depth (CTD) sensor package and a chlorophyll *a* (Chl *a*) fluorometer (ECO-AFL/FL, WET Labs).

### Nutrient determination

Samples for nutrient analyses were filtered directly from the Niskin bottles using a 0.2 μm Supor membrane cartridge filter (AcroPak 1500™, PALL Corp.) into sterile 15 mL Falcon™ tubes and analyzed on-board. Nitrate plus nitrite, nitrite, and soluble reactive phosphorus (SRP) concentrations were determined colorimetrically using an Astoria Pacific nutrient autoanalyzer, according to manufacturer specifications (Parsons et al. 1984). The detection limits for N + N and SRP were 0.14 and 0.04 μM (3σ, *n* = 7), respectively.

### N<sub>2</sub> fixation rate measurements

N<sub>2</sub> fixation rates were quantified using a modified version of the <sup>15</sup>N<sub>2</sub> tracer method (Montoya et al. 1996) known as the bubble release technique (Klawoon et al. 2015; White et al. 2020). This approach addresses the issue of slow N<sub>2</sub> gas dissolution (White et al. 2020)—which may result in rate underestimation—by removing undissolved N<sub>2</sub> after a mixing period. The <sup>15</sup>N<sub>2</sub> enrichment of the dissolved N<sub>2</sub> pool is then measured directly. White et al. (2020) provide a thorough description and discussion of this approach.

### N<sub>2</sub> fixation incubations

Whole water was collected from Niskin bottles directly into 10 L LDPE carboys. Carboys were covered with mesh screens during daytime sample collection so that microbes were not exposed to potentially harmful ultraviolet radiation. To measure NFR, triplicate PETG bottles (0.5–2 L) were rinsed and filled completely from carboys. Incubations were initiated by injecting tracer-level (~ 10% by vol.) <sup>15</sup>N<sub>2</sub> gas (~ 99% <sup>15</sup>N, Cambridge Isotope Laboratories, Inc., Tewksbury, MA) through a silicon septa using a gas-tight syringe (VICI Precision Sampling, Baton Rouge, LA). Samples were mixed gently for 15 min to aid gas dissolution as described in Selden et al. (2019). At the end of mixing, any remaining bubble was removed using a syringe. Incubation bottles were then transferred to on-deck incubators equipped with flow-through seawater and screens to maintain approximate temperature and light conditions.

While concerns have been raised regarding contamination by some commercially available <sup>15</sup>N<sub>2</sub> stocks leading to false positives (Dabundo et al. 2014; White et al. 2020), this issue has never been reported for <sup>15</sup>N<sub>2</sub> gas from Cambridge Isotope Laboratories. Furthermore, for each <sup>15</sup>N<sub>2</sub> gas tank used in this study, <sup>15</sup>N enrichment of the particulate N (PN) pool was undetectable in some samples despite significant plankton biomass, which would not have been possible if <sup>15</sup>N-labeled DIN were present. We are therefore confident that the rates

presented here did not result from <sup>15</sup>N contamination of gas stocks. All incubations lasted approximately 24 h and so represent daily rates.

### Sample processing, storage, and analysis

After the ~ 24 h incubation period, a 6 mL aliquot from each incubation bottle was transferred to a 12 mL He-purged Exetainer™ using a gas-tight syringe (Hamilton Co., Reno, NV). To prevent microbial activity in exetainers, 50 μL ZnCl<sub>2</sub> (50%, w/v) was also injected. Samples were stored upside-down, such that the liquid rather than the gas was in contact with the septum, and ground-shipped to Princeton University for analysis. The isotopic enrichment of the N<sub>2</sub> gas was measured on a Europa 20/20 Isotope Ratio Mass Spectrometer (IRMS) in continuous flow mode within 2.5 months of sample collection. The remaining sample was filtered onto pre-combusted (450°C for 2 h) 0.3 μm GF75 glass fiber filters (Advantec MFC, Inc., Dublin, CA) and stored frozen (4°C) until analysis at Old Dominion University.

Analyte can be lost from gas samples stored in Exetainer™ vials due to leakage on monthly timescales (Laughlin and Stevens 2003). Absolute N<sub>2</sub> mass in the exetainers is not pertinent to Eq. 1. However, the intrusion of isotopically light atmospheric N<sub>2</sub> into vials can dilute the isotopic enrichment of sample N<sub>2</sub> (White et al. 2020). While <sup>15</sup>N<sub>2</sub> samples in this study were not stored submerged, as is advisable (White et al. 2020), exetainers were over-pressurized relative to the atmosphere. Consequently, gas intrusion was likely minor.

Triplicate samples were collected to measure the ambient <sup>15</sup>N enrichment and mass of PN. These samples were filtered onto pre-combusted (450°C for 2 h) 0.3 μm GF75 glass fiber filters and stored frozen (– 20°C) until analysis. Ambient PN sample were processed and analyzed separately from isotopically-enriched samples in order to avoid contamination.

Samples for initial and final PN enrichment and mass were analyzed on a Europa 20/20 IRMS equipped with an automated nitrogen and carbon preparation module. The linearity of the IRMS response was evaluated daily using standards ranging from 1.17 to 100 μg N. Samples with filter mass below the lowest standard in the linear range on the day they were run were discarded. PN filter mass ranged from 3.3 to 87.1 μg (mean 15.7 μg). See Table S4 for these values.

### Rate calculation

N<sub>2</sub> fixation rates (NFR) were calculated using Eq. 1 (Montoya et al. 1996).

$$\text{NFR} = \frac{A_{\text{PN}t=f} - A_{\text{PN}t=0}}{A_{\text{N}_2} - A_{\text{PN}t=0}} \times \frac{[\text{PN}]}{\text{time}} \quad (1)$$

where  $A_{\text{PN}t=0}$  and  $A_{\text{PN}t=f}$  represent the atom-% enrichment of <sup>15</sup>N in the PN pool at the initial and final time points of the incubation period,  $A_{\text{N}_2}$  represents the <sup>15</sup>N enrichment of the N<sub>2</sub> pool, and [PN] is the PN concentration. Here, we used the

initial PN mass to calculate PN concentration rather than the mean value across time because initial PN volume was measured with greater accuracy and precision than that at the final time point. If the  $^{15}N_2$  sample enrichment was  $< 2.0$  atom-% for a given incubation, no rate was calculated. If the  $^{15}N_2$  sample for a given incubation was lost or damaged, the mean enrichment achieved across the study ( $5.14 \pm 2.03$  atom-%) was used to calculate NFR. The range of  $^{15}N_2$  enrichments in this study was 2.1–15.4 atom-%.

As shown in Eq. 1, NFRs are a function of both the relative importance of  $N_2$  to PN turnover and the absolute amount of PN in a given sample. To facilitate intersystem comparison of diazotroph activity, the specific uptake rate of  $N_2$  by particles in the incubation bottle is therefore also reported (Eq. 2).

$$\text{Specific uptake rate} = \frac{A_{PNt=f} - A_{PNt=0}}{A_{N_2} - A_{PNt=0}} \times \frac{1}{\text{time}} \quad (2)$$

This value can be conceptualized as the relative activity of diazotrophs within a water parcel, i.e., the fraction of total community PN that is derived from  $N_2$  fixation, or as the inverse of PN turnover time if, hypothetically, all PN were diazotroph-derived (Glibert and Capone 1993).

The error reported is the standard deviation of three replicate incubations. Limits of detection (LOD) were calculated for each individual rate following Eq. 3. The minimum detectable difference between initial and final PN  $^{15}N$  atom-% ( $\min\Delta A_{PN}$ ) is assumed to be equal to three times the standard deviation of eight  $12.5 \mu\text{g N}$  standards measured daily alongside samples (White et al. 2020).

$$\text{LOD} = \frac{\min\Delta A_{PN}}{(A_{N_2} - A_{PNt=0})} \times \frac{[PN]}{\text{time}} \quad (3)$$

### Statistics

Replicated rate measurements were grouped by locality and compared via one-way ANOVA. Significance was determined by comparing the computed test statistics ( $F$  values) to those derived from 10,000 random permutations of the data. This approach is robust regardless of whether the residuals of the data are normally distributed (Manly 2007).

### Sequencing and quantification of *nifH*

The diazotroph communities present during the study period were investigated by sequencing and quantifying the abundance of *nifH*, a structural gene of the enzyme that mediates  $N_2$  fixation. For *nifH* quantification, water was collected from Niskin bottles into acid-cleaned 4 L Cubitainers® (Qorpak, Clinton, PA) and filtered onto  $1.2 \mu\text{m}$  polyethersulfone membrane filters (Sterlitech, Kent, WA) using acid-cleaned tubing. Filters were immediately transferred to sterile 2 mL tubes and submerged in RLT Plus buffer (Qiagen). Samples were frozen in liquid nitrogen immediately after filtering and, once ashore, stored at  $-80^\circ\text{C}$  until analysis. As

many non-colonial and non-symbiotic diazotrophs are smaller than  $1.2 \mu\text{m}$ , *nifH* was sequenced from smaller size-fraction samples ( $0.03 \mu\text{m}$ ) collected at a subset of stations. The sampling protocol was as described above except that water was collected from Niskin bottles into acid-cleaned 10 L LDPE carboys and samples were flash-frozen dry.

Nucleic acids (DNA and RNA) were co-extracted from samples using the AllPrep DNA/RNA Mini Kit (Qiagen) following manufacturer's instructions with the addition of a bead-beater step. To quantify the abundance of *nifH* transcripts, indicative of diazotrophic activity, RNA was treated with amplification-grade DNase I (Invitrogen, Carlsbad, CA) and reverse-transcribed using SuperScript IV First-Strand Synthesis System (Invitrogen, Carlsbad, CA). We note that the efficiency of reverse transcription reactions can vary with RNA concentration and the reverse transcriptase employed (Bustin et al. 2014). To minimize intra-study variability, we maximized the concentration of RNA in our reactions, and used the same enzyme for reverse transcription with all samples from which transcript abundance was quantified. For small size-fraction samples (used to determine relative abundance of *nifH* transcripts), reverse-transcription was performed using the SuperScript III First-Strand Synthesis System (Invitrogen, Carlsbad, CA) and the *nifH3* primer (Zehr and Turner 2001) in place of random hexamers.

To assess diazotroph community diversity, DNA and the product of RNA reverse-transcription (cDNA) were amplified via nested polymerase chain reaction (PCR) using degenerate primers (Zehr and Turner 2001), with the adjustment that Illumina overhang adapter sequences for two-step amplicon sequencing ([http://www.illumina.com/content/dam/illumina-support/documents/documentation/chemistry\\_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf](http://www.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf)) were added to the second round PCR primers (*nifH1* and *nifH2*). PCR products were then gel-purified using a QIAquick Gel Extraction Kit (Qiagen, Germany), subject to index PCR ([http://www.illumina.com/content/dam/illumina-support/documents/documentation/chemistry\\_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf](http://www.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf)), and sequenced on an Illumina MiSeq platform using a  $2 \times 300$  bp kit. Sequences were subsequently demultiplexed, imported into CLC Genomics Workbench (Qiagen, Germany) as pairs, trimmed, and merged. To standardize the number of reads per sample, 11,000 random reads were extracted from each prior to *nifH* community composition analysis using the minimum entropy decomposition pipeline (Eren et al. 2015). The nucleotide sequences for MED-defined operational taxonomic units were identified by BLAST (Altschul et al. 1990) against an in-house *nifH* sequence database. Raw sequence data is available in the Sequence Read Archive (National Center for Biotechnology Information) under the BioProject identifier PRJNA683637.

The abundances of dominant diazotrophic phylotypes were quantified in surface waters via quantitative PCR (qPCR) using

2 × TaqMan Fast Advanced Master Mix (Applied Biosystems) and custom primer/probe sets (Table S1). These targeted *nifH* sequences specific to *Trichodesmium* spp., *Richelia intracellularis* (Het-1, a symbiont of the diatom *Rhizosolenia* sp.) and two UCYN-A sublineages, UCYN-A1 (Church et al. 2005a) and UCYN-A2 (Thompson et al. 2014). The abundance of *Braarudosphaera bigelowii*, a known host for UCYN-A, was also quantified using an 18S rRNA sequence (Thompson et al. 2014). Primer/probe sets for *Trichodesmium* spp. and Het-1 were redesigned for this study based on previously published gene sequences (Table S1).

DNA samples were diluted 1 : 10 with nuclease-free DEPC-treated water (Millipore, Germany) before qPCR. Reactions (20  $\mu$ L total volume) were prepared by combining the following: 10  $\mu$ L 2 × TaqMan Fast Advanced Master Mix, 4  $\mu$ L nuclease-free water, 1.6  $\mu$ L 5  $\mu$ M forward and reverse primers, 0.8  $\mu$ L 5  $\mu$ M probe, and 2  $\mu$ L 1 : 10 DNA or cDNA. Serially-diluted (5 × 10<sup>-2</sup> – 10<sup>-9</sup> ng DNA reaction<sup>-1</sup>) synthetic plasmids (GeneWiz, USA) and no-template controls were run in triplicate alongside samples. Reactions proceeded using a StepOnePlus Real-Time PCR System and StepOne Software v. 2.2.2 (Applied Biosystems) following manufacturer's specifications. Effective LODs and limits of quantification were determined for qPCR data by assuming that the minimum detectable and quantifiable concentrations are 3 and 10 copies reaction<sup>-1</sup>, respectively (Bustin et al. 2020), as follows:

$$\text{Effective limit of detection} = \frac{3}{V_t} \times \frac{V_d}{V_f} \quad (4)$$

$$\text{Effective limit of quantification} = \frac{10}{V_t} \times \frac{V_d}{V_f} \quad (5)$$

where  $V_t$ ,  $V_d$ , and  $V_f$  represent the volume of the DNA template ( $\mu$ L extract reaction<sup>-1</sup>), the DNA extract ( $\mu$ L), and the total seawater filtered to obtain the extract (L).

### Supervised machine learning approach

Empirical models of NFR, as functions of a varied suite of hydrographic predictors, were derived via random forest regression (Breiman 2001) using the scikit-learn package (Pedregosa et al. 2011) in Python. Random forest regression is an ensemble machine learning approach. Briefly, the training data are iteratively sub-sampled (with replacements) and each sub-sample is used to fit a decision tree. Their predicted values are then averaged, improving model accuracy and reducing over-fitting (Breiman 2001). The environmental predictors applied included sample depth, seafloor depth, temperature, salinity, Chl *a* concentration derived from in situ fluorescence, and N + N and SRP concentrations. These parameters were chosen based on the depth of their coverage and theoretical applicability to predicting NFRs. With this approach, NFRs are predicted from multi-dimensional trends in the input parameters, and thus accounts for relationships

among input variables (e.g., N + N : SRP). Seafloor depth at each sampling point was taken as the mean depth within 0.1° from the ETOPO 1-Arc Minute Relief Model (Amante and Eakins 2009).

A model trained on temperature, Chl *a* concentration and seafloor depth was used to predict sea surface NFRs from an independent, external dataset composed of Moderate Resolution Sea Imaging Spectroradiometer (MODIS) sea surface temperature and Chl *a* concentrations (NASA Goddard Space Flight Center 2018 Reprocessing) and ETOPO 1-Arc Minute Global Relief Model seafloor depth (Amante and Eakins 2009). Salinity was not included here because it did not significantly augment model performance and available satellite salinity data (e.g., Meissner et al. 2019) were at significantly lower resolution than the other data. To estimate NFR from CTD depth profiles, the model was trained on depth, seafloor depth, temperature, salinity and Chl *a* concentration. The relative importance of N + N and SRP in predicting NFRs was assessed by adding these parameters into the latter model and predicting NFR from Niskin bottle files where nutrient concentrations were measured.

### Data accession and transformation

Models were trained on NFRs calculated for each individual incubation. This ensured that the model incorporated rate measurement variability and served to maximize training data volume. Each NFR was associated with the mean hydrographic parameters of all Niskin bottles from which sample water was collected. Rates were logarithm-transformed (base 10), following Sammartino et al. (2018), to balance the importance of extreme values among the training data. All rates below the LOD were assumed to be equal to the median LOD (0.54 nmol N L<sup>-1</sup> d<sup>-1</sup>). The median value was used rather than the individual LODs for each rate measurement to avoid scaling undetectable rates with PN concentration. As some detectable values in the training set were below the median LOD, trends among low NFRs (on the order of 10<sup>-1</sup> nmol N L<sup>-1</sup> d<sup>-1</sup>) may have been obscured. Data were subsequently split at random into training/testing (80%) and validation (20%) subsets. Validation data were withheld from model construction. All predictive data were scaled based on the training/testing data.

To reduce noise in the depth profile predictor dataset, hydrographic parameters were binned at 1 m intervals.  $N_2$  fixation rate profiles were predicted from the average water column profile at each station, calculated as the average of all binned casts ( $n = 1-3$ ). Casts at each station were collected within ~ 3 h of one another. To reduce the computational expense of running the model, NFRs were predicted for every other depth bin (e.g., every 2 m).

Processed MODIS satellite data were accessed from <https://oceandata.sci.gsfc.nasa.gov> on March 19, 2020. Data are mean values from August 12 to 19, 2016. This time period was selected because it is the 8-day mean available from NASA Goddard Space Flight Center (2018 Reprocessing) that most

closely aligned with the study period (August 7–16, 2016). Satellite data and bathymetry model (Amante and Eakins 2009) output were associated based on latitude/longitude and binned at 0.05° resolution.

### Random forest regression

For each model variant, hyperparameters (e.g., total number of decision tree estimators, maximum tree depth) were tuned using a grid search approach with K-fold cross validation (Pedregosa et al. 2011). With this approach, the data are divided into equal “folds” ( $n = 9$ ) and each is used in turn to test the performance of a model built on the other  $n-1$  folds. Here, ninefolds were used to ensure that each represented > 20 rate measurements. For each run, model hyperparameters were varied. The best combination of hyperparameters was selected as that with the highest coefficient of determination ( $R^2$ ). This process was repeated  $n$  times such that each fold was used to test the model and obtain optimized hyperparameters once (Table S2). The final model was then refit using the best hyperparameters for the entire dataset. Model performance was assessed using the validation dataset withheld at the beginning via linear regression. Model output is available in Table S3.

### Areal (depth-integrated) $N_2$ fixation rate calculations

Areal NFRs were calculated using (1) measured NFRs and (2) predicted NFRs from Model 1. Rates were trapezoidally depth-integrated through the upper 100 m of the water column or to the seafloor depth, whichever was shallower. Areal rates were only calculated from direct measurements if at least two replicated NFRs were available within the integration depth range. The 100 m threshold was chosen because the density of direct measurements made below 100 m was relatively low. While areal rates are often calculated to a given isolume rather than a given depth, we deemed this approach to be inappropriate here due to the presence of non-cyanobacterial (potentially non-phototrophic) diazotrophs (see below).

The following assumptions were made to calculate areal NFRs: (1)  $N_2$  fixation was assumed to be constant from the surface (0 m) to the depth of the shallowest rate, which was always within the surface mixed layer. (2) At very shallow stations (seafloor depth < 50 m), which were typically well-mixed,  $N_2$  fixation was assumed to be constant from the depth of the deepest rate to the seafloor. At all other stations, NFR was assumed to be equal to 0  $\text{nmol N L}^{-1} \text{d}^{-1}$  at the bottom of the integration depth range if there were no rates at or below this depth. We note that detectable NFRs were often observed below this depth at deep stations, particularly those within the Gulf Stream. At these stations, rates were available below 100 m and this assumption was consequently unnecessary. (3) To avoid over-estimation, any measured or predicted NFRs below the detection limit or the mean measured rate LOD (1  $\text{nmol N L}^{-1} \text{d}^{-1}$ ), respectively, were assumed to be equal to 0  $\text{nmol N L}^{-1} \text{d}^{-1}$ .

## Results and discussion

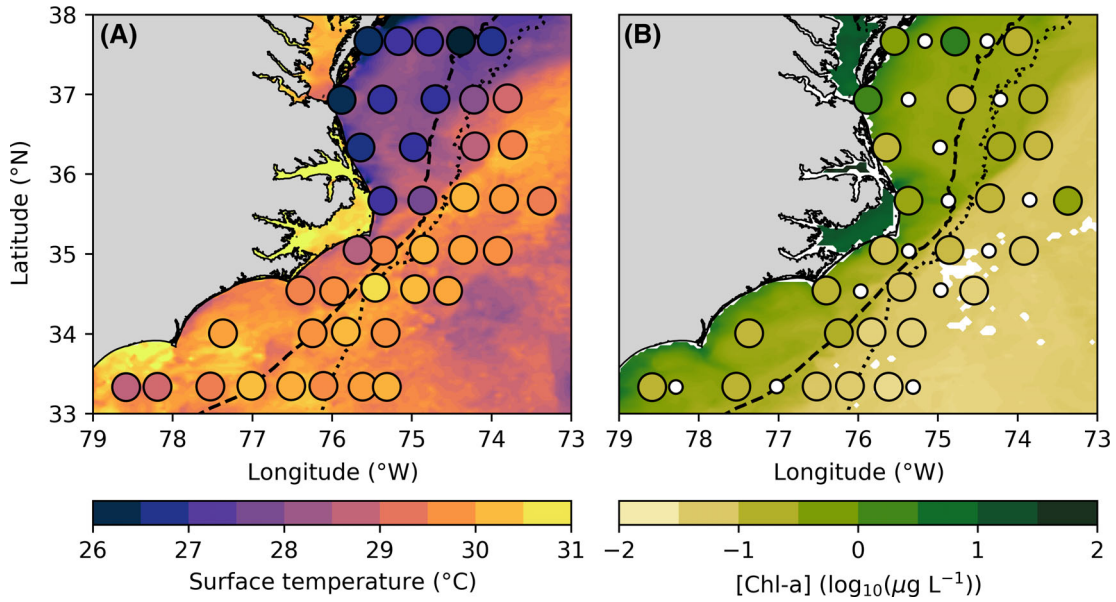
### Regional hydrography

Three water masses converge near Cape Hatteras, NC (Fig. 1B), which is designated as the boundary between the Mid- and South Atlantic Bights. Cool (surface temperature < 29°C), low salinity (31.5–33.5) shelf waters formed in the Mid-Atlantic Bight flow southward (Churchill and Berger 1998). This mass meets with warm (surface temperature > 29°C) and high salinity (> 34.5) northward-flowing waters from the South Atlantic Bight continental shelf and Gulf Stream (Fig. 2A). The Gulf Stream separates from the continental boundary at Cape Hatteras—thus ceasing to be a western boundary current—and flows to the northeast (Csanady and Hamilton 1988). The exact point at which these water masses meet and mix is variable, as is the separation point of the Gulf Stream. At their convergence point, shelf waters are entrained by the Gulf Stream, transporting shelf material seaward (Churchill and Berger 1998). This flow pattern results in the formation of a gyre along the continental slope between the Mid-Atlantic Bight shelf break and the Gulf Stream (Csanady and Hamilton 1988). At the time of this study, surface Chl *a* was elevated within Mid-Atlantic Bight slope waters relative to the Gulf Stream—on par with concentrations observed across the shelf (Fig. 2B).

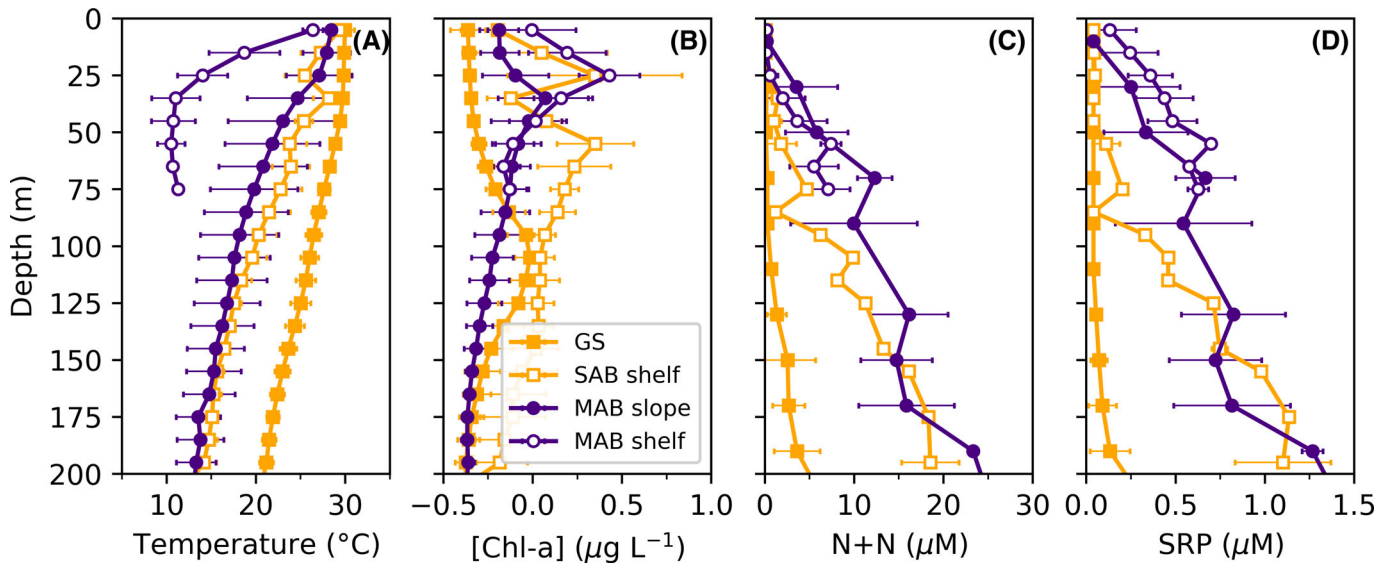
Both Mid- and South Atlantic Bight shelf waters are subject to inputs from the Gulf Stream and freshwater sources. Gulf Stream waters intrude directly on the South Atlantic Bight, or may be transported to the Mid- and South Atlantic Bight shelf via eddies and meanders (Atkinson 1977). At the time of this study, low surface salinities (28.6 and 30.6, respectively) were observed at Stas. 14 and 15 on the inner shelf north of Cape Hatteras. These anomalies were likely due to outflow from Albemarle and Pamlico Sounds via Oregon Inlet (Fig. 1B), which can deliver terrestrial carbon and nutrients to the shelf (Churchill and Berger 1998).

Mid-Atlantic Bight shelf waters are characterized by a sharp, shallow thermocline, which occurred around 25 m at the time of this study (Fig. 3A). Below the mixed layer on the Mid-Atlantic Bight shelf, the water temperature was typically ~ 10°C. South Atlantic Bight shelf waters and Mid-Atlantic Bight slope waters were warmer by comparison, reaching ~ 15°C in deeper (> 100 m) waters (Fig. 3A), though significant variability was observed in mean hydrographic characteristics of Mid-Atlantic Bight slope water (Fig. 3). This variability was likely the result of significant water mass mixing in the portion of the Mid-Atlantic Bight slope captured during our cruise. Gulf Stream waters were warm (Fig. 3A) and oligotrophic (Fig. 3C,D), typically exhibiting deeper Chl *a* maxima than in shelf or Mid-Atlantic Bight slope waters (Fig. 3B).

Nitrate plus nitrite was drawn down in surface waters across the region (Fig. 3C). However, significant concentrations (< 5  $\mu\text{M}$ ) were typically observed in bottom waters at



**Fig 2.** Mean MODIS (NASA Goddard Space Flight Center 2018 Reprocessing) sea surface temperature (A) and chlorophyll *a* (Chl *a*) fluorescence (B) from August 12th to 19th, 2016, overlain by surface (< 3 m) temperature and extracted Chl *a* values at each station. White dots represent stations where Chl *a* concentration was not measured directly. The dashed and dotted lines indicate the 200 and 2000 m isobaths, representing the edges of the shelf and slope, respectively.



**Fig 3.** Mean temperature (A), chlorophyll *a* (Chl *a*; B), nitrate plus nitrite (N + N; C) and soluble reactive phosphorus (SRP; D) for Gulf Stream (GS) stations (sea surface temperature > 29°C, seafloor depth > 200 m; closed orange squares), South Atlantic Bight (SAB) shelf stations (latitude < 35.5°N, seafloor depth < 200 m; open orange squares), and Mid-Atlantic Bight (MAB) shelf (latitude > 35.5°N, seafloor depth < 200 m; open indigo circles) and slope (latitude > 35.5°N, sea surface temperature < 29°C, seafloor depth < 200 m; closed indigo circles) stations. All values were binned at 10 m intervals except for SRP and N + N in slope waters, which were binned at 20 m intervals. Horizontal bars represent the standard deviation of values within a given bin, i.e., the variability of each parameter within a given bin.

Mid-Atlantic Bight shelf stations, below ~ 25 m in Mid-Atlantic Bight slope waters, and below ~ 75 m in South Atlantic Bight shelf waters (Fig. 3C). In contrast, SRP was detectable in surface waters along the Mid-Atlantic Bight shelf, and elevated

(< 0.25  $\mu\text{M}$ ) throughout the water column at Mid-Atlantic Bight shelf and slope stations (Fig. 3D). In Gulf Stream waters, SRP—like N + N—remained low throughout the upper water column at slope stations and until ~ 100 m at shelf stations (Fig. 3C).

### Direct $N_2$ fixation rate measurements

Undoubtedly the most striking finding of this study was an  $N_2$  fixation hotspot near Cape Hatteras, NC, that appeared to extend across the shelf to the slope (Fig. 4B,F,I). Stations nearest the coast along the 36.3 and 35.6°N transects (Stas. 14 and 15), where low sea surface salinity anomalies were observed (see above), exhibited NFRs exceeding  $100 \text{ nmol N L}^{-1} \text{ d}^{-1}$  (mean =  $300 \pm 160 \text{ nmol N L}^{-1} \text{ d}^{-1}$ ,  $n = 7$  surveyed depths) throughout the water column (see Table S4). Here,  $N + N$  was undetectable and SRP was low but typically measurable ( $> 0.5 \mu\text{M}$ ; Table S4). Rates remained elevated at the outer shelf and shelf break (Stas. 12, 13, 17; mean =  $220 \pm 180 \text{ nmol N L}^{-1} \text{ d}^{-1}$ ,  $n = 9$  depths). Rates were lower in slope waters along hotspot transects, reaching  $7.1 \pm 2.0$  and  $51 \pm 18 \text{ nmol N L}^{-1} \text{ d}^{-1}$  ( $n = 3$ ) in surface waters at Stas. 11 and 19, respectively. Both  $N + N$  and SRP were undetectable in the upper water column at these stations.

In comparison, the two transect to the north of the hotspot (37.7 and 36.9°N) were characterized by undetectable to low NFRs (Fig. 4A,E; Table S4). Elevated NFRs were, however, observed in surface waters near the shelf break (5.6 and  $18.3 \text{ nmol N L}^{-1} \text{ d}^{-1}$  at Stas. 1 and 10, respectively). In South Atlantic Bight shelf waters south of the hotspot, NFRs in the upper 150 m averaged  $13.9 \pm 6.2$  ( $n = 16$  surveyed depths). Rates in Gulf Stream waters off the South Atlantic Bight shelf break averaged  $10.3 \pm 8.8$  ( $n = 25$ )  $\text{nmol N L}^{-1} \text{ d}^{-1}$  ( $< 150$  m). Too few NFR measurements have, to date, been reported from the South Atlantic Bight to comment on whether these rates represent an annual peak. However, NFRs in the Mid-Atlantic Bight are typically highest in the Spring (Mulholland et al. 2019). As this study was performed in August, diazotrophy in the northern Mid-Atlantic Bight waters may already have commenced their seasonal decline.

Rates of  $N_2$  fixation (Eq. 1) are the product of PN concentration and the rate of specific  $N_2$  uptake (Eq. 2), i.e., the rate at which N atoms are transferred from the  $N_2$  pool to the PN pool. Consequently, NFRs may increase if either biomass or diazotroph community activity increases. Changes in the latter may be driven by changes in the proportion of diazotrophs within a given community, an increase in cell-specific diazotrophic activity, or both. Examining the rate of specific  $N_2$  uptake thus offers a means of assessing diazotroph activity that is independent of biomass, which is advantageous when analyzing  $N_2$  fixation across steep gradients in PN concentration. In this study, PN was higher across the continental shelf and in Mid-Atlantic Bight slope waters (Fig. S1; Table S4), contributing to higher rates inshore throughout the region (Fig. 4). However, the relative contribution of diazotroph communities to PN turnover was an order of magnitude greater in N-deplete offshore waters than in shelf waters to the north and south of the hotspot (Table 1). This finding suggests that the relative importance of diazotrophy to community N supply is greater in more oligotrophic waters, and thus supports the historical paradigm (Carpenter and Capone 2008).

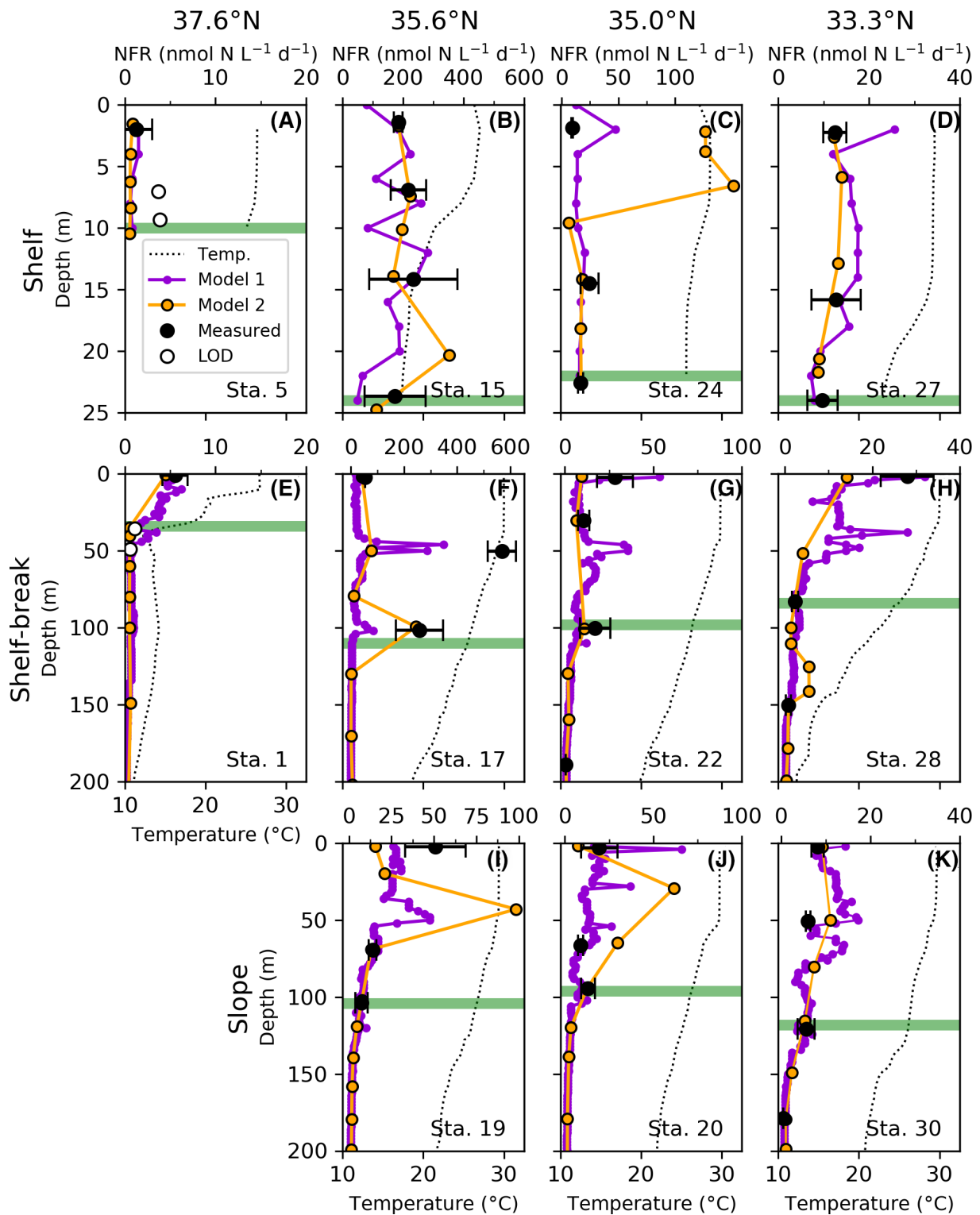
Specific uptake rates at hotspot stations (12–15, 17) ranged from  $0.026 \pm 0.016 \text{ d}^{-1}$  (Sta. 13 at 25 m,  $n = 3$ ) to  $0.57 \pm 0.05 \text{ d}^{-1}$  (Sta. 17 at 50 m,  $n = 3$ ). These values are more than an order of magnitude greater than anywhere else (Tables 1, S4). This observation suggests that high NFRs at hotspot stations were not driven by high PN concentrations (Table S4), but rather by an increase in the relative importance of  $N_2$  fixation to PN turnover. Along the hotspot transects, there was no significant difference between specific uptake rates on and off the shelf (Table 1). Thus, despite the strong hydrographic gradients here (e.g., temperature; Fig. 2A), diazotrophy remained favorable across the shelf and slope adjacent to Cape Hatteras. This observation may have resulted from water mass mixing associated with the frontal zone here. For example, if  $N_2$  fixation were locally iron-limited, mixing could theoretically have alleviated this limitation as follows: If terrestrially-derived iron (from freshwater/estuarine inputs) is maintained in shelf waters by humics and other weak ligands (e.g., Laglera and van den Berg 2009), and if oligotrophic Gulf Stream waters support siderophore production, then strong iron-binding siderophores would ligate iron from the shelf, making it more bioavailable.

Reduced  $N + N$  : SRP ratios along hotspot transects may have enhanced cell-specific  $N_2$  uptake rates (Knapp et al. 2012) or the relative ecological success of diazotrophs and diazotroph-hosts. Mean  $N + N$  : SRP ratios in the upper 150 m were significantly lower at hotspot vs. non-hotspot stations on both the shelf (mean  $N + N$  : SRP : hotspot =  $0.77 \pm 0.87$ , other shelf stations =  $16.5 \pm 14.6$ ; one-way ANOVA,  $n_1 = 26$ ,  $n_2 = 278$ ,  $F = 15.93$ ,  $p = 0.0002$  based on 10,000 random permutations) and the slope (mean  $N + N$  : SRP : hotspot =  $10.5 \pm 9.5$ , other slope stations =  $17.8 \pm 14.4$ ; one-way ANOVA,  $n_1 = 50$ ,  $n_2 = 468$ ,  $F = 12.21$ ,  $p = 0.0005$  based on 10,000 random permutations). Reduced  $N + N$  : SRP ratios along hotspot transects could have resulted from a combination of ecological (e.g., altered community structure or activity) and chemico-physical/biological (e.g., altered  $N + N$  or SRP uptake due to relief of co-limitation by an essential element) factors.

Throughout the region, NFR vertical profiles displayed relatively little variability within shelf waters (Fig. 4A–D). At the shelf break and slope, where the water column exhibited greater structure, rates were typically highest in surface waters and decreased with depth (Fig. 4E–K), although NFR maxima were frequently observed near Chl *a* maxima. These trends were driven by both changes in PN concentration and specific  $N_2$  uptake rates (Table S4).

At select stations (10, 20, 34, 39), NFRs were measured in deep waters (200–500 m depth).  $N_2$  fixation was detectable in all deep waters surveyed (Table S4). The mean NFR measured below 200 m was  $0.90 \pm 0.57 \text{ nmol N L}^{-1} \text{ d}^{-1}$  ( $n = 6$  surveyed depths). For these deep samples,  $N + N$  concentrations exceeded  $5 \mu\text{M}$ ; particulate C concentrations ranged from 1.9 to  $7.1 \mu\text{M}$  (Table S4). These values are consistent with the observation of low NFRs ( $\sim 1 \text{ nmol N L}^{-1} \text{ d}^{-1}$ ) elsewhere in the





**Fig 4.** Representative  $N_2$  fixation rate (NFR) profiles arranged left-to-right by latitude and top-to-bottom by distance from shore. Black dots = measured NFRs. Error bars =  $\pm 1$  standard deviation. White dots = the limit of detection (LOD) where  $N_2$  fixation was measured but not detected. Model 1 (purple) predicts NFR from depth, seafloor depth, temperature, salinity and chlorophyll *a* (Chl *a*) concentration. Model 2 (orange) predicts NFR from  $N + N$  and SRP concentrations in addition to the parameters included in Model 1. The dotted line indicates temperature. The green bar indicates the depth of the primary Chl *a* maximum.

**Table 1.** Comparison of shelf (< 200 m) and offshore (> 200 m) replicated specific N<sub>2</sub> uptake rates (NFRs) in the upper 150 m (one-way ANOVA).

Region	Mean specific N <sub>2</sub> uptake rates on shelf (d <sup>-1</sup> )*	Mean specific N <sub>2</sub> uptake rates offshore (d <sup>-1</sup> )*	Degrees of freedom	Test statistic (F)	p value <sup>†</sup>
North of hotspot (36.5–38°N)	3.55 ± 5.94 × 10 <sup>-4</sup>	5.87 ± 7.51 × 10 <sup>-3</sup>	16	6.14	0.047
NFR hotspot (36.5–35.5°N)	1.14 ± 0.79 × 10 <sup>-1</sup>	1.59 ± 0.19 × 10 <sup>-1</sup>	20	0.47	0.498
South of hotspot (33.0–35°N)	9.85 ± 4.57 × 10 <sup>-3</sup>	2.19 ± 1.86 × 10 <sup>-2</sup>	36	5.70	0.017

\*A value of zero was used where N<sub>2</sub> fixation was not detected. Mean error given as ± 1 standard deviation.

<sup>†</sup>Based on 10,000 random permutations.

interior ocean, particularly under organic-carbon replete conditions (Moisander et al. 2017 and references therein). Supplementing organic carbon in incubations has been observed to increase NFRs in mesopelagic (e.g., Benavides et al. 2015) and coastal (e.g., Rahav et al. 2016) waters. Organic carbon additions may provide a substrate for heterotrophic respiration, support the formation of high-metabolic microenvironments, or both. Recently, Pedersen et al. (2018) demonstrated that estuarine non-cyanobacterial diazotrophs can colonize natural particles, although pre-colonization by other bacterioplankton appears to be requisite. Similarly, Farnelid et al. (2019) found that suspended particles are loci for diazotrophs in the North Pacific Subtropical Gyre. We hypothesize that the low NFRs observed in deep waters during this study were supported by organic carbon inputs from the North Atlantic continental shelf.

### Diazotroph diversity and abundance

Different diazotrophic clades display disparate ecophysiologicals. To determine the diazotroph taxa present in the study region, we sequenced *nifH* in a subset of surface samples (Fig. S2). We found that *Trichodesmium* spp. dominated the *nifH* DNA pool (> 70% of reads) in Gulf Stream (Stas. 19 and 22) and Mid-Atlantic Bight slope (Stas. 10 and 12) surface waters (Fig. S2A). Most of these reads were identified as *Trichodesmium thiebautii*; however, *T. erythraeum* dominated the DNA pool at a South Atlantic Bight shelf break station (22). In the NFR hotspot on the shelf (Sta. 15), *nifH* gene associated with *Anabaenopsis* sp. and unidentifiable Cyanobacteria abounded (Fig. S2A) in the low salinity (28.6) surface waters. These organisms were likely transported from the nearby Albemarle and Pamlico Sounds on to the shelf (Mulholland et al. 2012).

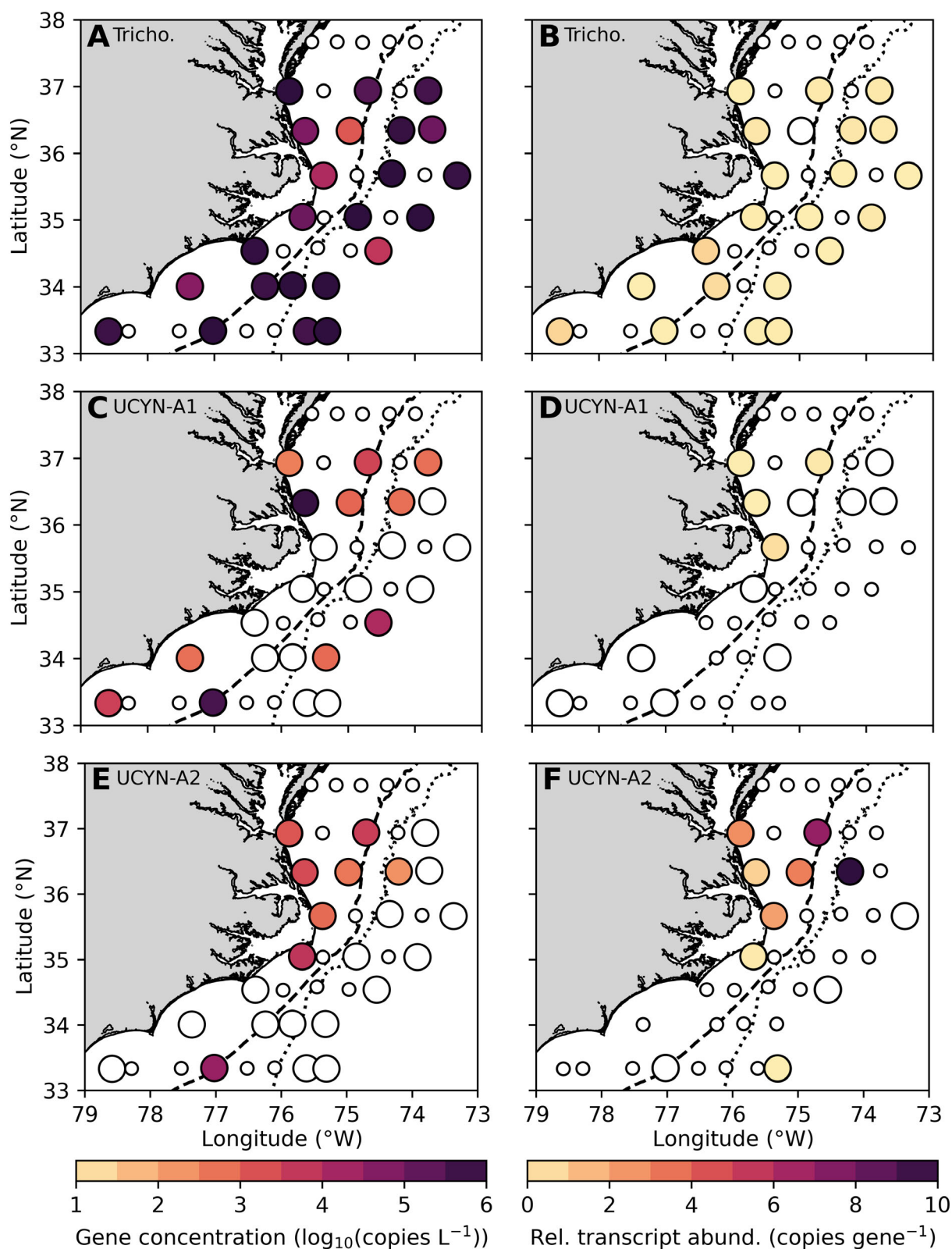
We also sequenced the *nifH* transcript pool (Fig. S2B). UCYN-A dominated *nifH* transcript reads at both shelf and shelf break hotspot stations (Stas. 15 and 12; Fig. S2B). At Sta. 15, the relative abundance of UCYN-A transcripts was greater at the Chl *a* maximum (97%) than at the surface (64%). At the surface, *nifH* transcripts associated with the freshwater diazotroph

*Anabaenopsis* sp. (22%), gammaproteobacteria (7%), and *T. thiebautii* (6%) were also present. *Anabaenopsis* sp. is known to continue fixing N<sub>2</sub> at relatively high salinity (~20; Moisander et al. 2002), and consequently may have contributed to the high NFRs observed at Stas. 14 and 15, which were slightly brackish (28.6–30.6). However, no sequences from this preferentially freshwater organism were recovered in the more marine waters offshore. Consequently, it is unlikely that it contributed substantially to the observed hotspot.

In Mid-Atlantic Bight slope waters to the north (Sta. 10), *Cyanothece* sp. sequences dominated the transcript pool (67%), although gammaproteobacterial *nifH* transcripts were present as well (11%). However, it should be noted that this sample was collected at night when *Cyanothece* sp. are known to be active (Colón-López et al. 1997) and UCYN-A and *Trichodesmium* spp. are not (Church et al. 2005b; Muñoz-Marín et al. 2019). In South Atlantic Bight shelf break waters (Sta. 22) collected in the evening, UCYN-A (27%), gammaproteobacteria (36%), *T. erythraeum* (21%) and *T. thiebautii* (16%) were all actively transcribing *nifH*.

Given the prevalence of *Trichodesmium* spp. throughout the region and apparent activity of UCYN-A at the hotspot, the absolute abundance of *Trichodesmium* spp. and UCYN-A sublineages 1 and 2 *nifH* genes and transcripts were quantified in surface waters across the study area via qPCR (Table S5). *Trichodesmium* spp. were highly abundant throughout the region (mean *nifH* gene abundance = 10<sup>5</sup> copies L<sup>-1</sup>, *n* = 23 triplicated samples; Fig. 5A). *Trichodesmium* spp. *nifH* expression tended to be low relative to its abundance except in South Atlantic Bight shelf waters (Fig. 5B). UCYN-A1 and 2 were both abundant at hotspot stations (Fig. 5C,E). While the relative expression of UCYN-A1 *nifH* was low at these stations (Fig. 5D), the relative expression of UCYN-A2 *nifH* was high, ~650 transcripts gene<sup>-1</sup> on average (*n* = 6 replicated samples).

Diatom community analysis of surface waters detected a high relative abundance of *Rhizosolenia* sp., a potential diazotroph host (Villareal 1992), concentrated near the N<sub>2</sub> fixation hotspot on the shelf (Chappell, unpublished). Consequently, the



**Fig 5.** Abundance of *Trichodesmium* spp. (**A** and **B**) UCYN-A1 (**C** and **D**) and UCYN-A2 (**E** and **F**) *nifH* genes and transcripts (normalized to gene abundance) in surface (< 3 m) waters. Large open circles indicate stations where sequences were not detected. White dots represent stations where diazotroph abundances were not measured. Where transcripts but not their corresponding genes were detectable, transcript abundance was normalized to the effective limit of detection (Eq. 4). The dashed and dotted lines represent the 200 and 2000 m isobaths, respectively.

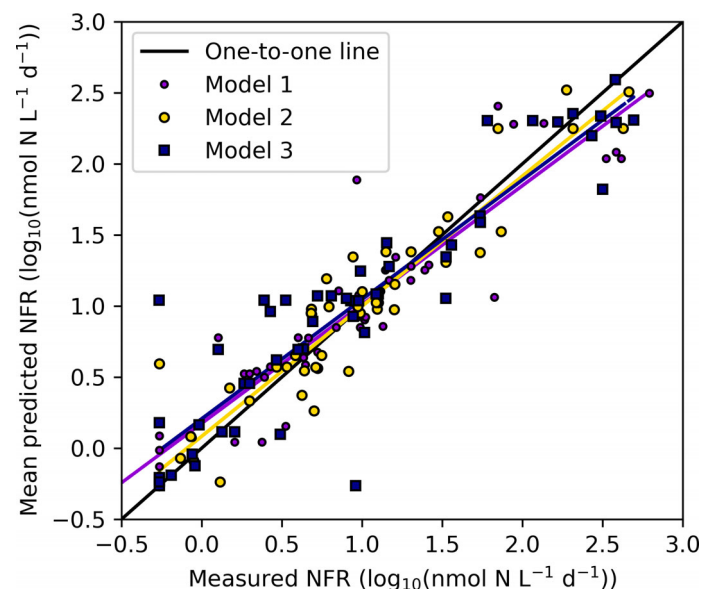
abundance of *nifH* sequences diagnostic of the *Rhizosolenia* sp. symbiont (Het-1) was also quantified. However, Het-1 was not detected at hotspot shelf stations and was at low abundance or undetectable throughout the entire study region (Fig. S3).

Additionally, UCYN-A can become dislodged from its host during filtration (Thompson et al. 2014). As free UCYN-A cells are small ( $< 1 \mu\text{m}$ ; Thompson et al. 2014), they may not have been caught on the filters that we used for qPCR ( $1.2 \mu\text{m}$ ). For this reason, we also assessed the abundance of the only identified UCYN-A host, *B. bigelowii* ( $7\text{--}10 \mu\text{m}$ ; Thompson et al. 2014). *B. bigelowii* was most abundant ( $> 10^4$  18S rRNA gene copies  $\text{L}^{-1}$ ) at coastal stations along the Outer Banks and at the shelf break along hotspot transects (Fig. S4; Table S5). Additionally, *B. bigelowii* was detected at some South Atlantic Bight shelf stations and in Gulf Stream waters.

### Machine learning-based estimates of $N_2$ fixation

#### $N_2$ fixation rate model performance

Three random forest regressor models were built to predict NFRs in the region during the study period from (1) depth profiles of physical (sample depth, seafloor depth, temperature and salinity) and biological (Chl *a* concentration) parameters, (2) depth profiles of physical, biological and biogeochemical (N + N and SRP concentrations) parameters, and (3) sea surface parameters (temperature and Chl *a* concentration) and seafloor depth. All three models performed well ( $R^2 > 0.80$ ) when tested against validation data withheld from model training (Fig. 6; Table S2), but tended to overpredict low values and underpredict high values, as has been previously observed in



**Fig 6.** Performance of  $N_2$  fixation rate (NFR) models on discrete validation datasets. Model 1 (physical and biological predictors):  $y = 0.84x + 0.18$ ,  $df = 50$ ,  $R^2 = 0.86$ ,  $p < 10^{-6}$ . Model 2 (physical, biological, and biogeochemical predictors):  $y = 0.92x + 0.08$ ,  $df = 45$ ,  $R^2 = 0.89$ ,  $p < 10^{-6}$ . Model 3 (sea surface variables and seafloor depth):  $y = 0.84x + 0.21$ ,  $df = 50$ ,  $R^2 = 0.83$ ,  $p < 10^{-6}$ .

machine learning-based NFR estimates (Tang et al. 2019). Greater fidelity at NFR extremes was achieved when nutrient concentrations (N + N and SRP) were included (Fig. 6). However, the coverage of nutrient data was low relative to the other parameters. Consequently, NFR predictions by Model 2 were at lower depth resolution than those from Model 1 (Fig. 4).

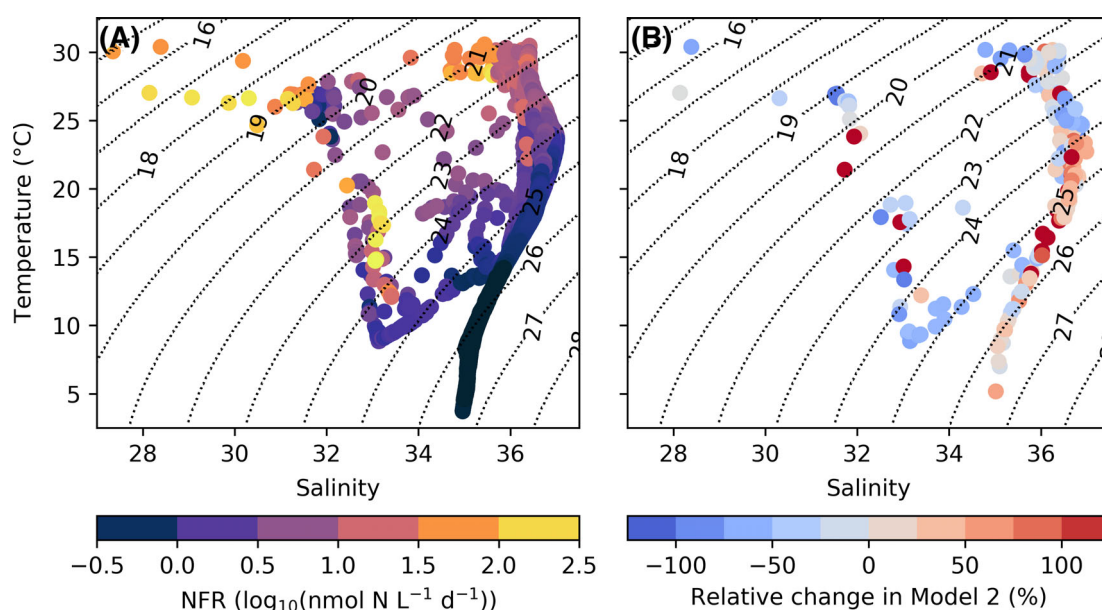
#### Modeled $N_2$ fixation profiles

Models 1 and 2 offer greater depth coverage of NFRs and suggest complexities in NFR profiles that could not have been resolved from direct measurements alone (Fig. 4). As observed from in situ measurements, NFRs at shallow shelf stations were typically uniform through the water column. At deeper stations, however, NFRs displayed significant variability. Narrow maxima were often observed at the bottom of the mixed layer and near the primary Chl *a* maximum (Fig. 4E–K). Below this point, both predicted and measured rates were typically low ( $< 1 \text{ nmol N L}^{-1} \text{ d}^{-1}$ ).

The highest rates predicted by Model 1 occurred in (1) warm, low salinity ( $< 31.5$ ) waters associated with freshwater outflow from Oregon Inlet (Stas. 14 and 15; Table S4), and (2) under cool ( $< 29^\circ\text{C}$ ), moderate salinity ( $31.5\text{--}33.5$ ) conditions characteristic of the Mid-Atlantic Bight shelf waters (Figs. 3, 6A). High rates ( $> 30 \text{ nmol N L}^{-1} \text{ d}^{-1}$ ; Fig. 7A) were also observed in the cooler ( $27\text{--}29^\circ\text{C}$ ) surface waters of the Mid-Atlantic Bight slope (salinity  $< 34.5$ ; Fig. 7A). Rates were moderate in Gulf Stream surface waters ( $\sim 10 \text{ nmol N L}^{-1} \text{ d}^{-1}$ ) and decreased with depth until they were at the detection limit (Fig. 7A), typically just below the Chl *a* maximum (Fig. 4I–K). In mid-salinity mixing waters, rates were low to moderate ( $\sim 1\text{--}10 \text{ nmol N L}^{-1} \text{ d}^{-1}$ ). These findings are highly consistent with direct observations (described above).

Model 2, which included N + N and SRP as NFR predictors, generally agreed well with Model 1 (Fig. 7B). Both models were able to predict NFR to the order of magnitude at shelf stations within the NFR hotspot (Stas. 14 and 15; Fig. 4B), except at the thermocline at Sta. 14 where Model 1 underpredicted  $N_2$  fixation. However, the inclusion of nutrients in Model 2 generally resulted in lower predicted NFRs in mid-salinity mixing waters (Fig. 7B), though SRP was in slight excess of N + N here. Interestingly, Model 2 also predicted higher NFRs than Model 1 in Gulf Stream waters below the nutricline. These waters bore N + N : SRP ratios exceeding 20, suggesting that dissolved inorganic N was not limiting. This finding contrasts with the longstanding paradigm that  $N_2$  fixation occurs predominantly in waters where N + N is limiting, and suggests that this paradigm was not reflected in the larger training dataset (i.e., the measured NFRs).

We hypothesize that the inclusion of waters where UCYN-A was highly active in the training set skewed the relationship of between NFRs and nutrients. Unlike *Trichodesmium* spp. and other cyanobacterial diazotrophs, UCYN-A continues to fix  $N_2$  under both N-limited and N-replete conditions (Mills et al. 2020). Though ostensibly



**Fig 7.** Temperature-salinity diagrams. Colored dots depict (A) Model 1  $N_2$  fixation rate (NFR) results and (B) results of  $((\text{Model 2} - \text{Model 1})/\text{Model 1}) \times 100$ , i.e., the relative change associated with the inclusion of nutrients in the model. Dotted lines indicate isopycnals ( $\sigma_t$ ,  $\text{kg m}^{-3}$ ).

paradoxical given the high energetic cost of  $N_2$  vs.  $N + N$  assimilation, both UCYN-A1 and UCYN-A2 hosts appear unable to utilize ambient nitrate and, consequently, require their symbionts to fix  $N$  constitutively (Mills et al. 2020). Furthermore, recent work by Mills et al. (2020) suggests that UCYN-A NFRs can be enhanced by  $N + N$  additions, potentially due to secondary ecological effects (e.g., enhanced production of siderophores and thus iron availability by bacterioplankton). Future machine learning-based models of  $N_2$  fixation may be able to better tease apart the complex relationships among different diazotrophic clades and ambient  $N + N$  concentrations using larger training datasets, or by parameterizing the abundance of distinct groups.

#### Areal $N_2$ fixation rate estimation

Measured NFRs and those predicted from Model 1 (depth, seafloor depth, temperature, salinity, Chl  $a$  concentration) were trapezoidally depth-integrated from 0 to 100 m. At stations exhibiting more complex water column structure, the fine-scale resolution of the model often served to reduce or increase depth-integrated NFRs, depending on the locality of the measured depths (Table 2). Thin but significant NFR peaks were often predicted by Model 1 at the surface, thermocline, and at or above the Chl  $a$  maximum (Fig. 4). Areal rates calculated from direct measurements may thus over- or underestimate NFRs if sampling was biased toward or failed to resolve these peaks, assuming that these peaks represent real features. For example, model-based areal rates were more than double measurement-based estimates at Stas. 10 and 11 on the Mid-Atlantic Bight slope. Model 1 results (Table S3) suggest that there were NFR peaks ( $20\text{--}30 \text{ nmol N L}^{-1} \text{ d}^{-1}$ ) at the base of

the mixed layer that were not resolved by direct measurements. Conversely, depth-integrated measured rates at shelf break stations along the hotspot transects (12 and 17) exceeded those calculated from Model 1 output by about an order of magnitude. Though exacerbated by the tendency of our models to underestimate very high NFRs (Fig. 6), direct measurements at these stations were made at water column features where Model 1 predicted relatively thin NFR peaks. Consequently, the areal rates calculated from these direct measurements likely overestimate water column NFR.

Our work highlights the significant impact of sampling strategy on areal rate calculations when coverage of direct measurements is low. Discrepancies among studies in sampling strategy and depth coverage must be accounted for when comparing rates across regions and when using areal rates to build predictive or descriptive models.

#### Sea surface $N_2$ fixation

Surface NFRs were predicted from regional bathymetry and mean sea surface conditions (temperature, Chl  $a$  concentration) at the time of the study (Fig. 8). The model suggests that NFRs are undetectable to low ( $< 4 \text{ nmol N L}^{-1} \text{ d}^{-1}$ ) in the high Chl  $a$  waters that are immediately adjacent to the coastline (mean MODIS Chl  $a$  concentration =  $3.9 \mu\text{g L}^{-1}$  where seafloor depth  $< 15 \text{ m}$ ; Figs. 2B, 8, 9). These conditions were present at the coastal stations on the two northernmost transects (Stas. 5 and 6; seafloor depth =  $11 \text{ m}$ , Chl  $a$  concentration  $> 1 \mu\text{g L}^{-1}$ ) where measured NFRs were low or undetectable (Table S4).

Predicted NFRs peaked away from the coast, near the 20 m isobath where Chl  $a$  was moderate (mean MODIS Chl  $a$

**Table 2.** Predicted (Model 1) and measured N<sub>2</sub> fixation rates (NFRs) depth-integrated over the upper 100 m of the water column.\*

Sta.	Lat. (°N)	Long. (°W)	Mean temp. At 2 m (°C)	Mean salinity at 2 m	Seafloor depth (m)	Areal NFR calculated from predicted rates (μmol N m <sup>-2</sup> d <sup>-1</sup> )	Areal NFR calculated from measured rates (μmol N m <sup>-2</sup> d <sup>-1</sup> )
1	37.67	73.99	26.7	33.3	1261	292	101
3	37.66	74.78	26.9	31.5	42	169	ND
5	37.67	75.54	26.4	31.7	14	ND	5.64
6	36.93	75.88	26.3	31.3	14	10.4	11.9
8	36.94	74.70	26.9	32.1	88	143	74
10	36.93	73.81	28.9	35.6	2651	1030	375
11	36.35	73.75	29.4	35.8	2947	1130	465
12	36.34	74.21	28.4	34.7	2327	2950	25,500
13	36.33	74.97	26.9	31.3	41	891	5220
14	36.33	75.64	26.6	30.3	21	3890	9090
15	35.67	75.37	27.0	28.1	22	3730	4580
17	35.69	74.35	30.2	36.3	2312	4150	42,600
19	35.66	73.37	29.4	35.9	3682	3030	2820
20	35.04	73.92	29.6	36.1	3488	1510	1840
22	35.04	74.86	30.2	34.8	2715	2240	2200
24	35.05	75.68	28.5	34.9	30	493	521
25	33.33	78.58	28.7	35.9	21	372	483
27	33.34	77.53	29.4	30.2	34	389	387
28	33.34	77.03	30.4	28.4	217	913	1380
29	33.34	76.52	30.0	36.1	615	1370	1930
30	33.34	76.10	29.6	35.7	1650	1070	895
31	33.34	75.62	NaN	NaN	3298	1010	296
35	34.01	75.84	30.1	35.3	1106	2910	1460
36	34.01	76.24	29.7	34.7	256	1170	408
37	34.01	77.37	29.9	35.7	26	548	278
41	34.54	76.40	29.5	35.7	18	262	258

\*Or to the seafloor when seafloor depth was less than 100 m.

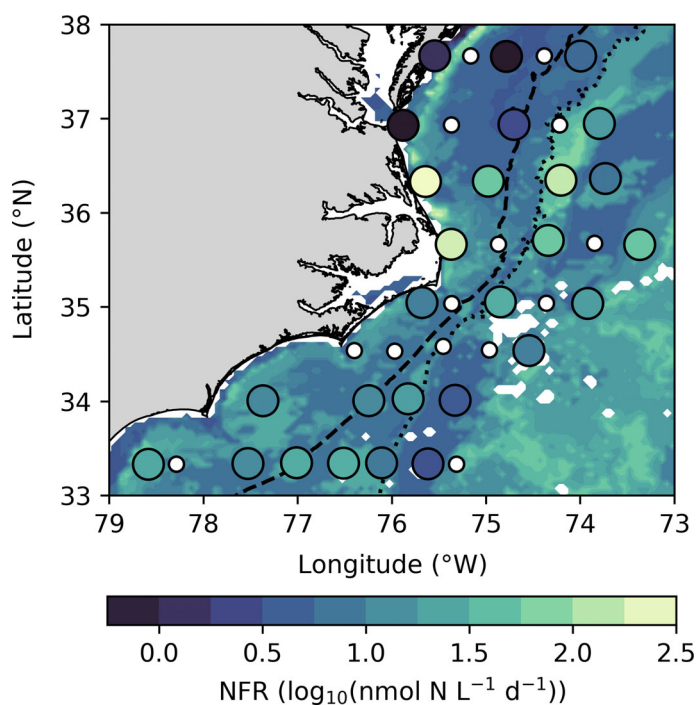
concentration = 0.87 μg L<sup>-1</sup> where seafloor depth was 15–25 m; Fig. 9A), then declined. The magnitude of this peak was greater in Mid-Atlantic Bight waters (> 35.5°N; maximum 5 m running mean: 44.6 nmol N L<sup>-1</sup> d<sup>-1</sup>), which included the hotspot, than in the South Atlantic Bight (< 35.5°N; maximum 5 m running mean: 24.6 nmol N L<sup>-1</sup> d<sup>-1</sup>). From the 40 m isobath to the shelf break (200 m), mean surface NFRs predicted by Model 3 were 5.3 ± 2.9 (*n* = 387) and 9.4 ± 5.8 (*n* = 455) nmol N L<sup>-1</sup> d<sup>-1</sup> in the Mid- and South Atlantic Bights, respectively. These middle to outer shelf waters were warmer in the South Atlantic Bight and bore lower Chl *a* concentrations than in the Mid-Atlantic Bight (Fig. 2). The abundance of *Trichodesmium* spp. *nifH* transcripts relative to genes was elevated in these South Atlantic Bight waters, as well (Fig. 5B).

Model 3 also predicted high NFRs along the offshore frontal zone (~ 29°C) between the Mid-Atlantic Bight shelf waters and the Gulf Stream, which occurs just beyond the 2000 m isobath (Figs. 8, 9B). This finding is consistent with NFR measurements in surface waters (Stas. 10, 12, 17; Fig. 8; Table S4) and results

from Models 1 and 2 (Table S3). Our sequencing (Fig. S2) and qPCR results (Figs. 5, S3, S4; Table S5) suggest that a myriad of diazotrophs may have been present and/or active here, including *Trichodesmium* spp., UCYN-A, *Cyanothece* sp., gammaproteobacteria, and Het-1. Het-1 sequences were recovered from surface waters via transcript sequencing at Sta. 10 (< 1% relative abundance), and they were observed at low abundance at Stas. 10 (120 ± 64 gene copies L<sup>-1</sup>) and 12 (147 ± 108 gene copies L<sup>-1</sup>), which were within the offshore band of high predicted NFRs.

High rates (~ 30 nmol N L<sup>-1</sup> d<sup>-1</sup>) were also predicted in oligotrophic open-ocean waters to the southeast. This finding fits within the existing conceptual model of the factors governing *Trichodesmium* activity (Carpenter and Capone 2008). However, these waters were poorly represented in the training dataset (i.e., among direct measurements) and should be interpreted with caution.

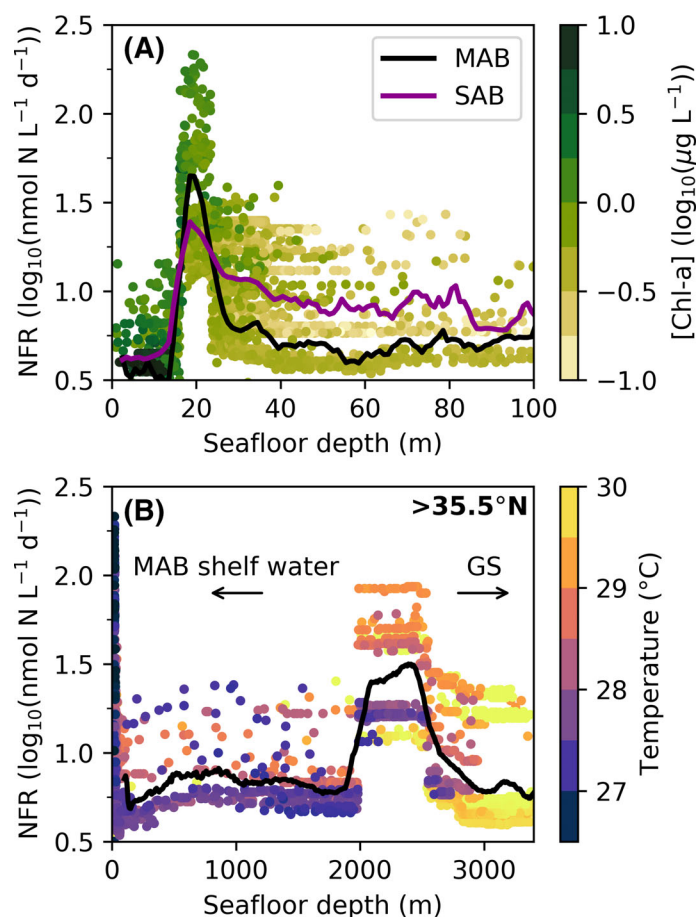
We note that measured rates as well as Models 1 and 2 suggest the presence of high NFRs deeper in the water column



**Fig 8.**  $N_2$  fixation rates (NFRs) predicted by Model 3 from mean MODIS sea surface temperature and chlorophyll *a* concentrations from August 12 to 19, 2016 (NASA Goddard Space Flight Center 2018 Reprocessing) and model-derived bathymetry (Amante and Eakins 2009) overlain by in situ NFR measured within the surface mixed layer. The dashed line indicates the 200 m isobath, i.e., the shelf break. The dotted line indicates the 2000 m isobath. White dots represent stations where surface NFRs were not measured directly.

along hotspot transects in the Mid-Atlantic Bight (36.3 and 35.6°N; e.g., Fig. 4F; Tables S3, S4). Indeed, there was no significant difference between  $N_2$  specific uptake rates at shelf and slope stations along these transects, indicating that diazotroph activity was elevated throughout these waters relative to others in the region. These high subsurface rates would not have been captured by Model 3 as the satellite-based measurements it used to predict NFRs lack the necessary depth-resolution. The extension of high NFRs across the shelf, following the net flow, suggest that inner shelf waters may have provided seed communities or resources for  $N_2$  fixation offshore.

The findings presented here suggest that the consumption of continental inputs by phytoplankton creates a coastal niche for diazotrophs. These likely include UCYN-A in the colder Mid-Atlantic Bight waters, *Trichodesmium* spp., particularly in the warmer South Atlantic Bight waters, and potentially some recently flushed freshwater cyanobacteria (see “Diazotroph diversity and abundance” section). Additionally, our results imply that conditions in the frontal regions between Mid-Atlantic Bight shelf waters and South Atlantic Bight shelf and Gulf Stream waters create conditions that are favorable for diazotrophs—predominantly UCYN-A2 based on our diazotroph community analysis (Figs. 5, S2).



**Fig 9.** Panel **A**:  $N_2$  fixation rates (NFRs) predicted by Model in surface shelf waters as a function of seafloor depth. Colors correspond to MODIS chlorophyll *a* (Chl *a*) concentrations. The lines represent the 5 m running mean for predicted NFRs in the Mid- (MAB, black) and South Atlantic Bights (SAB, purple), respectively. Panel **B**: NFRs predicted by Model 3 in surface waters across the MAB shelf and slope (>35.5°N). Colors correspond to MODIS temperature. The black line represents the 200 m predicted NFR running mean. “GS” indicates Gulf Stream waters.

## Conclusions and implications

Nitrogen fixation has long been hypothesized to occur in warm, oligotrophic surface waters and decline where fixed  $N$  is no longer limiting and phototrophic diazotrophs are out-competed by other phytoplankton (Carpenter and Capone 2008). Consequently, NFRs are generally expected to decline when Chl *a* concentrations are high. Research from the western North American continental shelf, including the findings presented here, paint a very different picture. Nitrogen fixation rates on the shelf (Mulholland et al. 2019; Tang et al. 2019; this study) frequently exceed those measured in the tropical North Atlantic basin (e.g., Martinez-Perez et al. 2016). Indeed, surface NFRs were positively correlated with Chl *a* concentrations in August 2015 and 2016 (Tang et al. 2019). There are two explanations for these observations: (1) diazotroph abundance and activity scale with biomass, or

(2) conditions in coastal waters favor N<sub>2</sub> fixation and diazotrophs are consequently more active, which should be reflected by higher rates of diazotroph-driven PN turnover (i.e., specific N<sub>2</sub> uptake rates). N<sub>2</sub> fixation rates may scale with biomass if a subset of the total community is constitutively fixing N<sub>2</sub>, as has been demonstrated for UCYN-A (Mills et al. 2020), or if lower per cell NFRs are compensated by increased diazotroph abundance resulting from heightened nutrient availability, as shown in *Trichodesmium* sp. and *Crocospaera* sp. cultures (Knapp et al. 2012).

We found that specific N<sub>2</sub> uptake rates were significantly greater in N-deplete offshore waters than on the continental shelf, except along transects aligned with and just north of Cape Hatteras (Table 1). N<sub>2</sub> fixation rates on the shelf that are high relative to measurements from tropical, oligotrophic waters (e.g., Martinez-Perez et al. 2016) thus likely arise because conditions at the continental margin are favorable for phytoplankton growth in general, rather than for diazotrophs specifically. This finding supports both the traditional paradigm that oligotrophic ocean waters offer a more significant niche for diazotrophs than coastal environments, and the idea that some (presumably symbiotic) diazotrophs may be constitutively fixing N<sub>2</sub>. However, empirical model results hint at greater nuance. Model 3 predicts enhanced NFRs along coastline, just seaward of the coastal Chl *a* maximum (Figs. 8, 9A). It is unlikely that this peak was driven by high PN concentrations because PN tends to scale with Chl *a* (Figs. 2, S1). Rather, it indicates that the conditions present here favor diazotrophic metabolism.

Both direct measurements and model results indicate that the frontal mixing zone between cool Mid-Atlantic Bight shelf waters and warm South Atlantic Bight shelf and Gulf Stream waters created conditions conducive to exceptionally high levels of diazotrophy (> 100 nmol N L<sup>-1</sup> d<sup>-1</sup>) both on and off the shelf during the study period (e.g., Figs. 4, 8). Based on the distribution of high NFRs, we hypothesize that high NFRs offshore were supported by shelf water properties (e.g., communities, nutrients) advected there. This N<sub>2</sub> fixation hotspot likely resulted from the proliferation and activity of the haptophyte-symbiont UCYN-A (Figs. 5, S2), though other diazotrophs were present and active as well. Mills et al. (2020) have hypothesized that N<sub>2</sub> fixation by UCYN-A may be linked to the activity of their larger community. We proffer that either altered ecological dynamics or nutrient availability may have played a role in the formation of the N<sub>2</sub> fixation hotspot documented here. As the Gulf Stream transports a significant fraction of shelf matter offshore (Churchill and Berger 1998), organic carbon produced as a result of these high NFRs may be advected into the basin. Thus, if the documented N<sub>2</sub> fixation hotspot is a regular feature near Cape Hatteras, it may affect North Atlantic carbon cycling.

By applying a supervised machine learning approach, this study offers a nuanced view of N<sub>2</sub> fixation and hints at unresolved complexities in its spatial distribution. Our work

highlights the possibly confounding role of sampling strategy on calculating areal rates, which are frequently used to resolve global patterns and build models (e.g., Tang et al. 2019), and the potential for dynamic frontal systems to augment fixed N inputs. Understanding the spatial heterogeneity of key biogeochemical processes in the ocean will facilitate the ongoing efforts to constrain N budgets on regional to global scales.

## References

- Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**: 403–410.
- Amante, C., and B. W. Eakins. 2009. ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis. NOAA Technical Memorandum NESDIS NGDC-24.
- Atkinson, L. P. 1977. Modes of Gulf Stream intrusion into the South Atlantic Bight shelf waters. *Geophys. Res. Lett.* **4**: 583–586.
- Benavides, M., P. Moisaner, H. Berthelot, T. Dittmar, O. Grosso, and S. Bonnet. 2015. Mesopelagic N<sub>2</sub> fixation rates related to organic matter composition in the Solomon and Bismarck Seas (Southwest Pacific). *PLoS One* **10**: e0143775. doi:10.1371/journal.pone.0143775
- Blais, M., et al. 2012. Nitrogen fixation and identification of potential diazotrophs in the Canadian Arctic. *Global Biogeochem. Cycles* **26**: GB3022. doi:10.1029/2011gb004096
- Breiman, L. 2001. Random forests. *Machine Learning* **45**: 5–32.
- Bustin, S. A., H. Dhillon, S. Kirvell, C. Greenwood, M. Parker, G. Shipley, and T. Nolan. 2014. Variability of the reverse transcription step: Practical implications. *Clin. Chem.* **61**: 1–212. doi:10.1373/clinchem.2014.230615
- Bustin, S. A., R. Mueller, and T. Nolan. 2020. Parameters for successful PCR primer design, p. 5–22. *In* R. Biassoni and A. Raso [eds.], *Quantitative real-time PCR: Methods and protocols*. Methods in molecular biology. Springer Science and Business Media, LLC, part of Springer Nature.
- Capone, D. G., and others. 2005. Nitrogen fixation by *Trichodesmium* spp.: An important source of new nitrogen to the tropical and subtropical North Atlantic Ocean. *Global Biogeochem. Cycles* **19**: GB2024. doi:10.1029/2004GB002331
- Carpenter, E. J., and D. G. Capone. 2008. Nitrogen fixation in the marine environment, p. 141–198. *In* D. G. Capone, D. A. Bronk, M. R. Mulholland, and E. J. Carpenter [eds.], *Nitrogen in the marine environment*. Academic Press.
- Church, M. J., B. D. Jenkins, D. M. Karl, and J. P. Zehr. 2005a. Vertical distributions of nitrogen-fixing phylotypes at Stn ALOHA in the oligotrophic North Pacific Ocean. *Aquat. Microb. Ecol.* **38**: 3–14. doi:10.3354/ame038003
- Church, M. J., C. M. Short, B. D. Jenkins, D. M. Karl, and J. P. Zehr. 2005b. Temporal patterns of nitrogenase gene (*nifH*) expression in the oligotrophic North Pacific Ocean. *Appl.*



- Environ. Microbiol. **71**: 5362–5370. doi:[10.1128/aem.71.9.5362-5370.2005](https://doi.org/10.1128/aem.71.9.5362-5370.2005)
- Churchill, J. H., and T. J. Berger. 1998. Transport of middle Atlantic Bight shelf water to the Gulf Stream near Cape Hatteras. *J. Geophys. Res.: Oceans* **103**: 30605–30621.
- Colón-López, M. S., D. M. Sherman, and L. A. Sherman. 1997. Transcriptional and translational regulation of nitrogenase in light-dark-and continuous-light-grown cultures of the unicellular cyanobacterium *Cyanothece* sp. strain ATCC 51142. *J. Bacteriol.* **179**: 4319–4327.
- Csanady, G., and P. Hamilton. 1988. Circulation of slopewater. *Cont. Shelf Res.* **8**: 565–624.
- Dabundo, R., and others. 2014. The contamination of commercial  $^{15}N_2$  gas stocks with  $^{15}N$ -labeled nitrate and ammonium and consequences for nitrogen fixation measurements. *PLoS One* **9**: e110335. doi:[10.1371/journal.pone.0110335](https://doi.org/10.1371/journal.pone.0110335)
- Eren, A. M., H. G. Morrison, P. J. Lescault, J. Reveillaud, J. H. Vineis, and M. L. Sogin. 2015. Minimum entropy decomposition: Unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J.* **9**: 968–979. doi:[10.1038/ismej.2014.195](https://doi.org/10.1038/ismej.2014.195)
- Farnelid, H., and others. 2019. Diverse diazotrophs are present on sinking particles in the North Pacific Subtropical Gyre. *ISME J.* **13**: 170–182. doi:[10.1038/s41396-018-0259-x](https://doi.org/10.1038/s41396-018-0259-x)
- Fonseca-Batista, D., and others. 2019. Evidence of high  $N_2$  fixation rates in the temperate Northeast Atlantic. *Biogeosciences* **16**: 999–1017. doi:[10.3929/ethz-b-000332762](https://doi.org/10.3929/ethz-b-000332762)
- Glibert, P., and D. Capone. 1993. Mineralization and assimilation in aquatic, sediment, and wetland systems, p. 243–272. *In* R. Knowles and T. H. Blackburn [eds.], *Nitrogen isotope techniques*. San Diego, CA: Academic Press.
- Grosse, J., D. Bombar, H. N. Doan, L. N. Nguyen, and M. Voss. 2010. The Mekong River plume fuels nitrogen fixation and determines phytoplankton species distribution in the South China Sea during low and high discharge season. *Limnol. Oceanogr.* **55**: 1668–1680. doi:[10.4319/lo.2010.55.4.1668](https://doi.org/10.4319/lo.2010.55.4.1668)
- Hunt, B. P., S. Bonnet, H. Berthelot, B. J. Conroy, R. A. Foster, and M. Pagano. 2016. Contribution and pathways of diazotroph-derived nitrogen to zooplankton during the VAHINE mesocosm experiment in the oligotrophic New Caledonia lagoon. *Biogeosciences* **13**: 3131–3145. doi:[10.5194/bg-13-3131-2016](https://doi.org/10.5194/bg-13-3131-2016)
- Klawoon, I., G. Lavik, P. Böning, H. K. Marchant, J. Dekaezemaeker, W. Mohr, and H. Ploug. 2015. Simple approach for the preparation of  $^{15-15}N_2$  enriched water for nitrogen fixation assessments: Evaluation, application and recommendations. *Front. Microbiol.* **6**: 769 doi: [10.3389/fmicb.2015.00769](https://doi.org/10.3389/fmicb.2015.00769).
- Knapp, A. N. 2012. The sensitivity of marine  $N_2$  fixation to dissolved inorganic nitrogen. *Front. Microbiol.* **3**: 1–14. doi: [10.3389/fmicb.2012.00374](https://doi.org/10.3389/fmicb.2012.00374)
- Knapp, A. N., J. Dekaezemaeker, S. Bonnet, J. A. Sohm, and D. Capone. 2012. Sensitivity of *Trichodesmium erythraeum* and *Crocospaera watsonii* abundance and  $N_2$  fixation rates to varying  $NO_3^-$  and  $PO_4^{3-}$  concentrations in batch cultures. *Aquat. Microb. Ecol.* **66**: 223–236. doi:[10.3354/ame01577](https://doi.org/10.3354/ame01577)
- Laglera, L. M., and C. M. G. van den Berg. 2009. Evidence for geochemical control of iron by humic substances in seawater. *Limnol. Oceanogr.* **54**: 610–619.
- Laughlin, R. J., and R. J. Stevens. 2003. Changes in composition of nitrogen-15-labeled gases during storage in septum-capped vials. *Soil Sci. Soc. Am. J.* **67**: 540–543. doi:[10.2136/sssaj2003.5400](https://doi.org/10.2136/sssaj2003.5400)
- Manly, B. F. 2007. Randomization, p. 1–28. *In* B. P. Carlin, C. Chatfield, M. Tanner and Z. Jim [eds.], *Randomization, bootstrap and Monte Carlo methods in biology*. 3, **70**: Texts in Statistical Science, 3rd ed, V. 70. Chapman & Hall/CRC, Taylor & Francis Group.
- Martinez-Perez, C., and others. 2016. The small unicellular diazotrophic symbiont, UCYN-A, is a key player in the marine nitrogen cycle. *Nat. Microbiol.* **1**: 16163. doi:[10.1038/nmicrobiol.2016.163](https://doi.org/10.1038/nmicrobiol.2016.163)
- Meissner, T., F. J. Wentz, A. Manaster, and R. Lindsley. 2019. Remote Sensing Systems SMAP Ocean Surface Salinities [Level 2C, Level 3 Running 8-day, Level 3 Monthly], Version 4.0 validated release. doi:[10.5067/SMP40-3SPCS](https://doi.org/10.5067/SMP40-3SPCS)
- Mills, M. M., and others. 2020. Unusual marine cyanobacteria/haptophyte symbiosis relies on  $N_2$  fixation even in N-rich environments. *ISME J.* **14**: 2395–2406. doi:[10.1038/s41396-020-0691-6](https://doi.org/10.1038/s41396-020-0691-6)
- Moisander, P., E. McClinton, and H. Paerl. 2002. Salinity effects on growth, photosynthetic parameters, and nitrogenase activity in estuarine planktonic cyanobacteria. *Microb. Ecol.* **43**: 432–442. doi:[10.1007/s00248-001-1044-2](https://doi.org/10.1007/s00248-001-1044-2)
- Moisander, P. H., and others. 2010. Unicellular cyanobacterial distributions broaden the oceanic  $N_2$  fixation domain. *Science* **327**: 1512–1514. doi:[10.1126/science.1185468](https://doi.org/10.1126/science.1185468)
- Moisander, P. H., M. Benavides, S. Bonnet, I. Berman-Frank, A. E. White, and L. Riemann. 2017. Chasing after non-cyanobacterial nitrogen fixation in marine pelagic environments. *Front. Microbiol.* **8**: 1736 doi:[10.3389/fmicb.2017.01736](https://doi.org/10.3389/fmicb.2017.01736).
- Montoya, J. P., M. Voss, P. Kahler, and D. G. Capone. 1996. A simple, high-precision, high-sensitivity tracer assay for  $N_2$  fixation. *Appl. Environ. Microbiol.* **62**: 986–993.
- Moore, C., and others. 2013. Processes and patterns of oceanic nutrient limitation. *Nat. Geosci.* **6**: 701–710. doi:[10.1038/ngeo1765](https://doi.org/10.1038/ngeo1765)
- Mulholland, M., and others. 2012. Rates of dinitrogen fixation and the abundance of diazotrophs in North American coastal waters between Cape Hatteras and Georges Bank. *Limnol. Oceanogr.* **57**: 1067–1083. doi:[10.4319/lo.2012.57.4.1067](https://doi.org/10.4319/lo.2012.57.4.1067)
- Mulholland, M., and others. 2019. High rates of  $N_2$  fixation in temperate, western North Atlantic coastal waters expands the realm of marine diazotrophy. *Global Biogeochem. Cycles* **33**: 826–840. doi:[10.1029/2018GB006130](https://doi.org/10.1029/2018GB006130)

- Muñoz-Marín, M. C., I. N. Shilova, T. Shi, H. Farnelid, A. M. Cabello, and J. P. Zehr. 2019. The transcriptional cycle is suited to daytime N<sub>2</sub> fixation in the unicellular cyanobacterium “*Candidatus* Atelocyanobacterium thalassa” (UCYN-A). *mBio* **10**: e02495–e02418. doi:10.1128/mBio.02495-18
- NASA Goddard Space Flight Center, O. B. P. G. 2018. Reprocessing. Moderate-resolution imaging spectroradiometer (MODIS) aqua chlorophyll data. NASA Ocean Biology Distributed Active Archive Center (OB.DAAC). <https://oceancolor.gsfc.nasa.gov/>
- Nixon, S., and others. 1996. The fate of nitrogen and phosphorus at the land-sea margin of the North Atlantic Ocean. *Biogeochemistry* **35**: 141–180.
- Palter, J. B., and others. 2020. High N<sub>2</sub> fixation rates near the Gulf Stream consistent with a circulation control on diazotrophy. *Geophys. Res. Lett.* **47**: e2020GL89103. doi:10.1029/2020GL89103
- Parsons, T. R., Y. Maita, and C. M. Lalli. 1984. A manual of biological and chemical methods for seawater analysis. Publ. Pergamon Press.
- Pedersen, J. N., D. Bombar, R. W. Paerl, and L. Riemann. 2018. Diazotrophs and N<sub>2</sub> fixation associated with particles in coastal estuarine waters. *Front. Microbiol.* **9**: 2759. doi:10.3389/fmicb.2018.02759
- Pedregosa, F., and others. 2011. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**: 2825–2830.
- Prufert-Bebout, L., H. W. Paerl, and C. Lassen. 1993. Growth, nitrogen fixation, and spectral attenuation in cultivated *Trichodesmium* species. *Appl. Environ. Microbiol.* **59**: 1367–1375.
- Rahav, E., and others. 2013. Dinitrogen fixation in aphotic oxygenated marine environments. *Front. Microbiol.* **4**: 227. doi:10.3389/fmicb.2013.00227
- Rahav, E., M. Giannetto, and E. Bar-Zeev. 2016. Contribution of mono and polysaccharides to heterotrophic N<sub>2</sub> fixation at the eastern Mediterranean coastline. *Sci. Rep.* **6**: 27858. doi:10.1038/srep27858
- Sammartino, M., S. Marullo, R. Santoleri, and M. Scardi. 2018. Modelling the vertical distribution of phytoplankton biomass in the Mediterranean Sea from satellite data: A neural network approach. *Remote Sens. (Basel)* **10**: 1666. doi:10.3390/rs10101666
- Selden, C., and others. 2019. Dinitrogen fixation across physico-chemical gradients of the Eastern Tropical North Pacific oxygen deficient zone. *Global Biogeochem. Cycles* **33**: 1187–1202. doi:10.1029/2019GB006242
- Tang, W., and others. 2019. Revisiting the distribution of oceanic N<sub>2</sub> fixation and estimating diazotrophic contribution to marine production. *Nat. Commun.* **10**: 831. doi:10.1038/s41467-019-08640-0
- Tang, W., Z. Li, and N. Cassar. 2019. Machine learning estimates of global marine nitrogen fixation. *J. Geophys. Res.: Biogeosci.* **124**: 717–730. doi:10.1029/2018JG004828
- Thompson, A., B. J. Carter, K. Turk-Kubo, F. Malfatti, F. Azam, and J. P. Zehr. 2014. Genetic diversity of the unicellular nitrogen-fixing cyanobacteria UCYN-A and its prymnesiophyte host. *Environ. Microbiol.* **16**: 3238–3249. doi:10.1111/1462-2920.12490
- Villareal, T. A. 1992. Marine nitrogen-fixing diatom-cyanobacteria symbioses, p. 163–175. *In* E. J. Carpenter, D. G. Capone, and J. G. Rueter [eds.], *Marine pelagic cyanobacteria: Trichodesmium and other diazotrophs*. Springer.
- White, A. E., and others. 2020. A critical review of the <sup>15</sup>N<sub>2</sub> tracer method to measure diazotrophic production in pelagic ecosystems. *Limnol Oceanogr. Methods* **18**: 129–147. doi:10.1002/lom3.10353
- Zehr, J., and P. Turner. 2001. Nitrogen fixation: Nitrogenase genes and gene expression, p. 271–286. *In* J. H. Paul [ed.], *Methods in microbiology*. Academic Press Ltd.
- Zehr, J. P., and D. G. Capone. 2020. Changing perspectives in marine nitrogen fixation. *Science* **368**: eaay9514. doi:10.1126/science.aay9514

### Acknowledgments

As is often the case in oceanography, this work was a truly collaborative effort. We would like to thank the captain and crew of the *R/V Hugh R. Sharp* for their dedication and support, as well as the science crew who assisted in data collection. In particular, we thank Brittany Winder for her role in cruise planning and preparation, and her stellar managerial efforts during field operations. We thank Bess Ward for the use of her facilities to analyze <sup>15</sup>N<sub>2</sub> samples. We thank David Shelley for assisting with sample collection and processing, Zuzanna Abdala for collecting samples for molecular analysis, Kim Powell for offering her expertise and aid in sequencing, and Katherine Crider for her assistance in qPCR analysis. Finally, we thank Denni Einarsson, Michael Echevarria, and Austin Abbott for their thoughts on the machine learning aspects of this study. This work was supported by awards from the NSF Chemical Oceanography (OCE-1459698) to MRM, the Jeffress Trust Awards Program in Interdisciplinary Research to PDC, and the Jacques S. Zaneveld and Neil and Susan Kelley Endowed Scholarships to CRS.

### Conflict of Interest

None declared.

Submitted 04 August 2020

Revised 08 December 2020

Accepted 25 January 2021

Associate editor: Ilana Berman-Frank