Improved estimation of oceanographic climatology using empirical orthogonal functions and clustering

Karl Thomas Hjelmervik
Norwegian Defence Research Establishment (FFI)
Horten, Norway
Email: karl.thomas.hjelmervik@ffi.no

Karina Hjelmervik
Faculty of Technology and Maritime Sciences
Vestfold University College, Norway
Email: karina.hjelmervik@hive.no

***Abstract*—Vertical profiles of temperature, salinity, and sound speed velocity are used in numerous applications where accurate vertical profiles are crucial. Conventional climatological representations of vertical oceanographic profiles are based on mean or median profiles of historic data in a rectangular area containing the position in question. In areas containing oceanographic fronts mean profiles may not be representative for the profiles in the area and may even be unphysical.**

We propose a different approach to generate more realistic climatological estimates of the vertical profiles at a given time and position. The depth-dependent behaviours of all historic temperature and salinity profiles are classified by combining Empirical Orthogonal Function (EOF) analysis with K-means clustering. All profiles with similar EOF-coefficients are sorted into a single cluster and averaged to find a representative profile for that cluster. The geographical extent and temporal validity of the cluster are given by the positions and measurement times of the contained profiles.

The method is here illustrated using ARGO temperature profiles from the North Atlantic from 2001 to 2012. The proposed method automatically allocates a high density of clusters in areas with large oceanographic variability, such as areas with oceanographic fronts.

On the eastern coast of North America cold water from the Labrador Sea runs southwards between the coastline and the warmer Gulf Stream running northeast, resulting in strong fronts. The depth–dependent behaviour of an average profile from all profiles contained in a rectangular, geographic window may differ strongly from the present oceanographic profiles. The profiles representing the nearby clusters, on the other hand, better represent the general depth-dependent behaviour of the profiles in this region.

I. Introduction

In areas dominated by different water masses seperated by fronts, a typical situation in the littorals [1]–[3], estimating representative climatological profiles is a challenging task. Conventional estimates are based on mean or median profiles of historic data in a rectangular area containing the position in question. An example of a climatology database is World Ocean Atlas [4], [5] which uses geographical boxes of either 1^o or 5^o for either annual, seasonal, or monthly temporal resolutions.

A geographical box used for estimating climatological profiles may contain several distinctly different profiles, and since fronts are dynamic [2], the water masses present in a small geographical box may change in the course of a month. The non-Gaussality of the profiles present in such a geographical box may result in an averaged profile which is statistically improbable in that area or even nonphysical.

In the acoustic community, climatological oceanographic profiles are widely used to estimate sound speed profiles for acoustic propagation modeling. The modeled acoustic field is sensitive to errors in the sound speed [6], [7], and particularly to the vertical sound speed gradient [8]. It is therefore vital in these applications that the climatological estimate of the vertical profile has a depth–dependent behaviour similar to the oceanographic profiles expected in the region of interest.

Here we show that a newly proposed method [9] captures the essence of all present types of water. The method employs Empirical Orthogonal Functions (EOF) [10] and k-means clustering [11] to divide a set of historic profiles into different clusters. The clusters replace the rectangularly shaped geographical boxes and are then each associated with average temperature profiles and an averaged position. When a sufficient amount of clusters is used, the statistics for each cluster will be approximately Gaussian [12], and thus the average profiles are more representative for their respective clusters. Similar methods have earlier been demonstrated on modelled oceanography [13], [14] in an area in the Norwegian Trench dominated by fronts due to the interaction of Atlantic water and the Norwegian Coastal Current.

The method is here tested on approximately 87 600 measured oceanographic profiles from the North Atlantic Ocean. The data are ARGO profiles collected and made freely available by the Coriolis project and programmes contributing to it (http://www.coriolis.eu.org).

Comparisons of the proposed and conventional methods are made in order to assess the ability of the proposed method to generate valid climatology for areas dominated by fronts. The presented analysis focuses on an area along the eastern coast of North America. In this area cold water from the Labrador Sea runs southwards between the coastline and the warmer Gulf Stream running northeast, resulting in strong fronts [1], [15], [16, and more].

II. Theory

The depth-dependent behaviour of historic temperature profiles are classified by combining EOF analysis [10] with K-means clustering [11] following [9].

Consider a set of N measured temperature profiles. The temperature measurements are interpolated to selected depth steps and given by the vector $\mathbf{T_n} = [T_n^{(1)}, T_n^{(2)}, \cdots, T_n^{(J)}]$, where n is the measurement number and each element corresponds to a single depth. The geographical latitude and longitude coordinates of the measurements are given by $\mathbf{x_n} = [x_n^{(1)}, x_n^{(2)}]$.

Let the weighted temperature profile and weighted position be given by:

$$\widehat{T}_n^{(j)} = w_T^{(j)} T_n^{(j)} \quad (1)$$

$$\widehat{x}_n^{(i)} = w_x^{(i)} x_n^{(i)} \quad (2)$$

where j corresponds to depth steps of the temperature profile. The weights, w_T and w_x, may be selected freely, but the choice will have a strong influence on the resulting clusters. If a single weight is selected far higher than all other weights, then the corresponding measurement (either a single depth step in the temperature profile or a single geographical coordinate) will have a significantly higher influence on the shape of the resulting clusters. An interesting note is that if the inverted standard deviation $\left(\frac{1}{\sigma}\right)$ for each measurement is used instead, then all depth steps and geographical coordinates will have equal influence on the resulting clusters.

Let the data matrix $\mathbf{P}$ begiven by:

$$\mathbf{P} = \begin{bmatrix} \mathbf{p_1} \\ \mathbf{p_2} \\ \vdots \\ \mathbf{p_N} \end{bmatrix} \quad (3)$$

where $\mathbf{p_n} = [\widehat{\mathbf{T}}_\mathbf{n}, \widehat{\mathbf{x}}_\mathbf{n}]$. EOF and K-means clustering method described in [9] is applied on the matrix $\mathbf{P}$. Each of the measurements may then be described as a sum of the mean value and the computed EOFs times their corresponding coefficients:

$$p_n^{(j)} = \overline{p}^{(j)} + \sum_{k=1}^{K} \kappa_{nk} u_k^{(j)}. \quad (4)$$

$\overline{p}^{(j)}$ is the jth value of the mean data vector. $K = J + 2$ is the length of the data vector $\mathbf{p_n}$. $u_k^{(j)}$ is the jth depth step in the kth EOF with correspoinding coefficient κ_{nk}.

The resulting coefficients are entered into the K-means clustering algorithm in order to group the measurements into clusters. A useful property of EOFs is that the majority of the variance in the data set is represented by the first few coefficients. The clusteringing may be simplified by reducing the number of coefficients used. The Bayesian information criteria [12] is employed to determine both the number of coefficients and clusters used.

The K-mean clustering algorithm sorts the profiles with similar EOF-coefficients into a single cluster and their temperature profiles are averaged to find a representative profile for that cluster. The geographical extent and temporal validity of the cluster are given by the positions and measurement times of the contained profiles.

TABLE I: Number of ARGO profiles from the North Atlantic Ocean. Q1 to Q4 indicate different seasons from Winter (Jan – Mar) to Autumn(Oct – Dec).

Year	Q1	Q2	Q3	Q4	Total
2001	203	320	429	575	1 527
2002	640	937	1 270	1 295	4 142
2003	1 200	1 169	1 175	1 368	4 912
2004	1 341	1 282	1 317	1 342	5 282
2005	1 277	1 272	1 337	1 584	5 470
2006	1 572	1 668	1 898	2 126	7 264
2007	2 138	2 230	2 359	2 567	9 294
2008	2 599	2 588	2 604	2 678	10 469
2009	2 625	2 753	2 531	2 561	10 470
2010	2 479	2 515	2 565	2 905	10 464
2011	3 013	2 717	2 157	1 997	9 884
2012	1 918	1 882	2 008	2 587	8 395
Sum	21 005	21 333	21 650	23 585	87 573

III. Data set

The data set used was collected and made freely available by the Coriolis project and programmes that contribute to it (http://www.coriolis.eu.org). The data set consists of 87 573 ARGO profiles from the North Atlantic Ocean from 2001 to 2012, (see Tab. I).

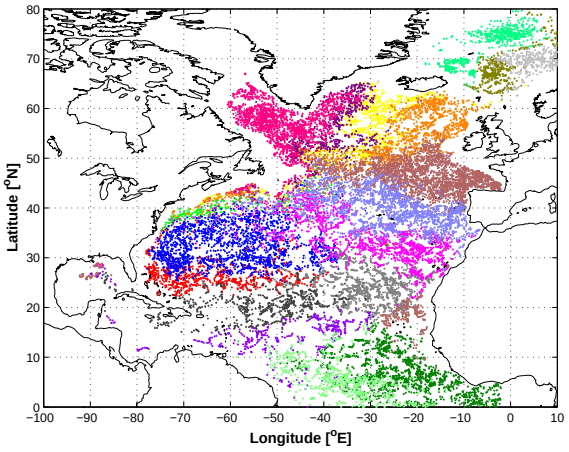
Nonphysical and incomplete profiles are removed. A profile is considered incomplete if it does not contain measurements shallower than 10 m depth and deeper than 500 m depth. Profiles containing temperature measurements below -10^{o}C and above 40^{o}C are considered nonphysical. Likewise for profiles containing salinity measurements below 15 PSU and above 50 PSU. Also, profiles with spikes in temperature (more than 5^{o}C) or salinity (more than 2 PSU) between neighbouring depth samples are considered nonphysical. The remaining profiles are interpolated linearly to the following depths (in meters): 10, 20, 30, 50, 75, 100, 125, 150, 200, 250, 300, 400, and 500.

IV. Results

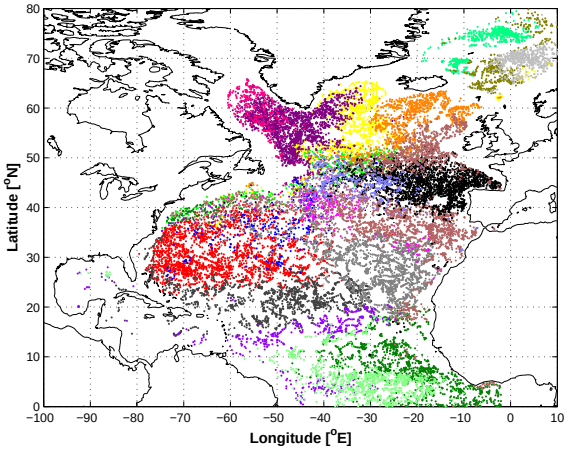
The proposed method is employed on the ARGO data set. 20 clusters and four EOF–coefficients were used in the clustering as determined by the Bayesian information criteria [12]. The geographical distribution of the clusters for the first and third quarter of the year are shown in Fig. 1. The geographical distribution of the clusters vary in time due to heating of the sea surface during the summer season and cooling during winter season.

The proposed method automatically allocates a high density of clusters in areas with large oceanographic variability, such as areas with oceanographic fronts [9]. There is a strong temperature front outside the eastern coast of North America [1], [15], [16, and more]. Due to the presence of many different clusters in a secluded area this front is easily seen in Fig. 1.

The clusters dominating the ocean outside the easter coast of North America vary in time (see Fig. 2). During winter season the vertical profiles closest to the coast are almost constant

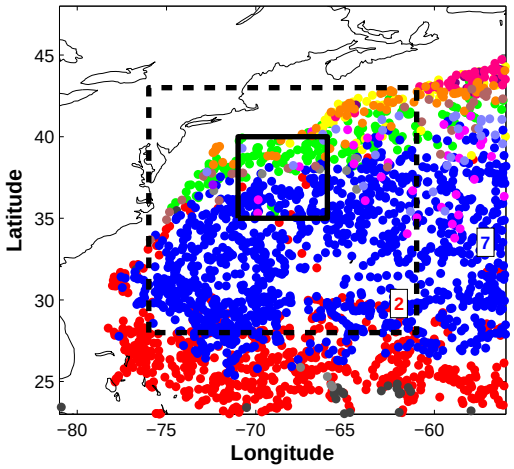


(a) First quarter

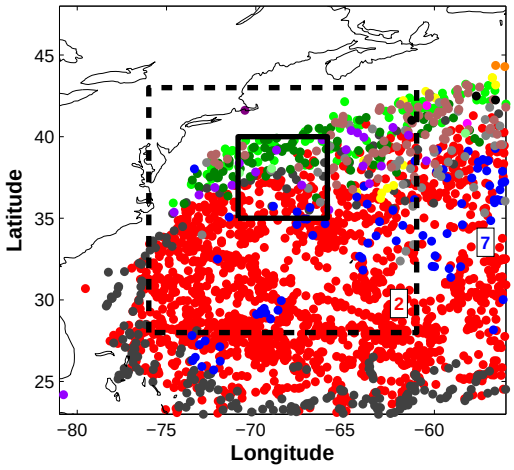


(b) Third quarter

Fig. 1: The geographical distribution of clusters in the North Atlantic Ocean.



(a) First quarter



(b) Third quarter

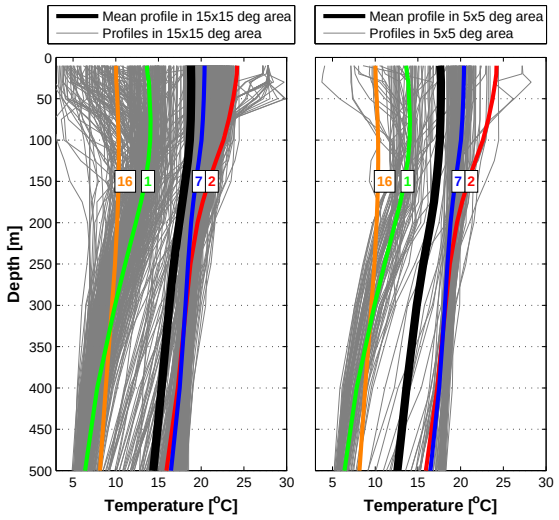
Fig. 2: The cluster distribution at the East Coast of North America. The profiles inside a 15^o x 15^o (dashed) and a 5^o x 5^o (solid) window are studied.

due to mixing of the upper layers, and therefore more similar to the profiles in the cold water in the Labrador Sea than the water further away from the coast (see Fig. 3). During summer season the surface water is more heated and therefore more similar to the water further south.

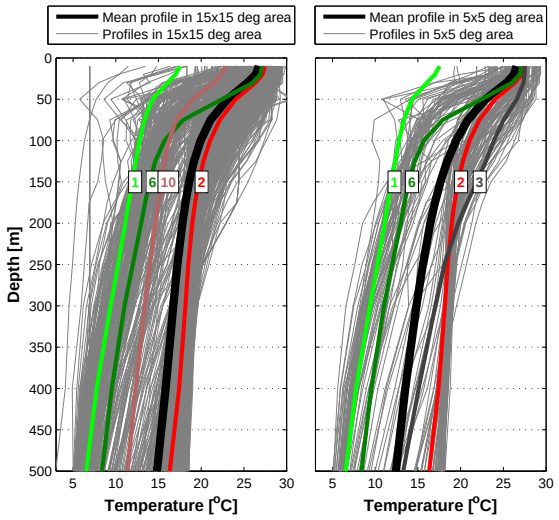
Estimating vertical profiles in an area dominated by fronts is a challenging task because fronts are dynamic. Conventional climatological methods average profiles measured in the vicinity of the desired location. However, the average profile in an area dominated by fronts may fall between different types of profiles (see Fig. 3). During the first quarter cluster number 1 (light green) and 7 (blue) dominate both the 15^o x 15^o and the 5^o x 5^o window outside the eastern coast of North America (see Fig. 4). During the third quarter the surface waters are more heated and thus cluster number 2 (red) and 6 (dark green) dominate. Cluster number 2 and 7 represent the warmer profiles further from the coast and cluster number 1 and 6 represent the colder water closer to the coast. The representative profiles of the two dominant clusters in the area are more probable and physically correct representations of the oceanographic profile at a given location than the mean profiles (see Fig. 3).

Acoustic modeling is an example of an application where correct representation of the vertical oceanographic profile is important. The modeled acoustic field is very sensitive to errors in the sound speed profile [6], [7], which is calculated from temperature, salinity, and pressure profiles.

Fig. 5 shows the modeled, incoherent transmission loss [8] for a source at 50 m depth for sound speed profiles estimated from the temperature profiles of the two most dominant clusters and the mean temperature profile in the $5^o \times 5^o$ window shown in Fig. 3. The acoustic model LYBIN [17] is used to model the transmission loss. For this analysis the salinity value is assumed constant in depth, which makes the sound speed dependent on the hydrostatic pressure and the

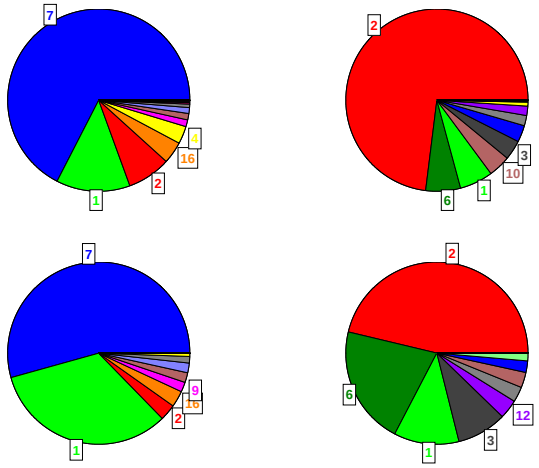


(a) First quarter



(b) Third quarter

Fig. 3: The profiles in the 15^o x 15^o (left) and the 5^o x 5^o (right) window together with the mean profile and representative profiles of the dominant clusters in the area. Only profiles from the relevant season are used to estimate the mean and representative profiles.



(a) First quarter

(b) Third quarter

Fig. 4: The amount of profiles belonging to each cluster in the 15^o x 15^o (upper) and 5^o x 5^o (lower) window at the East Coast of North America.

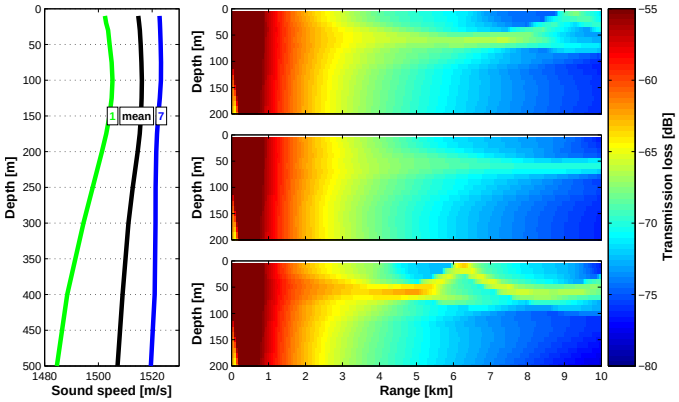


Fig. 5: The sound speed profile (left) and transmission loss (right) corresponding to the mean sound speed profile (black, upper), and the representative profiles for the most (blue, midle) and second most (green, lower) dominant cluster in the $5^o \times 5^o$ window during the first quarter.

temperature profile alone.

The observed differences in the modeled transmission loss, Fig. 5, are due to the differences in the vertical sound speed gradient. Since the source is placed in a surface duct, the calculated transmission loss depends strongly on the local vertical sound speed gradient near the surface [8]. There are also some subtle differences below the surface duct, due to the differences in the vertical gradient in the lower half of the profiles. The representative profile for the most dominant cluster has a local minimum in the sound speed close to the surface, resulting in a surface duct and strong surface interaction. The representative profile for the second most dominating cluster has close to constant sound speed in the upper 100 m, resulting in far less surface interaction. The mean profile contains a weak surface duct with some surface interaction. The transmission loss modeled using the mean profile deviates from the transmission loss modeled using the representative profiles of the two most dominant clusters. This illustrates that the acoustic field is sensitive to the differences observed in the mean profile and the profiles representing the clusters.

V. Conclusion

The proposed method for dividing measured profiles into groups with equivalent statistical attributes have been demonstrated on measured ARGO data. The method employs empirical orthogonal functions and k-means clustering for the grouping. The method is particularly good at estimating physically correct and statistically probable vertical profiles in areas dominated by fronts. Here demonstrated on the northeastern coast of North America, where it is shown to outperform conventional methods.

An application which requires a physically correct representation of the vertical gradient is modeling of underwater propagation of acoustic waves. A simple example using the acoustic model LYBIN demonstrates that the modeled acoustic transmission loss is very sensitive to the choice of temperature profile used in the modeling.

References

[1] C. O. D. Iselin, "A study of the circulation of the western north atlantic," *Papers in Physical Oceanography and Meteorology*, vol. 4, p. 101 pp, 1936.

[2] M. Mork, "Circulation phenomena and frontal dynamics of the norwegian coastal current," *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, vol. 302, no. 1472, pp. 635–647, 1981.

[3] D. S. Ullman and P. C. Cornillon, "Satellite-derived sea surface temperature fronts on the continental shelf off the northeast u.s. coast," *J. Geophys. Res.*, vol. 104(C10), pp. 23 459 – 23 478, 1999.

[4] S. Levitus, Ed., *World Ocean Atlas 2009*. U.S. Government Printing Office, Washington, D.C., 2010, vol. 2: Salinity.

[5] ——, *World Ocean Atlas 2009*. U.S. Government Printing Office, Washington, D.C., 2010, vol. 1: Temperature.

[6] S. E. Dosso, "Environmental uncertainty in ocean acoustic source localization," *Inverse Problems*, vol. 19, no. 2, p. 419, 2003.

[7] K. LePage, "Modeling propagation and reverberation sensitivity to oceanographic and seabed variability," *IEEE J. Oceanic Eng.*, vol. 31, pp. 402–412, 2006.

[8] F. B. Jensen, W. A. Kuperman, M. B. Porter, and H. Schmidt, Eds., *Computational Ocean Acoustics*. Springer Verlag, 2000.

[9] K. T. Hjelmervik and K. Hjelmervik, "Estimating climatological temperature and salinity profiles using empirical orthogonal functions and clusterings," *Ocean Dynamics*, 2012, in review.

[10] R. W. Preisendorfer, *Principal Component Analysis in Meteorology and Oceanography*. Elsevier, 1988.

[11] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipies*, 3rd ed. Cambridge University Press, 2007.

[12] C. Fraley and A. E. Raftery, "How many clusters? which clustering method? answers via model-based cluster analysis," *The Computer Journal*, vol. 41, no. 8, pp. 578–588, 1998.

[13] J. K. Jensen, K. T. Hjelmervik, and P. Østenstad, "Finding acoustically stable areas through empirical orthogonal function (eof) classification," *Oceanic Engineering, IEEE Journal of*, vol. 37, no. 1, pp. 103 –111, jan 2012.

[14] K. T. Hjelmervik, J. K. Jensen, P. Østenstad, and A. Ommundsen, "Classification of acoustically stable areas using empirical orthogonal functions," *Ocean Dynamics*, vol. 62, pp. 253–264, 2012, 10.1007/s10236-011-0499-z.

[15] G. Bearman, Ed., *Seawater: Its composition, properties and behaviour*. Open University, 1997.

[16] M. S. McCartney and C. Mauritzen, "On the origin of the warm inflow to the nordic seas," *Progress in Oceanography*, vol. 51, pp. 125–214, 2001.

[17] E. Dombestein and T. Jenserud, "Improving underwater surveillance: Lybin sonar performance prediction," *Proceeedings of MAST 2010*, 2010.