KNN Regression Model-Based Refinement of Thermohaline Data

Yu Gou
Jilin University
Changchun City, China
gouyu18@mails.jlu.edu.cn

Jun Liu
Jilin University
Changchun City, China
liujun1509@jlu.edu.cn

Tong Zhang†
Jilin University
Changchun City, China
tongzhang18@mails.jlu.edu.cn

ABSTRACT
This paper carries out a refinement on the basis of existing data sets, whose level of granularity is not available for some experimental analysis such as thermocline research. The thermocline is sensitive to thermohaline data granularity for sudden sea temperature changes. We refined the data with the KNN regression method and managed to choose the optimal parameters for the construction of a prediction model. We also refined the temperature and salinity data in BOA_Argo using the regression forecast model. The original data, whose horizontal resolution is 1°×1° and vertically divided into uneven 58 layers from the sea surface to 1,975 meters underwater, has been refined into a new set with the resolution of 1°×1° horizontally and 1-meter interval vertically. At each point, we refined the previously uneven 58 temperature data samples into 1,976 evenly distributed data samples. The refined data sets can be used in experimental analysis, and the validity of this method has been verified by regional data.

KEYWORDS
KNN, BOA_ARGO, thermohaline data, granularity, thermocline

1 INTRODUCTION
With the increasing demand for ocean exploration and the continuous development of computer technology, we come to combine oceanography with computer science. An increasing number of computer science methods have been applied to ocean research and such combinations are of great significance. The reasons will be discussed in detail below.

For one thing, the oceans cover more than 71% of the earth, and the proportion is likely to rise to 72% as the greenhouse effect increases. China has nearly 300 square kilometers of sea area and has clearly realized that it is unrealistic to develop the land without considering the ocean. Such a vast area has great research value to each coastal country, and its strategic significance cannot be neglected [5]. The overall developments of land, sea, and air must go hand in hand to construct an integrated information space in the sky and the sea to promote the ability of wide-space and long-time communication. To study the ocean, we need to rely on the ocean data. We collect data from the ocean and transmit data to the data center. All data are transferred through communication channels. Therefore, underwater communication is necessary and extremely important.

But we still have to understand the difficulties that exist in underwater communications. First, the bandwidth of the underwater acoustic channel is limited, and there are various losses in the process of information transmission. Second, there is considerable noise interference in the sea, which is caused by natural weather phenomena such as wind, tsunamis, and earthquakes, as well as human shipping and drilling. In addition, the multipath effect and Doppler effect greatly influence underwater acoustic communication. Research on underwater communications can be traced back to the 1820s, when countries such as France and Switzerland measured the speed of sound transmission in the water in Geneva. Existing transmission methods are divided into wire and wireless transmission. Since the 1980s, with the aim of constructing a marine information system, international organizations, together with Europe, the United States, Japan, and other world ocean associations, have invested huge amounts of money to launch a series of large marine information system construction projects, including the Global Ocean Real-Time Observation network program (Argo) and the United States Integrated Ocean Observation System (IOOS)[2]. NEPTUNE submarine observation network program[1], European underwater observation program (ESONET, http://www.esonet-noe.org/),
will result in many small thermoclines that are not screened out so with the real value. When the data are collected and returned, the
Take the thermocline, for example. A thermocline (sometimes met-
alimnion in lakes) is a thin but distinct layer in a large body of
liquid (e.g., water, such as an ocean or lake) in which temperature
change continuously with depth under water. The experimental
operating environment in this paper is Intel(R) Core(TM) i7-6700
CPU @3.40ghz 3.41ghz, 8G RAM, Windows10 64-b, Python 3.5.KNN
regression by scikit-learn[10][5]. The experimental results were
visualized by PLOTLY and MATLAB.
In the second part, we will show how to use the machine learning
method to conduct the data refinement. The results are presented
and discussed in the third part. The fourth part contains the conclu-
sion of this paper and discusses some directions for future work.

2 METHOD

2.1 Notations

Here we introduce some of the notations in this article. First we
denote \( k \) as the number of neighbors assigned to a query point.
\textbf{Train} and \textbf{Test} denote training sets, and testing sets respectively. \( \mathcal{X} \)
refers to the original data sets from BOA_Argo, where \( x_i \in \mathcal{X} \subseteq \mathbb{R} \).

2.2 Experimental Design

This paper adopts data from BOA_Argo, which are provided by the
China Argo Real-Time Data Center (http://www.argo.org.cn). The
original data’s horizontal resolution is 1° x 1°, and it is vertically
divided into uneven 58 layers from the sea surface to 1,975 meters
underwater. As the depth goes deeper, the gap between the two
layers increases. This makes thermocline research difficult. As men-
tioned previously, there are many 1- to 2-meters-thin thermo-layers
in the ocean that need to be incorporated as we process them. If the
gap between layers is too great, many thin thermal layers will be
ignored, which will have a great influence on the final determina-
tion. According to the Ocean Survey Specification[5], thermal layers
that have thickness less than 5 meters need to be filtered out. If the
interval between two layers is more than 5 meters, they cannot be
merged and need to be quarantined to two different thermoclines.
Therefore, on the basis of the original data set, we preprocessed and
processed the data and interpolated the data vertically using the
k-nearest neighbors (KNN) method[13][14], making it a data set
with 1 m intervals and 501 even layers in the vertical direction[7].
According to a large number of visualization experiments, when
the depth is deeper than 500 meters, the seawater enters a static
layer with almost constant temperature, which rarely produces (or
even no longer produces) a thermocline. Because the data obtained
are not comprehensive enough (Because we do not yet fully under-
stand the ocean floor, or are unsure if there is a submarine volcano
and turbulence), we only research the area from sea surface to 500
meters below. The study of the characteristics of deeper regions
will be carried out after better oceanic data are available in the
future.
In seawater, the temperature and salinity values change continu-
ously among different points, so the prediction of temperature
and salinity is a regression problem. Common regression methods
include linear regression, SVM (Support Vector Machine), decision
tree regression, KNN (k-nearest neighbor), integration methods,
and so on. KNN, as an example based on ‘lazy learning’ method,
KNN Regression Model-Based Refinement of Thermohaline Data

and it predicts information according to the given test sets and the average value of k nearest "neighbors". The changes in temperature, salinity, and depth in the ocean are all continuous, and this method of predicting based on samples applies to marine scenarios. Algorithm 1 represents the procedure of how the refinement is done.

**Algorithm 1** Procedure of KNN Regression Model-Based Refinement

**Input:**
- X is the data sets;
- x is the data from the data set;
- x\text{norm} is the normalized value;
- x_{\text{min}} is the minimum value of the data set;
- x_{\text{max}} is the maximum value of the data set;

**Output:**
- Evenly distributed data sets;
- 1: Load the data sets;
- 2: Preprocess the data;
- 3: Drop missing data;
- 4: Data normalization;
- 5: \( x_{\text{norm}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \);
- 6: Training sets \( \text{Train} \) and testing sets \( \text{Test} \) are split in different proportion.
- 7: Training KNN model
- 8: Train the model with different parameter \( k \)
- 9: Select the optimal model with the least MSE.
- 10: Predict the target for the provided data.
- 11: Scale back the data to the original representation.
- 12: Thermohaline data of each sample point are refined into vertically evenly distributed data sets by using the optimal model.

The distance between a query point and its neighbor is usually calculated by Manhattan and Euclidean distance. In this paper, we use Euclidean distance to assign nearest neighbors to a query point. Manhattan distance can be represented as,

\[
L_1(x_i, x_j) = \sum_{i=1}^{n} |x_i^{(l)} - x_j^{(l)}| \quad (1)
\]

Euclidean distance can be represented as,

\[
L_2(x_i, x_j) = \left( \sum_{i=1}^{n} (x_i^{(l)} - x_j^{(l)})^2 \right)^{1/2} \quad (2)
\]

Euclidean distance is the most common measuring method, and it measures the absolute distance between each point in the multi-dimensional space. Because the calculation is based on the absolute value of each dimensional feature, Euclidean measurement needs to make sure that each dimensional index is at the same scale level, and the use of Euclidean distance for two different units may invalidate the result.

The example above in Figure 1 shows how the KNN method can be used to perform the regression prediction of data based on cases. There are 12 instances in the picture, 11 of which are known for classification and one of which needs to be classified. Of the 11 instances of the known classification, there are 5 “purple triangle” category and 6 “pink square” category. The whole picture is roughly divided into three areas: (1) solid line circle area; (2) dotted circle area; (3) solid line box area. When the sample falls in the region of “solid line circle,” the one which is to be classified is determined as "purple triangle" according to the sample features within certain distances. When the sample falls within the "dotted line circle" area, the prediction point is judged as "pink square" according to the five sample features closest to it. When the sample falls within the ‘solid line circle’ area, the predicted point was also determined as a “pink square.” It can be seen that in the KNN problem, the most important is the choice of parameter \( k \), which is the number of samples to be considered for the classification. The following section describes how to select the KNN parameters.

**2.3 Parameter Selection and Experimental Results Analysis**

When using the KNN method for the regression prediction problem, the similarity between the predicted value by the KNN method and the real data is the premise of all follow-up studies. In this paper, mean squared error (MSE) is adopted for performance evaluation. MSE is the mean of the squared error between the predicted data and the original data.

\[
\text{SSE (Sum of squares due to error)} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2
\]

\[
\text{MSE} = \frac{\text{SSE}}{n} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (3)
\]

The closer the MSE value is to 0, the better the data fitted by the model and the more successful the data prediction. In the contrast, the closer the MSE value is to 1 indicates that the model is not suitable for the data prediction of the current scenario.

In the calculation of thermocline, according to the Ocean Survey Specification, we mainly use temperature and salinity to analyze the thermocline. Because the variation patterns of temperature and salinity are different, we will separately choose the regression model parameters for the two hydrological features.

**2.4 Data Resources**

In this paper, the data mainly come from the BOA_Argo and WOA13. Argo is a global array of 3,800 free-drifting profiling floats that measure the temperature and salinity of the upper 2,000 m of the ocean. This allows, for the first time, continuous monitoring of the temperature, salinity, and velocity of the upper ocean, with all
The changes in temperature directly affect the formation of the thermocline, and temperature is one of the most important factors for the study of the thermocline. In seawater, temperature changes with depth, which is continuous, so we chose KNN as the model for the data regression prediction. In this method, the temperature assigned to the query point is calculated based on the average temperature of the nearest \( k \) "neighbor" points. In this paper, we allocate the points with weights, known as "Ker" in the following table. To make the experimental process easier to understand, here is a brief explanation of some of the variables defined in the experiment. The YearMonth represents the year and month of the experimental data being collected. Target is the feature to be predicted by the KNN regression model. In this experiment, the value set, Target is \( \text{temp, sal} \), which represent temperature and salinity, respectively. Model is the selected prediction model, and the model in this experiment is the KNN model. Ker determines how the model allocates weight to nearby neighbor nodes, and the set of values is \( \text{uniform, distance} \). The default value is \( \text{Ker} = \text{uniform} \). At this point, the program assigns equal weight to all points. When \( \text{Ker} = \text{distance} \), the program assigns weight to nearby points in proportion to the reciprocal of the query point distance. In other words, the closer to the query point, the higher its weight. This weight allocation rule is useful for making temperature data predictions. Therefore, this weight allocation method is used to process temperature data in this paper. The refinement of salinity data in the next section also adopts this method. \( K \) sets the number of nodes used as a reference to predict the value of the target node, and the parameter \( k \) is also a decisive factor for the performance of the KNN regression mode. The MSE has been introduced above and will not be repeated here. Next, we will use KNN regression models with different \( k \) values to carry out the refinement of the 2016 temperature data in the experimental area. The following figure shows the MSE comparison from different KNN regression models with different \( k \) values. The smallest MSE represents the best regression performance. As can be seen from the above figure, the best experimental performance occurs when \( k = 3 \). Next, we must determine whether the KNN model, which applies to annual (long-term) predictions, applies to January (short-term) temperature forecasts as well. To verify the applicability of the experimental model to short-term data, we then took data from June 2016 as an example to compare the MSE between the predicted value and real value at different \( k \) values. Table 1 shows the results.

### EXPERIMENTS

#### 3.1 Temperature

The changes in temperature directly affect the formation of the thermocline, and temperature is one of the most important factors for the study of the thermocline. In seawater, temperature changes with depth, which is continuous, so we chose KNN as the model for the data regression prediction. In this method, the temperature assigned to the query point is calculated based on the average temperature of the nearest \( k \) "neighbor" points. In this paper, we allocate the points with weights, known as "Ker" in the following table. To make the experimental process easier to understand, here is a brief explanation of some of the variables defined in the experiment. The YearMonth represents the year and month of the experimental data being collected. Target is the feature to be predicted by the KNN regression model. In this experiment, the value set, Target is \( \text{temp, sal} \), which represent temperature and salinity, respectively. Model is the selected prediction model, and the model in this experiment is the KNN model. Ker determines how the model allocates weight to nearby neighbor nodes, and the set of values is \( \text{uniform, distance} \). The default value is \( \text{Ker} = \text{uniform} \). At this point, the program assigns equal weight to all points. When \( \text{Ker} = \text{distance} \), the program assigns weight to nearby points in proportion to the reciprocal of the query point distance. In other words, the closer to the query point, the higher its weight. This weight allocation rule is useful for making temperature data predictions. Therefore, this weight allocation method is used to process temperature data in this paper. The refinement of salinity data in the next section also adopts this method. \( K \) sets the number of nodes used as a reference to predict the value of the target node, and the parameter \( k \) is also a decisive factor for the performance of the KNN regression mode. The MSE has been introduced above and will not be repeated here. Next, we will use KNN regression models with different \( k \) values to carry out the refinement of the 2016 temperature data in the experimental area. The following figure shows the MSE comparison from different KNN regression models with different \( k \) values. The smallest MSE represents the best regression performance. As can be seen from the above figure, the best experimental performance occurs when \( k = 3 \). Next, we must determine whether the KNN model, which applies to annual (long-term) predictions, applies to January (short-term) temperature forecasts as well. To verify the applicability of the experimental model to short-term data, we then took data from June 2016 as an example to compare the MSE between the predicted value and real value at different \( k \) values. Table 1 shows the results.
The experimental results show that even within a month, the optimal model still appears when $k = 3$. By applying the KNN regression model at $k = 3$ to the data of other months, the following results can be obtained.

It can be seen from the above figure and the above table that the MSE of each month of the whole year of 2016 is between 0.001 and 0.003, and the predicted value and the real value can be well fitted, indicating that when $k = 3$, the model is applicable to the prediction of the temperature value within the month range (short-term). Therefore, we can conclude that the KNN regression prediction model of $k = 3$ is the most suitable model for temperature prediction.
3.2 Salinity

Salinity is another key factor in thermocline analysis. Changes in salinity can lead to changes in the electrical conductivity of seawater, disturb the balance of the original ecosystem, and also cause abnormal climate changes. In Peru, for example, in an abnormal year for seawater salinity, the week trade (including the southeast trade winds, and the northeast trade wind) forces lead to weaker forces of the south and north equatorial current, and equatorial counter-current strengthen at the same time. All above together block the Humboldt current, causing Peruvian coastal salinity and temperature rises, resulting in the precipitation acting abnormally eventually leading to the west coast of the Pacific being in drought. These have a lot to do with El Nino and La Nina.

As mentioned in the previous section, salinity value is also a continuous value varying with depth, so we also adopt the KNN regression model with the weight allocation method of "distance". Next, we adopt the KNN regression model with different $k$ values to carry out the refinement of the salinity data of the experimental area in 2016. The following figure shows the MSE comparison of the KNN regression model with different $k$ values. Similarly, we believe that the smaller the MSE, the better the regression performance. As with temperature, we also need to know whether the KNN model, which is applicable to the annual salinity (long-term) prediction, is applicable to the salinity value prediction of January (short-term). We then took data from June 2016 as an example to compare the MSE between the predicted value and real value at different $k$ values. Table 3 shows the results. The experimental results show that even within a month, the optimal model still appears when $k = 3$. By applying the KNN regression model at $k = 3$ to the data from the other months, the following results can be obtained. It can be seen from the above tables and figures that the MSE of each month in 2016 is between 0.000001

### Table 3: MSE at different $k$ values between the predicted temperature and the original salinity in June 2016

| Year | Month | Target | Model | Ker | k | MSE     |
|------|-------|--------|-------|-----|---|---------|
| 2016 | 06    | sal    | KNN   | distance | 1 | 0.000025 |
| 2016 | 06    | sal    | KNN   | distance | 3 | 0.000014 |
| 2016 | 06    | sal    | KNN   | distance | 5 | 0.000021 |
| 2016 | 06    | sal    | KNN   | distance | 7 | 0.000033 |
| 2016 | 06    | tem    | KNN   | distance | 9 | 0.000046 |

3.2 Salinity

Salinity is another key factor in thermocline analysis. Changes in salinity can lead to changes in the electrical conductivity of seawater, disturb the balance of the original ecosystem, and also cause abnormal climate changes. In Peru, for example, in an abnormal year for seawater salinity, the week trade (including the southeast trade winds, and the northeast trade wind) forces lead to weaker forces of the south and north equatorial current, and equatorial counter-current strengthen at the same time. All above together block the Humboldt current, causing Peruvian coastal salinity and temperature rises, resulting in the precipitation acting abnormally eventually leading to the west coast of the Pacific being in drought. These have a lot to do with El Nino and La Nina.

As mentioned in the previous section, salinity value is also a continuous value varying with depth, so we also adopt the KNN regression model with the weight allocation method of "distance". Next, we adopt the KNN regression model with different $k$ values to carry out the refinement of the salinity data of the experimental area in 2016. The following figure shows the MSE comparison of the KNN regression model with different $k$ values. Similarly, we believe that the smaller the MSE, the better the regression performance. As with temperature, we also need to know whether the KNN model, which is applicable to the annual salinity (long-term) prediction, is applicable to the salinity value prediction of January (short-term). We then took data from June 2016 as an example to compare the MSE between the predicted value and real value at different $k$ values. Table 3 shows the results. The experimental results show that even within a month, the optimal model still appears when $k = 3$. By applying the KNN regression model at $k = 3$ to the data from the other months, the following results can be obtained. It can be seen from the above tables and figures that the MSE of each month in 2016 is between 0.000001

### Table 3: MSE at different $k$ values between the predicted temperature and the original salinity in June 2016

| Year | Month | Target | Model | Ker | k | MSE     |
|------|-------|--------|-------|-----|---|---------|
| 2016 | 06    | sal    | KNN   | distance | 1 | 0.000025 |
| 2016 | 06    | sal    | KNN   | distance | 3 | 0.000014 |
| 2016 | 06    | sal    | KNN   | distance | 5 | 0.000021 |
| 2016 | 06    | sal    | KNN   | distance | 7 | 0.000033 |
| 2016 | 06    | tem    | KNN   | distance | 9 | 0.000046 |
Table 4: MSE between the predicted salinity and the original salinity in 2016 when k=3

| Year | Month | Target | Model  | Ker   | k   | MSE      |
|------|-------|--------|--------|-------|-----|----------|
| 01   | sal   | KNN    | distance | 3     | 0   | 0.000032 |
| 02   | sal   | KNN    | distance | 3     | 0   | 0.000026 |
| 03   | sal   | KNN    | distance | 3     | 0   | 0.000023 |
| 04   | sal   | KNN    | distance | 3     | 0   | 0.000025 |
| 05   | sal   | KNN    | distance | 3     | 0   | 0.002012 |
| 06   | sal   | KNN    | distance | 3     | 0   | 0.000014 |
| 07   | sal   | KNN    | distance | 3     | 0   | 0.000015 |
| 08   | sal   | KNN    | distance | 3     | 0   | 0.000031 |
| 09   | sal   | KNN    | distance | 3     | 0   | 0.000031 |
| 10   | sal   | KNN    | distance | 3     | 0   | 0.000033 |
| 11   | sal   | KNN    | distance | 3     | 0   | 0.000054 |
| 12   | sal   | KNN    | distance | 3     | 0   | 0.000055 |

Figure 8: Original data distribution and the refined data distribution of temperature, and 0.000055, and the predicted value and the real value can be well fitted, indicating that when k=3, the model is applicable to the prediction of salinity value of the month range (short-term). Therefore, we can conclude that the KNN regression prediction model of k=3 is the most suitable one for salinity prediction. Conclusions can be drawn from the above two experiments that despite whether it is temperature or salinity or short-term or long-term, the predicted effect is best when k = 3. Therefore, all the data in the following experiments in this paper are predicted by the KNN method with the parameters of k=3.

3.3 Refinement Results

In the previous section, we selected the model for the data refinement process. And when we applied the model to the original data, the refined data were obtained. Figure 8 and 9 show the original data distribution and the refined data distribution of temperature and salinity, respectively.

After the process, we refined the effective data amount of each sample point from 3,800 and 4,060 to 35,070. From the above two groups of graphs, it can be seen that the original data are coarser in granularity, and the refined data are smoother and more fine-grained. However, it is not difficult to see that the data after refinement are still not smooth enough. The refinement of thermohaline data serves as the basis of subsequent thermocline research. Thermoclines change with seasons.
and are generally classified as the following four periods: no thermocline period, growth period, strong period and fade period. We chose the temperature profiles of February, May, August and November to represent these four temperature distribution periods, which reflect the variation tendency of temperature throughout the year. The following conclusions can be obtained by observing the temperature profiles of the original data and the temperature profiles of refined data. First of all, the thermocline occurs all year round in the experimental area, and most of the thermocline occurs at a depth of 50 meters. The highest sea surface temperatures occur in summer, and, in the rest of the year, the sea surface temperature falls with the decrease of thermal radiation. We can also clearly see that the original data sample points are very limited. At each point, there are only 35 unevenly distributed temperature samples from the sea surface to 500 meters below the surface vertically, and the intervals between each sample are large. As mentioned above, the appearance of the thermocline, namely the salutation of temperature, is sudden, and the excessive spacing may cause some thin thermoclines to be ignored. After refinement, the previously uneven 35 temperature data samples are refined into 501 evenly distributed data samples. However, it is not difficult to find that the degree of the processed data is still not fine enough. In future research, we should refine the data to a granularity of 0.5 m or even 0.1 m, and release the processed data sets to the website as a data product.

4 CONCLUSION

This paper discusses the data prediction problem based on the KNN regression method and selects the most suitable parameters for the studied ocean scenario to construct the refined model. The temperature and salinity data in BOA_Argo are refined using the regression prediction model. The original data, whose horizontal resolution is 1°x1° and vertically divided into uneven 58 layers from sea surface to 1,975 meters underwater, are refined into new sets with a resolution of 0.1°x0.1°horizontally and 1-meter interval vertically. The refined data sets can be used in experimental analyses and the validity of this method has been verified by regional data.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (Grants No.61631008), the Fundamental Research Funds for the Central Universities 2017TD-18, and the State Key Laboratory of Robotics No. 2015-O06. These data were collected and made freely available by the International Argo Program and the national programs that contribute to it (http://www.argo.ucsd.edu, http://argo.jcommops.org). The Argo Program is part of the Global Ocean Observing System.

REFERENCES

[1] Christopher R. Barnes, Mairi M. R. Best, Fern R. Johnson, Lucie Pautet, and Benoît Pireyre. 2013. Challenges, Benefits, and Opportunities in Installing and Operating Cabled Ocean Observatories: Perspectives From NEPTUNE Canada. *Journal of Oceanic Engineering*, 38, 1 (2013), 144–157.
[2] Jeff De La Beaudariere. 2009. The NOAA IOOS Data Integration Framework: Initial implementation report. In *Oceans*. 1-8.
[3] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Nticu, Peter Prettenhofer, Alexandre Gramfort, and Jacques Grobler. 2013. API design for machine learning software: experiences from the scikit-learn project. *Eprint Arxiv* (2013).
[4] Thierry Carval, Robert Keeley, Yasushi Takatsuki, Takashi Yoshida, Claudia Schmid, Roger Goldsmith, Anne Wong, Arun Thresher, Anh Tran, and Stephen Loch. 2015. Argo user’s manual V3.2. (10.99 ms) (2015). https://doi.org/10.17782/42182
[5] Chengquan Hu, Yu Gou, Tong Zhang, Kai Wang, Lili He, and Yu Jiang. 2017. Analysis of Thermocline Influencing Factors Based on Decision Tree Methods. (2017).
[6] Chengquan Hu, Yu Gou, Tong Zhang, Kai Wang, Lili He, and Yu Jiang. 2017. Thermocline Analysis Based on Entropy Value Methods. (2017).
[7] Chengquan Hu, Tong Zhang, Jin Wang, Yu Gou, Kai Wang, Hongtao Bai, and Yu Jiang. 2017. Construction of High Resolution Thermocline Grid Data Sets. (2017).
[8] Li Jiahao, Li Peiyi, Chen Shang, Zhang Yijun, Yu Xiaoguo, Ji Weidong, Zhang Yusheng, Yao Yong, Wang Weiyang, and Wu Ziyin. 2011. Ocean survey specification. China Standardization 2 (2011).
[9] Y. Jiang. 2017. A Machine Learning Approach to Argo Data Analysis in a Thermocline. Sensors 17, 10 (2017), 2225.
[10] Fabian Pedregosa, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Oliver Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and Jake Vanderplas. 2012. Scikit-Learn. Machine Learning in Python. *Journal of Machine Learning Research*, 12, 10 (2012), 2825–2830.
[11] Getting Started. 2011. Ocean Data View. (2011).
[12] Y. U. Yongqiang, Weipeng Zheng, Bin Wang, Hualong Liu, and Jiping Liu. 2011. Versions g1.0 and g1.1 of the LASGIAP Flexible Global Ocean-Atmosphere-Land System Model. *Advances in Atmospheric Sciences*. 28, 1 (2011), 99–117.
[13] Shichao Zhang, Xuelong Li, Ming Zong, Xiaofeng Zhu, and Debo Cheng. 2017. Learning k for KNN Classification. *ACM Transactions on Intelligent Systems & Technology*. 8, 3 (2017), 43.
[14] ZhiHua Zhou. 2016. *Machine learning*. Tsinghua University Press.