Title

Riley J. Hale1\*, Jesse M. Wilson1, Elizabeth J. Connors1, Ewa Merz1, Samantha M. Clements1,2, Melissa L. Carter1, Andrew D. Barton1,2, Jeff S. Bowman1,3,4

1 Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA 92092, United States

2 Center for Marine Biodiversity and Conservation, University of California San Diego 92092, La Jolla, CA, United States

3 Department of Ecology, Behavior and Evolution, University of California San Diego, La Jolla, CA 92092, United States

4 Center for Microbiome Innovation, University of California San Diego, La Jolla, CA 92092, United States

\*Correspondence: [rhale@ucsd.edu](mailto:rhale@ucsd.edu)

# ABSTRACT

# INTRODUCTION

Oceanic fluxes of carbon dioxide are a major component of the global carbon cycling, and marine microbiota are major modulators of these fluxes (e.g. Azam et al., 1983; Ducklow, 1994; Pomeroy et al., 2007). The basis of the microbial loop is the remineralization of organic matter by heterotrophic microbes, primarily bacteria. This, combined with primary production by phytoplankton, respiration by other organisms, vertical pumps, and the viral shunt, describes the cycling of carbon and other key nutrients in marine ecosystems.

Net trophic status can be described as the difference between primary production and community respiration of organic carbon in the ocean. The system is considered net autotrophic when primary production is larger than respiration, and is considered net heterotrophic when respiration is larger than production (Smith & Mackenzie, 1987). Extreme autotrophic and heterotrophic events have been observed in both offshore areas throughout the California Current region (Bograd et al., 2008; Siedlecki et al., 2015) and in nearshore, coastal Southern California waters as recently as the spring of 2020 following a harmful algal bloom (Kahru et al., 2021; Wilson et al., 2022). These events have acute effects throughout trophic webs, including consequences for coastal fisheries in eastern boundary upwelling zones (Chan et al., 2008; Trainer et al., 2010). Furthermore, nearshore ecosystems are an increasingly important portion of marine primary production (Kahru et al., 2015, Zhao et al., 2020) and are predicted to become increasingly important in the ocean’s response to global climate change (Bernhardt & Leslie, 2013; Spalding et al., 2014). However, there is a current lack of consensus about whether near-shore zones are net autotrophic or net heterotrophic (Bauer et al., 2013; Regnier et al., 2013), given the high levels of both primary production and respiration. Thus, accurate methods that can quantify the biological and physical factors that drive changes in net trophic status are vital for understanding the function of coastal ecosystems and predicting future change.

While net trophic status is defined in terms of carbon, it is often estimated using dissolved oxygen concentrations due to the converse cycling of oxygen in photosynthesis and respiration (Dodds & Cole, 2007). The term “apparent oxygen utilization” (AOU) is often used to describe oxygen cycling in the ocean, defined as the difference between dissolved oxygen concentrations at saturation and measured dissolved oxygen. An oxygen anomaly with negative values refers to oversaturation and positive values refers to undersaturation (Najjar & Keeling, 1997). However, AOU does not differentiate between oxygen cycled biologically through photosynthesis and respiration and oxygen cycled physically through ocean-atmosphere interactions (e.g., via bubble injection). Previous studies have observed and predicted shifts in dissolved oxygen concentrations across various marine ecosystems, but these studies do not assess net trophic status, as dissolved oxygen alone does not indicate net trophic status (Chen et al., 2024).

What is BOU, why is it important

previous methods for estimation (our paper, general productivity estimates)

Why is the biological aspect/microbial community important

# METHODS

Seawater samples were collected twice-weekly between 4 January 2018 and 30 October 2024 from the Ellen Browning Scripps Memorial Pier at the Scripps Institution of Oceanography (La Jolla, CA, USA) as a part of the Scripps Ecological Observatory (<https://ecoobs.ucsd.edu/>) in tandem with the Southern California Coastal Ocean Observing System (SCCOOS) Scripps Pier Shore Station sampling efforts (<https://sccoos.org/>). Samples were obtained by lowering a bucket using a winch system at approximately 11:00 AM and collecting surface seawater. Samples were brought back to the lab and filtered through a sterile 47 mm 0.2 μM Supor membrane disc filter (Pall Corporation, Port Washington, New York, USA). Filters were stored at –80 °C until DNA extraction. From 4 January 2018 to 8 January 2019, DNA was extracted using MoBio DNEASY PowerWater Kit (Qiagen, Venlo, Netherlands); after this point, DNA was extracted using a KingFisherTM Flex automated extraction instrument and the MagMax Microbiome Ultra Nucleic Acid Extraction Kit (ThermoFisher Scientific, Waltham, Massachusetts, USA). DNA concentrations were then quantified using the Qubit HS DNA quantification kit (Invitrogen). For library preparation, extracted DNA samples were sent to Argonne National Laboratory Environmental Sample Processing Center for the amplification using 515F & 806R primers (Walters et al. 2016), followed by 2 x 151 paired end sequenced on the Illumina Miseq platform.

Reads were first filtered, denoised, and merged with dada2 (Callahan et al. 2016). We then determined microbial community structure and performed metabolic pathway inference using the paprica pipeline v0.70 (Erazo, Dutta, and Bowman 2021), a phylogenetic placement approach for microbial community structure and metabolic inference (<https://github.com/bowmanjeffs/paprica>). The paprica pipeline uses rRNA gene amplicon sequencing data and provides taxonomic resolution at the strain and species level for select regions on the prokaryotic phylogenetic tree and provides an estimate of gene and metabolic pathway abundance.

Relative abundance data were transformed with a Hellinger transformation using the ‘decostand’ function in the ‘vegan’ R package (Oksanen et al. 2001). Hellinger transformations calculates the square root of a taxon’s relative abundance in a given sample and allows for a reduction in the effects of proportionally large relative abundance values. This makes dynamics of lower-abundance taxa considerable and reduces the overpowering effects of highly abundant taxa such those that form large blooms.

We included all 16S and 18S ASVs (amplicon sequence variants), including bacteria, archaea, and eukaryotes, as potential predictive features in our model. We then used ‘Boruta’ function in the R package ‘Boruta’, a feature selection algorithm, to select which ASVs to provide our model as predictors of BOU (Kursa and Rudnicki 2010). The Boruta algorithm provides a method for reducing predictors of little or no importance for BOU prediction. This function iteratively replaces each prediction feature, here ASV Hellinger transformed relative abundance data, with randomized versions of itself and assesses impacts on response quality, here BOU prediction. Then, we remove all predictors that get rejected as important for model prediction, reducing the total number of ASV predictors from over 10000 to less than 100 while maintaining nearly identical predictive power.

Here we propose a random forest model to predict the estimated BOU time series from microbial community structure. Random forest models create many decision trees (a “forest”) and aggregate them into a single averaged decision tree model. Here, we use the ‘ranger’ function in the ‘ranger’ R package to create the random forest model (Wright, Wager, and Probst 2015).

Following our previous methods, we set aside a three month subset of the time series for model validation (Hale et al., 2025?). This time period spanned from May 1, 2021 to August 31, 2021. This method provides a means to validate the model while taking into account the lack of independence of time series data. We also implement a similar temporal subset cross-validation method, following Hale et al., 2025. After creating a final model, we iteratively set each continuous 3-month period within the full time series and retrain the model with the same hyperparameters. Then we compare model performance on each iterative validation dataset to assess model performance given a variety of different training data across the full time series. This allows for confirmation of model performance under different conditions than it was originally trained on.

Once the trained model was cross-validated, we created our final model using all available microbial community time series data and final tuned hyperparameters. This resulted in a prediction of BOU across the full microbial community time series. We then calculated the relative influence of each ASV as predictor variables in our model using XXXXX function in the XXXXX package.

# RESULTS

# DISCUSSION

# CONCLUSIONS

# REFERENCES

Callahan, Benjamin J., Paul J. McMurdie, Michael J. Rosen, Andrew W. Han, Amy Jo A. Johnson, and Susan P. Holmes. 2016. “DADA2: High-Resolution Sample Inference from Illumina Amplicon Data.” *Nature Methods* 13 (7): 581–83. https://doi.org/10.1038/nmeth.3869.

Erazo, Natalia G., Avishek Dutta, and Jeff S. Bowman. 2021. “From Microbial Community Structure to Metabolic Inference Using Paprica.” *STAR Protocols* 2 (4): 101005. https://doi.org/10.1016/j.xpro.2021.101005.

Kursa, Miron B., and Witold R. Rudnicki. 2010. “Feature Selection with the **Boruta** Package.” *Journal of Statistical Software* 36 (11). https://doi.org/10.18637/jss.v036.i11.

Oksanen, Jari, Gavin L. Simpson, F. Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R. Minchin, R.B. O’Hara, et al. 2001. “Vegan: Community Ecology Package.” https://doi.org/10.32614/CRAN.package.vegan.

Walters, William, Embriette R. Hyde, Donna Berg-Lyons, Gail Ackermann, Greg Humphrey, Alma Parada, Jack A. Gilbert, et al. 2016. “Improved Bacterial 16S rRNA Gene (V4 and V4-5) and Fungal Internal Transcribed Spacer Marker Gene Primers for Microbial Community Surveys.” *mSystems* 1 (1): e00009-15. https://doi.org/10.1128/mSystems.00009-15.

Wright, Marvin N., Stefan Wager, and Philipp Probst. 2015. “Ranger: A Fast Implementation of Random Forests.” https://doi.org/10.32614/CRAN.package.ranger.