Article ID: 1006-8775(2013) 01-0059-08

INTERPOLATION TECHNIQUE FOR SPARSE DATA BASED ON INFORMATION DIFFUSION PRINCIPLE—ELLIPSE MODEL

ZHANG Ren (张 韧)[1], HUANG Zhi-song (黄志松)[1], LI Jia-xun (李佳讯)[1], LIU Wei (刘 巍)[2]

(1. College of Meteorology and Oceanography, PLA University of Science and Technology, Nanjing 211101 China; 2. School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031 China)

Abstract: Addressing the difficulties of scattered and sparse observational data in ocean science, a new interpolation technique based on information diffusion is proposed in this paper. Based on a fuzzy mapping idea, sparse data samples are diffused and mapped into corresponding fuzzy sets in the form of probability in an interpolation ellipse model. To avoid the shortcoming of normal diffusion function on the asymmetric structure, a kind of asymmetric information diffusion function is developed and a corresponding algorithm-ellipse model for diffusion of asymmetric information is established. Through interpolation experiments and contrast analysis of the sea surface temperature data with ARGO data, the rationality and validity of the ellipse model are assessed.

Key words: information diffusion; interpolation algorithm; sparse data; ellipse model

CLC number: P731/O241.3 **Document code:** A

1 INTRODUCTION

Accurate and reliable oceanic observational data are very important for marine research and ocean-atmosphere numerical forecast. However, the most difficult obstacle nowadays is the lack of oceanic environmental data and the lack of data analysis and information extraction technique.

ARGO (Array for Real-time Geostrophic Oceanography) is a global oceanic observation and investigation project proposed and supported by the United States, Great Britain, France, Japan, China, and other countries. The goal of the ARGO project is to establish a global real-time observational network which consists of about 3000 automation floats and can quickly and accurately collect the global temperature and salinity profiling data and their characteristics. The global ARGO observational network can provide more than 10,000 temperature and salinity real-time observational profiles monthly. The ARGO data has such advantages as vast spatial range, long duration and sufficient observational factors and supports whole weather detection, which can efficiently make up for the shortage of the oceanic environment data, especially the observations in the deep layers[1].

However, there are some inherent limitations in the ARGO data: the average array interval of ARGO floats was about 300 km, and there is only one profile about every 10 days for each ARGO float. For regional oceanic and atmospheric research, the ARGO data is sparse in space and discrete in time. For partial key sea areas, such as the South China Sea and the northwest Pacific, the temporal-spatial resolution of ARGO data is far from enough. Considering the factors of floats shift and the difference of measuring periods and regional distribution, the available ARGO data are even rare[2].

Interpolation algorithms are common techniques for estimating and approaching the missing information with nearby data. The current interpolation algorithms contain the Newton method, Lagrange method, spline interpolation, polynomial interpolation, finite element method, weighing function method, variational method, spectrum method, successive correction method, optimal interpolation and so on[3, 4], which can basically meet the needs of the interpolation and fitting of large-scale atmospheric and oceanic data.

However, a key premise involved is that the above interpolated techniques need to be provided with sufficient data and related information. If the observational data is too rare, the precision and reliability of the interpolation algorithm will be

Received 2011-04-21; **Revised** 2012-10-23; **Accepted** 2013-01-15
Foundation item: Project of Natural Science Foundation of China (41276088)
Biography: ZHANG Ren, Ph.D., professor, primarily undertaking research on air-sea interaction.
Corresponding author: ZHANG Ren, e-mail: zren63@126.com

restricted greatly. There is universal scattered and sparse observational data such as ARGO data in the oceanic environment, but the general interpolation techniques have serious limitations in dealing with such type of data[5]. From the analysis above, it has important scientific meaning and practical value to develop new interpolation techniques to address the issue of sparse observation data. Ellipse interpolation algorithm model—a new interpolation technique for dealing with sparse data based on the information diffusion principle—was proposed in this paper.

2 THE PRINCIPLE OF INFORMATION DIFFUSION INTERPOLATION

2.1 *The Principle of information diffusion*

The technique of information diffusion is a concept of research and a mathematic model put forward for solving the imperfect information existing in evaluating strong natural disasters, such as earthquake, storm surge, mudslides, and etc[6], which are characterized by high degree of severity and small samples. Information diffusion is a fuzzy mathematic method by optimizing the fuzzy information of samples in order to make up for the missing information and can effectively deal with the imperfect sample information by transforming single samples into fuzzy-set samples in the form of probability[7]. Now, the information diffusion technique is only used for the risk assessment of the small-sample events[8-10], and its idea is suitable for solving the imperfect data such as sparse data interpolation.

Suppose that $W=\{W_1 \quad W_2 \quad \cdots \quad W_n\}$ is the knowledge sample series, L the underlying domain, and the observational value of W_i is l_i. Let $x=\phi\left(l-l_i\right)$, then if W is imperfect, there is a function $\mu\left(x\right)$ which can make the l_i point information (value-1) diffuse into l with $\mu\left(x\right)$. The diffused information distribution pattern $Q\left(l\right)=\sum_{j=1}^{n}\mu\left(x\right)=\sum_{j=1}^{n}\mu\left(\phi\left(l-l_i\right)\right)$ can better describe the whole structure of W, which is called the information diffusion principle[6].

Matrix probability density function estimation through the information diffusion principle is called diffusion estimation. The exact definition of the diffusion estimation is defined as follows. If $\mu\left(x\right)$ is defined on a Borel measurable function in $\left(-\infty,+\infty\right)$, $d>0$ is a constant, $x=\frac{l-l_i}{d}$, then

$$\hat{f}\left(l\right)=\frac{1}{nd_i}\sum_{i=1}^{n}\mu\left[\frac{l-l_i}{d}\right] \qquad (1)$$

is the diffusion estimation of the matrix probability density function $f(l)$ where $\mu\left(x\right)$ is the diffusion function and d is the window width.

2.2 *The idea of information diffusion*

In this paper, the idea of information diffusion is introduced into the fitting and interpolation of sparse data and a new interpolation technique—algorithm for information diffusion and interpolation, which is suitable for sparse data and small samples, is proposed.

2.2.1 FUZZY MAPPING RELATIONSHIP BETWEEN INPUT AND OUTPUT

For an input-output system, Ω is denoted as the matrix, x as the input variable, y as the output variable, X as the input set, and Y as the output set, viz. $x\in X$, $y\in Y$, $\Omega=X\times Y$.

Let $f(x,y)$ be the probability density function of matrix Ω, then the density of condition probability of y is described as follows with $x=u$.

$$f_{Y|X}(y\,|\,u)=f(u,y)\Big/\int_{v\in Y}f(u,v)dv\,. \qquad (2)$$

Based on the fuzzy-set idea, the Ω input-output system is defined to be an output fuzzy set $\tilde{B}$ under a given input, the membership function of fuzzy set $\tilde{B}$ is corresponding to the probability density function of output, and the normalized results of the probability density function are a membership function, i.e., the membership function of fuzzy set $\tilde{B}$ is described as follows,

$$\mu_{\tilde{B}}(y)=\frac{f(u,y)\Big/\int_{v\in Y}f(u,v)dv}{\max\limits_{y\in Y}\left\{f(u,y)\Big/\int_{v\in Y}f(u,v)dv\right\}}=\frac{f(u,y)}{\max\limits_{y\in Y}\left\{f(u,y)\right\}}\,. \quad (3)$$

Therefore, in the input-output system, for any given input x ($x\in X$), its whole potential output can be denoted by the fuzzy set $\tilde{B}$, which is the fuzzy mapping relationship between input and output.

Generally, the probability density function $f(x,y)$ of the matrix Ω is hardly acquired at hand but currently estimated by mass statistics of samples. Under the condition of rare data information, it will still be given an approximated estimation of the distribution of whole probability density by the information diffusion based on a few sample data obtained.

2.2.2 Estimation of information diffusion function

S is a set of small sample series for a constructed Ω input/output system, denoted as

$$S=\{(x_1,y_1),(x_2,y_2),\cdots(x_n,y_n)\}.$$

For the lack of samples, the probability density function is often hardly constructed by current common statistical analysis methods.

In the principle of information diffusion, small samples series S can be regarded as the information points scattering in the input/output phase space $X\times Y$. By point-set mapping, each sample data can be diffused as a fuzzy set with points of multiple samples. Because of the unclear, blur and flexible surrounding borderline, the information collectivity provided by each sample is a fuzzy information set. Addressing the unclear and blur borderline surrounding the sample point (x_i,y_i), a controlling point is introduced in the input space X and output space Y respectively, u_j ($j=1,2,\cdots s$) and v_k ($k=1,2,\cdots t$), which are normalized discrete information points in the input/output space. The sets for the input and output controlling points are separately marked as follows,

$$U=\{u_1,u_2,\cdots u_s\},\quad V=\{v_1,v_2,\cdots v_t\}.$$

Thus, the space of controlling points, $U\times V$, is structured in gridded distribution in the input-output space. Through a point of information infusion, information is reasonably and efficiently diffused into the whole space of controlling points by a suitable form (of information diffusion formula) to effectively capture and exploit the imperfect sample information.

Denote $A=\Omega\times U\times V$, define a mapping $\mu:A\to[0,1]$ for an argument field A to domain [0, 1], so that

$$\left((x,y),u,v\right)\in A\to\mu_{u_jv_k}\left(x_i,y_i\right)\in[0,1]$$

where $(x,y)\in\Omega$, $u\in U$, $v\in V$, $j=1,2,\cdots s$, $k=1,2,\cdots t$, $i=1,2,\cdots n$, and $\mu_{u_jv_k}\left(x_i,y_i\right)$ is called the information diffusion function.

Let $q_{u_jv_k}=\sum_{i=1}^{n}\mu_{u_jv_k}\left(x_i,y_i\right)$, marked as q_{jk} , $t=\sum_{j=1}^{s}\sum_{k=1}^{t}q_{jk}$. By information diffusion, the estimation of probability density function at point (u_j,v_k) in the Ω input-output system can be denoted as

$$\hat{f}(u_j,v_k)=\frac{q_{jk}}{t}. \tag{4}$$

2.2.3 Mapping interpolation with information diffusion

Substitute the information diffusion estimation $\hat{f}(u_j,v_k)$ of the probability density function into the input-output mapping relationship formula (3). When the input is u, the output is as follows:

$$\tilde{B}=\int_{y\in Y}\mu_{\tilde{B}}(y)/y=\int_{y\in Y}\frac{\hat{f}(u,y)}{\max\limits_{y\in Y}\left\{\hat{f}(u,y)\right\}}\Bigg/y. \tag{5}$$

(Note that the "$\int$" here is not the integration symbol and "/" is not for operation of division, but a basic fuzzy-set symbol)

Thus, based on the information diffusion principle, sparse sample data and limited sample information available, a fuzzy mapping relationship is established between the input and the output to determine interpolation results through de-fuzzy operation from the output.

2.3 *Algorithm of 2-D information diffusion interpolation*

There are many interpolation methods but current interpolation algorithms can hardly achieve the expected effect when the data sample is sparse or rare. For 2-dimensional data field interpolation, if the longitude and latitude information are regarded as the input and the variable values as the output, the two-dimensional data field can be achieved by the technique of information diffusion interpolation.

The main idea of the information diffusion is as follows. For an imperfect data sample, it can obtain more information of the sample by using a suitable diffusion function $\mu(x)$. For the information diffusion interpolation of sparse data, its goal is to seek an objective, accurate and efficient diffusion function and to achieve a reasonable diffusion and optimized approach for imperfect sparse data sample. In mathematic principle, the goal of the above information diffusion interpolation technique is to solve the small sample problem of the imperfect data, which is the chief difference from other interpolation methods.

Detailed operational steps are described as follows.

Let sparse sample series $S=[(x_1,y_1,v_1),(x_2,y_2,v_2),\cdots,(x_n,y_n,v_n)]$ and interpolate gridded coordinate: $X=[X_1,X_2,\cdots,X_s]$, $Y=[Y_1,Y_2,\cdots,Y_t]$, which are also known as controlling points of the input factor. Set a controlling point for the input sample $V=[V_1,V_2,\cdots,V_r]$ and then the interpolated field can be expressed as

$$F_{i,j}=R(\tilde{B})=R(\int_{v\in V}\mu_{\tilde{B}}(v)/v)=R\left(\int_{v\in V}\frac{\hat{f}((X_i,Y_j),v)}{\max\limits_{v\in V}\left\{\hat{f}((X_i,Y_j),v)\right\}}\Big/v\right) \tag{6}$$

where $\hat{f}((X_i,Y_j),V_k)=\frac{q_{ij,k}}{t}$, $t=\sum_{i=1}^{s}\sum_{j=1}^{t}\sum_{k=1}^{r}q_{ij,k}$, $q_{ij,k}=\sum_{l=1}^{n}\mu_{X_iY_j,V_k}\left(x_l,y_l,v_l\right)$ ($i=1,2,\cdots,s$, $j=1,2,\cdots,t$), R denotes the de-fuzzy operation, and $\mu_{X_iY_j,V_k}\left(x_l,y_l,v_l\right)$ is the diffusion function.

3 INFORMATION DIFFUSION INTERPOLATION MODEL

3.1 *Symmetrical information diffusion function—normal interpolation model*

The key of information diffusion is to construct an exact and reasonable diffusion function. Huang[6] deduced a simple but practical diffusion function by simulating the molecule diffusion.

$$\mu_{u_j}(x_i)=\frac{1}{\sqrt{2\pi}h}e^{-\frac{(u_j-x_i)^2}{2h^2}} \tag{7}$$

which is called the normal information diffusion function, where x_i is the sample point, u_j the controlling point, and h diffusion coefficient. Based on an "average space model" and "two-point handy principle", a simple calculation formula of the diffusion coefficient can be deduced as follows.

$$h=\begin{cases}0.8146(b-a), & n=5\\ 0.5690(b-a), & n=6\\ 0.4560(b-a), & n=7\\ 0.3860(b-a), & n=8\\ 0.3362(b-a), & n=9\\ 0.2986(b-a), & n=10\\ 2.6851(b-a)/(n-1), & n\geq 11\end{cases} \tag{8}$$

where $b=\max\limits_{1\leq i\leq n}\{x_i\}$, $a=\min\limits_{1\leq i\leq n}\{x_i\}$, n is the sample amount, and the diffusion coefficient is related with the amount and the maximum /minimum of the sample.

In 2-D information diffusion interpolation of sparse data, a 3-D (2-D input and 1-D output) diffusion function is needed and the 3-D normal information diffusion function is as follows.

$$\mu_{X_iY_j,V_k}\left(x_l,y_l,v_l\right)=\frac{1}{(\sqrt{2\pi})^3h_xh_yh_v}\exp[-\frac{(X_i-x_l)^2}{2h_x^2}-\frac{(Y_j-y_l)^2}{2h_y^2}-\frac{(V_k-v_l)^2}{2h_v^2}]. \tag{9}$$

The normal information diffusion function shows that the information diffuses and weakens symmetrically with space around the data point. Take a 2-D normal information diffusion for example. Its function is

$$\mu=\frac{1}{2\pi h_xh_y}\exp(-\frac{\Delta x^2}{2{h_x}^2}-\frac{\Delta y^2}{2{h_y}^2}) \tag{10}$$

where h_x and h_y are diffusion coefficients.

Make $x'=\frac{\Delta x}{\sqrt{2}h_x}$, $y'=\frac{\Delta y}{\sqrt{2}h_y}$. It means that the removal of the influence of the dimension and unit will then transform the 2-D normal information diffusion function as follows.

$$\mu=\frac{1}{2\pi h_xh_y}\exp[-(x'^2+y'^2)]. \tag{11}$$

The exponential part –($x'^2+y'^2=r^2$) shows that the information of sample points diffuses symmetrically along all directions after the removal of the influence of the dimension and unit in the 2-D data field, and the 3-D diffusion function is also the same. Therefore, the normal information diffusion is homogeneous and symmetrical.

As the actual data samples (such as the atmospheric and oceanic observational data) often have complicated structures and asymmetric characteristics, the normal diffusion model is only suitable for the ideal condition and the asymmetric information diffusion model of approaching to the factual condition must be taken into account to objectively describe and express the generalized, abnormal and asymmetric real data sample structure.

3.2 *Asymmetric information diffusion—ellipse algorithm model*

To overcome the shortcomings of the normal information diffusion model, a new ellipse algorithm model based on the asymmetric information diffusion is proposed. The data point's information often diffuses and extends not in a round symmetrical form but in an ellipse asymmetric form. That is, in some directions, the information diffuses quickly, which is defined as the ellipse long axis, while in other directions it diffuses slowly (defined as the ellipse short axis). Thus, a "round" symmetrical normal information diffusion model can be developed into an "ellipse" asymmetric information diffusion model, shown in Figure 1.

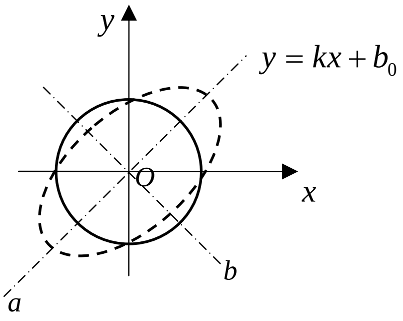

Figure 1. Sketches for normal (solid line, round) and asymmetric (dashed line, ellipse) information diffusion.

3.2.1 2-D "ELLIPSE" DIFFUSION FUNCTION

In the 2-D data structure shown in Figure 1, if the direction of fast diffusion is corresponding with the ellipse long axis (line a) and the direction of slow diffusion is corresponding with the ellipse short axis (line b), the information diffusion function can be transformed as follows:

$$\mu = \frac{1}{2\pi h_x h_y}\exp\{-\frac{1}{k^2+1}[\frac{1}{\lambda}(\frac{x}{\sqrt{2}h_x}+k\frac{y}{\sqrt{2}h_y})^2+(k\frac{x}{\sqrt{2}h_x}-\frac{y}{\sqrt{2}h_y})^2]\} \qquad (12)$$

where k is the slope rate of the ellipse long axis, called rotation coefficient, λ the rate square of the ellipse long axis, called a flex coefficient, and the function is called the ellipse asymmetric information diffusing function.

3.2.2 k AND λ PARAMETERS

In the 2-D ellipse asymmetric information diffusing function, two important parameters are introduced, i.e., rotation coefficient k and flex coefficient λ. k is the slope rate of the ellipse long axis, which is related with the xOy plane distribution of the sample and the sample information diffuses more quickly along the direction of the ellipse long axis. Thus the probability of the sample point around the long axis is the largest and the sum of the square distance from each sample point (x_i, y_i) to line $y = kx + b_0$ can be regarded as minimum, viz.

$$Q = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n}\frac{(kx_i + b_0 - y_i)^2}{k^2+1} = \min. \qquad (13)$$

Thus,

$$\begin{cases} \dfrac{\partial Q}{\partial b_0} = \sum\limits_{i=1}^{n}\dfrac{2(kx_i + b_0 - y_i)}{k^2+1} = 0 \\ \dfrac{\partial Q}{\partial k} = \dfrac{\sum\limits_{i=1}^{n}\left[2(kx_i + b_0 - y_i)(k^2+1) - 2k(kx_i + b_0 - y_i)^2\right]}{(k^2+1)^2} = 0 \end{cases}. \qquad (14)$$

And it can be simplified into

$$\begin{cases} b_0 = \dfrac{1}{n}(\sum\limits_{i=1}^{n} y_i - k\sum\limits_{i=1}^{n} x_i) \\ k^2 \cdot \sum \hat{x}_i\hat{y}_i + k\cdot\sum(\hat{x}_i^2 - \hat{y}_i^2) - \sum\hat{x}_i\hat{y}_i = 0 \end{cases} \qquad (15)$$

where $\hat{x}_i = x_i - \frac{1}{n}\sum x_i$, $\hat{y}_i = y_i - \frac{1}{n}\sum y_i$, and b_0 and k can be solved.

The flex coefficient λ denotes the fat/thin degree of the ellipse, which can be expressed as the rate of the average (or maximum) distance from each of the sample points to the short axis (line b) and to the long axis (line a), as shown in Figure 1.

3.2.3 MULTI-DIMENSIONAL ELLIPSE DIFFUSION FUNCTION

The non-dimensional form of the multi-dimensional (m+*1*–D) normal information diffusing function is presented as follows.

$$\mu = \frac{1}{(2\pi)^{\frac{m+1}{2}} h_y \prod\limits_{i=1}^{m} h_{x_i}}\exp\left[-\left(\sum_{i=1}^{m} x_i'^2 + y'^2\right)\right]. \qquad (16)$$

By introducing the flex coefficient, the "round" normal symmetrical information diffusing function can be developed as a “ellipse” asymmetric information diffusing function.

$$\mu = \frac{1}{(2\pi)^{\frac{m+1}{2}} h_y \prod_{i=1}^{m} h_{x_i}} \exp\left[-\left(\sum_{i=1}^{m} \frac{x_i'^2}{\lambda_i} + y'^2 \right) \right] . \quad (17)$$

The rotation factor (rotation coefficient) in the asymmetric information diffusing function $-\sum_{i=1}^{m} \frac{x_i'^2}{\lambda_i} - y'^2$ is introduced to make the direction of fast/slow diffusion corresponding with the long/short axis of the projection plane. Then, the coordinate of $x_i'(i = 1 \cdots m)$ is transformed in the following steps:

$$\begin{cases} x_i' = x_i' \cos\theta_i + y' \sin\theta_i \\ y' = -x_i' \sin\theta_i + y' \cos\theta_i \end{cases} . \quad (18)$$

Then,

$$\mu = \frac{1}{(2\pi)^{\frac{m+1}{2}} h_y \prod_{i=1}^{m} h_{x_i}} \exp\{-\sum_{i=1}^{m} \frac{1}{\lambda_i}[x_i' \cos\theta_i - \sum_{j=i+1}^{m} (x_j' \sin\theta_j \sin\theta_i \prod_{k=i+1}^{j-1} \cos\theta_k) +$$

$$y' \sin\theta_i \prod_{j=i+1}^{m} \cos\theta_j]^2 - [-\sum_{i=1}^{m} (x_i' \sin\theta_i \prod_{j=1}^{i-1} \cos\theta_j) + y' \prod_{i=1}^{m} \cos\theta_i]^2\} \quad (19)$$

where

$x_i' = \frac{x_i}{\sqrt{2} h_{x_i}}$, $y' = \frac{y}{\sqrt{2} h_y}$, $\cos\theta_i = \frac{1}{\sqrt{k_i^2 + 1}}, \sin\theta_i = \frac{k_i}{\sqrt{k_i^2 + 1}}$, λ_i is the flex coefficient, k_i the rotation coefficient. Thus the multi-dimensional ($m+1$–D) asymmetric information diffusing function was obtained.

If there is a relationship between the input item x_i' and x_j' $(i \neq j)$, the formula above should be transformed as follows:

$$\begin{cases} x_i' = x_i' \cos\theta_{ij} + x_j' \sin\theta_{ij} \\ x_j' = -x_i' \sin\theta_{ij} + x_j' \cos\theta_{ij} \end{cases} . \quad (20)$$

4 ALGORITHM EXPERIMENT

To examine the practical effectiveness and reliability of the information diffusion interpolation algorithm, the sea surface temperature (SST) reanalysis data provided by the U.S. National Centers for Environmental Prediction (NCEP) / National Center for Atmospheric Research (NCAR) is used to make an interpolation test and comparative analysis between different algorithms.

The year of the reanalysis data is 2009 and the temporal resolution is monthly. Data coverage is the marine area from 100 to 250°E and from 0 to 60°N; the spatial resolution is 2°×2°, and there are a total of 75 (zonal)×30 (meridional)=2,250 grids and nearly 2,000 data grids are left after deducting the land.

Experiment One:

Data: 23 samples (about 1%) are randomly extracted from about 2,000 data points as “observed” data (the rest is treated as missing data); interpolation experiments and comparative analysis are conducted using the Kriging interpolation, the normal interpolation model and the asymmetric "ellipse" diffusion model proposed in this work, respectively.

Objective: To examine the accuracy and reliability of different methods using the 1% “observation” points to interpolate the original sea surface temperature data.

Figure 2 shows the results of the correlation coefficient and the mean square error being compared between the interpolated SST field and the actual field; the information diffusion interpolation is more accurate than the Kriging interpolation method, especially the asymmetric "ellipse" diffusion model that has the maximum correlation coefficient and minimal mean square error, meaning that the effect of the asymmetric "ellipse" diffusion model is the best.

Experiment Two:

Data: The area and time are the same as that in Experiment One but with the interpolated sample points increased by 10% by extracting 230 samples (10%) randomly from all data points as the “observed” data (the rest is treated as missing data), conducting interpolation experiments and comparative analysis using the Kriging model, the proposed normal interpolation model and the asymmetric “ellipse” diffusion model respectively, and examining the accuracy and reliability of different methods using the 10% “observation” points to interpolate the original data SST.

Figure 3 shows the results of the correlation coefficient and the mean square error being compared between the interpolated SST field and the actual field; the effectiveness of different interpolation methods is close and the difference is small (the maximum difference of the correlation coefficient is 0.002 and the mean square error is 0.2, about an order of magnitude smaller compared with that of Experiment One). In contrast, the Kriging interpolation method is more effective than the information diffusion interpolation, meaning that the Kriging interpolation, as a classical and mature interpolation algorithm, is reliable and effective.

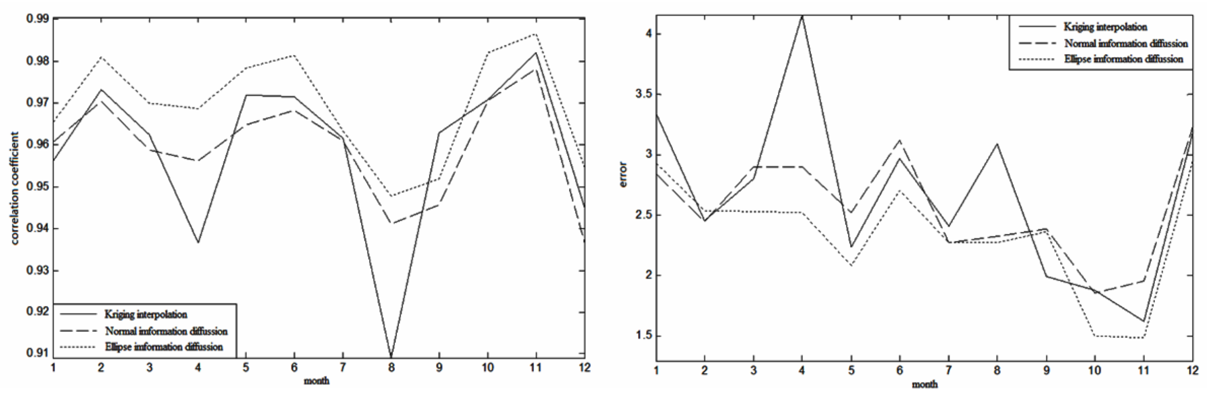

Figure 2. Comparison of different interpolation algorithms (1% sample): Correlation coefficient (left panel); mean square error (right panel).

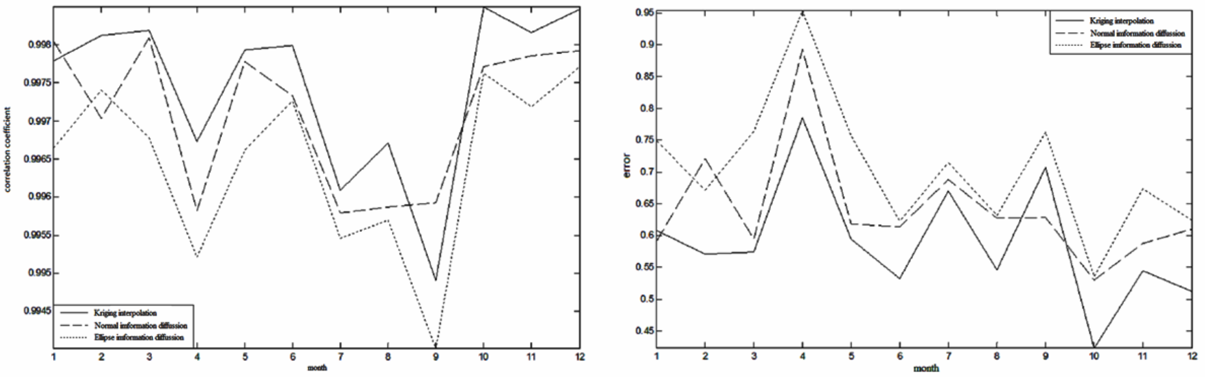

Figure 3. Comparison of different interpolation algorithms (10% sample): Correlation coefficient (left panel); mean square error (right panel).

Comparisons of the results of Experiment One and Two are shown as follows. The interpolation methods based on the small-sample information diffusion proposed in this paper (particularly the asymmetric ‘ellipse’ diffusion model) have shown obvious advantages in dealing with the interpolation of very sparse samples. With the increase of sample data, their advantages gradually diminish (some degree of superiority still exists in tests with 2% to 5% of the samples, figure not shown). In the interpolation experiments above, 5% can be considered as a critical point. For the sparse data field that is less than this criterion, the information diffusion interpolation model is more preferential than other techniques. For other types of sparse data set, the criteria should be a little different and evaluated according to actual condition, but their conjunct principle and core are suitable for analyzing and processing the interpolation calculation of imperfect information situations such as sparse data.

Algorithm application:

The information diffusion interpolation is applied to the standardized processing of the actual field of ARGO scattered elements. As ARGO floats observe once every 10 days on average and much less ARGO floats data can be received within a day, ARGO observational data belong to the data field distributed sparsely, which means that it is suitable to interpolate by the information diffusion interpolation method.

Data: ARGO floats at 10-m depth on January 1, 2009; range: 100 to 250°E, 0 to 60°N.

Figure 4 shows the ARGO buoy data (The small dots in the figure are for the buoy location, about 80 of them) and the 1°×1° grid of the temperature field respectively using the symmetrical information diffusion and the asymmetric ‘ellipse’ model interpolation. There are more than 4,000 data grid points in the region (excluding the land), so the result is equivalent to interpolation of 2% of ARGO sparse samples. In contrast, the ellipse model has a richer description of the details of the results.

As there is no operational reanalysis grid data set of ARGO for contrast tests, the corresponding correlation analysis and error analysis were not done.

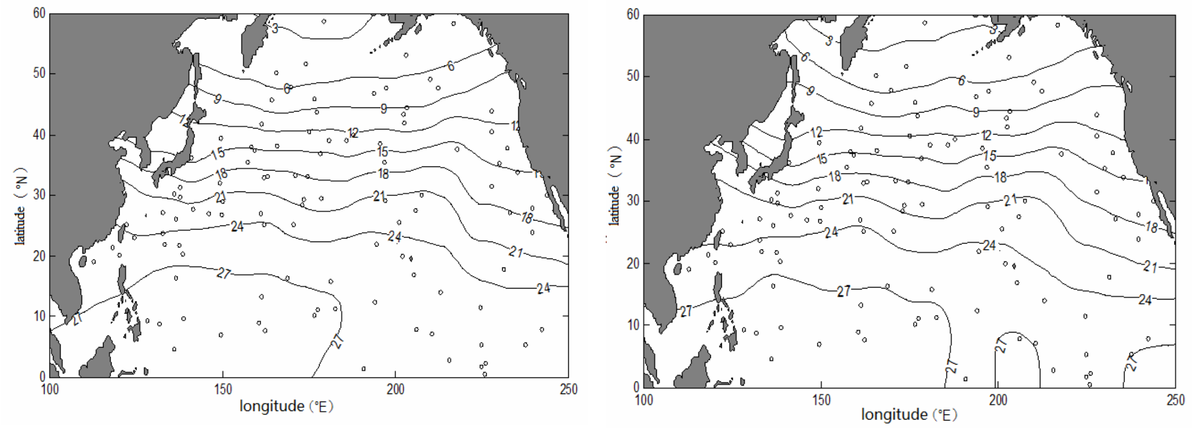

Figure 4. Grid point interpolation field of ARGO observational data of sea surface temperature: Symmetrical information diffusion (left panel); asymmetric ellipse model interpolation model (right panel).

5 SUMMARY

Addressing the difficulties of the scattered and sparse observational data in marine science, a new interpolation technique based on the information diffusion idea is proposed in this paper. Based on the idea of fuzzy sets and by constructing the spread function approximate to the goal data's distribution structure, the limited data samples are diffused and mapped into the corresponding fuzzy sets in the form of probability in the information diffusion interpolation model. To avoid the shortcomings of asymmetrical data structure in the normal diffusion function, a type of asymmetric information diffusion function is developed and a corresponding asymmetric information diffusion algorithm-ellipse model is established. Contrastive experiment analysis shows that the chief advantages of the information diffusion interpolation technique outdo other methods and that it is suitable for dealing with the sparse observational data (usually with less than 5% samples). It provides reference to the analysis and processing of small-sized samples and sparse data that actually exist in the natural sciences.

Acknowledgement: The authors would like to extend their gratitude to the Chinese center of ARGO Data for providing the research data.

[1] XU Jian-ping. ARGO for Global Ocean Observation [M]. Beijing: Ocean Press, 2002: 16-18.

[2] SU Ji-lan. Understand Correctly to ARGO Plan [J]. Ocean Technol., 2001, 20(2): 1-2.

[3] WU Hong-bao, WU Lei. Methods for Diagnosing and Forecasting Climate Variability [M]. Beijing: Meteorological Press, 2005: 186-188.

[4] FENG Guo-lin, DONG Wen-jie. Nonlinear Space-Time Distribution Theory and Methods of Observational Data [M]. Beijing: China Meteorological Press, 2006: 115-116.

[5] ZHANG Ren, WAN Qi-lin, LIANG Jian-yin, et al. Scattered data optimization and imperfect information recovery in geoscience [J]. J. Data Acquis. Process., 2006, 21(2): 209-216.

[6] HUANG Chong-fu. Risk Assessment of Natural Disaster Theory and Practice [M]. Beijing: Science Press, 2006: 63-66.

[7] HUANG C F. Information Diffusion Technique and Small Sample Problem [J]. Inform. Technol. Decision Making, 2002, 1(2): 229-249.

[8] ZHANG Ji-quan, LI Ning. Quantitative Methods and Applications of Risk Assessment and Management on Main Meteorological Disasters [M]. Beijing: Beijing Normal University Press, 2008: 213-215.

[9] ZHANG Ren, XU Zhi-Sheng, HONG Mei. Decision-making of marine environment risk for combined operations based on imperfect data samples [J]. Milit. Operat. Res. Sys. Eng., 2009, 23(1): 48-52.

[10] ZHANG Ren, XU Zh-sheng, HUANG Zhi-song. Information diffusion and its application to evaluating the influence of atmospheric oceanic environment on shipborne missile [J]. Fire Contr. Comm. Contr., 2010, 35(2): 41-44.

REFERENCES:

Citation: ZHANG Ren, HUANG Zhi-song, LI Jia-xun et al. Interpolation technique for sparse data based on information diffusion principle—ellipse model. *J. Trop. Meteor.*, 2013, 19(1): 59-66.