A Fuzzy-Based Four-Dimensional Data Assimilation Algorithm for Comprehensive Ocean Observation Information System

Hua Han[1] Fengming Liu[1] Yongsheng Ding[1, 2]

[1]College of Information Sciences and Technology, Donghua University, Shanghai 201620, P. R. China
[2]Engineering Research Center of Digitized Textile & Fashion Technology, Ministry of Education, Donghua University, Shanghai 201620, P. R. China

Abstract

Comprehensive Ocean Observation Information Systems (COOIS) comprises a massive and complex resource data, which is vital to an increasingly important application of Integrated Intelligent Ocean Information Management System (IIOIMS). IIOIMS is just complex, but uses a holistic approach to deal with the sophistication. The application domain and its resource require a tool of matching characteristics, which is facilitated by the current wide availability of high performance computing. In this paper, we first build the COOIS network and experimental platform to aggregate the comprehensive ocean observation information. Then, we present a fuzzy-based data assimilation framework and its algorithm for doing with the sophistication of COOIS. Data assimilation is widely applied for those different type data, but we firstly propose to solve this problem with fuzzy set. Theoretically, this will efficiently reduce the complexity and enhance the accuracy and utility of COOIS. Finally, we provide a demonstrated example to demonstrate the efficiency of the fuzzy-based data assimilation algorithm.

Keywords: COOIS, Fuzzy Set, Data Assimilation, Intelligent System

1. Introduction

Integrated Intelligent Ocean Information Management System (IIOIMS) emphasizes both on the complexity of physical oceanography, but also on the recognized user for monitoring and forecasting information [1]. Oceanographic services have already been established in many ocean facing countries, and they do provide forecast for winds, waves, surges, tides, ice coverage etc., as such with value added services concerning adjacent sectors such as oil spill prediction, eutrophication, coastal erosion, etc. [2].

The development of comprehensive ocean observation information systems (COOIS) may be viewed a platform for oceanographic services, which get a vast quantity and diversity of ocean information from various observation equipments, instruments, and system, such as Buoy system, earthquake observation system, weather observation system, and satellite.

Databases concerned with the natural environment may be grouped into several broad categories, and special attention has also to be paid to the relationships of these databases with land databases at the coast. Oceanographic databases concerned with the sea itself: Surface and water column, are also extensive historically and similarly complemented by modern remote sensing data [3]. Tidal observations in particular are of long standing. Systematic large-scale data sets are based on scientific surveys, and deal with temperature, salinity and other variables. Again the purposes of data collection have been a combination of prediction and pure science considerations [4]. The amount of information is prodigious, much of which originates in ports at local level from whence national statistics are derived, while scientific data are collected by the IIOIMS. The vast quantity and diversity of ocean information can appear bewildering at first sight. Making sense of it in management terms involves specifying how it is used [5].

Ocean information system development on the Internet provides a large number of benefits, such as increased information provision and accessibility, enhanced communications and networking within the community [6]. Current developments fall into two categories. The first one concerns metadatabases, where the systems available provide the pointers to databases. The second one serves as a platform for the community to exchange ideas and information.

However, in order to provide oceanographic services and value added services, the IIOIMS firstly deal with the various data in the databases. One of the development technologies is data assimilation. Many assimilation techniques have been developed for oceanography. They are different in their numerical cost, optimality, and suitability for real-time data assimilation. Direct insertion method or re-initialization method are two simple methods for data assimilation [7]. Crossman analysis and related methods is a common method assumed to be univariate and represented as grid-point values [8]. Other methods include Simulated Annealing [9], Genetic Algorithm, and Hybrid algorithm, etc.. The recent trend in data assimilation is to combine the advantages of 4D-Var and the Kalman filter

techniques[10].

In order to assimilate those data, the data assimilate system should firstly fuzzifiy their verges for efficient utilization. Data unified facilitate system supports much more services concerning adjacent sectors. So, inspired from fuzzy set theory [11], in this paper, we build the COOIS network and experimental platform to aggregate the comprehensive ocean observation information. Then we present a fuzzy-based data assimilation framework and its algorithm. In order to demonstrate its efficiency, we provide a demonstrated example. Finally, we conclude the paper and point out the future work of COOIS.

2. The COOIS Network and Experimental Platform

In order to appreciate the scope of these database developments, it is worth considering several examples databases: Ocean and coastal information systems on the Internet. A distributed information exchange infrastructure GENIE (Global Environmental Network for Information Exchange) is being developed for data sets on global environmental change, including meteorological records, sea level changes, population growth and migration, hydrological records, agricultural statistics, and satellite imagery. The project aims to simplify the search for finding what data are available in these fields, and to some extent facilitate the interchange of information between the researchers using WWW. One fundamental technique of this system is data assimilation.

So, in order to aggregate the comprehensive ocean observation information, we build the COOIS network and experimental platform as shown in Fig. 1. This network comprises with buoy observation system, earthquake observation system, weather observation system, etc. communicating with data center through the satellite. Each observation system includes lots of subsystems, equipments and instruments. All of observation data will be sent to data center for further processing. One of the processing steps is fuzzy-based data assimilation as shown in Fig. 2.

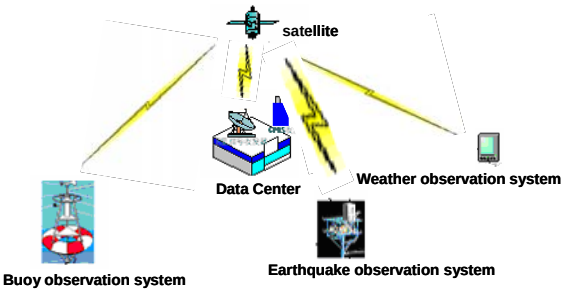

Fig. 1: The COOIS experimental platform.

3. Fuzzy-Based Four-Dimensional Data Assimilation

We present a fuzzy-based data assimilation framework as shown in Fig. 2. Fuzzy set is used for assimilation while processing greatly different information.

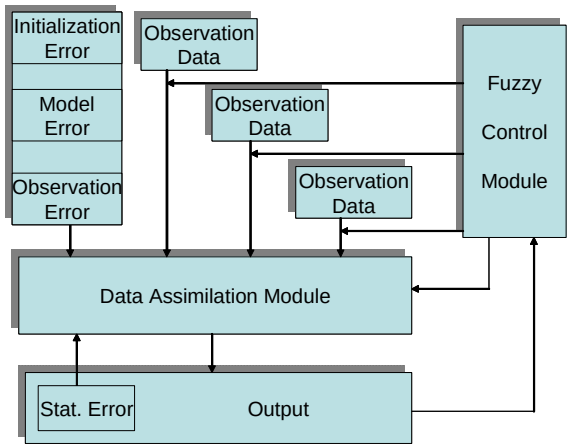

Fig. 2: Fuzzy-based data assimilation framework.

3.1. Data assimilation

Data assimilation is an analysis technique in which the observed information is accumulated into the model state by taking advantage of consistency constraints with laws of time evolution and physical properties [12]. There are two basic approaches to data assimilation: Sequential assimilation and non-sequential. Sequential assimilation only considers observation made in the past until the time of analysis, which is the case of real-time assimilation systems. Non-sequential, or retrospective assimilation can use observation from the future, for instance in a reanalysis exercise. Another distinction can make between methods that are intermittent or continuous in time [13]. In an intermittent method, observations can be processed in small batches, which is usually technically convenient. In a continuous method, observation batches over longer periods are considered, and the correction to the analyzed state is smooth in time, which is physically more realistic.

In a real-time assimilation system, 4D-Var over a short time interval is a very efficient analysis method [14, 15]. A Hessian estimation method can provide a good estimation of the analysis error covariance matrix. A simplified forecast step based on the extended Kalman filter is then used to estimate the forecast error covariance at the time of the next analysis, which must be combined with an empirical, more static model of the background error covariance. It is hoped that a good compromise between these algorithms can be achieved. There can be some constructive interactions with the problems of ensemble prediction, and specific studies of analysis

quality like sensitivity studies and observation targeting. These new methods provide many by-products which still remain to be used as diagnostic tools for improving the assimilation and forecast system.

The 4D-Var assimilation method can be described in the following way. First, a cost function is conceived to measure the error between the forecast and the observations. Then the adjoint equations, which are used to evaluate the gradient of this cost function, are obtained by applying a variation procedure to the Lagrangian problem. The cost function and its gradient are minimized by a minimization algorithm, such as the steepest descent, in order to find the optimal initial conditions that will give the optimal forecast.

Compared to a 3D analysis algorithm in a sequential assimilation system, 4D-Var has the following characteristics:

1) It works only under the assumption that the model is perfect. Problems can be expected if model error is large.

2) It requires the implementation of the rather special operators, the so-called adjoins model. This can be a lot of work if the forecast model is complex.

3) In a real-time system, it requires the assimilation to wait for the observations over the whole 4D-Var time interval to be available before the analysis procedure can begin, whereas sequential systems can process observations shortly after they are available.

4) It used as the initial state for a forecast, then by construction of 4D-Var one is sure that the forecast will be completely consistent with the model equations and the 4D distribution of observations until the end of the 4D-Var time interval (the cutoff time). This makes intermittent 4D-Var a very suitable system for numerical forecasting.

5) 4D-Var is an optimal assimilation algorithm over its time period. It means that it uses the observations as well as possible, even if is not perfect, to provide in a much less expensive way than the equivalent Kalman Filter.

Over a given time interval, under the assumption that the model is perfect, with the same input data (initial background and its covariance, distribution of observations and their covariance), the 4D-Var analysis at the end of the time interval is equal to the Kalman filter analysis at the same time. A special property of the 4D-Var analysis in the middle of the time interval is that it uses all the observations simultaneously, not just the ones before the analysis time. It is said that 4D-Var is a smoothing algorithm.

3.2. Fuzzy-based data assimilation

In oceanography, there are many linguistic fuzzy words, some of which are warm, cloudy, foggy, dense, high, low, dry, wet, small, etc. For instance, any statement about the weather temperature includes uncertainty in the forms of vagueness or ambiguity. If the temperature at a place changes between almost T_0 and T_1 ^{0}C, then this domain of change should have linguistically some subsets by considering everyday conservation. In general, the temperature is either cold or cool or warm or hot. Hence, there are four subsets of the temperature universal set at a location. Within the whole universal set, it is not possible to define the delimitation of these linguistic words with certainty.

Accordingly constructed triangles represent the approximate properties of cold, cool, warm and hot fuzzy subsets. Any meteorological factor can be subdivided into fuzzy sets that interfere with each other. However, a subjective point in delimiting the fuzzy subsets can be avoided by employing actual data and/or expert opinions as will be explained in the application section of this paper.

In any diagnostic or prognostic study in oceanography for the application of fuzzy-based data assimilation, there are four interdependent steps. A successful execution of these steps leads to the solution of the problem in a fuzzy environment, i.e., the solution procedure digests any type of uncertainty in the basic evolution of the event concerned. The algorithm is as shown in Fig. 3.

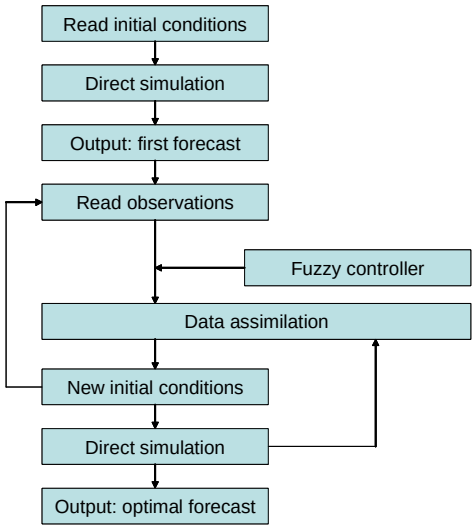

Fig. 3: Algorithm for fuzzy-based data assimilation.

In Fig. 3, the fuzzy controller includes the following components.

1) **Fuzzification:** All data are considered as having ambiguous characteristics and therefore their domain of change are divided into many fuzzy subsets which are complete, normal and consistent with each other. Hence the domain of change is fuzzified. This stem is applied to each oceanography factor considered in the solution of the problem.

2) **Inference:** This step, in fact, relates systematically pair wise all the data that take place in the solution depending on the purpose of the problem. This part includes many fuzzy conditional statements to describe a certain situation. For instance, if two

type data *X* and *Y* are interactive then they are dependent on each other. Conditional statements express the dependence as the following verbally without any equation as used in the classical approaches,

R_i: If X^i is $A^i(n)$ Then Y^i is $B^i(n)$ (1)

where $A^i(n)$ and $B^i(n)$ are the linguistic description of *X* and *Y*, respectively, and they are fuzzy subsets of *X* and *Y* that cover the whole domain of change of *X* and *Y*. The fuzzy conditional statements in Eq.(1) can be formalized in the form of the fuzzy relation $R(X, Y)$ =Also (R_1; R_2; R_3; ⋯; R_N) where Also represents a sentence connective which combines R_i's into the fuzzy relation $R(X, Y)$, and R_i denotes the fuzzy relation between *X* and *Y* determined by the *i*-th fuzzy conditional statement. After having established the fuzzy relationship $R(X, Y)$, the compositional rule of inference is applied to infer the fuzzy subset *B* for *Y*, given a fuzzy subset *A* for *X* as $B=A$ o $R(X,Y)$, where "o" is a compositional operator.

3) Defuzzification: The result from the previous step is in the form of fuzzy statement and in order to calculate the deterministic value of a linguistic variable *Y* the defuzzification method must be applied as

$$y = \frac{\sum_i^L y_i}{L} \tag{2}$$

or center-average method according to using fuzzy basis expansion, expressed as

$$P(x) = \frac{\prod_{i=1}^M \mu A_{ij}(x_i)}{\sum_{j=1}^L \prod_{i=1}^M \mu A_{ij}(x_i)} \tag{3}$$

$$y = f(x) = \sum_{j=1}^L p_j(x) y_j \tag{4}$$

where $P(x)$ is a fuzzy basis function and *y* is particular value of the linguistic variable *Y*. y_j is the support value in which the membership function reaches its maximum grade of membership, and finally *L* is number of rules and *M* the number of inputs.

Fuzzy controller by means of the first order structural dependence along a given time series provides simple prediction process provided that the fuzzy subsets of the variability domain are divided into meaningful fuzzy intervals. Once the fuzzy data subsets are provided, it is then possible to train sequentially the data time series so as to find the steady state percentages, i.e., probabilities in the transitional matrix. These are final fuzzy associative matrix elements. After training period the following fuzzy rule base is obtained for the maximum data prediction.

This fuzzy rule-base is the main tool in prediction the future likely maximum data values. It is obvious that the periodic pattern in the daily data sequences is modeled successfully with the proposed fuzzy logic model, because the relative error appears less than 10 percent. In the non-training part, the actual values and predicted ones do not fall on each other and consequently the error amount is $X - \hat{X}$. However, in order to asses the validity of fuzzy prediction, it is necessary to have an overall measure of the individual errors in the form of average performance error (*APE*) defined as follows

$$APE = \frac{\sum_i^n \left| x_i - \hat{x}_i \right|}{\sum_i^n |x_i|} \times 100 \tag{5}$$

Data assimilation: 4D is a simple generalization of 3D for observations that are distributed in time. The equations are the same, provided the observation operators are generalized to include a forecast model that will allow a comparison between the model state and the observations at the appropriate time.

Over a given time interval, the analysis being at the initial time, and the observations being distributed among *n* times in the interval, we denote by the subscript i the quantities at any given observation time i. Hence, y_i, x_i and x_{ti} are the observations, the model and the true states at time i, and R_i is the error covariance matrix for the observation errors $y_i - H_i(x_{ti})$. The observation operator H_i at time i is linearized as H_i. The background error covariance matrix *B* is only defined at initial time, the time of the background x_b and of the analysis x_a.

So, in its general form, it is defined as the minimization of the following cost function:

$$J(x) = (x - x_b)^T B^{-1} (x - x_b) + \sum_{i=0}^n (y_i - H_i[x_i])^T R_i^{-1} (y_i - H_i[x_i]) \tag{6}$$

Which can be proven, like in the 3D case detailed previously, to be equivalent to finding the maximum likelihood estimation of the analysis subject to the hypothesis of Gaussian errors?

The 4D assimilation problem is by convention defined as the above minimization problem subject to the strong constraint that the sequence of model states x_i must be a solution of the model equations:

$$\forall i \quad x_i = M_{0 \to i}(x)$$

Where $M_{0 \to i}$ is a predefined model forecast operator from the initial time to i. The 4D assimilation problem is thus a nonlinear constrained optimization problem which is very difficult to solve in the general case.

4. A Demonstration Example

Fuzzy controller by means of the first order structural dependence along a given time series provides simple prediction process provided that the fuzzy subsets of the variability domain are divided into meaningful fuzzy intervals. The application of the

methodology proposed in the previous sections is presented for daily temperature records. In this study, only the daily maximum temperature records of the most recent two years duration are used. First of all, the maximum temperature domain is divided into 7 triangular subsets that are normal, consistent and complementary. Here, normality implies that fuzzy subset has membership value equal to 1 at least for one of the members. They are complementary in the sense that at any temperature value there are distinctive fuzzy temperature subsets and their membership degrees summation at a given temperature is equal to 1. On the other hand, these 7 fuzzy subsets, namely A_i (i=1,2,. . . ,7) and they are treated equivalently for the input and output maximum temperature values. Herein, the input is the maximum temperature of any day and the output the maximum temperature for the following day.

Once the fuzzy temperature subsets are provided, it is then possible to train sequentially the temperature time series so as to find the steady state percentages, i.e., probabilities in the transitional matrix. The first 365 daily temperature values are employed for determining the transition matrix elements from the fuzzy subsets in Table 1. These are final fuzzy associative matrix elements.

Table 1 shows an example of the table.

	*A*1	*A*2	*A*3	*A*4	*A*5	*A*6	*A*7
*A*1	0.3	0.5	0.2	0	0	0	0
*A*2	0.05	0.29	0.43	0.21	0.01	0	0
*A*3	0.01	0.13	0.36	0.35	0.12	0.02	0.01
*A*4	0	0.03	0.23	0.37	0.24	0.06	0
*A*5	0	0	0.04	0.20	0.26	0.22	0.01
*A*6	0	0	0	0.03	0.24	0.01	0
*A*7	0	0	0	0	0.36	0.5	0.14

Table 1: Relative transition matrix

After training period the following fuzzy rule base is obtained for the maximum temperature prediction. The fuzzy rule-base is the main tool in prediction of the future likely maximum temperature values. It is obvious that the periodic pattern in the daily temperature sequences is modeled successfully with the proposed fuzzy logic model, because the relative error appears less than 10 percent. In the non-training part, the actual values and predicted ones do not fall on each other and consequently the error amount is $X - \hat{X}$. However, in order to asses the validity of fuzzy prediction it is necessary to have an overall measure of the individual errors in the form of average performance error (APE) defined as follows. For the daily temperature series calculated in this study APE=7.12%. This is less than practically acceptable limit of 10%.

That is, those observation temperature data are fuzzified for assimilating with other data. The processing process of the other data is very similar. These fuzzified data are assimilated together to provide oceanographic services and value added services.

5. Conclusions and Future Work

The ocean information management systems are characterized by three major themes: the information database, provision of access, and networks. Databases are useless without their applications. In considering applications, pure science may be thought of as important, even though the resources for the provision for most information have been provided to meet the requirements related to specific ocean industries. Access and overall use will benefit from continued development of networks-based integrated information management. However, all of uses depend on all the observation information disposed. Data assimilation is the important step of processes of disposal. In the future, we will further study on the method of data assimilation. Meanwhile, in order to efficiently utilize the data and obtain accuracy information products, we will adopt intelligent algorithms such as genetic algorithm and immune algorithm to research the information management system.

Acknowledgment

This work was supported in part by the Key Project of the National Nature Science Foundation of China (No. 60534020), Program for New Century Excellent Talents in University from Ministry of Education of China (No. NCET-04-415), the Cultivation Fund of the Key Scientific and Technical Innovation Project from Ministry of Education of China (No. 706024), and International Science Cooperation Foundation of Shanghai (No. 061307041).

References

[1] A. D. Terwisscha van Scheltinga, and H. A. Dijkstra, Nonlinear data-assimilation using implicit models, *Nonlinear Processes in Geophysics*, 12: 515-525, 2005.

[2] J. C. Derber and F. Bouttier, A Reformulation of the Background Error Covariance in the ECMWF Global Data Assimilation System, *Tellus*, Vol. 51A, pp: 195-221, 1999.

[3] G. J. Han, J. Zhu, and G. Q. Zhou, Salinity Estimation Using the T-S Relation in the Context of Variational Data Assimilation, *J. Geophys. Res.*, pp: 173-186, 2003.

[4] Masao Nagasaki, Rui Yamaguchi, Ryo Yoshida,Seiya Imoto,Atsushi Doi,Yoshinori Tamada,Hiroshi Matsuno, Satoru Miyano, and Tomoyuki Higuchi, Genomic Data Assimilation for Estimating Hybrid Functional Petri Net from Time-Course Gene Expression Data, *Genome Informatics*, 17: 46-61, 2006.

[5] J. Zhu and M. Kamachi, An Adaptive Variational Method for Data Assimilation with Imperfect Models, *Tellus*, 52: 265-279, 2000.
[6] J. Zhu, W. Hui, and G. Zhou, SST Data Assimilation Experiments Using an Adaptive Variational Method, *Chinese Science Bulletin*, 47(23): 2010-2013, 2002.
[7] R. F. Li, X. B. You, and W. Prunchan, A Three-Dimensional Coastal Ocean Circulation Model, *Bangkok, Thailand*, pp: 1-16, 2002.
[8] J. A. Carton, G. Chepurin, X. Cao, and B. S. Giese, A Simple Ocean Data Assimilation Analysis of the Global upper Ocean 1950-1995, Part 1: Methodology, *J. Phys. Oceanogr*, 30: 294-309, 2000.
[9] Emanuele Di Lorenzo, Andrew M. Moore, Hernan G. Arango, Bruce D. Cornuelle, Arthur J. Miller, Brian Powell, Boon S. Chua, and Andrew F. Bennett, Weak and strong constraint data assimilation in the inverse Regional Ocean Modeling System (ROMS): Development and application for a baroclinic coastal upwelling system, *Ocean Modelling*, 16(1):160-187, 2007.
[10] A. Ridgwell, J. C. Hargreaves, N. R. Edwards, J. D. Annan, T. M. Lenton, R. Marsh, A. Yool, and A. Watson, Marine geochemical data assimilation in an efficient Earth System Model of global biogeochemical cycling, *Biogeosciences*, 4: 87-104, 2007.
[11] Zhilan Xiong, Yanling Hao, Jinchen Wei and Lijuan Li, Fuzzy Adaptive Kalman Filter for Marine INS/GPS Navigation, *Proceedings of the IEEE International Conference on Mechatronics & Automation Niagara Falls*, Canada, July 2005.
[12] W. Fu, G. Zhou, and H. Wang, Estimating Background Error Covariance from Model Outputs for Oceanic Data Assimilation, *Advances in Atmospheric Science*, 2003.
[13] M. Kamachi, T. Kuragano, N. Yoshioka, J. Zhu, and F. Uboldi, Ocean Data Assimilation of Satellite Altimetry and Predictability in the Western North Pacific, *Advances in Atmospheric Science*, 18(9): 767-786, 2001.
[14] Z. D. Luo, J. Zhu, and Y. J. Wu, A Decouple Conjugate Gradient-Gauss-Newton's Iterative Scheme for Assimilating Altimetry Data Problems, *Science in China*, 2003.
[15] X. You, L. Rongfeng, Z. Zhou, J. Zhu, and Q. Zeng, Sea Temperature Variational Data Assimilation in the China Sea and its Adjacent Areas, *China Science Bulltine*, 2004.