Science of the Total Environment 677 (2019) 205–214

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Modelling the influence of environmental parameters over marine planktonic microbial communities using artificial neural networks

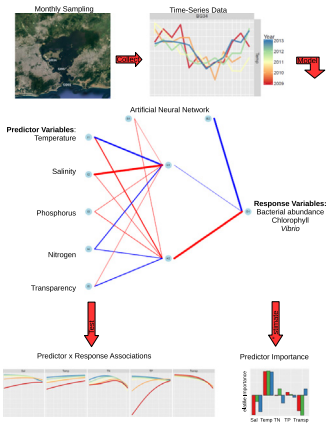
F.H. Coutinho [a,b,c,*], C.C. Thompson [a], A.S. Cabral [a], R. Paranhos [a], B.E. Dutilh [a,b,c], F.L. Thompson [a,d,**]

[a] Universidade Federal do Rio de Janeiro (UFRJ), Instituto de Biologia (IB), Rio de Janeiro, Brazil
[b] Radboud University Medical Centre, Radboud Institute for Molecular Life Sciences, Centre for Molecular and Biomolecular Informatics (CMBI), Nijmegen, the Netherlands
[c] Utrecht University, Theorethical Biology and Bioinformatics, Utrecht, the Netherlands
[d] Universidade Federal do Rio de Janeiro (UFRJ), COPPE, SAGE, Rio de Janeiro, Brazil

HIGHLIGHTS

- Artificial Neural Networks accurately modelled the microbiome from Guanabara Bay.
- Temperature and Salinity are main factors controlling the Microbiome.
- Phosphorus, Nitrogen and Transparency are of lesser importance.
- Small increases in temperature are predicted to promote the growth of pathogens.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 18 January 2019
Received in revised form 1 April 2019
Accepted 1 April 2019
Available online 26 April 2019

Editor: Frederic Coulon

Keywords:
Tropical
Estuary
Eutrophication
Pollution
Machine learning

ABSTRACT

Guanabara Bay is a tropical estuarine ecosystem that receives massive anthropogenic impacts from the metropolitan region of Rio de Janeiro. This ecosystem suffers from an ongoing eutrophication process that has been shown to promote the emergence of potentially pathogenic bacteria, giving rise to public health concerns. Although previous studies have investigated how environmental parameters influence the microbial community of Guanabara Bay, they often have been limited to small spatial and temporal gradients and have not been integrated into predictive mathematical models. Our objective was to fill this knowledge gap by building models that could predict how temperature, salinity, phosphorus, nitrogen and transparency work together to regulate the abundance of bacteria, chlorophyll and *Vibrio* (a potential human pathogen) in Guanabara Bay. To that end, we built artificial neural networks to model the associations between these variables. These networks were carefully validated to ensure that they could provide accurate predictions without biases or overfitting. The estimated models displayed high predictive capacity (Pearson correlation coefficients ≥0.67 and root mean square error ≤ 0.55). Our findings showed that temperature and salinity were often the most important factors regulating the

* Correspondence to: F.H. Coutinho, Radboud University Medical Centre, Radboud Institute for Molecular Life Sciences, Centre for Molecular and Biomolecular Informatics (CMBI), Nijmegen, the Netherlands.
** Correspondence to: F.L. Thompson, Av. Carlos Chagas Filho. S/N - CCS - IB - BIOMAR - Laboratório de Microbiologia - BLOCO A (Anexo) A3 - sl 102, Cidade Universitária, Rio de Janeiro, RJ 21941-599, Brazil.
E-mail addresses: felipehcoutinho@gmail.com (F.H. Coutinho), fabianothompson1@gmail.com (F.L. Thompson).

https://doi.org/10.1016/j.scitotenv.2019.04.009

Artificial neural networks
Time-series

abundance of bacteria, chlorophyll and *Vibrio* (absolute importance ≥5) and that each of these has a unique level of dependence on nitrogen and phosphorus for their growth. These models allowed us to estimate the Guanabara Bay microbiome's response to changes in environmental conditions, which allowed us to propose strategies for the management and remediation of Guanabara Bay.

1. Introduction

Guanabara Bay is a tropical estuarine bay surrounded by the metropolitan area of Rio de Janeiro. This is the second largest urban area in Brazil, inhabited by approximately twelve million people in 2016. As one of the first settlements in the country, Guanabara Bay (GB) has a long history of anthropogenic pollution delivered directly into the bay or indirectly into the water basin of the surrounding rainforest biome. GB has an area of approximately 380 km^2 and average depth of 7 m. The bay has a mixed semidiurnal tide regime with a range of 0.7 m, and renewal of 50% of the water volume occurs in an estimated 11.4 days (Kjerfve et al., 1997) (Fig. 1). The innermost sites of the north-western section receive freshwater input from the surrounding basin as well as sewage from domestic, agricultural and industrial sources. This ecosystem receives daily an estimated 1.5×10^6 m^3 of untreated domestic sewage and 150 tons of industrial waste water (Fistarol et al., 2015). The inner sections of GB receive a small input of oceanic water but massive inputs of pollution. Therefore, in these regions, salinity levels are low and nutrient (i.e., phosphorus and nitrogen) concentrations are high. Thus, heavily polluted sites within the innermost region have undergone extreme eutrophication (Kjerfve et al., 1997; Mayr et al., 1989; Paranhos et al., 1998), leading to algal blooms and fish mortality events (Santos et al., 2007; Villac and Tenenbaum, 2010). Towards the outermost regions of the bay, the greater input of oceanic waters leads to an increase in salinity and transparency accompanied by lower nutrient concentrations. Mixing of waters brought by tides and currents creates a gradient of water quality conditions between these two extremes. (Mayr et al., 1989; Paranhos et al., 1998, 2001; Vieira et al., 2007, 2008).

Previous investigations have sought to characterize how the taxonomic composition of microbial communities is affected by water quality conditions at GB using metagenomics (Gregoracci et al., 2012; Thompson et al., 2011; Vieira et al., 2008). These studies have reported that *Pelagibacter*, *Prochlorococcus*, *Synechococcus* and other marine oligotrophs dominate the communities in less polluted sites. Meanwhile, microbial communities in sites with higher degrees of pollution

Fig. 1. Map of Guanabara Bay showcasing the location of sampling sites.

are dominated by copiotrophic bacteria, such as *Gammaproteobacteria*, *Bacteroidetes* and *Firmicutes*, which are taxa that include many species of opportunistic pathogens (Gregoracci et al., 2012; Vieira et al., 2008).

Many inhabitants of Rio de Janeiro have no access to treated water or sewage collection systems, which puts them at increased risk of contact with opportunistic aquatic pathogens, the causative agents of waterborne diseases. Across Guanabara Bay, the abundance of pathogenic bacteria is positively correlated with the degree of pollution (Gregoracci et al., 2012; Vieira et al., 2008), which has raised public health concerns associated with the local population interacting with this ecosystem. These concerns have further intensified following the discovery of antibiotic-resistant bacteria in this habitat (Coutinho et al., 2014). Among the pathogens that dwell in the bay, the genus *Vibrio* is of special relevance. Members of this genus are abundant across GB (Gregoracci et al., 2012), and many of them (e.g., *V. cholerae*, *V. parahaemolyticus* and *V. vulnificus*) are potentially pathogenic (Coutinho et al., 2014; Salloto et al., 2012; Vieira et al., 2008). The presence of these bacteria represents a threat to the local population because many inhabitants use these habitats as a source of food and leisure (Fistarol et al., 2015).

Understanding the mechanisms by which environmental parameters influence microbial communities from urban estuaries is fundamental for developing policies to manage these environments. This knowledge can be applied to minimize the threat they pose to public health. Decades of studies at Guanabara Bay have revealed that nutrient concentrations (i.e., nitrogen and phosphorus sources) and temperature positively correlate with the abundance of bacterial cells, chlorophyll concentrations and *Vibrio* abundance, while the opposite effect has been observed for salinity (Andrade et al., 2003; Fistarol et al., 2015; Gonzalez et al., 2000; Gregoracci et al., 2012; Paranhos et al., 2001; Vieira et al., 2008). Nevertheless, these findings have been limited to short time scales and have not been integrated into models that could account for the synergistic effects of environmental variables.

Artificial neural networks (ANNs) have become increasingly common within the field of microbial ecology due to their ability to yield accurate predictive models that can account for non-linear associations and can incorporate interactions between variables (Flombaum et al., 2013; Kuang et al., 2016; Larsen et al., 2012; Santos et al., 2014; Tromas et al., 2017). An artificial neural network is a machine learning algorithm that can be trained to predict the value of a response variable based on the values of one or more predictor variables, much like in linear regression analysis. Our rationale was that data from a long-term ecological study of Guanabara Bay could be used to train artificial neural networks to accurately predict the abundances of biotic variables in response to abiotic factors.

In this study, our objective was to build artificial neural networks to model the response of the GB microbial community to environmental variables. We expected to obtain models capable of accurately predicting how environmental parameters regulate the abundances of three response variables: bacterial abundance, chlorophyll *a*, and *Vibrio*. Our findings allowed us to quantify the relative importance of different predictors; to model changes in the abundance of response variables across seasonal and pollution gradients; and, finally, to use these findings to propose strategies for the management and recovery of the Guanabara Bay ecosystem.

2. Methods

2.1. Sampling site description

Three sampling sites were assessed to quantify their physical, chemical and biological characteristics (Fig. 1). Surface water samples were collected monthly at each site over a period of five years (January 2009 until December 2013). Thus, 60 sampling events took place at each site, for a total of 180 samples. Each of the three sampling sites differed in terms of their proximity to the ocean and their degree of anthropogenic impact. Site 1 was located at the interface between Guanabara Bay and the Atlantic Ocean. Intense oceanic influence has made this habitat the least impacted among all three sites. Site 7 was at an intermediate zone located at the inner portion of the bay. Site 34 was located at the innermost location, with the highest degree of pollution among the three sites. Poor mixing of water and intensive sewage discharge has led to heavy eutrophication at site 34 (Gregoracci et al., 2012; Vieira et al., 2008).

2.2. Sampling and data collection

Water samples were assessed for levels of physical and chemical parameters by using oceanographic methods previously described by Grasshoff et al. (2009). We measured the water temperature (Temp) and salinity (Sal) using a CTD device. Transparency (Transp) was determined with a Secchi disc. Concentrations of total phosphorus (TP) were determined by acid digestion to phosphate, and concentrations of total nitrogen (TN) were determined by digestion with potassium persulfate following nitrate determination. Three biological parameters were also measured for each sample: 1) bacterial abundance (BacAbund), 2) chlorophyll *a* (Chloro) and 3) abundance of *Vibrio* cells (Vibrio). Bacterial abundance was measured using flow cytometry as described by (Cabral et al., 2017). Chlorophyll *a* analyses were performed following vacuum filtration (<25 cm of Hg). Chlorophyll was extracted from cellulose membrane Millipore HAWP 0.45 μm filters overnight in 90% acetone solution at 4 °C. Analyses were performed with a UV-VIS Perkin Elmer Lambda 20 spectrophotometer (Perkin Elmer, USA). The abundance of *Vibrio* cells was measured with counts of colony forming units (CFU) following inoculation of 100 μl of sample water onto *Vibrio* selective thiosulfate-citrate-bile-salts-sucrose (TCBS) agar plates and growth for 24 h at 30 °C (as described by Gregoracci et al. (2012)). These measurements of biotic and abiotic variables are displayed in Fig. S1.

2.3. Artificial neural network training and validation

The correlation coefficients estimated between predictors and response variables demonstrated that many of the predictor-response associations were non-linear, as evidenced by the values of the Pearson correlation scores being lower than the Spearman correlation scores (Fig. S2). This result suggested that such associations between predictor and response variables should not be modelled with linear models. Therefore, ANNs are an ideal alternative to model these relationships as they are better suited to account for the non-linear associations between variables and the interactions among predictors. Artificial neural network models were built to predict the responses of three biological variables: bacterial abundance, chlorophyll concentration and *Vibrio* abundance. Five predictor variables were used as inputs (Temp, Sal, TN, TP and Transp), resulting in a variable to sample ratio of 1:36, which was considered adequate to avoid overfitting (Dreiseitl and Ohno-Machado, 2002). Predictors and response variables were log10 transformed prior to training. This was done to reshape the association of data points, consequently improving the fitting capacity of the ANNs. All analyses were carried out in R (R Core Team, 2016) using functions from the 'nnet' package (Venables and Ripley, 2002) to train networks ('nnet' function) and from the 'NeuralNetTools' package (Beck, 2018) for the Lek profile and Olden importance analyses (using the 'lekprofile' and 'olden' functions, respectively). Training was performed setting the maximum number of iterations to 1000; reltol to 0.01; and weight decay to 0.001 (Krogh and Hertz, 1992). These values were selected in order to avoid overfitting. The 'reltol' setting was responsible for stopping the training process if the network reached a point where the optimization did not reduce the fit criterion by a factor of at least 1 – reltol. Simultaneously, over each iteration, weights were multiplied by the value set to weight decay, which prevented any weights from reaching extremely high values that lead to unrealistic models.

We tested the influence, on network performance, of the starting weights, the number of neurons, the number of training iterations and the number and set of samples used for training (Supplementary Text 1). During this step, samples were randomly and evenly divided into training and validation sets. ANN performance was evaluated by comparing the measured values of response variables against those predicted by the ANNs. We calculated the Pearson correlation coefficients for the training (PCCt) and validation (PCCv) sets and the root mean square errors for the training (RMSEt) and validation (RMSEv) sets. The results from these analyses (Supplementary Text 1) indicated that the parameters selected to train the networks were adequate for achieving good performance (i.e., high PCCt and low RMSEt values) without overfitting (i.e., adequate fits for both training and validation sets).

Thus, we proceeded to build the final ANNs using all samples from all sites ($n = 180$) for training. Pooling all samples together regardless of sampling site maximized the number of samples used for training and led to general models that could generate predictions regardless of sampling site or time point. The location and date of sampling were not used as predictor variables for training the networks; thus, any autocorrelation between samples collected close together in space or time could not bias the output or performance of the models. Effectively, not including the location and date of sampling during training meant that temporal trends in the associations between predictor and response variables were disregarded and that no spatial interactions were considered.

Ten thousand ANNs were built for each response variable to obtain independent models for bacterial abundance, chlorophyll and *Vibrio* abundance. ANNs were scored based on their RMSEt and PCCt values, and the relative importance of the predictor variables was estimated in all valid networks based on neuron weights as described by Olden et al., (2004). Networks were considered valid if they were characterized by PCCt >0.5 and RMSE <1. Networks with values of importance assigned to predictors below −10 or above +10 were not considered valid, in order to exclude over simplistic models that did not include the influence of all predictors. A single best ANN (Fig. S7) was selected for each response variable based on the lowest RMSEt. The 'Lek-profile method' (Lek et al., 1995) was used with the best network as a means of performing a sensitivity analysis. The end goal of the sensitivity analysis was to obtain predictions for each response variable across the range of values for each given predictor variable while holding all other predictor variables constant. Predictor variables were held constant at different values ranging from their 10th to 90th percentiles.

3. Results

3.1. Long-term description of GB microbial communities and environmental parameters

The five years of time-series data collected allowed for a characterization of physical, chemical and biological parameters in Guanabara Bay (Fig. S1): Sites GB1 and GB7 exhibited more stable values of predictors (Temp, Sal, Transp, TP, and TN) throughout the year, while Site GB34 exhibited the most notable monthly and yearly oscillations. Sites GB1 and GB7 were characterized by the lowest average values of TP and TN and the highest average Transp and Sal. The opposite pattern was observed for GB34, which displayed the highest average values of TP and TN and lowest Transp and Sal. Water temperature was comparable between sites but tended to be slightly higher at GB34.

We investigated the presence of associations between both predictor and response variables by calculating Pearson and Spearman correlation coefficients between all possible combinations of variables (Fig. S2). Salinity and transparency were each negatively correlated with the response variables, while temperature, TP and TN were each positively correlated with the response variables. The strong and significant associations observed between predictor and response variables was consistent with our rationale that these predictors could be used to model the values of the response variables. Yet, the absolute values of most Pearson correlation scores observed between the response variables (BacAbund, Chloro and Vibrio) and the predictors were below 0.6 (range: −0.69 to +0.67), suggesting that the variation within no single predictor could linearly account for the values of the response variables alone. Thus, the use of artificial neural networks was justified, since ANN models can incorporate multiple variables, and interactions between them, to maximize accuracy. Likewise, the lower yet significant correlation scores observed between predictors suggested little redundancy between them, justifying the inclusion of all of them in the models.

3.2. Artificial neural networks accurately model associations between predictors and response variables

Neural networks were successful in modelling the responses of bacterial abundance, chlorophyll concentration and *Vibrio* abundance to the five predictor variables (Table 1). The most accurate ANN, obtained for bacterial abundance, had the highest PCCt value (0.87) and lowest RMSEt (0.23) value, followed by the chlorophyll ANN (PCCt = 0.82 and RMSEt = 0.34), while the *Vibrio* ANN presented the lowest accuracy (PCCt = 0.67 and RMSEt = 0.55). Because to obtain these networks, all of the data were used for training, the values of PCCt = PCCv and RMSEt = RMSEv. Higher PCCt and lower RMSEt values implied better network performance. Thus, the highest precision was observed for the ANN built for bacterial abundance, followed by chlorophyll and lastly *Vibrio*. The differences in precision observed for the three final networks could have been a consequence of different levels of noise in the measurements of the response variables, or else they could have been due to the need for additional predictors to model the response variables. Nevertheless, network performance was comparable to that of other ANNs built to model microbial communities from other ecosystems (Flombaum et al., 2013; Larsen et al., 2012; Santos et al., 2014). The neural networks outperformed linear regression models previously estimated using time-series data from Guanabara Bay for bacterial abundance (linear regression PCC = 0.79) and *Vibrio* (PCC = 0.35) (Gregoracci et al., 2012). The superior performance of the ANNs compared to the linear regression models likely resulted from the fact that the neural networks incorporated interactions between predictors and because they modelled non-linear associations between predictor and response variables.

3.3. Predictor variables differ in their relative importance

The relative importance of the predictor values in the best performing ANNs was estimated using the Olden method (Fig. 2). This approach assigned importance values to the predictors that were proportional to their influence on the output of the model. These values were above zero if the predictor had a positive association with the response variable and below zero if the predictor had a negative association with the response variable. The further away from zero, the more important the predictor was, regardless of sign. For all three response variables, temperature consistently ranked as the most important predictor and had a positive association with all the response variables. Salinity was predicted to have a negative association with the response variables and was the second most important predictor of bacterial abundance and *Vibrio* but not for chlorophyll. Transparency was an important predictor of bacterial abundance and chlorophyll and was predicted to have a negative effect on these variables. Finally, TP and TN were often among the least important predictors in all cases. These trends of importance observed for the best networks were consistent with those observed for the 5% top-scoring ANNs. Although predictor importance varied among these top-scoring networks, the overall ranking, absolute values and sign of the predictors was conserved (Fig. S8).

Table 1
Performance and variable importance for the best performing networks for bacterial abundance, chlorophyll and *Vibrio*. The rank of predictor variable importance is included in parentheses.

	Performance		Predictor importance (rank)				
	PCC	RMSE	Importance Temp	Importance Sal	Importance TP	Importance TN	Importance Transp
Bacterial Abundance	0.87	0.23	9.14 (1)	−7.54 (2)	1.14 (4)	−0.17 (5)	−5.98 (3)
Chlorophyll	0.82	0.34	9.43 (1)	−2.43 (4)	0.35 (5)	2.48 (3)	−7.72 (2)
Vibrio	0.67	0.55	8.95 (1)	−6.39 (2)	−0.81 (5)	−3.01 (3)	2.43 (4)

3.4. Predictor variables displayed unique trends of association with the response variables

Sensitivity analysis was performed by obtaining predictions from the ANNs while changing the values of one predictor at a time, while all the other predictors were held constant at their 10th, 30th, 50th, 70th and 90th percentile values (Fig. 3). The shape and direction (negative or positive) of these curves demonstrated how each predictor variable individually influenced the output of the neural networks. The sensitivity analysis also revealed that the associations inferred by the ANNs were unique for each combination of response variable and predictor. Additionally, these curves showed how the predictor-response association could shift depending on the combination of the values of the remaining predictors, as highlighted by the differences between the curves depending on the percentile values at which the remaining predictor variables were held constant.

The sensitivity analysis indicated a slightly negative association between salinity and bacterial abundance, which shifted to slightly positive when the other predictors were above their 70th percentile values. The associations between salinity and chlorophyll and between salinity and *Vibrio* were steeply negative regardless of the percentile values of the other predictors. Meanwhile, temperature was positively associated with all of the response variables, independent of the percentile values of the other predictors. The association between bacterial abundance and TN levels shifted from slightly positive to slightly negative with increasing TN levels, tending to peak at intermediate values. Associations between response variables and TN were the most varied, often shifting from positive to negative with increasing TN concentrations, especially when the other predictors were set at their 70th or 90th percentile values. TP showed steep positive associations with all the response variables, but this effect was even more pronounced for bacterial abundance. Finally, transparency had a negative association with bacterial abundance and had almost no association with *Vibrio*, but the association of this predictor with chlorophyll shifted from positive to negative with increasing transparency.

3.5. Predictions for the GB microbiome in simulated scenarios

Next, ANNs were used to estimate the values of the response variables in simulated scenarios. ANNs were input with the monthly median value of each predictor at each of the three sampling sites. To investigate the isolated effect of changes in each predictor at every month and sampling site, we varied the values of one predictor at a time using different multiplication factors (0.8, 0.9, 0.95, 1, 1.05, 1.1 and 1.2) (Fig. 4). These simulations re-captured the peaks of bacterial abundance, chlorophyll and *Vibrio* that occur during the rainy season, i.e., the austral summer (December to March), in Guanabara Bay (Fig. S1). Consistent with the sensitivity analysis, these simulations predicted drastic changes in the levels of BacAbund, Chloro and *Vibrio* following changes in temperature of as little as 5% above or below the monthly median. These were even more pronounced using values 20% above or below the monthly median. Likewise, small changes in salinity

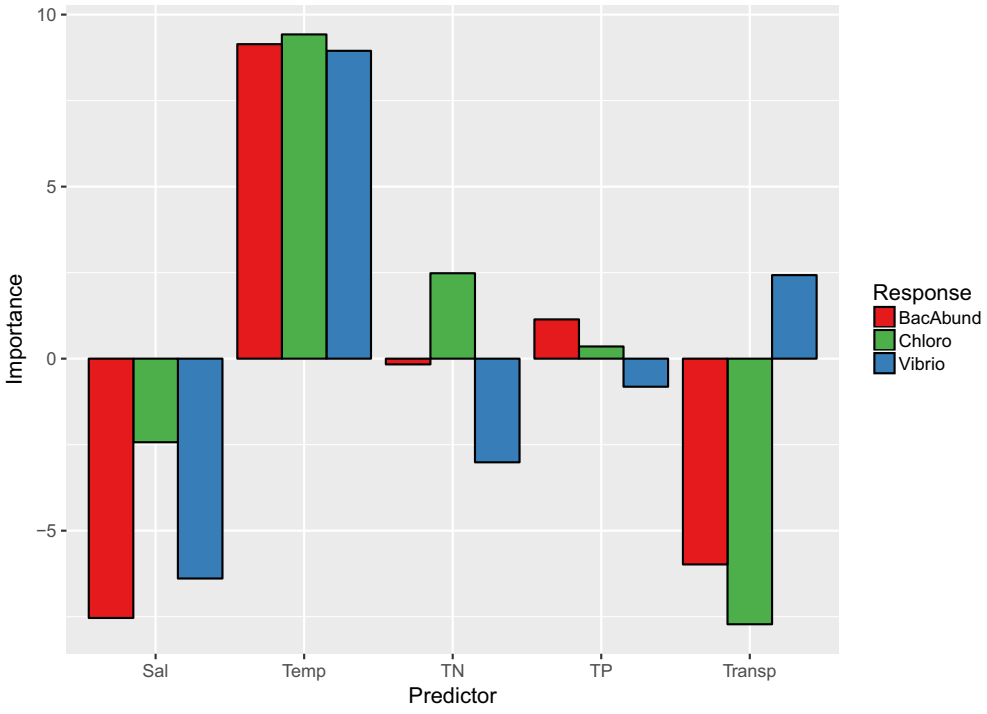


Fig. 2. Bar plot depicting the relative importance of predictors for the three response variables estimated from ANN weights through the Olden method.

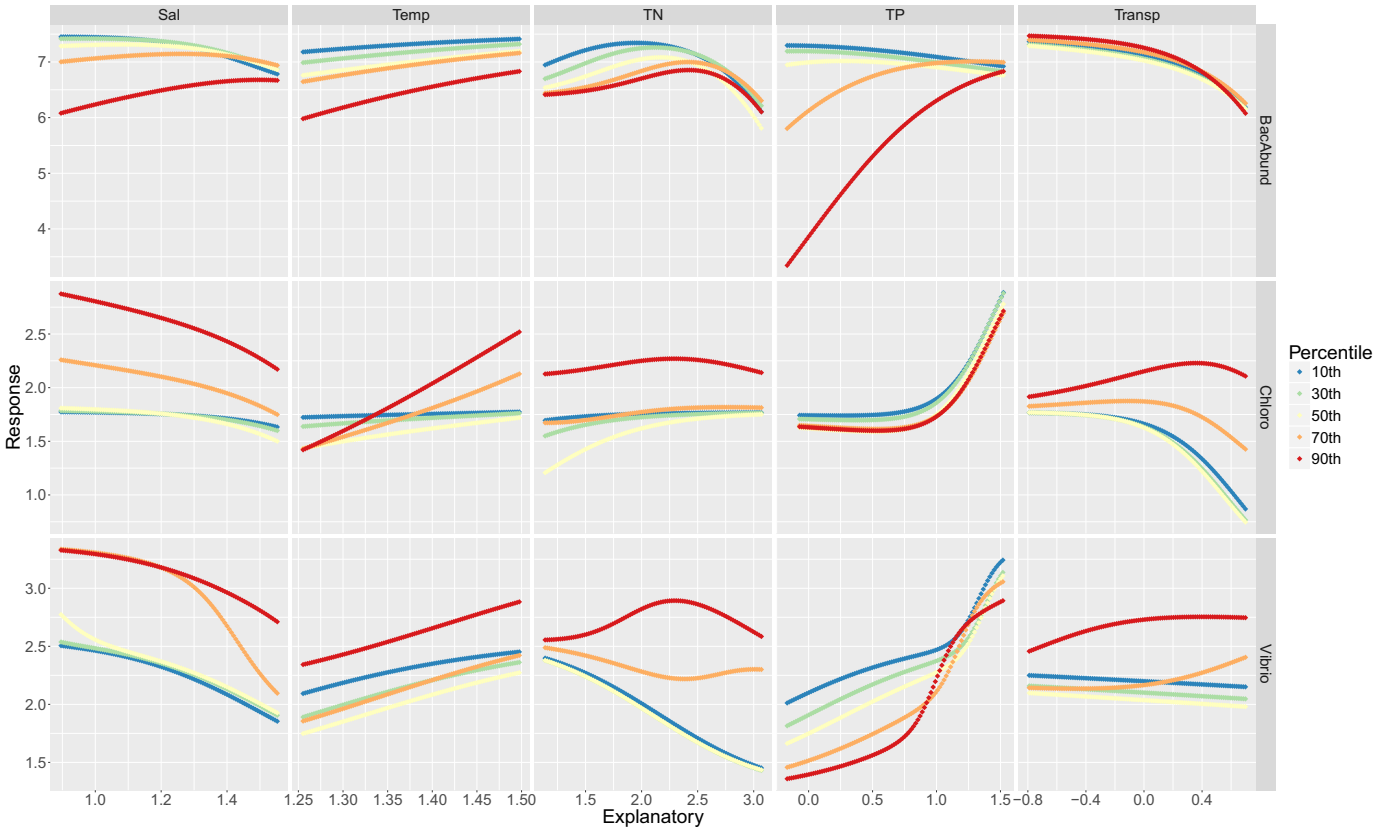


Fig. 3. Sensitivity of the response variables to changes in the predictor variables. The influence of each predictor (columns) over the response variable (rows) was modelled with all other input parameters held constant at values ranging from their 10th to 90th percentiles. The X axis depicts the log10 transformed value of the predictor variable, while the Y axis depicts the log10 transformed value of the response variable. Scales of the X and Y axes vary between rows and columns, respectively.

were predicted to have a strong effect on the abundance of bacterial cells and *Vibrio* but a small effect on chlorophyll. These associations between response variables and salinity and temperature were consistent across all sampling sites and months (Fig. 4). Changes in TP and TN affected the ANN predictions for the abundance of *Vibrio* across all sampling sites and months. However, these two predictors had little influence on the output of bacterial abundance and chlorophyll ANNs. Site 34 was an exception to this trend, where changes in TP were predicted to affect chlorophyll concentrations more drastically. Changes of up to 20% in transparency were expected to have no influence on *Vibrio* abundance. Nevertheless, the same changes in transparency were expected to affect bacterial abundance and chlorophyll. Interestingly, for chlorophyll these effects were expected to be much more pronounced at sites 01 and 07, which often have had lower values of transparency than site 34 (Fig. S1).

4. Discussion

4.1. Temperature and salinity regulate microbial community abundances in Guanabara Bay

ANN models allowed for three components of GB microbial communities to be evaluated: Total abundance of prokaryotic cells, photosynthesizers as measured through chlorophyll *a* concentrations, and *Vibrio*, a genus of aquatic copioheterotrophs. Temperature was ranked among the most important factors regulating the levels of the response variables (Fig. 2), and strong positive associations between temperature and these variables were observed (Fig. 3). Salinity ranked among the most important predictors but had a negative association with the abundances of bacterial cells, chlorophyll and *Vibrio*. Analyses of microbial communities spanning multiple ecosystems have suggested that salinity and temperature are major factors shaping microbial community composition across aquatic habitats (Lozupone and Knight, 2007; Sunagawa et al., 2015; Thompson et al., 2017). Our results demonstrate that these factors not only influence the taxonomic composition of microbial communities but also how these factors shape their abundances. Moreover, our results indicated that the importance of temperature and salinity markedly varies among the three components of the aquatic microbiome. The positive associations with temperature were likely a reflection of the increase in microbial metabolism brought by higher temperatures. Although the three variables were predicted to respond negatively to salinity (Figs. 2 and 3), this trend was much less pronounced for chlorophyll. This result suggests that, although shifts in salinity have affected the taxonomic composition of these sites (Gregoracci et al., 2012; Vieira et al., 2008), this has had a small effect on the overall chlorophyll levels of the communities.

Transparency was the second most important predictor of chlorophyll but was not as relevant for the abundance of bacteria and *Vibrio*. Interestingly, lower transparency levels were positively associated with the chlorophyll concentration but negatively associated at higher levels. This result suggests that the ideal levels of transparency that maximized chlorophyll production were between 0.8 m and 1.25 m, depending on the levels of the other predictor variables (Fig. 3). This trend is consistent with the photoinhibition process that affects the photosynthetic apparatus at very high levels of light irradiance (Dechatiwongse et al., 2014; Murata et al., 2007; Roach and Krieger-Liszkay, 2014). This phenomenon can be intensified by salt or temperature stresses (Lesser and Farrell, 2004; Murata et al., 2007; Sudhir and Murthy, 2004; Takahashi and Murata, 2008; Tang et al., 2007). Both of these are faced by the bay's photosynthetic microbiome. Importantly, other forms of chlorophyll besides the one that was measured in this study

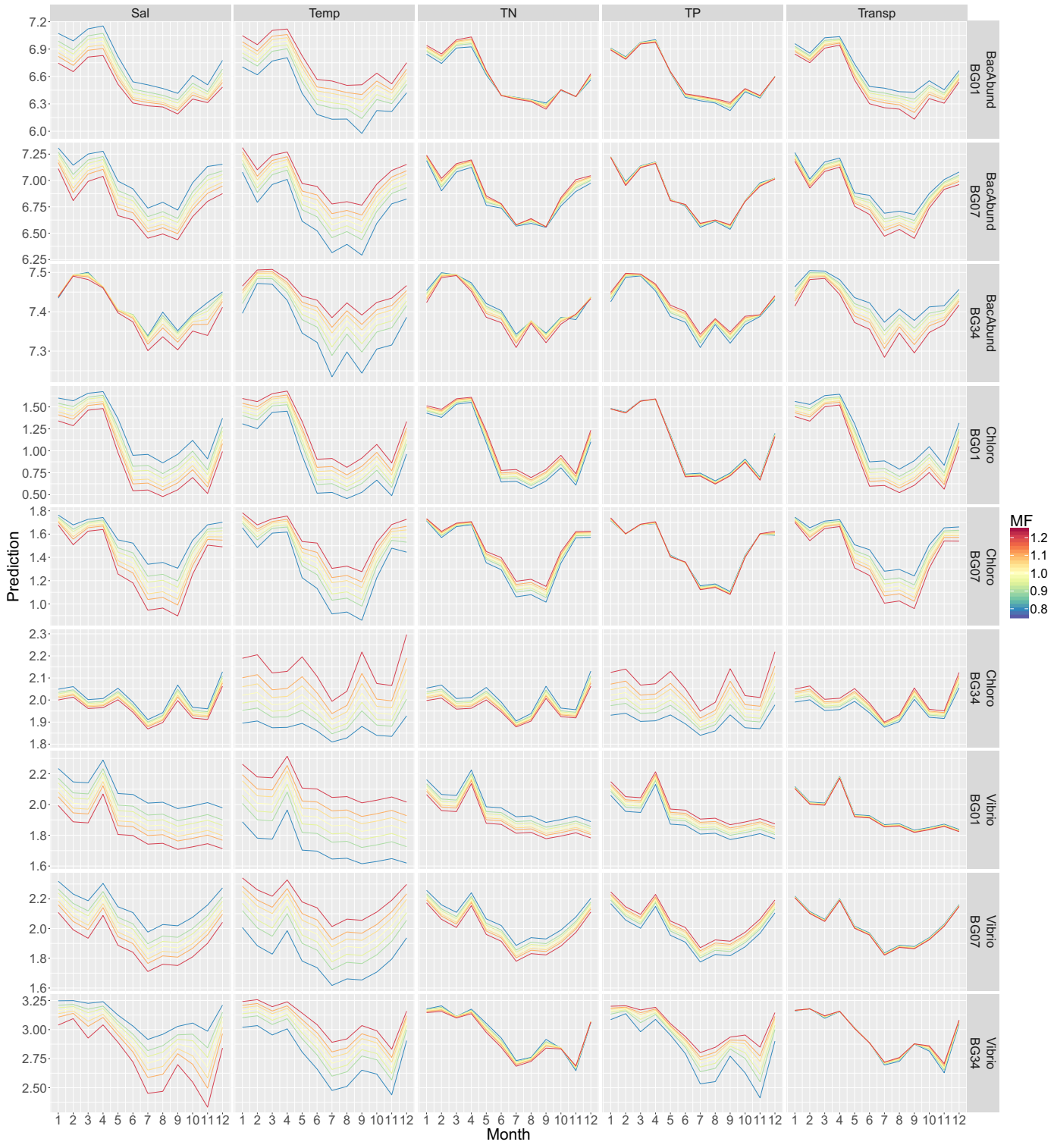


Fig. 4. ANN predictions for GB microbial communities for simulated scenarios of changing environmental parameters for each month and sampling site. Each panel represents a combination of predictors (columns) and response variable/sampling site (rows). Predictions were obtained by supplying neural networks with median values of each input variable at each month of the year. Coloured lines depict the predictions returned by the neural networks when varying the value of the tested predictor variable from 80 to 120% of the median of that specific month and sampling site, while the values of all the other predictor variables were held constant. The scales of the Y axes vary between rows. The networks predicted markedly distinct responses from the microbial community driven by increases or decreases in predictors depending on the sampling sites.

(chlorophyll *a*) could have had associations with predictor variables that were different from those presented here.

TN and TP ranked among the least important predictors for all response variables (Fig. 3). Nevertheless the importance values obtained for these predictors suggested that each of the three response variables have depended differently on nitrogen and phosphorus concentrations for growth. TN had almost no importance to BacAbund but had slightly higher importance for chlorophyll and *Vibrio*, while the opposite pattern was observed for TP, which was more important to BacAbund than it was for chlorophyll and *Vibrio*. Based on these observations, we

postulated that bacterial abundance has been much more dependent on TP concentrations than the other two response variables. Meanwhile, chlorophyll and *Vibrio* have depended more on TN concentrations than BacAbund.

The sensitivity analysis for TN and TP revealed unique trends for the associations between these predictors and the response variables, particularly when the other predictors were set to their 70th and 90th percentile values. Examples of such cases included the associations between TN-*Vibrio*, TP-*Vibrio* and TP-BacAbund. The shapes of these curves pointed to associations that were neither consistently positive nor consistently negative. One possible explanation for this was that these patterns were artefacts resulting from the fact that these variables were ranked among the least important, which allowed the ANNs to create such unusual patterns when modelling these associations without affecting performance. Considering that TN and TP did not rank as the most important factors for any of the response variables, we postulated that, at the time, physical parameters (i.e., temperature, salinity and transparency) were more relevant for determining the abundance of bacteria, chlorophyll and *Vibrio* than nutrients (i.e., TP and TN). Based on this evidence, we postulated that, due to the intense eutrophication at the bay, the microbial community had reached its maximum capacity for taking up and utilizing nutrients. Thus, the GB microbiome growth might have no longer been limited by nutrient availability. Instead, temperature, salinity and transparency have acted together in determining the abundance of bacteria, *Vibrio* and photosynthesizers through three major mechanisms: altering the taxonomic composition of the community, affecting their growth rates, and regulating the rates of photosynthesis and consequently primary productivity. In the original pristine conditions, the growth of the microbial community was likely limited by the availability of TP and TN.

4.2. Environmental factors and public health in Guanabara Bay

Our results suggested that *Vibrio* levels were primarily regulated by temperature and salinity. These results were in accordance with previous findings that explored the associations between *Vibrio* and these parameters (Constantin de Magny et al., 2008; Haley et al., 2014; Höfle et al., 2015; Vezzulli et al., 2016). Our results corroborated these findings while also elucidating the associations between *Vibrio* abundance with nitrogen, phosphorus and transparency. Therefore, the results suggest that the growth of *Vibrio* and of other potentially pathogenic bacteria has been fuelled by sewage input that releases phosphorus sources into the bay. This explanation concurs with previous analyses that have indicated that *Vibrio* and other copiotrophic and potentially pathogenic bacteria depend on phosphorus (Coutinho et al., 2015; Frost et al., 2008).

Members of the *Vibrio* genus thrive in eutrophic waters with ranges of temperature and salinity similar to those observed in Guanabara Bay (Constantin de Magny et al., 2008; Lobitz et al., 2000; Russek-cohen et al., 2003; Vezzulli et al., 2016). Moreover, the abundances of members of the genus *Vibrio* in aquatic habitats have been shown to be positively correlated with that of other genera of potentially pathogenic bacteria such as *Escherichia*, *Pseudomonas*, *Bacillus* and *Clostridium* (Coutinho et al., 2015). This means that whenever high abundances of *Vibrio* are observed, the abundance of other pathogenic bacteria is also expected to be high and vice-versa. Thus, the observed interactions between *Vibrio* and the predictor variables could likely be extrapolated to other opportunistic aquatic pathogens in Guanabara Bay. This was relevant because understanding the associations between pathogens and environmental conditions has been a fundamental step in predicting, preventing and mitigating the impacts of disease outbreaks associated with aquatic habitats (Lobitz et al., 2000; Russek-cohen et al., 2003). Our ANNs predicted an increased abundance of *Vibrio* with higher temperatures and lower salinity (Figs. 2 and 3). Together, these findings indicated that warmer and less saline waters of the innermost sections of GB have posed a higher risk for the local population than the colder and more saline waters typical of the regions that receive higher inputs of oceanic waters. Thus, the innermost regions of the bay likely have been the most threatening because the high sewage input and low influence of oceanic waters in these regions has created an ideal environment for the proliferation of *Vibrio* and other potential pathogens. Therefore, developing strategies for minimizing pollution into these regions of Guanabara Bay that represent the biggest threat to public health should be a priority.

4.3. Future changes and management of Guanabara Bay

Despite decades of investments, the water pollution at Guanabara Bay is increasing. The continuous growth of the metropolitan area of Rio de Janeiro will likely lead to increasing input of sewage into GB in the future. This will lead to increased nutrient concentrations and reduced salinity at the impacted sites. Furthermore, climate change is expected to increase water temperatures in Guanabara Bay. Our models cannot predict into the future as they were not built to incorporate associations between time and the response variables nor temporal changes in the associations between predictor and response variables. Nevertheless, the ANNs did reveal strong associations among salinity, temperature, nutrients and the microbial community of this habitat. Assuming that the observed associations between predictors and response variables remain stable through time, our results suggest that the aforementioned changes expected to affect this ecosystem in the future would lead to higher densities of bacteria and potential pathogens.

Our findings provide insights for developing strategies for reversing the impacts to Guanabara Bay. This could be achieved by a combination of proper sewage treatment, reduction of nutrient loads, minimizing deforestation and recovery of the surrounding and aquatic vegetation. Additionally, bioremediation strategies capable of reducing nutrient availability could be applied as well (Boesch et al., 2001; Greening and Janicki, 2006; Little et al., 2000; McGann et al., 2003; Paerl, 2009; Walker et al., 2013). Furthermore, our results demonstrate that ANN models could serve as tools to assess the threat level to public health posed by the aquatic ecosystem throughout changing environmental conditions, which could be used to predict and mitigate disease outbreaks (Constantin de Magny et al., 2008; Russek-cohen et al., 2003).

4.4. Network performance and limitations

Although the ANNs performed adequately they did not achieve perfect precision. This could have happened for multiple reasons, including the presence of noise in the data, the number of samples available for training, or the need to include more predictors. Our models did not incorporate some parameters that likely influence the microbial community of Guanabara Bay such as some physical (e.g. tides and precipitation) and biological (e.g., abundance of grazers and viruses) variables that could affect the microbial community. For the sake of simplicity, we also did not include temporal or spatial associations in the ANNs, which could also have hampered the performance of networks. We refrained from using time-series-specific neural networks because many of the phenomena that influence the microbiome of Guanabara Bay do not follow straightforward temporal trends. For example, rain, sewage discharge, and tides all impact the predictor variables in the bay, but those do not occur in any specific frequency that could have been resolved with our sampling strategy. Thus, the networks were built to model the associations between biological variables and abiotic parameters regardless of how these associations change through time. More complex models that incorporate the aforementioned variables would require a much larger number of samples but could provide a more comprehensive understanding of the dynamics taking place within the microbial community that resides in Guanabara Bay. Likewise, the advancement of machine learning approaches and perhaps the use of algorithms designed specifically to incorporate temporal trends (e.g., recurrent neural networks) could improve the precision of these

models. Nevertheless, our work provides a stepping stone for future studies that aim to understand the dynamics of the GB microbiome through ecological modelling approaches.

5. Conclusions

We generated ANNs with strong predictive capacities that could model the response of microbial communities to environmental parameters in Guanabara Bay. These models provided supporting evidence of how the abundances of bacteria, chlorophyll and *Vibrio* have been regulated by environmental parameters; ranked their relative importance over the response variables; and characterized the synergistic effects between variables. Furthermore, ANNs allowed us to infer the response of the microbial communities to changes in water quality conditions throughout seasonal and pollution gradients. Our findings provide insightful information on the dynamics of microbial communities for tropical estuarine ecosystems, for which little information is currently available. This approach could be easily applied to other similar datasets from other ecosystems and has served as a proof-of-principle of the usefulness of neural networks in the field of microbial ecology. This outcome is especially relevant considering the current scenarios of global climate changes and increasing environmental impacts. Future studies with a larger sample set, using more predictor variables, and incorporating associations between biotic variables will likely provide an even better description of the ecological associations taking place within the GB microbiome. Our findings have implications for the development of strategies to reduce the burden of waterborne diseases and for the remediation of urban aquatic ecosystems, to preserve their biodiversity as well as their economic, historical and aesthetic values.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2019.04.009.

Acknowledgements

The authors wish to thank Dr. Can Kesmir for insightful discussions. Authors were supported by CNPq, CAPES, FAPERJ, and NWO. FHC was supported by a grant from “Ciência sem fronteiras”. BED was supported by the Netherlands Organization for Scientific Research (NWO) Vidi grant 864.14.004.

Competing interests

The authors declare no competing interests.

References

Andrade, L., Gonzalez, A., Araujo, F., Paranhos, R., 2003. Flow cytometry assessment of bacterioplankton in tropical marine environments. J. Microbiol. Methods 55, 841–850.
Beck, M., 2018. NeuralNetTools: Visualization and Analysis Tools for Neural Networks.
Boesch, D.F., Brinsfield, R.B., Magnien, R.E., 2001. Chesapeake bay eutrophication. J. Environ. Qual. 30, 303–320.
Cabral, A.S., Lessa, M.M., Junger, P.C., Thompson, F.L., Paranhos, R., 2017. Virioplankton dynamics are related to eutrophication levels in a tropical urbanized bay. PLoS One 12. https://doi.org/10.1371/journal.pone.0174653.
Constantin de Magny, G., Murtugudde, R., Sapiano, M.R.P., Nizam, A., Brown, C.W., Busalacchi, A.J., Yunus, M., Nair, G.B., Gil, A.I., Lanata, C.F., Calkins, J., Manna, B., Rajendran, K., Bhattacharya, M.K., Huq, A., Sack, R.B., Colwell, R.R., 2008. Environmental signatures associated with cholera epidemics. Proc. Natl. Acad. Sci. 105, 17676–17681. https://doi.org/10.1073/pnas.0809654105.
Coutinho, F.H., Silveira, C.B., Pinto, L.H., Salloto, G.R.B., Cardoso, A.M., Martins, O.B., Vieira, R.P., Clementino, M.M., 2014. Antibiotic resistance is widespread in urban aquatic environments of Rio de Janeiro, Brazil. Microb. Ecol. 68, 441–452. https://doi.org/10.1007/s00248-014-0422-5.
Coutinho, F.H., Meirelles, P.M., Moreira, A.P.B., Paranhos, R.P., Dutilh, B.E., Thompson, F.L., 2015. Niche distribution and influence of environmental parameters in marine microbial communities: a systematic review. PeerJ 3, e1008. https://doi.org/10.7717/peerj.1008.
Dechatiwongse, P., Srisamai, S., Maitland, G., Hellgardt, K., 2014. Effects of light and temperature on the photoautotrophic growth and photoinhibition of nitrogen-fixing cyanobacterium Cyanothece sp. ATCC 51142. Algal Res. 5, 103–111. https://doi.org/10.1016/j.algal.2014.06.004.
Dreiseitl, S., Ohno-Machado, L., 2002. Logistic regression and artificial neural network classification models: a methodology review. J. Biomed. Inform. 35, 352–359. https://doi.org/10.1016/S1532-0464(03)00034-0.
Fistarol, G.O., Coutinho, F.H., Moreira, A.P.B., Venas, T., Cánovas, A., de Paula, S.E.M., Coutinho, R., de Moura, R.L., Valentin, J.L., Tenenbaum, D.R., Paranhos, R., do Valle, R. de A.B., Vicente, A.C.P., Amado Filho, G.M., Pereira, R.C., Kruger, R., Rezende, C.E., Thompson, C.C., Salomon, P.S., Thompson, F.L., 2015. Environmental and sanitary conditions of Guanabara Bay, Rio de Janeiro. Front. Microbiol. 6, 1–17. https://doi.org/10.3389/fmicb.2015.01232.
Flombaum, P., Gallegos, J.L., Gordillo, R.A., Rincón, J., Zabala, L.L., Jiao, N., Karl, D., Li, W., Lomas, M., Veneziano, D., Vera, C., Vrugt, J.A., Martiny, A.C., 2013. Present and future global distributions of the marine Cyanobacteria Prochlrococcus and Synechococcus. Pnas 110, 9824–9829. https://doi.org/10.1073/pnas.1307701110/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1307701110.
Frost, P.C., Ebert, D., Smith, V.H., 2008. Responses of a bacterial pathogen to phosphorus limitation of its aquatic invertebrate host. Ecology 89, 313–318. https://doi.org/10.1890/07-0389.1.
Gonzalez, A.M., Paranhos, R., Andrade, L., Valentin, J.L., 2000. Bacterial production in Guanabara Bay (Rio de Janeiro, Brazil) evaluated by 3 H-leucine incorporation. Braz. Arch. Biol. Technol. 43, 493–500. https://doi.org/10.1590/S1516-89132000000500008.
Grasshoff, K., Kremling, K., Ehrhardt, M., 2009. Methods of Seawater Analysis. John Wiley & Sons.
Greening, H., Janicki, A., 2006. Toward reversal of eutrophic conditions in a subtropical estuary: water quality and seagrass response to nitrogen loading reductions in Tampa Bay, Florida, USA. Environ. Manag. 38, 163–178. https://doi.org/10.1007/s00267-005-0079-4.
Gregoracci, G.B., Nascimento, J.R., Cabral, A.S., Paranhos, R., Valentin, J.L., Thompson, C.C., Thompson, F.L., 2012. Structuring of bacterioplankton diversity in a large tropical bay. PLoS One 7, e31408. https://doi.org/10.1371/journal.pone.0031408.
Haley, B.J., Kokashvili, T., Tskshvediani, A., Janelidze, N., Mitaishvili, N., Grim, C.J., de Magny, G.C., Chen, A.J., Taviani, E., Eliashvili, T., Tediashvili, M., Whitehouse, C.A., Colwell, R.R., Huq, A., 2014. Molecular diversity and predictability of Vibrio parahaemolyticus along the Georgian coastal zone of the Black Sea. Front. Microbiol. 5, 1–9. https://doi.org/10.3389/fmicb.2014.00045.
Höfle, M., Pezzati, E., Vezzulli, L., Brettar, I., Pruzzo, C., 2015. Effects of global warming on Vibrio ecology. Microbiol. Spectr. 3. https://doi.org/10.1128/microbiolspec.ve-0004-2014.
Kjerfve, B., Ribeiro, C.H.A., Dias, G.T.M., Filippo, A.M., Da Silva Quaresma, V., 1997. Oceanographic characteristics of an impacted coastal bay: Baía de Guanabara, Rio de Janeiro, Brazil. Cont. Shelf Res. 17, 1609–1643. https://doi.org/10.1016/S0278-4343(97)00028-9.
Krogh, A., Hertz, J.A., 1992. A simple weight decay can improve generalization. Adv. Neural Inf. Proces. Syst. 4, 950–957.
Kuang, J., Huang, L., He, Z., Chen, L., Hua, Z., Jia, P., Li, S., Liu, J., Li, J., Zhou, J., Shu, W., 2016. Predicting taxonomic and functional structure of microbial communities in acid mine drainage. ISME J. 10, 1–13. https://doi.org/10.1038/ismej.2015.201.
Larsen, P.E., Field, D., Gilbert, J.A., 2012. Predicting bacterial community assemblages using an artificial neural network approach. Nat. Methods 9, 621–625. https://doi.org/10.1038/nmeth.1975.
Lek, S., Belaud, A., Dimopoulos, I., Lauga, J., Moreau, J., 1995. Improved estimation, using neural networks, of the food consumption of fish populations. Mar. Freshw. Res. 46, 1229–1236.
Lesser, M.P., Farrell, J.H., 2004. Exposure to solar radiation increases damage to both host tissues and algal symbionts of corals during thermal stress. Coral Reefs 23, 367–377. https://doi.org/10.1007/s00338-004-0392-z.
Little, J.L., Hall, R.I., Quinlan, R., Smol, J.P., 2000. Past trophic status and hypolimnetic anoxia during eutrophicaton and remediation of Gravenhurst Bay, Ontario: comparison of diatoms, chironomids, and historical records. Can. J. Fish. Aquat. Sci. 57, 333–341. https://doi.org/10.1139/f99-235.
Lobitz, B., Beck, L., Huq, A., Wood, B., Fuchs, G., Faruque, A.S., Colwell, R., 2000. Climate and infectious disease: use of remote sensing for detection of Vibrio cholerae by indirect measurement. Proc. Natl. Acad. Sci. U. S. A. 97, 1438–1443. https://doi.org/10.1073/pnas.97.4.1438.
Lozupone, C.A., Knight, R., 2007. Global patterns in bacterial diversity. Proc. Natl. Acad. Sci. U. S. A. 104, 11436–11440.
Mayr, L.M., Tenenbaum, D.R., Villac, M.C., Paranhos, R.P., Nogueira, C.R., Bonecker, S.L.C., Bonecker, A.C.T., 1989. Hydrobiological Characterization of Guanabara Bay. Coastlines of Brazil. pp. 124–138.
McGann, M., Alexander, C.R., Bay, S.M., 2003. Response of benthic foraminifers to sewage discharge and remediation in Santa Monica Bay, California. Mar. Environ. Res. 56, 299–342. https://doi.org/10.1016/S0141-1136(02)00336-7.
Murata, N., Takahashi, S., Nishiyama, Y., Allakhverdiev, S.I., 2007. Photoinhibition of photosystem II under environmental stress. Biochim. Biophys. Acta Bioenerg. 1767, 414–421. https://doi.org/10.1016/j.bbabio.2006.11.019.
Olden, J.D., Joy, M.K., Death, R.G., 2004. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecol. Model. 178, 389–397. https://doi.org/10.1016/j.ecolmodel.2004.03.013.
Paerl, H.W., 2009. Controlling eutrophication along the freshwater-marine continuum: dual nutrient (N and P) reductions are essential. Estuar. Coasts 32, 593–601. https://doi.org/10.1007/s12237-009-9158-8.
Paranhos, R., Pereira, A.P., Mayr, L.M., 1998. Diel variability of water quality in a tropical polluted bay. Environ. Monit. Assess. 50, 131–141.
Paranhos, R., Andrade, L., Mendonça-Hagler, L.C., Pfeiffer, W.C., 2001. Coupling bacterial abundance with production in a polluted tropical coastal bay. In: Faria, B.M., Farjalla, V.F., Esteves, F.A. (Eds.), Aquatic Microbial Ecology in Brazil. Series Oecologia Brasiliensis. Oecologia Bras, Rio de Janeiro, Brazil.

R Core Team, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Roach, T., Krieger-Liszkay, A., 2014. Regulation of photosynthetic electron transport and photoinhibition. Curr. Protein Pept. Sci. 15, 351–362. https://doi.org/10.2174/1389203715666140327105143.

Russek-cohen, E., Choopun, N., Rivera, I.N.G., Gangle, B., Jiang, S.C., Rubin, A., Patz, J. a, Huq, A., Colwell, R.R., 2003. Predictability of Vibrio cholerae in Chesapeake Bay Vale. Appl. Environ. Microbiol. 69, 2773–2785. https://doi.org/10.1128/AEM.69.5.2773.

Salloto, G.R.B., Cardoso, A.M., Coutinho, F.H., Pinto, L.H., Vieira, R.P., Chaia, C., Lima, J.L., Albano, R.M., Martins, O.B., Clementino, M.M., 2012. Pollution impacts on bacterioplankton diversity in a tropical urban coastal lagoon system. PLoS One 7, e51175. https://doi.org/10.1371/journal.pone.0051175.

Santos, V.S., Villac, M.C., Tenenbaum, D.R., Paranhos, R., 2007. Auto-and heterotrophic nanoplankton and filamentous bacteria of Guanabara Bay (RJ, Brazil): estimates of cell/filament numbers versus carbon content. Braz. J. Oceanogr. 55, 133–143. https://doi.org/10.1590/S1679-87592007000200006.

Santos, E.C., Armas, E.D., Crowley, D., Lambais, M.R., 2014. Artificial neural network modeling of microbial community structures in the Atlantic Forest of Brazil. Soil Biol. Biochem. 69, 101–109. https://doi.org/10.1016/j.soilbio.2013.10.049.

Sudhir, P., Murthy, S.D.S., 2004. Effects of salt stress on basic processes of photosynthesis. pdf. Photosynthetica 42, 481–486.

Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D.R., Alberti, A., Cornejo-castillo, F.M., Costea, P.I., Cruaud, C., Ovidio, F., Engelen, S., Ferrera, I., Gasol, J.M., Guidi, L., Hildebrand, F., Kokoszka, F., Lepoivre, C., 2015. Structure and function of the global ocean microbiome. Science 348, 1–10. https://doi.org/10.1126/science.1261359 80-..

Takahashi, S., Murata, N., 2008. How do environmental stresses accelerate photoinhibition? Trends Plant Sci. 13, 178–182. https://doi.org/10.1016/j.tplants.2008.01.005.

Tang, D., Shi, S., Li, D., Hu, C., Liu, Y., 2007. Physiological and biochemical responses of Scytonema javanicum (cyanobacterium) to salt stress. J. Arid Environ. 71, 312–320. https://doi.org/10.1016/j.jaridenv.2007.05.004.

Thompson, F.L., Bruce, T., Gonzalez, A., Cardoso, A., Clementino, M., Costagliola, M., Hozbor, C., Otero, E., Piccini, C., Peressutti, S., Schmieder, R., Edwards, R., Smith, M., Takiyama, L.R., Vieira, R., Paranhos, R., Artigas, L.F., 2011. Coastal bacterioplankton community diversity along a latitudinal gradient in Latin America by means of V6 tag pyrosequencing. Arch. Microbiol. 193, 105–114. https://doi.org/10.1007/s00203-010-0644-y.

Thompson, L.R., Sanders, J.G., McDonald, D., Amir, A., Ladau, J., Locey, K.J., Prill, R.J., Tripathi, A., Gibbons, S.M., Ackermann, G., Navas-Molina, J.A., Janssen, S., Kopylova, E., Vázquez-Baeza, Y., González, A., Morton, J.T., Mirarab, S., Zech Xu, Z., Jiang, L., Haroon, M.F., Kanbar, J., Zhu, Q., Jin Song, S., Kosciolek, T., Bokulich, N.A., Lefler, J., Brislawn, C.J., Humphrey, G., Owens, S.M., Hampton-Marcell, J., Berg-Lyons, D., McKenzie, V., Fierer, N., Fuhrman, J.A., Clauset, A., Stevens, R.L., Shade, A., Pollard, K.S., Goodwin, K.D., Jansson, J.K., Gilbert, J.A., Knight, R., Rivera, J.L.A., Al-Moosawi, L., Alverdy, J., Amato, K.R., Andras, J., Angenent, L.T., Antonopoulos, D.A., Apprill, A., Armitage, D., Ballantine, K., Bárta, J., Baum, J.K., Berry, A., Bhatnagar, A., Bhatnagar, M., Biddle, J.F., Bittner, L., Boldgiv, B., Bottos, E., Boyer, D.M., Braun, J., Brazelton, W., Brearley, F.Q., Campbell, A.H., Caporaso, J.G., Cardona, C., Carroll, J., Cary, S.C., Casper, B.B., Charles, T.C., Chu, H., Claar, D.C., Clark, R.G., Clayton, J.B., Clemente, J.C., Cochran, A., Coleman, M.L., Collins, G., Colwell, R.R., Contreras, M., Crary, B.B., Creer, S., Cristol, D.A., Crump, B.C., Cui, D., Daly, S.E., Davalos, L., Dawson, R.D., Defazio, J., Delsuc, F., Dionisi, H.M., Dominguez-Bello, M.G., Dowell, R., Dubinsky, E.A., Dunn, P.O., Ercolini, D., Espinoza, R.E., Ezenwa, V., Fenner, N., Findlay, H.S., Fleming, I.D., Fogliano, V., Forsman, A., Freeman, C., Friedman, E.S., Galindo, G., Garcia, L., Garcia-Amado, M.A., Garshelis, D., Gasser, R.B., Gerdts, G., Gibson, M.K., Gifford, I., Gill, R.T., Giray, T., Gittel, A., Golyshin, P., Gong, D., Grossart, H.-P., Guyton, K., Haig, S.-J., Hale, V., Hall, R.S., Hallam, S.J., Handley, K.M., Hasan, N.A., Haydon, S.R., Hickman, J.E., Hidalgo, G., Hofmockel, K.S., Hooker, J., Hulth, S., Hultman, J., Hyde, E., Ibáñez-Álamo, J.D., Jastrow, J.D., Jex, A.R., Johnson, L.S., Johnston, E.R., Joseph, S., Jurburg, S.D., Jurelevicius, D., Karlsson, A., Karlsson, R., Kauppinen, S., Kellogg, C.T.E., Kennedy, S.J., Kerkhof, L.J., King, G.M., Kling, G.W., Koehler, A.V., Krezalek, M., Kueneman, J., Lamendella, R., Landon, E.M., Lane-deGraaf, K., LaRoche, J., Larsen, P., Laverock, B., Lax, S., Lentino, M., Levin, I.I., Liancourt, P., Liang, W., Linz, A.M., Lipson, D.A., Liu, Y., Lladser, M.E., Lozada, M., Spirito, C.M., MacCormack, W.P., MacRae-Crerar, A., Magris, M., Martín-Platero, A.M., Martín-Vivaldi, M., Martínez, L.M., Martínez-Bueno, M., Marzinelli, E.M., Mason, O.U., Mayer, G.D., McDevitt-Irwin, J.M., McDonald, J.E., McGuire, K.L., McMahon, K.D., McMinds, R., Medina, M., Mendelson, J.R., Metcalf, J.L., Meyer, F., Michelangeli, F., Miller, K., Mills, D.A., Minich, J., Mocali, S., Moitinho-Silva, L., Moore, A., Morgan-Kiss, R.M., Munroe, P., Myrold, D., Neufeld, J.D., Ni, Y., Nicol, G.W., Nielsen, S., Nissimov, J.I., Niu, K., Nolan, M.J., Noyce, K., O'Brien, S.L., Okamoto, N., Orlando, L., Castellano, Y.O., Osuolale, O., Oswald, W., Parnell, J., Peralta-Sánchez, J.M., Petraitis, P., Pfister, C., Pilon-Smits, E., Piombino, P., Pointing, S.B., Pollock, F.J., Potter, C., Prithiviraj, B., Quince, C., Rani, A., Ranjan, R., Rao, S., Rees, A.P., Richardson, M., Riebesell, U., Robinson, C., Rockne, K.J., Rodriguezl, S.M., Rohwer, F., Roundstone, W., Safran, R.J., Sangwan, N., Sanz, V., Schrenk, M., Schrenzel, M.D., Scott, N.M., Seger, R.L., Seguin-Orlando, A., Seldin, L., Seyler, L.M., Shakhsheer, B., Sheets, G.M., Shen, C., Shi, Y., Shin, H., Shogan, B.D., Shutler, D., Siegel, J., Simmons, S., Sjöling, S., Smith, D.P., Soler, J.J., Sperling, M., Steinberg, P.D., Stephens, B., Stevens, M.A., Taghavi, S., Tai, V., Tait, K., Tan, C.L., Tas, N., Taylor, D.L., Thomas, T., Timling, I., Turner, B.L., Urich, T., Ursell, L.K., van der Lelie, D., Van Treuren, W., van Zwieten, L., Vargas-Robles, D., Thurber, R.V., Vitaglione, P., Walker, D.A., Walters, W.A., Wang, S., Wang, T., Weaver, T., Webster, N.S., Wehrle, B., Weisenhorn, P., Weiss, S., Werner, J.J., West, K., Whitehead, A., Whitehead, S.R., Whittingham, L.A., Willerslev, E., Williams, A.E., Wood, S.A., Woodhams, D.C., Yang, Y., Zaneveld, J., Zarraonaindia, I., Zhang, Q., Zhao, H., 2017. A communal catalogue reveals Earth's multiscale microbial diversity. Nature https://doi.org/10.1038/nature24621.

Tromas, N., Fortin, N., Bedrani, L., Terrat, Y., Cardoso, P., Bird, D., Greer, C.W., Shapiro, B.J., 2017. Characterising and predicting cyanobacterial blooms in an 8-year amplicon sequencing time course. ISME J. 11, 1746–1763.

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S. Fourth ed. Springer, New York.

Vezzulli, L., Grande, C., Reid, P.C., Hélaouët, P., Edwards, M., Höfle, M.G., Brettar, I., Colwell, R.R., Pruzzo, C., 2016. Climate influence on Vibrio and associated human diseases during the past half-century in the coastal North Atlantic. Proc. Natl. Acad. Sci. 113, E5062–E5071. https://doi.org/10.1073/pnas.1609157113.

Vieira, R.P., Clementino, M.M., Cardoso, A.M., Oliveira, D.N., Albano, R.M., Gonzalez, A.M., Paranhos, R., Martins, O.B., 2007. Archaeal communities in a tropical estuarine ecosystem: Guanabara Bay, Brazil. Microb. Ecol. 54, 460–468. https://doi.org/10.1007/s00248-007-9261-y.

Vieira, R.P., Gonzalez, A.M., Cardoso, A.M., Oliveira, D.N., Albano, R.M., Clementino, M.M., Martins, O.B., Paranhos, R., 2008. Relationships between bacterial diversity and environmental variables in a tropical marine environment, Rio de Janeiro. Environ. Microbiol. 10, 189–199. https://doi.org/10.1111/j.1462-2920.2007.01443.x.

Villac, M.C., Tenenbaum, D.R., 2010. The phytoplankton of Guanabara Bay, Brazil: I. Historical account of its biodiversity. Biota Neotrop. 10, 271–293. https://doi.org/10.1590/S1676-06032010000200030.

Walker, T.R., MacAskill, D., Rushton, T., Thalheimer, A., Weaver, P., 2013. Monitoring effects of remediation on natural sediment recovery in Sydney Harbour, Nova Scotia. Environ. Monit. Assess. 185, 8089–8107. https://doi.org/10.1007/s10661-013-3157-8.