Workshop on Visualisation in Environmental Sciences (EnvirVis) (2021)
S. Dutta and K. Feige and K. Rink and D. Zeckzer (Editors)

Probabilistic Principal Component Analysis Guided Spatial Partitioning of Multivariate Ocean Biogeochemistry Data

Subhashis Hazarika[1], Ayan Biswas[1] , Earl Lawrence[1], and Phillip J. Wolfram[2]

[1]Computer, Computational, and Statistical Sciences, Los Alamos National Laboratory, New Mexico, USA
[2]Advanced Engineering Analysis, Los Alamos National Laboratory, New Mexico, USA

Abstract

Farm-scale cultivation of macroalgae for the production of renewable biofuel depends on complex ocean hydrodynamics and also on the availability of different essential nutrients. To better understand such conditions that are conducive for the growth of macroalgae, scientists implement large-scale computational models, simulating several physical variables (essential nutrients, and other chemical compounds), relevant to study oceanic biogeochemistry (BGC). Visualizing and analysing the different physical variables and their inter-variable relationships across the spatial domain is crucial to form concrete understanding of the underlying physical phenomenon. To facilitate such multivariate analyses for large-scale simulation data, a popular and effective way is to decompose the spatial domain into smaller local regions based on the variable relationships. However, spatial decomposition of multivariate data is not trivial. In this paper, we propose a novel multivariate spatial data partitioning approach using probabilistic principal component analysis. We also perform detailed study of other prospective multivariate partitioning schemes and compare them with our proposed method. To demonstrate the efficacy of our approach, we studied nutrient relationships across different regions of the ocean using a high-resolution Ocean BCG simulation data set, which comprises of multiple physical variables essential for macroalgae cultivation. We further validate the results of our analyses by getting feedback from domain experts in the field of ocean sciences.

Categories and Subject Descriptors (according to ACM CCS): Human-centered computing → Visualization → Visualization application domains → Scientific visualization

1. Introduction

Seaweeds or macroalgae are important sources for a variety of biofuels, and thus can potentially meet the global demand for alternative sources of renewable energy. Large-scale cultivation of macroalgae can supply 10% of the United States' transportation energy demands [WMB*19]. However, successful fuel production from such farms depends on ambient ocean hydrodynamics and the availability of the right proportions of essential nutrients. Scientific studies of these factors can be performed using large-scale computational simulations, modeling ocean hydrodynamics and concentration of essential nutrient and chemical compounds. Within the Energy Exascale Earth System Model (E3SM) [PADB*19], scientists are trying to study ocean biogeochemistry (BGC) [MDK*01, WMB*19] by using the high-resolution Model for Prediction Across Scales Ocean (MPAS-O) [RPH*13]. The high-resolution spatial data generated by the simulation model is comprised of multiple physical variables such as concentration of essential nutrients like nitrate, ammonium, phosphate, iron, silicon as well as other organic and inorganic biogeochemical compounds along with various hydrodynamic variables as illustrated in Figure 1. Understanding how these variables (e.g., different nutrient concentrations) are related is a crucial step in many multivariate analysis tasks which involve correlation analysis, association analysis, and query-driven analysis, just to name a few [STpS06, FH09]. However, analyzing

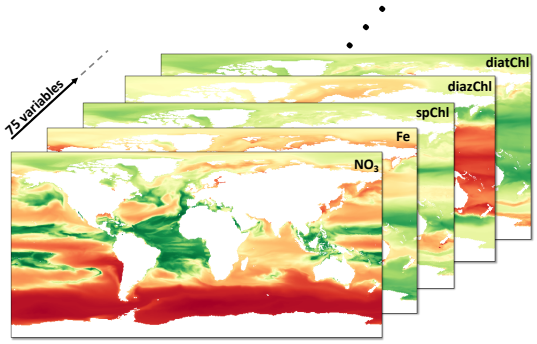

Figure 1: *Multivariate Ocean BGC data set comprising of 75 physical variables including concentration of essential nutrients, various organic and inorganic compounds along with other measurements from the simulation.*

DOI: 10.2312/envirvis.20211078

and visualizing large-scale multivariate simulation data can be a challenging and non-trivial task because of the high-resolution spatial dimensions, coupled with a large number of physical variables to take into consideration during analysis.

An effective strategy to address the challenges associated with large-scale simulation data is to first decompose the high-resolution spatial domain into smaller contiguous regions [DWS*17, WKW*17, HDSC19]. This can help to model complex global non-linear functions by using piecewise locally linear and simpler models for each spatial partitions [RS00, KL97, KKW*15]. While multiple data-driven spatial partitioning schemes (both regular and irregular partitioning) exist for univariate data, to the best of our knowledge, not many partitioning schemes exist for multivariate data with large number of variables. In this paper, we propose a novel spatial data partitioning scheme for multivariate data which takes into consideration the inter-variable relationships to create partitions where the corresponding multivariate data can be modeled and analyzed by using linear models.

Our proposed partitioning scheme is built on top of the *simple linear iterative clustering* (SLIC) algorithm [ASS*12] to create irregular shaped partitions and utilizes *probabilistic principal component analysis* (PPCA) [TB99] to capture the multivariate relationships. SLIC algorithm operates on local spatial regions of the data to generate irregular partitions such that the data within each partition can be modeled by a linear model such as principal component analysis (PCA). Among its many other advantages, probabilistic formulation of classical PCA allows us to evaluate the likelihood of a PCA model. We utilize this property of PPCA to design our new spatial decomposition scheme for multivariate data to create partitions with homogeneous variable relationships. We also performed extensive evaluation of our proposed scheme with other possible choices of regular and irregular partitioning schemes to identify the advantages and disadvantages of these methods across different multivariate analysis tasks. We applied our partitioning scheme on multivariate Ocean BCG data with 75 physical variables to study the local variable relationship among essential nutrients in different regions of the ocean and validated the analysis results with expert feedbacks. To summarize, the main contribution of our work are as follows:

1. Proposed a multivariate relationship guided irregular spatial domain partitioning scheme.
2. Performed extensive evaluation of different spatial partitioning schemes for multivariate data analyses.
3. Applied our approach on multivariate Ocean BCG data to study nutrient relationships across different regions of the ocean.

2. Related Works

Multivariate data analysis and visualization for scientific simulation data is a well researched topic. Extensive reviews and surveys about the state-of-the-art for multivariate scientific visualization exist in literature. Noteworthy among them are the works of Wong and Bergeron [WB97], and Fuchs and Hauser [FH09]. Many multivariate analysis and visualization tasks are fundamentally based on studying the variable relationships across the spatial domain. Sauber et al. [STpS06] computed local correlation coefficient between the variables in a multivariate spatial data and created multifield graphs to effectively visualize the variable relationships. Similar correlation-based analysis was performed to enable different query-driven visualizations for multivariate data [BGJA07]. Nagaraj et al. [NNN11] utilized local gradient-based analysis to measure variable relationships for the purpose of comparative visualization. Gosnik et al. [GGA*11] derived special correlation fields by performing normalized dot product between the gradient fields of different variables to visualize their relationships. Wang et al. [WYG*11] studied causal relationship among variables by using transfer entropy methods. Biswas et al. [BDSW13] proposed specific mutual information metrics to study surprise and predictability of different variables in multivariate data and thereby enrich corresponding multivariate visualizations. Jänicke et al. [JWSK07] proposed different local statistical measures quantifying variable relationships to highlight informative regions in their visualizations.

Intelligent decomposition of the spatial domain into smaller regions have various applications and use-cases for scientific data analysis and visualization. Particularly for large-scale high-resolution simulation data sets, spatial partitions help to make the analysis tasks more manageable and scalable. Hazarika et al. [HDSC19] used regular partitioning schemes to decompose the spatial domain such that linear relationship models like Gaussian copula function [HBS18] can be employed to model variable relationships. Dutta et al. [DCH*17] utilized regular block-wise local Gaussian mixture models to study flow instability and track features for jet turbine simulation data. They also proposed an irregular spatial data partitioning scheme [DWS*17] to model univariate data based on simple linear iterative clustering algorithm [ASS*12] to create homogeneous region for distribution-based data modeling. Besides, there are multiple such examples of spatial data partitioning to achieve scalable analysis of large-scale simulation data analysis and visualization [WKW*17, HBW*20, DS15, WHLS19]. However, most of this methods perform a naive partitioning of the domain or use some univariate data property to create intelligent partitions. Our proposed probabilistic PCA based partitioning scheme is an attempt to come up with a variable relationship guided spatial decomposition method to facilitate multivariate analyses.

3. Proposed Approach

Our proposed multivariate data partitioning scheme combines probabilistic principal component analysis (PPCA) with an irregular spatial decomposition algorithm called simple linear iterative clustering (SLIC). The resulting partitions are such that the multivariate data within each partition can be modeled by simple linear models for detail analyses. In this section, we first briefly explain the concept of PPCA and how it serves our data partitioning objectives. We then elaborate on the details of our partitioning schemes.

3.1. Probabilistic Principal Component Analysis (PPCA)

PPCA [TB99] is a probabilistic formulation of the classical PCA [Jol86]. While the classical PCA is based on mapping high-dimensional observed data space to low dimensional latent space, the PPCA framework is based on mapping from a latent space to the data space. The graphical representation of this probabilistic model is shown in Figure 2 for a dataset of N observations, where $\mathbf{x}$ is

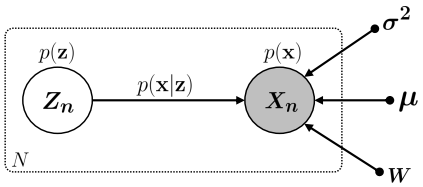

Figure 2: *Graphical model representation of PPCA for a data set with N high-dimensional observations of X, where each observation x_n is associated with value z_n of the low-dimensional latent variable Z*

high dimensional (d-dimensional) multivariate observed data and $\mathbf{z}$ is the corresponding low-dimensional (q-dimensional, and $q < d$) latent space data. $\mathbf{x_n}$ and $\mathbf{z_n}$ in Figure 2 represent the random variables for individual instances of N data-points.

The latent variable model for PPCA is framed over the general formulation of PCA as factor analysis, which can be stated as,

$$\mathbf{x} = \mathbf{W}\mathbf{z} + \mu + \varepsilon \tag{1}$$

where, the parameter $\mathbf{W}$ is a $d \times q$ matrix relating the two set of variables, the parameter μ permits the model to have non-zero mean and ε models the noise in this linear transformation. In PPCA, the prior over the latent variable $\mathbf{z}$ is given by a zero-mean unit-covariance Gaussian, i.e., $\mathbf{z} \sim \mathcal{N}(0, I)$. The noise is modeled using an isotropic Gaussian noise model, i.e., $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$, where the parameter σ^2 governs the noise variance.

Given these Gaussian priors over $\mathbf{z}$ and ε, the probability distribution of the observed variable $\mathbf{x}$ conditioned over the latent variable $\mathbf{z}$ in conjunction with Equation 1 can be stated as,

$$\mathbf{x}|\mathbf{z} \sim \mathcal{N}(\mathbf{W}\mathbf{z} + \mu, \sigma^2 I) \tag{2}$$

Likewise, the marginal distribution over the observed variable is Gaussian, which can be readily obtained by integrating out the latent variable in Equation 2 and is given by,

$$\mathbf{x} \sim \mathcal{N}(\mu, \mathbf{C}) \tag{3}$$

where, the observation covariance model is specified by $\mathbf{C} = \mathbf{W}\mathbf{W}^\mathbf{T} + \sigma^2\mathbf{I}$. The corresponding log-likelihood for this probabilistic model for N data-points is given as;

$$\begin{aligned} \mathcal{L}_{pca} &= \sum_{n=1}^{N} \ln p(\mathbf{x_n}) \\ &= -\frac{N}{2}\{d\ln(2\pi) + \ln|\mathbf{C}| + tr(\mathbf{C}^{-1}\mathbf{S})\} \end{aligned} \tag{4}$$

where, $\mathbf{S}$ is the sample covariance matrix of the observed data given by,

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x_n} - \mu)(\mathbf{x_n} - \mu)^T \tag{5}$$

The parameters $\mathbf{W}$, μ, and σ^2 can be estimated by using maximum likelihood estimation. To contrast this with classical PCA, these same parameter values can be obtained in a deterministic manner by eigen decomposition. One of the main advantage of the probabilistic formulation [TB99] of PCA is the existence of a likelihood function (Equation 4). This allows direct comparison with other probabilistic models and has been used creatively to create mixture of PCAs to model complex nonlinear relationships as well as help work with missing data.

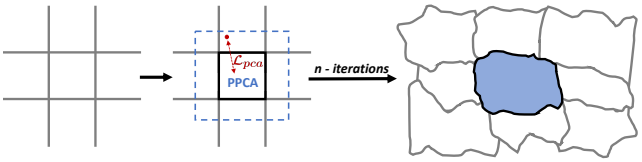

Figure 3: *High-level illustration of our PPCA guided irregular partitioning scheme.*

In this paper, we utilize this likelihood estimation of PPCA to propose a variable relationship-driven irregular spatial partitioning scheme for multivariate data to study intricate nutrient relationships essential for macroalgae growth.

3.2. PPCA guided Irregular Partitioning Scheme

To create a multivariate data property driven spatial decomposition scheme, we apply the popular simple linear iterative clustering (SLIC) algorithm [ASS*12]. SLIC is widely used in the field of image processing and segmentation to create homogeneous superpixels (a spatially contiguous partition with more than one pixels) for further downstream image analysis. This has also been adopted by the scientific data visualization community to create homogeneous data partitions for spatial data modeling [DWS*17]. However, SLIC works well with univariate data sets or data with limited variable fields (e.g. the 3 RGB channels in images). For multivariate data with large number of variable fields there is no unique approach to decompose the spatial domain into regions with homogeneous variable relationships.

The SLIC algorithm can be interpreted as a spatially constrained variant of the k-means clustering where each spatial location (i.e., pixels for images) is mapped to a neighbouring group of locations with similar data properties. The goal of our proposed approach is to utilize the log-likelihood of PPCA (Equation 4) as a distance measure in SLIC to identify groups of spatial locations whose underlying multivariate data can be modeled by a simple linear model.

SLIC is an iterative algorithm that starts with regular equal-sized non-overlapping spatial partitions and incrementally redefines the partition boundaries based on desired data property to create irregular data-drive partitions. The steps involved in our PPCA based partitioning algorithm is outlined below and illustrated in Figure 3.

Step-1: Initialization: Create k non-overlapping clusters or partitions by regularly partitioning the spatial domain. Let C_k be the centers of these clusters and the initial size of each partition is $S \times S$. We maintain a label image l and a log-likelihood image $\mathcal{L}$ to keep track of the computation. For a given spatial location x, $l(x)$ returns the current assigned label based on which partition/cluster x belongs (i.e., $l(x) \in \{1, 2, ..k\}$). $\mathcal{L}(x)$ gives the log-

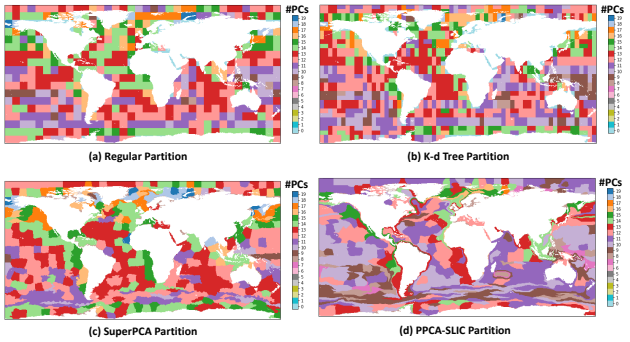

Figure 4: *Minimum number of principal components (PCs) required to capture 99% explained variance of the original multivariate data for individual partitions .*

likelihood value of the point x to the PPCA model of the current partition that it belongs.

Step-2: Iterate till Termination: We then iterate over all k partitions. For each partition we do the following:

- **Update label and log-likelihood images:** For all location x within a $2S \times 2S$ window around the center C_k, we compute the log-likelihood value (Equation 4) of x to the PPCA model of the current partition data. If this log-likelihood value is more than $\mathcal{L}(x)$ and $l(x) \neq k$, we update the current label and log-likelihood images at location x to store k and the current maximum log-likelihood value.
- **Update partitions:** Compute the center C_k based on the updated labels of the label image l in the previous step.

Termination: Terminate when clusters remain unchanged after an iteration or when the threshold of maximum number of iterations is reached.

Step-3: Enforce Spatial Connectivity: Sometimes the assigned locations of a partition may not be fully connected because the connectivity is not enforced in the above steps. Just like the traditional SLIC algorithm, we perform a post-processing step at the end to enforce spatial connectivity. We examine the labeled connected components not connected to their partition center C_k and relabel them so that they are connected to the spatially nearest partition labels.

The multivariate data for the locations within each resulting partition can be modeled by simple linear models, including PCAs for further downstream analyses.

4. Evaluation Study and Results

To evaluate the performance of our proposed partitioning algorithm, we compared our approach with other possible spatial data decomposition schemes for multivariate data. Here is a brief outline of the three spatial data partitioning schemes that we compared our approach with.

P1: Regular Blockwise Partitioning: This is a simple yet effective spatial partitioning scheme widely used as part of many large-scale data analysis tasks. The spatial domain is divided into equal-sized non-overlapping blocks of user-defined dimensions. This is a completely data-agnostic approach and does not take into consideration the underlying data properties while decomposing the spatial domain. Therefore, they are great at easily breaking down large-scale problems into smaller sub-problems but the analysis results may not be necessarily optimal at all times. In our work, we perform regular blockwise spatial partitioning and evaluate the multivariate data within each partition.

P2: K-d Tree Partitioning: K-d tree is a popular data-driven spatial partitioning scheme that employ a top-down sub-division scheme to decompose the spatial domain. It recursively partitions the spatial domain till a particular data property is achieved for the underlying data in a partition. We improvised this partitioning to address multivariate data and match with our proposed irregular partitioning scheme. In our redesigned K-d tree partitioning approach for multivariate data, we first create a local PCA model with the multivariate data of a partition. Next, the decision to further decompose a partition is made based on if a certain q number of principal components (PCs) can capture say 99% variance of the data. If this criteria is not satisfied for a partition, we further sub-divide the current partition till either the criterion is met or the size of the partition has reached the minimum dimension set for a partition.

P3: SuperPCA: While the above two approaches produce regular axis-aligned partitions, there are very few irregular partitioning schemes for multivariate data. Jiang et al. [JMC*18] proposed a SLIC based irregular partitioning scheme for hyperspectral image classification. While their approach was applied for images with different spectral bands, for our evaluation purpose we treat them as multivariate data. Their proposed approach is called superpixelwise PCA or SuperPCA for short. SuperPCA first creates a global PCA model on the full resolution multivariate data and then apply the SLIC based partitioning scheme for univariate data on the spatial field of the first principal component that captures the maximum variance of the original data. While this approach may work well for variables with less spatial variations, for data sets with varying variable fields, using the first PC field for spatial decomposition may not be a good solution. In this work, we apply SuperPCA approach to our multivariate data and compare with our proposed partitioning approach.

4.1. Evaluation Criteria

To compare how well the multivariate relationships are captured by the individual partitions from different partitioning schemes, we designed a set of evaluation criteria. The primary goal of the partitioning task is to identify spatial regions whose multivariate data can be modeled and hence analyzed by simple linear models. Utilizing classical PCA models for individual spatial partitions, we created the following criteria to understand how well the partitioning schemes operate on multivariate data.

C1: Capturing Maximum Variance: From the perspective of dimensionality reduction, PCA essentially projects the multiple variables to a new set of uncorrelated variables/dimensions in the latent space, called principal components (PC's). The latent variables or PCs are ordered in such a way that the first few retain most of the variation in all of the original variables. The

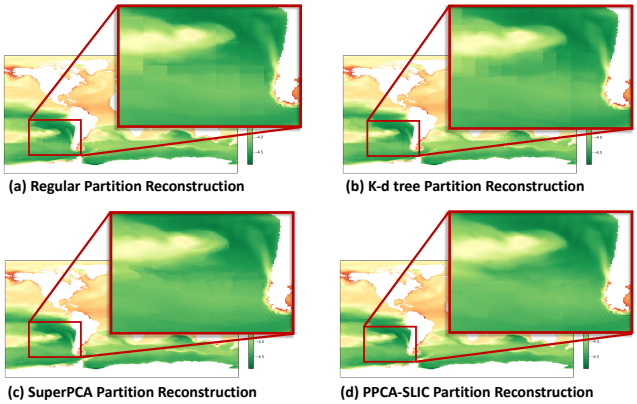

Figure 5: *Visual artifacts of different partitioning schemes after reconstructing the full multivariate data from the the PCA models of individual partitions. Here we show the results for the Fe concentration field.*

amount of variation captured by a set of PCs is also referred to as *explained variance*. This is a useful way to judge how well the multivariate data is modeled by linear models such as PCA. The lower the number of PCs required to capture a certain fraction (say 99%) of the explained variance the better the data is modeled by the PCA model. We use this metric to evaluate the individual partitions in our proposed partitioning scheme.

C2: Multivariate Reconstruction: Another evaluation metric is how well the original high-dimensional (d-dimensional) multivariate fields can be reconstructed back from the low dimensional latent space representations. In our study, as explained above in criteria **C1**, we first identify how many PC's (say q, where $q < d$) are required to capture 99% explained variance of the original data. Next, we reconstruct the full d-dimensional multivariate field from the q latent variable space for each partition. We evaluate the reconstructed fields quantitatively by using measures like root mean-square errors and qualitatively by looking for any visual artifacts induced by the partitioning scheme.

4.2. Evaluation Results

We performed this study on the Ocean BGC data, which comprises of 75 physical variables pertaining information of essential nutrient concentrations besides other organic and inorganic compounds. The spatial resolution of the data is 720×360. We compared the performance of the three partitioning schemes (**P1**, **P2**, **P3**) with our proposed approach based on the evaluation criteria **C1** and **C2** described above. Figure 4 shows the number of PC's required to capture 99% of the explained variance of individual partitions. As can be seen in Figure 4(d), majority of the partitions created by our proposed PPCA guided SLIC approach requires about 10 or 11 PC's (light and deep purple colors respectively) to capture 99% variance of 75 variables. On the otherhand, majority of the partitions for **P1**, **P2**, and **P3** (Figure 4a, 4b, 4c respectively) have relatively higher number of PCs to capture the same variance in each partition. Quantitatively, the respective modes (highest frequency value) for all the four partitioning schemes, **P1**, **P2**, **P3**, and our proposed approach are 13, 13, 13, and 10 respectively. This indicates that the resulting partitions produced by our approach perform well in modeling the underlying multivariate data.

Partitioning Scheme	Avg. Error (RMSE)	Std. Deviation of Error
Regular Partition (P1)	**0.0185**	**0.0022**
K-d Tree (P2)	**0.0171**	**0.0018**
Super PCA (P3)	**0.0155**	**0.0024**
PPCA-SLIC (proposed)	**0.0138**	**0.0020**

Table 1: *Quantitative evaluation of the reconstructed multivariate scalar fields for the four different partitioning schemes. We report the average root mean square error (RMSE) across all the 75 variables and their corresponding standard deviations.*

As mentioned in criteria **C2**, another important factor is how well the multivariate data can be reconstructed from the low dimensional PCA space. Reconstructing the full scalar fields by performing multivariate reconstruction for individual partitions often introduces visual artifacts near the partition boundaries. This is a measure of how much *data-aware* the partitioning scheme was while decomposing the spatial domain. Figure 5 shows the reconstructed field of Fe (iron) concentration for the four partitioning schemes in this work. The zoomed-in images show regions with visual artifacts in the reconstructed fields introduced by the partition boundaries. As shown in Figure 5(a) and 5(b) for **P1** and **P2** respectively, both the regular partitioning schemes display the maximum visual artifacts. Among the irregular partitioning schemes, our proposed approach (Figure 5d) performs slightly better than **P3** (Figure 5c), and much better than **P1** and **P2**.

To get a quantitative understanding of how the partitioning schemes performed during multivariate reconstruction task we calculated the normalized root mean squared error (RMSE) of reconstructed scalar fields for all the variables. The 75 variable fields had different values of RMSE after reconstruction. Table 1 shows the average RMSE scores across all the variables for the four different partitioning schemes along with the standard deviation of these error values. As can be seen, the multivariate reconstruction error is less for our proposed partitioning scheme as compared to the other methods. This is another measure to show that the regions identified by our method can be better modeled by linear models such as PCA.

4.3. Case Study: Linear Correlation Analysis of Essential Nutrients

The resulting partitions generated by our proposed multivariate partitioning scheme can be used to conveniently study variable relationships using simple linear models and measures. We used our proposed partitioning scheme to understand how two of the essential nutrients, namely iron (Fe) and nitrate (NO_3), crucial for macroalgae cultivation, are related (correlated) across the spatial domain i.e, different regions of the ocean mass. For this, we used the Pearson Correlation Coefficient (PCC) between Fe and NO_3 for the individual partitions. PCC is a popular measure of linear correlation between two variables and is effective only when the two variables have a linear relationship. Our proposed partitioning scheme decomposes the spatial domain in such a way that the

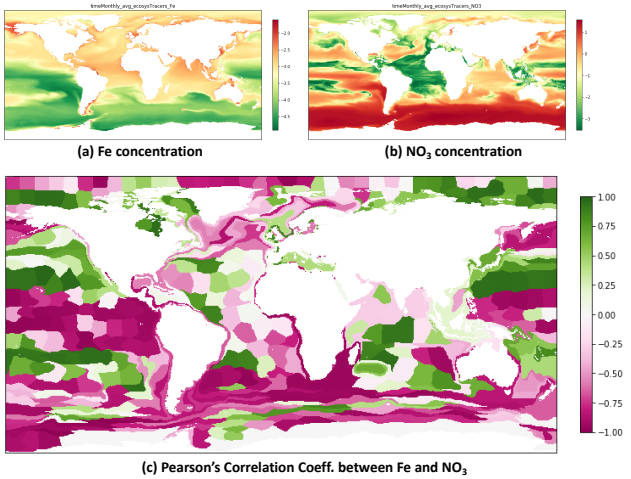

Figure 6: *Case Study: Linear correlation between Fe and NO_3 concentrations using the Pearson Correlation Coefficient for the individual partitions. (c) shows how the two essential nutrients are related across different spatial regions of the ocean.*

variable relationships within each partition can be represented by linear models. Therefore, we can reliably use popular linear measures such as PCC to study variable relationship across the spatial domain without using complex non-linear models.

Figure 6(a) and 6(b) show the scalar fields of Fe and NO_3 concentrations respectively. Figure 6(c) shows the correlation between these two nutrients across different spatial regions via the partitions created by our proposed scheme. Pink color indicates regions with negative correlation between the nutrients and green color highlights regions with strong positive correlations.

Domain Expert Feedback: Our experts feel that strong correlations (both negative and positive) in this case indicate ocean regions where there is no nutrient limitation and hence conducive conditions for growth that are not limited by NO_3 or Fe. They feel that the success of this metric illustrates the capacity to identify key relationships within the physical-BGC system that can be ideally used in the future for change detection algorithms to identify climate change effects on BGC processes. Obviously, correlation is not causation. However, identification of nonlinear tipping points, e.g., as shown via a rapid change in a correlation metric over time, is an important need facilitated via the analysis approach presented herein. As Earth System model process resolution and fidelity increases, our spatial partitioning based analysis will enable identification of key relationships needed to understand Earth System evolution under climate changes.

5. Discussion and Limitations

In this paper, we have highlighted the use-case of studying the linear correlation among variables using the proposed spatial partitioning scheme for multivariate data. Other linear models can also be utilized to perform analysis of the individual partition data. How well such linear models can represent the underlying data of a partition also depend on the size of the partition. Generally, smaller the partition sizes, the better are the chances that the partition data can be modeled by a linear model. In our experiments with different partitioning schemes, we tried to have consistent number of partitions and similar average partition sizes. For the regular partitioning scheme, each block was of size 20×20. With the K-d tree partitioning scheme we stopped dividing the partition further if one of the spatial dimension is less than 20. For both of the two irregular SLIC based methods (superPCA and our proposed method) we initialize with regular partitions of size 20×20. These ensures that the overall partition sizes are not very different across the partitioning schemes during our evaluation study.

We implemented our proposed partitioning scheme using Python programming language on a regular workstation machine with 2.8 GHz Intel Core i7 processor and 16 GB of memory. The overall task of decomposing the spatial domain using **P1**, **P2**, **P3**, and our proposed methods took *0.2 seconds*, *4.1 minutes*, *3.2 minutes*, and *5.8 minutes* respectively. Therefore, the naive regular partitioning schemes is much faster than the other partitioning schemes which are data property driven. However, the data modeling quantity of the regular partitions are not up to the mark with the data-driven schemes as was highlighted in the evaluation studies in Section 4. One limitation of our proposed partitioning scheme is that at the cost of better multivariate modeling quality, the computation time can be slightly higher than other partitioning schemes. However, there is a scope to make our approach faster by parallelizing the algorithm because we adopted a bottom-up approach where we initialize with regular partitions. We can then apply the partition update step of SLIC independently in parallel across the different initial partitions. Parallel execution may not be a straight-forward implementation for both **P2** and **P3** because they involve computing the global PCA of the data which can be a computational overhead if the data is distributed across computational nodes, as is often the case for large-scale simulation models.

6. Conclusion and Future Work

In this paper, we have proposed a new spatial partitioning scheme for multivariate data. We utilized the properties of probabilistic formulation of PCA along with SLIC-based irregular partitioning algorithm to create variable relationship-aware spatial partitions. We extensively evaluated our proposed scheme with other possible multivariate data partitioning schemes on a multivariate Ocean BGC data set.

In future, we plan to improve the computational time of our proposed scheme by looking at ways to parallelize different steps of the algorithm. We also plan to apply our spatial decomposition approach to perform multivariate data reduction for large-scale simulation data by storing summary statistics for the individual partitions.

Acknowledgment

This research was supported by the Laboratory Directed Research and Development program of Los Alamos National Laboratory under project number 20200065DR. PJW was supported as part of the Energy Exascale Earth System Model (E3SM) project, funded by

the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research.

References

[ASS*12] ACHANTA R., SHAJI A., SMITH K., LUCCHI A., FUA P., SÜSSTRUNK S.: Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence 34*, 11 (Nov 2012), 2274–2282. 2, 3

[BDSW13] BISWAS A., DUTTA S., SHEN H.-W., WOODRING J.: An information-aware framework for exploring multivariate data sets. *IEEE Trans. Vis. Comput. Graph. 19*, 12 (2013), 2683–2692. 2

[BGJA07] BETHEL W., GOSINK L., JOY K., ANDERSON J.: Variable interactions in query-driven visualization. *IEEE Transactions on Visualization and Computer Graphics 13* (09 2007), 1400–1407. 2

[DCH*17] DUTTA S., CHEN C. M., HEINLEIN G., SHEN H. W., CHEN J. P.: In situ distribution guided analysis and visualization of transonic jet engine simulations. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (Jan 2017), 811–820. 2

[DS15] DUTTA S., SHEN H.-W.: Distribution driven extraction and tracking of features for time-varying data analysis. *IEEE transactions on visualization and computer graphics 22*, 1 (2015), 837–846. 2

[DWS*17] DUTTA S., WOODRING J., SHEN H. W., CHEN J. P., AHRENS J.: Homogeneity guided probabilistic data summaries for analysis and visualization of large-scale data sets. In *2017 IEEE Pacific Visualization Symposium (PacificVis)* (April 2017), pp. 111–120. 2, 3

[FH09] FUCHS R., HAUSER H.: Visualization of multi-variate scientific data. In *Computer Graphics Forum* (2009), vol. 28, Wiley Online Library, pp. 1670–1690. 1, 2

[GGA*11] GOSINK L., GARTH C., ANDERSON J., BETHEL E., JOY K.: An application of multivariate statistical analysis for query-driven visualization. *IEEE Trans. on Vis. and Comp. Graphics 17*, 3 (2011), 264–275. 2

[HBS18] HAZARIKA S., BISWAS A., SHEN H.: Uncertainty visualization using copula-based analysis in mixed distribution models. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (2018), 934–943. 2

[HBW*20] HAZARIKA S., BISWAS A., WOLFRAM P. J., LAWRENCE E., URBAN N.: Relationship-aware multivariate sampling strategy for scientific simulation data. In *2020 IEEE Visualization Conference (VIS)* (2020), pp. 41–45. 2

[HDSC19] HAZARIKA S., DUTTA S., SHEN H., CHEN J.: Codda: A flexible copula-based distribution driven analysis framework for large-scale multivariate data. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (2019), 1214–1224. 2

[JMC*18] JIANG J., MA J., CHEN C., WANG Z., CAI Z., WANG L.: Superpca: A superpixelwise pca approach for unsupervised feature extraction of hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing 56*, 8 (2018), 4581–4593. 4

[Jol86] JOLLIFFE I. T.: Principal components in regression analysis. In *Principal component analysis*. Springer, 1986, pp. 129–155. 2

[JWSK07] JÄNICKE H., WIEBEL A., SCHEUERMANN G., KOLLMANN W.: Multifield visualization using local statistical complexity. *IEEE Transactions on Visualization and Computer Graphics 13*, 6 (Nov 2007), 1384–1391. 2

[KKW*15] KOCH S., KASTEN J., WIEBEL A., SCHEUERMANN G., HLAWITSCHKA M.: 2d vector field approximation using linear neighborhoods. *The Visual Computer 32* (2015), 1563–1578. 2

[KL97] KAMBHATLA N., LEEN T. K.: Dimension reduction by local principal component analysis. *Neural Computation 9*, 7 (1997), 1493–1516. 2

[MDK*01] MOORE J., DONEY S. C., KLEYPAS J. A., GLOVER D. M., FUNG I. Y.: An intermediate complexity marine ecosystem model for the global domain. *Deep Sea Research Part II: Topical Studies in Oceanography 49*, 1 (2001), 403 – 462. The US JGOFS Synthesis and Modeling Project: Phase 1. 1

[NNN11] NAGARAJ S., NATARAJAN V., NANJUNDIAH R. S.: A gradient-based comparison measure for visual analysis of multifield data. In *Computer Graphics Forum* (2011), vol. 30, Wiley Online Library, pp. 1101–1110. 2

[PADB*19] PETERSEN M. R., ASAY-DAVIS X. S., BERRES A. S., CHEN Q., FEIGE N., HOFFMAN M. J., JACOBSEN D. W., JONES P. W., MALTRUD M. E., PRICE S. F., RINGLER T. D., STRELETZ G. J., TURNER A. K., VAN ROEKEL L. P., VENEZIANI M., WOLFE J. D., WOLFRAM P. J., WOODRING J. L.: An evaluation of the ocean and sea ice climate of e3sm using mpas and interannual core-ii forcing. *Journal of Advances in Modeling Earth Systems 11*, 5 (2019), 1438–1458. 1

[RPH*13] RINGLER T., PETERSEN M., HIGDON R., JACOBSEN D., JONES P., MALTRUD M.: A multi-resolution approach to global ocean modeling. *Ocean Modelling 69* (09 2013), 211–232. 1

[RS00] ROWEIS S. T., SAUL L. K.: Nonlinear dimensionality reduction by locally linear embedding. *Science 290*, 5500 (2000), 2323–2326. 2

[STpS06] SAUBER N., THEISEL H., P. SEIDEL H.: Multifield-graphs: An approach to visualizing correlations in multifield scalar data. *IEEE Transactions on Visualization and Computer Graphics 12*, 5 (Sept 2006), 917–924. 1, 2

[TB99] TIPPING M. E., BISHOP C. M.: Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61*, 3 (1999), 611–622. 2, 3

[WB97] WONG P. C., BERGERON R. D.: 30 years of multidimensional multivariate visualization. In *Scientific Visualization, Overviews, Methodologies, and Techniques* (Washington, DC, USA, 1997), IEEE Computer Society, pp. 3–33. 2

[WHLS19] WANG J., HAZARIKA S., LI C., SHEN H.-W.: Visualization and visual analysis of ensemble data: A survey. *IEEE Transactions on Visualization and Computer Graphics 25*, 9 (2019), 2853–2872. 2

[WKW*17] WANG K., KEWEI LU, WEI T., SHAREEF N., SHEN H.: Statistical visualization and analysis of large data using a value-based spatial distribution. In *2017 IEEE Pacific Visualization Symposium (PacificVis)* (2017), pp. 161–170. 2

[WMB*19] WOLFRAM P. J. J., MALTRUD M. E., BRADY R., BRUS S. R., YANG Z., WANG T.: Multi-resolution, multi-scale modeling of ocean biogeochemistry for scalable macroalgae production. 1

[WYG*11] WANG C., YU H., GROUT R. W., MA K., CHEN J. H.: Analyzing information transfer in time-varying multivariate data. In *2011 IEEE Pacific Visualization Symposium* (2011), pp. 99–106. 2