Global Biogeochemical Cycles

RESEARCH ARTICLE
10.1029/2018GB005992

A Machine Learning (kNN) Approach to Predicting Global Seafloor Total Organic Carbon

Taylor R. Lee[1], **Warren T. Wood**[1], **and Benjamin J. Phrampus**[2]

[1]U.S. Naval Research Laboratory, John C. Stennis Space Center, Hancock County, MS, USA, [2]ASEE Postdoctoral Program, U.S. Naval Research Laboratory, John C. Stennis Space Center, Hancock County, MS, USA

Key Points:
- Total organic carbon in the shallow seafloor is a fundamental quantity for many subsurface processes but is only very sparsely sampled
- We use machine learning techniques to predict total organic carbon for the entire seafloor, with uncertainty, a 5 × 5-arc minute grid
- Parameter space proximity indicates where and what kinds of future measurements are the most optimal for reducing prediction uncertainty

Supporting Information:
- Supporting Information S1
- Data Set S1
- Data Set S2
- Data Set S3
- Data Set S4
- Data Set S5
- Data Set S6
- Data Set S7

Correspondence to:
T. R. Lee and W. T. Wood,
taylor.lee@nrlssc.navy.mil;
warren.wood@nrlssc.navy.mil

Citation:
Lee, T. R., Wood, W. T., & Phrampus, B. J. (2019). A machine learning (kNN) approach to predicting global seafloor total organic carbon. *Global Biogeochemical Cycles*, *33*, 37–46. https://doi.org/10.1029/2018GB005992

Received 4 JUN 2018
Accepted 20 DEC 2018
Accepted article online 4 JAN 2019
Published online 23 JAN 2019

Abstract Seafloor properties, including total organic carbon (TOC), are sparsely measured on a global scale, and interpolation (prediction) techniques are often used as a proxy for observation. Previous geospatial interpolations of seafloor TOC exhibit gaps where little to no observed data exists. In contrast, recent machine learning techniques, relying on geophysical and geochemical properties (e.g., seafloor biomass, porosity, and distance from coast), show promise in making comprehensive, statistically optimal predictions. Here we apply a nonparametric (i.e., data-driven) machine learning algorithm, specifically k-nearest neighbors (kNN), to estimate the global distribution of seafloor TOC. Our results include predictor (feature) selection specifically designed to mitigate bias and produce a statistically optimal estimation of seafloor TOC, with uncertainty, at 5 × 5-arc minute resolution. Analysis of parameter space sample density provides a guide for future sampling. One use for this prediction is to constrain a global inventory, indicating that just the upper 5 cm of the seafloor contains about 87 ± 43 gigatons of carbon (Gt C) in organic form.

1. Introduction

Total organic carbon (TOC) in seafloor sediment is a commonly made measurement in the marine sciences, and a database of historic measurements in the upper 5 cm has been made widely available (Seiter et al., 2004). The accumulation and degradation of this carbon pool provide strong controls on environmental and carbon cycling processes (Schoepfer et al., 2015). Accurate estimations of seafloor TOC serve as the basis for many empirical and theoretical models estimating organic matter transformation and degradation, particularly methane (CH_4) (Burdige, 2007; Colwell et al., 2008; Middelburg, 1989; Pinero et al., 2013; Rothman & Forney, 2007). Seafloor TOC is the principle source of shallow biogenic methane gas production (Arndt et al., 2013), while deeply buried organic carbon provides the source material for virtually all hydrocarbons, including those used for fossil fuels, that is, oil and thermogenic methane.

If buried quickly enough below the oxic zone, TOC forms biogenic and thermogenic methane gas within the subsurface that can escape the seafloor through conduits (seafloor seeps), allowing gas to enter the water column (Hovland & Judd, 2007). Methane cold seeps facilitate reactions that contribute to the formation of authigenic carbonate and support chemosynthetic biota (Levin et al., 2016). Furthermore, methane produced in the subsurface under the appropriate pressure and temperature conditions may solidify as methane hydrate (Sloan, 2003). Natural methane hydrates have been studied as a potential energy resource (Max et al., 2006), as well as a potential agent of climate change (Ruppel & Kessler, 2017).

Despite the importance of seafloor TOC, direct observations are sparse, and large areas of the seafloor remain virtually unsampled (Figure 1). Acquisition of TOC and similar seafloor property data is particularly difficult and expensive due to familiar difficulties in marine science data acquisition, for example, distant locations, deep water, and challenging environments. As a result, we are left with a global data set largely inadequate for addressing fundamental issues, namely, assessing global inventories or determining likely locations of methane or other hydrocarbon accumulations.

Historically, estimates in areas where data acquisition is limited or nonexistent have used some form of interpolation from existing data. In a global distribution analysis of seafloor TOC, Seiter et al. (2004) interpretively delineated regional geological provinces by various physical and chemical properties. Each geological province was predicted using spatial interpolation (i.e., kriging) to determine the distribution of seafloor TOC. In regional provinces lacking sufficient observed data (e.g., Western Pacific Ocean and portions of the Arctic), no prediction was possible. Similarly, Goutorbe et al. (2011) estimated global heat flux, using several categorical predictors such as basin and rift types.

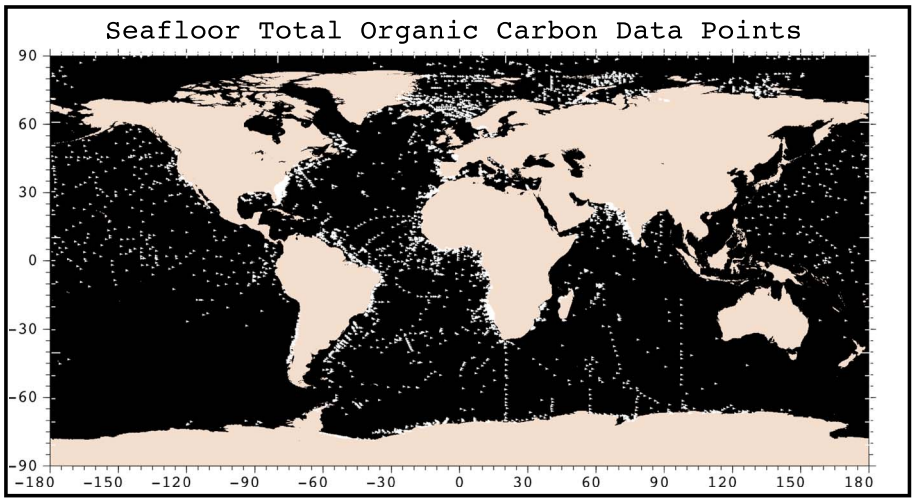

Figure 1. Locations of sampled seafloor total organic carbon (white) representing the upper 5 cm of sediment using a 5-arc minute resolution (data from Seiter et al., 2004; Beazley, 2003).

Recently, machine learning techniques have provided accurate global estimations of geological properties where little to no data presently exist. Sediment porosity at a 5-arc minute resolution was predicted using a random forest algorithm in an analysis by Martin et al. (2015). This machine learning technique uses a series of regression trees to determine new outcomes (i.e., predictions) based on associations of global properties with previously observed data. Martin et al. (2015) showed this technique to be more accurate in predictions than strictly geospatial interpolation techniques. Additionally, other data sets have been predicted from machine learning algorithms; one estimated global seafloor lithology (Dutkiewicz et al., 2015) and the other estimated seafloor biomass distribution (Wei et al., 2010).

In several of these examples, significant prior knowledge, beyond what was strictly observed, influenced the final estimate, mostly by interpretively restricting the number and types of predictors used or by constructing interpretive (as opposed to observed) predictors. Our intent here is to employ a decidedly different paradigm; instead of actively incorporating our best intuitive knowledge directly into the prediction, we are first making a purely data-driven prediction using only the available data and predictors. This prediction can later be reinterpreted as required for any given purpose. Thus, we make a clear distinction between purely data-driven and data-informed estimates.

For TOC, as a first step, we want to maximize the influence of all available Earth science observations and minimize the influence of any theories or assumptions about TOC distribution. Our intent is that our estimates and uncertainties represent only what is present in observations. Data-driven, machine learning estimates specifically lack the intuition and experience of decades of research, but they also lack potential misconceptions and unintended bias. For this reason, data-driven estimates and uncertainties are likely more faithful to the direct observations. This system for prediction is also very amenable to the addition of new observations and predictors—Updates can be made quickly with essentially no expert reinterpretation.

2. Methods and Materials

2.1. k-Nearest Neighbors Algorithm

The selection of a machine learning algorithm is generally guided by performance on a particular problem and is therefore problem specific. In this case we chose a single algorithm, k-nearest neighbors (kNN), because its simplicity allows for (1) minimal user inputs to influence the results and (2) by our methods a direct (if empirical) estimate of uncertainty. The kNN implementation used here is specific to seafloor prediction and has not been published, so we describe it here. The Python code is built upon scikit-learn (Pedregosa et al., 2011).

kNN, where "k" represents the number of nearest neighbors, uses proximity in parameter space (predictor space) as a proxy for similarity. kNN is nonparametric, making no prior assumptions about the probability

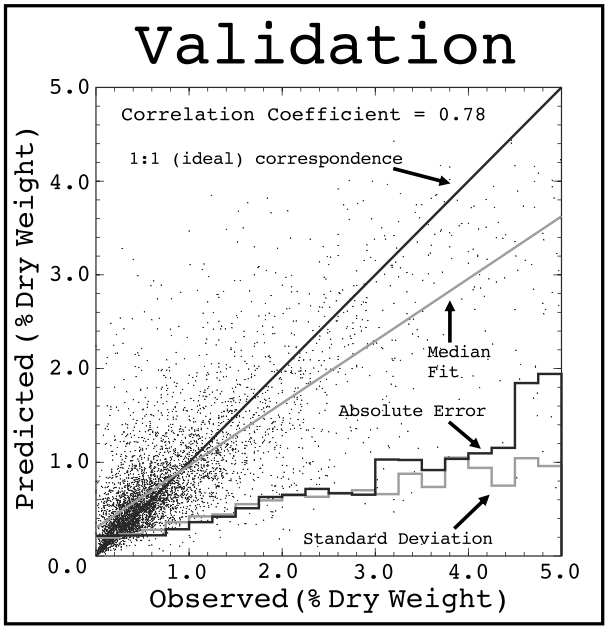

Figure 2. Validation plot of observed seafloor TOC (x) versus predicted seafloor TOC (y), both expressed in percent dry weight. The gray line is the median fit for the data. Black line is a perfect (1:1) fit to the data. Thick black and gray lines represent absolute error and standard deviation per 0.25% bin of dry weight seafloor TOC, respectively. TOC = total organic carbon.

distribution of the observed data, and is arguably the simplest machine learning algorithm. kNN is essentially a formalization of the intuitive notion that if two locations on the seafloor are similar in many ways that we have observed, then they are also similar in a way that we have not observed. The "ways we have observed" are the predictors, quantities known (or estimated) everywhere on the seafloor such as water depth, distance from shore, and bottom water temperature. The quantity we have observed in some places, but want to predict elsewhere, we refer to as the predictand. When predicting a value at a point where no observation exists, we find the k observations, where the predictor values are most similar to those at the location we are trying to predict. The value of the predictand is assumed to be the mean value of the kNN. The standard deviation of those same k observed values is our measure of uncertainty.

To quantitatively determine which are the nearest neighbors, we must first calculate a distance in parameter space (equation (1)).

$$D_j = \sqrt{\sum_{i=1}^{\mathrm{ndim}} (x_i - y_i)^2} \quad (1)$$

Distances in parameter space are calculated using L2-normalized (i.e., Euclidean) distance where x_i and y_i represent the predictor values at observed and unknown locations, respectively, and D_j is the total distance in parameter space from the jth observed datum to the point we are trying to predict. Machine learning methods, including kNN, typically use either L1-normalized or Euclidean (i.e., L2-normalized) distance, each having their own advantages. The primary advantage of L1-normalized distance is the robustness to outliers in the observed data set. Euclidean distance provides an analytical solution but is not as robust in the presence of outliers. We have conditioned the observed data set such that outliers are eliminated (section 3.1). Therefore, we have selected to use the analytic solution (i.e., Euclidean distance) as our metric of distance. Predictor values are normalized to a mean of 0 and a variance of 1. After the distances to the entire set of observed data are calculated, the distances are ranked so the k neighbors with the smallest value of D or the nearest observation can be identified. The predicted value at any given point is the average of the value of the nearest neighbors, where each neighbor is weighted by the inverse of its relative (scaled) distance.

One of the key advantages of this kind of prediction is that points that are very distant geographically may be very close in parameter space and therefore very helpful in predicting values. Also important is that kNN (categorized as a lazy or instance-based learner) is only able to predict values from within the range of the observed data; that is, it can only predict from experience, and cannot predict wild, or geologically unreasonable values. However, this means kNN also cannot predict values outside the range of sampled data.

2.2. Quantifying Neighborhood Size (k) and Predictor Selection

A key aspect to our estimates is predictor selection, which is based on validation. In this study, we use tenfold validation; withholding a random 10% of the data and using the remaining 90%, we predict the value at the withheld points. This is repeated with a different random 10%, and so on until each point has been withheld and predicted. This technique, at least for large enough data sets, mitigates the likelihood of overfitting, because one cannot over fit withheld data. A validation plot of the observed (x) versus predicted (y) TOC is shown in Figure 2.

Essentially, the only interpretive bias we cannot avoid is the choice of the number of nearest neighbors, k. Selection of k is a common limitation. In predictions where k is small, particularly k = 1, the variance per prediction is higher such that the nearest neighbor in the parameter space defines the unknown instance. In noisy data sets, where the nearest neighbor is based on data that are of poor quality, the unknown instance will result in noisy predictions. However, in predictions where a large neighborhood is used, the predictions begin to become biased (i.e., "over smoothing") by creating prediction values that approach

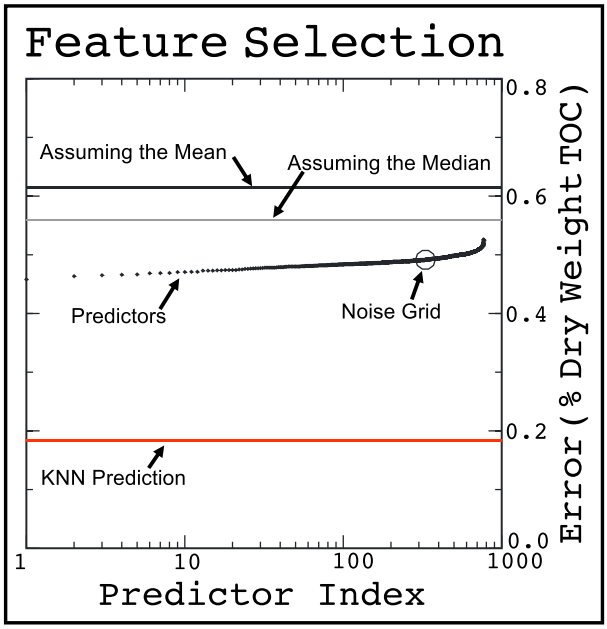

Figure 3. Predictor index versus error. Black diamonds represent individual error of predictor grids ranked from lowest to highest. The circle represents the rank and error of the uniform random noise grid. Solid black and gray lines represent the mean error if we assumed the TOC was everywhere the mean or median of all the observations, respectively. The solid red line indicates the median error in our best predicted seafloor TOC using the kNN algorithm. TOC = total organic carbon; kNN = k-nearest neighbors.

the mean of the observed data set (Zhang, 2016). As a result of these constraints, and after several trials of k ranging from 1 to 21, we chose k = 5, interpretively balancing k to be small enough to mitigate over smoothing and large enough for noise reduction.

The kNN methodology employed here is a two-step processes, the first being feature selection. Predictions are sensitive to predictor (feature) selection and as discussed earlier can be used to impart interpretive bias into the prediction, which we are actively seeking to avoid. To mitigate predictor selection bias, we use a univariant selection process in which we determine the median prediction error in the kNN validation using each predictor individually to predict the withheld observations. The errors from each predictor are then ranked (black diamonds in Figure 3), from lowest to highest.

We assume that using a predictor consisting only of uniform random noise (large circle in Figure 3) has no predictive value, and so too every predictor with an individual prediction error higher than that of the noise. Therefore, only the predictors whose individual prediction error was less than that of the noise were used in the final prediction of TOC. The final prediction uses all of the best predictors less than that of the noise grid simultaneously.

3. Data

3.1. Observed Data and Predictor Grids

Direct observations of seafloor TOC by Seiter et al. (2004) are available for public download through Pangaea (www.pangaea.de). This is not intended to represent a comprehensive inventory of all TOC data to date, but rather the single largest source of those that have been systematically accumulated, that is, amenable to machine learning prediction. For demonstration purposes this data set has been augmented to include Beazley (2003) Gulf of Mexico (GoM) data points.

Our global database of TOC observations consists of 5,623 measurements expressed in percent dry weight sediment representing the upper 5 cm of the seafloor. We chose to use only TOC data points with less than 5% dry weight to mitigate outliers in predictions, noting that only 127 points, or 2.2% of the data set, are greater than 5% dry weight TOC. Sampled observations are commonly clustered around areas of geological interest resulting in an irregularly spaced grid. For compatibility in our kNN algorithm, we group and average observations per 5 × 5-arc minute grid cell resulting in a uniformly spaced grid consisting of 4,913 cell-centered observed data points.

kNN requires geologic predictors (i.e., features) to determine correlations among observed data. We compiled as many predictors as possible, with global or almost global coverage, from a variety of widely available sources. Predictor grids lacking global coverage or at inappropriate resolutions are resampled, cell centered, and interpolated as needed using various techniques, including machine learning (e.g., extending some predictors to higher latitudes). Our final predictors are two-dimensional global gridded geologic measurements and calculated radius statistics (mean, natural log of the mean, absolute average deviations at each grid cell over radii of 1,000, 500, 250, 125, 50, and 10 km) at 5 × 5-arc degree resolution. At present, our predictor database consists of more than 600 global grids at 5 × 5-arc minute resolution available for public download at https://doi.org/10.5281/zenodo.1471638 (Lee et al., 2018).

4. Results and Discussion

4.1. Validation and Uncertainty

Using the above methods with 5 nearest neighbors and 397 feature selected predictors at a 5 × 5-arc minute resolution results in a predicted global distribution of seafloor TOC (Figure 4). We used tenfold cross validation to calculate a median prediction error of 0.18% dry weight (red line in Figure 3). For comparative

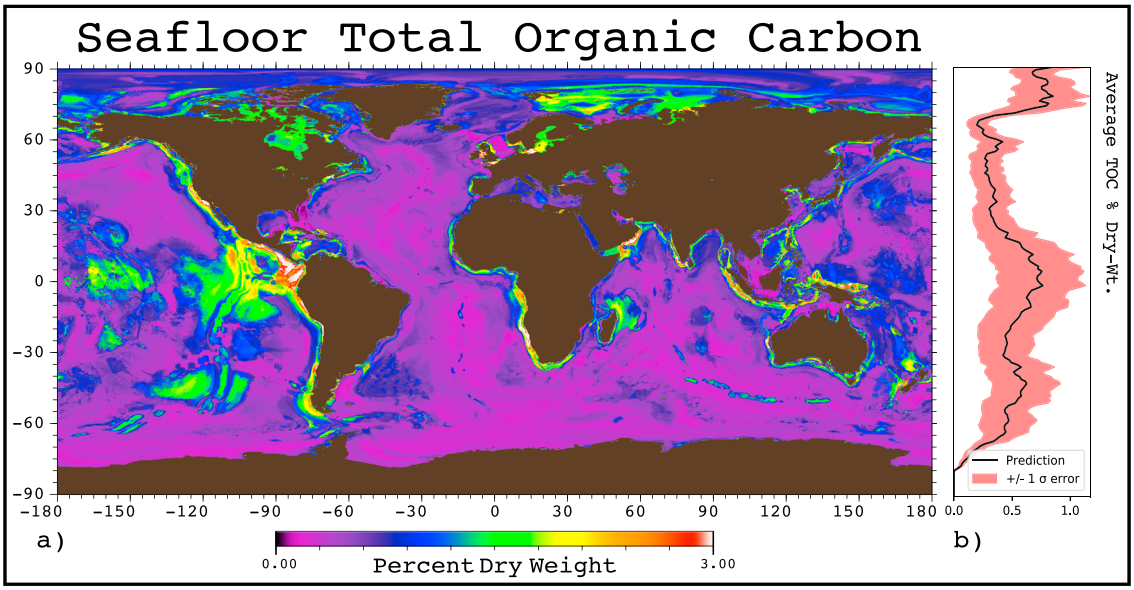

Figure 4. (a) Seafloor TOC prediction produced using a k-nearest neighbor algorithm with 397 predictors and 5 nearest neighbors. (b) The average TOC content per latitude ±1 standard deviation (shaded red). TOC = total organic carbon.

purposes, a prediction where all unknown grid cells are assigned the mean of the observed data set (0.88% dry weight) results in a prediction error of 0.61% dry weight (thin black line in Figure 3). The validation plot for our final prediction is shown in Figure 2, with the 1:1 correspondence (perfect correlation and perfect predictive skill) and a median fit of our predictions. The median fit shows strong positive correlation between our predicted (x) and observed (y) with a correlation coefficient (r) of 0.78.

Additionally, for direct comparison we interpolate data points of seafloor TOC using the Generic Mapping Tools (Smith & Wessel, 1990) surface function with a tension of 0.25 (Figure 5a). We perform tenfold cross validation (Figure 5b) and determined a correlation coefficient of 0.39. This r value provides quantitative measure that kNN (correlation: 0.78) performs statistically better than a standard interpolation procedure based solely on geographic location.

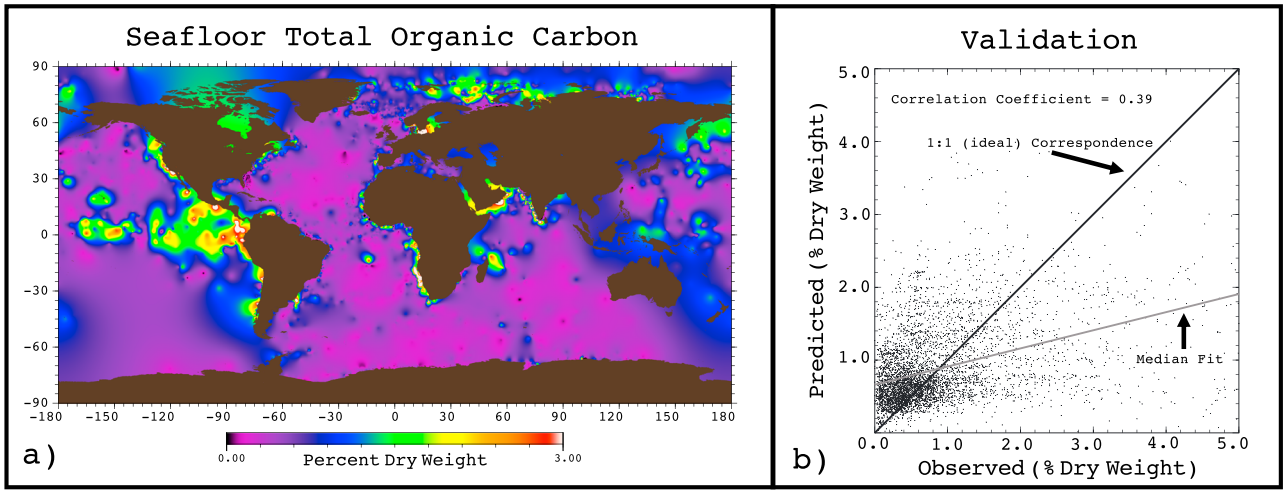

Figure 5. (a) Seafloor total organic carbon prediction from GMT surface command with tension of 0.25. (b) Tenfold cross validation of seafloor total organic carbon using GMT surface command with tension of 0.25.

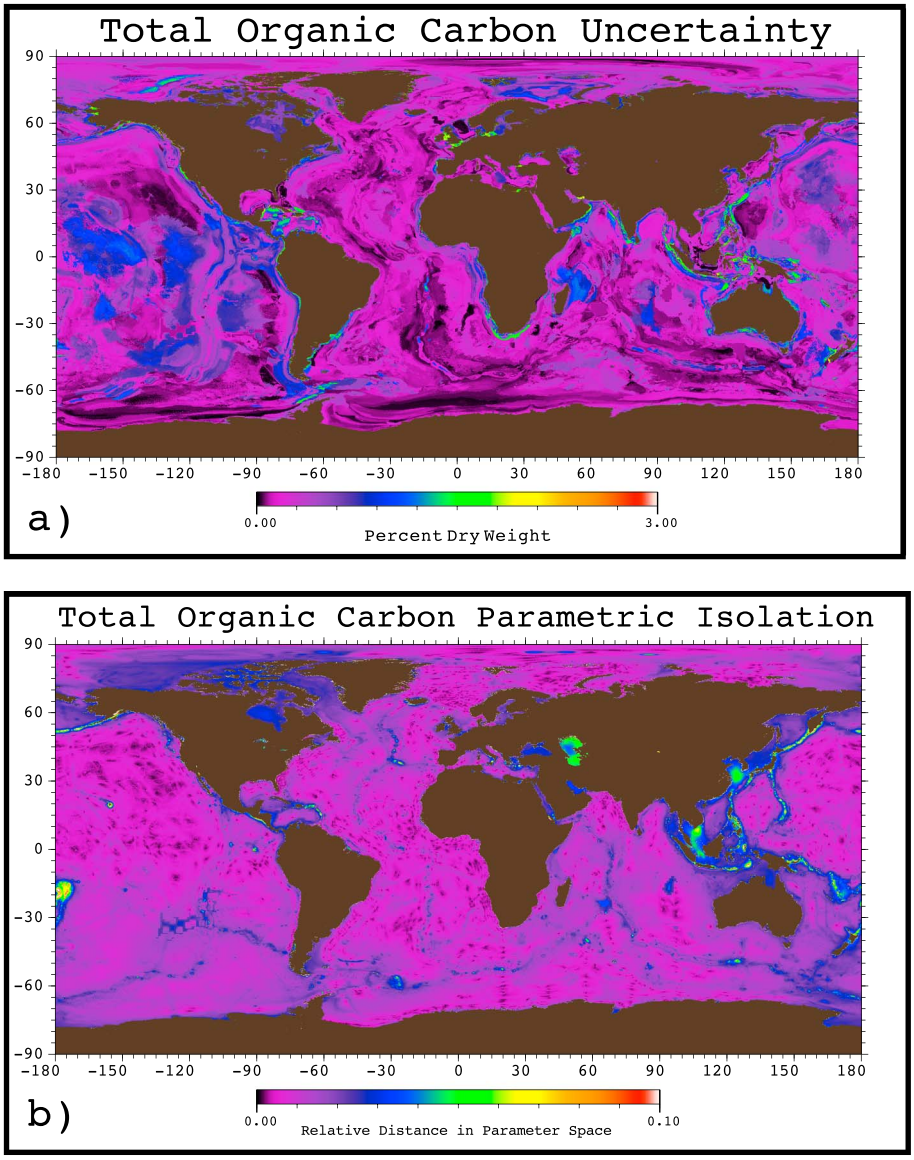

Figure 6. (a) Standard deviation, which we assume is a proxy for uncertainty, per final seafloor total organic carbon prediction; (b) parametric isolation per final seafloor total organic carbon prediction, measuring relative distance to single nearest neighbor in parameter space.

Predictions such as those presented here are far more useful when accompanied by an uncertainty, but to our knowledge there is no standard means of estimating uncertainty from a kNN prediction. Instead, we have used, as a proxy for uncertainty, the standard deviation of the values of the nearest neighbors (Wood et al., 2018). Figure 2 shows the average standard deviation (heavy gray line) and average absolute error (heavy black line) averaged over bins of width 0.25% dry weight. At lower values of seafloor TOC, binned absolute error very closely mimics binned standard deviation, suggesting that standard deviation may serve as a reasonable proxy for prediction uncertainty.

The parallelism between absolute error and standard deviation makes sense intuitively. If all the nearest neighbors in parameter space came from the same location, the standard deviation of the TOC values would yield the best estimate of uncertainty. In the case of our implementation of kNN prediction, we have simply lifted the requirement that the nearest neighbors are geographically colocated—It is the geological similarity, not the geographic similarity that drives the prediction. While our metric of uncertainty makes intuitive sense, it lacks a proof of theory, and more analysis is required to fully justify it mathematically. Figure 6a shows a global distribution of standard deviation in our TOC prediction.

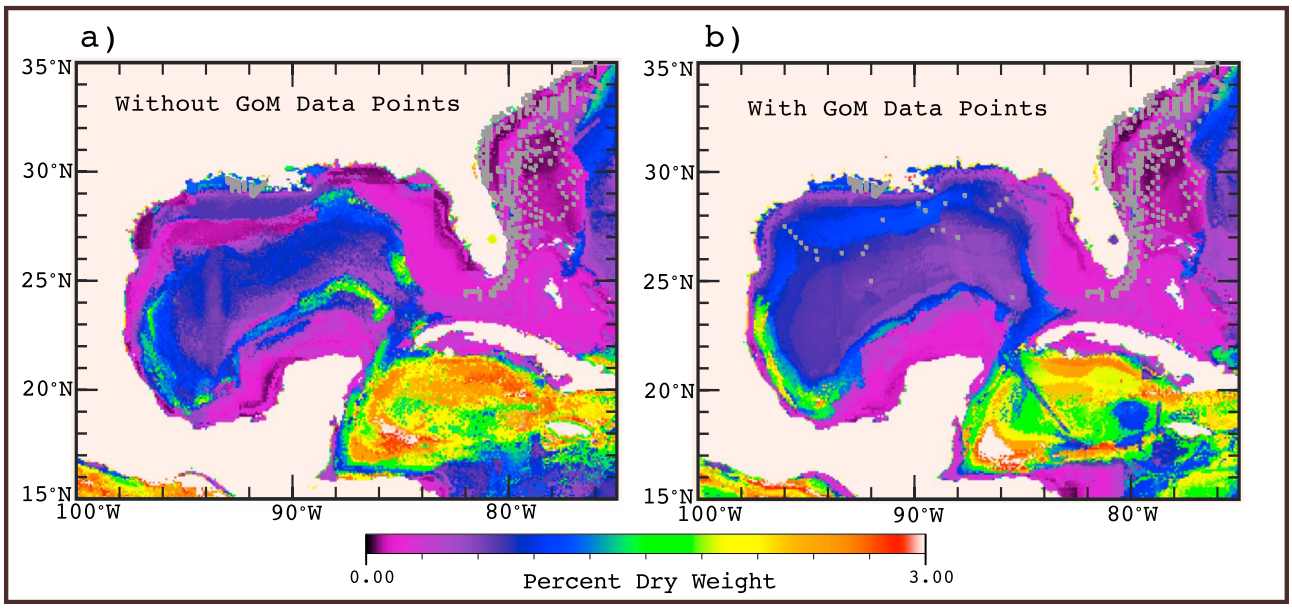

Figure 7. (a) The GoM seafloor total organic carbon upon removal of 23 data points (Beazley, 2003); (b) the GoM seafloor total organic carbon prediction using 23 data points from Beazley (2003). GoM = Gulf of Mexico.

Qualitatively, standard deviations (i.e., uncertainty) reflect the density and variability of observed data in parameter space. Low uncertainty is attributed to high data density (many near neighbors) and low variability (neighbors are similar). This will only happen in a well-sampled portion of parameter space. Likewise, higher uncertainty is attributed to low observed data density (neighbors are sparse in parameter space) and/or highly variable, indicative of poorly sampled parameter space. We therefore expect reductions in standard deviation with denser sampling of observed data in parameter space (not necessarily geographic space).

With a strictly data-driven approach, new observations can very quickly provide updated predictions. For example, we remove the 23 observed data points from Beazley (2003) in the GoM and perform the same kNN prediction using the same predictor grids. Figures 7a and 7b show the difference made by removing these points. The change in global error was negligible, but the 23 points have noticeably changed the prediction in the GoM.

Predictor grids can be similarly updated. Adding a new and/or updated predictor grid may result in a different feature selection, while subsequently updating the final value of the TOC prediction.

4.2. Parametric Isolation

kNN is an instance-based learner; therefore, predictions are most accurate where parameter (i.e., predictor) space is well sampled by observations. Here we define parametric isolation as the distance in parameter space to the single nearest neighbor. A plot of parametric isolation for our TOC prediction is shown in Figure 6b. Parametric isolation indicates how well parameter space is sampled by observations and therefore identifies the best place to acquire new data (geographically, warm colors in Figure 6b) to improve sampling in parameter space (geologically); that is, a guide indicating where to sample next to best improve the prediction.

Areas where observations exist (black spots in Figure 6b) exhibit the lowest parametric isolation because they are the closest (i.e., most alike) in parameter space (as well as geographic space) to their nearest neighbor, indicating a well-sampled parameter space. Areas with cooler colors, but geographically distant from black spots in Figure 6b, indicate areas that are geologically similar to areas where we already have samples, that is, also well sampled in parameter space.

Points that are furthest in parameter space (warm colors in Figure 6b) represent values that are least like any observed data point. Areas of high parametric isolation, namely, data-deficient areas (for the data set we

used this includes the Arctic Ocean and seas of the Western Pacific) indicate locations where acquiring more samples would most likely significantly improve the prediction (i.e., these areas are poorly sampled in parameter space). Data acquisition at locations with high parametric isolation may improve predictions, not only for that geographic area but also for areas geologically similar (i.e., similar parameter values).

It is possible to have high uncertainty (i.e., large spread in nearest neighbors) but still have minimal parametric isolation (i.e., one nearby neighbor). A high uncertainity would mean there is large spread in the nearest neighbors; a small parametric isolation would indicate at least one nearest neighbor in parameter space is capable of explaining the geological envrionment predicted. All this considered, uncertainity (i.e., standard deviation) and parametric isolation cannot be directly compared since they are not explicitly correlated. Uncertainity can, however, be directly related to the TOC prediction.

4.3. Comparison to Non-Data-Driven Results

Our final data-driven prediction is geologically consistent with previous analyses (e.g., Seiter et al., 2004), but our analysis provides estimates of seafloor TOC in regions where there were previously large data deficiencies (e.g., Western Pacific Ocean and portions of the Arctic). Additionally, we average seafloor TOC (Figure 4a) and uncertainty ($\pm$1 standard deviation; Figure 4b) in percent dry weight per latitude. Our estimate of seafloor TOC yields similarities in areas where an expected high accumulation of TOC is likely (e.g., equatorial latitudes and continental shelf). In particular, along the equator upwelling of nutrient-rich waters results in increased primary productivity and thus particulates organic matter, thereby increasing sediment organic matter (Archer et al., 1997). High latitudes exhibit higher average concentrations of TOC, which is likely attributed to ocean current transport of nutrients and terrestrial output from nearby rivers (Birgel et al., 2004).

In general, the mechanisms that result in deposition of seafloor TOC are reasonably well understood scientifically (Arndt et al., 2013). Therefore, we have included an additional prediction of seafloor TOC using only predictors and their associated statistics, which are deterministically known to influence the deposition of TOC. Grids (i.e., predictors) selected as controls on the accumulation of seafloor TOC include river mouth ocean fluxes (e.g., total suspended sediment, particulate organic carbon, and bicarbonate), distance from coast, bathymetry, biomass (Wei et al., 2010), porosity (as a grain size proxy), and bottom water currents (magnitude and direction). The deterministic prediction of seafloor TOC, available in the supporting information, used the five nearest neighbors and in total 146 predictors (listed in the supporting information). Many of the 146 predictors were based on statistics (e.g., mean or deviation of values within a given radius of each grid point) applied to the more fundamental quantity.

As a result, the median prediction error using these well-established deterministic predictors is 0.19% dry weight with a correlation coefficient of 0.76 between observed and predicted values. Quantitatively, minimal differences exist between our data-driven approach and the deterministic prediction where the average residual is 0.24% dry weight TOC. Qualitatively, the deterministic prediction is more geographically variable (i.e., speckled), while the data-driven approach results in a smoother prediction. Similarly, in the data-driven approach, predicted TOC is high in equatorial regions and along the continental shelf. The prediction error and correlation coefficient of the deterministic prediction using tenfold cross validation are consistent with our purely data-driven approach and, more importantly, a general understanding on the processes of seafloor TOC accumulation. We find no significant benefit on the use of more deterministic grid over our data-driven prediction. The prediction and standard deviation grid are available in the supporting information.

4.4. Global Inventory

Collectively ocean sediments are known to be one of, if not the largest, pool of global carbon. We estimate the total amount of organic carbon stored in the upper 5 cm of seafloor sediments by calculating the volume of sediment for the upper 5 cm of seafloor sediments based on our 5 $\times$ 5 arc minute prediction. Since TOC is expressed a function of dry weight, we use a porosity grid (see the supporting information; Martin et al., 2015) to account for only the solid portion of sediment volume. Assuming the dry density of sediment to be 2.65 g/cm^3, we calculate the dry mass of the sediment in each grid cell, and then the corresponding mass of carbon (TOC is given in percent dry weight). Integrating over the entire seafloor yields an estimated global

inventory of 87 gigatones of organic carbon (Gt C) held within upper 5 cm of seafloor sediments. Similarly, we integrate uncertainty (i.e., standard deviation) to be 43 Gt C.

By comparison, the carbon stored in the atmosphere is approximately 867 Gt C, assuming ~410 ppm CO_2 (IPCC, 2013) whereby one ppm CO_2 by atmospheric volume is approximately 2.13 Gt C. The Intergovernmental Panel on Climate Change (IPCC, 2013) estimates that fossil fuel and cement production releases approximately ~8 Gt C annually. Therefore, only the top 5 cm of the seafloor represents approximately 10% as much carbon as is sequestered in the atmosphere and more than 10 times that which is released into the atmosphere annually. Quantifying the carbon withheld in the marine sediments is fundamental to other carbon cycling calculations as each source/sink system are interrelated.

5. Conclusion

Accurate estimations of seafloor TOC are required for a wide variety of modeling applications and estimation of global carbon inventories. Previous geospatial predictions based on sparse data relied on strictly geospatial interpolation, provided no associated uncertainties, and failed to predict large portions of the global seafloor. Our application of kNN machine learning algorithm with univariant predictor selection results in a geologically consistent, easily updatable data-driven prediction of TOC at every point on the global seafloor. Standard deviation serves as a proxy for absolute error performed via tenfold cross validation allowing for uncertainty estimation at each predicted location.

A by-product of this kind of machine learning is parametric isolation—the distance in parameter space between the prediction value and the single nearest neighbor. This metric indicates geographic locations that are dissimilar in parameter space (i.e., geologically) to any other observed point, thereby indicating which locations are most advantageous to sample for more accurate predictions. It is effectively a guide to where to sample next.

Potential uses of a geospatial prediction of seafloor TOC are wide ranging, including furthering global and regional modeling efforts. One use of this prediction is making a data-driven estimate (with uncertainty) of the global inventory of organic carbon stored in the upper 5 cm of the seafloor, namely, 87 $\pm$ 43 Gt.

Acknowledgments
This research was funded by the U.S. Naval Research Laboratory base program. Data in this study are all publicly available and are included in the supporting information and online at https://doi.org/10.5281/zenodo.1471638. We thank the editor, Peter Raymond, and two anonymous reviewers for their thoughtful reviews, which have improved the quality of the analysis and manuscript.

References

Archer, D. E., Peltzer, E. T., & Kirchman, D. L. (1997). A timescale for dissolved organic carbon production in equatorial Pacific surface waters. *Global Biogeochemical Cycles*, *11*(3), 435–452. https://doi.org/10.1029/97GB01196

Arndt, D., Jørgensen, B. B., LaRowe, D. E., Middelburg, J. J., Pancost, R. D., & Regnier, P. (2013). Quantifying the degradation of organic matter in marine sediments: A review and synthesis. *Earth-Science Reviews*, *123*, 53–86. https://doi.org/10.1016/j.earscirev.2013.02.008

Beazley, M. J. (2003). The significance of organic carbon and sediment surface area to the benthic biogeochemistry of the slope and deep water environments of the Northern Gulf of Mexico, (master's thesis). Retrieved from OAKTrust. (http://hdl.handle.net/1969.1/534). College Station, TX: Texas A&M University.

Birgel, D., Stein, R., & Hefter, J. (2004). Aliphatic lipids in recent sediments of the Fram Strait/Yermak Plateau (Arctic Ocean): Composition, sources and transport processes. *Marine Chemistry*, *88*(3–4), 127–160. https://doi.org/10.1016/j.marchem.2004.03.006

Burdige, D. J. (2007). Preservation of organic matter in marine sediments: Controls, mechanisms, and an imbalance in sediment organic carbon budgets? *Chemical Reviews*, *107*(2), 467–485. https://doi.org/10.1021/cr050347q

Colwell, F. S., Boyd, S., Delwiche, M. E., Reed, D. W., Phelps, T. J., & Newby, D. T. (2008). Estimates of biogenic methane production rates in deep marine sediments at Hydrate Ridge, Cascadia Margin. *Applied and Environmental Microbiology*, *74*(11), 3444–3452. https://doi.org/10.1128/AEM.02114-07

Dutkiewicz, A., Muller, R. D., & O'Callaghan, S. (2015). Census of seafloor sediments in the world's ocean. *Geology*, *43*(9), 795–798. https://doi.org/10.1130/G36883.1

Goutorbe, B., Poort, J., Lucazeau, F., & Raillard, S. (2011). Global heat flow trends resolved from multiple geological and geophysical proxies. *Geophysical Journal International*, *187*(3), 1405–1419. https://doi.org/10.1111/j.1365-246X.2011.05228.x

Hovland, A., & Judd, M. (2007). Migration and seabed features. In *Seabed fluid flow* (pp. 189–247). New York: Cambridge University Press.

IPCC (2013). *Climate change 2013: The physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, UK: Cambridge University Press.

Lee, T. R., Wood, W. T., & Phrampus, B. J. (2018). Global derived datasets for use in k-NN machine learning prediction of global seafloor total organic carbon (Version 1) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.1471639

Levin, L. A., Baco, A. R., Bowden, D. A., Colaco, A., Cordes, E. E., Cunha, M. R., et al. (2016). Hydrothermal vents and methane seeps: Rethinking the sphere of influence. *Frontiers in Marine Science*, *3*. https://doi.org/10.3389/fmars.2016.00072

Martin, K. M., Wood, W. T., & Becker, J. J. (2015). A global prediction of seafloor sediment porosity using machine learning. *Geophysical Research Letters*, *42*, 10,640–10,646. https://doi.org/10.1002/2015GL065279

Max, M. D., Johnson, A. H., & Dillon, W. P. (2006). Why gas hydrate? In M. D. Max, A. H. Johnson, & W. P. Dillon (Eds.), *Economic geology of natural gas hydrates, coastal systems and continental margins* (Vol. 9, pp. 17–44). Dordrecht, Netherlands: Springer.

Middelburg, J. J. (1989). A simple rate model for organic matter decomposition in marine sediments. *Geochimica et Cosmochimica Acta*, *53*(7), 1577–1581. https://doi.org/10.1016/0016-7037(89)90239-1

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pinero, E., Marquardt, M., Hensen, C., Haeckel, M., & Wallmann, K. (2013). Estimation of global inventory of methane hydrates in marine sediments using transfer functions. *Biogeosciences*, *10*(2), 959–975. https://doi.org/10.5194/bg-10-959-2013

Rothman, D. H., & Forney, D. C. (2007). Physical model for the decay and preservation of marine organic carbon. *Science*, *316*(5829), 1325–1328. https://doi.org/10.1126/science.1138211

Ruppel, C. D., & Kessler, J. D. (2017). The interaction of climate change and methane hydrates: Climate-hydrates interactions. *Reviews of Geophysics*, *55*, 126–168. https://doi.org/10.1002/2016RG000534

Schoepfer, S. D., Shen, J., Wei, H., Tyson, R. V., Ingall, E., & Algeo, T. J. (2015). Total organic carbon, organic phosphorus, and biogenic barium fluxes as proxies for paleomarine productivity. *Earth-Science Reviews*, *149*, 23–52. https://doi.org/10.1016/j.earscirev.2014.08.017

Seiter, K., Hensen, C., Schröter, J., & Zabel, M. (2004). Organic carbon content in surface sediments—Defining regional provinces. *Deep Sea Research Part I: Oceanographic Research Papers*, *51*(12), 2001–2026. https://doi.org/10.1016/j.dsr.2004.06.014

Sloan, E. D. (2003). Fundamental principles and applications of natural gas hydrates. *Nature*, *426*(6964), 353–359. https://doi.org/10.1038/nature02135

Smith, W. H. F., & Wessel, P. (1990). Gridding with continuous curvature splines in tension. *Geophysics*, *55*(3), 293–305. https://doi.org/10.1190/1.1442837

Wei, C.-L., Rowe, G. T., Escobar-Briones, E., Boetius, A., Soltwedel, T., Caley, M. J., et al. (2010). Global patterns and predictions of seafloor biomass using random forests. *PLoS One*, *5*(12), e15323. https://doi.org/10.1371/journal.pone.0015323

Wood, W. T., Runyan, T. R., & Obelcz, J. (2018). Practical quantification of uncertainty in seabed property prediction using geospatial KNN machine learning, Abstract EGU2018–9760 presented at 2018 General Assembly, EGU, Vienna, Austria.

Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. *Annals of Translational Medicine*, *4*(11), 218. https://doi.org/10.21037/atm.2016.03.37