Supplementary: An Interactive Detection and Visualization of Ocean Carbon Regimes

Sweety Mohanty[1,2], Daniyal Kazempour[1], Lavinia Patara[2], Peer Kröger[1]
1. Christian-Albrechts-University, 2. GEOMAR Helmholtz Centre for Ocean Research, Kiel, Germany
smo, dka, pkr@informatik.uni-kiel.de | smohanty, lpatara@geomar.de

I. More on Related Work

When it comes to the design of interactive tools for specific domains, questions that arise are: Do we need such tools in general and what can we expect from them? In the particular field of marine science, the authors of [1] state that web-based interactive tools are valuable and of paramount importance to "...explore and better understand environmental changes across habitats and entire ecosystems as well as the responses of living resources within them to multiple drivers." [1]. In a previous work of [2] the authors conducted a survey in which they interviewed 76 climate impact researchers of different experience levels and scientific backgrounds. From the 76 individuals, the authors derived requirements that should be met in an interactive data analysis tool. Among them were aspects such as (1) dynamic filtering/adjustment mechanisms being mandatory (2) visual encoding should be interactively adjustable and (3) visualizing on two or three-dimensional geographical maps being of utmost importance. Therefore our proposed tool is also constructed with these aspects in mind.

Within the scope of our application, it is also important to consider related work in the context of climate and ocean apps. In a more recent contribution, the authors in [3] propose an interactive web tool based on simulation data to visualize and inspect regional ocean model (ROM) data from southeast Asia. It provides functionalities such as computing climatology, mean, anomaly, and trends. The focus of the application is set on sea surface height (SSH) development and trends. This tool however does not discover regions with similar properties. One does also not obtain local models and does not enable the domain experts to inspect clusters for the discovery of potentially interesting patterns. A related work that actually does perform clustering is in [4]. Here the authors utilize unsupervised learning methods for obtaining global marine regions from plankton community structure and nutrient flux data. They capture non-linear ecosystem models tailored to high-dimensional cases in which the challenge is to prune irrelevant features. To achieve this the authors rely on t-SNE [5] for dimensionality reduction followed by clustering via the density-based clustering method DBSCAN [6]. The purpose of the approach in [4] is to facilitate the comparison between models and to augment the understanding and observation of maritime environments. The proposed method, however, does not come with an interactive tool. Furthermore, the data is clustered and not, as in our case the obtained regression models to identify similar regions w.r.t. the local relationships between different drivers. An ocean data exploration tool with emphasis on spatial and temporal patterns is introduced in the work[7]. The purpose of the interactive web tool OceanTEA is to facilitate the analysis of climate-relevant ocean observations. It provides the domain scientists with four views encompassing time series management, data exploration, spatial analysis, and temporal pattern discovery. Furthermore, it provides a rich set of filters and settings that can be applied. However, this tool does not provide any cluster-analysis capabilities and thus does not augment in the detection of regions with similar regression patterns between the drivers. Lastly in [8] the authors propose an interactive climate data visualization tool with a strong emphasis on high-resolution and fast computation through parallelization techniques. This contribution also does not provide any cluster-analysis functionalities and is furthermore not tailored to the ocean-climate context.

One last aspect in the context of related work is pointing toward the feature of computing cuts within the dendrogram at different heights. In [9] the authors propose DESPOTA to achieve this goal. In their approach, however, the authors do not elaborate on the runtime complexity and state that one could "...study ... the computational complexity ...for further research". A more detailed view on their code basis

[1] reveals with "The current version of the code is not very fast in case of big data, but it works." that this approach is on large datasets, due to its reliance on permutation tests, computationally prohibitive and therefore not feasible, especially not in an interactive setting. To meet the demand for fast computation we use a simpler, yet computationally faster method in this demo.

II. System Architecture in Detail

A. Ocean Model Output Data

We develop our project on the output of a global ocean sea ice model, NEMO-LIM2[10] coupled to a biogeochemistry based model simulating the carbon cycle [11]. The model has a global resolution of 0.5 degrees (resulting in 511 x 722 latitude and longitude indices) and 46 vertical levels. In this study, we analyze only the uppermost ocean level in contact with the atmosphere. The ocean model is forced by an atmospheric reanalysis product covering the period 1958 to 2018. From the model output we extract SST, DIC, ALK, and fCO_2 at a monthly resolution and save it as pickle (.pkl) files. We utilize these .pkl files to conduct further computations and bypass repeated execution of the data extraction technique.

B. Backend functions

We define a carbon regime based on the relationship of CO_2 concerning its drivers - SST, DIC, and ALK. Because the relation between fCO_2 and its drivers exhibits strong spatial variations, We explore the local relationships instead of global relationships. Since the biogeochemical ocean model follows very complex methodologies to calculate the carbon dioxide fugacity, we begin our analysis by building multiple piece-wise local linear relationships. As our next step to detect the regimes, we use a hierarchical clustering procedure to batch similar relationships.

1) ***1. Grid-based Multivariate Linear Regression****:* We partition the global ocean into 2 x 2 degree grid cells and run multivariate linear regressions between CO_2 and SST, DIC, and ALK in each grid cell on a monthly scale. As all grids do not contain a similar number of data points, we only use regressions output whose P-value is less than 0.04.

2) ***2. Agglomerative Clustering****:* The agglomerative Clustering algorithm initiates by considering each data point as an individual cluster. Then sets of two singleton clusters combine sequentially till all clusters are integrated into one large one enclosing all data points. We use **Ward Linkage** and **Euclidean distance** to build the dendrogram. Ward linkage gives us the distance between two clusters, equal to the increase in the sum of squares of individual clusters after they merge. The link heights of the dendrogram represent this distance of merge.

3) ***Distance-variance cluster selection methodology****:* Our approach builds upon top-down dendrogram traversal. As we move down on the dendrogram on the distance axis, from top to the bottom, the distance between one merge to the next decreases. This signifies that the cost to merge two different clusters also decreases, i.e., similarity between clusters increases. During this traversal, variance of the individual clusters also changes. In our proposed approach, we measure this change in similarity or distance and variance between two different levels of any link on the dendrogram. We set two threshold parameters - `delta_dist` and `delta_var`. We start from the topmost dendrogram link, which represents the entire dataset, and move down the tree. At each link or merge position (hereafter, "head"), we calculate the variance and note the height of the merge. We separately compute the variance of the left and right sub-dendrograms immediately below it and note their respective distances on the dendrogram. We determine the change in distance and change in variance in the individual left and right branches with respect to their merge head. We compare the changes detected with our two threshold variables. If the change in both distance and variance is low (compared with a threshold) between a merge link and a sub-tree, we consider such merge to be inexpensive, i.e., the merge did not create significant change of distance and variance after the clustering. So, we stop the traversal for that particular branch and declare the sub-tree (left or right) of the head as a cluster. If both the difference in distance and variance are high, i.e., they surpass the thresholds, it implies that two dissimilar clusters are merged. We traverse further down to detect lower sub-trees where change in distance and variance is minor.

4) ***BIC Score computation****:* We run our analysis for different sets of `delta_dist` and `delta_var`. To understand how fitting the resulting clusters for each pair of thresholds are, we determine BIC (Bayesian Information Criterion) Score [**?**]. BIC Scores specify the threshold selection under the framework of maximum likelihood estimation. The lower the BIC, the better the selected method [**?**]. Hence, the score emphasizes the values of `delta_dist` and `delta_var` of the clusters over given data distribution. We tra-

[1]https://github.com/domenicovistocco/despota

verse the dendrogram and compute the BIC scores of the cluster output by iterating through different collections of `delta_dist` and `delta_var`.

C. Demonstration Tool Development

We are using Dash, a python framework, to build both the interactive user interface and the server. The Dash framework facilitates both frontend and backend workflow as it is based on Flask, Plotly. js and React. js. The total end-to-end runtime of our tool in a rudimentary scenario is between five to six minutes. Users of our tool can query particular months and years of ocean model output. They will be able to see the generated dendrogram after clustering grid-based multivariate linear regression. They will have access to BIC scores generated for different values of distance and variance thresholds in the form of a graph. Then they can select the parameters from preset values or input their own. The server will then return potential clusters by spotlighting corresponding links of the dendrogram. The users can also catch the carbon regimes on a world map, visualize the clustered distribution of linear regression coefficients, check associated mean regression slopes in a table, and download the dataset with the cluster labels.

III. Carbon regimes Evaluation

In this section we highlight how we assess the carbon regimes detected by our methodology. The tool generates a tabular overview of different clusters that imparts crucial insights into the detected carbon regimes and steers us to conduct evaluations through the visualizations. Let us consider the clusters of January 2017 for threshold parameters set at `delta_dist` = 20 and `delta_var` = 0.2. We receive 6 clusters as shown in figures-1 and 2.

The table in figure-3 shows the mean values of regression coefficients of SST, DIC, and ALK in each cluster. We see that the detected carbon regimes are a "subtropical regime" (cluster 2), where fCO_2 shows a positive dependence on SST and negative dependence on DIC, a "transition regime" (cluster 1), where fCO_2 shows a weak dependence on all drivers (possibly because they are all relevant but push in different directions), and an "upwelling regime" (cluster 5) where fCO_2 shows a strong positive dependence on DIC. Then there is a set of high-latitude regimes which cover small areas and characterized by extreme responses to environmental conditions. For instance clusters 3 and 4 are characterized by a strong fCO_2 decrease with increasing SST (possibly because increased temperature melts the sea ice which in turn fuels fCO_2 uptake by phytoplankton) an "ice regime" (cluster 6) characterized by a strong positive dependence on both SST and DIC, and another high-latitude regime where ALK gains importance.

IV. Future Works

A fascinating future application of this tool is to track the time-changing patterns of carbon regimes in response to changing environmental conditions, which are expected to be drastically impacted by anthropogenic climate change.

References

[1] A. Benson, T. Murray, G. Canonico, E. Montes, F. E. Muller-Karger, M. T. Kavanaugh, J. Trinanes, and L. M. Dewitt, "Data management and interactive visualizations," *Oceanography*, vol. 34, no. 2, pp. 130–141, 2021.

[2] C. Tominski, J. F. Donges, and T. Nocke, "Information visualization in climate research," in *2011 15th International Conference on Information Visualisation*. IEEE, 2011, pp. 298–305.

[3] F. R. Muhammad, I. W. Amanullah, and A. Faqih, "Seacoapp: A web app to analyze, download, and visualize regional ocean model (rom) datasets in southeast asia," in *IOP Conference Series: Earth and Environmental Science*, vol. 893, no. 1. IOP Publishing, 2021, p. 012077.

[4] M. Sonnewald, S. Dutkiewicz, C. Hill, and G. Forget, "Elucidating ecological complexity: Unsupervised learning determines global marine eco-provinces," 2020.

[5] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[6] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[7] A. Johanson, S. Flögel, C. Dullo, and W. Hasselbring, "Oceantea: Exploring ocean-derived climate data using microservices," 2016.

[8] D. N. Williams, "Visualization and analysis tools for ultra-scale climate data," *Eos, Transactions American Geophysical Union*, vol. 95, no. 42, pp. 377–378, 2014.

[9] D. Bruzzese and D. Vistocco, "Cutting the dendrogram through permutation tests," *Proceedings of COMPSTAT'2010*, pp. 847–854, 2010.

[10] M. Gurvan, R. Bourdallé-Badie *et al.*, "Nemo ocean engine," Mar. 2022.

[11] C.-T. Chien, J. V. Durgadoo *et al.*, "Foci-mops v1–integration of marine biogeochemistry within the flexible ocean and climate infrastructure version 1 (foci 1) earth system model," *Geoscientific Model Development Discussions*, pp. 1–58, 2022.

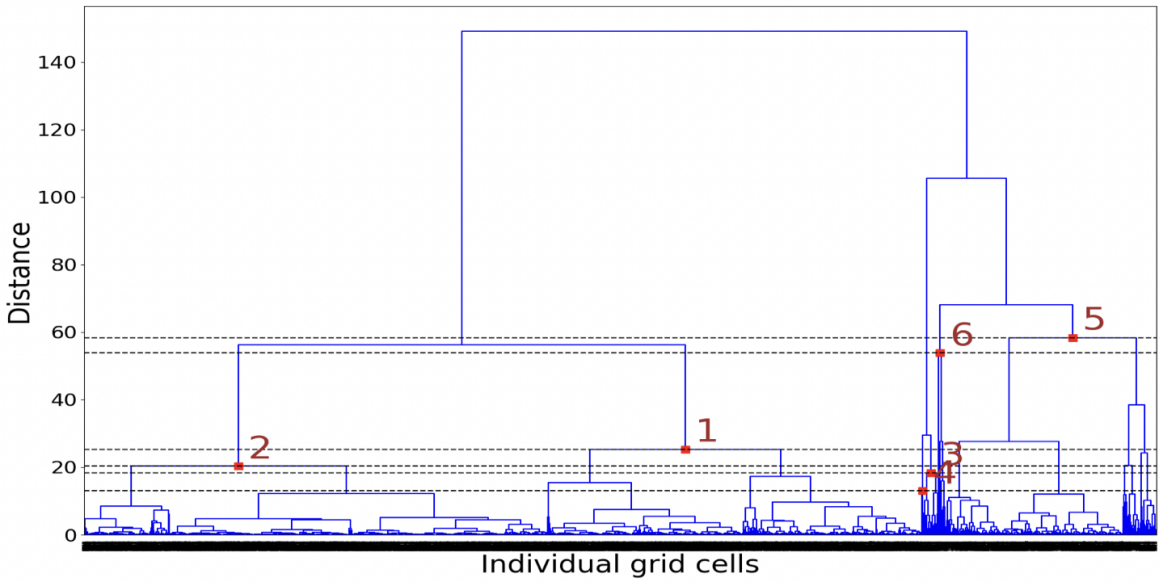

Figure 1. The dendrogram with 6 detected clusters of January, 2017

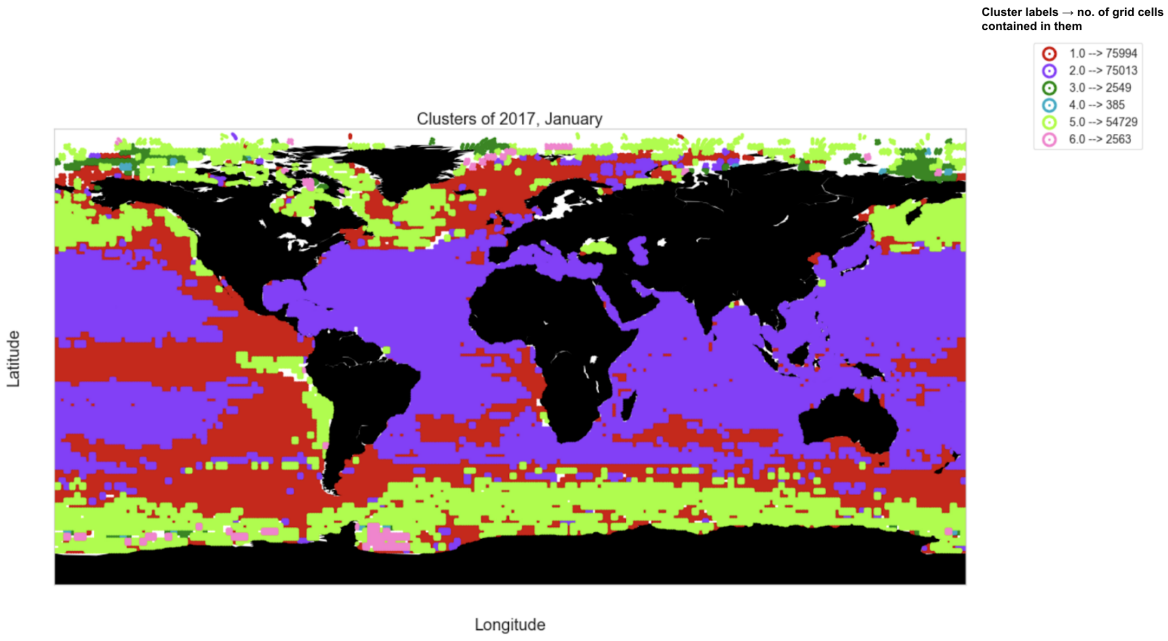

Figure 2. The map shows the 6 detected clusters of January, 2017

Cluster Label	SST Slope	DIC Slope	ALK Slope
1	0.09	0.06	0.07
2	0.04	-0.7	0.65
3	-5.44	1.7	-1.91
4	-11.83	1.36	-1.69
5	0.12	1.23	-1.34
6	4.99	3.42	-3.29

Figure 3. A tabular overview of our clustering output.