Noname manuscript No.
(will be inserted by the editor)

Estimating temperature and salinity profiles using empirical orthogonal functions and clustering on historical measurements

Karl Thomas Hjelmervik · Karina Hjelmervik

Received: date / Accepted: date

Abstract Oceanographic climatology is normally estimated by dividing the world's oceans into geographical boxes of fixed shape and size, where each box is represented by a climatological salinity and temperature profile. The climatological profile is typically an average of historical measurements from that region. Since an arbitrarily chosen box may contain different types of water masses both in space and time, an averaged profile may be a statistically improbable, or even nonphysical representation.

This paper proposes a new approach that employs empirical orthogonal functions in combination with a clustering technique to divide the world's oceans into climatological regions. Each region is represented by a cluster that is determined by minimising the variance of the state variables within each cluster. All profiles contained in a cluster are statistically similar to each other, and statistically different from profiles in other clusters. Each cluster is then represented by mean temperature and salinity profiles and a mean position.

Methods for estimating climatological profiles from the cluster information are examined and their performances are compared to a conventional method of estimating climatology. The comparisons show that the new methods outperform conventional methods and are particularly effective in areas where oceanographic fronts are present.

Keywords Oceanography · Climatology · Empirical orthogonal functions · Clustering

K. T. Hjelmervik
Norwegian Defence Research Establishment, 2027 Kjeller, Norway
Tel.: +123-45-678910
Fax: +123-45-678910
E-mail: Karl-Thomas.Hjelmervik@ffi.no

K. Hjelmervik
Faculty of Technology and Maritime Sciences, Vestfold University College, 3103 Tønsberg, Norway

DOI til publisert versjon / DOI to published version: 10.1007/s10236-013-0623-3

1 Introduction

Climatological underwater temperature and salinity profiles are estimated from historic data. Conventional methods divide the world's oceans into geographical boxes of fixed size and shape and average. Climatological profiles are estimated by averaging, or finding the median of, all historic profiles measured within each box. Typically, the historic data set is also divided in time in order to produce an estimate for each month or season. An example of a climatology database is World Ocean Atlas [12,13] which uses geographical boxes of either 1^o or 5^o for either annual, seasonal, or monthly temporal resolutions.

There are obvious advantages of using such rigid methods, particularly in terms of robustness, but also some disadvantages. For example, consider areas dominated by two or more different types of water masses separated by fronts, a not unusual situation in the littorals [7,16,20]. A geographical box used for estimating climatological profiles may contain several distinctly different profiles, and since fronts are dynamic [16], the water masses present in a small geographical box may change in the course of a month. The temperature and salinity distributions in such areas will typically have multiple peaks and high skewness and kurtosis resulting in a statistically improbable or even unphysical averaged profile.

Some applications of climatology require accurate and physical representations of the oceanographic profile rather than a conventional averaged profile. For example, modeling of acoustic wave propagation requires the present sound speed profile [8]. The modelled acoustic field is highly sensitive to errors in the sound speed profile [4,11], which is derived from temperature and salinity profiles [3].

We propose a new method for estimating climatological profiles where the physical and statistical behaviour is preserved. The method employs empirical orthogonal functions (EOF) [19] and clustering [18] 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 and salinity profiles and an averaged position. When a sufficient amount of clusters are used, the statistics of each cluster will be approximately Gaussian [5], which makes the average profile a good representative for that cluster and if used in an acoustic model, the predicted field will be representative for the entire cluster.

For a given geographic position the method outputs several estimates of the climatological salinity and temperature profiles and the probability that they apply for the specified position. This way, the user is made aware if the area is dominated by statistically different types of waters (several nearby clusters with comparable probability), or a single dominant water type (one cluster with very high probability).

The proposed method is tested on temperature and salinity profiles collected and made freely available by the Coriolis project and programmes that contribute to it (http://www.coriolis.eu.org). Comparisons are made to conventional methods of estimating climatology.

EOFs are popular tools in oceanography and have been used extensively in the literature since the 70s [17]. LeBlanc and Middleton [10] employed EOFs to complete sound speed profiles with missing data points using climatological data. EOFs are easily combined with clustering techniques. This combination is much used for classification purposes, *e. g.* in seabed classification [15], and has also been used on modelled oceanographic data [6,9]. Bunkers et al [2] have shown that EOF

Dette er en postprint-versjon / This is a postprint version.
DOI til publisert versjon / DOI to published version: 10.1007/s10236-013-0623-3

and clustering may be used to improve climatological estimates of meteorological data.

2 Method

Let a set of N measured oceanographic profiles, with positions given by $\mathbf{x_n} = (x_n^{(1)}, x_n^{(2)})$, contain both measurements of salinity, $s_n^{(j)}$, and temperature, $t_n^{(j)}$, as functions of depth, where j is the depth step and n is the profile number.

Let the entire set of profiles be split into M clusters with mean position, $\widehat{\mathbf{x}}_\mathbf{m} = (\widehat{x}_m^{(1)}, \widehat{x}_m^{(2)})$, and let mean temperature and salinity at each depth be given by $\widehat{t}_m^{(j)}$ and $\widehat{s}_m^{(j)}$, respectively. The clusters may be of different sizes and each contains N_m profiles, where m indicates the cluster number.

According to Bayes' law the probability that the mth cluster contains the nth profile is given by:

$$P(m|n) = \frac{P(n|m)P(m)}{P(n)}, \tag{1}$$

$P(m)$ is the probability that the mth cluster contains a profile and is simply estimated by:

$$P(m) = \frac{N_m}{N}. \tag{2}$$

$P(n)$ is a normalising factor given by:

$$P(n) = \sum_{m=1}^{M} P(n|m). \tag{3}$$

For a given profile n, $P(n|m)$ may be interpreted as a function of the attributes of the mth cluster. Assume that the cluster positions, temperature profiles, and salinity profiles are independent of each other, then:

$$P_d(n|m) = f_\mathbf{x}(\mathbf{x_m}, \mathbf{x_n}) \prod_{j=1}^{J} f_\mathbf{t}(t_m^{(j)}, t_n^{(j)}) f_\mathbf{s}(s_m^{(j)}, s_n^{(j)}), \tag{4}$$

where $P_d(n|m)$ is the likelihood function; the probability distribution corresponding to $P(n|m)$. $f_\mathbf{x}(\mathbf{x_m}, \mathbf{x})$, $f_\mathbf{t}(t_m^{(j)}, t^{(j)})$, and $f_\mathbf{s}(s_m^{(j)}, s^{(j)})$ are the probability distributions for the positions, temperature profiles, and salinity profiles for the mth cluster. Furthermore, assume independent Gaussian distributions, then:

$$P_d(n|m) = \left((2\pi)^{J+1} \prod_{j=1}^{2} \sigma_{xm}^{(j)} \prod_{j=1}^{J} \sigma_{sm}^{(j)} \prod_{j=1}^{J} \sigma_{tm}^{(j)} \right)^{-1} \tag{5}$$
$$\exp\left[-\frac{1}{2} \sum_{j=1}^{2} \left(\frac{x_n^{(j)} - \widehat{x}_m^{(j)}}{\sigma_{xm}^{(j)}} \right)^2 - \frac{1}{2} \sum_{j=1}^{J} \left(\left(\frac{s_n^{(j)} - \widehat{s}_m^{(j)}}{\sigma_{sm}^{(j)}} \right)^2 + \left(\frac{t_n^{(j)} - \widehat{t}_m^{(j)}}{\sigma_{tm}^{(j)}} \right)^2 \right) \right].$$

Dette er en postprint-versjon / This is a postprint version.
DOI til publisert versjon / DOI to published version: 10.1007/s10236-013-0623-3

By requiring that each profile is initially assigned to a single cluster only, the distribution of clusters may be determined. For our purposes the optimal distribution of clusters is the distribution that maximises the product $\prod_{n=1}^{N} P_d(n|m)$. This may be approximated by minimising the log–likelihood function:

$$\min_{\mathbf{C}} \left[\frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{M} c_{mn} \left(2 \log \left(\prod_{j=1}^{2} \sigma_{xm}^{(j)} \prod_{j=1}^{J} \sigma_{sm}^{(j)} \prod_{j=1}^{J} \sigma_{tm}^{(j)} \right) \right. \right. \tag{6}$$
$$\left. \left. + \sum_{j=1}^{2} \left(\frac{x_n^{(j)} - \widehat{x}_m^{(j)}}{\sigma_{xm}^{(j)}} \right)^2 + \sum_{j=1}^{J} \left(\frac{s_n^{(j)} - \widehat{s}_m^{(j)}}{\sigma_{sm}^{(j)}} \right)^2 + \sum_{j=1}^{J} \left(\frac{t_n^{(j)} - \widehat{t}_m^{(j)}}{\sigma_{tm}^{(j)}} \right)^2 \right) \right].$$

where c_{mn} is unity when the nth profile initially is part of the mth cluster and zero otherwise. $\mathbf{C}$ is a vector containing all cluster parameters, *e. g.* cluster centroids and standard deviations. This nonlinear optimisation problem is not easily solvable, and some simplifying assumptions must be made to process the amount of data needed to generate useful climatology. In our case we assume that the standard deviations of the temperature and salinity profiles are approximately equal for all clusters:

$$\sigma_{xm}^{(j)} \approx \sigma_x^{(j)}, \qquad \sigma_{tm}^{(j)} \approx \sigma_t^{(j)}, \qquad \sigma_{sm}^{(j)} \approx \sigma_s^{(j)},$$

The approximated standard deviations, $\sigma_x^{(j)}$, $\sigma_t^{(j)}$, and $\sigma_s^{(j)}$, are selected according to the application. Since the first term in (6) becomes a constant, the minimisation problem is reduced to:

$$\min_{\mathbf{C}} \left[\sum_{n=1}^{N} \sum_{m=1}^{M} c_{mn} \left(\sum_{j=1}^{2} \left(\frac{x_n^{(j)} - \widehat{x}_m^{(j)}}{\sigma_x^{(j)}} \right)^2 + \sum_{j=1}^{J} \left(\frac{s_n^{(j)} - \widehat{s}_m^{(j)}}{\sigma_s^{(j)}} \right)^2 + \sum_{j=1}^{J} \left(\frac{t_n^{(j)} - \widehat{t}_m^{(j)}}{\sigma_t^{(j)}} \right)^2 \right) \right]. \tag{7}$$

Let the vector $\mathbf{p_n}$ with elements $p_n^{(j)}$ be given by:

$$\mathbf{p_n} = \left[\frac{x_n^{(1)}}{\sigma_x^{(1)}}, \frac{x_n^{(2)}}{\sigma_x^{(2)}}, \frac{s_n^{(1)}}{\sigma_s^{(1)}}, \frac{s_n^{(2)}}{\sigma_s^{(2)}}, \frac{s_n^{(3)}}{\sigma_s^{(3)}}, ..., \frac{s_n^{(J)}}{\sigma_s^{(J)}}, \frac{t_n^{(1)}}{\sigma_t^{(1)}}, \frac{t_n^{(2)}}{\sigma_t^{(2)}}, \frac{t_n^{(3)}}{\sigma_t^{(3)}}, ..., \frac{t_n^{(J)}}{\sigma_t^{(J)}} \right]^T, \tag{8}$$

and

$$\widehat{p}_m^{(j)} = \frac{1}{N_m} \sum_{n=1}^{N} c_{mn} p_n^{(j)}, \tag{9}$$

then (7) may be written as:

$$\min_{\mathbf{C}} \left[\sum_{n=1}^{N} \sum_{m=1}^{M} c_{mn} \sum_{j=1}^{2J+2} \left(p_n^{(j)} - \widehat{p}_m^{(j)} \right)^2 \right]. \tag{10}$$

Let $p_n^{(j)}$ be represented by the weighted sum of a set of EOFs [17], such that:

$$p_n^{(j)} = \overline{p}^{(j)} + \sum_{k=1}^{K} \kappa_{nk} u_k^{(j)}, \tag{11}$$

Dette er en postprint-versjon / This is a postprint version.
DOI til publisert versjon / DOI to published version: 10.1007/s10236-013-0623-3

$u_k^{(j)}$ are the EOFs and κ_{nk} their corresponding weights, also called coefficients. $K = 2J + 2$ is the number of elements of the EOFs. The EOFs are orthonormal, thus inserting (11) into (10) yields:

$$\min_{\mathbf{C}} \left[\sum_{n=1}^{N} \sum_{m=1}^{M} c_{mn} \sum_{k=1}^{K} \left(\kappa_{nk} - \widehat{\kappa}_{mk} \right)^2 \right], \tag{12}$$

$\widehat{\kappa}_{mk}$ is the averaged coefficients representing the profiles in cluster m:

$$\widehat{\kappa}_{mk} = \frac{1}{N_m} \sum_{n=1}^{N} c_{mn} \kappa_{nk}. \tag{13}$$

(12) is a minimization over the sum of all variances of the EOF coefficients for all clusters and may be written as:

$$\min_{\mathbf{C}} \left[\sum_{m=1}^{M} \sum_{k=1}^{K} N_m \sigma_{km}^2 \right], \tag{14}$$

where the standard deviations of the EOF coefficients in each cluster are given by:

$$\sigma_{km} = \sqrt{\frac{1}{N_m} \sum_{n=1}^{N} c_{mn} (\kappa_{kn} - \widehat{\kappa}_{km})^2}. \tag{15}$$

Considering the approximation in (7) and that each of the parameters are normalised by their approximated standard deviation, then c_{mn} is simply reduced to:

$$c_{mn} = \begin{cases} 1, & \sum_{k=1}^{K} \left(\kappa_{kn} - \widehat{\kappa}_{km} \right)^2 < \sum_{k=1}^{K} \left(\kappa_{kn} - \widehat{\kappa}_{ki} \right)^2 \; \forall \; i \neq m \\ 0, & \text{otherwise.} \end{cases}, \tag{16}$$

which is equivalent to placing a profile in the cluster whose centroid is at the lowest Euclidean distance from the profile in coefficient space. K–means clustering is a fast clustering algorithm that solves this problem, but is prone to finding local minima rather than the actual optimum. The clustering algorithm should be repeated several times with different random seeds and the best solution should be chosen. In the example in the following sections, clustering was repeated 120 times and the cluster distribution that resulted in the most parameters ($x_m^{(j)}$, $s_m^{(j)}$, and $t_m^{(j)}$) with Gaussian statistics was selected. The Kolmogorov-Smirnov test with a significance level of 5% was employed to determine whether a parameter was Gaussian.

For a given cluster distribution, the expected values and standard deviations may be approximated from the data. The expected values are then approximated as:

$$\widehat{t}_m^{(j)} \approx \frac{1}{N_m} \sum_{n=1}^{N} c_{mn} t_n^{(j)}, \tag{17}$$

Dette er en postprint-versjon / This is a postprint version.
DOI til publisert versjon / DOI to published version: 10.1007/s10236-013-0623-3

and similar for position, $\widehat{x}_m^{(j)}$, and salinity, $\widehat{s}_m^{(j)}$. The standard deviations, which then replace the approximations in (7), are estimated as follows:

$$\sigma_{tm}^{(j)} \approx \sqrt{\frac{1}{N_m} \sum_{n=1}^{N} c_{mn} \left(t_n^{(j)} - \widehat{t}_m^{(j)} \right)^2}. \tag{18}$$

and similarly for position, $\sigma_{xm}^{(j)}$, and salinity, $\sigma_{sm}^{(j)}$.

2.1 Estimating climatology

Given a set of clusters determined by the method described in the previous section, climatology for a geographic position, $\mathbf{x}$, is estimated. Since the salinity and temperature profiles at $\mathbf{x}$ are unknown, the marginal probability distribution for position must be used to determine the probability that cluster m represents position $\mathbf{x}$. From (1) the marginal distribution for position is given by:

$$P_d(m|\mathbf{x}) = \frac{N_m}{N} \frac{\left(\prod_{j=1}^{2} \sigma_{xm}^{(j)} \right)^{-1} \exp\left(-\left(\frac{x^{(1)} - \widehat{x}_m^{(1)}}{\sqrt{2}\sigma_{xm}^{(1)}} \right)^2 - \left(\frac{x^{(2)} - \widehat{x}_m^{(2)}}{\sqrt{2}\sigma_{xm}^{(2)}} \right)^2 \right)}{\sum_{m=1}^{M} \left(\prod_{j=1}^{2} \sigma_{xm}^{(j)} \right)^{-1} \exp\left(-\left(\frac{x^{(1)} - \widehat{x}_m^{(1)}}{\sqrt{2}\sigma_{xm}^{(1)}} \right)^2 - \left(\frac{x^{(2)} - \widehat{x}_m^{(2)}}{\sqrt{2}\sigma_{xm}^{(2)}} \right)^2 \right)}. \tag{19}$$

Five different methods for estimating the climatological profile at position $\mathbf{x}$ are employed:

1. *Nearest proximity.* The mean salinity and temperature profiles from the cluster with a centroid closest in Euclidean distance to $\mathbf{x}$.
2. *Weighted proximity.* A weighted average of the mean salinity and temperature profiles of all clusters, where the squared Euclidean distance is used as weights.
3. *Most probable.* The mean salinity and temperature profiles from the most probable cluster (maximize $P_d(m|\mathbf{x})$).
4. *Weighted probability.* A weighted average of the mean salinity and temperature profiles of all clusters, where $P_d(m|\mathbf{x})$ are used as weights.
5. *Best fit.* The mean salinity and temperature profiles from the three most probable clusters.

Note that only the latter three methods use the marginal distribution in (19), the two first methods only apply the Euclidean distance in geographic coordinates to determine which cluster to use. Note also that the fifth method actually yields three climatological profiles. In later comparisons, the profile resulting in the best fit with data is used. Clearly, in an operational scenario the user does not know which profile gives the best fit, but the added information of knowing the three most typical *types* of profiles is useful, particularly in areas where fronts are present.

Each of the above mentioned methods are compared to a conventional method for estimating climatology. The area of interest is divided into equally sized geographic boxes and the average of all profiles within each box is the conventional climatological estimate for that box.

Dette er en postprint-versjon / This is a postprint version.
DOI til publisert versjon / DOI to published version: 10.1007/s10236-013-0623-3

3 Example 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 19 701 ARGO profiles from the North Atlantic Ocean from 1. of January to 31. of March between 2001 and 2012, (see Tab. 1).

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°C and above 40°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°) 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, 500} following [12,13]. The depth steps have lower density in the deeper regions since most of the variability is closer to the surface, see Fig. 1. The method requires the same depth sampling in all profiles. A maximum depth of 500 m was chosen. Most of the variance is then included and the profiles in relatively shallow regions are preserved. A maximum depth of *e. g.* 1500 m would result in the loss of approximately 40% of the data due to exclusion of shallower profiles.

The geographical variability is illustrated by the sea surface temperature in Fig. 2. The sea surface temperature in general decreases with increasing latitude, but at some positions the measured sea surface temperature deviates from the surrounding measurements. A stricter filtering of the data would have removed these outliers. A temperature front is observed along the East Coast of North America, which is in agreement with earlier literature on the subject [1,7,14, and more].

Table 1 Number of ARGO profiles from the North Atlantic Ocean during the first quarter of each year

Year	No. of profiles
2001	203
2002	596
2003	1 136
2004	1 320
2005	1 214
2006	1 385
2007	2 034
2008	2 386
2009	2 470
2010	2 370
2011	2 876
2012	1 711
Total	19 701

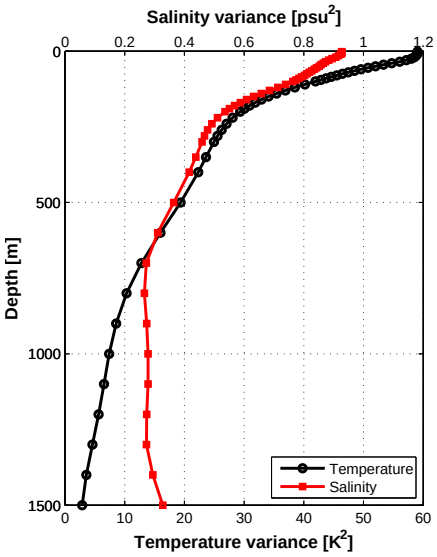

Fig. 1 Variance of temperature and salinity as a function of depth for the entire data set.

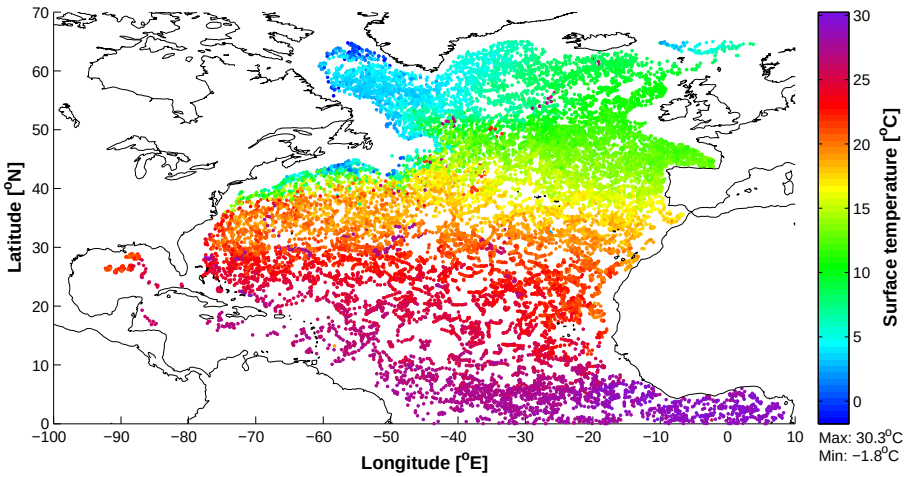

Fig. 2 Sea surface temperature at the positions of the applied profiles.

4 Validation

The validation scheme is divided into two steps. The first step analyses how well the climatology represents the data foundation, while the second step assesses the method's ability to predict future profiles.

The data set is split into two parts. The first part of the data set, henceforth called *historic data*, consists of all data measured in January to March each year

Dette er en postprint-versjon / This is a postprint version.
DOI til publisert versjon / DOI to published version: 10.1007/s10236-013-0623-3

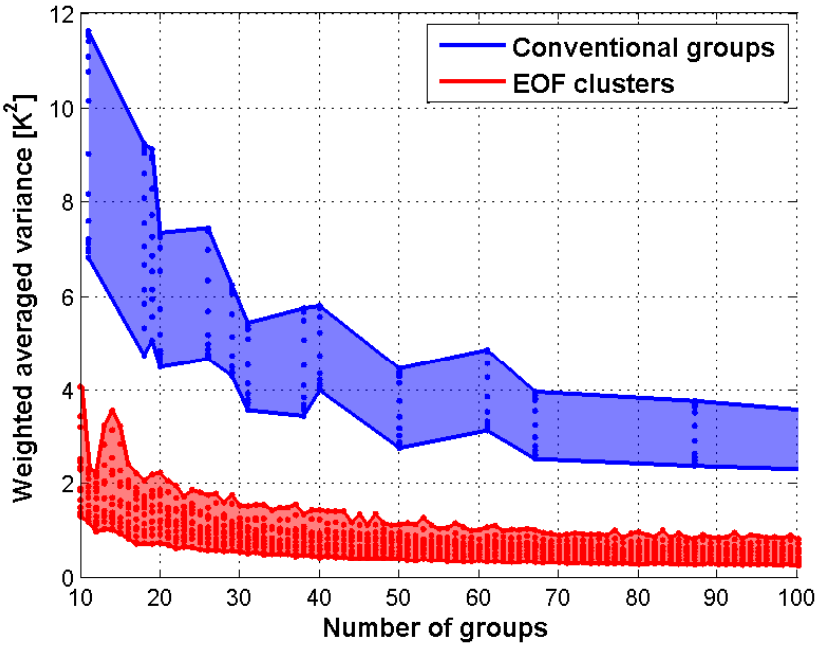

Fig. 3 Weighted averaged variance for temperature as a function of number of groups using both conventional grouping and EOF combined with clustering. The dots represent the weighted averaged variance for each depth separately.

from 2001 to 2011. The second part of the data set, henceforth called *present data*, consists of data measured in January to March 2012.

The EOF and clustering technique described in Sect. 2 is applied on the historic data set. The number of clusters used is varied from 10 to 100. Climatological profiles from the historic data are also estimated using a conventional method as described in section 2.1. The size of the geographic boxes used is varied in order to find climatological estimates comparable to the ones using the proposed method.

Fig. 3 shows the weighted average variance from all groups using both the EOF and clustering technique, and the conventional method. The weighted averaged variance for temperature at a given depth is given by:

$$(\overline{\sigma}_{tm}^{(j)})^2 = \frac{1}{N}\sum_{m=1}^{N_m} N_m(\sigma_{tm}^{(j)})^2 \tag{20}$$

where the standard deviation in temperature for each group, $\sigma_{tm}^{(j)}$, is given in (18). Fig. 3 clearly shows that oceanographic variations in each group are significantly lower when using the EOF and clustering technique than when using the conventional method. The average profile in a cluster is far more representative for the profiles in the cluster, than the average profile in a geographical box is for the profiles contained in that box. Notice also that the variance is larger for shallower depths, which is expected since most geographical and short time scale variations are located in the upper layers.

Dette er en postprint-versjon / This is a postprint version.
DOI til publisert versjon / DOI to published version: 10.1007/s10236-013-0623-3

The variances seem to converge for an increasing number of conventional boxes. Even by reducing the size of each box to a single position, there would still be temporal variations present that will add to the total variance. The clustering on EOF coefficients connects similar profiles in the same cluster regardless of position and time and thus reduces the variance below this limit. The obvious disadvantage with clustering is that the geographical extent of a cluster becomes ambiguous as two profiles measured in the same location at different times may belong to two different clusters.

The methods described in Sect. 2.1 are used to generate a temperature and salinity profile for all profile locations in both the historic and the present data set. The following error function for temperature is used to evaluate the ability of the methods to represent the historic data set:

$$E_{Ht} = \sqrt{\frac{1}{N_H J} \sum_{n=1}^{N_H} \sum_{j=1}^{J} \left(t_{Hn}^{(j)} - t^{(j)}(\mathbf{x_n}) \right)^2}. \tag{21}$$

J is the number of depth steps and N_H is the number of historic profiles. $t^{(j)}(\mathbf{x_n})$ is the jth depth step of the estimated climatological profile and $t_{Hn}^{(j)}$ is the measured temperature profile from the historic data set. The error function used to evaluate the ability of the method to predict profiles is similar:

$$E_{Pt} = \sqrt{\frac{1}{N_P J} \sum_{n=1}^{N_P} \sum_{j=1}^{J} \left(t_{Pn}^{(j)} - t^{(j)}(\mathbf{x_n}) \right)^2} \tag{22}$$

N_P is the number of profiles in the present data set. Equivalent error functions for salinity are also applied.

The resulting error functions are compared in Figs. 4 and 5. Clearly, the methods that rely on spatial proximity (methods 1 and 2) give poorer estimates than the methods using the marginal distribution (methods 3 – 5). For both the historic and present data set the methods using marginal distributions outperform the conventional method for temperature estimates. The salinity estimates have performance equivalent to that of the conventional method. The best fit method (method 5) performs particularly well, but this method assumes that the user is able to pick the best profile out of three choices.

5 Results

Creating climatology based on clustering and EOF is here demonstrated using 26 clusters. The amount of clusters used was determined using the Bayesian Information Criteria [5]. For comparison the conventional climatological method with a resolution of 15° by 15° resulting in 26 groups is also employed.

The average standard deviations over the 26 conventional groups, see Fig. 6, are used as $\sigma_t^{(j)}$ and $\sigma_s^{(j)}$ in (7) applied in the EOF clustering. The standard deviations for latitude and longitude are set to 7° and 14°, respectively. A higher standard deviation in longitude is selected, because larger latitudinal than longitudinal oceanographic variations are observed in the data set, see Fig. 2, thus clusters with higher standard deviation in longitude than latitude provide better

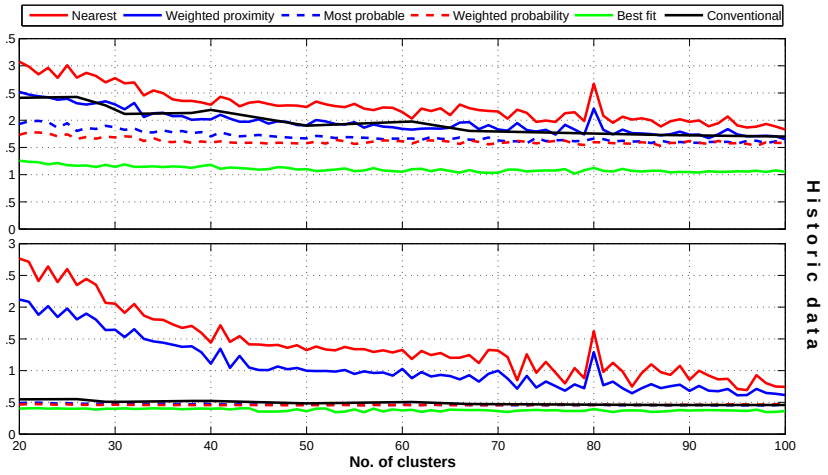

Fig. 4 The historic error function defined in (21) as a function of number of clusters for the different methods of estimating climatology as described in section 2.1.

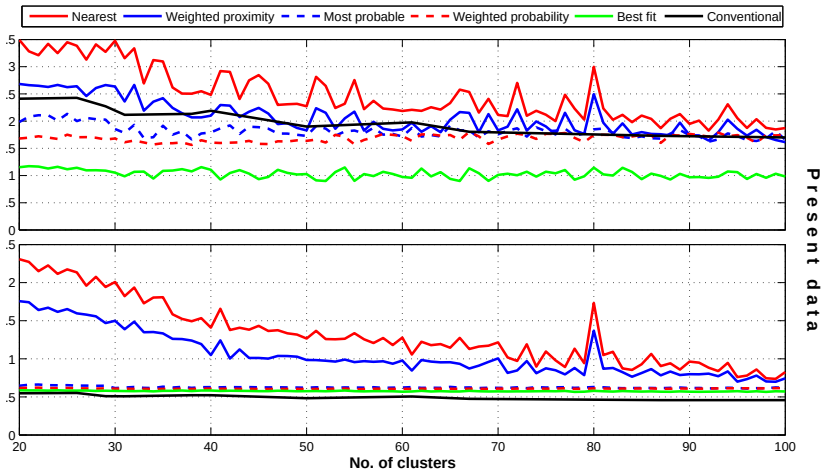

Fig. 5 The predicted error function defined in (22) as a function of number of clusters for the different methods of estimating climatology as described in section 2.1.

climatological estimates. Note that the minimisation in (7) depends on the relative magnitudes of the standard deviations only. By decreasing the selected standard deviation of one parameter only, the weight of that parameter in the minimisation is increased.

According to the proportion of variances five coefficients capture approximately 98% of the variance in the profiles, see Fig. 7. This is sufficient for the purpose of estimating climatology. The first EOF coefficient has the highest variance and contains approximately 75% of the variance. Fig. 8 shows the EOF's derived from the historical data set. Due to the standard deviation chosen in Sec. 4, the absolute value of the first EOF is higher for temperature than for salinity, see Fig. 8. Larger

Dette er en postprint-versjon / This is a postprint version.
DOI til publisert versjon / DOI to published version: 10.1007/s10236-013-0623-3

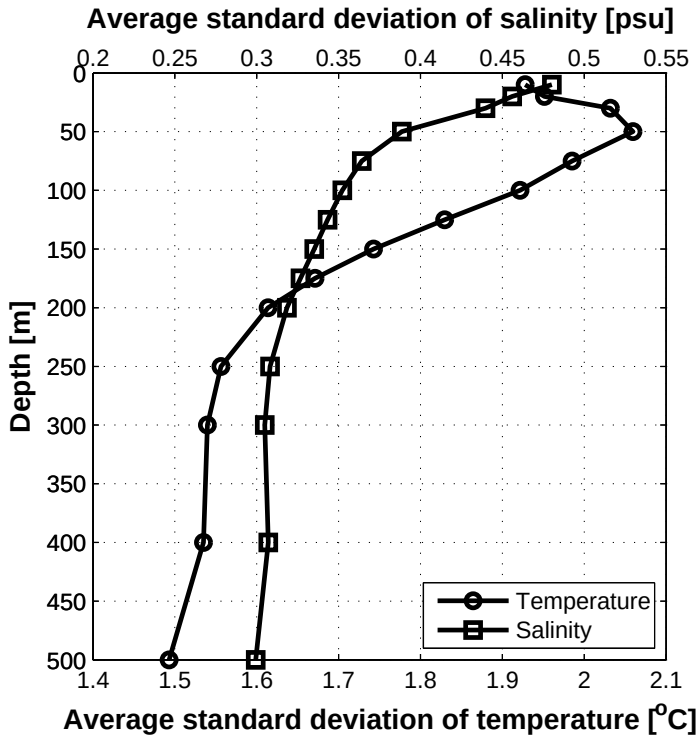

Fig. 6 The standard deviations of temperature and salinity as functions of depth. The standard deviations are averaged over all 26 groups when using conventional methods for dividing the area into boxes (15° by 15°).

EOF values for temperature causes the proposed method to have a larger focus on temperature than salinity when generating clusters, which is the main reason why temperature estimates are more accurate as observed in Sec. 4. The salinity estimates could be improved by decreasing the chosen standard deviation of the salinity, but that would in turn reduce the performance of the temperature estimates. A possible improvement, which is considered outside the scope of this work, is to perform a separate cluster analysis for temperature and salinity, resulting in two sets of clusters and possible improvements in both salinity and temperature estimates.

Fig. 9 shows the distribution of the two first EOF coefficients that represent all profiles in the historic data set. Each coefficient pair is coloured according to what cluster they belong to. For high latitudes the temperature varies less with depth than further south. In order to adjust the gradient of the mean temperature profile, the coefficient corresponding to the first EOF in Fig. 8 have larger values at higher latitudes. The geographical distribution of all profiles is given in Fig. 10. The position of each profile is coloured according to what cluster the profile belongs to. Observe that the clusters are spatially compact, but less compact than what is observed in EOF coefficient space. There is a strong geographical mixing between neighbouring groups, which explains some of the errors observed in

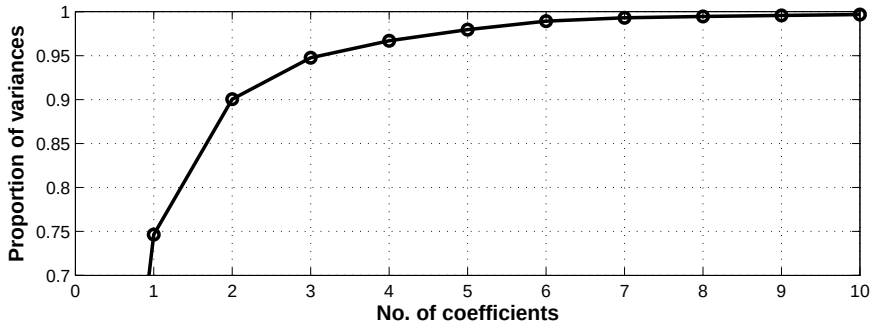

Fig. 7 The proportion of variances for the EOFs using 26 clusters.

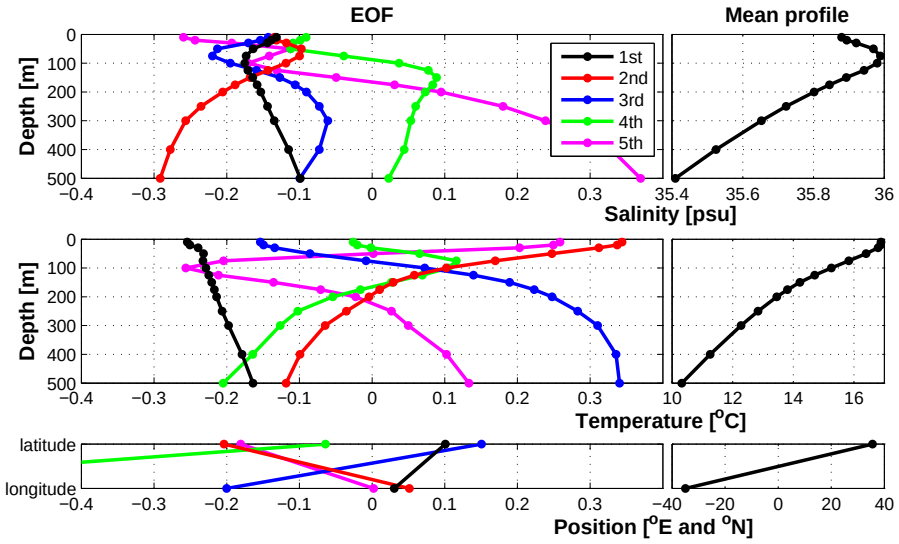

Fig. 8 Left: The contribution from salinity, temperature, and position to the five first EOFs. Right: The mean profiles of salinity and temperature and the mean position used in (11). Note that the EOF analysis is performed on normalised versions of the salinity, temperature, and position, as defined in equation (8).

Sec. 4. By decreasing the selected maximum spatial standard deviation, $\sigma_x^{(j)}$, the mixing could be reduced. Since the idea behind the proposed method is to find clusters characterised by oceanographic homogeneity, one must be careful not to overdo the requirement for the spatial standard deviations. There is a trade-off in the method between spatially contiguous clusters with low mixing (lower spatial standard deviations) and oceanographically homogeneous clusters (higher spatial standard deviations).

The contribution of each profile to the error function (22) for the present data set is plotted geographically in Fig. 11. The results using the conventional method with a $5^o \times 5^o$ box is included for reference. Note that the conventional method has an overall performance that is comparable to the "Best fit" method, except in difficult areas close to the coast, such as north of Great Britain and the East

Dette er en postprint-versjon / This is a postprint version.
DOI til publisert versjon / DOI to published version: 10.1007/s10236-013-0623-3

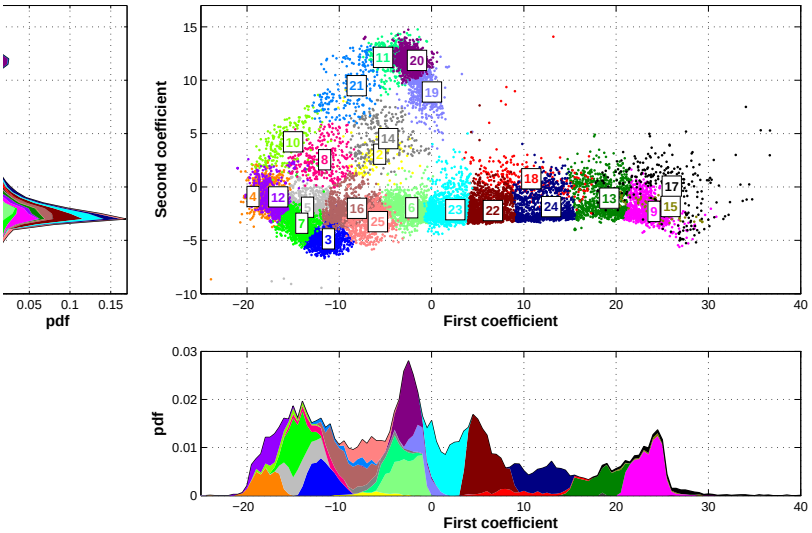

Fig. 9 The first and second EOF coefficients for all profiles in the historic data set with their corresponding probability density functions. The coefficients are here clustered into 26 groups represented by the colours. In the probability density functions, the area of each color represent the ratio the corresponding coefficient represents.

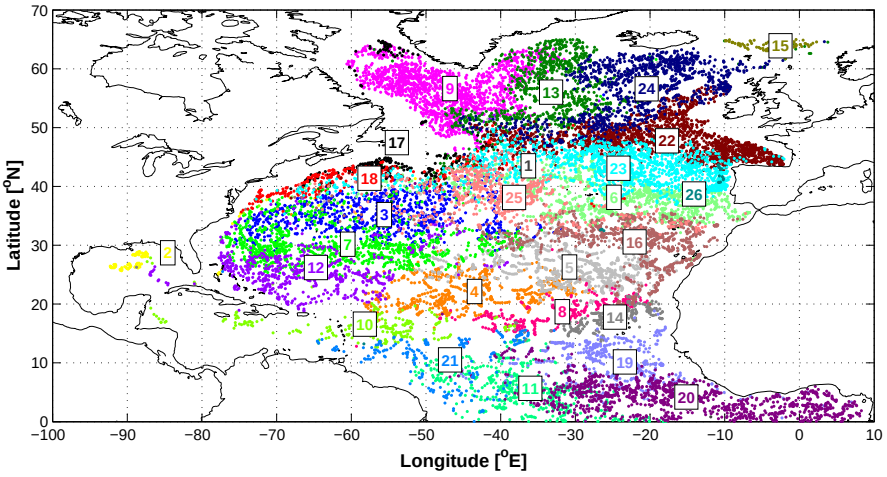

Fig. 10 The geographic position of the clustered profiles when the EOF coefficients are clustered into 26 groups.

Dette er en postprint-versjon / This is a postprint version.
DOI til publisert versjon / DOI to published version: 10.1007/s10236-013-0623-3

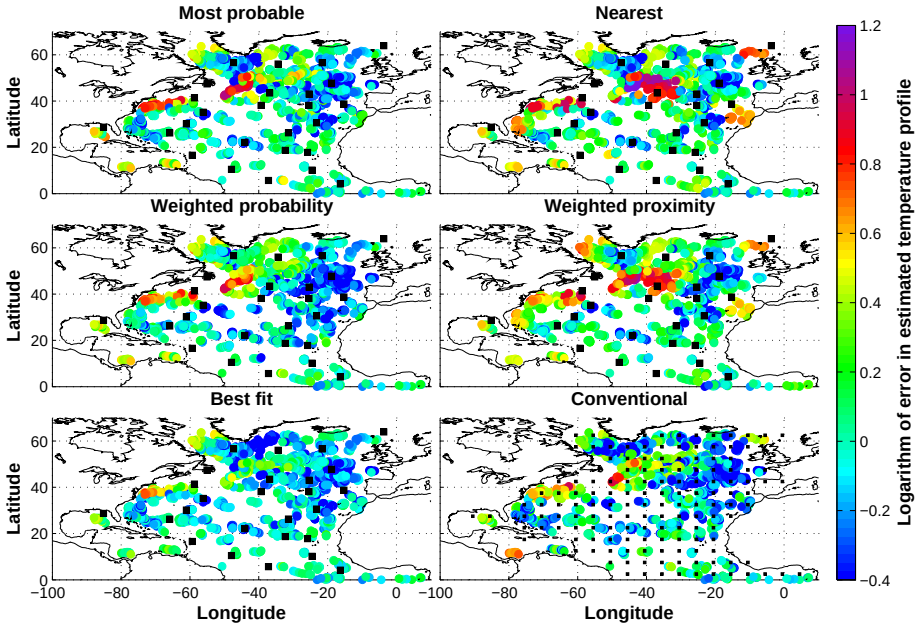

Fig. 11 The logarithm of the RMS error in estimated temperature for the present data set using the six methods described in Sec. 2.1. The profiles are here divided into 26 groups with the centre positions marked with a black square. The conventional method is used on $5^o \times 5^o$ squares whose centers are marked by small black squares.

Coast of North America. The conventional method employs 147 resolution cells, while the proposed method used only 26 clusters.

The three methods using the marginal distribution (methods 3 - 5) to estimate a climatological profile outperforms the methods based on spatial proximity (methods 1 and 2), which is in agreement with Sec. 4. The methods that rely on spatial proximity perform poorly in an area around 40^oW and 45^oN. The main reason is found in Fig. 10 which shows that the nearest cluster center is cluster 1 which consists of outliers, including the profiles with sea surface temperatures that deviate from the surrounding measurements observed in Fig. 2. These profiles all have extreme values in EOF coefficient space beyond the limits of Fig. 9. The methods that rely on proximity is sensitive to outliers when the geographical center of the outliers is close to the position in question. Since only 0.16% of the profiles are included in this cluster, the methods that rely on probabilites are much less affected. This explains some of the observed differences in RMS between the methods that rely on proximity and the methods that rely on probabilities, see Figs. 4 and 5.

In areas dominated by two or more fundamentally different types of profiles the averaging made by conventional methods may result in nonphysical and/or a statistically improbable climatological estimate of the profiles. Fig. 12 shows an example from the East Coast of North America where two different types of water masses are located in the same area. Cold water from the Labrador Sea runs southwards between the coast line and the warmer Gulf Stream running northeast. The average of all profiles inside a $15^o \times 15^o$ box does not represent

Dette er en postprint-versjon / This is a postprint version.
DOI til publisert versjon / DOI to published version: 10.1007/s10236-013-0623-3

the profiles in the area since it falls between the two groups of profiles. The non–Gaussality of the temperature and salinity distributions in the box makes a simple averaging method misleading. Even a reduction of the box size to $5^o \times 5^o$, does not improve the average estimate. Increasing the resolution further will not remove the problem since the fronts separating the different types of water masses in the area are dynamic and therefore measurements in a single position may in time change from one type of water mass to the other.

In areas dominated by fronts any single estimate of the profile would be misleading. The best fit method separates different types of water masses into clusters, and therefore gives reliable climatological estimates of the temperature and salinity profiles for all present water masses. In such cases, presenting different possible profiles with associated probabilities is clearly more useful than presenting a single, averaged profile, which has a very low probability of being an actual profile in such an area.

6 Conclusion

A method for dividing an ocean into climatological regions using empirical orthogonal functions and clustering has been presented and demonstrated on ARGO buoys data for the winter seasons from 2001 to 2011. A set of oceanographic profiles are divided into clusters, where each cluster is represented by a mean position, a mean salinity, and a mean temperature profile.

Different schemes for estimating climatology for a specific geographic position using these clusters were proposed and tested. ARGO buoy data from the winter season in 2012 were then used to validate the method by comparing its climatological estimates to estimates from conventional climatological methods. Some schemes were solely based on the Euclidean distance from the selected position to nearby clusters. These schemes had equal or poorer performance than the conventional method. The remaining schemes, however, employed the marginal probability distribution for geographic position in order to select the most probable clusters, rather than the nearest. These schemes had better performance than the conventional method.

One of the advantages of the proposed method is the ability to estimate different *types* of profiles, where each cluster represents a *type*. The method also estimates the probability that these profiles are representative for a specific geographic position. By offering the user not just a single, but several profiles, he may better understand the present oceanography. It is shown that if the user is able to select the correct type of profile then the performance of the proposed method to estimate climatological profiles far exceeds that of conventional methods, particularly for temperature profiles. An example of an area dominated by different types of waters is given, and in this area an averaged profile is a poor representative for the oceanography. In such areas it is better to present several possible water types, rather than an average profile that is statistically improbable and possibly even nonphysical.

DOI til publisert versjon / DOI to published version: 10.1007/s10236-013-0623-3

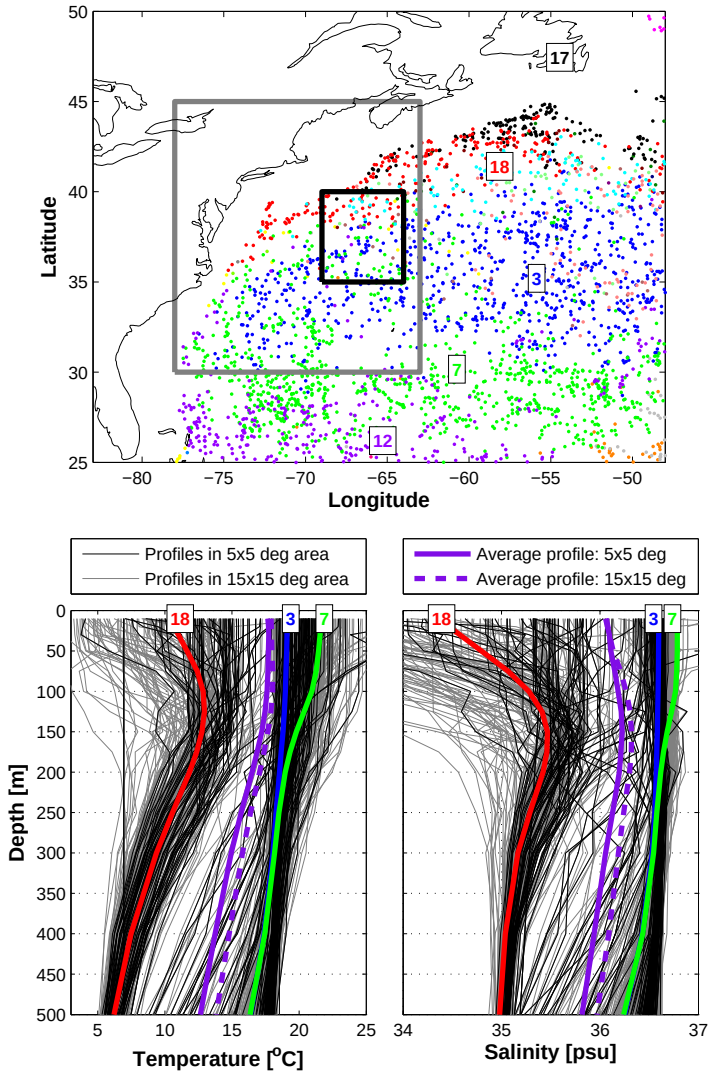

Fig. 12 The plot shows the profiles inside the 15^o x 15^o box (grey) and 5^o x 5^o box (black) and their respective average profiles (purple), together with the average profiles from nearby clusters (the cluster numbers are indicated in the plot).

References

1. Bearman, G. (ed.): Seawater: Its composition, properties and behaviour. Open University (1997)
2. Bunkers, M.J., Jr, J.R.M., Degaetand, A.T.: Definition of climate regions in the northern plains using an objective cluster modification technique. Journal of Climate **9**, 130–146 (1996)
3. Chen, C.T., Millero, F.J.: Speed of sound in seawater at high pressures. J. Acoust. Soc. Am. **62**, 1129 – 1135 (1977)

Dette er en postprint-versjon / This is a postprint version.
DOI til publisert versjon / DOI to published version: 10.1007/s10236-013-0623-3

4. Finette, S.: A stochastic representation of environmental undertainty and its coupling to acoustic wave propagation in ocean waveguides. J. Acoust. Soc. Am. **120**, 2567–2579 (2006)
5. Fraley, C., Raftery, A.E.: How many clusters? which clustering method? answers via model-based cluster analysis. The Computer Journal **41**(8), 578–588 (1998)
6. Hjelmervik, K.T., Jensen, J.K., Østenstad, P., Ommundsen, A.: Classification of acoustically stable areas using empirical orthogonal functions. Ocean Dynamics **62**, 253–264 (2012). 10.1007/s10236-011-0499-z
7. Iselin, C.O.D.: A study of the circulation of the western north atlantic. Papers in Physical Oceanography and Meteorology **4**, 101 pp (1936)
8. Jensen, F.B., Kuperman, W.A., Porter, M.B., Schmidt, H.: Computational Ocean Acoustics. Springer Verlag (2000)
9. Jensen, J.K., Hjelmervik, K.T., Østenstad, P.: Finding acoustically stable areas through empirical orthogonal function (eof) classification. Oceanic Engineering, IEEE Journal of **37**(1), 103 –111 (2012)
10. LeBlanc, L.R., Middleton, F.H.: An underwater acoustic sound velocity data model. J. Acoust. Soc. Am. **67 (6)**, 2055 – 2062 (1980)
11. LePage, K.: Modeling propagation and reverberation sensitivity to oceanographic and seabed variability. IEEE J. Oceanic Eng. **31**, 402–412 (2006)
12. Levitus, S. (ed.): World Ocean Atlas 2009, vol. 1: Temperature. U.S. Government Printing Office, Washington, D.C. (2010)
13. Levitus, S. (ed.): World Ocean Atlas 2009, vol. 2: Salinity. U.S. Government Printing Office, Washington, D.C. (2010)
14. McCartney, M.S., Mauritzen, C.: On the origin of the warm inflow to the nordic seas. Progress in Oceanography **51**, 125–214 (2001)
15. Milligan, S.D., LeBlanc, L.R., Middleton, F.H.: Statistical grouping of acoustic reflection profiles. J. Acoust. Soc. Am. **64**(3), 795–807 (1978)
16. Mork, M.: Circulation phenomena and frontal dynamics of the norwegian coastal current. Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences **302**(1472), 635–647 (1981)
17. Preisendorfer, R.W.: Principal Component Analysis in Meteorology and Oceanography. Elsevier (1988)
18. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipies, 3rd edn. Cambridge University Press (2007)
19. Therrien, C.W.: Discrete Random Signals and Statistical Signal Processing. Prentice Hall (1992)
20. Ullman, D.S., Cornillon, P.C.: Satellite-derived sea surface temperature fronts on the continental shelf off the northeast u.s. coast. J. Geophys. Res. **104(C10)**, 23,459 – 23,478 (1999)