Terr. Atmos. Ocean. Sci., Vol. 25, No. 5, 727-741, October 2014 doi: 10.3319/TAO.2014.03.17.01(Oc)

NOTES AND CORRESPONDENCE

Querying Similar Water Masses Visualization Tool Design and Implementation Based on Polynomial Regression

Jian-Heng Wu[1], Bor-Shen Lin[1, *], and Jia-Yu Kuo[2]

[1] *Department of Information Management, National Taiwan University of Science and Technology, Taipei, Taiwan*
[2] *Institute of Oceanography, National Taiwan University, Taipei, Taiwan*

Received 4 December 2013, revised 6 March 2014, accepted 17 March 2014

ABSTRACT

The temperature-salinity relationship is one of the most important characteristics used for identifying water masses in marine research. Nonetheless, it is not easy to search, compare or analyse the temperature-salinity characteristic efficiently in the ocean database of a wide ranging area. Since marine data are typically collected over a wide range area, how to represent, manage and share such data flexibly and responsively is a critical issue in marine research. Visualization techniques are powerful media for data presentation and knowledge discovery. The temperature-salinity relationship that signifies the characteristics of water mass is modelled in this study as a polynomial function whose coefficients can be estimated through statistical regression. Based on such representation, the distance between two temperature-salinity characteristics could be measured automatically, allowing the comparison of similar water masses for a wide range area to be efficiently performed. The proposed approach can effectively reduce the amount of computations by aggregating the data with seasonal and spatial variations, facilitating the comparison of different water masses through sampling the temperature-salinity characteristics without degrading their discriminating capabilities. With reduced scale data it becomes feasible to visualize or compare them in real time. This tool is helpful for querying geographic locations with similar temperature-salinity characteristic interactively and for tracking specific patterns of water masses, such as the Kuroshio near Taiwan or those in the South China Sea.

Key words: Ocean database, Temperature-salinity, Polynomial regression, Information visualization, Water mass, Taiwan

Citation: Wu, J. H., B. S. Lin, and J. Y. Kuo, 2014: Querying similar water masses visualization tool design and implementation based on polynomial regression. Terr. Atmos. Ocean. Sci., 25, 727-741, doi: 10.3319/TAO.2014.03.17.01(Oc)

1. INTRODUCTION

The ocean database is a common basis for marine relevant researches, including physical oceanography, chemical oceanography and so on. Ocean database management has long been difficult for several reasons. First, the collection and processing of marine data are expensive and laborious. Conventionally, marine data are observed and collected using instruments on cruises that navigate on the sea in a wide range area for a long period of time. The collected data are typically raw text files, which require further processing to ensure that they are accurate and consistent for merging into the database. Second, the huge amount of data makes it difficult to efficiently perform large-scale analysis, such as computing the statistics or comparing the characteristics of many locations. Even for the most advanced systems, finding the right pieces of information in a timely fashion in a large database still remains a difficult issue. Without appropriate data scale reduction it is usually infeasible to visualize or compare them in real time. Third, the ocean database has potential users from diverse groups from different research communities, including military, economic, social, ecological or academic users, with individual information needs. The database management system therefore needs to provide the users a flexible interface that can meet diverse requirements for potential users, from low-level raw data to high-level statistical or analytical data. As a consequence, designing and building a flexible and reusable management

* *Corresponding author*
E-mail: bslin@cs.ntust.edu.tw

system for the ocean database is a challenging issue.

Temperature-salinity (T-S) relationship that can express the equation state of sea water has long been an important characteristic in oceanology. It is often used to determine the nature of the transformation and interaction of different waters (Mamayev 1975), or to identify water masses. Conventional analysis of water masses requires manual comparison for a few figures of T-S relationship generated from the observed data. Provided that the figures for a wide range area could be automatically compared and visualized efficiently, it would be beneficial for large scale or cross-domain researches of water masses. Therefore, efficient search, comparison and analysis of the T-S characteristics for a wide range area are touchstones for an advanced ocean database system.

This paper presents a flexible ocean database management system that allows the user to query similar water masses for a wide range area. Regression analysis of T-S characteristics in marine data is studied first with a similarity measure for two T-S characteristics proposed to automatically compare different water masses. This makes it possible to efficiently identify similar water masses for a reference location in a wide range area. The proposed database management system was built based on service oriented architecture (SOA) that can achieve flexibility, scalability and interoperability (Garlan 2000). Data aggregation in the data layer is devised to effectively reduce the computational cost of statistical or analytical data such as the T-S characteristic, while web services of different types, such as primary web services and analytical web services, in the service layer can be exported for flexible integration. The web client further consumes the services, visualizes the data geographically and responsively interacts with the users. The analysis approach and the reference architecture were successfully applied to a web tool that presents the T-S characteristic interactively based on conductivity-temperature-depth (CTD) data near Taiwan. The proposed approach was finally validated on the Kuroshio distribution and it was found that similar water masses from a reference location within the Kuroshio region obtained using this approach coincide highly with the Kuroshio paths obtained from other ocean researches.

2. BACKGROUND TECHNOLOGIES AND METHODS

2.1 Temperature-Salinity Curves of Water Mass

According to Mamayev (1975), the T-S relationship analysis, together with the field expressing the state equation of the sea water, allow us to take into account the most important factors that determine the nature of the transformation and interaction of different waters. Water masses usually refer to persistent T-S characteristics. One benefit of the T-S relationship is that it allows us to map the geographic distribution of different water property groups (Kim et al. 1991). T-S relationship is therefore utilized in this study because it is an effective characteristic for identifying a water mass and has been widely used in oceanography. One example of the T-S relationship for CTD cast A6D1 of cruise ORIII 1470 is presented in Fig. 1. The red point in Fig. 1a indicates a location on the path of the Kuroshio region (Liang et al. 2008). The Fig. 1b is the T-S relationship corresponding to the red point on the left hand side. The known dataset of water masses for the Kuroshio region around Taiwan will be used as the target for T-S characteristic analysis in this study because the Kuroshio has characteristics so different from the other water masses prominently in this area that it can be identified visually.

2.2 Oceanography Data near Taiwan

Taiwan, located between the tropics and the subtropics,

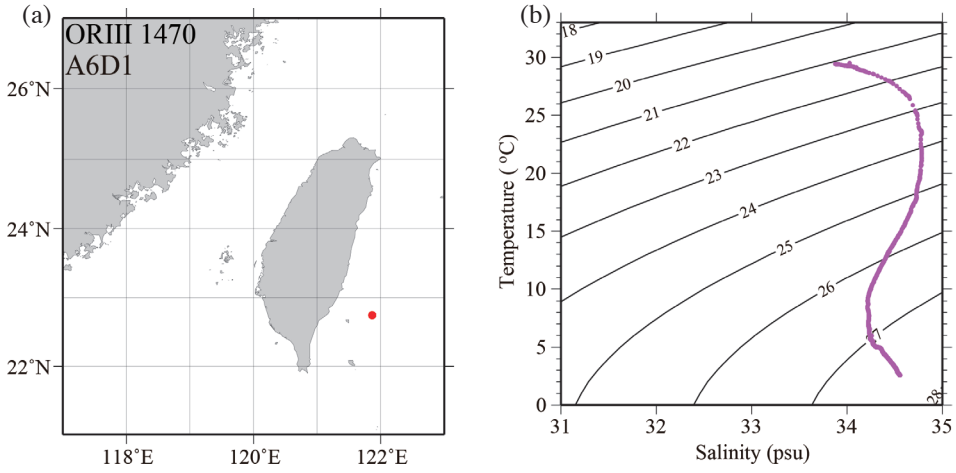

Fig. 1. T-S relationship for CTD cast A6D1 of cruise ORIII 1470, provided by the Ocean Data Bank of Taiwan's Ministry of Science and Technology operated by the Institute of Oceanography, National Taiwan University. The red point in (a) indicates a location on the path of the Kuroshio region (Liang et al. 2008); (b) is the T-S relationship corresponding to the red point on the left hand side.

lies on the border between the largest land mass and the largest ocean in the world, where the marine and atmospheric environments are complex and sensitive (Chien et al. 2010). A large scale area oceanographic database called the Ocean Data Bank (ODB) was adopted to analyze the water masses around Taiwan. The data bank's website is www.odb.ntu.edu.tw. This database belongs to the Ocean Data Bank of the Ministry of Science and Technology (ODB/MOST), initiated and operated by the Institute of Oceanography, National Taiwan University since 1986. The main data sources in the ODB are provided by R/V Ocean Researchers I, II, and III through long-term surveys around Taiwan over the past three decades. More than 20 million CTD data has been accumulated over the past 25 years from over forty thousand stations. ODB/MOST provides CTD data for relevant investigations in the East Asian Seas region. The ocean conductivity, temperature and depth distribution for the region around 10° - 30°N and 110° - 130°E are presented in this study. The temperature-salinity data were collected from 1986 - 2010 by Ocean Researchers I, II, and III using the CTD instrument. We use the annual CTD maximum depth data from 23709 Casts (between 0 - 5513 m) collected during the last 25 years, from 1986 till 2010. There are more than 20 million records collected and processed by ODB/MOST with strict quality control. The cruise track information is shown in Fig. 2, in which the pink line denotes the cruise tracks of Ocean Researcher I, the green line denotes the cruise tracks of Ocean Researcher II, and the blue line denotes the cruise tracks of Ocean Researcher III.

2.3 Polynomial Regression

Polynomial regression (Stigler 1971) is a form of regression in which the relationship between the independent variable x and the dependent variable y is modelled as a polynomial of degree n. The relationship can be used to describe nonlinear relationships, such as the growth rate of tissues (Shaw et al. 2006), or the distribution of carbon isotopes in lake sediments (Barker et al. 2001). We use a polynomial regression in this paper to obtain the T-S relationship of water mass in oceanography.

The general polynomial regression model form is depicted below.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \varepsilon_i \quad i = 1, \cdots, n \tag{1}$$

where the x_i's are the input variable samples and the y_i's are the output variable samples, β_p's are the regression coefficients, and ε_i's are the error terms. The above equations form a set of linear equations for parameter β_p's, which can be further represented in the matrix-vector form as follows (Cetisli and Kalkan 2011).

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^p \\ 1 & x_2 & x_2^2 & & x_2^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^p \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{2}$$

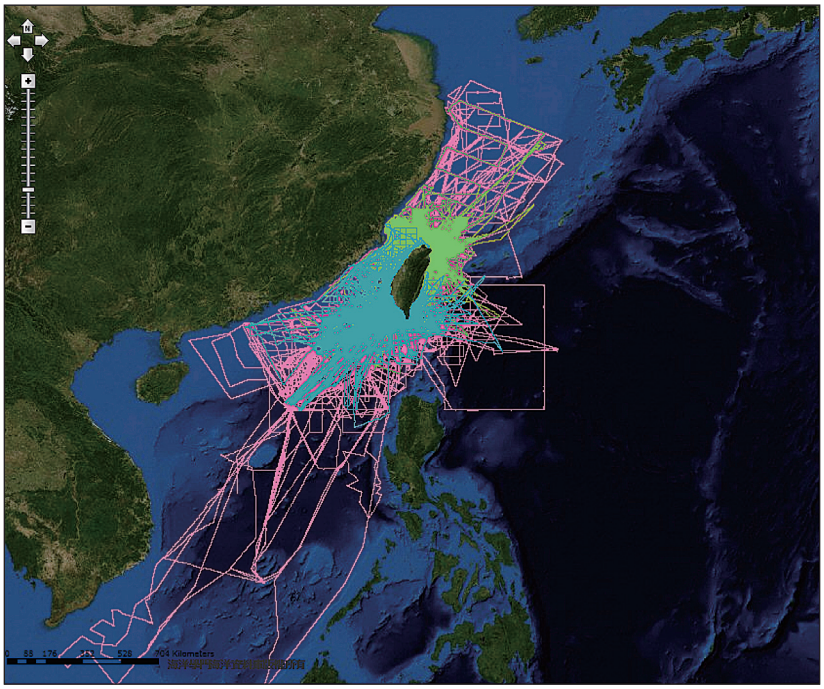

Fig. 2. Cruise tracks of R/V Ocean Researcher I, II, and III, provided by the Ocean Data Bank of Taiwan's Ministry of Science and Technology operated by the Institute of Oceanography, National Taiwan University.

Or alternatively,

$$y = X\beta + \varepsilon \tag{3}$$

When the training samples $\{(x_i, y_i)\}$ are given the optimal β that minimizes the square errors $|\varepsilon|^2$ can then be solved as below.

$$\beta^* = (X^T X)^{-1} X^T y \tag{4}$$

In statistics the mean squared error (MSE) signifies the expected difference between the value of an estimator (y_i') and the true value (y_i) of the quantity being estimated, and can be used to indicate prediction accuracy. The MSE is zero if the y_i and y_i' values are identical (Mielke et al. 1996). With the coefficients β^* for the polynomial the MSE can be computed as below,

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2, \quad y' = X\beta^* \tag{5}$$

The regression function in Eq. (1) can be used to represent the T-S relationship for a water mass where the temperature is the input variable x and the salinity is the output variable y. Given a set of training samples consisting of pairs of temperature and salinity values, i.e., $\{(x_i, y_i)\}$, the regression coefficients β can be computed and the derived polynomial can then be used to predict the salinity for a given temperature.

The choice of regression polynomial degree depends on the computational cost and available space. According to earlier research by Teague et al. (1990), the polynomial degree can be chosen as five for the top and middle salinity models and seven for the middle temperature model. We make use of the JSXGraph library, which is an open-source client-side web library for displaying interactive mathematics and drawings in a web browser, to generate the regression curve for an assigned degree. Examples of regression analysis for a set of data from the Kuroshio dataset with polynomials of degree 3 - 5 are shown in Figs. 3a - c, respectively, in which the points are the training samples and the solid green curve is the derived polynomial. The MSE with respect to the polynomial degree is further shown in Fig. 3d. As can be seen in Fig. 3d the MSE decreases as the degree increases and the estimation error is low when the degree is larger than 4. To simplify the analysis the degree of the polynomial in this paper is set at 5.

2.4 Distance Measure

In order to automatically compare the T-S characteristics of water masses the distance measured between two T-S

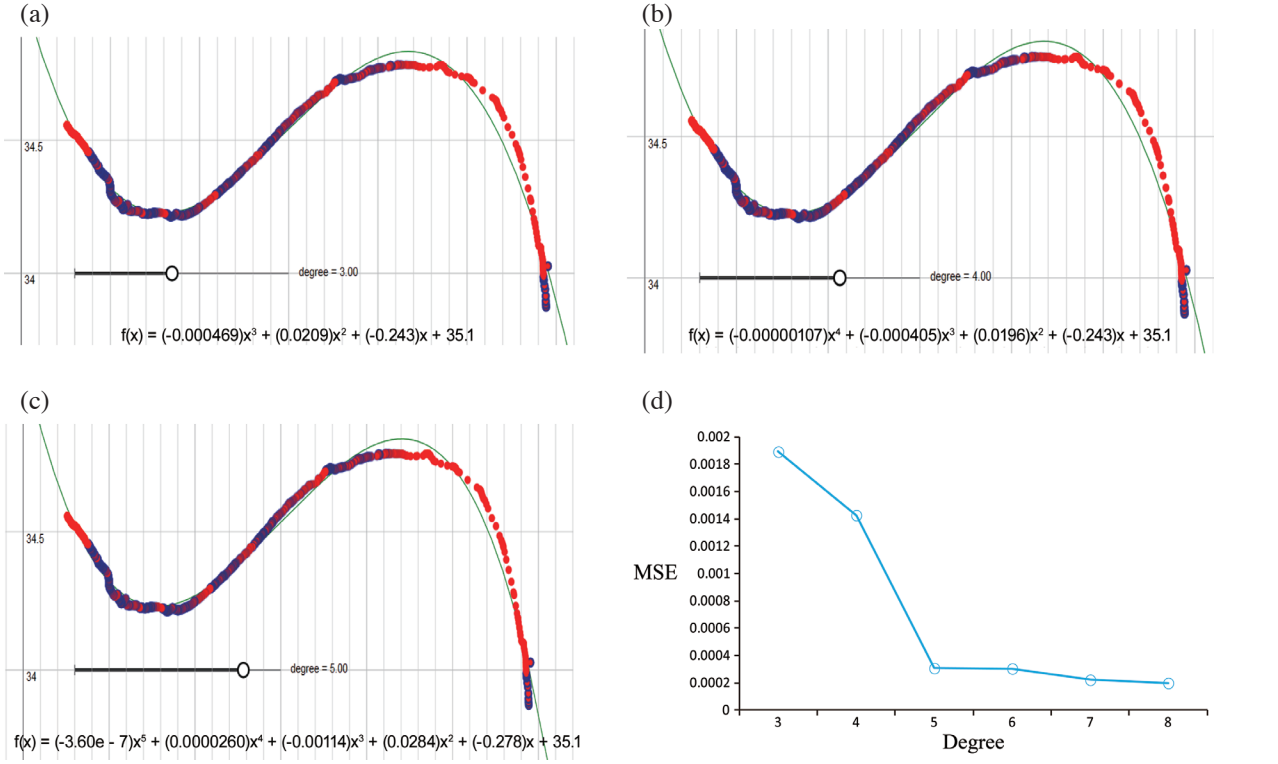

Fig. 3. Examples of regression analysis for a set of T-S data from the Kuroshio dataset with polynomials of (a) degree 3 (b) degree 4 (c) degree 5, respectively. The points in each figure are the training samples, and the solid green curve is the derived polynomial. (d) Further shows the mean square error (MSE) with respect to the degree.

characteristics needs to be defined beforehand. Euclidean distance is a classic and popular distance measure in mathematics (Patel and Mehta 2012). Suppose $p = (p_1, p_2, \ldots, p_n)$ and $q = (q_1, q_2, \ldots, q_n)$ are two vectors in an n-dimensional space. The Euclidean distance between the two vectors, p and q, is then

$$d(p, q) = d(q, p) = \left[\sum_{i=1}^{n} (p_i - q_i)^2\right]^{\frac{1}{2}} \quad (6)$$

Equation (6) can be generalized to measure the distance between two T-S characteristics, provided that the T-S characteristics such as those in Fig. 3 can be converted into n-tuple vectors. This can be achieved by sampling the salinity-temperature polynomial at certain temperatures. Assume that the regression function is $s = f(t; \beta)$, where t is the temperature, s is the salinity, and β is the coefficient of the polynomial. An n-dimensional vector $s = (s_1, s_2, \ldots, s_n)$ can then be obtained by sampling the polynomial function $f(t; \beta)$ uniformly as below,

$$\begin{aligned} &s_{k+1} = f(t_{\min} + k\Delta t; \beta), 0 \le k < n \\ &\Delta t = (t_{\max} - t_{\min}) / (n - 1) \end{aligned} \quad (7)$$

where $t_{\min}$ and $t_{\max}$ are respectively the lowest and the highest temperatures for the observed data and Δt is the space between any two adjacent temperatures. Accordingly, every regression polynomial (T-S characteristic) can be represented as an n-dimensional vector and the distance between two polynomials can be computed based on Eq. (7). With the distance measure a set of water masses can be compared automatically based on their T-S characteristics. It should be noticed, however, the meaning of distance is opposite to that of similarity. The smaller the distance is, the more similar the two T-S characteristics are. Therefore, for a water mass at the target location, finding the most similar water mass is in fact equivalent to finding those geographical locations whose T-S characteristics are closest to the T-S characteristic of the target location.

3. SYSTEM DESIGN

In the past decade ODB information systems were usually designed with some proprietary structure and focused on very limited functions. This resulted in a communication problem between the user and the developer, with high maintenance costs due to the lack of flexibility in the system architecture. New strategies for system design have been recently proposed to meet the demands for flexibility and adaptability, which triggered new ways of system implementation. SOA is such a solution that can facilitate the integration of various applications implemented with heterogeneous technologies as long as these applications release their services in the form of standardized interfaces (Guo et al. 2010). In SOA the application logics contained in various systems across the organization are exposed as services that can then be consumed by other applications (Chua and Lee 2009). This can help organizations mitigate the problems of legacy systems and increase the interoperability, reusability and flexibility while reducing the cost of development and maintenance. In addition, SOA can modularize the system to be extensible, scalable and compatible for future demands. Therefore, this is potentially a good solution for the ODB information system. In the system design that follows SOA is adopted as a reference model for the proposed ODB system.

3.1 System Architecture

According to SOA, the ODB information system can be divided into four layers, as shown in Fig. 4. Initially, marine observation data are collected automatically by the instruments on the cruises and stored as raw text files that are less structured and difficult to search. In the data layer the raw data are parsed, converted into the structural form and stored in the relational database for more efficient searches. The data in the ODB system includes Conductivity Temperature Depth (CTD) data, Acoustic Doppler Current Profiler (ADCP) data, EK500 acoustic image data, chemical data and biological data, as can be seen on the left hand side of Fig. 4. However, the amount of ocean data is too huge to be used efficiently, so the converted data might need to be further processed to obtain more aggregated or analytical data through statistical computation, data aggregation, regression analysis and so on. The Analytical Ocean Data shown in Fig. 4 is a type of analytical data derived from polynomial regression.

In the service component layer software components for accessing the data sources produced in the data layer can be built in order to provide basic CRUD (create, read, update, and delete) operations for the data tables. These components are exported as web services in the service layer, which can be discovered, described and accessed based on XML and standard Web protocols over intranets, extranets and the Internet (Alonso et al. 2004). The service layer can also provide interoperability among multiple components to accomplish more complex or integrated business logics when necessary. In the ODB system web services from four domains are implemented, including physical oceanography (PHY), marine geology geophysics (MGG), marine biology (BIO) and chemical oceanography (CHEM). In the application UI layer the web clients are devised to consume the services and interact with the user through such devices as desktop, tablet or mobile phone. The web client of the PHY domain finally consumes the PHY service using such web technologies as AJAX, JSXGraphic library and Google Maps API, to provide responsive interaction with the user.

3.2 Prototype Implementation

PHY service implementation is taken as an example in this section to illustrate how the SOA reference model in Fig. 4 is realized in the ODB system. The processes for implementing the PHY service are shown in Fig. 5 and presented in the following sections.

3.2.1 Data Conversion

As depicted in Fig. 5, the raw ocean data from shipboard were first parsed by the applications and then merged into the CTD database after quality control processes were applied. A few primary CTD data sample records in the database are displayed in Table 1. It can be seen from Table 1 that each record contains information on the salinity, temperature, time, and location in which the record was tracked. There are more than 20 million CTD records in the database, resulting in huge storage volume, poor indexing performance and longer search delay. Data aggregation is therefore necessary for dealing with such problems.

3.2.2 Data Aggregation

The primary CTD data are aggregated according to their geographical locations to increase the computational efficiency for TS characteristics, as shown in Fig. 5. The geographical area near Taiwan between 10° - 30° latitude and 110° - 130° longitude was uniformly divided into 15 × 15 minute squares, called grids. Each CTD record was then assigned uniquely to a grid according to its altitude and longitude, with the CTD data acquired for every grid. A few aggregated CTD data sample records with the same grid Centre ID are shown in Table 2. As can be seen from this table, the first two columns signify the longitude and the latitude of the observation location, respectively, while the third column signifies the observation depth. Columns 4 and 5 are the observed salinity and temperature, respectively, while column 6 contains the longitude and latitude of the grid centre that a record was assigned. Through data aggregation, 20 million CTD data records in ODB can be distributed into 1328 grids, and the aggregated data for every grid can further be used to compute the corresponding T-S characteristics through regression. Data aggregation is an effective pre-processing step for reducing regression computations based on the assumption that the T-S characteristic within a grid is stable.

3.2.3 Regression Analysis

With the aggregated data for every grid, the corresponding regression coefficients can be computed according to Eq. (4) depicted in section 2.3, and then stored in the analytical ocean data. Table 3 shows a few analytical ocean data sample records. In Table 3 the column Centre ID is the grid centre, while columns f0 - f5 are the coefficients of a regression polynomial of degree 5. The regression coefficients are used to produce the T-S characteristic for a grid. The distance between the T-S characteristics of two grids can be computed using Eq. (6) in section 2.4.

It is noteworthy that conventional T-S relationship displays the salinity on the x axis and the temperature on the y axis, as shown in Fig. 6a, where y is not a function of x since a value of x might correspond to multiple values of y. The problem can be solved using the generalized nonlinear regression in the form of $g(x, y) = 0$, such as the quadratic equation shown in Fig. 6b. However, to simplify the computation, the coordinates x and y for each record are swapped such that the salinity-temperature relationship, as shown in Fig. 6c, can be obtained through polynomial regression, as depicted in Fig. 6d.

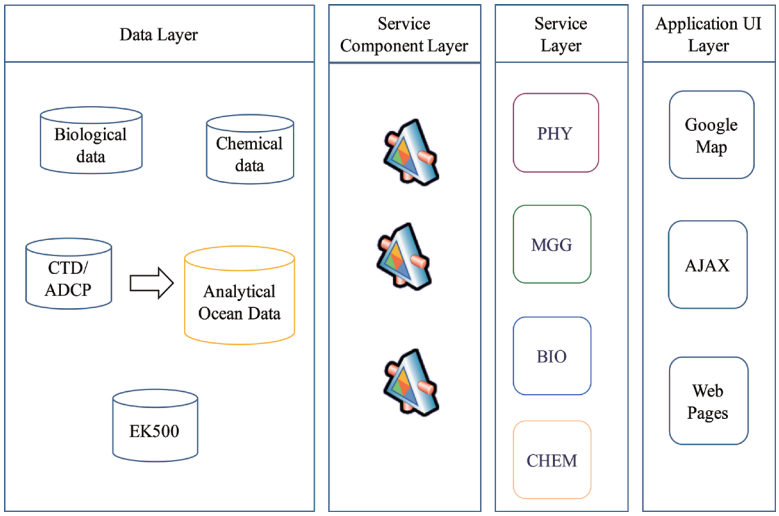

Fig. 4. The ODB information system architecture using SOA as the reference model.

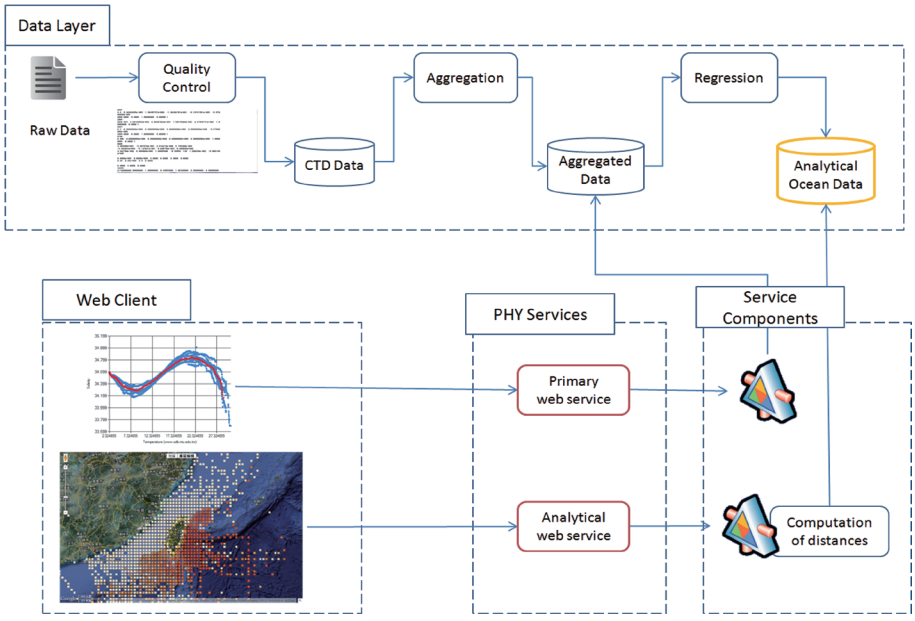

Fig. 5. Processes for the implementation of PHY service.

Table 1. Sample records of primary CTD data.

Cruise_ID	Cast_ID	Local_Time	Longitude_degree	Latitude_degree	Depth	Salinity	Temperature
010782	D6D1	2006-02-25 16:11:10	117754333	21488000	3	34.313	24.943
010782	D6D1	2006-02-25 16:11:10	117754333	21488000	4	34.313	24.943
010782	D6D1	2006-02-25 16:11:10	117754333	21488000	5	34.313	24.942
010782	D6D1	2006-02-25 16:11:10	117754333	21488000	6	34.313	24.942
010782	S7D1	2006-02-24 22:22:22	117263400	21595667	4	34.276	24.740
010782	S7D1	2006-02-24 22:22:22	117263400	21595667	5	34.277	24.740
010782	S7D1	2006-02-24 22:22:22	117263400	21595667	6	34.277	24.739

Table 2. The sample records of aggregated CTD data.

Longitude	Latitude	Depth	Salinity	Temperature	CentreID
117.3394	21.29633	2	33.344	29.651	117.25_21.25
117.3394	21.29633	3	33.344	29.61	117.25_21.25
117.3394	21.29633	4	33.343	29.596	117.25_21.25
117.3394	21.29633	5	33.342	29.586	117.25_21.25
117.3394	21.29633	6	33.343	29.566	117.25_21.25
117.3394	21.29633	7	33.343	29.562	117.25_21.25
117.3394	21.29633	8	33.344	29.554	117.25_21.25
117.3394	21.29633	9	33.343	29.546	117.25_21.25
117.3394	21.29633	10	33.344	29.54	117.25_21.25

Table 3. Sample records of analytical ocean data containing the coefficients of regression polynomials.

CentreID	f0	f1	f2	f3	f4	f5
117.25_21	38.53432	-1.27061	0.14652	-0.00797	0.000214	-2.37E-06
117.25_21.25	40.27099	-1.91005	0.235324	-0.01375	0.000389	-4.36E-06
117.25_21.5	70.47117	-10.3587	1.157567	-0.06297	0.001678	-1.76E-05
117.25_21.75	53.17164	-5.27742	0.574457	-0.0303	0.000785	-8.10E-06
117.25_22	-677.386	163.8514	-14.9604	0.677293	-0.0152	0.000135
117.25_22.25	-2586.52	590.4141	-52.8979	2.355771	-0.05214	0.000459
117.25_22.5	12063.78	-2440.18	197.1876	-7.93509	0.159036	-0.00127
117.25_22.75	-66638.4	13460.95	-1084.17	43.54345	-0.87207	6.97E-03
117.25_23	828.9836	-144.322	10.43299	-0.37599	0.006779	-4.92E-05
117.25_23.25	23197.14	-2994.93	119.0585	-0.33716	-0.07426	0.001185

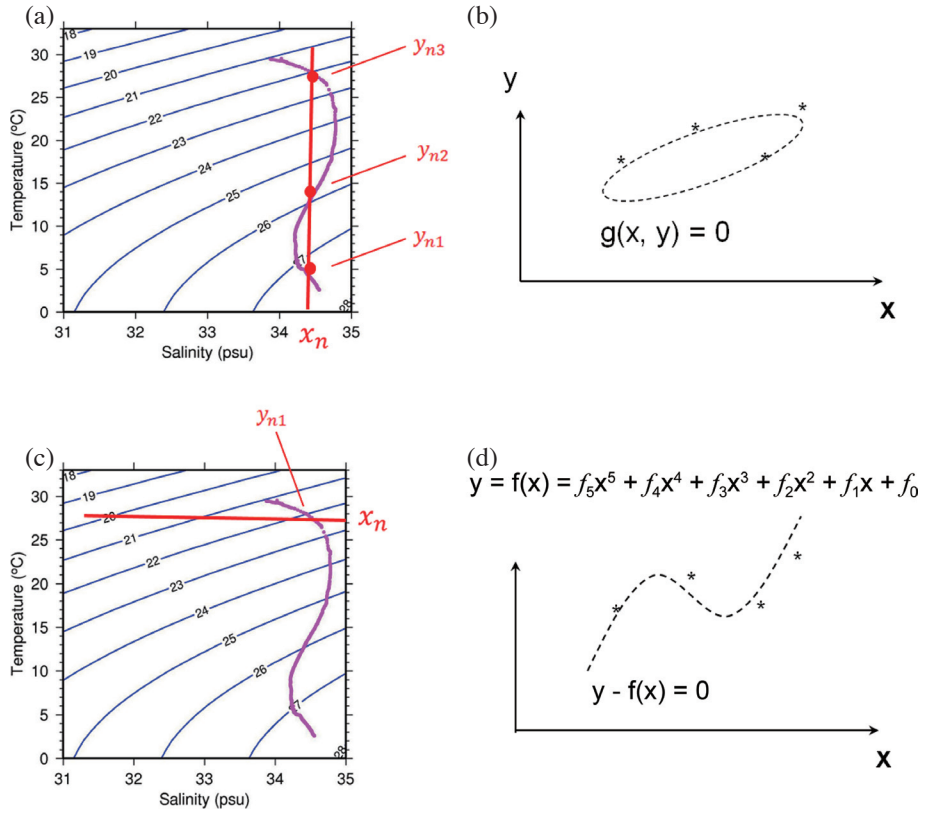

Fig. 6. T-S relationship with respect to regression model. Conventional T-S relationship displays the salinity on the x axis and the temperature on the y axis as shown in (a), where y is not a function of x. The relationship can be solved with nonlinear regression in the form of g(x, y) = 0, such as the quadratic equation shown in (b). To simplify the computation, the coordinates x and y for each record are swapped such that the salinity-temperature relationship, as shown in (c), can be obtained through polynomial regression, as depicted in (d).

3.2.4 Web Services and Web Client

The analytical ocean data produced in regression analysis can be further exported, either directly as a primary web service for retrieving the T-S characteristic at an assigned grid, or indirectly as an analytic web service for performing large-scale analysis, as shown in Fig. 5. In the primary web service the T-S characteristic is obtained based on the regression coefficients for an assigned grid. In the analytical web service, on the other hand, the distances between a reference grid and the other grids are computed according to the T-S characteristics obtained from the regression

coefficients in the analytical ocean data. The distances for all grids with respect to the reference grid are replied to the client for visual presentation to the user.

The steps for computing the distances are displayed in Fig. 7. Assume that we have a set of temperature and salinity data for a reference grid, as shown in Fig. 7b. The lowest and highest temperatures, t_{min} and t_{max}, can then be obtained from the data. The range between t_{min} and t_{max} is then uniformly divided according to Eq. (7) such that a vector of salinities sampled at equally spaced temperatures for the reference grid can be produced, as shown in Fig. 7a. Similarly, the vectors for the other grids can also be obtained, and accordingly the Euclidean distances denoting the dissimilarities between these grids and the reference grid can be computed based on Eq. (6). The distances can be further encoded with gradient colors such that similar water masses for the reference grid can be easily observed. Figures 7c - f display the T-S characteristics for four grids with the corresponding encoding colors. It can be observed that the T-S characteristic in Fig. 7c is the most similar (with low distance) to that of the reference grid, so this grid is encoded with dark orange. The T-S characteristic in Fig. 7f is least similar (with high distance) to that of the reference grid, so this grid is encoded with light pink. With the encoding colors corresponding to all grids, similar water masses for the reference grid can be visualized geographically using a similar map that will be shown later.

The PHY web services can be integrated to provide an intuitive visualization interface which the user can interact with flexibly to explore the T-S characteristics of water masses. The web client was built with the Google Maps public API, which is simple, pre-styled, open and interoperable (Freifeld et al. 2008). Figure 8 displays a few snapshots of the user interface for the ODB system for a query dataset. Figure 8a is the input field containing the query CTD data of a reference grid in the Kuroshio region. Figure 8b is the corresponding T-S diagram and Fig. 8c is the similarity map of that grid. The setting as shown in Fig. 8a is used to filter the CTD data based on season or range of years. From the similarity map in Fig. 8c, it is quite easy to track how the

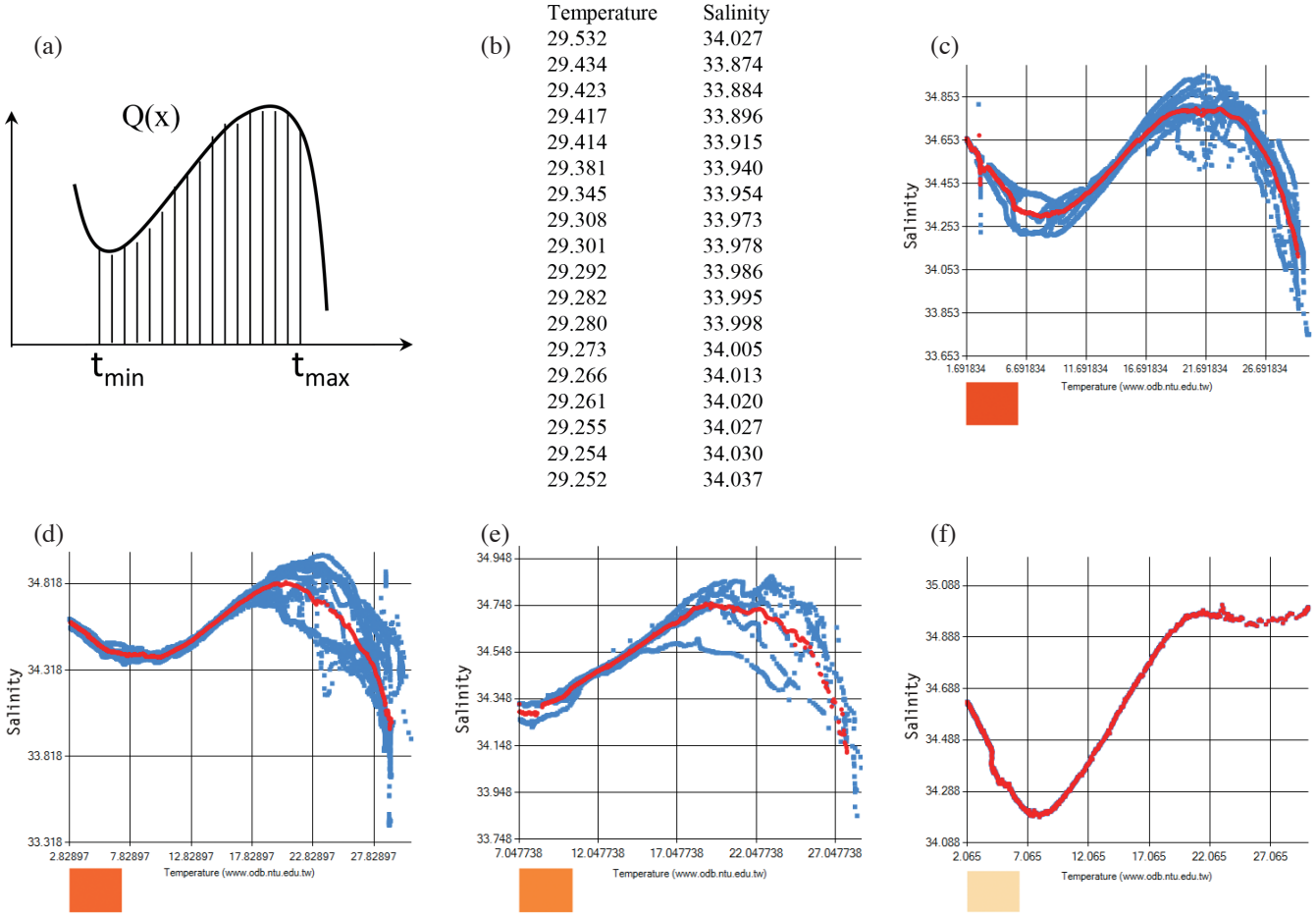

Fig. 7. Computation of distances and color coding. A set of temperature-salinity data for a reference grid is shown in (b). The vectors for all the grids can be obtained by sampling the polynomial. Accordingly, the distances between these grids and the reference grid can be computed and encoded with gradient colors such that similar water masses for the reference grid can be tracked. (c) through (f) display the T-S characteristics for four grids with the corresponding encoding colors.

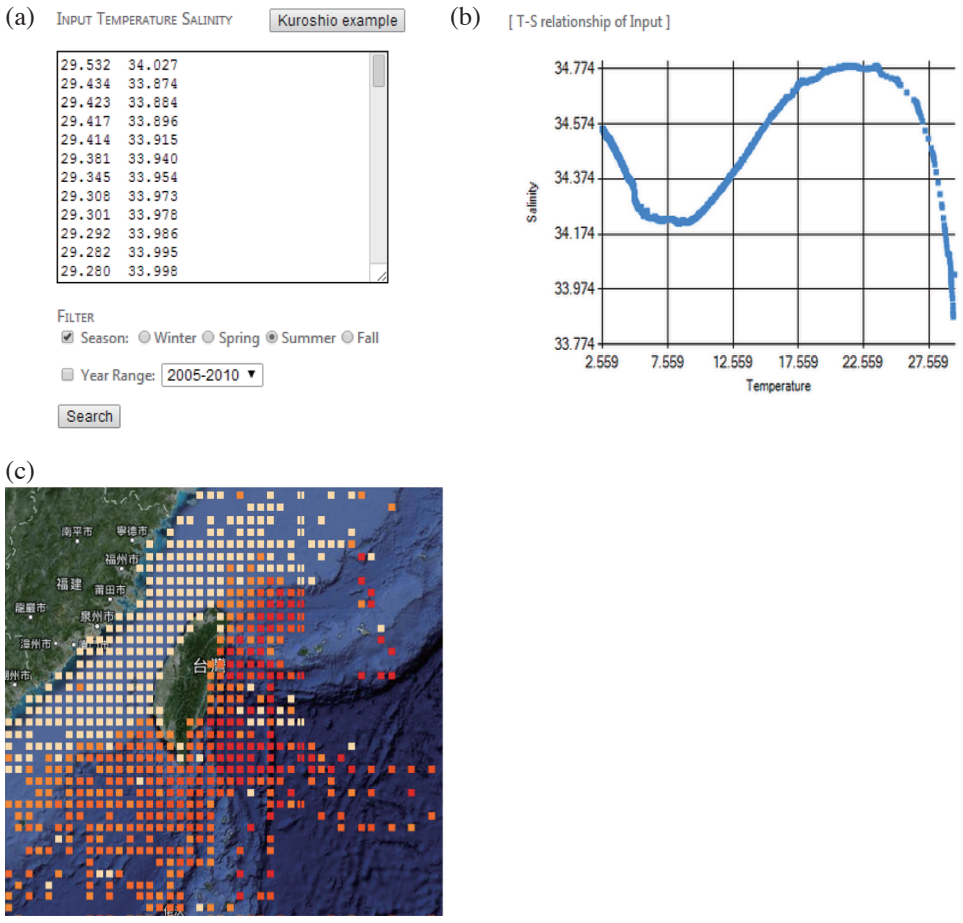

Fig. 8. Visualization interface for exploring T-S relationships. (a) is the input field of a query dataset in the Kuroshio region, area (b) is the T-S diagram of that dataset, and area (c) is the similarity map for the query. The setting on the bottom of (a) is for filtering the CTD data so as to see how the T-S characteristic and similarity map varies with respect to the season or the range of years.

characteristic of water masses varies with respect to the location. In addition, the system provides an interactive interface for querying the similarity map, as shown in Fig. 9. The user can click a grid on the similarity map displayed in Fig. 9a to see the corresponding T-S diagram, as shown in Fig. 9b, and compare it with that of the reference grid. The white cross in Fig. 9a is the location of the Kuroshio region in the reference grid. The red curve in Fig. 9b is the average value at the same depth in that grid. The information about the filter setting (season or range of year) and sample size (Data Count = 1734) is displayed on the top of Fig. 9b. Moreover, to estimate the appropriate sample size that can achieve acceptable quality in polynomial regression, the MSE with respect to the sample size is further plotted in Fig. 10. As can be seen from this figure, 200 is an appropriate sample size threshold for obtaining acceptable quality of regression. Therefore, those grids with the number of samples below the threshold will not be displayed.

In the next section, the similarity map will be compared with other ocean researches, while the seasonal and multi-year changes for the T-S characteristic or similarity map will be further discussed.

4. DISCUSSION

In the previous sections an approach comparing and discovering similar water masses based on T-S characteristic was proposed and realized on a visualization tool of ODB system. To verify the validity of this approach, the result from the Kuroshio example was compared with earlier studies of the Kuroshio path (Liang et al. 2008) and the Kuroshio path with an ADCP presentation from ODB, as shown in Figs. 11a - c, respectively. Figure 11a is the result of this study using the Kuroshio example data as input. Figure 11b is the climatological temperature figure and daily drifting trajectory at 30 m depth which shows the Kuroshio flows northward along the east coast of Taiwan (ODB 2010). Figure 11c is the Kuroshio path with an ADCP presentation from ODB. The Kuroshio paths in Fig. 11b or c coincide highly with the dark orange region in Fig. 11a, which is the result of the proposed approach. It implies that the distance measured based on the polynomial regression

function is successful for querying similar water masses in a wide range area. The result is consistent with those obtained from earlier researches. The proposed approach effectively reduces the amount of computations through data aggregation without losing the precision and facilitates the comparison of water masses through T-S characteristics sampling without degrading their discriminating capabilities. The ODB system has been deployed as a web site for promoting joint marine research, and is available publicly at the URL: http://app05.odb.ntu.edu.tw/phy/BlockQry.aspx.

In addition, seasonal variation was determined an important feature of current flow near Taiwan in earlier studies by Tang et al. 2000 and Jan et al. 2002. For comparison, the CTD data are aggregated individually according to the seasons in the data aggregation process and the T-S relationships for four seasons are shown in Fig. 12. As can be seen from this figure, seasonal change is prominent for the Kuroshio nearby Taiwan. In addition, the seasonal change in the similarity map for two reference grids, one in the Kuroshio region nearby Taiwan and the other in the South China Sea, are shown in Figs. 13 and 14, respectively. The difference in the patterns in different seasons could be helpful for illustrating the Kuroshio Intrusion phenomenon into the South China Sea in fall (Farris and Wimbush 1996). Similarly, the range of years can be set to filter the CTD data in the Kuroshio region, as shown in Fig. 8a, to obtain similarity maps for a long period of time. Four multi-year ranges from 1985 - 2010 are set for aggregating the CTD data, with the corresponding similarity maps displayed in Fig. 15. This figure shows that the Kuroshio patterns are broadly similar

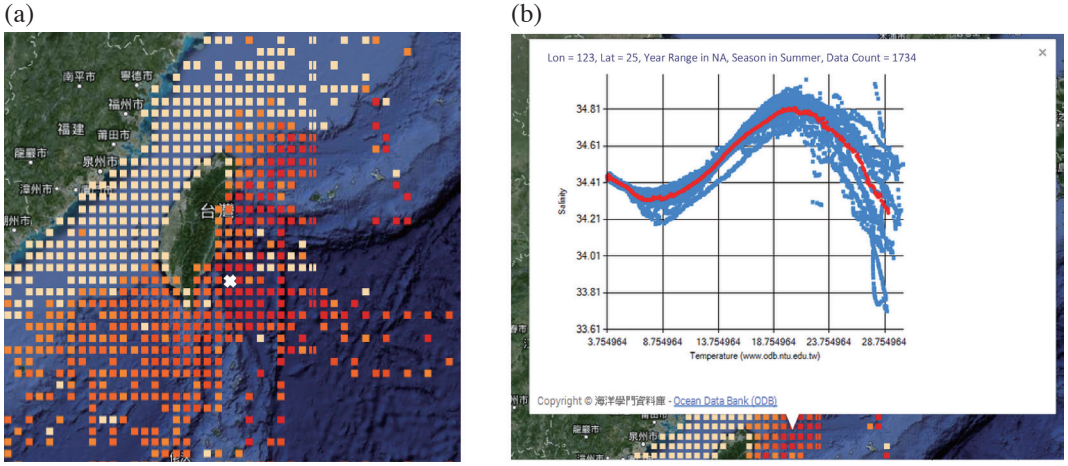

Fig. 9. Interactive query interface. The user can click a grid in (a) to see the corresponding T-S diagram, as shown in (b). On the similarity map, the white cross is the reference grid and the red curve is the T-S relationship of the grid. The information about filter setting (season or range of years) and sample size (Data Count = 1734) is shown on the top of (b).

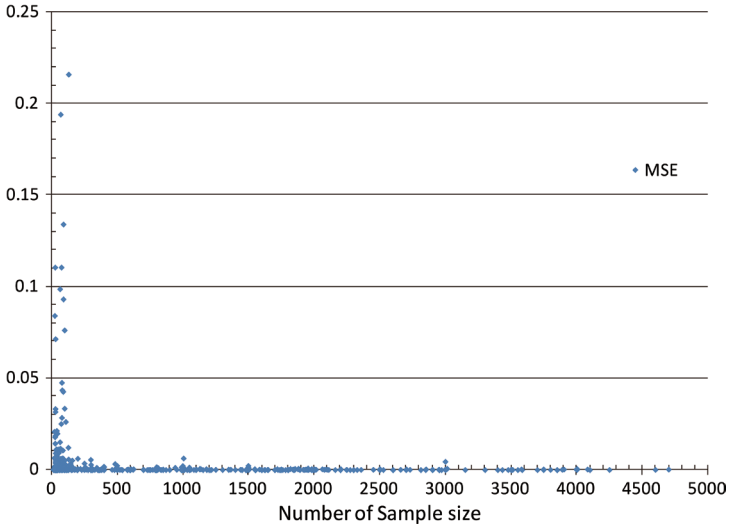

Fig. 10. The mean squared error (MSE) with respect to the sample size. 200 is an appropriate sample size threshold that can achieve acceptable quality in polynomial regression.

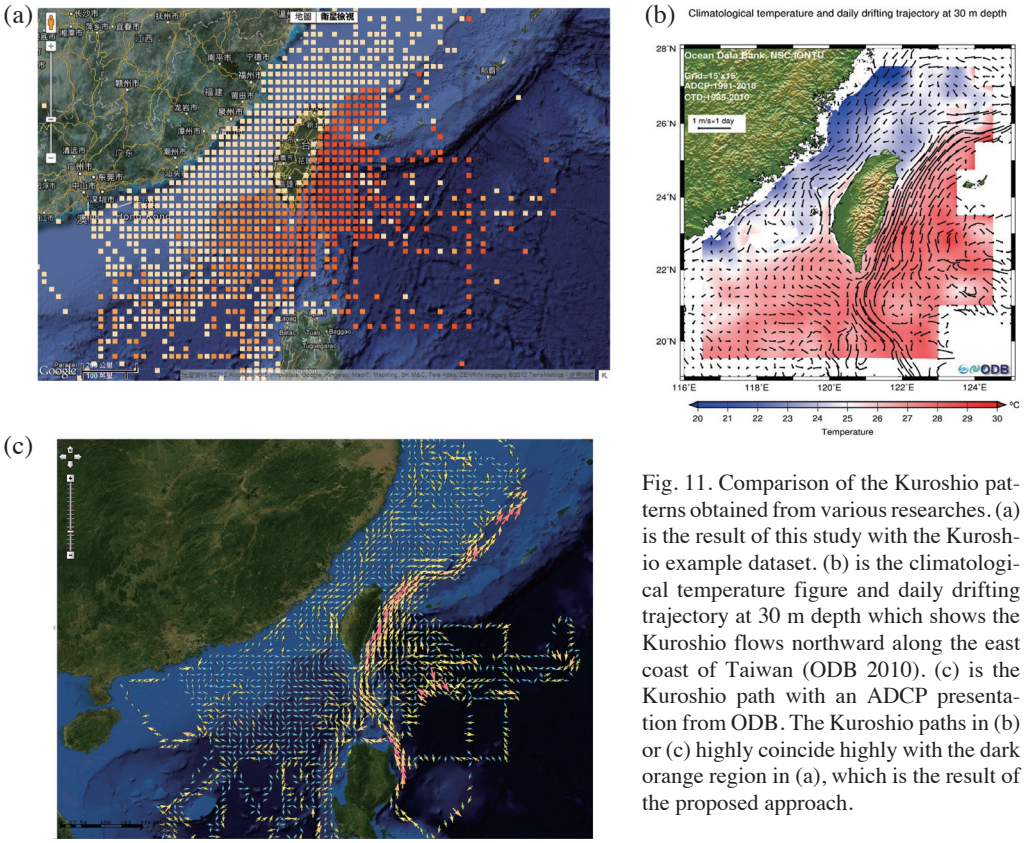

Fig. 11. Comparison of the Kuroshio patterns obtained from various researches. (a) is the result of this study with the Kuroshio example dataset. (b) is the climatological temperature figure and daily drifting trajectory at 30 m depth which shows the Kuroshio flows northward along the east coast of Taiwan (ODB 2010). (c) is the Kuroshio path with an ADCP presentation from ODB. The Kuroshio paths in (b) or (c) highly coincide highly with the dark orange region in (a), which is the result of the proposed approach.

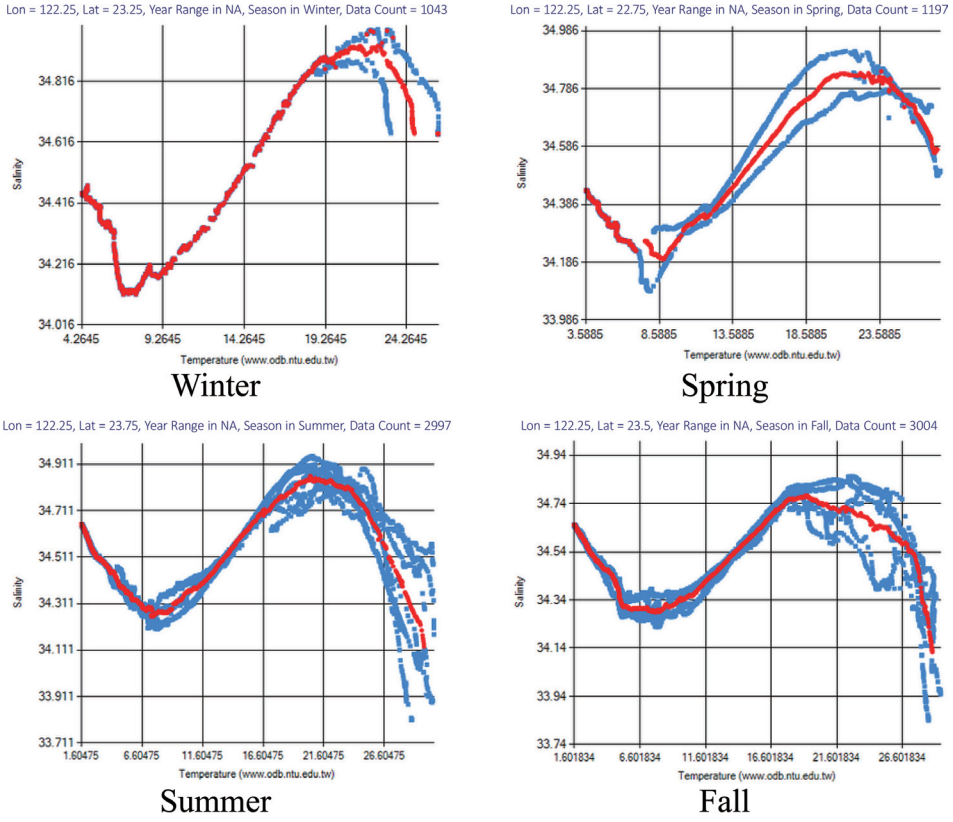

Fig. 12. Seasonal change of T-S characteristic for the Kuroshio nearby Taiwan.

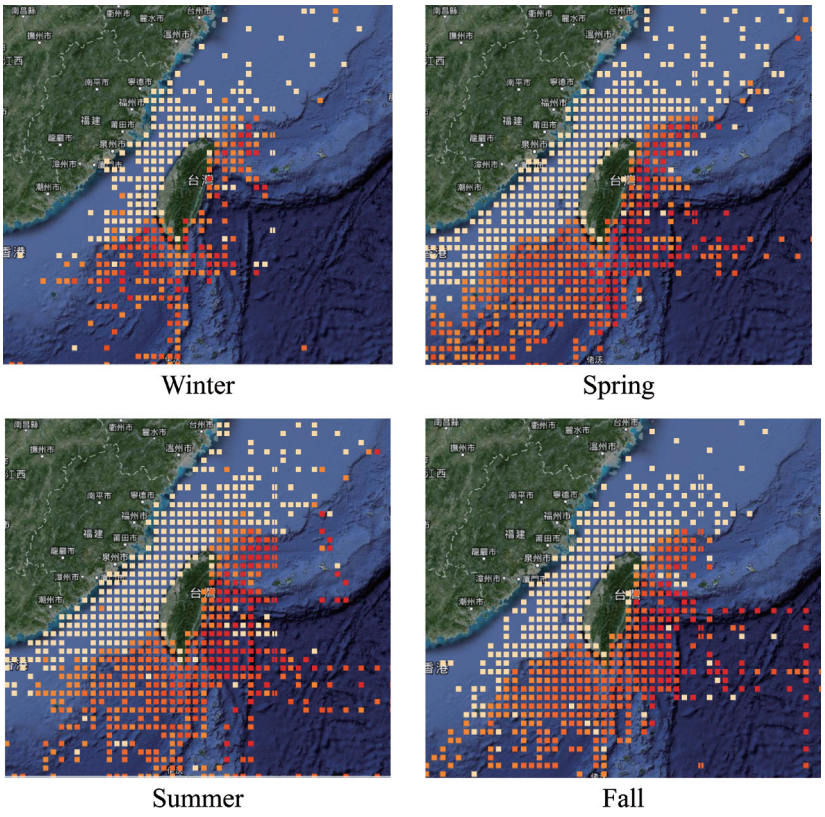

Fig. 13. Seasonal change of similarity map for the Kuroshio nearby Taiwan.

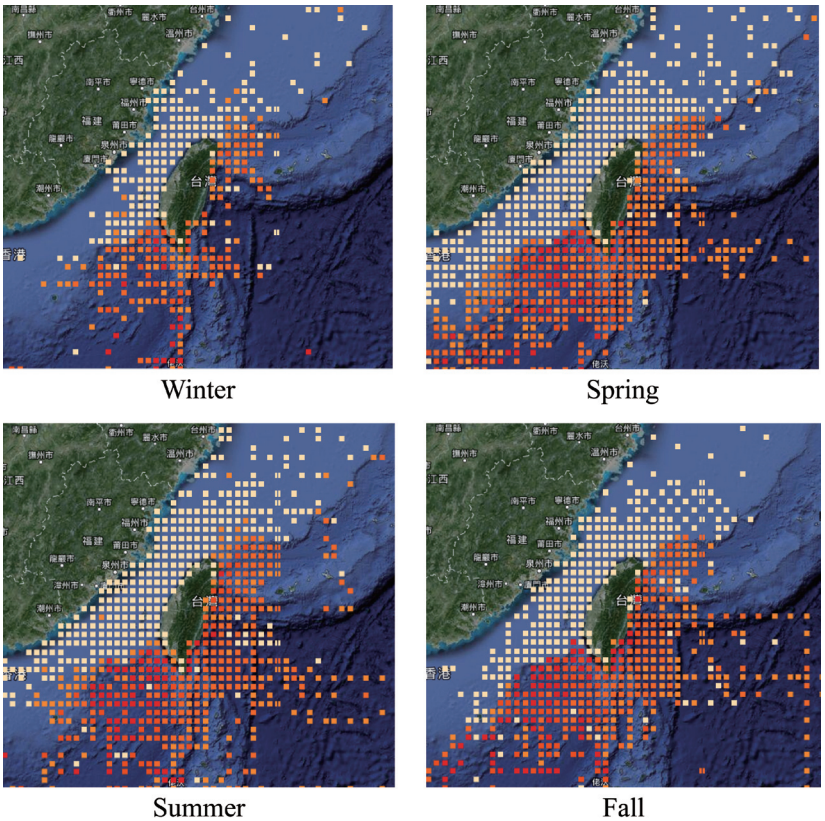

Fig. 14. Seasonal change of similarity map in the South China Sea. The differences in the patterns in different seasons could be helpful in illustrating the phenomenon of Kuroshio Intrusion into the South China Sea in fall (Farris and Wimbush 1996).

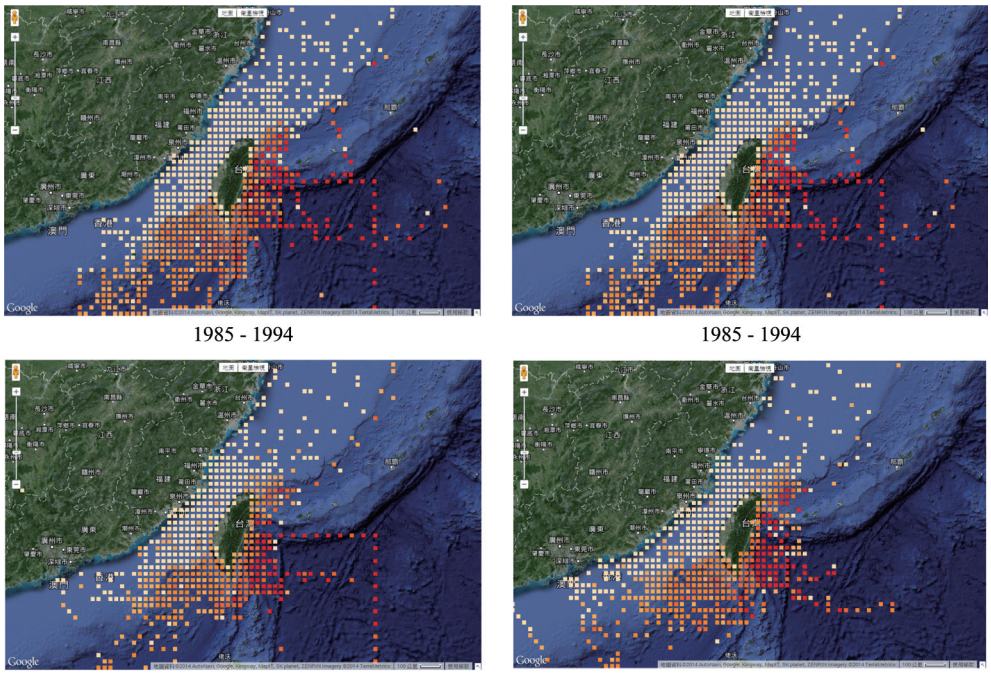

Fig. 15. Multi-year changes in similarity map in Kuroshio nearby Taiwan. The Kuroshio patterns are broadly similar for many years based on long-term observation.

for many years based on long-term observation. Notice that here the multi-year range is around 5 years in order to aggregate sufficient CTD data for computing T-S characteristics because the amount of data varies drastically year to year. Since this tool facilitates the comparison of water masses across seasons, multi-year ranges, or locations, explorations such as a long-term geographical change in T-S characteristic can be achieved efficiently, but the knowledge discovery based on them becomes more feasible.

5. CONCLUSIONS

This paper presented the T-S characteristic is modelled as a polynomial function with coefficients estimated using statistical regression. With such representation, the distance between two T-S characteristics could be measured and the search and comparison of similar water masses can then be conducted automatically and efficiently. This analysis approach can search geographic locations with similar T-S characteristics and also track water masses with similar T-S characteristics for a wide range area. In addition, the change in T-S characteristic across seasons or multi-year ranges can be easily observed by setting the filtering criteria. The validity of the proposed approach was successfully verified using the Kuroshio dataset. This approach can obtain the Kuroshio pattern that coincides highly with those obtained from earlier researches. This implies that the data aggregation using grids and the simplification for comparing regression polynomials can effectively reduce the computation without largely degrading the discriminating capabilities of T-S characteristics in water masses.

Moreover, with SOA as a reference model, the ODB information system can achieve good reusability and flexibility for integration and provide a visualization interface that can interact responsively with the user. This can effectively facilitate the knowledge exploration and discovery process. An implementation case using PHY service was used as an example to illustrate how primary and analytical data can be flexibly exported as web services and easily integrated based on this architecture. The layered SOA makes it easier to reuse the modules from various data sources to fulfil the information needs of different users, including primary data, statistical data and analytical data. This ODB information system is flexible and adaptable for future decision supporting systems for marine research and can promote joint research across domains.

Acknowledgements We would like to express our appreciation to the Ocean Data bank of Ministry of Science and Technology for providing the historical CTD data for this study (http://www.odb.ntu.edu.tw).

REFERENCES

Alonso, G., F. Casati, H. Kuno, and V. Machiraju, 2004: Web Services: Concepts, Architectures and Applications, Springer Science & Business Media, 354 pp.

Barker, P. A., F. A. Street-Perrott, M. J. Leng, P. B.

Greenwood, D. L. Swain, R. A. Perrott, R. J. Telford, and K. J. Ficken, 2001: A 14,000-year oxygen isotope record from diatom silica in two alpine lakes on Mt. Kenya. *Science*, **292**, 2307-2310, doi: 10.1126/science.1059612. [Link]

Cetisli, B. and H. Kalkan, 2011: Polynomial curve fitting with varying real powers. *Electron. Electr. Eng.*, **112**, 117-122, doi: 10.5755/j01.eee.112.6.460. [Link]

Chien, L. K., T. S. Feng, C. C. Yen, B. C. Lee, and H. W. Chang, 2010: The application of e-technology for marine information service. *J. Mar. Sci. Technol.*, **18**, 797-808.

Chua, F. F. and C. S. Lee, 2009: Collaborative learning using service-oriented architecture: A framework design. *Knowl. Base. Syst.*, **22**, 271-274, doi: 10.1016/j.knosys.2009.01.003. [Link]

Farris, A. and M. Wimbush, 1996: Wind-induced Kuroshio intrusion into the South China Sea. *J. Oceanogr.*, **52**, 771-784 , doi: 10.1007/BF02239465. [Link]

Freifeld, C. C., K. D. Mandl, B. Y. Reis, and J. S. Brownstein, 2008: HealthMap: Global infectious disease monitoring through automated classification and visualization of internet media reports. *J. Am. Med. Inform. Assn.*, **15**, 150-157, doi: 10.1197/jamia.M2544. [Link]

Garlan, D., 2000: Software architecture: A roadmap, Proceedings of the Conference on the Future of Software Engineering, 91-101, doi: 10.1145/336512.336537. [Link]

Guo, L., J. Gong, J. Sun, and X. Wei, 2010: Study on GIS architecture based on SOA and RIA. 2010 3rd International Conference on Information Sciences and Interaction Sciences (ICIS), IEEE, China, Chengdu, 620-625, doi: 10.1109/ICICIS.2010.5534675. [Link]

Jan, S., J. Wang, C. S. Chern, and S. Y. Chao, 2002: Seasonal variation of the circulation in the Taiwan Strait. *J. Mar. Syst.*, **35**, 249-268, doi: 10.1016/S0924-7963-(02)00130-6. [Link]

Kim, K., K. R. Kim, T. S. Rhee, H. K. Rho, R. Limeburner, and R. C. Beardsley, 1991: Identification of water masses in the Yellow Sea and the East China Sea by cluster analysis. *Elsevier Oceanogr. Ser.*, **54**, 253-267, doi: 10.1016/S0422-9894(08)70100-4. [Link]

Liang, W. D., Y. J. Yang, T. Y. Tang, and W. S. Chuang, 2008: Kuroshio in the Luzon Strait. *J. Geophys. Res.*, **113**, doi: 10.1029/2007JC004609. [Link]

Mamayev, O. I., 1975: Temperature-Salinity Analysis of World Ocean Waters, Elsevier Oceanography Series, Elsevier, Amsterdam, 247 pp.

Mielke, P. W., K. J. Berry, C. W. Landsea, and W. M. Gray, 1996: Artificial skill and validation in meteorological forecasting. *Weather Forecast.*, **11**, 153-169, doi: 10.1175/1520-0434(1996)011<0153:ASAVIM>2.0.CO;2. [Link]

Ocean Data Bank (ODB), 2010: Ocean Data Bank of Ministry of Science and Technology/National Taiwan University 1985-2010 Data Report, Institute of Oceanography, National Taiwan University, Taiwan, 47.

Patel, V. R. and R. G. Mehta, 2012: Data clustering: Integrating different distance measures with modified k-Means algorithm. In: Deep, K., A. Nagar, M. Pant, and J. C. Bansal (Eds.), Proceedings of the International Conference on Soft Computing for Problem Solving (SocProS 2011) December 20-22, 2011, Vol. 2, Springer India, 691-700, doi: 10.1007/978-81-322-0491-6_63. [Link]

Shaw, P., D. Greenstein, J. Lerch, L. Clasen, R. Lenroot, N. Gogtay, A. Evans, J. Rapoport, and J. Giedd, 2006: Intellectual ability and cortical development in children and adolescents. *Nature*, **440**, 676-679, doi: 10.1038/nature04513. [Link]

Stigler, S. M., 1971: Optimal experimental design for polynomial regression. *J. Am. Stat. Assoc.*, **66**, 311-318, doi: 10.1080/01621459.1971.10482260. [Link]

Tang, T. Y., J. H. Tai, and Y. J. Yang, 2000: The flow pattern north of Taiwan and the migration of the Kuroshio. *Cont. Shelf Res.*, **20**, 349-371, doi: 10.1016/S0278-4343(99)00076-X. [Link]

Teague, W. J., M. J. Carron, and P. J. Hogan, 1990: A comparison between the Generalized Digital Environmental Model and levitus climatologies. *J. Geophys. Res.*, **95**, 7167-7183, doi: 10.1029/JC095iC05p07167. [Link]