## Title

Application of clustering techniques to study environmental characteristics of microbialite-bearing aquatic systems



## Journal

Biogeosciences Discussions, 12(13)

## Authors

Dalinina, R
Petryshyn, VA
Lim, DS
et al.

## Publication Date

2015-07-09

## DOI

10.5194/bgd-12-10511-2015

Peer reviewed

Biogeosciences Discuss., 12, 10511–10544, 2015
www.biogeosciences-discuss.net/12/10511/2015/
doi:10.5194/bgd-12-10511-2015


Biogeosciences Discussions
Open Access



# Application of clustering techniques to study environmental characteristics of microbialite-bearing aquatic systems

**R. Dalinina[1], V. A. Petryshyn[2,7], D. S. Lim[3], A. J. Braverman[1,4], and A. K. Tripati[2,3,5,6,7]**

[1]Center for Applied Statistics, University of California, Los Angeles, CA, USA
[2]Department of Earth, Planetary, and Space Sciences, University of California, Los Angeles, CA, USA
[3]Institute of Planets and Exoplanets, NASA Ames Research Center, Moffet Field, CA, USA
[4]Jet Propulsion Laboratory, Pasadena, CA, USA
[5]Department of Atmospheric and Oceanic Sciences, University of California, Los Angeles, CA, USA
[6]Institute of the Environment and Sustainability, University of California, Los Angeles, CA, USA
[7]European Institute of Marine Sciences (IUEM), Université de Brest, UMR6538, Domaines Océaniques, Rue Dumont D'Urville, Plouzané, France



Received: 05 May 2015 – Accepted: 09 June 2015 – Published: 09 July 2015

Correspondence to: V. A. Petryshyn (vpetryshyn@ucla.edu)

Published by Copernicus Publications on behalf of the European Geosciences Union.




Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |

**Abstract**

Microbialites are a product of trapping and binding of sediment by microbial communities, and are considered to be some of the most ancient records of life on Earth. It is a commonly held belief that microbialites are limited to extreme, hypersaline settings. However, more recent studies report their occurrence in a wider range of environments. The goal of this study is to explore whether microbialite-bearing sites share common geochemical properties. We apply statistical techniques to distinguish any common traits in these environments. These techniques ultimately could be used to address questions of microbialite distribution: are microbialites restricted to environments with specific characteristics; or are they more broadly distributed? A dataset containing hydrographic characteristics of several microbialite sites with data on pH, conductivity, alkalinity, and concentrations of several major anions and cations was constructed from previously published studies. In order to group the water samples by their natural similarities and differences, a clustering approach was chosen for analysis. *k* means clustering with partial distances was applied to the dataset with missing values, and separated the data into two clusters. One of the clusters is formed by samples from atoll Kiritimati (central Pacific Ocean), and the second cluster contains all other observations. Using these two clusters, the missing values were imputed by *k* nearest neighbor method, producing a complete dataset that can be used for further multivariate analysis. Salinity is not found to be an important variable defining clustering, and although pH defines clustering in this dataset, it is not an important variable for microbialite formation. Clustering and imputation procedures outlined here can be applied to an expanded dataset on microbialite characteristics in order to determine properties associated with microbialite-containing environments.



## 1 Introduction

Microbialites, organo-sedimentary deposits formed by the trapping and binding of sediment by benthic microbial communities, or by microbially-induced precipitation of minerals (Burne and Moore, 1987), constitute a class of lithified structures that form as a result of microbial activities. Found in the rock record, microbialites (stromatolites) are thought to be the earliest evidence for life on Earth (Walter, 1976; Semikhatov et al., 1979; Hofmann et al., 1999; Riding and Awramik, 2000; Allwood et al., 2006). Despite their high profile in the geobiologic community, much about microbialites is enigmatic. Although they are perceived by some as being limited to harsh environments (i.e. high salinity which may exclude certain types of metazoa), a survey of the literature shows they frequently occur in aquatic systems that do not exhibit such extreme properties (e.g. Grotzinger and Knoll, 1999; Lim et al., 2009; Petryshyn et al., 2012) If salinity is not the controlling factor in microbialite distribution, what is? Is there a characteristic set of geochemical properties common to sites where microbialites are actively forming?

In order to discern whether clustering techniques can be used to address these questions, we have created a pilot dataset from published studies on bodies of water containing actively accreting microbialites. Each published study we used focused on a specific set of geochemical or physical characteristics of an aquatic environment, such as the concentration of certain chemical species in the water. Combining data from each publication led to the overall dataset, with variables consisting of geochemical measurements and observations from various bodies of water (hereafter referred to as sites). The resulting dataset has numerous "holes" due to differences in geochemical collection and measurement strategies. Disregarding incomplete observations from data analysis would ignore valuable information; a more preferable strategy is to impute the missing values. We therefore focus on clustering methods, which can reveal any natural structure or similarities between the sites and can also be used to impute missing values. Once clusters are created, missing items can be estimated based on the characteristics of the cluster to which they belong. We use this clustering tech-

nique to examine the role of salinity and pH in controlling microbialite formation. As described above, salinity is thought to be an important variable for microbialite growth. Additionally, pH is known to be relevant for carbonate mineral precipitation and can also influence the ecology/microbial population of an environment.

## 2 Locality information

### 2.1 Microbialite-forming sites

In order to develop a method that could be used in the future for a more detailed analysis, we selected sites with the most complete geochemical datasets or observations (i.e., sites that are very well characterized geochemically). Microbialite-bearing sites were chosen from three different environments for this investigation: Pavilion Lake (freshwater lacustrine), atoll Kiritimati (hypersaline marine), and German karst streams (alkaline riverine). Also, as a control, samples from Stinking Springs, Mono Lake, and seawater were added. The variable set was restricted to 12 most commonly measured features: concentrations of $Ca^{2+}$, $Mg^{2+}$, $SO_4^{2-}$, $K^+$, $Na^+$, Si, $Cl^-$, $Sr^{2+}$, $Ba^{2+}$, as well as pH, conductivity, and alkalinity. All sites are reported to have actively precipitating microbial carbonates. The lack of extensive hydrographic data for other lakes precluded their being used in this study.

#### 2.1.1 Pavilion Lake

Pavilion Lake is a groundwater fed freshwater lake in British Columbia, located 450 km northeast of Vancouver. The lake has an area of roughly 5.7 km × 0.8 km, and a mean pH of 8.3, (Lim et al., 2009; Brady et al., 2010). Several different morphologies of microbialites are found at a variety of depths in Pavilion Lake, ranging from 5 to 55 m depth (Laval et al., 2000; Lim et al., 2009, 2011; Petryshyn et al., 2015). Microbialites range from several centimeters to meters in height, and consist mainly of fine-grained carbonate (micrite) (Fig. 1a and b). Pavilion Lake microbialites are thought accrete



predominantly via in situ carbonate precipitation, with little to no input coming from the trapping and binding of detrital carbonate (Laval et al., 2000), indicating that these microbialites should both be influenced by and record the geochemical conditions of the lake. Extensive geochemical characterization of the lake has been carried out by the Pavilion Lake Research Project (PLRP).

#### 2.1.2 Atoll Kiritimati

Microbial carbonates were discovered to be actively accreting in several lakes on the atoll Kiritimati, an equatorial raised coral atoll located in the Central Pacific Ocean (formerly called Christmas Island; Arp et al., 2011). The atoll, the largest in the world by land area (Valencia, 1977), is part of the Republic of Kiribati, and is located in the equatorial dry belt (1°55′ N, 157°25′ W). Several lakes ranging from nearly fresh to hypersaline occupy the inner atoll (Saenger et al., 2006), though for this study we are focusing on those sampled by Arp et al. (2011).

Most of the microbialites in atoll Kiritimati accrete in hypersaline and hydrologically closed lakes which are cut off from seawater reflux, but are susceptible to spillover during periods of high sea level (Arp et al., 2011). The thicknesses of the microbialites vary. All are typically found underneath 3–5 cm thick microbial mats in shallow (20–50 cm depth) water. Near shore, microbialites are thin (< 1 cm), but they can grow to ∼ 50 cm offshore (20 m from the water line). Microbialites are often interbedded with 2–5 cm thick gypsum layers, which are most prominent in the middle of the structure (Arp et al., 2011). The structures are made up of 1–4 cm thick layers of vertical aragonite sheets and appear to form within the exopolymeric matrix of the microbial mats (Fig. 1d). While the lakes are supersaturated with respect to aragonite, only trace amounts of the mineral precipitate on top of the microbial mats. Instead, the bulk of the microbialites form lower in the mat, at the transition to the anoxic zone. This is quite a different formation environment from the other environments characterized in this study.

### 2.1.3 German Karst Streams

Four German streams consisting of alkaline karst water were also included in the data set (see Figs. 1 and 2 in Arp et al., 2010):

1. The Westerhöfer Bach, located west of the Harz Mountains, ~ 27 km northeast of Göttingen. The stream is less than 2 m wide, and is fed by a spring discharging from an aquifer in the Middle-Triassic Muschelkalk Group.
2. The Deinschwanger Bach on the western maring of the Franconian Alb (roughly 30 km southeast of Nürnberg). Its main springs discharge from the base of the Weißjura Group aquifer (Upper Jurassic limestones underlain by clays of the Middle Jurassic Ornatenton Formation). The maximum width of the stream is two meters.
3. The Reinsgraben, near the eastern margin of Göttingen. The stream is similar to the Westerhöfer Bach, in that it is also fed from the Middle Triassic Muschelkalk Group aquifer.
4. The Steinerne Rinne, 1.3 km south of Erasbach, southern Franconian Alb. The stream is fed by the Weißjura Group aquifer, like the Deinschwanger Bach.

Each stream is home to a diverse community of cyanobacteria and eukaryotes (mostly diatoms), and is actively precipitating "tufa stromatolites" (laminated microbialites). The spring waters feeding these streams all have a higher $pCO_2$ than the atmosphere, and thus rapidly degas when leaving the spring site. This results in a rapid rise in pH, as well as an increase in carbonate ion activity in the water, leading to calcium carbonate supersaturation. However, $CaCO_3$ does not readily precipitate in the streams until a ten-fold supersaturation is reached. Photosynthetic bacteria in these streams aid in the precipitation of calcium carbonate by either providing an organic template on which to nucleate, or locally increasing alkalinity to the point where precipitation proceedes on its own (Arp et al., 2010).



"Tufa stromatolites" vary from site to site, but generally have a pattern of alternating porous and dense lamination. This dense/porous couplet is thought to represent one year of deposition, with porous laminae being deposited in the winter/spring, and the dense laminae accreting in the summer/autumn (Arp et al., 2010). The pairs of laminations vary in thickness between sites. The lamination couplets range from 1.6–5.4 mm in the Westerhöfer Bach, and 3.9–7.6 mm thick in the Deinschwanger Bach. In all cases, the porous layers are thinner than their dense counterparts, and included organic and quartz detritus. The lamination contacts from porous to dense are gradational rather than sharp, while contacts from dense to porous are unstable and were often broken during analysis (Fig. 1c, Fig. 13 in Arp et al., 2010).

## 2.2 Non-microbialite-forming sites

In addition to microbialite-forming sites, several "outgroup" sites are included in the data set: a warm saline spring that has microbial carbonate, but no microbialites; a highly alkaline lake with inorganic tufa towers, and, as a control, data for average seawater.

### 2.2.1 Warm saline springs – Stinking Springs, Utah, USA

The Stinking Springs is a warm (~ 48 °C), sulfur-rich saline bicarbonate spring in Boxelder County, Utah, USA (Bonny and Jones, 2007). There is microbially-mediated carbonate production in Stinking Springs, however it all takes place within the ubiquitous bacterial mats that line the spring channels, and true microbialites are not formed. Geochemical data was taken from Bonny and Jones (2007). Additional data was collected by the International Geobiology Course in 2012 and 2013 (Metzger et al., 2013; Monteverde et al., 2013).

### 2.2.2 Highly alkaline lake – Mono Lake, California, USA

Mono Lake is a highly alkaline, closed basin lake located east of the Sierra Nevada mountain range in central California. The lake is fed by groundwater and freshwater

streams. Near the shore, calcium carbonate (calcite and aragonite) precipitates where calcium-rich groundwater seeps in an mixes with the high-pH, $CO_3^{2-}$ rich lake water (Nielson and DePaolo, 2013), forming large tufa towers that grow upward from the lake bottom. While Mono Lake is known to be a microbially-rich environment, the formation of its extensive tufa towers is understood to be a purely inorganic process (Dunn, 1953; Scholl and Taft, 1964). The lake itself is highly oversaturated with respect to calcite (Saturation index > 20), most likely due to carbonate inhibition by high phosphate content (Bischoff et al., 1993).

#### 2.2.3 Average seawater

Aside from places such as Shark Bay in Australia and the Bahamian bank, microbialites are not reported to form in modern marine environments. In order to have a control, we have chosen to include the values for average open ocean water in our dataset.

## 3 Statistical analysis

By statistically comparing the geochemistry of the above sites, some of which form microbialites, some of which form microbial carbonate (but not microbialites), some of which have inorganic carbonate precipitation, and a control group, we aim to discern which characteristics are the most important to the formation of microbialites. We have specifically chosen these sites because they are all characterized extensively and represent a wide variety of depositional environments.

However, there is not perfect overlap in the geochemical parameters measured (i.e., not all parameters were measured at all of the sites), and therefore there are some gaps in the data or holes in our initial data set (Table 1). Traditional *k* means clustering method cannot be directly applied to dataset with missing values. In order to account for this, a different approach, suggested by Himmelspach and Conrad (2010), is applied here. The authors modified a *k* means algorithm to use a partial distance instead



of a traditional Euclidean distance measure. This requires little modification to the algorithm, and allows us to use all available information from both complete and partially observed items.

Simiarly, we apply *k* means clustering based on partial distance to our dataset which allows us to not omit observations with missing values. This technique has not yet been used on biological/chemical systems, as far as we are aware. A clustering approach is appropriate because it can aid in the identification of any common traits among the microbialite sites, and it serves as the basis for imputation or filling in missing geochemical parameters at a site. Here, missing values are imputed by the mean of their neighbors in a cluster.

### 3.1 Exploratory analysis

First, to visualize any obvious patterns in the data, some exploratory plots were made. All variables were standardized for comparison (see Table 1). To visualize patterns within sites for each variable, a parallel coordinates plot was generated (Fig. 2). The horizontal axis contains all 12 variables (each variable with its own vertical axis).

A few observations are immediately apparent. First, Kiritimati appears to stand apart from the rest of the sites. Within Kiritimati, sample measurements vary greatly, almost creating two separate clusters – one with high pH, and high concentrations of Mg, Ca, a high alkalinity, and one with low values for these parameters. The rest of the sites tend to follow more or less the same pattern, with the exception of Pavilion Lake having much lower alkalinity. Also, most of the variance in Pavilion Lake can be concentrated in two variables – in Si and Ba concentrations (Fig. 2).

At non-microbialite forming sites, some measurements also stand out from the rest of the observations. As expected, seawater (Fig. 2, light blue) differs from the rest in several parameters: pH, Ca, Ba, and K concentrations. Mono Lake (orange) measurements appear to be significantly different in most variables as well. Finally, Stinking Springs (red) does not appear to differ as much from the microbialite-bearing sites as the other two sites.

A different way to potentially assess similarity between observations is a dendrogram (a visualization of hierarchical agglomerative clustering; Fig. 3). Each observation starts out as its own cluster, and then is joined to its closest neighbor based on some distance measure. From Fig. 3, it is obvious that Pavilion Lake tends to separate from other sites by forming a cluster unto itself; German streams combine into another group, along with samples from atoll Kiritimati. Even though Kiritimati stands out as a separate cluster, it is joined with German streams sooner than with Pavilion Lake, suggesting more similarities with the streams than with the samples from Pavilion Lake. As it was with the parallel coordinate plot, Mono Lake and seawater samples separate out from the rest of the observations as having little similarities. Interestingly, the Mono Lake branch is not joined with any other sites until the highest level of the tree, indicating the lack of any significant similarities between this lake and other sites at hand.

### 3.2 Clustering

Once general trends were established, $k$ means clustering was applied in order to separate the data into a number of distinct clusters. This separation gives further insight into the differences and similarities between observations and will be later used for imputation of missing values. $k$ means is a divisive clustering algorithm that separates a variable amount of data into $k$ number of groups by minimizing the distance between observations and the cluster's center (within sum of squares, Johnson and Wichern, 2001). The algorithm requires $k$ (the number of clusters) to be known in advance, which is rarely the case in practice. A number of techniques have been developed to estimating $k$; here we use a simple silhouette plot as well as a pseudo-F statistic (described below).

To search for the best number of clusters, one can compare some measure of goodness of observations' classification across different values of $k$. One such measure, suggested by Kaufman and Rousseeuw (1987), is a cluster silhouette. It is a simple visual way to determine the optimal number of clusters based on minimizing distance



between observations in the same cluster while maximizing separation between the clusters.

Let observation $i$ be classified as a member of cluster $A$. The silhouette value $s(i)$ is defined as

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

where $a(i)$ is the average distance of $i$ to all other observations in $A$ and $b(i)$ is defined as minimum average distance of $i$ to all observations in clusters other than $A$. The number of clusters that maximize average silhouette can be used as the $k$ in $k$ means algorithm.

For our data, a silhouette plot was calculated for the number of clusters ranging from 2 to 10 (Fig. 4). Based on the silhouette plot, it appears that the optimal number of clusters is three. However, the silhouette width for $k = 3$ and $k = 2$ are similar (0.948 and 0.931), indicating that the data could be represented with either 2 or 3 clusters. This lack of distinct separation is explained when we look at cluster composition. Cluster 1 is composed of some observations from Kiritimati samples, cluster 2 is entirely comprised of samples from Mono Lake, and the rest of observations compose the third cluster. This confirms the exploratory analysis, which both identified Mono Lake as the most divergent from all the other sites, and separated the Kiritimati sites into two distinct groups.

### 3.3 Hypothesis test on number of clusters

In addition to determining the optimal number of clusters, it is of interest to ask whether any clustering is beneficial/relevant, or if all water samples should be treated as one pool of observations ($k = 1$). This is equivalent to testing the following hypothesis:

$H_0: k = 1 \quad H_a: k > 1.$

In particular, we first consider the hypothesis with both $k = 3$ and $k = 2$.

A number of techniques for such a test have been suggested in literature. We chose a test based on a pseudo $F$ test statistic. Let "WSS" and "BSS" be defined as "Within" and "Between-cluster Sum of Squares", respectively. Here $n$ is the number of points being clustered. Then, Calinski and Harabazs (1974) define:

$$F = \frac{\mathrm{BSS}/(k-1)}{\mathrm{WSS}/(n-k)} \quad (1)$$

and suggest using it as an informal test statistic (also referred to as the CH index). Cavalli-Sforza (1965) also used this criteria in a context of multivariate cluster analysis.

The pseudo $F$ statistic above for both $k = 3$ and $k = 2$ was found to be less than 0.0001. Thus, there is evidence to reject the null hypothesis; separating data into 3 (or 2) clusters appears to be more beneficial than considering all observations as a uniform pool.

Naturally, a question arises that if $k = 3$, 4, or more clusters would better fit the data than $k = 2$. Calinski and Harabazs (1974) suggest choosing $k$ for which the CH index reaches a global or a local maximum, or at least has a rapid increase. A plot of the CH index as a function of the number of the clusters reveals that 2 clusters appear to be a significantly better choice than any other number up to 10 (Fig. 5).

A natural separation of data into 3 clusters makes sense intuitively – we have observed from the parallel coordinate and dendrogram plots that Kiritimati and the non-microbialite bearing sites (Mono Lake, seawater, Stinking Springs) appear to be different from the rest (and from each other). Looking at average silhouette widths for 2 cluster shows a high silhouette width for $k = 2$ indicating well placed observations to a cluster. On the other hand, for $k = 3$, cluster separation is less defined (Table 2). This hints that one of the three clusters might be formed by observations from the non-microbialite group, and the other two by actual pattern of dissimilarity within data.



### 3.4 *k* means clustering

We applied a modified version of a $k$ means algorithm with $k = 2$. Since a traditional $k$ means analysis is based on Euclidean distance and cannot handle data with missing values, we modify the algorithm to use partial distance instead, and therefore utilize all available information (Himmelspach and Conrad, 2010). Note that a more robust version of $k$ means analysis based on clustering around medoids is used here (Kaufman and Rosseauw, 1987). A medoid is an object representative of a cluster such that total dissimilarity of all objects to their nearest medoid is minimal. An example of a medoid is a cluster mean.

Partial distance ($d_{\mathrm{part}}$) between observations $x = x_1 \ldots x_i \ldots x_p$ and $y = y_1 \ldots y_i \ldots y_p$ in a $p$ dimensional space is computed as follows (Himmelspach and Conrad, 2010):

$$d_{\mathrm{part}}(x,y) = \frac{p}{p - \sum_{i=1}^{p} b_i} \sum_{\text{for all } i:\ b_i=0} (x_i - y_i)^2$$

where

$$\begin{cases} b_i = 0, & \text{if } x_i, y_i \text{ are observed} \\ b_i = 1, & \text{if } x_i, y_i \text{ are observed} \end{cases}$$

and $i = 1, \ldots, p$.

$k$ is the pre-determined number of clusters and $n$ is the number of observations. The clustering algorithm proceeds as follows:

1. Construct initial medoids $m_1, \ldots, m_k$ (i.e. cluster centers) to minimize the sum of all distances between observations in a cluster and cluster's center. Assign observation $i$ to cluster $k$ based on the minimum $d_{\mathrm{part}}(i, m_k)$.

2. Let $i$ be an observation belonging to cluster $C$, $j$ observation not in $C$.

Swap $i$ and $j$ cluster membership if doing so will decrease the objective function. Repeat until convergence. After applying this algorithm to the data, three clusters emerge.

One cluster is composed of 6 observations from atoll Kiritimati, the second of Mono Lake samples, and the third cluster consists of all other observations. In particular, the 6 observations from Kiritimati are characterized by high concentration of Ca and Mg. Separation of one of the control groups into its own cluster confirms our intuition.

### 3.5 Clustering a subset of data

We explored which data features (i.e. variables) might be responsible for the observed clustering of Kiritimati samples. In order to find out what variables contain the most information useful for clustering, *k* means clustering was applied to the set of 6 most complete variables and the resulting cluster assignment was compared to that of the whole dataset. This new set was narrowed in order to determine how many and which variables of the original set could be removed while still retaining the original clustering assignment.

The analysis began by identifying the most complete variables (pH, Ca, Mg, Conductivity, K, Na), deleting incomplete observations (resulting number of rows = 144) and applying clustering to the subsequent dataset. The clustering resulting from these six variables is identical to that of the whole dataset with partial distances (two clusters, six observations from Kiritimati forming its own cluster). This suggests that the deleted variables that were left out of this clustering analysis did not add any significant information to the structure.

Next, we tested to see if any of the six complete variables could also be omitted without changing the clustering structure. In fact, clustering each combination of two out of the six variables (but excluding pH) lead to the clustering assignment identical to that of the whole data set. In other words, restricting the dataset to contain only the information of two variables in Table 2 will produce two clusters, with one of the groups being the 6 observations from Kiritimati. The same holds for combinations of three variables: any combination of three variables, except for those combinations containing pH, produces the same clustering assignment. Combinations containing pH produce



three clusters: two original clusters and a third cluster containing 26 observations from all sites.

When clustering any four variables at a time, resulting cluster assignment is identical to that of clustering six variables (or the whole dataset), even for combinations containing pH. So any third variable combined with pH and another variable (or with any two variables) will produce two clusters as when clustering the whole dataset. Using correlation between the six most complete variables, a dendrogram was constructed to show the similarity between the variables (Fig. 6). pH once again appears separate from the other five variables. In the remainder of this paper, we use the clustering information to fill in the gaps in the dataset (i.e., impute the missing values), explore the meaning of the variability of pH clustering, and interpret our findings in the context of the original research question.

### 3.6 Imputation/filling in data gaps

Now that the data are separated into two distinct groups, cluster information can be used to fill in data gaps by imputing the missing values. A number of imputation procedures have been developed and used extensively in the literature. In this case, we use the basic idea that observations in a given cluster are similar to each other, and therefore missing values can be imputed by a simple cluster mean of that variable. For this data, a non-parametric *k* nearest neighbor imputation (KNN) (Dixon, 1979; Troyanskaya, 2001) was chosen. This method does not make any assumptions about distribution of the data and whether observations are missing at random.

KNN imputation borrows from the simple idea of similarity between the features based on some metric, such as Euclidean distance. If a value in column is missing, then *k* nearest neighbors are found based on Euclidian distance and the average of these neighbors are used to fill in the missing value. This technique produces a complete dataset that can be used.

## 4 Discussion of controls on microbialite formation

One would assume when first considering these vastly different environments, that each site would form a distinct cluster, or perhaps that clusters would change based on which variables in the analysis were considered. We suspected that a pattern would emerge, converging on a variable or set of variables that unite microbialite-forming environments. However, this was not the case.

A few surprising results came out of the statistical analysis of this dataset:

1. Despite the wide range of variables and environments in the initial dataset, three distinct clusters were formed.
2. All of the sites with microbialites (or at least significant microbial carbonate) clustered together, away from Mono Lake.
3. It was found that pH is almost solely responsible for determining the clustering pattern of the observations.
4. Mono Lake is the most distinct group of those analyzed, with seawater and a cluster of Kiritimati samples falling out as the next most distinct group.
5. Samples from atoll Kiritimati show the biggest variation within their own site. One cluster of Kiritimati samples is completely separate from all other sites (grouped with seawater), while the others form a cluster with the German karst streams
6. Stinking Springs, which was included as a control, grouped closely with Pavilion Lake, a wildly different setting.

### 4.1 Clustering

These very distinct environments, when put through rigorous statistical analysis, only seem to be differentiated significantly through pH. Mono Lake naturally separates out



because it has the highest pH of all the sites. However, this does not make pH the dominant control on microbialite formation.

Quite the opposite, this result indicates that pH has *no* control on microbialite formation, so long as reasonable conditions for the precipitation of calcium carbonate are met. This may seem an unexpected finding, as high pH and the exclusion of grazing metazoa is thought to be one of the main reasons microbialites are found in alkaline systems today. It is of note that microbialites are known to form in waters with a very wide spectrum of pH including highly alkaline lakes, in normal seawater, and in acid mine drainages.

Table 1 includes the range of each variable by microbialite site. Samples from atoll Kiritimati not only cluster on their own, but have the widest range of across all variables, as is seen in Fig. 2 (explained by local conditions of the aquatic environment from which the samples were taken in the original study). Samples within this cluster separating loosely into two distinct groups; one with low concentration of Ca, Mg, and salinity, and one with high values for those given variables. The overall distinct clustering of the samples from atoll Kiritimati should be explainable by the fact that these microbialites are aragonite, and marine (and thus cluster with seawater), while the rest are terrestrial, and calcite. The high Mg concentration of the marine environment leads to the precipitation of aragonite. However, as seen in the Fig. 6, Mg concentration does not affect the clustering relationship of the sites, only pH does. Figure 2 shows that a subset of Kiritimati samples have high contents of many major ions (Ca, Sr, K, Na). The combination of all of those variables may be enough to distinctly separate this group.

The location of Stinking Springs in the clustering analysis was unexpected. The warm saline spring clusters most closely to Pavilion Lake, a slightly alkaline freshwater system. From Fig. 2, it appears that their clustering similarity is caused by both pH and their silica content.

### 4.2 Biological processes as possible variables of importance

There could be other variables not included in our analysis that account for microbialite formation. Are the types of bacteria or the community structures similar in all the environments? Are a combination of certain phyla needed to produce microbialites?

Pavilion Lake microbialites are covered in mats dominated by filamentous cyanobacteria (*Oscillatoria sp., Calothrix* sp., *Pseudoanabaena* sp., and *Fisherella* sp.). Heterotrophs and diatoms are also present in large quantities (*Gomphonema*, *Cyclotella,* and *Achnanthes*) (Laval et al., 2000). Microbialite-produce mats from atoll Kiritimati are layered and diverse, with some sections dominated by cyanobacteria (such as *Leptolyngbya, Cyanothece, Entophysalis* and *Spirulina*; Arp et al., 2011). Given the hypersaline nature of the environment, there is a bias towards more salt-tolerant species. Empty diatom test (genus *Navicula*) are found in the mat, though no eukaryote-specific assays were performed (Arp et al., 2011). German stream tufas were coated in biofilms dominated by the cyanobactera *Leptolyngbya* sp., *Phormidium incrustatum/calcareum*, and *Pseudoanabaena.* Diatoms are extrememly diverse in the area, with thirteen different lineages representing eight genera and thirteen species found in the Westerhöfer Bach site alone (Arp et al., 2010). (Other bacteria such as *Proteobacteria, Acidobacteria, Bacteriodetes, Actinobacteria,* and *Nitrospira* are also found in association with the microbialites.)

The non-microbialte forming localities also have interesting biological components. At Stinking Springs, near the spring source, *Oscillatoria* sp. dominate the communities. Along the outflows, layered orange, green and red mats ring the spring water. These mats are composed of layers of cyanobacteria (e.g. *Oscillatoria. psueudoanabaena*), sulfur bacteria (*Desulfobacterales*) and diatoms, among many others (Bonny and Jones, 2007; Gong et al., 2012; Monteverde et al., 2013). The biology of Mono Lake has been of great interest recently, although it is not thought to be a factor in tufa formation.



While certain trends do appear (the ubiquitous presence of *Pseudoanabaena* and *Leptolynbya*, for instance), they can hardly account for microbialite formation. Such cyanobacterial occurrence is extremely common in a variety of environments. It would be telling if there was a novel bacterium or eukaryote (such as the rare "stromatolite builder" strain isolated by Pepe-Raney et al., 2012) that was common to all sites, however this is not the case. All of the reported sequences are common in these sites, as well as in numerous environments that do not harbor microbialites. Clearly, the presence of these certain communities alone cannot account for the building of microbialites.

Results from this work suggest that microbialites are broadly distributed across the environments with a wide spectrum of geochemical characteristics. None of the variables studied here are readily responsible for the formation of microbialites. However, statistical analysis conducted here was restricted to the set of six microbialite sites and three control sites. In the future, *k* means clustering with partial distance can be easily applied to a bigger dataset, particularly one including data from environments that do not form microbialites. If the new dataset contains missing observations/data gaps, KNN imputation can be used to fill in the missing values and conduct further statistical analysis.

### 4.3 Other factors to explore through future research

Several variables that we were not able to consider are likely also important for the formation of microbialites. For example, we cannot exclude the possibility that physical characteristics of the tested sites control microbialite distribution, such as grazing activity, water agitation, flow rate, and clarity. Additionally trapping and binding does not depend exclusively on the microbial community but also on the availability of detrital particles and the energy of the sedimentary system. Extracellular polymeric substances (EPS, major components of a microbial mat biomass) are known to play a key role for both trapping and binding and for influencing and promoting authigenic nucleation of minerals. Differences in EPS abundance and their chemical composition may

be key factors controlling the occurrence of microbialites. Quantification of metabolic rate may also be a key factor in "microbially induced" precipitation of authigenic minerals within microbial mats.

Therefore although this clustering analysis shows no clear control of geochemical characteristics on the distribution of microbialites and counters the premise that microbialites are limited to settings of a particular pH or salinity range, further study is needed to fully elucidate the controls on microbialite distribution. An issue is that many variables that control microbialite distribution are only rarely reported in the literature, and/or are very difficult and time consuming to obtain. Additionally the examination of a larger number of sites, as data become available, will also allow for a more complete assessment on the factors influencing microbialite formation.

## 5 Conclusions

In this work, we have explored properties of geochemical characteristics of several different microbialite-forming environments, ranging from freshwater to hypersaline. The initial dataset with missing values was clustered via *k* means algorithms using partial distances. The dataset was narrowed to the 6 most complete variables (pH, Ca, Mg, conductivity, K, Na) and analysis was repeated to determine whether there is a subset of variables that produces the same clustering results as with the whole dataset.

This analysis resulted in pH being separated out as particularly different from the rest of the variables, and being almost solely responsible for the patterns of clustering. Clustering distinguished samples from atoll Kiritimati to be particularly distinct from the other sites, which is most likely due to the mineralogy of the microbialites (aragonite vs. calcite). When considering the biology of the sites as a potential variable that could explain the pattern, no distinct trends readily emerged.

Observations that were omitted originally were then clustered using partial distances as the measure of similarity. The resulting clustering assignment was used to impute missing values using *k* nearest neighbors procedure. This paper can be used as a general outline of methods that could be applied to an expanded multivariate dataset with missing values. Also, different imputation techniques could be applied and compared against the one presented here.



These results indicate that, contrary to commonly held beliefs about microbialite formation, salinity and high pH are not important variables. It is clear that as long as the conditions for carbonate precipitation are met, microbialites can form at a range of pH. Moving forward, the methods outlined in this study can be used to construct a larger dataset which compares these results to those from other microbialite-forming environments, and non-microbialite forming environments.

*Acknowledgements.* We wish to thank the participants and instructors of the International Geobiology Summer Course (2012 and 2013), which was funded through The Agouron Institute and the Gordon and Betty Moore Foundation, as well as NASA and NSF. AKT acknowledges support from DOE grant DE-FG0213ER16402, ACS grant 51182-DNI2, and NSF grants OCE-1437166, EAR-1352212, EAR-1325054, and EAR-0949191. This work was supported by the "Laboratoire d'Excellence" LabexMER (ANR-10-LABX-19) and co-funded by a grant from the French government under the program "Investissements d'Avenir".

## References

Allwood, A. C., Walter, M. R., Kamber, B. S., Marshal, C. P., and Burch, I. W.: Stromatolite reef from the Early Archean era of Australia, Nature, 441, 714–718, 2006.

Arp, G., Hoffman, J., and Reitner, J.: Microbial fabric formation in spring mounds ("microbialites") of alkaline salt lakes in the Badain Jaran Sand Sea, PR China, Palaios, 13, 581–592, 1998.

Arp, G., Reimer, A., and Reitner, J.: Calcification in cyanobacterial biofilms of alkaline salt lakes, Eur. J. Phycol., 34, 393–403, 1999a.

Arp, G., Thiel, V., Reimer, A., Michaelis, W., and Reitnery, J.: Biofilm exopolymers control microbialite formation at thermal springs discharging into the alkaline Pyramid Lake, Nevada, USA, Sediment. Geol., 126, 159–176, 1999b.

Arp, G., Reimer, A., and Reitner, J.: Photosynthesis-induced biofilm calcification and calcium concentrations in Phanerozoic oceans, Science, 292, 1597–1784, 2001.

Arp, G., Reimer, A., and Reitner, J.: Microbialite formation in seawater of increased alkalinity, Satonda crater lake, Indonesia, J. Sediment. Res., 73, 105–127, 2003.
Arp, G., Bissett, A., Brinkmann, N., Cousin, S., De Beer, D., Friedl, T., Mohr, K. I, Neu, T. R., Reimer, A., Shiraishi, F., Stackebrandt, E., and Zippel, B.: Tufa-forming biofilms of German karstwater streams: microorganisms, exopolymers, hydrochemistry and calcification, Geological Society, London, Special Publications, 336, 83–118, 2010.
Arp, G., Helms, G., Karlinska, K., Schumann, G., Reimer, A., Reitner, J., and Trchet, J.: Photosynthesis versus exopolymer degradation in the formation of microbialites on the Atoll of Kiritimati, Republic of Kiribati, Central Pacific, Geomicrobiol. J., 29, 29–65, 2012.
Biscoff, J. L., Stine, S., Rosenbauer, R. J., Fitzpatrick, J. A., and Stafford, T. W.: Ikaite precipitation by mixing of shoreline springs and lake water, Mono Lake, California, USA, Geochim. Cosmochim. Ac., 51, 1413–1423, 1993.
Bonny, S. M. and Jones, B.: Diatom-mediated barite precipitation in microbial mats calcifying at Stinking Springs, a warm sulfur system in Northwestern Utah, USA, Sediment. Geol., 194, 223–244, 2007.
Brady, A. L., Slater, G. F., Omelon, C. R., Southam, G., Druschel, G., Andersen, D. T., Hawes, I., Laval, B., Lim, D. S. S.: Photosynthetic isotope biosignatures in laminated micro-stromatolitic and non-laminated nodules associated with modern, freshwater microbialites in Pavilion Lake, B.C., Chem. Geol., 274, 56–67, 2010.
Burne, R. V. and Moore, L. S.: Microbialites: organosedimentary deposits of benthic microbial communities, Palaios, 2, 241–254, 1987.
Calinski, T. and Harabazs, J.: A dendrite method in cluster analysis, Commun. Stat., 3, 1–27, 1973.
Dixon, J. K.: Pattern recognition with partly missing data, IEEE T. Syst. Man Cyb., 9, 617–621, 1979.
Dunn, J. R.: Origin of the deposits of tufa in Mono Lake, J. Sediment. Petrol., 23, 18–23, 1953.
Gong, J., Edwardson, C., Mackey, T. J., Dzaugis, M., Ibarra, Y., Course 2012, G., Frantz, C. M., Osburn, M. R., Hirst, M., Williamson, C., Hanselmann, K., Caporaso, J., Sessions, A. L., and Spear, J. R.: Microbial Diversity and Lipid Abundance in Microbial Mats from a Sulfidic, Saline, Warm Spring in Utah, USA. American Geophysical Union, Fall Meeting 2012, abstract No. B51D-0593, 2012.
Grotzinger, J. P. and Knoll, A. H.: Stromatolites in precambrian carbonates: evolutionary mileposts or environmental dipsticks?, Annu. Rev. Earth Pl. Sc., 27, 313–358, 1999.





Hawes, I., Symner, D. Y., Andersen, D. T., and McKey, T. J.: Legacies of recent environmental change in the benthic communities of Lake Joyce, a perennially ice-covered Antarctic Lake, Geobiology, 9, 394–410, 2011.
Himmelspach, L. and Conrad, S.: Clustering Approaches for Data with Missing Values: Comparison and Evaluation, Fifth International Conference on Digital Information Management, (ICDIM) 2010, 19–28, 2010.
Hofmann, H. J., Grey, K., Hickman, A. H., and Thorpe, R. I.: Origin of 3.45 Ga coniform stromatolites in Warrawoona Group, Western Australia, Geol. Soc. Am. Bull., 111, 1256–1262, 1999.
Johnson, R. A. and Wichern, D. W.: Applied Multivariate Statistical Analysis, Prentice Hall, Upper Saddle, NJ, 2001.
Kaufman, L. and Rousseeuw, P.: Clustering by Means of Medoids, 1987.
Kazmierczak, J. and Kempe, S.: Satonda Crater Lake, Indonesia: hydrogeochemistry and biocarbonates, Facies, 28, 1–32, 1993.
Kazmierczak, J. and Kempe, S.: Genuine modern analogues of Precambrian stromatolites from caldera lakes of Niuafo'ou Island, Tonga, Naturwissenschaften, 93, 119–126, 2006.
Kazmierczak, J., Kempe, S., Kremer, B., Lopez-Garcıa, P., Moreira, D., and Tavera, R.: Hydrochemistry and microbialites of the alkaline crater lake Alchichica, Mexico, Facies, 57, 543–570, 2011.
Laval, B., Cady, S. L., Pollack, J. C., McKay, C. P., Bird, J. S., Grotzinger, J. P., Ford, D. C., and Bohm, H. R.: Unique assemblage of modern freshwater microbialites, Pavilion Lake, British Columbia, Canada, Nature, 407, 626–629, 2000.
Lim, D., Laval, B. E., Slater, G., Antoniades, D., Forrest, A. L., Pike, W., Pieters, R., Saffari, M., Reid, D., Schulze-Makuch, D., Andersen, D., and McKay, C. P.: Limnology of Pavilion Lake, B. C., Canada – characterization of microbialite forming environment, Fund. Appl. Limnol., 173, 329–351, 2009.
Lim, D. S. S., Brady, A. L., Abercromby, A. F., Andersen, D. T., Andersen, M., Arnold, R. R., Bird, J. S., Bohm, H. R., Cady, S. L., Cardman, Z., Chan, A. M., Chan, O., Chénard, C., Cowie, B. R., Davila, A., Deans, M. C., Dearing, W., Delaney, M., Downs, M., Fong, T., Forrest, A., Gernhardt, M. L., Gutsche, J. R., Hadfield, C., Hamilton, A., Hawes, I., Hansen, J., Heaton, J., Imam, Y., Laval, B., Lees, D., Leoni, L., Looper, C., Love, S., Marinova, M. M., McCombs, D., McKay, C. P., Mullins, G., Nebel, S. H., Nuytten, P., Pendery, R., Pike, W., Pointing, S. P., Pollack, J., Raineault, N., Reay, M., Reid, D., Sallstedt, T., Schulze-Makuch, D.,

Seibert, M., Shepard, R., Slater, G. F., Sumner, D. Y., Suttle, C. A., Trembanis, A., Turse, C., Wilhelm, M., Wilkinson, N., Williams, D., Winget, D. M., and Winter, C.: A historical overview of the Pavilion Lake Research Project – analog science and exploration in an underwater environment, Geol. S. Am. S., 483, 85–116, 2011.

Metzger, J. G., Monteverde, D., Kelly, H., Bournod, C., Wang, D. T., Frantz, C. M., Osburn, M. R., Berelson, W., Sessions, A. L., Hanselmann, K., Johnson, H., Stamps, B. W., Vuono, D., Shapiro, R. S., and Spear, J. R.: Biogeochemistry of Stinking Springs, UT Part I: Inorganic carbon dynamics and constraints on nutrient fluxes in a warm, salty, sulfidic spring. American Geophysical Union, Fall Meeting 2013, abstract No. B13E-0556, 2013.

Monteverde, D., Metzger, J. G., Bournod, C., Kelly, H., Johnson, H., Sessions, A. L., Osburn, M., Shapiro, R. S., Rideout, J., Johnston, D. T., Stevenson, B., Stamps, B. W., Vuono, D., Hanselmann, K., and Spear, J. R.: Biogeochemistry of Stinking Springs, Utah. Part II: Microbial Diversity and Photo- and Chemo-Autotrophic Growth Rates in a Layered Microbial Mat. American Geophysical Union, Fall Meeting 2013, abstract No. B13E-0550, 2013.

Nielson, L. C. and DePaolo, D. J.: Ca fractionation in a high-alkalinity lake system: Mono Lake, California, Geochim. Cosmochim. Ac., 118, 276–294, 2013.

Pepe-Raney, C., Berelson, W. M., Corsetti, F. A., Treants, M., and Spear, J. R.: Cyanobacterial construction of hot spring siliceous stromatolites in Yellowstone National Park, Environ. Microbiol., 14, 1182–1197, 2012.

Petryshyn, V. A., Corsetti, F. A., Berelson, W. M., Lund, S. P., and Beaumont, W.: Stromatolite lamination frequency, Walker Lake, Nevada: implications for stromatolites as biosignatures, Geology, 40, 499–502, 2012.

Petryshyn, V. A., Lim, D., Laval, B. L., Brady, A., Slater, G., and Tripati, A. K.: Reconstruction of limnology and microbialite formation conditions from carbonate clumped isotope thermometry, Geobiology, 13, 53–67, 2015.

Riding, R. E. and Awramik, S. M.: Microbial Sediments, Springer-Verlag, Cambridge, Massachusettes, 2000.

Saenger, C., Miller, M., Smittenberg, R. H., and Sachs, J.: Physio-chemical survey of inland lakes and saline ponds: Christmas Island (Kritimati) and Washington (Teraina) Islands, Republic of Kiribati, Aquatic Biosystems, 2, 1–15, doi:10.1186/1746-1448-2-8, 2006.

Scholl, D. and Taft, W.: Algae, contributors to the formation of calcareous tufa, Mono Lake, California, J. Sediment. Res., 34, 309–319, 1964.



Semikhatov, M. A., Gebelein, C. D., Cloud, P., Awramik, S. M., and Benmore, W. C.: Stromatolite morphogenesis – progress and problems, Can. J. Earth Sci., 16, 992–1015, 1979.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, B., and Altman, R.: Missing value estimation methods for DNA microarrays, Bioinformatics, 17, 520–525, 2001.

Valencia, M. J.: Christmas Island (Pacific Ocean): reconnaissance geologic observations, Atoll Research Bulletin, 197, 1–17, 1977.

Walter, M. R.: Stromatolites: Developments in Sedimentology, 20, Elsevier Scientific Publishing Company, Amsterdam, 1976.

**Table 1.** Reported variables for study sites: Pavilion Lake (PL), Kiritimati (K), Westerhöfer Bach (WB), Deinschwanger Bach (DB), Reinsgraben (R), and Steinern Rinne (SR), Mono Lake (M), seawater (S), Stinking Springs (SP).

| Variable | PL | K | WB | DB | R | SR | M | S | SP | Units |
|---|---|---|---|---|---|---|---|---|---|---|
| pH | 7.12–9.13 | 7.16–9.64 | 7.32–8.3 | 7.36–8.5 | 7.31–8.28 | 7.1–8.03 | 9.8–9.8 | 8.3–8.3 | 6.3–7.4 | |
| Ca | 0.92–10.63 | 0.22–38.6 | 3.58–3.95 | 1.87–2.28 | 4.89–5.2 | 2.86–3.52 | 0.106–0.106 | 10–10 | 16.5–23 | $mmol\,L^{-1}$ |
| Mg | 0.22–255.09 | 0.1–279.7 | 1.67–1.72 | 0.86–1.27 | 1.11–1.13 | 0.13–0.14 | 1.54–1.54 | 0.05–0.05 | 10.4–18.35 | $mmol\,L^{-1}$ |
| Conductivity | 0.27–26.6 | 0.2–156.3 | 0.90–1.04 | 0.58–0.64 | 1.14–1.23 | 0.52–0.65 | 77–91 | 50–50 | 30.9–52.4 | $mS\,cm^{-1}$ |
| Alkalinity | 0–0.018 | 0.19–14.2 | 4.8–5.4 | 4.68–5.22 | 4.63–5.2 | 4.42–6.1 | 498.4–498.4 | 2.3–2.3 | 2.8–8.7 | $meq\,L^{-1}$ |
| $SO_4$ | 0.075–321.66 | 0.05–139.17 | 2.81–2.95 | 0.17–0.19 | 3.63–3.8 | 0.27–0.27 | 102.9 | 0.03 | 0.615–4.9 | $mmol\,L^{-1}$ |
| K | 0.018–22.30 | 0–48 | 0.052–0.055 | 0.028–0.058 | 0.043–0.046 | 0.014–0.022 | 37.4–37.4 | 9.74–9.74 | 12.6–23.3 | $mmol\,L^{-1}$ |
| Na | 0.061–211.83 | 1–2334 | 0.33–0.34 | 0.31–0.35 | 0.606–0.623 | 0.13–0.20 | 1187–1187 | 0.46–0.46 | 274–652 | $mmol\,L^{-1}$ |
| Si | 0.0018–0.584 | 0.002–0.19 | 0.157–0.159 | 0.097–0.102 | 0.148–0.155 | 0.084–0.084 | 0.0675–0.0675 | 0.225–0.225 | 0.1475–0.225 | $mmol\,L^{-1}$ |
| Cl | 0.0169–17.347 | 0–2637 | 0.291–0.299 | 0.55–0.65 | 0.546–0.596 | 0.14–0.14 | 494.3–494.3 | 0.54–0.54 | 342.9–742.9 | $mmol\,L^{-1}$ |
| Sr | 0.0009–0.0944 | 0.0013–0.32 | 0.0175–0.019 | 0.00032–0.00055 | n/a | 0.00151–0.00156 | 0.0002–0.0002 | 0.075–0.075 | 0.137–0.217 | $mmol\,L^{-1}$ |
| Ba | $7.28 \times 10^{-5}$–0.0012 | n/a | 0.00021–0.00025 | 0.00011–0.00013 | n/a | n/a | n/a | n/a | 0.0035–0.07 | $mmol\,L^{-1}$ |
| Source | Lim et al. (2009) | Arp et al. (2011) | Arp et al. (2010) | Arp et al. (2010) | Arp et al. (2010) | Arp et al. (2010) | Nielsen and DePaolo (2013) | | Bonny and Jones (2007) | |





**Table 2.** Average Silhouette Widths for 2 and 3 clusters.

| 2 Clusters ($k$ = 2) | 3 Clusters ($k$ = 3) |
|---|---|
| 0.950 | 0.844 |
| 0.919 | 0.104 |
| | 0.914 |

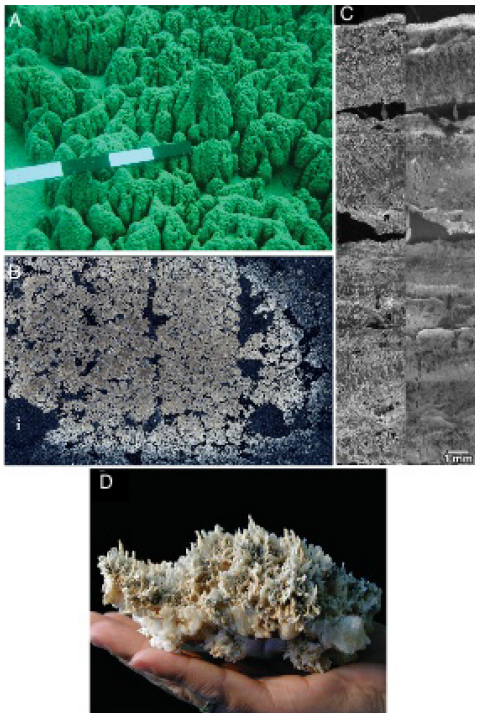


**Figure 1.** Samples of microbialites from environments used in this study. **(a)** Field photo of Pavilion Lake microbialites at 26 m water depth (from Petryshyn et al., 2015). **(b)** Thin section of same microbialite (from Petryshyn et al., 2015). **(c)** Figure 13a from Arp et al. (2010). Thin section of weakly-laminated microbialite from the German karst stream Westerhöfer Bach. **(d)** Figure 13a from Arp et al. (2011). Hand sample of a reticulate microbialite from atoll Kiritimati.





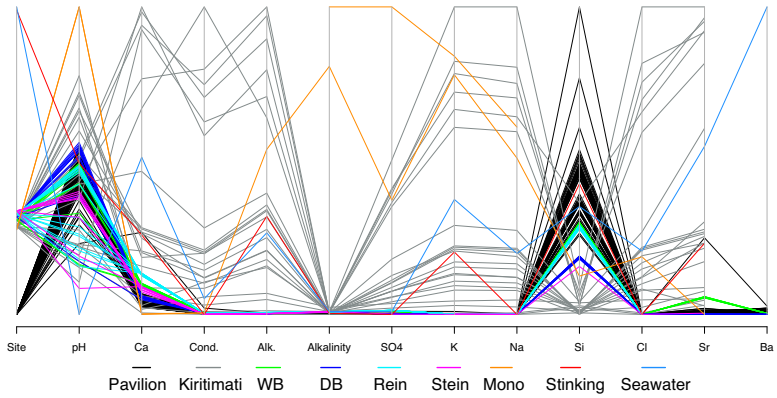


**Figure 2.** Parallel coordinates plot for the geochemical data of the sites. This exploratory plot allows for the initial visualization of similarities and differences between sites. The 12 variables considered in this analysis are listed on the horizontal axis. Kiritimati (grey) appears to stand apart from the rest of the sites. Seawater (light blue) differs from the rest in pH, Ca, Ba, and K concentrations. Mono Lake (orange) measurements appear to be significantly different in most variables as well. Finally, Stinking Springs (red) does not appear to differ as much from the microbialite-bearing sites as the other two non-micribialite-bearing sites.

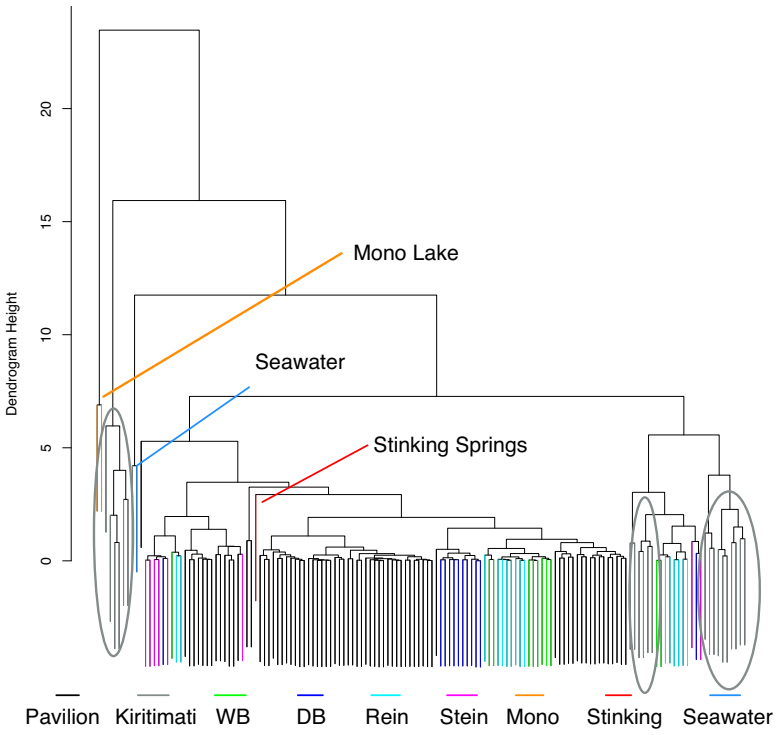


**Figure 3.** Hierarchical clustering of all observations. Control Groups (Stinking Springs, Mono Lake, and Seawater) are noted. Ovals highlight the clustering relationship of the Kirtimati sites.





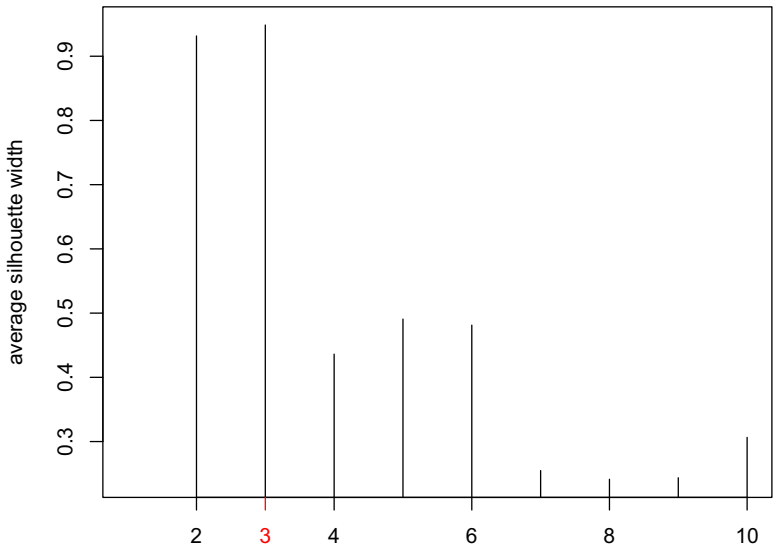


**Figure 4.** silhouette Plot for determining the optimal number of clusters. The number of clusters that maximize average silhouette can be used as the $k$ in $k$ means algorithm. Based on the plot, it appears that the optimal number of clusters is three. However, the silhouette width for $k = 3$ and $k = 2$ are similar (0.948 and 0.931), indicating that the data could be represented with either 2 or 3 clusters.

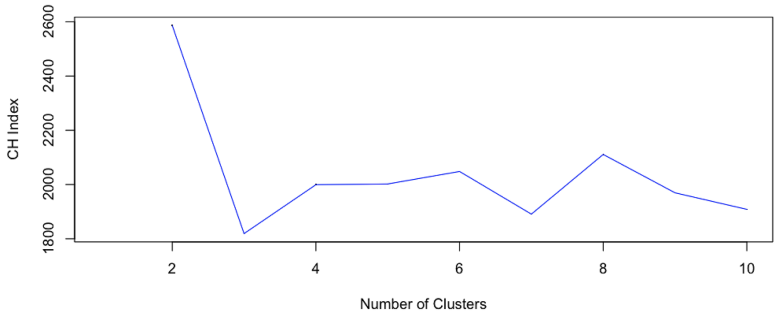


**Figure 5.** A plot of the CH index as a function of the number of the clusters. This test determines whether it is beneficial to treat the data as clusters ($k$ = 2, 3, ..., 10), or if all water samples should be treated as one pool of observations ($k$ = 1). The plot reveals that 2 clusters appear to be a significantly better choice than any other number up to 10.





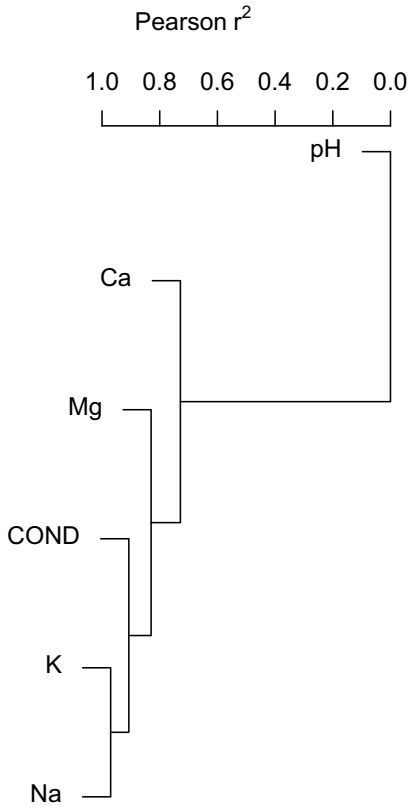


**Figure 6.** Dendrogram of 6 most complete variables, showing the similarity between these measured parameters. pH appears separate from the other five variables.