Building an Expert System Module for World Ocean Thermocline Analysis: Thermocline Qualitative Modeling

Marin Trošelj[1] , Maja Matetić[2] and Gordan Grgurić[3]

Department of Polytechnics[1]
University of Rijeka, Faculty of Arts and Sciences
Omladinska 14, 51000 Rijeka, Croatia
Phone: (051) 345 046 E-mail: marin.troselj@gmail.com

Department of Informatics[2]
University of Rijeka,
Omladinska 14, 51000 Rijeka, Croatia
Phone: (091) 503 13 27 E-mail: maja.matetic@ri.t-com.hr

Marine Science Program[3]
The Richard Stockton College of New Jersey
PO Box 195, Pomona, NJ 08240-0195, USA
Phone: (609) 652-4492, E-mail: iaprod730@stockton.edu

Abstract - As the amount of useful data is evolving, we need to create adequate tools which could be used in data analysis and interpretation. We are dealing with the knowledge mining problem concerning World Ocean Atlas 2005 (WOA) which is a data product of the National Oceanographic Data Center (U.S.). In our approach we propose methods such us qualitative modeling and conceptual clustering. The aim is to create an expert system module for the analysis of world ocean thermocline. Expert background knowledge guides the modeling process to the desired thermocline qualitative model which incorporates interesting chunks of knowledge. Learned qualitative thermocline models can be further used in analysis of other data from WOA 2005 related to other oceans and seas.

I. INTRODUCTION

As the data gained by measurements is evolving, we need to develop tools which could be used for data analysis and interpretation. The aim is to analyze world ocean thermocline at different geographic locations. Thermocline analysis requires model building of varying thermocline types. Learning of thermocline models is based on data from the World Ocean Atlas 2005 (WOA) which is a data product of the National Oceanographic Data Center (U.S.) [2,11].

Many water ecosystems are endangered by human actions, in spite of their importance for all living systems. Qualitative models of water ecosystem and related physical phenomena may be useful for understanding such systems, for predicting values of variables and for combining such understanding with restoration and proactive management [8].

In our approach we propose modeling methods such as qualitative modeling and conceptual clustering [1, 3, 5, 6, 7].

First we present thermocline definition and the motivation for thermocline modeling. The third and fourth paragraphs give data description and describe data pre-processing procedure. The fifth paragraph introduces the ideas regarding conceptual clustering procedure and learning of varying thermocline types models. At the end we give an overview of the conceptual clustering and the analysis of the obtained results. We conclude giving the plan for future work.

II. DEFINITION OF THERMOCLINE

Surface sea layer is an essential element of heat and freshwater transfer between the atmosphere and the ocean. This is the mixed layer which absorbs heat during spring and summer by storing it until the following autumn and winter, and thus moderating the earth's seasonal temperature extremes (Fig.1, Fig 2 [9]).

The deep mixed layer from the previous winter is covered by a shallow layer of warm water. Mixing is achieved by the action of wind waves, which cannot reach much deeper than a few tens of meters [9]. Below the layer of active mixing is a zone of rapid transition, where temperature decreases rapidly with depth. This transition layer is called the seasonal thermocline.

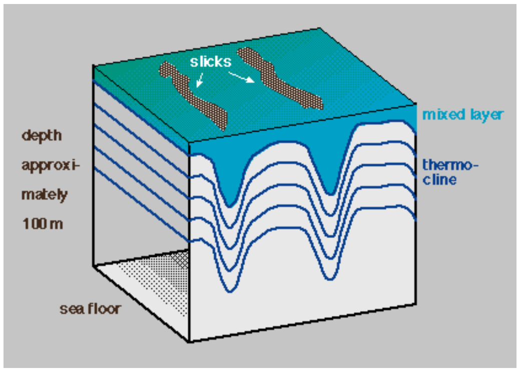

Fig. 1. An illustration of a thermocline [9].

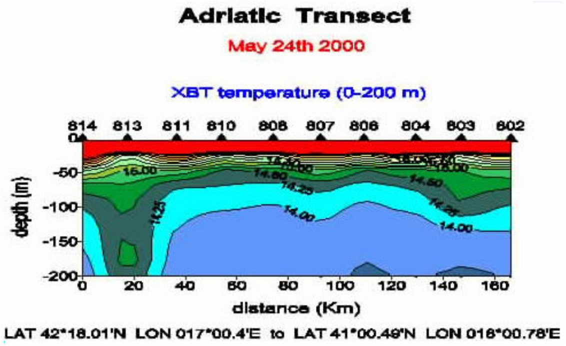

Fig. 2. Surface heat flux data show that the winter was strong in the year 2000 in the Adriatic region [4]

The depth range from below the seasonal thermocline to about 1000 m is known as the permanent or oceanic thermocline [9]. It is the transition zone from the warm waters of the surface layer to the cold waters of great oceanic depth. The temperature at the upper limit of the permanent thermocline depends on latitude, reaching from well above 20°C in the tropics to just above 15°C in temperate regions; at the lower limit temperatures are rather uniform, around 4 - 6°C depending on the particular ocean [9].

North Atlantic Deep Water is the product of a process that involves deep convection in the Arctic Ocean, the Greenland Sea and the Labrador Sea.

Deep ocean is a large part of the ocean characterized by restricted water exchange with the surface ocean. This results in different hydrodynamics and sets it apart from the surface ocean. While the circulation in surface ocean is dominated by wind-driven currents, the circulation in deep ocean is determined by thermohaline processes. Water renewal below 1000 m is achieved by currents which are driven by density differences produced by temperature (thermal) and salinity (haline) effects. The associated circulation is therefore referred to as the thermohaline circulation [9].

III. Data Description

The knowledge mining problem we deal with concerns the World Ocean Atlas 2005 (WOA 2005) which is a data product of the National Oceanographic Data Center (U.S.) [2]. WOA 2005 is available on web pages [11] where we can use the select tool and choose parameter and the region of interest. We build the thermocline model for North Atlantic region and temperature is the parameter which determines a thermocline [Fig.3].

Temperature values are measured at different depths so we have a sequence of measured temperature values at every geographic location. Ocean depths at which measurements are taken are included in set D and are expressed in meters:

$$D = \{0, 10, 20,30, 50, 75, 100, 125, 150, 200, 250, 300, 400, 500, 600, 700, 800, 900, 1000, 1100, 1200, 1300, 1400, 1500, 1750, 2000, 2500, 3000, 3500, 4000, 4500, 5000, 5500\} \qquad (1)$$

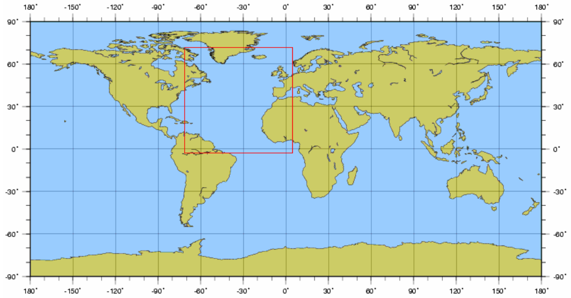

Fig. 3. North Atlantic region [11].

The length of temperature sequence depends on total depth at the given location. A temperature sequence at a geographic location lat_i, long_j is represented with an n-tuple of the following form:

$$((t1, d1), (t2, d2), ..., (tn-1, dn-1), (tn, dn)), \text{ where } d1,..., dn \in D \qquad (2)$$

where i=-70,..., -3 deg. latitude, j=-72,..., -5 deg. longitude, only for geographic locations representing North Atlantic.

So thermocline is a depth at which the rate of temperature decrease with the increase of depth is the largest. Implementing the transformation of temperature sequences we obtain sequences of vertical temperature gradients $\nabla t = \partial T/\partial z$ which incorporate thermocline property (∂z is increase of depth):

$$((\nabla t2, d2), (\nabla t3, d3), ..., (\nabla tn-1, dn-1), (\nabla tn, dn)), \text{ where } d2,..., dn \in D$$

$$\text{where } \nabla ti =(ti - ti-1)/(di - di-1), i = 2,..., n \qquad (3)$$

IV. Data Preprocessing

Besides evident knowledge concerning physical world, qualitative model also represents associated abstractions used by an expert in creating models [3]. Qualitative representation of temperature gradient is symbolic and uses discrete value spaces. Discretization must be relevant for object being modeled, namely discrete values are imposed only if they are essential in modeling of a distinct domain aspect regarding the foreground task.

We base thermocline modeling and analysis on quantization of the temperature gradient. The use of approximate temperature gradient values instead of exact numeric values is the first step anticipating the procedure of thermocline classes formation. Approximate values have to be conceptually relevant for creating thermocline models. In the first step data preprocessing algorithm computes temperature gradient minimum and maximum value searching through all temperature gradient sequences regarding the North Atlantic basin data.

Thermocline model building is based on qualitative temperature gradient as the modeling attribute. The qualitative transformation of temperature gradient sequences is performed in two steps.

A. Transformation of temperature sequences

On the base of temperature sequences at different geographic locations (NODC data, Eq.2, 4), pre-processing procedure results with temperature gradient sequences (Eq. 5).

```
Deg.Latitude: -3.5  Deg.Longitude: -34.5
Temperature sequence:
27.312  27.313  27.284  27.25  26.961  25.633
22.442  18.172  15.359  12.968  11.735  10.731
8.703  7.013  5.787  5.026  4.545  4.29  4.265
4.341  4.427  4.441  4.38  4.246  3.856  3.499
3.003  2.704  2.53  1.722                    (4)
```

```
Deg.Latitude: -3.5  Deg.Longitude: -34.5
Temperature gradient sequence:
9.99451e-005   -0.00289993     -0.00340004
-0.01445       -0.05312        -0.12764
-0.1708 -0.11252       -0.04782       -0.02466
       -0.02008       -0.02028       -0.0169
-0.01226       -0.00761        -0.00481
-0.00255       -0.000250001    0.000760002
       0.00086 0.000139999     -0.000609999
-0.00134       -0.00156        -0.001428
-0.000992      -0.000598       -0.000348
-0.001616
                                              (5)
```

B. Quantization of temperature gradient value

Temperature gradient value is quantisied in six intervals. The set of discrete values is defined as V = {s, w, n, f, i, r}.

Interval ranges are defined as follows ([-0.28, 0.28] → V):

1. interval [0.18, 0.28), denoted as 's';
2. interval [0.09, 0.18), denoted as 'w',
3. interval [0, 0.09), denoted as 'n',
4. interval [-0.023, 0), denoted as 'f',
5. interval [-0.046, -0.023), denoted as 'i',
6. interval [-0.08, -0.046), denoted as 'r'. (6)

For illustration we give a transformation example. On the base of temperature gradient sequence at the given location (Eq.5), qualitative temperature gradient sequence is:

fnnnnwwwnnnnnnnnnnnfffnnnnnnnn (7)

C. Transformation of qualitative temperature gradient sequences

We are interested only in thermocline type dominating at a vertical location. In order to emphasize the thermocline type, the temperature gradient sequences are transformed. The algorithm pseudocode description follows.

```
for i=1:3842 {
 for j=1:32 {
  if z[i,j]==s {
   p1=i p2=j p3=i p4=j

   while (z[p1,p2-1]==w) {
   p2=p2-1
   z[p1,p2]=s
   }

   while (z[p3,p4+1]==w) {
   p4=p4+1
   z[p3,p4]=s
   }
  }

else

  if z[i,j]==r {
   p1=i p2=j p3=i p4=j

   while (z[p1,p2-1]==i) {
   p2=p2-1
   z[p1,p2]=r
   }

   while (z[p3,p4+1]==i) {
   p4=p4+1
   z[p3,p4]=r
   }
  }
 }
}
```

For the North Atlantic basin we have 3842 qualitative temperature gradient sequences consisting of symbols 'n', 'w', 'i', 'r', 's'. Symbols 's' and 'r' represent dominant sequence symbols determining thermocline type. Transformation examples are:

wsw --> sss
swwwwws--> ssssss
wwwws --> ssss

and

iirrii --> rrrrrr
irrrrr --> rrrrr
iiiiiir --> rrrrrrr

For input qualitative temperature gradient sequences

nnnnnnnnnnnnwsswnnnnn
iiiiirnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnwww

the resulting sequences are:

nnnnnnnnnnnnssssnnnnn
rrrrrrnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnwww

V.THERMOCLINE CONCEPTUAL CLUSTERING AND LEARNING OF THERMOCLINE QUALITATIVE MODELS

So on the base of expert background knowledge and data from WOA 2005 for the North Atlantic region, we want to learn models of varying thermocline types. Selected North Atlantic region is large enough to include all thermocline types encompassing regions from the North Sea to tropic regions [Fig. 3].

Data pre-processing includes data transformation and quantization based on expert background knowledge described in earlier paragraphs. In this way we get qualitative temperature gradient sequences representing the base for qualitative thermocline models learning.

The next modeling step is clustering of qualitative thermocline patterns [Fig. 4]. Since qualitative patterns incorporate expert knowledge, the results are clusters representing concepts – characteristic thermocline types descriptions. Depending on clusters quality and expert expectations, qualitative model parameters as the base for transformation and quantization, are tuned until clusters are of acceptable quality (Fig 4). Qualitative model tuning can be performed by changing qualitative model parameters and implementing different classifiers. Our intention is to use hierarchical clustering based on Levenshtein distance as similarity measure [1, 6, 10].

The resulting clusters representing varying thermocline types, are the base for thermocline qualitative models learning [Fig. 5]. In the case of insufficient number of patterns representing some rare thermocline instance, additional data from WOA 2005 can be processed in order to solve the sparse data problem.

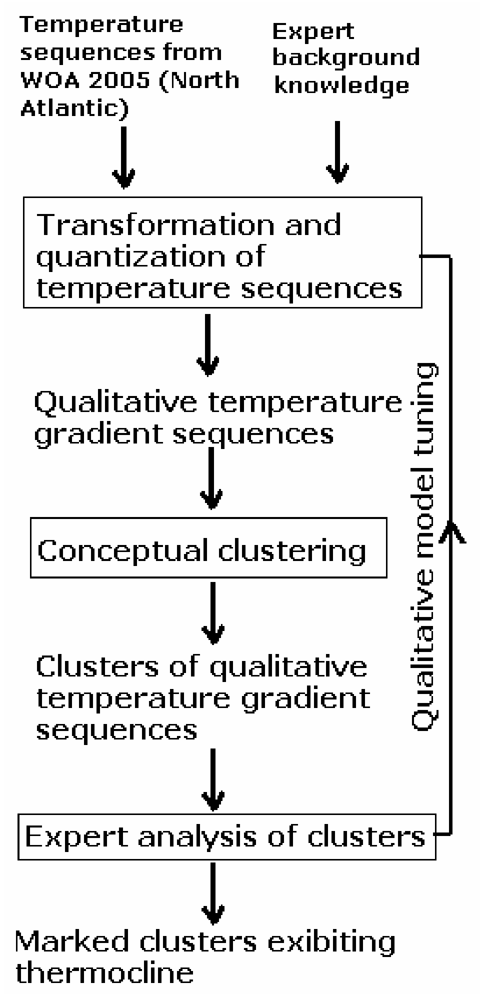
Temperature sequences from WOA 2005 (North Atlantic)
Expert background knowledge
Transformation and quantization of temperature sequences
Qualitative temperature gradient sequences
Conceptual clustering
Clusters of qualitative temperature gradient sequences
Expert analysis of clusters
Marked clusters exibiting thermocline
Qualitative model tuning

Fig. 4. Thermocline conceptual clustering.

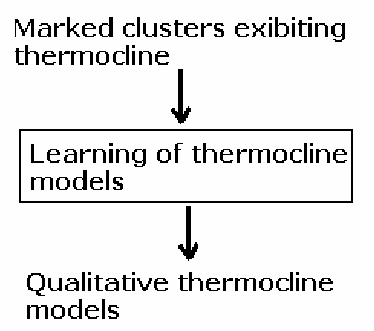
Marked clusters exibiting thermocline
Learning of thermocline models
Qualitative thermocline models

Fig. 5. Learning of varying thermocline types models.

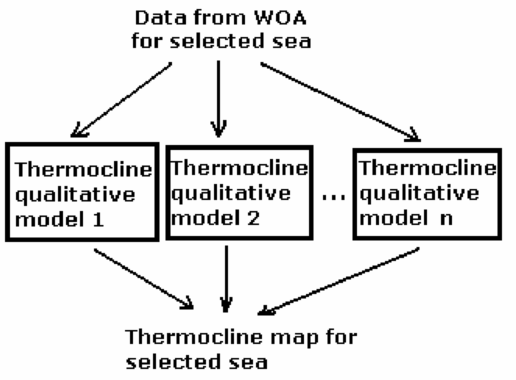
Data from WOA for selected sea
Thermocline qualitative model 1
Thermocline qualitative model 2
...
Thermocline qualitative model n
Thermocline map for selected sea

Fig. 6. Thermocline analysis.

Learned qualitative thermocline models can be further used in the data analysis of unseen data from WOA 2005 regarding other oceans and seas [6]. Possible implementation of qualitative models could be making of geographic maps.

VI. ALGORITHM DESCRIPTION AND EXPERIMENTAL RESULTS

Conceptual clustering of qualitative temperature gradient sequences is implemented as hierarchical clustering [1].

The input for the hierarchical clustering method is the set of 3842 sequences. The result of the hierarchical clustering is a tree representing all possible resulting clusters - a dendogram (Fig. 8).

Hierarchical clustering method has three steps [1,7]:

1. Step: Compute triangular matrix of distances between input sequences (patterns).

It includes computing of similarities between pairs of sequences using Levenshtein distance DL(x,y) [10]. Since these sequences of symbols are of different length, the dynamic programming method is used in order to determine the similarity.

2. Step: If the distance between clusters Si and Sj (in the first iteration every cluster contains only one pattern) is the smallest (the greatest similarity), we merge them in one cluster. We achieve cluster Si+Sj. Distances between a new cluster and the remaining clusters D(Si+Sj, Sk) are computed as:

D(Si+Sj, Sk) = 1/2 D(Si, Sk) + 1/2 D(Sj, Sk) + 1/2 |D(Si, Sk) - D(Sj, Sk)| (8)

Distance D(Si+Sj, Sk) is equal to the distance between the most distant patterns from the clusters Si+Sj and Sk.

3. Step: If new distance matrix has more than one column we repeat step 2; otherwise, we end the process. The result is a dendogram of patterns. An example of a dendogram is shown in Fig. 8.

A part of cluster number 3 is given for illustration in Fig.7.

We chose the dendogram cut which gives 50 clusters among which we can find comprehensive clusters. Only the clusters containing more than 50 sequences are significant. Table I shows cluster prototypes representing cluster center.

VII. ANALYSIS OF THE RESULTS

The result of hierarchical clustering is dendogram of clusters containing thermocline patterns. Dendogram cut is made at the tree level with 50 clusters. Only 14 of those clusters contain a significant number of patterns which has to be greater than the total number of clusters. Table 1 shows 14 cluster prototypes representing characteristic cluster thermocline patterns where n represents cluster number.

```
nnwwwnnnnnnnnnnnnnfnnnnnnnnn
nnwwwnnnnnnnnnnnnnfnnnnnnnnn
nnwwwnnnnnnnnnnnnnfnnnnnnnn
nnwwwnnnnnnnnnnnnnfnnnnnnnn
nnwwwnnnnnnnnnnnnffnnnnnnnn
nnwwwnnnnnnnnnnnnffnnnnnnnnn
nnwwwnnnnnnnnnnnnnfnnnnnnnn
nnwwnnnnnnnnnnnnnffnnnnnnnn
nnwwnnnnnnnnnnnnnffnnnnnnnnn
nnwwnnnnnnnnnnnnnffnnnnnnnnn
nnnwwnnnnnnnnnnnnnfnnnnnnnnn
```

Fig. 7. A part of sequences from the third cluster.

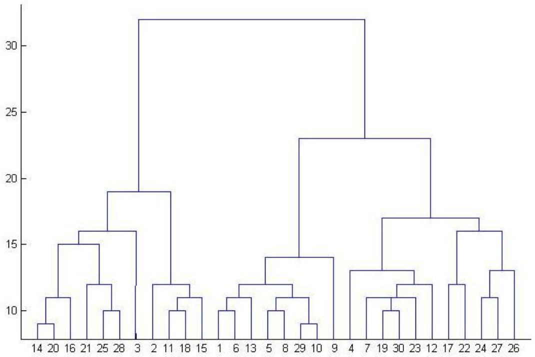

Fig. 8. An example: The dendogram of thermocline patterns for 30 clusters.

TABLE I
CLUSTER PROTOTYPES

n	Cluster prototype
3	nnwwwnnnnnnnnnnnnffnnnnnnnnnn
4	nnwwwnnnnnnnnnnnnnnnnnnnnnnnnnf
5	nnnnnnnnnnnnnnnnnnnnnfnfnnn
14	nnnnnnnnnnnfn
21	nnnsssnnnnnnnnnnnnfnnnnnnnnn
22	nsssnnnnnnnnnnnnnnnnnnnnnnnnnn
23	nnnnnwwnnnnnnnnnnffnnnnnnnn
25	nnnnnnnnnnnnnnnnnnnnnnnnnnnnf
27	nnnnnnfffnnnnnnnnnnnnn
39	nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
40	fffffffnnnfn
45	nnnnnnnnnnnnnnnnnnnnnnnnnn
49	nwwn
50	nnnnnnnnnnnnnnnnnnnnnnnfnffffff

The following groups were selected for further analysis, based on the significance of their thermocline conditions. Group 3, which exhibits a weak thermocline at the surface. Groups 5 and 14, which exhibit no thermocline but have a small temperature inversion (increase in temperature with depth) close to the bottom. Groups 21 and 22 best describe strong thermoclines at or close to the surface, and as such should be indicative of low latitude locations. Finally, group 39 is the most comprehensive group that exhibits no thermocline and will serve as a reference.

VIII. CONCLUSION

Expert analysis of clustering results shows that qualitative model needs to be tuned in order to satisfy some of the expert expectations. Significant clusters obtained are not enough precise representations of all interesting thermocline types. Future work will focus on that problem in effort to get more complete conceptual cluster descriptions.

The paper introduces our approach in an effort to model thermocline as a physical phenomenon. The resulting clusters obtained by the conceptual clustering are the base for the thermocline model building. In our future work we intend to tune initial parameters and implement different classifiers which determine the formation and the quality of clusters representing thermocline types. The next step is the implementation of machine learning methods [1,7] in order to learn different thermocline models for representative clusters. Learned thermocline models could then be used in thermocline analysis for other data from WOA 2005.

REFERENCES

[1] E. Alpaydin, *Introduction to Machine Learning*, MIT Press, 2004.

[2] T.P.Boyer, J.I. Antonov, H. Garcia, D.R. Johnson, R.A. Locarnini, A.V. .Mishonov, M.T. Pitcher, O.K. Baranova, I. Smolyar, World Ocean Database 2005, Chapter 1: Introduction, *NOAA Atlas NESDIS 60*, Ed. S. Levitus, U.S. Government Printing Office, Washington, D.C, 2006:182 pp.

ftp://ftp.nodc.noaa.gov/pub/WOD05/DOC/wod05_intro.pdf (24.1.2009)

[3] B. Kuipers, *Qualitative Reasoning: Modeling and Simulation with Incomplete Knowledge*, MIT Press, 1994.

[4] L´Istituto Nazionale di Oceanografia e di Geofisica Sperimentale – OGS, http://doga.ogs.trieste.it/cgibin/lipa_tra?Survey=A000524ADR (24.1.2009.)

[5] M. Matetic, S. Ribaric, I. Ipsic, "Qualitative Modelling and Analysis of Animal Behaviour", *Applied Intelligence Journal,* vol.21:p.25-44. Kluwer Academic Publishers, 2004.

[6] M. Matetic, S. Ribaric, "Conceptual Clustering of Object Behaviour in the Dynamic Vision System Based on Qualitative Spatio -Temporal Model", Proceedings of MIPRO 2002 conference, Section CIS, Opatija, 2002; p. 15-20.

[7] R.S. Michalski, I. Bratko, M. Kubat editors, *Machine Learning and Data Mining: Methods and Applications*. London, John Wiley Sons, 1998.

[8] P. Salles, B. Bredeweg, S. Araujo, "Qualitative models about stream ecosystem recovery: Exploratory studies", *Ecological Modelling*, Vol. 194, Issues 1-3, p. 80-89, 2006.

[9] M. Tomczak, "An Introduction to Physical Oceanography", http://www.es.flinders.edu.au/~mattom/IntroOc/ (24.1.2009.)

[10] R.A Wagner, M.J. Fisher, "The String to String Correction Problem", *Journal of the Association for Computing Machinery*, 21(1), 168-173, 1974.

[11] WORLD OCEAN ATLAS 2005, The National Oceanographic Data Center, http://www.nodc.noaa.gov/OC5/WOA05/pr_woa05.html (24.1.2009.)