A Novel Distance Measure for Ocean Reconstruction from Sparse Observations Demonstrated on the Atlantic

Markus Kronenberger[*] TU Kaiserslautern
Christopher Weber UC Davis
Geoffrey Gebbie WHOI
Oliver Kreylos UC Davis
Louise H. Kellogg UC Davis
Lorraine E. Lisiecki UC Santa Barbara
Carlye Peterson UC Santa Barbara
Howard J. Spero UC Davis
Bernd Hamann UC Davis
Hans Hagen TU Kaiserslautern

Abstract

We introduce a distance measure for use in scattered data approximation. Reconstruction from sparse, non-uniformly distributed data should utilize application-specific knowledge to produce high-quality results. Our distance measure is considering the specific problem of computing reconstructions from sparse observational paleoceanography data, where it is possible to consider certain problem-specific knowledge to produce reconstructions of scientific value. Our approach to the problem combines a new distance measure with the well-known moving least squares (MLS) method. We demonstrate that our approach produces high-quality results, by contrasting our distance measure against Euclidean and geodesic distances. We have used our method to generate reconstructions from data in the Atlantic Ocean.

1 Introduction

One of the fundamental challenges in the study of climate change throughout Earth's history is how to combine models of past ocean circulation, as reconstructed from sparse geochemical data collected from deep sea sedimentary cores, with modern ocean circulation data to yield insight into the processes governing ocean circulation in the past. We approach this challenge through the analysis of carbon isotope datasets that have been generated from analyses of microfossils collected from deep sea cores [12].

Most previous synthesis studies of circulation changes during the Last Glacial Maximum (LGM, 20,000 years before present time) focused on geochemical data from benthic foraminifera in only two dimensions, water depth and latitude (e.g., [3]). Compiled and synthesized data in the third (longitude) and fourth (time) dimensions would provide important constraints on overturning rates and water mass boundary variations along flow paths that are not yet available.

Continuous records of foraminifera data are mainly found at ocean depths less than 4.5km, which corresponds to the continental margins along the ocean basin periphery, elevated topographic areas and the mid ocean ridges. With carbon isotope records from ∼480 deep sea cores [12], combining these data to produce a three-dimensional perspective of ocean current circulation through time requires an interpolation methodology that links data across thousands of kilometers in the latitude/longitude domain, and hundreds of meters across the vertical water column. Such a reconstruction would allow to extract topological/structural information. Further, it could be combined with a human-in-the-loop analysis to extract new knowledge about features and processes.

The characteristics of the data (few data points, sparse and non-uniform distribution, differences in length ratios, data points available only at the border of the interpolation space) are challenging for commonly used interpolation schemes, therefore providing the motivation for the research described in this paper. To address the pressing need for specialized schemes handling such a challenging data set, we designed a method that combines a reliable interpolation scheme with a novel distance measure. This distance measure is learned via training performed on the data itself in a preprocessing step. The most important advantages of the method are that it does not need additional information (e.g., a flow field) and has self-optimizing parameters (e.g., compensating differences in length ratios).

2 Related Work

Scattered Data Interpolation: Franke et al. [6] provided an overview of existing interpolation methods and evaluates them on some examples for the reconstruction of surfaces. They noted that real data sets with characteristics similar to those we have (sparse and scattered), for which some of the considered methods perform poorly, exist. Unfortunately they have not investigate this aspect further.

A well-known interpolation scheme for irregularly distributed data is MLS, which was introduced by Levin in [10]. This method produces high-quality results for the approximation of shapes, even when the data is noisy. A comparison of different approaches of MLS is given in [2]. The main reasons that we have chosen this method to demonstrate our novel distance measure are its robust behavior against noise, its flexibility and that it is proven to be useful in a wide range of applications.

Interpolation in the field of Oceanography: Due to the uncertainty in ocean observations, oceanographers recognized early that linear interpolation methods, such as Aitken-Lagrange interpolation [4], were unsatisfactory as they were fitting observational noise. In increasing order of complexity, oceanographers have implemented nudging techniques (e.g., [11]), temporally sequential recursive least-squares methods (i.e., Kalman filter [5]), and whole-domain least-squares methods (i.e., four-dimensional variational data assimilation). The primary drawback with the latter methods is the computational expense of representing oceanic fields and the uncertainties.

More recently, least-squares methods have permitted the reconstruction of global, four-dimensional (i.e., time-varying and spatial, globally) property fields over the last 30,000 years [7] and the LGM [8]. Such methods have succeeded in reproducing the observations within their uncertainty, obey ocean dynamics and boundary conditions including bathymetry, and simultaneously reproduce multiple property fields, but require large memory and computational resources.

3 Method Description

In order to present our novel distance measure we have chosen MLS as underlying fundamental reconstruction scheme. We first briefly outline this method, for the sake of completeness and then give a detailed description of the proposed distance measure.

[*] e-mail: m_kronenbe09@cs.uni-kl.de

3.1 Moving Least Squares

MLS is a commonly used method for the approximation of scattered data in computer graphics [10]. Since it is a well-known technique and not the focus of this work we provide only a short introduction of what we need for the description of our distance measure and the demonstrations in Section 4.

For the purposes of this work we use the general matrix form of MLS given by Shen in [14] to find a function f that approximates the values F of a set of unorganized sample points:

$$f(\mathbf{x}) = b^T(\mathbf{x})(B^T(W(\mathbf{x}))^2B)^{-1}B^T(W(\mathbf{x}))^2F \tag{1}$$

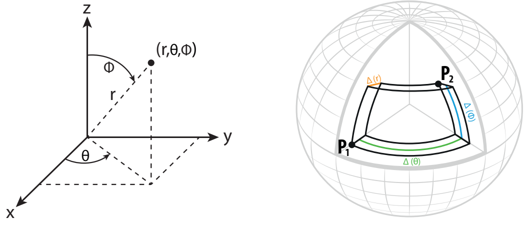
The matrix B contains a number of basis functions b. For example, a linear basis for a three dimensional problem has the functions 1, x, y, and z. W is a diagonal weight matrix depending on weighted distances, which we calculate using the inverse distance function as given in [14]:

$$w(d) = 1/(d^2 + \lambda^2), \tag{2}$$

where λ is a parameter to control the approximation behavior and d is a distance between two points. Obviously the distance measure has a significant influence on the weights and thereby the reconstruction.

3.2 Proposed Distance Measure

A new data-driven distance measure for two points P_1 and P_2, given in spherical coordinates, is presented. It is used in our proposed approach, but could also be integrated in another interpolation scheme that must calculate distances between sample points of spherical nature.



(a) (b)

Figure 1: (a) In the spherical coordinate system the position of a point is described by three variables r, θ and ϕ. (b) Our proposed distance measure is a weighted combination of differences of those components ($\Delta(\theta)$, $\Delta(\phi)$ and $\Delta(r)$).

In spherical coordinates, a point is described by a radius r and two angles θ, ϕ as shown in Figure 1a. If two points $P_1 = (r_1, \theta_1, \phi_1)$ and $P_2 = (r_2, \theta_2, \phi_2)$ are located on the surface of the same sphere, which means that r_1 is equal to r_2, the shortest path between them is part of a great circle. The geodesic distance can for example be computed by the Haversine formula [15]:

$$d_{geodesic}(P_1, P_2) = 2r_1 atan2(\sqrt{Q}, \sqrt{1-Q}) \tag{3}$$

where

$$Q = sin^2((\phi_2 - \phi_1)/2) + cos(\phi_1)cos(\phi_2)sin^2((\theta_2 - \theta_1)/2).$$

Eq. 3 is correct as long as the points lie on the same sphere. If r_1 and r_2 differ one could for example compute the resulting distance as follows:

$$d_{avgGeodesic}(P_1, P_2) = \frac{d_{geodesic}(P_1, P^{122}) + d_{geodesic}(P^{211}, P_2)}{2} \tag{4}$$

where

$$P^{122} = (r_1, \phi_2, \theta_2) \text{ and } P^{211} = (r_2, \phi_1, \theta_1).$$

In applications like oceanography it is not always clear what a distance between two points should be. This is due to the fact that water cannot flow from one position to another by following a straight line or a great circle when this is not implied by the flow field. To this end, a high-quality distance measure would be one that considers the flow as proposed by Streletz et al. [16]. Our method assumes that flow information is not given, and we therefore must restrict our distance measure purely on the data available.

In the absence of flow field information, we therefore consider a weighted combination of differences in the three spherical coordinates r, θ, and ϕ (see Fig. 1b). Our distance between two points P_1 and P_2 is defined by Eq. 5. We consider the following generic points $P^{112} = (r_1, \phi_1, \theta_2)$, $P^{211} = (r_2, \phi_1, \theta_1)$, $P^{212} = (r_2, \phi_1, \theta_2)$ and $P^{122} = (r_1, \phi_2, \theta_2)$. Our new distance measure is defined as:

$$d_{proposed}(P_1, P_2) = \alpha\Delta(\phi) + \beta\Delta(\theta) + \gamma\Delta(\mathrm{r}) \tag{5}$$

where

$$\Delta(\phi) = (d_{avgGeodesic}(P_2, P^{212}) + d_{avgGeodesic}(P^{122}, P^{112}))/2,$$
$$\Delta(\theta) = (d_{avgGeodesic}(P_1, P^{112}) + d_{avgGeodesic}(P^{211}, P^{212}))/2,$$
$$\Delta(\mathrm{r}) = |r_1 - r_2|.$$

A machine-learning pre-processing step is performed to estimate adequate values for the weights α, β and γ, which is similar to a leave-p-out cross-validation. Specifically, this step proceeds as follows: We repeatedly partition the input point set S randomly in two subsets, a training set T and a validation set V. In every iteration i we train α_i, β_i and γ_i on T, so that the resulting reconstruction using MLS has a small root mean square (RMS) error for V. We terminate the process of parameter optimization when n iterations have been carried out. At this point, we assume that the values of the three weights are properly adjusted for the data we are concerned with.

Our optimization procedure has three input parameters: the set S of data points, an integer p that is used to partition S and an integer n that specifies the number of iterations. Additionally a matrix ABG is initialized to store the weights α_i, β_i, γ_i of the individual iteration i. The process is performed n times. In each iteration the following steps are executed. S is randomly partitioned in a training set T and a validation set V so that V has p elements. In our experiments a good choice for p was $|S|/3$. When p is chosen too small it results in oscillating weights of the individual iterations and the method does not converge. After initializing the weights α_i, β_i and γ_i with 1 they are optimized on T to minimize the RMS error for V, using MLS as our approximation method of choice.

The simultaneous optimization of the weight parameters leads to a multidimensional optimization problem and is in general hard to solve. Although more sophisticated approaches could be used, in our case it suffices to optimize α_i, β_i and γ_i sequentially. This is done by repeating the following steps for several times: The values of each parameter are decreased or increased and overwritten by the value that produces the smallest RMS error for the current validation set. This simple procedure suffices to demonstrate the viability of our approach, but it should be improved further. At the end of every iteration the optimized values are added to ABG.

Finally, the medians of the weights considering all iterations that are stored in AGB are determined. This extracts the weights that have produced the smallest errors in the most iterations and therefor are seen as meaningful for S. In contrast parameters leading to the smallest error in one iteration are often only optimal for the current training set T. Using the median reduces the influences of the random configurations.

4 Results

In the following, the performance of our distance measure in combination with MLS is demonstrated for two examples.

4.1 Analytically Defined Data

For our first demonstration we consider a volume element E of a thin spherical shell defined by $6,360\text{km} \leq r \leq 6,371\text{km}$, $0^\circ \leq \theta \leq 45^\circ$ and $0^\circ \leq \phi \leq 45^\circ$. 6,371km is approximately the Earth's radius, ignoring its ellipsoidal shape, and the thickness of this shell approximates the depth of the Mariana Trench, which is the lowest location of the ocean. In summary, we consider the four layers in this volume:

$$g(r,\theta,\phi) = \begin{cases} 1 & r < 6,363\text{km} \\ 2 & 6,363\text{km} \leq r < 6,364\text{km} \\ 4 & 6,364\text{km} \leq r < 6,367\text{km} \\ 5 & r \geq 6,367\text{km} \end{cases} \tag{6}$$

We consider the function g in E and distribute randomly points within this volume. Our goal is to reconstruct all reference points on a $0.5\text{km} \times 1^\circ \times 1^\circ$ grid from these sparse data. We call the set of our input points test data set and the set of points that should be reconstructed, and for which the exact values are available, reference data set. Finally, the RMS error over all grid points except those lying on the border of E is computed.

For the reconstruction we use MLS that relies on distances in the dataset. We compare the Euclidean, the geodesic and the proposed one. Only for our approach a pre-processing step is performed once for every test data set. As described in Section 3 it uses three input parameters. In this case S is the set of the randomly distributed points. The second parameter p divides S in two subsets. An empirically confirmed choice for p is $|S|/3$, which has produced high-quality results in our experiments. The last parameter n defines the number of iterations to perform the training for weights. We set n to 100 for this experiment, which is sufficient considering the small amount of data points. In Section 4.2.2 we investigate how this parameter has to be chosen for a real data set.

Our experiment is divided in two parts. The first part investigates the behavior of the reconstruction with decreasing number of sample points and the second part analyzes the robustness of the results when changing the distribution of a fixed number of points.

For the first part we randomly pick 30 points in E and generate progressively smaller subsets with 30, 25, 20, 15 and 10 data points. MLS together with all three distance measures is used to reconstruct the reference test data set, and the resulting RMS errors are computed. For this test data set our proposed distance measure produces the smallest RMS errors (< 0.3377) in all cases. This is due to the specialization of our method for such problems (sparse data, differences in length ratios, functions with a layer characteristic). The RMS errors (> 1.9376) produced by using Euclidean and geodesic distance are close to each other and relatively stable, but about one order of magnitude larger than the RMS error values obtained with our distance measure.

The question remains whether the results produced with the proposed distance measure are relatively invariant under a change of point distribution. To examine this aspect we generated 50 test data sets with 10 randomly distributed points in each case and computed the resulting RMS errors. It is remarkable that our proposed approach works well in this scenario. Even for extreme distributions it is consistently better when compared to the results computed with the other two distance measures. Only the variability of the resulting errors is larger than in the case of Euclidean and geodesic distance.

4.2 Atlantic Ocean Data

As shown in the previous section our proposed approach works well for an analytically defined data set. However, the actual goal is to reconstruct the LGM δ^{13}C sediment core data set presented in Section 1. This task is more challenging due to large gaps in the data as well as clusters resulting from data collection.

Since the actual LGM data set does not provide us with a ground truth against which the reconstruction results can be evaluated, we cannot use it directly to perform a proof of concept. For this reason, we construct a data set of the modern ocean based on the locations given by the LGM data set. Further, we limit the data set to the Atlantic Ocean, because it is the best-observed ocean basin during glacial times.

4.2.1 Data Preparation

As reference data set, we use a subset of the World Ocean Circulation Experiment (WOCE) oceanography gridded field [9] that was released with uncertainty estimates. WOCE organized, collected, and compiled global ocean observations for the decade from 1988 until 1998. We consider the part between 53°S to 60°N latitude and 98°W to 18°E longitude as representing the Atlantic Ocean. This subset contains points of the Pacific Ocean as well as Mediterranean Sea, which we do not need for the reconstruction of the Atlantic and for that reason we removed them. Since we only have data located at the sea floor it would be challenging to interpret the part between 0m and 1,000m depth, due to the extremely strong mixing in this interval and the interaction with the atmosphere. To this end, we do not include this part in our experiment. Furthermore, we had to use phosphate as substitute tracer instead of δ^{13}C, due to lack of a gridded climatology of modern-day δ^{13}C. Using phosphate makes sense, because it has nearly a linear relation to δ^{13}C [1]. Our resulting reference data set includes 419,623 grid points.

For our test data set we selected all nearest neighbors of the data points included in the LGM data set within our reference data set, resulting in 186 data points. The number of points is smaller than the actual number of core samples, because several points of the LGM data set are mapped to the same neighbors, as well as gaps in the reference data set. This new subset has characteristics similar to the original LGM data set. We use our proposed approach to generate a reconstruction based on these 186 representative core locations of the modern ocean and compare it against the WOCE reference data set, which we assume as ground truth.

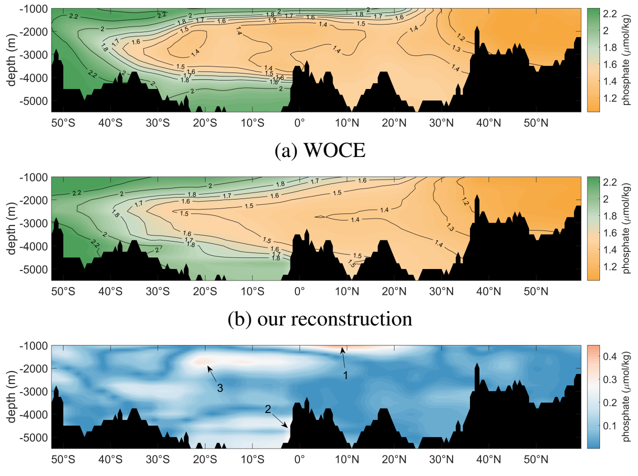
4.2.2 Proposed Approach

A pre-processing step is performed to obtain appropriate values for the weights that are tuned to the actual data set. As described in Section 3 three input parameters are considered for this purpose. The first parameter S is our test data set that we have described in Section 4.2.1 with an amount of 186 data points. For the second parameter p that defines the division of S we already obtained good results with $|S|/3$ in our prior example and for this reason choose it as $186/3 = 62$. The last parameter n, the number of trainings iteration, depends on the complexity of the function that is covered by S. Since this is difficult to capture, we use the following procedure.

We performed our optimization with a relatively high value for n. For all three weight parameters after at least 200 iterations a convergence of the values was recognizable. Therefore, we set n to 200 for this experiment. Note that in the actual experiment we do not train the weights again using n = 200. Should new data points be added to the test data set, this value can be used as a reference. The optimized values for the test data set are α = 0.015, β = 0.011 and γ = 35.57. This clearly shows that the weights are specialized for our problem. The length ratio in the ocean between differences in depth and differences in latitude and longitude, respectively, are very high. Our distance measure compensates for this effect by stretching in depth (= γ) direction.

4.2.3 Discussion

In order to investigate the performance of our reconstruction against the reference data set we do an visual comparison using section plots that cover significant parts of the reconstruction. For the Atlantic Ocean a popular visualization in literature is a North-South



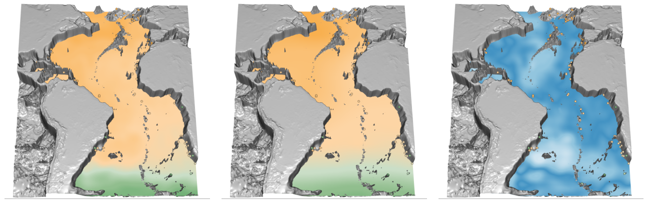
(a) WOCE

(b) our reconstruction

(c) absolute difference

Figure 2: The distribution of phosphate is compared between the WOCE data set and our reconstruction, along a North-South section at 26°W in the modern Atlantic Ocean.

(meridional) section. For this reason we have generated comparable plots for our reconstruction as well as for the reference data set by cutting out a section at 26°W and from 50°S to 60°N (see Fig. 2). To highlight the largest deviations also the absolute difference is given in Fig. 2c. The major part of this plot presents a relatively small difference below 0.1μmol/kg. There are only three areas, denoted with numbers 1 to 3, that exhibit a large difference above 0.4μmol/kg. In the cases of 1 and 2 the reconstruction leaves the convex hull of the test data set and therefore we are extrapolating there. This could be the reason for larger errors in these cases. The other location 3 shows a high difference in the middle of the volume and for this reason we assume that our data points, located only at the ocean floor, did not capture this feature.



(a) WOCE

(b) our reconstruction

(c) absolute difference

Figure 3: Comparison of phosphate distribution at a depth level of 3,000m. Data locations are indicated as spheres. The same colormap as in Fig. 2 is used.

In addition to the meridional section we have evaluated our results at a depth level of 3,000m (see Fig. 3). This visualization also shows the bathymetry and the data locations. As before, the differences overall are small, except one location in the south-west region of the Atlantic. There are two reasons for this, first, this area is lacking sample points at the considered depth level and, second, in the southern Atlantic numerous water masses meet and mix, making a reconstruction based on see floor samples even more challenging.

As a further quality criterion we computed the RMS error between our result and all 419,623 reference points. The reconstruction was computed in ~46 minutes on a laptop (Intel i7 @2.4Ghz, 16GB RAM). We calculated an error of 0.1183μmol/kg. Since our main interest is the reconstruction of δ^{13}C from the sediment core data set [12] we converted the error to pseudo-δ^{13}C using the linear relation given in [1], Eq. 3. This results in an error of 0.1275 per mil, which we seem to be a reasonable result, because it is close to the typical isotope ratio mass spectrometer (IRMS) error on the measurement of $\delta^{13}C$ of 0.1 [13].

5 Conclusions and Future Work

This paper has introduced a novel distance measure that helps to provide accurate reconstructions for sparse and scattered data. Its performance was demonstrated on an analytically defined data set, as well as a real data set. In the first experiment significant improvements of the reconstruction compared to results using Euclidean and geodesic distance were achieved. In the second a relatively small RMS error for the modern Atlantic Ocean was obtained, despite the very few data. These results indicate a promising application of our approach to data sets of other fields e.g., well logging.

It is planned to apply the approach to the actual LGM δ^{13}C sediment core data set and discuss the results in the context of paleoceanography. The primary motivation is to improve the insight into the data and to guide the future collection of cores.

References

[1] W. S. Broecker and E. Maier-Reimer. The influence of air and sea exchange on the carbon isotope distribution in the sea. *Global Biogeochemical Cycles*, 6(3):315–320, 1992.

[2] Z.-Q. Cheng, Y.-Z. Wang, B. Li, K. Xu, G. Dang, and S.-Y. Jin. A survey of methods for moving least squares surfaces. In *Proceedings of the Fifth Eurographics / IEEE VGTC Conference on Point-Based Graphics*. Eurographics Association, 2008.

[3] W. B. Curry and D. W. Oppo. Glacial water mass geometry and the distribution of δ^{13}C of $\sum CO_2$ in the western Atlantic Ocean. *Paleoceanography*, 20(1), 2005.

[4] P. J. Davis, I. Polonsky, M. Abramowitz, and I. A. E. Stegun, editors. *Handbook of Mathemaitc Functions*. Dover, New York, 1965.

[5] J. Derber and A. Rosati. A global oceanic data assimilation system. *J. Phys. Oceanogr.*, 19:1333–1347, 1989.

[6] R. Franke. Scattered data interpolation: tests of some methods. *Math. Comp.*, 38:181–200, 1982.

[7] G. Gebbie. Tracer transport timescales and the observed Atlantic-Pacific lag in the timing of the Last Termination. *Paleoceanography*, 27(3), 2012.

[8] G. Gebbie. How much did Glacial North Atlantic Water shoal? *Paleoceanography*, 29(3):190–209, 2014.

[9] V. Gouretski and K. Koltermann. WOCE Global Hydrographic Climatology. *Berichte des Bundesamt für Seeschifffart und Hydrographie (BSH)*, (35/2004):0–52, 2004.

[10] D. Levin. The approximation power of moving least-squares. *Math. Comput.*, 67(224):1517–1531, Oct. 1998.

[11] P. Malanotte-Rizzoli and W. R. Holland. Data constraints applied to models of the ocean general circulation, Part I: The steady case. *J. Phys. Oceanogr.*, 16:1665–1687.

[12] C. D. Peterson, L. E. Lisiecki, and J. V. Stern. Deglacial whole-ocean δ^{13}C change estimated from 480 benthic foraminiferal records. *Paleoceanography*, 29(6):549–563, 2014.

[13] P. Quay, R. Sonnerup, T. Westby, J. Stutsman, and A. McNichol. Changes in the ^{13}C/^{12}C of dissolved inorganic carbon in the ocean as a tracer of anthropogenic CO_2 uptake. *Global Biogeochemical Cycles*, 17(1):4–1, 2003.

[14] C. Shen. *Building Interpolating and Approximating Implicit Surfaces Using Moving Least Squares*. PhD thesis, Berkeley, CA, USA, 2006.

[15] R. W. Sinnott. Virtues of the Haversine. *Sky and Telescope*, 68(2):159+, 1984.

[16] G. J. Streletz, G. Gebbie, H. J. Spero, O. Kreylos, L. H. Kellogg, and B. Hamann. Interpolating Sparse Scattered Oceanographic Data Using Flow Information. *AGU Fall Meeting Abstracts*, page A1576, Dec. 2012.