LIMNOLOGY and OCEANOGRAPHY

ASLO

Limnol. Oceanogr. 66, 2021, 4001–4016

doi: 10.1002/lno.11938

Dissolved organic compounds with synchronous dynamics share chemical properties and origin

Julian Merder,[1,2*] **Heidelinde Röder**,[1] **Thorsten Dittmar**,[1,3] **Ulrike Feudel**,[1] **Jan A. Freund**,[1] **Gunnar Gerdts**,[4] **Alexandra Kraberg**,[5] **Jutta Niggemann**[1*]

[1]Institute for Chemistry and Biology of the Marine Environment (ICBM), University of Oldenburg, Oldenburg, Germany
[2]Department of Global Ecology, Carnegie Institution for Science, Stanford, California
[3]Helmholtz Institute for Functional Marine Biodiversity (HIFMB), University of Oldenburg, Oldenburg, Germany
[4]Biosciences Division, Shelf Sea System Ecology, Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Sciences (AWI), Helgoland, Germany
[5]Biosciences Division, Polar Biological Oceanography, Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Sciences (AWI), Bremerhaven, Germany

Abstract

The temporal dynamics of dissolved organic matter (DOM) are inherently linked with the functioning of aquatic ecosystems. Because DOM represents a complex mixture of millions of different compounds, the statistical analysis of DOM dynamics poses a huge challenge. Here, we present a statistical approach based on hierarchical clustering of time series that groups DOM compounds with synchronous dynamics. We applied this approach to time series of Fourier-transform ion cyclotron resonance mass spectrometry data of DOM sampled over a period of 26 months near Helgoland, an island in the Southern North Sea. We identified three DOM clusters, which represented a total of 1392 different molecular formulae and showed distinct chemical properties and noticeably compound matches within the PubChem database. Correlations of the three DOM clusters with abundance data of prokaryote and phytoplankton species and with environmental parameters provided consistent indications on the potential origin of the clustered compounds. The first cluster integrated terrestrial DOM originating from riverine discharge reaching Helgoland waters. The second cluster was attributed to DOM related to phytoplankton and microbial activity, whereas the third cluster was interpreted as representing the marine refractory DOM background. Accordingly, while further partitioning divided each of the first two clusters into five sub-clusters with distinct temporal dynamics and molecular characteristics, the third cluster persisted as a stable feature. Applying a purely mathematical approach, we thus confirmed the differential dynamics of individual DOM compounds and compound groups and showed that temporal dynamics of dissolved molecules are linked to their origin and transformation history.

Dissolved organic matter (DOM) represents one of the largest active carbon pools on earth (700 Gt) comparable in size to the Earth's atmospheric CO_2 or all land plant biomass (Hedges 1992; Hansell et al. 2009). The amount of dissolved organic carbon (DOC) is more than 200 times higher than that of organic particulate carbon in the ocean, underpinning its significance in the microbial loop, for the remineralization of nutrients and the marine food web (Azam 1998; Hansell and Carlson 2015). An improved understanding of DOM dynamics is thus essential for a complete comprehension of the global carbon cycle (Hansell and Carlson 2015). DOM in the ocean does not stem from a single source but is a mélange of substances produced by marine organisms and terrestrial compounds introduced by rivers (e.g., Raymond and Bauer 2001; Moran et al. 2016). Processes such as photochemical alteration at the sea surface further modify the composition of DOM (Stubbins and Dittmar 2015). Eventually, this leads to highly complex mixtures consisting of thousands of different substances with fluctuating occurrence and concentrations. This molecular diversity of DOM poses major analytical challenges, not only for the correct determination and distinction of the different compounds of DOM but

*Correspondence: julian.merder@uni-oldenburg.de; jutta.niggemann@uni-oldenburg.de

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

also for devising appropriate methods for a statistical analysis of DOM data.

With regard to the precise identification of compounds, analytical methods for DOM characterization have improved a lot over the last decades. Fourier-transform ion cyclotron resonance mass spectrometry (FT-ICR-MS) has established itself as state-of-the-art analytical method in marine DOM geochemistry (Nebbioso and Piccolo 2013). FT-ICR-MS and subsequent software tools enable the synchronous detection of exact masses and identification of thousands of molecular formulae within the complex DOM mixtures (D'Andrilli et al. 2010; Riedel and Dittmar 2014). The software tools applied for processing of FT-ICR-MS data have improved over the last years as well (Tolić et al. 2017; Leefmann et al. 2019; Merder et al. 2020*a*). Due to these improvements, the follow-up statistical data analyses of the ever-growing data sets have become the new bottleneck for the interpretation of extensive DOM molecular data sets, such as long-term time series. Ideally, statistical data analyses are not only able to identify the dynamics of DOM compounds but also to shed light on the factors controlling these dynamics. To achieve this goal, the dynamics of DOM compounds must be linked to factors that potentially affect DOM composition, that is, abundances of prevailing organisms, especially microbial species, and environmental parameters.

As DOM compounds are both produced and consumed by marine biota, classical regression models, including generalized linear or additive models, are not appropriate because they presume a clear-cut separation into response and explanatory variables which is challenged by feedback mechanisms in the complex interaction network. Furthermore, it is unfeasible to interpret regression results for each of the thousands of formulae independently. There are several approaches tackling these issues (Kujawinski et al. 2016; Lucas et al. 2016; Osterholz et al. 2016*b*), all of them focusing on the collective compositional change instead of exploring dynamics of every compound on its own. For example, beta-diversity indices such as Bray–Curtis dissimilarity (Bray and Curtis 1957) yield a single value that describes compositional differences of two samples. Such approaches are indispensable tools to identify the processes that affect DOM composition (Osterholz et al. 2016*a*; Hawkes et al. 2018); however, they bear the risk of losing valuable information because the compounds that predominantly cause the change are often not easily identifiable. Moreover, it is difficult to separate true compositional changes from left over noise or contaminations remaining in the data sets. Furthermore, compounds that are produced and consumed at high rate, and as such drive microbial life in the ocean, may be present in seawater at very low concentration. These compounds may largely be hidden behind an invariable background of compounds. As such they are not appropriately assessed by statistical approaches that condense rich molecular data into bulk parameters.

Here, we introduce a new approach combining correlation (synchronous variation), hierarchical clustering and machine learning techniques applied to highly complex time series data. The time series analyzed in this study is unique with respect to its multivariate dimensions and, to our knowledge, one of the longest and best-resolved molecular DOM time series existing to date. It is based on FT-ICR-MS data of DOM sampled between March 2009 and May 2011 from surface waters off Helgoland Island in the German Bight. We included complementary time-series on abundances of prokaryotes and phytoplankton species (Wiltshire et al. 2010; Teeling et al. 2012), to identify covariations of microbial communities and DOM composition. The data set was further amended by time series of environmental variables, which provide necessary information on abiotic conditions that directly or indirectly affect DOM dynamics.

Selected aspects of DOM dynamics such as turnover of labile substrates and production of refractory compounds (Amon et al. 2001; Ogawa et al. 2001; Osterholz et al. 2015) or photodegradation (Vähätalo and Wetzel 2004; Stubbins and Dittmar 2015) have been studied in short- and long-term laboratory incubations. To understand DOM dynamics in natural environments, recent studies characterized and interpreted changes in molecular DOM composition along spatial gradients, for example, salinity or ocean currents (Flerus et al. 2012; Jørgensen et al. 2014; Osterholz et al. 2016*b*). In most cases, such spatial gradients also involve temporal scales, for example, aging of water masses, but environmental studies explicitly covering temporal variation of DOM components are scarce and limited in duration (Lucas et al. 2016) or in temporal and analytical resolution, like the many studies that report seasonal variations of optical DOM characteristics (e.g., Galletti et al. 2019).

The main goals of this study were (1) the identification and characterization of groups of DOM compounds with similar temporal dynamics through hierarchical cluster analysis of multivariate time series and (2) the interpretation of the temporal dynamics of the identified clusters in the context of environmental conditions and co-occurring microbial communities in order to assess potential sources and transformation histories of the respective DOM compounds.

Materials and methods

Study area, sampling, and data sets

All samples were obtained at Helgoland Island, which is located in the Southeastern North Sea ~ 50 km offshore in the German Bight. The sampling station Helgoland Roads ("Kabeltonne") is located (54°11.3′N, 7°54.0′E) between the main island and the minor island "Düne" (Fig. 1). Surface water (~ 2 m water depth) was sampled between March 2009 and May 2011, up to twice a week. DOM was solid-phase extracted (SPE; Dittmar et al. 2008). In brief, 2 L of filtered (Whatman GF/F), acidified seawater (pH 2; HCl, p.a.) was processed via 1 g of PPL resin (Agilent) and the SPE-DOM was

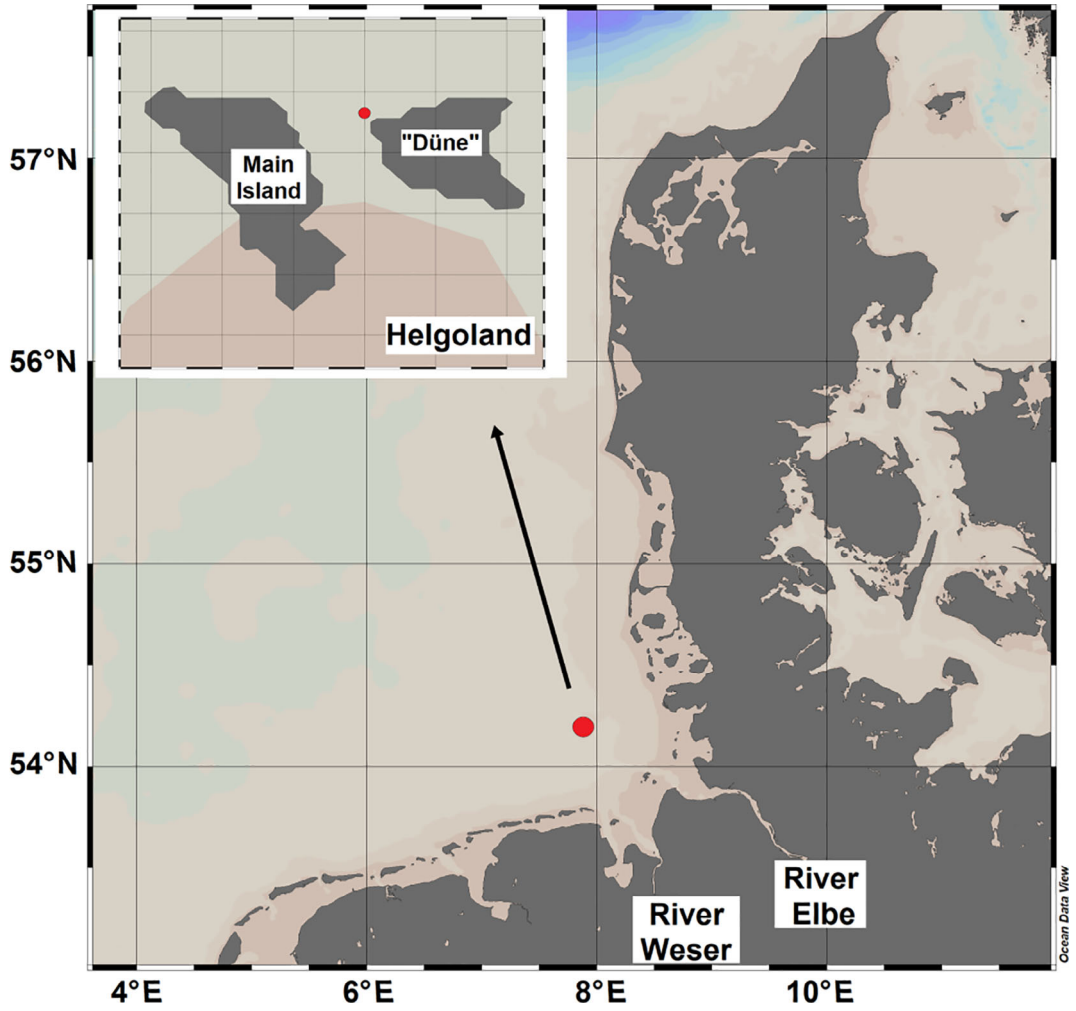

Fig. 1. Location of the study site Helgoland Roads indicated by red symbol ("Kabeltonne"; 54°11.3′N, 7°54.0′O) in the German Bight (North Sea).

eluted with 6 mL of methanol (MS grade) after desalting with acidified ultrapure water.

Physical and chemical parameters were routinely determined as part of the Helgoland Roads time series (Wiltshire et al. 2010; Wiltshire 2013; Wiltshire et al. 2015), including temperature, Secchi depth, salinity as well as concentrations of silicate, phosphate, nitrate, nitrite, and ammonium. Chlorophyll *a* (Chl *a*) concentrations were determined by two independent methods, one based on chlorophyll fluorescence measured with an AlgaeLabAnalyser (bbe Moldaenke), whereas the other quantified chlorophyll in the particulate (< 0.7 μm) fraction via high-pressure liquid chromatography (Wiltshire et al. 2010; Teeling et al. 2012). Concentrations of DOC and total dissolved nitrogen (TDN) were quantified after high-temperature catalytic combustion with a Shimadzu TOC analyzer. Accuracy of the DOC and TDN determinations was tested with help of the deep-sea reference provided by Dennis A. Hansell (University of Miami). Dissolved organic nitrogen (DON) concentrations were calculated as the difference of TDN and dissolved inorganic nitrogen (DIN: nitrate, nitrite, and ammonium) concentrations.

Abundance data of 95 phytoplankton taxa were obtained from the Helgoland Roads time series database (Wiltshire 2013). For comparison with the microbial (prokaryote) community, we used a freely available data set (Fuchs et al. 2016) that originated from the same location as the DOM data set and covered the period from January 2009 to June 2012. Sampling details and the structure of the microbial data are outlined in previous studies (Teeling et al. 2012; Fuchs et al. 2016). In brief, the microbial (prokaryote) community composition was analyzed using catalyzed reporter deposition fluorescence in situ hybridization of formaldehyde fixed cells on 0.2-μm pore-sized filters, and reported as relative abundances of individual taxa (Fuchs et al. 2016; Supporting Information Table S1). Occurrence and relative abundances of individual taxa follow distinct succession patterns (Teeling et al. 2012; Fuchs et al. 2016).

Molecular characterization of DOM

For analysis on the solariX 15 Tesla FT-ICR-MS (Bruker Daltonics), DOM extracts were diluted with ultrapure water and methanol (MS grade) to yield a DOC concentration of 20 mg C L^{-1} and a methanol to water ratio of 1 : 1 (v/v). Instrument settings were as specified in Seidel et al. (2014). For each mass spectrum, 500 scans were accumulated in the scanning range of 150–2000 Da. The instrument was externally calibrated with arginine clusters, and spectra were internally calibrated with a list of more than 100 known $C_xH_yO_z$ molecular formulae covering the mass range of the samples to achieve a mass error of < 0.1 ppm. A reference SPE-DOM sample from North Equatorial Pacific Intermediate Water (NEqPIW; Green et al. 2014) was frequently analyzed to control for instrument stability.

FT-ICR-MS data processing

Data generated by FT-ICR-MS analyses were processed with ICBM-OCEAN, an open platform for DOM mass spectra processing (Merder et al. 2020*a*). Analytical noise was defined based on the method detection limit and removed (Riedel and Dittmar 2014). The resulting data were used for a sample wise generalized additive model-based recalibration along the mass axis to reduce the systematic error (Merder et al. 2020*a*). Masses were aligned across all spectra resulting in improved mass precision and consequently subsequent molecular formula attribution (Merder et al. 2020*b*). For molecular formula attribution, we included the following elements (abundance ranges): C (1–100), ^{13}C (0–1), H (1–200), O (0–100), ^{18}O (0–1), N (0–6), ^{15}N (0–1), S (0–3), ^{34}S (0–1), P (0–3), and searched for matches in a tolerance range of 0.5 ppm. From these data, we filtered molecular formulae that were isotope verified and used them for further analysis. Isotope verified means that the molecular formula is accompanied by at least one isotopologue with a correct intensity ratio deviance (Merder et al. 2020*b*) using tolerances of $\pm$ 1000 permille from natural abundance for Δ^{13}C, Δ^{18}O, Δ^{15}N, and Δ^{34}S. If for a single mass more than one formula suggestion was isotope verified, we chose the molecular formula with the largest homologous series network, considering CH_2 and O (for details, *see* Merder et al. 2020*b*). Intensities were normalized by dividing each intensity by its respective sample mean intensity. Normalization was done as the last step of raw data processing, as removal of selected formulae can have a large impact on the calculation of the mean.

Hierarchical cluster analysis of DOM time series

All statistical analyses were performed with the Software "R" (R Core Development Team 2017) using the packages "vegan," "cluster," "ggplot2," "party," and "visNetwork."

For cluster analysis, only DOM molecular formulae that were detected in at least 50% of the investigated time points were considered. For a given pair of molecular formulae, "Spearman's rank correlation" (r_s) was computed from related intensity time series and used as a similarity metric. We preferred Spearman's rank correlation coefficient over Pearson's correlation coefficient because it is more robust and not restricted to linear correlations. Molecular formulae were ranked by aligning the formula-wise *z*-scored intensities in descending order. The *z*-scoring (subtracting the mean and dividing by the standard deviation) has no effect on the r_s calculation as it is a monotonic, hence, order preserving transformation but it brings formulae intensities to a comparable scale (Fig. 2). It should be emphasized that r_s was based only on contemporaneous variation, and lagged synchrony of time series is not recognized by this approach.

Subsequent hierarchical clustering was based on the resulting distance matrix, collecting all pairwise distances computed according to $1 - r_s$ as r_s is a measure of similarity. In hierarchical clustering, pairs of molecular formulae with the smallest distance in the distance matrix form clusters on a first level of the hierarchy. On higher hierarchical levels, smaller clusters merge into larger clusters building a tree-like structure often referred to and visualized as dendrograms, until all molecular formulae are agglomerated into a single cluster. Where the branches of two clusters merge into a bigger cluster is defined by the linkage method chosen for the cluster algorithm. Here, we applied the average linkage algorithm (Hahs-Vaughn 2016) that merges two clusters when their average distance falls below a rising threshold. A detailed description of hierarchical clustering and the most common linkage algorithms can be found in Legendre and Legendre (2012). In the final step, we decided on the optimal number of clusters or, in other words, defined where to cut the dendrogram, which is a crucial step for the resulting cluster composition and interpretation. For this purpose, we used the silhouette value (Rousseeuw 1987) that is calculated for each molecular formula as the average ratio of distances to members of its own cluster and the distance to members of its nearest-neighbor cluster. This silhouette value is a measure of how well a molecular formula matches its own cluster in relation to the nearest-neighbor cluster. For every molecular formula, the silhouette value can range from -1 to 1. A negative value indicates a bad attribution to a cluster. For every possible number of clusters "k," we calculated the average silhouette value (ASV) of all molecular formulae and chose the number of clusters "k" that maximizes the ASV, often referred to as silhouette coefficient for further analysis (Legendre and Legendre 2012).

A consensus time series was calculated for every cluster as the median intensity of all molecular formulae of that cluster for every point in time, respectively. For this, we included only time series of molecular formulae that had higher silhouette values than the ASV of the respective cluster. That way we obtained a better representation of the pattern, because the remaining subset was restricted to components that were most representative for the cluster. Conversely, the excluded components that were more distant from the cluster centroid than average were more likely to be derived from a unique origin like point sources or represent artifacts like contaminations.

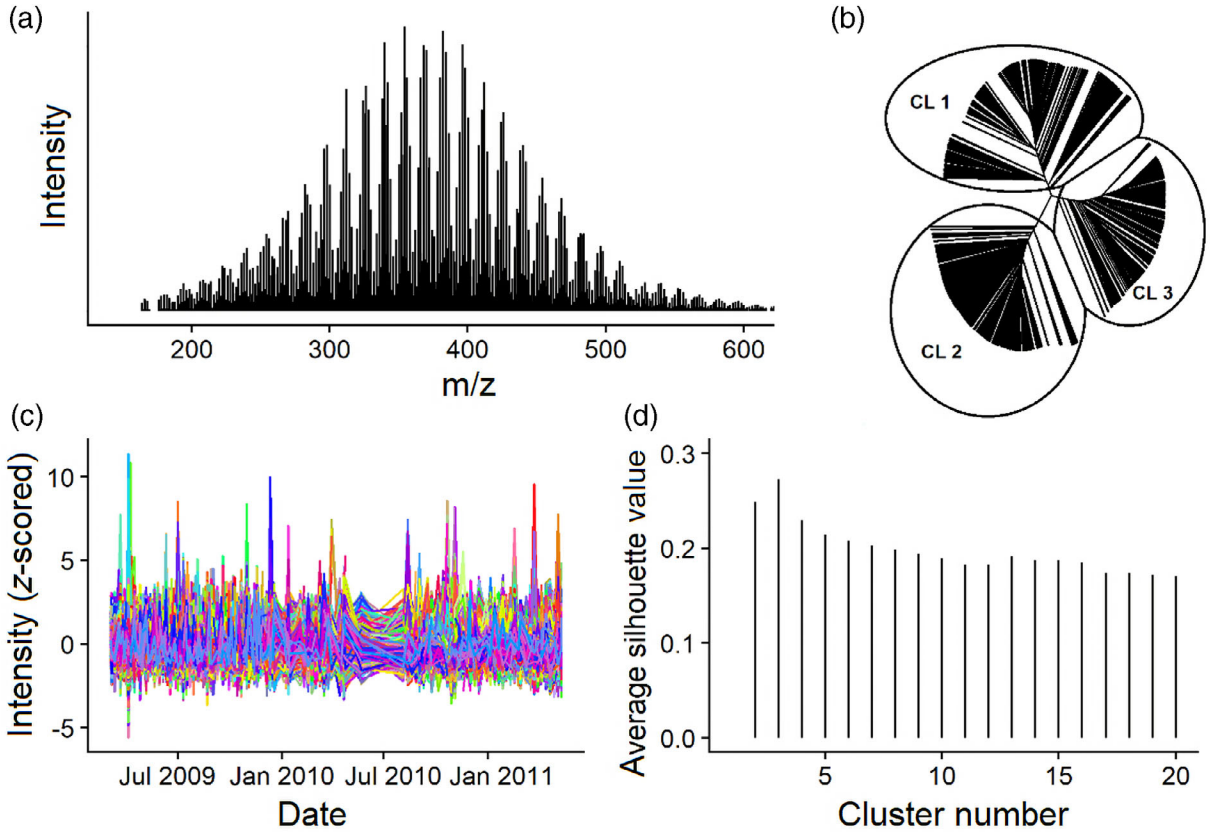

Fig. 2. Time series and clustering results of the Helgoland Roads molecular DOM time series. (**a**) Reconstructed average mass spectrum of Helgoland Roads DOM time series data, (**b**) unrooted dendrogram of the DOM time series (separation indicates the classification into three clusters: CL1 = cluster 1, CL2 = cluster 2, CL3 = cluster 3), (**c**) *z*-scored DOM time series of FT-ICR-MS signal intensities for 2109 molecular formulae, (**d**) the ASV for the range of cluster numbers 2–20.

Therefore, we decided not to include those for aggregating a consensus cluster time series.

We compared chemical properties across clusters using the above-determined final cluster representatives, considering the relative number of nitrogen, sulfur, and phosphorus of the molecular formulae, O/C, H/C ratios as well as average masses and the modified aromaticity index AI_{mod} (Koch and Dittmar 2006). In addition, we assessed the overlap of molecular formulae of the DOM clusters with molecular formulae from the NEqPIW reference sample, representing refractory deep ocean DOM, with t-peaks (Medeiros et al. 2016), which are molecular formula markers of terrigenous DOM, and with molecular formulae assigned to the "island of stability," which is also a measure for refractory marine DOM (Lechtenfeld et al. 2014).

Furthermore, we tested if the clusters resulting solely from synchronous dynamics can statistically be separated based on their chemical characteristics alone. For this, we used nonparametric classification trees (Hothorn et al. 2006) that, based on permutation tests, maximize separation of the response (here cluster assignments) with binary splits of the exploratory variables (here chemical parameters). In a nutshell, classification trees follow a top-down approach. They take a predictor variable and test for the best threshold of this predictor to split the data into two subsets. It recursively does this for each predictor and selects predictor and threshold that maximize between-cluster separation. For both subsets resulting from this split, the procedure is repeated until the optimal separation of clusters is reached. Classification trees have the advantage that they are nonparametric, but they are computationally expensive and tend to overfit when the trees become too large, so further splitting needs to be restricted based on reproducible and objective criteria. Here we stopped growing of the tree, if an additional split was statistically not significant ($p \geq 0.001$) based on the test described in Hothorn et al. (2006), so that the additional two resulting subsets would not improve the classification anymore.

Finally, we tested if the clusters found by maximizing ASV can be further divided into representative sub-clusters. For this, we only included the cluster representative formulae, which we considered a denoised representation of the cluster compounds. We again applied hierarchical

clustering with average linkage on this reduced formula subset, but this time did not determine the optimal cluster number by maximizing ASV. In contrast, we intended to split the original clusters into as many sub-clusters as possible. As during this process very small clusters emerge, for example, the segregation of a single formula from a cluster, we only interpreted clusters with a minimum of 20 members (molecular formulae) as valid sub-clusters.

Correlation of DOM clusters with environmental and microbiological data

To reveal potential connections of DOM clusters with environmental conditions and microbial communities, we performed correlation analyses, again using Spearman's rank correlation coefficient between the consensus DOM cluster time series and time series of environmental parameters, and prokaryote and phytoplankton abundances, including only pairwise complete data. Statistical significance of pairwise correlation coefficients was assessed via permutation tests, including the Benjamini–Hochberg correction for multiple testing (Benjamini and Hochberg 1995) for the final *p*-values. The Benjamini–Hochberg correction (control of false discovery rate) is less conservative than the classical Bonferroni correction (control of family-wise error rate). The results were visualized inside a special network assembly related to association networks in microbiology (Steele et al. 2011), but here with the DOM clusters forming the network centers. Prokaryote, phytoplankton, or environmental parameters significantly correlating with one or more of the clusters were linked radially to the respective cluster center. Because of its appearance, we refer to this network representation as "dandelion plot" in the following (https://icbm.de/komplsyst/helgoland-network). A two-dimensional version is displayed in a colored correlation matrix. We did not estimate correlations including time lags (i.e., the cross-correlation function) because of the unevenly distributed and in some cases even missing sampling points. Moreover, we abstained from a general additive model (Wood 2017) for the reconstruction of missing data, as the highly fluctuating concentrations even between consecutive sampling points potentially introduce unpredictable bias.

Ranking and database search of cluster representative DOM compounds

For each cluster, we ranked the molecular formulae by their silhouette value to assess how representative each formula is for the respective cluster (Supporting Information Data S1, Table S2 [Top 100]). For all cluster representative molecular formulae, we searched for matches with substances in the PubChem database (Kim et al. 2016) based on identical molecular formulae (https://pubchem.ncbi.nlm.nih.gov). We are aware that each molecular formula can represent multiple isomers with very different structures. Consequently, this database analysis was purely exploratory and, therefore, resultant statements are tentative. We also did not assess possible bias within the database toward substances in pharmaceutical or industrial usage or substances found within cultured or laboratory species. Therefore, we did not include a detailed statistical analysis of the molecular formula matches and instead only name noticeable accumulation of certain substance classes in relation to the identified DOM clusters.

Results

Time series of DOM molecular formulae

In total, we attributed 11,110 different molecular formulae to masses detected by FT-ICR-MS in any of the time series samples. After excluding all assigned isotopologues (formulae containing ^{13}C, ^{15}N, ^{34}S, ^{18}O) that are already represented by their respective ^{12}C, ^{14}N, ^{32}S, ^{16}O formulae and including only formulae detected in at least 50% of the time series samples, the final data set consisted of 2109 molecular formulae. The reconstructed mass spectrum (Fig. 2a) for the 2109 molecular formulae covered a mass range between 164 and 668 Da with a bell-shaped envelope of the intensity pattern (averaged intensities over time) typically observed for natural marine DOM (Zark and Dittmar 2018).

None of the considered molecular formulae contained phosphorus, 24% contained at least one nitrogen atom, and 7.5% included sulfur. The mean AI_{mod} was 0.3 with an interquartile range of 0.2, indicating that most of the molecular formulae represented aliphatic compounds (Koch and Dittmar 2006). The number of molecular formulae detected at the different time points was between 2050 and 2100, with less than 5% deviation and few distinct exceptions (Supporting Information Fig. S1a). Most of the 2109 molecular formulae were ubiquitous, that is, detected at all time points. Molecular formulae that were detected less frequently had generally low signal intensities (Supporting Information Fig. S1b,c).

The complete *z*-scored time series of all 2109 molecular formulae showed strong fluctuations even on a timescale of days (Fig. 2c). There was no overall visible trend. Instead, there was an accumulation of molecular formulae with increased intensities during certain time periods, especially in spring and autumn as well as in mid-summer 2009, which coincided with the occurrence of phytoplankton blooms (Teeling et al. 2012).

Clustering of DOM compounds

Based on Spearman's Rank correlation distance, the ASV was maximum for separation into three clusters (Fig. 2b,d) although the maximum value was low (ASV = 0.28). This is mainly due to the fact that during the hierarchical clustering all molecular formulae were assigned to one of the clusters, and therefore, all clusters included some molecular formulae with temporal dynamics strongly deviating from the cluster

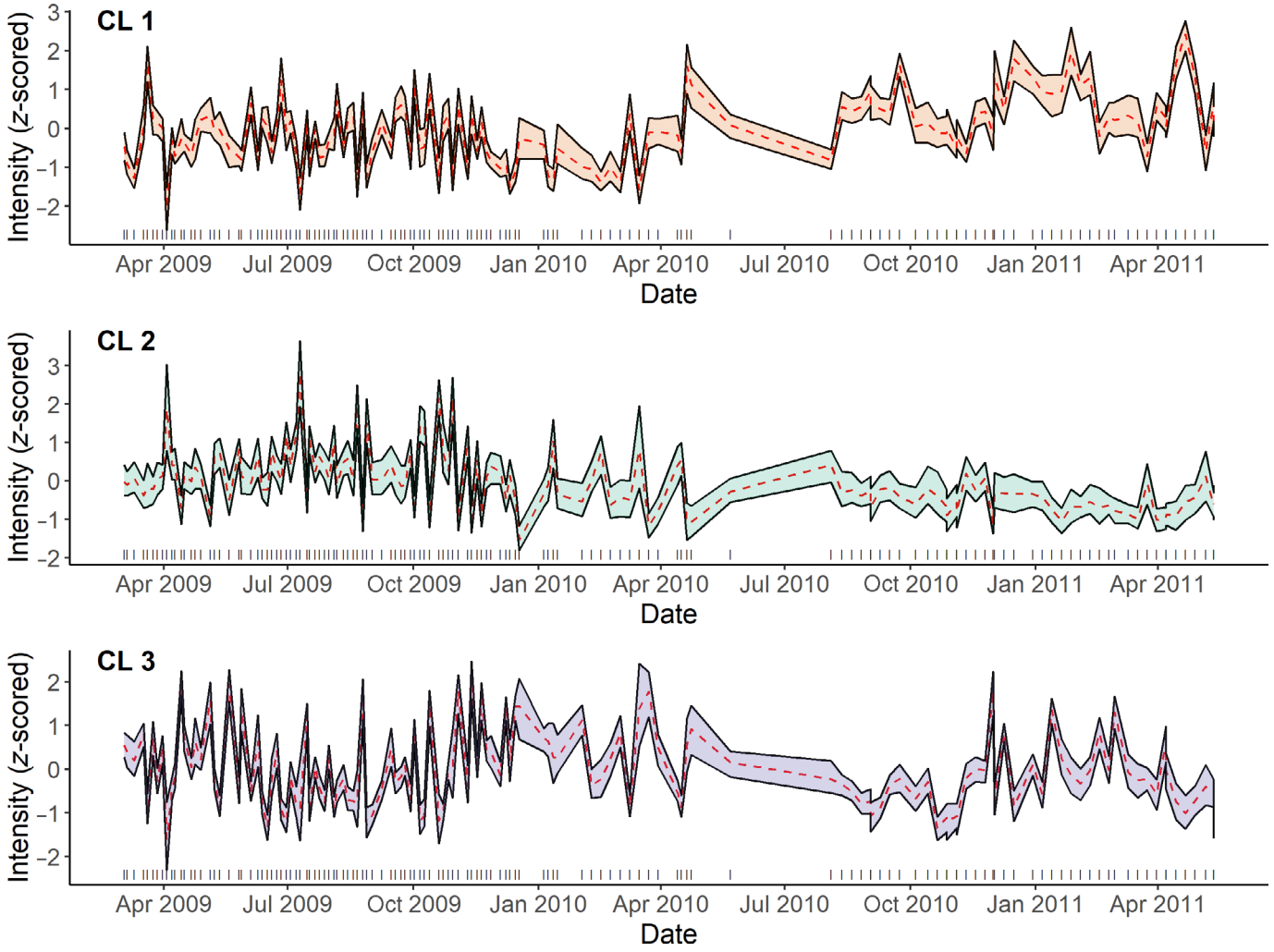

Fig. 3. Time series clusters reconstructed from Helgoland Roads molecular DOM time series data. Dotted red lines represent median values at each time point, colored areas highlight the related interquartile range. Calculations are based on molecular formulae exhibiting silhouette values above their respective cluster average (cluster 1 [CL1] = 546 of 835 formulae, cluster 2 [CL2] = 420 of 613 formulae, cluster 3 [CL3] = 426 of 661 formulae). Uneven tick marks on *x*-axis indicate sampling points.

mean. The three identified DOM clusters (cluster 1, cluster 2, and cluster 3) contained a similar number of molecular formulae each (n = 835, n = 613, and n = 661, respectively), so no cluster was overrepresented or underrepresented. Restriction to those molecular formulae that had higher silhouette values than the ASV of the respective cluster reduced the size of the clusters to 546, 420, and 426 for cluster 1, cluster 2, and cluster 3, respectively. Thus, the temporal dynamics represented by the three clusters (Fig. 3) integrated 66% of the 2109 molecular formulae, suggesting the existence of few major regulators that determine the bulk of DOM dynamics.

Most of the molecular formulae considered in the three confined clusters were detected at all time points (Supporting Information Fig. S2). Omnipresence of these compounds is consistent with the interpretation that few major regulators control their dynamics. The top 100 molecular formulae per cluster, ranked according to the silhouette value, are listed in Supporting Information Table S2, and the complete data of all 1392 formulae is supplied in Supporting Information Data S1.

The three clusters were clearly distinguishable based on the chemical composition of the contained molecular formulae (Fig. 4). The differential mass distribution is obvious from visual inspection of the reconstructed average mass spectra of the three clusters (Supporting Information Fig. S3). The molecular formulae of each of the three clusters were also clearly distinguishable by their elemental ratios, as visualized in their respective van Krevelen diagrams (Fig. 5). The separation of the molecular formulae into the three clusters that was purely based on their synchronous dynamics was largely reproduced by statistical analyses based on the associated chemical characteristics (Fig. 6). The classification tree based on the explanatory variables mass, O/C, and H/C ratios achieved a clear and statistically significant distinction between the three clusters. Around 90% of molecular formulae with masses < 432 and O/C

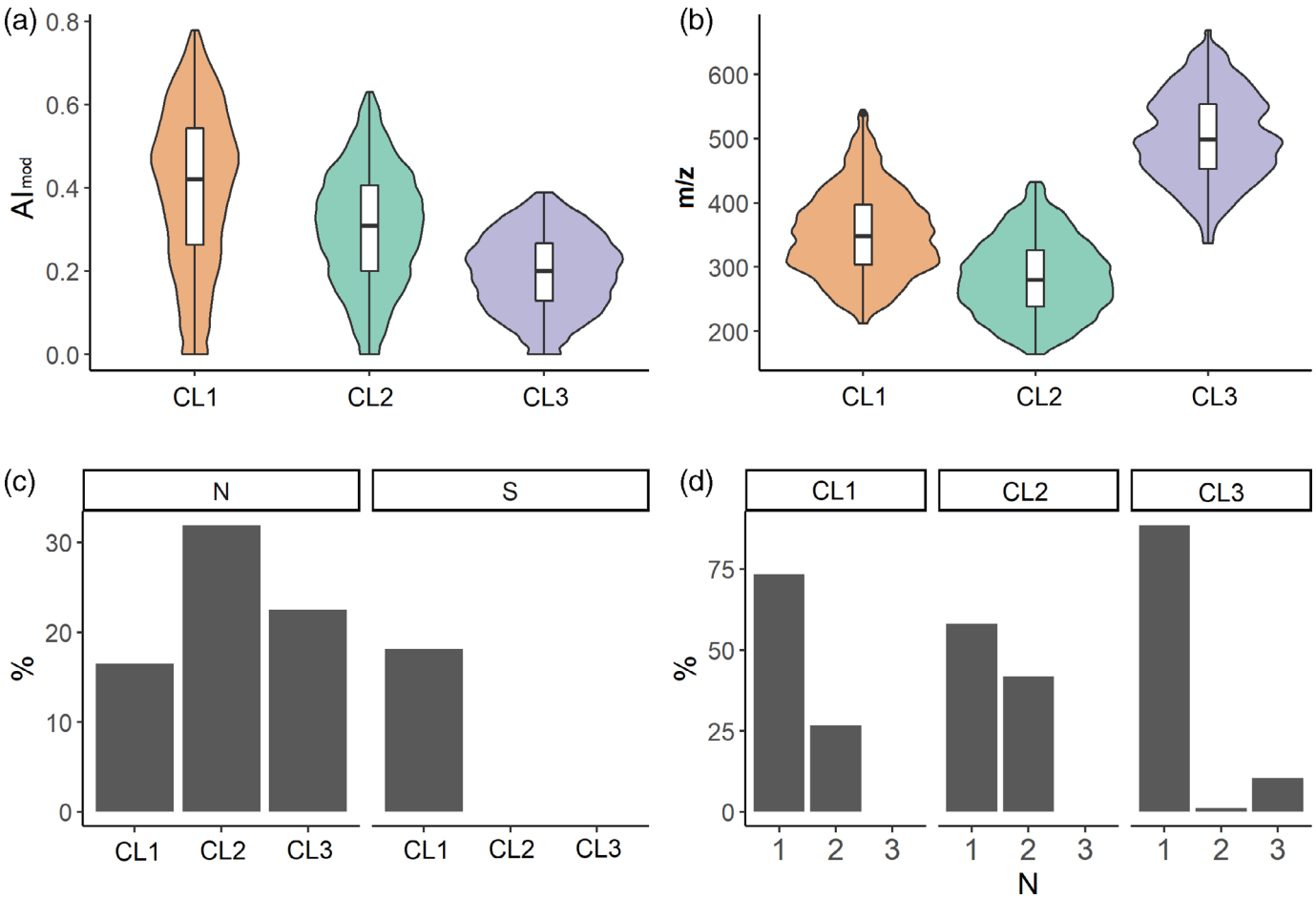

Fig. 4. Chemical properties of the three molecular DOM clusters. (**a**) Distribution of the modified aromaticity index (AI_{mod}) presented by violin plots and boxplots, (**b**) same as (**a**) but for molecular mass, (**c**) percentage of molecular formulae containing nitrogen and sulfur, (**d**) relative frequencies of number of nitrogen atoms observed for nitrogen-containing molecular formulae. CL1 = cluster 1, CL2 = cluster 2, CL3 = cluster 3.

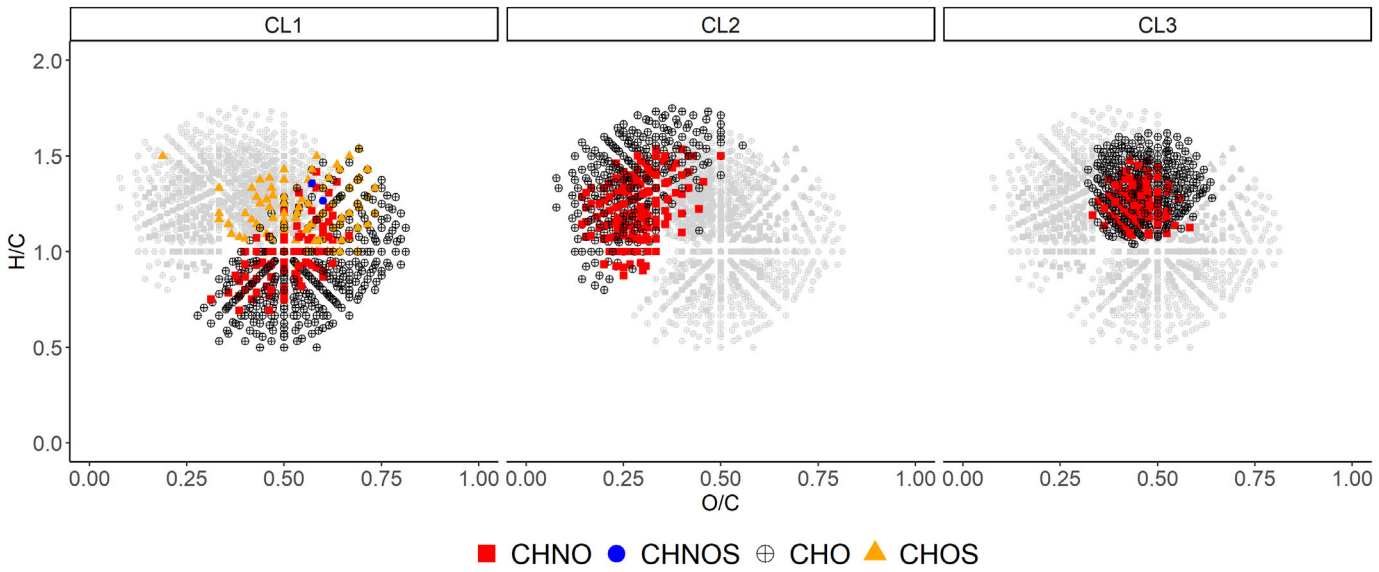

Fig. 5. Van Krevelen diagrams of the three molecular DOM clusters. Gray points depict all 1392 molecular formulae contained in any of the three clusters, colored symbols represent elemental composition of compounds of the respective cluster (CL1 = cluster 1, CL2 = cluster 2, CL3 = cluster 3).

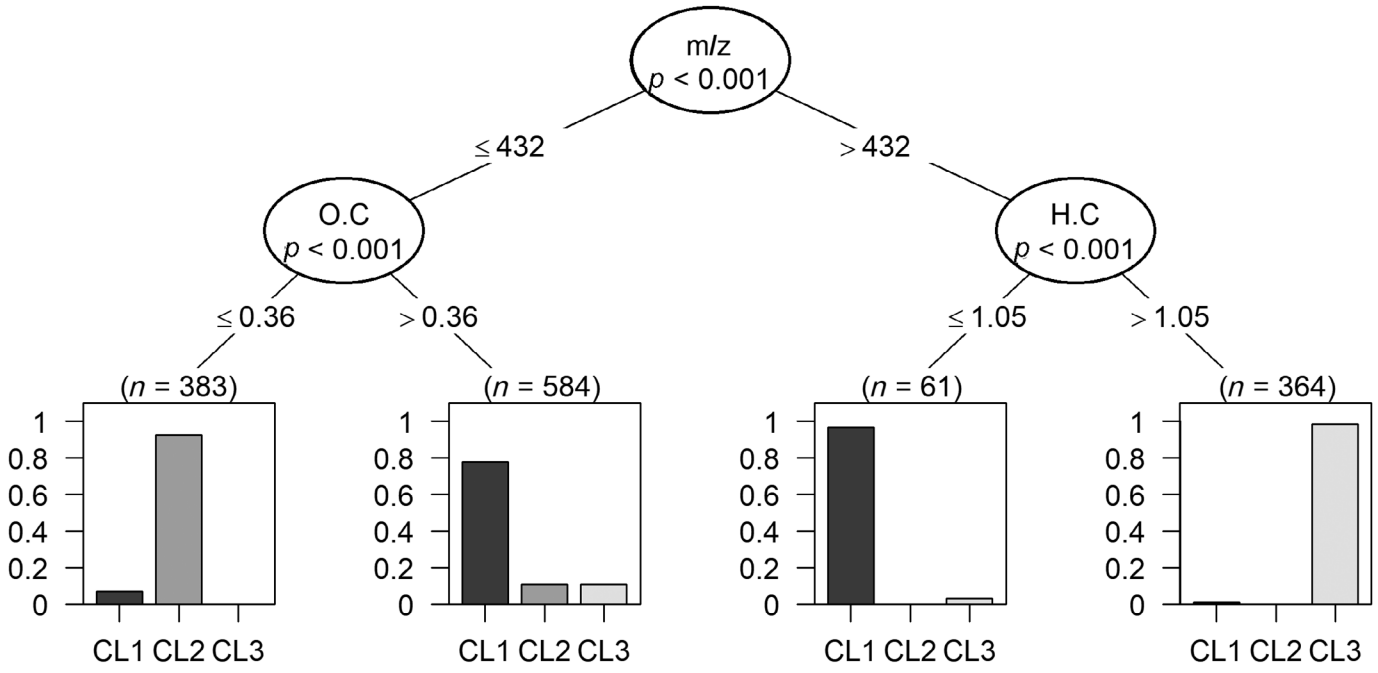

Fig. 6. Classification tree results dividing the three molecular DOM clusters based on chemical characteristics. A decision tree is a machine learning technique that uses binary splits of chemical parameters to optimally distinguish the DOM clusters (CL1 = cluster 1, CL2 = cluster 2, CL3 = cluster 3). The number "*n*" corresponds to the total number of molecular formulae, the *y*-axis in the barplot represents relative abundances, numbers on the branches at each split represent the threshold to split the data into two distinct subsets.

ratios < 0.36 were attributed to cluster 2. Almost 100% of molecular formulae with masses > 432 and H/C ratios > 1.05 were members of cluster 3. Cluster 1 contained almost 80% of molecular formulae with masses < 432 and O/C > 0.36 and almost 100% of molecular formulae with masses > 432 and H/C < 1.05. The additional chemical characteristics AI_{mod}, N, and S were tested for improved separation, but did not yield any additional statistically significant split of the decision tree.

For all clusters, there was a considerable but differential overlap with molecular formulae detected in the DOM reference sample from the North Pacific (cluster 1 = 60%, cluster 2 = 75%, cluster 3 = 90%). Molecular formulae assigned to t-peaks (Medeiros et al. 2016) and the "island of stability" (Lechtenfeld et al. 2014) showed characteristic distributions over the three clusters (Supporting Information Fig. S4). The three clusters were also characterized by distinct correlation patterns with environmental and microbiological parameters (Fig. 7; Supporting Information Fig. S5, Data S2). Further partitioning of the three clusters yielded a total of 11 sub-clusters, each characterized by distinct chemical properties (Fig. 8; Supporting Information Figs. S6, S7).

Discussion

Cluster properties and relation to environmental and microbiological parameters

Cluster 1 related to terrestrial origin

Cluster 1 was mainly characterized by highly oxygenated unsaturated molecular formulae and low H/C ratios of nitrogen-bearing molecular formulae (Fig. 5). It was also the only cluster including molecular formulae with $AI_{mod} \geq 0.67$ (Fig. 4), providing unequivocal evidence for the presence of condensed aromatic structures (Koch and Dittmar 2006). The predominantly highly unsaturated to aromatic character of the molecular formula in cluster 1 suggests a terrestrial origin (Koch and Dittmar 2006; Schmidt et al. 2009; Lu et al. 2015). A potential terrestrial origin of this cluster is consistent with the highest overlap with t-peaks (Medeiros et al. 2016; Supporting Information Fig. S4) that represent terrigenous molecular formulae. Cluster 1 also shared a considerable number of molecular formulae that form the "island of stability" (Lechtenfeld et al. 2014; Supporting Information Fig. S4), which has been proposed to represent the refractory core of marine DOM.

Compared to the other two clusters, cluster 1 included by far the highest number of sulfur-containing compounds, accounting for a proportion of 15% of all molecular formulae in cluster 1 (Figs. 4,5). Potential explanations for the high relative abundance of sulfur-containing compounds include introduction of sulfurized DOM from sulfidic environments (Schmidt et al. 2009; Gomez-Saez et al. 2017), excretions from algae (Cunha and Grenha 2016), release of cell wall components of archaea (Deatherage and Cookson 2012), wastewater input, and agricultural discharge (Gonsior et al. 2011; Wagner et al. 2015). Over 98% of the sulfur-containing molecular formulae in cluster 1 had an O/S ratio ≥ 4, which has been suggested as indicative for organosulfates enriched in humic-rich river DOM

(Lu et al. 2015). In general, riverine DOM is not enriched in sulfur compared to ocean DOM (Riedel et al. 2016), but anthropogenically impacted rivers show enrichments in sulfur-containing formulae (Wagner et al. 2015). For the Delaware estuary, increasing concentrations of dissolved organic sulfur were observed along a transect from the ocean to the river (Osterholz et al. 2016*a*). In tidally influenced coastal areas such as the coastal North Sea, freshwater inflow is often associated with input from sulfidic pore water discharging from submerged sediments that is enriched in sulfur-containing DOM (Seidel et al. 2014).

The sampling area around Helgoland represents a transition zone between coastal waters influenced mainly by the rivers Elbe and Weser and the marine waters of the North Sea (Hickel 1998; Fig. 1). Rivers are a well-recognized contributor of terrestrial DOM to the ocean, reflected in decreasing DOC concentrations as well as characteristic changes in the molecular DOM composition along increasing salinity gradients (Abdulla et al. 2013; Medeiros et al. 2015; Osterholz et al. 2016*b*). The strongest support for a terrestrial origin of cluster 1 is provided by the highly statistically significant negative correlation with salinity and the highly statistically significant positive correlation with DOC (Fig. 7, https://icbm.de/komplsyst/helgoland-network), reflecting that molecular formulae of cluster 1 were enriched at time points when freshwater inflow associated with increased DOC concentrations was high. In addition, cluster 1 was negatively correlated with concentrations of inorganic nitrogen species (nitrate, nitrite, DIN) while it was positively correlated with DON and Chl *a*, indicating that times of increased river discharge coincided with characteristic features of phytoplankton blooms, such as depletion of nutrients, enrichments of DON and Chl *a*.

Several phytoplankton taxa were positively correlated with cluster 1 (Fig. 7, Supporting Information Fig. S5). Most of these positive correlations were shared with cluster 2, indicating that the occurrence of these specific taxa coincided with river water inflow (cluster 1) and the conditions controlling cluster 2. There was no shared correlation with cluster 3; in contrast, all taxa positively correlating with cluster 1 showed no or anticorrelation with cluster 3. Unique positive correlations for cluster 1 were observed for the diatoms *Asterionellopsis glacialis*, *Ditylum brightwellii*, and *Bacillaria paxillifera*. All these species are considered tolerant to salinity fluctuations or reduced salinity (Rijstenbil et al. 1989). Cluster 1 did not show strong correlations with any of the prokaryotic taxa included in this time series study, only a moderate anticorrelation with the heterotroph Gram-negative *Marinoscillum* (phylum Bacteroidetes) targeted as CYT-734. Anticorrelation of cluster 1 with this marine genus is consistent with a terrestrial origin of this cluster. A likely reason why we did not detect positive correlations with any of the prokaryotes is that the data set on prokaryote abundance mainly includes typical marine taxa, while characteristic terrestrial organisms are not covered systematically (Fuchs et al. 2016; Supporting Information Table S1).

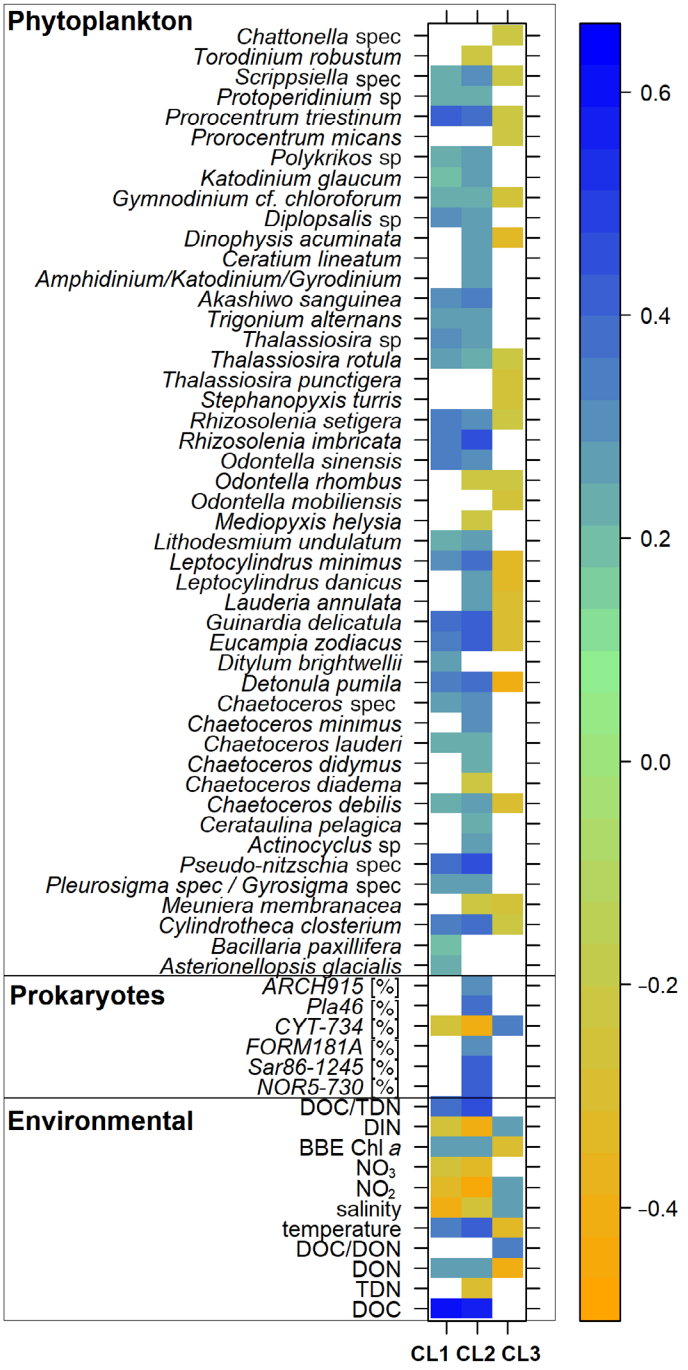

Fig. 7. Heatmap of Spearman correlations between time series of DOM clusters and phytoplankton, prokaryote, and environmental data. Nonsignificant ($p > 0.05$) correlations are displayed in white. All taxa or environmental parameters not included in this plot exhibited no significant correlation after multiple testing corrections. CL1 = cluster 1, CL2 = cluster 2, CL3 = cluster 3. Exact numbers of displayed data are summarized in Supporting Information Data S2.

Possible substances included in the PubChem database that matched the molecular formulae in CL 1 include hydroxycinnamic acids, for example, $C_{13}H_{12}O_9$ (caftaric acid), $C_{13}H_{12}O_8$

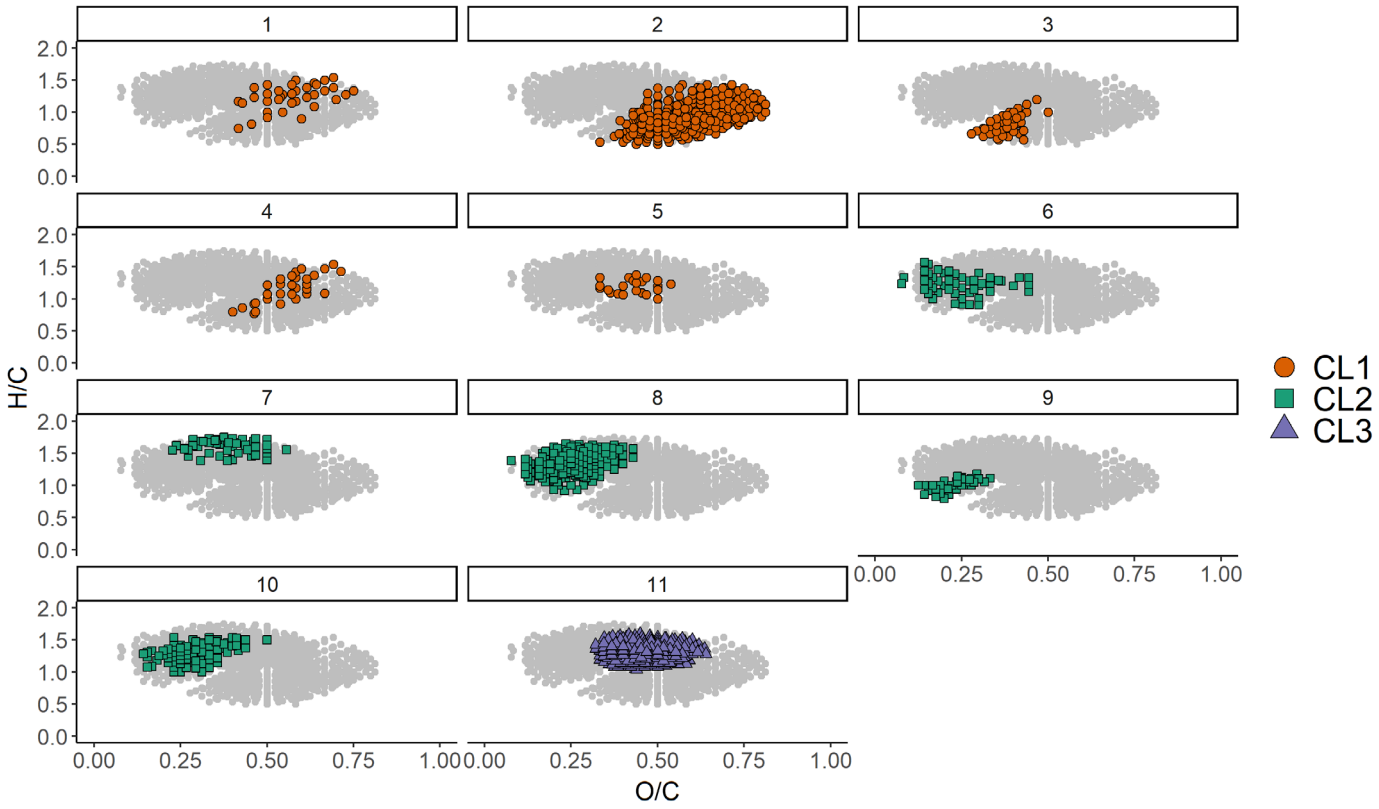

Fig. 8. Van Krevelen diagram of refined cluster partitioning (splitting) of the three molecular DOM clusters into 11 clusters. Gray points depict all 1392 molecular formulae contained in any of the clusters (same as for the three clusters; Fig. 5), colored symbols represent affiliation with the three original clusters (CL1 = cluster 1 [1–5], CL2 = cluster 2 [6–10], CL3 = cluster 3 [11]).

(coutaric acid, caffeoylmalic acid), and $C_{14}H_{14}O_9$ (fertaric acid), which are found in grapes and plant tissues (Vrhovšek 1998). Other molecular formulae like $C_{16}H_{18}O_9$ (chlorogenic acid) are intermediates in lignin biosynthesis (Boerjan et al. 2003). Furthermore, cluster 1 exhibited matches with secondary metabolites of vascular plants like the flavonoids catechin ($C_{15}H_{14}O_6$), quercetin ($C_{15}H_{10}O_7$), or kaempferol ($C_{15}H_{10}O_6$). We are aware that each molecular formula integrates thousands of potential structural isomers and we do not claim that the detected molecular formulae represent their respective matches in the database. Nonetheless, the fact that many of the database matches are related to higher plants is consistent with the proposed terrestrial origin of cluster 1.

Cluster 2 related to phytoplankton blooms and microbial activity

All molecular formulae of cluster 2 had an $AI_{mod} < 0.67$ (Fig. 4a), indicating the absence of compounds with purely condensed aromatic structures. Compared to cluster 1, the molecular formulae of cluster 2 were less oxygenated and more saturated (Fig. 5). Cluster 2 included the highest proportion of nitrogen-containing molecular formulae (Fig. 4) and a significantly higher number ($p < 0.001$, permutative resampling [10,000 repetitions]) of peptide-like molecular formulae (Rivas-Ubach et al. 2018) compared to the other two clusters (cluster 1 = 30, cluster 2 = 55, cluster 3 = 30). The more aliphatic character and the relative enrichment in molecular formulae containing one or more nitrogen atoms are characteristic for freshly produced phytoplankton DOM (Medeiros et al. 2015). A recent autochthonous origin of cluster 2 is supported by the small overlap with terrestrial compounds represented by the t-peaks (Medeiros et al. 2016) and with molecular formulae forming the "island of stability" of refractory marine compounds (Lechtenfeld et al. 2014; Supporting Information Fig. S4).

During phytoplankton blooms, inorganic nutrients typically become depleted and this is reflected in negative correlations of cluster 2 with concentrations of all dissolved nitrogen species (nitrate, nitrite, DIN, TDN; Fig. 7). Cluster 2 showed positive correlations with DOC, DON, and Chl *a* concentrations, providing strong evidence for a relation of cluster 2 with phytoplankton blooms and the associated release of microbial-derived DOM. Cluster 2 was also correlated to temperature, which is consistent with a predominant occurrence of plankton blooms and enhanced microbial activity during warmer seasons.

In support of a planktonic origin, cluster 2 was the cluster with the highest number of correlations with individual

phytoplankton taxa (Fig. 7). Positive correlations unique for this cluster included the diatoms *Cerataulina pelagica*, *Chaetoceros didymus*, *Chaetoceros minimus*, *Actinocyclus* sp., *Leptocylindrus danicus*, and *Lauderia annulata*, as well as the dinoflagellates *Amphidinium/Katodinium/Gyrodinium* and the toxic *Dinophysis acuminata*. We note that cluster 1 also showed positive correlations with phytoplankton species (Fig. 7; Supporting Information Fig. S5), but under different environmental conditions. This might indicate that an alternation of phytoplankton subcommunities can take place at Helgoland influenced by river discharge. Cluster 2 was also the cluster with by far the most positively and statistically significant correlations with the abundance of specific prokaryotes (Fig. 7, dandelion plot: https://icbm.de/komplsyst/helgoland-network), including those targeted by Sar86-1245 (Gammaproteobacteria), NOR5-730 (Gammaproteobacteria), Pla46 (Planctomycetes), ARCH915 (Archaea), and FORM181A (Flavobacteria). Some of these bacteria have been observed during and after phytoplankton blooms at the same sampling site off Helgoland, when microbial communities exhibit distinct succession of individual taxa (Teeling et al. 2012).

The molecular formulae most representative for the temporal dynamics of cluster 2 match PubChem entries of plant hormones, for example, $C_{15}H_{20}O_4$ which might correspond to abscisic acid, a growth-inhibiting phytohormone found not only in plants but also in microalgae and macroalgae (Stirk et al. 2009; Guajardo et al. 2016). Molecular formulae of intermediates of abscisic acid biosynthesis like $C_{15}H_{20}O_3$ (abscisic aldehyde) and $C_{15}H_{22}O_3$ (xanthoxin) were all highly ranked matches for cluster 2, as well as potential transformation products $C_{15}H_{20}O_5$ (phaseic acid) and $C_{15}H_{22}O_5$ (dihydrophaseic acid). The nitrogen-containing molecular formulae $C_{12}H_{13}NO_2$ that was detected at lower signal intensities matched indole-3-butyric acid, a growth inducing hormone of the auxin family widely distributed in plants and algae (Piotrowska-Niczyporuk and Bajguz 2014). Auxin synthesis involves the amino acid tryptophan (Amin et al. 2015) which derivatives also had molecular formula matches in cluster 2 as hydroxytryptophan or kynurenine ($C_{10}H_{12}N_2O_3$), as well as other amino acids such as *N*-acetyltyrosine ($C_{11}H_{13}NO_4$) or the dipeptide glycyl-L-tyrosine ($C_{11}H_{14}N_2O_4$). Other hormone matches such as melatonin ($C_{13}H_{16}N_2O_2$) are known to act against oxidative stress in plants and have also been found in algae (Arnao and Hernández-Ruiz 2006). Furthermore, the formula matches of cluster 2 include domoic acid ($C_{15}H_{21}NO_6$) a neurotoxin produced by algae and known to cause amnesic shellfish poisoning (Delegrange et al. 2018).

Bacteria also produce plant hormones to induce growth of diatoms (Amin et al. 2015; Ajani et al. 2018), and there is some evidence for the production of auxins by marine microalgae (Mazur et al. 2001; Labeeuw et al. 2016). In a recent study, the molecular formulae of the auxin indole-3-acetic acid and its precursor tryptophan have been detected in the exometabolome of pure cultures of *Dinoroseobacter shibae*, a representative of the globally abundant and often plankton-bloom associated marine Roseobacter group (Wienhausen et al. 2017). Overall, in contrast to cluster 1, with database matches predominantly associated to higher land plants, cluster 2 includes matches with compounds that are also produced by marine organisms. A microbial origin of many of the database matches is consistent with the proposed relation of cluster 2 with increased microbial activity in the course of phytoplankton blooms.

Cluster 3 representing the refractory marine DOM background

Cluster 3 was significantly different from the other two clusters in that it comprised molecular formulae with the highest masses and the lowest AI_{mod} (Fig. 4). The elemental ratios O/C and H/C covered a comparably small and well confined area in the van Krevelen diagram (Fig. 5), matching typical characteristics of carboxyl-rich alicyclic molecules (CRAM), that have been proposed as a ubiquitous constituent of natural DOM (Hertkorn et al. 2006). Cluster 3 showed the highest overlap with refractory molecular formulae found in deep Pacific DOM (90% of the molecular formulae in cluster 3 were also found in the NEqPIW reference), and the highest overlap with the "island of stability" (Lechtenfeld et al. 2014), indicating that this cluster represents the refractory marine DOM background. A ubiquitous refractory marine DOM pool has been proposed as the ultimate result of the multiple production, mixing, degradation, and transformation processes acting on DOM (Koch et al. 2005; Zark and Dittmar 2018). This hypothesis implies that the refractory DOM background does not contain any source-specific signature. The finding that cluster 3 contained a negligible number (only two matched) of the terrigenous compounds represented by t-peaks (Medeiros et al. 2016) is consistent with such a scenario and the interpretation of cluster 3 as representing the marine DOM background.

Cluster 3 exhibited statistically significant positive correlations with salinity, DOC/DON ratio, as well as with concentration of DIN, and anticorrelations with temperature and concentrations of Chl *a* and DON (Fig. 7). High salinity, low temperature, and high inorganic nutrient concentrations are typical for winter conditions when the North Sea around Helgoland is well mixed. Increased DOC/DON ratios are a typical feature of refractory marine DOM compared to freshly produced marine DOM (Hansell and Carlson 2015), providing further support for cluster 3 representing the marine DOM background. Low concentrations of Chl *a* and DON are consistent with the absence of phytoplankton growth and associated recent production of fresh DOM during winter. Accordingly, cluster 3 did not show any positive correlation to the abundance of phytoplankton and prokaryotes, except for a moderate positive correlation with the marine bacterium *Marinoscillum*, targeted as CYT-734.

Matches of cluster 3 molecular formulae in the PubChem database include a wide range of terpenoids, with a mentionable number of iridoid glycosides like $C_{24}H_{32}O_{10}$ (acevaltrate), $C_{22}H_{28}O_{10}$ (davisioside), $C_{26}H_{38}O_{12}$ (dihydrofoliamenthin), $C_{27}H_{38}O_{12}$ (sioriside), $C_{24}H_{30}O_{11}$ (harpagoside), $C_{23}H_{34}O_{11}$ (hookerinoid B), $C_{25}H_{36}O_{12}$ (nemoroside). Terpenoids are structurally very similar to proposed structures of CRAM (Hertkorn et al. 2006; Lam et al. 2007). Thus, as for the other two clusters, prominent PubChem database matches are consistent with the proposed origin and character of cluster 3. Iridoid glycoside like substances can become toxic in their metabolization (Yamane et al. 2010) making them unattractive to most microorganisms and additionally support the refractory character of this cluster.

Beyond the three-cluster partitioning

The three-cluster partitioning yielded an optimal separation of the data as indicated by the ASV (Fig. 2d) and the resulting three clusters were clearly distinguishable with respect to the chemical characteristics of their members as well as their proposed origin. A deeper partitioning into sub-clusters could provide valuable information on joint dynamics of smaller groups of compounds, that is, reveal DOM dynamics with higher molecular resolution.

Interestingly, while cluster 1 and cluster 2 split up into five sub-clusters each, cluster 3 persisted as a stable cluster. This corroborates the interpretation that cluster 3 represents a homogenous mixture of refractory marine background DOM, with all compounds exhibiting similarly low temporal dynamics that is not influenced by sporadic events like freshwater inflow or phytoplankton blooms. For all 11 sub-clusters we tested if they also achieved the distinct separation of chemical characteristics observed for the three clusters (Figs. 4, 5; Supporting Information Fig. S3). All 10 sub-clusters of cluster 1 and cluster 2 showed a clear separation based on their elemental O/C and H/C ratios in the van Krevelen space (Fig. 8). This is a remarkable finding, as it demonstrates that by statistical analysis of DOM time series, temporal dynamics of DOM compounds can be linked to distinct chemical composition. Thus, without knowing the exact structures of the molecular formulae, we can infer environmental behavior. The robust separation of compounds with differential chemical composition achieved by the hierarchical clustering was confirmed by classification trees purely based on information derived from molecular formulae (Supporting Information Fig. S6). Differences in nitrogen content were identified as an important factor for a deeper classification of substructures in cluster 1 and cluster 2. The 11 clusters also exhibited distinct mass distributions (Supporting Information Fig. S7), although the separation by mass was not as pronounced as for the three clusters.

Conclusion

The multitude of compounds that make up DOM are not only highly diverse in molecular composition but also exhibit very different dynamics. Here, we demonstrated that basic statistical approaches such as cluster and correlation analysis are valuable tools to identify synchronous dynamics and to reveal distinct coherencies between DOM molecular formulae and environmental conditions influencing the dynamics. The analyses performed in this study are not limited to time series data but can also be applied to deciphering dynamics and respective controls in studies of spatial gradients.

For the DOM time series presented here, we provide conclusive evidence that the DOM pool included three major groups of compounds that were clearly constrained based on their shared temporal dynamics. Although the time series analysis did not include any information on the nature of the DOM compounds, the resulting DOM clusters were clearly distinguishable based on their chemical composition. Correlations with abiotic and biotic environmental parameters were consistent with potential sources and history of the individual clusters.

Disclosing the structural identity of molecular formulae detected in DOM is still a major challenge. Targeted analysis of specific compounds is restricted to few known substances while the exact structure of the majority of DOM molecules remains uncharacterized. Assigning molecular formulae to matches in databases such as PubChem remains tentative. Our study shows that by combining information on dynamics and potential origin and transformation history with information on characteristics of respective database matches, structural identity of individual compounds can be narrowed down. Such confined tentative assignments can then be tested by more targeted studies, including chemical analytics (e.g., fragmentation experiments in FT-ICR-MS), biological (cultivation experiments), and environmental studies (specific conditions or gradients).

References

Abdulla, H. A., E. C. Minor, R. F. Dias, and P. G. Hatcher. 2013. Transformations of the chemical compositions of high molecular weight DOM along a salinity transect: Using two dimensional correlation spectroscopy and principal component analysis approaches. Geochim. Cosmochim. Acta **118**: 231–246. doi:10.1016/j.gca.2013.03.036

Ajani, P. A., T. Kahlke, N. Siboni, R. Carney, S. A. Murray, and J. R. Seymour. 2018. The microbiome of the cosmopolitan diatom *Leptocylindrus* reveals significant spatial and temporal variability. Front. Microbiol. **9**: 2758. doi:10.3389/fmicb.2018.02758

Amin, S. A., and others. 2015. Interaction and signalling between a cosmopolitan phytoplankton and associated bacteria. Nature **522**: 98–101. doi:10.1038/nature14488

Amon, R. M. W., H.-P. Fitznar, and R. Benner. 2001. Linkages among the bioreactivity, chemical composition, and diagenetic state of marine dissolved organic matter. Limnol. Oceanogr. **46**: 287–297.

Arnao, M. B., and J. Hernández-Ruiz. 2006. The physiological function of melatonin in plants. Plant Signal. Behav. **1**: 89–95.

Azam, F. 1998. Microbial control of oceanic carbon flux: The plot thickens. Science **280**: 694–696. doi:10.1126/science.280.5364.694

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. Roy. Stat. Soc. B **57**: 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

Boerjan, W., J. Ralph, and M. Baucher. 2003. Lignin biosynthesis. Annu. Rev. Plant Biol. **54**: 519–546. doi:10.1146/annurev.arplant.54.031902.134938

Bray, J. R., and J. T. Curtis. 1957. An ordination of the upland forest communities of Southern Wisconsin. Ecol. Monogr. **27**: 325–349.

Cunha, L., and A. Grenha. 2016. Sulfated seaweed polysaccharides as multifunctional materials in drug delivery applications. Mar. Drugs **14**: 42. doi:10.3390/md14030042

D'Andrilli, J., T. Dittmar, B. P. Koch, J. M. Purcell, A. G. Marshall, and W. T. Cooper. 2010. Comprehensive characterization of marine dissolved organic matter by Fourier transform ion cyclotron resonance mass spectrometry with electrospray and atmospheric pressure photoionization. Rapid Commun. Mass Spectrom. **24**: 643–650.

Deatherage, B. L., and B. T. Cookson. 2012. Membrane vesicle release in bacteria, eukaryotes, and archaea: A conserved yet underappreciated aspect of microbial life. Infect. Immunol. **80**: 1948–1957.

Delegrange, A., A. Lefebvre, F. Gohin, L. Courcot, and D. Vincent. 2018. *Pseudo-nitzschia* sp. diversity and seasonality in the southern North Sea, domoic acid levels and associated phytoplankton communities. Estuar. Coast. Shelf Sci. **214**: 194–206. doi:10.1016/j.ecss.2018.09.030

Dittmar, T., B. P. Koch, N. Hertkorn, and G. Kattner. 2008. A simple and efficient method for the solid-phase extraction of dissolved organic matter (SPE-DOM) from seawater. Limnol. Oceanogr. Methods **6**: 230–235. doi:10.4319/lom.2008.6.230

Flerus, R. O., J. Lechtenfeld, B. P. Koch, S. L. McCallister, P. Schmitt-Kopplin, R. Benner, K. Kaiser, and G. Kattner. 2012. A molecular perspective on the ageing of marine dissolved organic matter. Biogeosciences **9**: 1935–1955. doi:10.5194/bg-9-1935-2012

Fuchs, B. M., C. M. Bennke, G. Reintjes, M. Kassabgy, and R. I. Amann. 2016. Microbial community composition and bacterioplankton at time series station Helgoland Roads, North Sea. PANGAEA.

Galletti, Y., M. Gonnelli, S. Retelletti Brogi, S. Vestri, and C. Santinelli. 2019. DOM dynamics in open waters of the Mediterranean Sea: New insights from optical properties. Deep Sea Res Part I **144**: 95–114. doi:10.1016/j.dsr.2019.01.007

Gomez-Saez, G. V., A. M. Pohlabeln, A. Stubbins, C. M. Marsay, and T. Dittmar. 2017. Photochemical alteration of dissolved organic sulfur from sulfidic porewater. Environ. Sci. Technol. **51**: 14144–14154.

Gonsior, M., and others. 2011. Molecular characterization of effluent organic matter identified by ultrahigh resolution mass spectrometry. Water Res. **45**: 2943–2953. doi:10.1016/j.watres.2011.03.016

Green, N. W., E. M. Perdue, G. R. Aiken, K. D. Butler, H. Chen, T. Dittmar, J. Niggemann, and A. Stubbins. 2014. An intercomparison of three methods for the large-scale isolation of oceanic dissolved organic matter. Mar. Chem. **161**: 14–19.

Guajardo, E., J. A. Correa, and L. Contreras-Porcia. 2016. Role of abscisic acid (ABA) in activating antioxidant tolerance responses to desiccation stress in intertidal seaweed species. Planta **243**: 767–781.

Hahs-Vaughn, D. L. 2016. Applied multivariate statistical concepts, 1st ed. Taylor & Francis.

Hansell, D., C. Carlson, D. J. Repeta, and R. Schlitzer. 2009. Dissolved organic matter in the ocean: A controversy stimulates new insights. Oceanography **22**: 202–211. doi:10.5670/oceanog.2009.109

Hansell, D. A., and C. A. Carlson. 2015. Biogeochemistry of marine dissolved organic matter, 2nd ed. Academic Press.

Hawkes, J. A., N. Radoman, J. Bergquist, M. B. Wallin, L. J. Tranvik, and S. Löfgren. 2018. Regional diversity of complex dissolved organic matter across forested hemiboreal headwater streams. Sci. Rep. **8**: 16060. doi:10.1038/s41598-018-34272-3

Hedges, J. I. 1992. Global biogeochemical cycles: Progress and problems. Mar. Chem. **39**: 67–93.

Hertkorn, N., R. Benner, M. Frommberger, P. Schmitt-Kopplin, M. Witt, K. Kaiser, A. Kettrup, and J. I. Hedges. 2006. Characterization of a major refractory component of marine dissolved organic matter. Geochim. Cosmochim. Acta **70**: 2990–3010.

Hickel, W. 1998. Temporal variability of micro- and nanoplankton in the German Bight in relation to hydrographic structure and nutrient changes. ICES J. Mar. Sci. **55**: 600–609.

Hothorn, T., K. Hornik, and A. Zeileis. 2006. Unbiased recursive partitioning: A conditional inference framework. J. Comput. Graph. Stat. **15**: 651–674.

Jørgensen, L., C. A. Stedmon, M. A. Granskog, and M. Middelboe. 2014. Tracing the long-term microbial production of recalcitrant fluorescent dissolved organic matter in seawater. Geophys. Res. Lett. **41**: 2481–2488.

Kim, S., and others. 2016. PubChem substance and compound databases. Nucleic Acids Res. **44**: 1202–1213. doi:10.1093/nar/gkv951

Koch, B. P., M. Witt, R. Engbrodt, T. Dittmar, and G. Kattner. 2005. Molecular formulae of marine and terrigenous dissolved organic matter detected by electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. Geochim. Cosmochim. Acta **69**: 3299–3308.

Koch, B. P., and T. Dittmar. 2006. From mass to structure: An aromaticity index for high-resolution mass data of natural organic matter. Rapid Commun. Mass Spectrom. **20**: 926–932.

Kujawinski, E. B., K. Longnecker, K. L. Barott, R. J. M. Weber, and M. C. Kido Soule. 2016. Microbial community structure affects marine dissolved organic matter composition. Front. Mar. Sci. **3**: 45. doi:10.3389/fmars.2016.00045

Labeeuw, L., J. Khey, A. R. Bramucci, H. Atwal, A. P. de la Mata, J. Harynuk, and R. J. Case. 2016. Indole-3-acetic acid is produced by *Emiliania huxleyi* coccolith-bearing cells and triggers a physiological response in bald cells. Front. Microbiol. **7**: 828. doi:10.3389/fmicb.2016.00828

Lam, B., A. Baer, M. Alaee, B. Lefebvre, A. Moser, A. Williams, and A. J. Simpson. 2007. Major structural components in freshwater dissolved organic matter. Environ. Sci. Technol. **41**: 8240–8247.

Lechtenfeld, O. J., G. Kattner, R. Flerus, S. L. McCallister, P. Schmitt-Kopplin, and B. P. Koch. 2014. Molecular transformation and degradation of refractory dissolved organic matter in the Atlantic and Southern Ocean. Geochim. Cosmochim. Acta **126**: 321–337.

Leefmann, T., S. Frickenhaus, and B. P. Koch. 2019. UltraMassExplorer: A browser-based application for the evaluation of high-resolution mass spectrometric data. Rapid Commun. Mass Spectrom. **33**: 193–202.

Legendre, P., and L. Legendre. 2012. Numerical ecology, 3rd ed. Elsevier.

Lu, Y., X. Li, R. Mesfioui, J. E. Bauer, R. M. Chambers, E. A. Canuel, and P. G. Hatcher. 2015. Use of ESI-FTICR-MS to characterize dissolved organic matter in headwater streams draining forest-dominated and pasture-dominated watersheds. PLoS One **10**: e0145639. doi:10.1371/journal.pone.0145639

Lucas, J., I. Koester, A. Wichels, J. Niggemann, T. Dittmar, U. Callies, K. H. Wiltshire, and G. Gerdts. 2016. Short-term dynamics of North Sea bacterioplankton-dissolved organic matter coherence on molecular level. Front. Microbiol. **7**: 321. doi:10.3389/fmicb.2016.00321

Mazur, H., A. Konop, and R. Synak. 2001. Indole-3-acetic acid in the culture medium of two axenic green microalgae. J. Appl. Phycol. **13**: 35–42.

Medeiros, P. M., and others. 2015. Fate of the Amazon River dissolved organic matter in the tropical Atlantic Ocean. Global Biogeochem. Cycl. **29**: 677–690. doi:10.1002/2015GB005115

Medeiros, P. M., and others. 2016. A novel molecular approach for tracing terrigenous dissolved organic matter into the deep ocean. Global Biogeochem. Cycl. **30**: 689–699. doi:10.1002/2015GB005320

Merder, J., and others. 2020*a*. ICBM-OCEAN: Processing ultrahigh-resolution mass spectrometry data of complex molecular mixtures. Anal. Chem. **92**: 6832–6838. doi:10.1021/acs.analchem.9b05659

Merder, J., J. A. Freund, U. Feudel, J. Niggemann, G. Singer, and T. Dittmar. 2020*b*. Improved mass accuracy and isotope confirmation through alignment of ultrahigh-resolution mass spectra of complex natural mixtures. Anal. Chem. **92**: 2558–2565.

Moran, M. A., and others. 2016. Deciphering ocean carbon in a changing world. Proc Nat Acad Sci USA **113**: 3143–3151. doi:10.1073/pnas.1514645113

Nebbioso, A., and A. Piccolo. 2013. Molecular characterization of dissolved organic matter (DOM): A critical review. Anal. Bioanal. Chem. **405**: 109–124.

Ogawa, H., Y. Amagai, I. Koike, K. Kaiser, and R. Benner. 2001. Production of refractory dissolved organic matter by bacteria. Science **292**: 917–920.

Osterholz, H., J. Niggemann, H.-A. Giebel, M. Simon, and T. Dittmar. 2015. Inefficient microbial production of refractory dissolved organic matter in the ocean. Nat. Commun. **6**: 7422. doi:10.1038/ncomms8422

Osterholz, H., D. L. Kirchman, J. Niggemann, and T. Dittmar. 2016*a*. Environmental drivers of dissolved organic matter molecular composition in the Delaware Estuary. Front. Earth Sci. **4**: 95. doi:10.3389/feart.2016.00095

Osterholz, H., G. Singer, B. Wemheuer, R. Daniel, M. Simon, J. Niggemann, and T. Dittmar. 2016*b*. Deciphering associations between dissolved organic molecules and bacterial communities in a pelagic marine system. ISME J. **10**: 1717–1730.

Piotrowska-Niczyporuk, A., and A. Bajguz. 2014. The effect of natural and synthetic auxins on the growth, metabolite content and antioxidant response of green alga *Chlorella vulgaris* (Trebouxiophyceae). Plant Growth Regul. **73**: 57–66.

R Core Development Team. 2017. R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. Available from https://www.R-project.org.

Raymond, P., and J. E. Bauer. 2001. Riverine export of aged terrestrial organic matter to the North Atlantic Ocean. Nature **409**: 497–500. doi:10.1038/35054034

Riedel, T., and T. Dittmar. 2014. A method detection limit for the analysis of natural organic matter via Fourier transform ion cyclotron resonance mass spectrometry. Anal. Chem. **86**: 8376–8382.

Riedel, T., M. Zark, A. Vähätalo, J. Niggemann, R. G. M. Spencer, P. J. Hernes, and T. Dittmar. 2016. Molecular signatures of biogeochemical transformations in dissolved organic matter from ten world rivers. Front. Earth Sci. **4**: 85. doi:10.3389/feart.2016.00085

Rijstenbil, J. W., J. A. Wijnholds, and J. J. Sinke. 1989. Implications of salinity fluctuation for growth and nitrogen metabolism of the marine diatom *Ditylum brightwellii* in comparison with *Skeletonema costatum*. Mar. Biol. **101**: 131–141.

Rivas-Ubach, A., Y. Liu, T. S. Bianchi, N. Tolić, C. Jansson, and L. Paša-Tolić. 2018. Moving beyond the van Krevelen

Diagram: A new stoichiometric approach for compound classification in organisms. Anal. Chem. **90**: 6152–6160.

Rousseeuw, P. J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**: 53–65.

Schmidt, F., M. Elvert, B. P. Koch, M. Witt, and K.-U. Hinrichs. 2009. Molecular characterization of dissolved organic matter in pore water of continental shelf sediments. Geochim. Cosmochim. Acta **73**: 3337–3358.

Seidel, M., and others. 2014. Biogeochemistry of dissolved organic matter in an anoxic intertidal creek bank. Geochim. Cosmochim. Acta Theriol. **140**: 418–434. doi:10.1016/j.gca.2014.05.038

Steele, J. A., and others. 2011. Marine bacterial, archaeal and protistan association networks reveal ecological linkages. ISME J. **5**: 1414–1425. doi:10.1038/ismej.2011.24

Stirk, W. A., O. Novák, V. Hradecká, A. P□nčík, J. Rolčík, M. Strnad, and J. Van Staden. 2009. Endogenous cytokinins, auxins and abscisic acid in *Ulva fasciata* (Chlorophyta) and *Dictyota humifusa* (Phaeophyta): Towards understanding their biosynthesis and homoeostasis. Eur. J. Phycol. **44**: 231–240. doi:10.1080/09670260802573717

Stubbins, A., and T. Dittmar. 2015. Illuminating the deep: Molecular signatures of photochemical alteration of dissolved organic matter from North Atlantic Deep Water. Mar. Chem. **177**: 318–324.

Teeling, H., and others. 2012. Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom. Science **336**: 608–611. doi:10.1126/science.1218344

Tolić, N., and others. 2017. Formularity: Software for automated formula assignment of natural and other organic matter from ultrahigh-resolution mass spectra. Anal. Chem. **89**: 12659–12665. doi:10.1021/acs.analchem.7b03318

Vähätalo, A. V., and R. G. Wetzel. 2004. Photochemical and microbial decomposition of chromophoric dissolved organic matter during long (months–years) exposures. Mar. Chem. **89**: 313–326.

Vrhovšek, U. 1998. Extraction of hydroxycinnamoyltartaric acids from berries of different grape varieties. J. Agric. Food Chem. **46**: 4203–4208.

Wagner, S., T. Riedel, J. Niggemann, A. V. Vähätalo, T. Dittmar, and R. Jaffé. 2015. Linking the molecular signature of heteroatomic dissolved organic matter to watershed characteristics in world rivers. Environ. Sci. Technol. **49**: 13798–13806.

Wienhausen, G., B. E. Noriega-Ortega, J. Niggemann, T. Dittmar, and M. Simon. 2017. The exometabolome of two model strains of the *Roseobacter* group: A marketplace of microbial metabolites. Front. Microbiol. **8**: 1985. doi:10.3389/fmicb.2017.01985

Wiltshire, K. H. 2013. Phytoplankton abundance at time series station Helgoland Roads, North Sea, in 2012. PANGAEA.

Wiltshire, K. H., and others. 2010. Helgoland Roads, North Sea: 45 years of change. Estuaries Coast **33**: 295–310. doi:10.1007/s12237-009-9228-y

Wiltshire, K. H., M. Boersma, K. Carstens, A. C. Kraberg, S. Peters, and M. Scharfe. 2015. Control of phytoplankton in a shelf sea: Determination of the main drivers based on the Helgoland Roads time series. J. Sea Res. **105**: 42–52.

Wood, S. N. 2017. Generalized additive models: An introduction with R, 2nd ed. Chapman and Hall/CRC.

Yamane, H., K. Konno, M. Sabelis, J. Takabayashi, T. Sassa, and H. Oikawa. 2010. 4.08—Chemical defence and toxins of plants, p. 339–385. *In* L. Mander and H.-W. Liu [eds.], Comprehensive natural products II, 1st ed. Elsevier.

Zark, M., and T. Dittmar. 2018. Universal molecular structures in natural dissolved organic matter. Nat. Commun. **9**: 3178.

Acknowledgments

We thank Antje Wichels and Mirja Meiners for sampling, Katrin Klaproth for technical assistance with FT-ICR-MS analyses, and Matthias Friebe and Ina Ulber for DOC analyses. We are very grateful to Bernhard Fuchs for sharing his expert knowledge on bacterial communities. The helpful and constructive comments of two anonymous reviewers helped to improve the manuscript. This study was carried out in the framework of the PhD research training group "The Ecology of Molecules" (EcoMol) supported by the Lower Saxony Ministry for Science and Culture. Further funding was provided by Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center Roseobacter (TRR 51). Open access funding enabled and organized by Projekt DEAL.

Conflict of Interest

None declared.

Submitted 08 October 2020
Revised 13 April 2021
Accepted 27 July 2021

Associate editor: Thomas R. Anderson