IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 4, 2018, accepted August 9, 2018, date of publication August 21, 2018, date of current version September 7, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2866408

An Approach for Predicting Uncertain Spatiotemporal XML Data Integrated With Grey Dynamic Model

CHENGJIA SUN, LUYI BAI, LEI KANG, SHANHAO LI, AND NAN LI
Qinhuangdao Branch Campus, Northeastern University, Qinhuangdao 066004, China

Corresponding author: Luyi Bai (baily@neuq.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61402087, in part by the Natural Science Foundation of Hebei Province under Grant F2015501049, in part by the Fundamental Research Funds for the Central Universities under Grant N172304026, and in part by the Scientific Research Fund of the Hebei Education Department under Grant QN2014339.

ABSTRACT Query and prediction have been proved to be one of the most important operations for uncertain spatiotemporal data and deserve further study. In this paper, we propose an approach to predict uncertain spatiotemporal data, which is intended to integrate the grey dynamic model into the extensible markup language (XML). Our approach is unique in the predicting element nodes which are integrated into the position element node in uncertain spatiotemporal XML data tree, and at the same time, the other element nodes do not need to make any changes. In addition, we applied our method to a meteorological application and established a series of experimental models for testing. The experimental results show that our method is accurate and useful. The model of prediction with grey model based on XML (PGX), which is applied to uncertain spatiotemporal objects, is able to achieve the minimum mean accuracy of 0.5% in a short time. The experimental results show that PGX can effectively improve the efficiency of information storage and retrieval. The experimental prediction accuracy is guaranteed (the relative error is between 0.5% and 5%) and the query time based on XML is 89.2% shorter than that of SQL Server.

INDEX TERMS Uncertain spatiotemporal data, prediction, grey dynamic model.

I. INTRODUCTION

Spatiotemporal data are featured by spatial and temporal phenomena [1]. Previous efforts started with separate research in temporal and spatial database area. Along with the development of Mobile Devices and the advance of geographic information systems, a great deal of geoscientific data has been generated. These data lie in the continuous space and change with time and space, so they have a high degree of particularity and complexity [2]. Hence, spatiotemporal data management is becoming more and more important, especially spatiotemporal data prediction.

The spatial and temporal uncertainty exists widely in the complex and variational real world. For instance, the incompleteness of knowledge, the fuzziness of concepts and the derivation of data makes the entity possess the feature of uncertainty. Therefore, uncertainty is absolute while certainty is relative.

Recently, studies on predicting uncertain spatiotemporal objects have made great progress [3]. Jeung *et al.* [4] presented an approach which forecasted an object's future locations in a hybrid manner by using motion function and movement patterns. Cheng and Wang [5] used a dynamic recurrent neural network for spatial prediction, which is applied to the forest fire. The trajectory of moving target, which is limited by traffic network, can be predicted accurately [6]. The work of Boulila *et al.* [7] takes imperfection into account, relating to the spatiotemporal mining process in satellite imaging. The particularity and complexity of the spatiotemporal data are enhanced. Because spatiotemporal data have been applied increasingly and widely in the field of transportation, meteorology, earthquake rescue, criminal analysis, web applications, public health and medical services [8], [9]. Therefore, it is urgent to improve the predicted technology of uncertain spatiotemporal data.

Le Coz *et al.* [10] developed a method with a statistical representation of uncertainties. McMillan *et al.* [11] provided a comprehensive review of the uncertain values for gauging and rating curves. However, in the field of XML, there are few researches on the prediction of uncertain spatiotemporal data. Additionally, compared with the certain spatiotemporal data, uncertainty enhances spatiotemporal data particularity and complexity, making it more challenging to predict uncertain spatiotemporal data accurately. Therefore, a new modeling technology of predicting uncertain spatiotemporal data is required. Researches on grey dynamic model have been investigated widely in recent years. Original data in grey dynamic model are composed of both certain and uncertain information. By obtaining part of the data, the correlation of all factors is measured, the data are preprocessed, and finally the differences and similarities of the development trend of all factors are obtained. The predicted sequence can be obtained by processing the current information. Meanwhile, the predicted results are accepted when the accuracy of the predicted meets the reliability requirement. In this case, the grey dynamic model can accurately predict the changing trend of the uncertain state of the spatiotemporal data. Therefore, the grey dynamic model is integrated into the prediction of uncertain spatiotemporal data. There is a series of efforts to improve the perception of the grey theory [12], [13]. Hamzacebi and Es [14] show the superiority of Optimized Grey Modeling (1, 1) when compared with the results in literature. The grey relational analysis of the fuzzy sets also takes an important occupation in the fuzzy measure field [15]. Furthermore, Zhang *et al.* [16] propose a grey relational projection method for the MADM problems with intuitionistic trapezoidal fuzzy number attribute. However, the grey theory is studied on combination with other techniques rather than itself [13], [17]. In addition, it has been applied to other applications such as the electricity consumption. Actually, the theoretical basis of grey dynamic model is more beneficial to predict uncertain spatiotemporal data [18].

Extensible Markup Language (XML) has been applied to handle spatiotemporal data in the past few years [19]. There is a growing interest in the research of uncertain XML data [20]. In the field of uncertain spatiotemporal database, there are also several researchers modeling fuzzy spatiotemporal data based on XML and exploring algorithms for fixing inconsistencies of it in XML documents caused by changing operations [21]. Then Li and Ma [22] introduce the object-stack algorithm that outperformed the traditional XML keyword query algorithms significantly, which could get high quality of query results with high search efficiency on the fuzzy XML document.

In the field of spatiotemporal prediction, researches have been put forward which take advantages of grey dynamic model and XML. For example, Bai *et al.* [18] investigated spatiotemporal XML data with the grey dynamic model and presented an algorithm for interpolation and prediction of spatiotemporal data based on XML. Unfortunately, the work only considers certain spatiotemporal data, and the prediction of uncertain spatiotemporal data has not been further studied in their work. They also ignored the memory performance and other relevant problems on relational database and XML.

In this paper, a model is proposed that enables the accurate prediction of the uncertain spatiotemporal data. Besides, this model can be used to predict uncertain spatiotemporal data. For instance, the typhoon experiment can be strong proof to our theory. Our approach makes the following main contributions: Firstly, we integrate the grey dynamic model into XML to predict uncertain spatiotemporal data on the basis of previous achievement. In succession, the proposed predicted model (PGX) can be applied to reducing the uncertainty generated by the volatility of original data. Subsequently, our model can predict the motion of the uncertain spatiotemporal objects in the given time precisely. For example, in the meteorological experiments, this model can predict the future location region at a relatively accurate degree according to the center information of the hurricane. Finally, we make comparisons between relational database and XML to evaluate the effectiveness of storing uncertain spatiotemporal data. The experimental results demonstrate the performance advantages of our approach.

The rest of this paper is organized as follows. Section 2 presents related work. Section 3 proposes an approach for predicting uncertain spatiotemporal data and the model definition. In Section 4, our approach is applied to meteorological applications and the experimental results are presented. Section 5 concludes the paper.

II. RELATED WORK

A. SPATIOTEMPORAL DATA PREDICTION

In the field of spatiotemporal prediction, there was an early nonparametric regression method which used multivariate nonparametric regression to process spatiotemporal data by predicting traffic flow [23]. Through development, a nonparametric regression model was developed to discuss spatiotemporal prediction under the absence of data [24]. At the same time, the Markov model, which exhibits many years' superiority in predicted performance, also appears in the field of spatiotemporal prediction. Predicting future locations with Hidden Markov Model [25], using variable Hidden Markov Model [26], and modifying Hidden Markov Model to produce adaptive parameter selection trajectory predicted methods [27] are produced gradually.

In recent years, with the hot of artificial intelligence, in the field of spatiotemporal data prediction, the researches of neural network and in deep learning also occupy a territory. Some researchers have proposed convolution LSTM to construct the training model of end-to-end precipitation in spatiotemporal data prediction from the machine learning viewpoint [28]. In the latest research, a kind of deep convolution neural network model is used to deal with traffic network speed prediction [29]. Similarly, in deep learning, some scholars have raised the predicted problem of spatiotemporal data series [30]. Some scholars considered the correlation between space and time and used stack automatic encoder model to learn general traffic flow features in the field of deep learning. They trained the model in a greedy hierarchical

TABLE 1. Study on grey model parameters.

Grey Model	Methods	Ref.	Highlights
$X^{(0)}$	Function transformation	[37]	The consistency of the new comparison criterion of the smooth degree with the prevision criterion is proved.
	Buffer operators	[38]	Under the axiomatic system of buffer operator, strengthening buffer operators whose buffer intensity can be adjusted are constructed.
$X^{(1)}$	Accumulating generation	[39]	A new NDGM with the fractional-order accumulation is put forward.
$z^{(1)}(k)$	-	[40]	The sequences of exponential distribution are predicted through optimization of background value in grey differential equation.
	Convolution integral	[41]	GMC (1, n) has a control parameter u. n unknown interpolation coefficients are input into the background values of n variables so as to improve the adaptability of GMC (1, n).
	Box plot	[42]	Determine the background values by the box-plot membership function
$X^{(0)}$ (1)	-	[43]	The new initial condition is comprised of first and last items of a sequence generated from applying the first-order accumulative generation operator on sequence of raw data.
ε=□$^{(0)}(k)$−x$^{(0)}(k)$	-	[45]	Use the error sequence to remedy the original model to improve the accuracy.
	Neural network	[44]	Replace the traditional GM (1, 1) model with the NN-GM (1, 1) model for a grey residual modification model.
Regenerating parameter	Discrete equations	[48]	Generate the new parameter through Grey-Model-based discrete operation (DGM).
	Grey Lotka–Volterra model	[46]	A linear programming method is used to estimate the parameters of the grey Lotka–Volterra model under the criterion of the minimization of MAPE.
	Fractional order accumulation	[47]	The function of new pieces of information is greater than that of old pieces of information.
	Kernel-based	[35]	The KGM (1, n) model is introduced to do with nonlinear multivariate model.

manner and proposed a deep architecture model to represent the characteristics of the traffic flow used for prediction [31].

In recent years, the study of spatiotemporal data has shown a trend of combining with other methods and other fields. Trajectory prediction is one of the hot topics in the research of spatiotemporal data prediction. In the last ten years, more and more scholars have studied the trajectory prediction of moving objects. The problems of frequent and uncertain trajectory prediction of moving objects [32] and the spatiotemporal trajectory model [33] have been explored. Besides, the spatiotemporal behavior model-based method [34] for the prediction of moving position has been studied with the help of hidden Markov model [27], which shows that predicted methods have been applied to the research of spatiotemporal data continuously.

B. OTHER RECOMMENDATIONS

Significant effectiveness of grey model in time series forecasting in small samples has been widely concerned [35]. One of the most important aspects of grey system theory is the grey predicted model which is represented by the GM (1, 1) model with one order and one variable [36]. As a lot of methods have been put forward to improve the precision of grey model, some researchers begin to study the parameters in generation of grey model, shown in Table 1. The pre-processing of raw sequence $X^{(0)}$.

- The improvement of grey accumulating generation sequence $X^{(1)}$.
- Optimization of grey model background value $z^{(1)}(k)$.
- Optimization of grey model initial value $X^{(0)}(1)$.
- The residual error correction of grey predicted model $\varepsilon = \dot{x}^{(0)}\,(k) - x^{(0)}(k)$.

In Table 1, $X^{(0)}$ is generated by function transformation [37] and buffer operators [38]. Wei and Hu [37] proposed adjustable parameters to compute the relative optimal modeling parameters automatically, and the concept of the new smooth degree of the first-class grey model was studied [38]. $X^{(1)}$, the accumulating generation methods, was improved by the non-homogenous discrete grey model (NDGM) [39]. The methods of generating $z^{(1)}(k)$ include convolution integral [41] and box plot [42]. $X^{(0)}$ (1) is comprised of the first and the last item of a defined sequence. The residual error correction ε, which is used to judge the reasonability of prediction, is one of the most essential parts of grey model. Many scholars have examined the performance of ε [43], [45]. Besides, there are also other methods to regenerate the parameters [35], [46]–[48].

As can be seen in Table 1, many researchers studied some parameters in the generation of grey models which are essential to build an accurate predicting model.

In addition, some related work on grey model have combined it with other methods, such as Markov [49], [50], fuzzy mathematics [51], [52] and neural network [44], [53], [54]. Markov chain is suitable for long-term data sequences with large random fluctuations [49], fuzzy mathematics can deal with multivariable situation, and neural network has been applied when the raw series has the big fluctuation [54]. Those are presented in the following shown in Table 2.

III. UNCERTAIN SPATIOTEMPORAL XML DATA PREDICTION

A. UNCERTAIN SPATIOTEMPORAL DATA

Uncertain spatiotemporal objects can be studied from space and time. Moreover, the uncertainty of them is independent.

TABLE 2. Grey model with other methods.

Methods	Ref	Highlights
Grey-Markov predicted model	[49]	Grey-Markov forecasting model is a combination of grey predicted model and Markov chain which show obvious optimization effects for data sequences with characteristics of non-stationary and volatility.
	[50]	Grey-Markov model is used to time series models to forecast the consumption of conventional energy in India.
Fuzzy mathematics	[51]	The predicted performance of FTS- heuristic model, two-factor model, Markov model and GM- GM (1, 1), GM-Markov, GM-Fourier is investigated. The comparison of the models is based on forecasting error of time series.
	[52]	Predicted of multivariate interval-valued time series. Evolutionary algorithm for learning Fuzzy Grey Cognitive Maps (FGCMs).
Neural network	[53]	Combining the grey forecasting model with the GARCH to improve the estimated ability, the empirical evidence shows that the new hybrid GARCH model outperforms the other approaches in the neural network option-pricing model.
	[54]	Introducing the BP neural network to amend TFGM (1, 1) and propose NNTFGM (1, 1), the amendment to TFGM (1, 1) through the BP neural network or SVM is effective after the experiments.
	[43]	The NN-GM(1, 1) model is able to directly determine the developing coefficient and control variable using a SLP without requiring the background value.

TABLE 3. Types of time and types of space.

Types of time	Types of space
uncertain time point $X_{tinstant}$	uncertain point X_{spoint}
uncertain time interval $X_{tinterval}$	uncertain line X_{sline}
	uncertain region $X_{sregion}$

As a result, temporal types of uncertain spatiotemporal data can be divided into uncertain time point and uncertain time interval while spatial types of uncertain spatiotemporal data can be divided into uncertain point, uncertain line and uncertain region, shown in Table 3.

Definition 1: Temporal types of uncertain spatiotemporal data can be divided into uncertain time point $X_{tinstant}$ and uncertain time interval $X_{tinterval}$, where

- $X_{tinstant} = \{t, p_{tt} | \forall t \in R, p_{tt} \in (0, 1)\}$.
- $X_{tinterval} = \{t_s, t_e, p_{tl} | \exists t_s, t_e \in X_{tinstant}, p_{tl} \in (0, 1)\}$.

In Definition 1, p_{tt} represents the probability of $X'_{tinstant}s$ existence, t_s and t_e represent the starting time and ending time of the event; $X_{tinterval}$ represents the interval between two time points and p_{tl} represents the existent probability of $X'_{tinterval}s$. In this paper, the time point $X_{tinstant}$ is not regarded as the moment of reality, but the smallest fraction of discretized time.

Definition 2: Spatial types of uncertain spatiotemporal data can be divided into uncertain point X_{spoint}, uncertain line X_{sline}, and uncertain region $X_{sregion}$, where

- $X_{spoint} = \{patt, x_l, y_l, x_r, y_r, x, y\}$.
- $X_{sline} = \{p_l, ubound, lbound | \exists p_l \in X_{spoint}\}$.
- $X_{sregion} = \{p_r, x_l, y_l, x_r, y_r | \exists p_r \in X_{spoint}\}$.

In Definition 2, the value of *patt* is [0, 1], indicating the possibility of the point. When $patt \in [0, 1]$, the point is within *MBR* (minimum bounding rectangle) bounded by x_l, y_l, x_r, y_r, as shown in Fig. 1(a); p_l is a set of uncertain points within a strip area, and *ubound* and *lbound* are the two boundary lines which descript where the uncertain line lies, as shown in Fig. 1(b); p_r represents the set of uncertain points within a region, and (x_l, y_l) and (x_r, y_r) are the coordinates of the bottom left corner and the top right corner of the *MBR* as shown in Fig. 1(c).

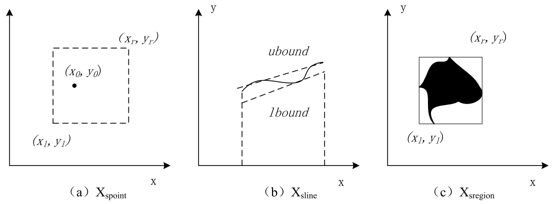


FIGURE 1. Uncertain spatial types of point, line and region.

Definition 3: Uncertain spatiotemporal type set T_u is defined as

$$\mathrm{T_u} = X_{tu} \times Y_{su} \tag{1}$$

where X_{tu} and X_{su} depict the uncertain temporal type and uncertain spatial type respectively.

According to Definition 3, uncertain spatiotemporal objects can be divided into six types as shown in Table 4 and Fig. 2, including uncertain point in uncertain time point, uncertain line in uncertain time point, uncertain region in uncertain time point, uncertain point in uncertain time interval, uncertain line in uncertain time interval, and uncertain region in uncertain time interval.

Definition 4: Types of uncertain spatiotemporal objects are U_{IP}, U_{LP}, U_{IL}, U_{LL}, U_{IR}, and U_{LR}, where

- $U_{IP} = (X_{tinstant}, X_{spoint})$ where $X_{tinstant} = t(t \in R)$ and $X_{spoint} = (patt, x_l, y_l, x_r, y_r, x, y)$.

TABLE 4. Types of uncertain spatiotemporal objects.

	Point	Line	Region
Time point	U_{IP}	U_{IL}	U_{IR}
Time interval	U_{LP}	U_{LL}	U_{LR}

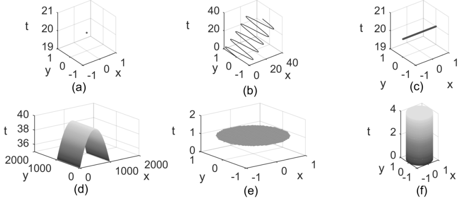


FIGURE 2. Types of uncertain spatiotemporal object.

- $U_{LP} = (X_{tinterval}, X_{spoint})$ where $X_{tinterval} = (t_s, t_e)$, $\exists t_s$, $t_e \in X_{tinstant}$ and $X_{spoint} = (patt, x_l, y_l, x_r, y_r, x, y)$.
- $U_{IL} = (X_{tinstant}, X_{sline})$ where $X_{tinstant} = t(t \in R)$ and $X_{sline} = (p_l, ubound, lbound)$, $\exists p_l \in X_{spoint}$.
- $U_{LL} = (X_{tinterval}, X_{sline})$ where $X_{tinterval} = (t_s, t_e)$, $\exists t_s$, $t_e \in X_{tinstant}$ and $X_{sline} = (p_l, ubound, lbound)$, $\exists p_l \in X_{spoint}$.
- $U_{IR} = (X_{tinstant}, X_{sregion})$ where $X_{tinstant} = t\ (t \in R)$ and $X_{sregion} = (p_r, x_l, y_l, x_r, y_r)$, $\exists p_r \in X_{spoint}$.
- $U_{LR} = (X_{tinterval}, X_{sregion})$ where $X_{tinterval} = (t_s, t_e)$, $\exists t_s$, $t_e \in X_{tinstant}$ and $X_{sregion} = (p_r, x_l, y_l, x_r, y_r)$, $\exists p_r \in X_{spoint}$.

According to the nature of spatiotemporal data, we give the definition of uncertain spatiotemporal data in the following.

Definition 5: The uncertain spatiotemporal (*USP*) data is a 4-tuple,$USP = (P, T, OID, ATTR)$, where

- P is the probability that an uncertain spatiotemporal object exists in the current situation.
- T is the time that an uncertain spatiotemporal object exists in the current situation.
- *OID* is the changing history of an uncertain spatiotemporal object.
- *ATTR* is the attributes of an uncertain spatiotemporal object.

Definition 6: The changing history of a spatiotemporal object is a 4-tuple, $OID = (type, pre, suc, T)$, where

- *type* is the type of the object's spatiotemporal change.
- *pre* is precursor.
- *suc* is successor.
- T is the changing time.

In Definition 6, *type* can represent *create*, *split*, *mergence*, *keep*, *expand*, *shrink*,*transform*, and *disappear*; *pre* and *suc* depict the changing state of a spatiotemporal object; T is the changing time, which can be time point or time interval.

Definition 7: Attribute of an uncertain spatiotemporal object is a 3-tuple, $ATTR = (general, spatial, forecast)$, where

- *general* denotes the general attribute.
- *spatial* denotes the spatial attribute.
- *forecast* denotes the forecast attribute.

As for Definition 7, firstly, *general* attribute contains (*num*, *val*), where *num* is the number of the *USP* and *val* is the value of the *USP*. Secondly, *Spatial* attribute contains *position* and *motion* which are position and movement trend of *USP* respectively. In particular, the *position* attribute represents the location of the spatiotemporal object in the form of *MBR*; *motion* represents the movement trend, having three attributes (T, *tend*, *range*). T is the time, *tend* is the direction, and *range* is the speed of the moving trend associated with a node *pro* to represent the probability of the trend, denoted by X_a and Y_a whose value can be l, r and k. l means that the spatiotemporal objects move left or down and r means that the spatiotemporal objects move right or up. k means that the spatiotemporal objects keep its position and will not move. X_g and Y_g represent the length on *x-axis* and length on *y-axis*. Lastly, *Forecast* attribute uses data from *general* attribute and *spatial* attribute to predict and store the predicted results by grey dynamic model.

From Definition 1 to Definition 7, the spatiotemporal data model can be represented as shown in Figure 3.

As we can see from Fig. 3, USP has three attributes as described from Definition 1 to Definition 7, where *Fposition* and *Fmotion* represent the forecast position and motion respectively.

B. UNCERTAIN SPATIOTEMPORAL XML DATA FOR PREDICTION

1) THE STRUCTURE OF USP BASED ON XML AND GREY MODEL

In XML data tree, every attribute is regarded as a node whose types are uncertain. With the unique feature of XML (extension and customize identification), we can develop custom models according to the characteristics of spatiotemporal data. In addition, the *forecast*, which has a special node called *patt*, has the ability to store the predicted results of other attributes. And *patt* describes the result of the prediction and has an uncertain type. This node represents the predicted results, and its value could be disjunctive (only one result from many possible ones) or conjunctive (two or more, or even the whole set of the possible results).

Based on the seven definitions above, the subsection will develop an uncertain spatiotemporal XML data model for prediction. The structure of constructing uncertain spatiotemporal XML data model integrated by grey dynamic model is shown in Fig. 4.

For a spatiotemporal object, the dynamic grey model and XML are used to study the existing form and predict its future changes. Firstly, we store uncertain spatiotemporal data in XML. Then we use grey dynamic model to process data in order to make predictions. Finally, we store the predicted results in XML database. Uncertain spatiotemporal XML data is quite flexible and can be updated based on the storage of the uncertain spatiotemporal objects. As a result, the storage of uncertain spatiotemporal objects here is an irregular multi-tree. This may inevitably cause troubles when querying

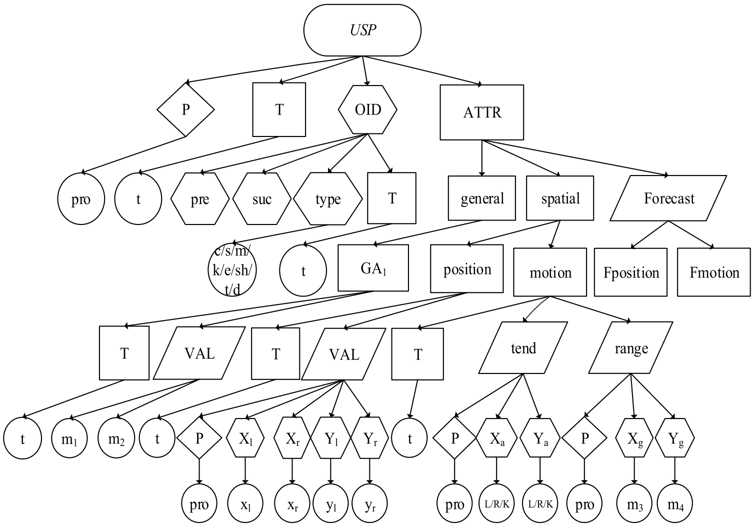


FIGURE 3. Spatiotemporal object model.

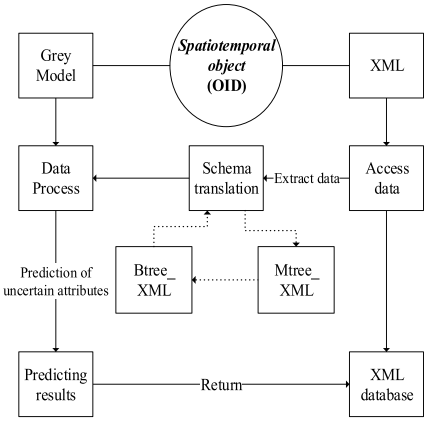
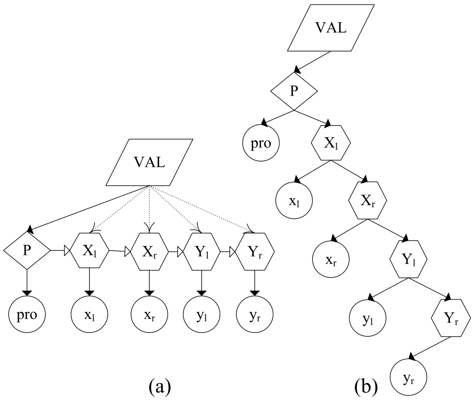


FIGURE 4. The structure of constructing uncertain spatiotemporal XML data model integrated by grey dynamic model.

data, so we encode it as a binary tree before processing it by using grey model.

2) PREPARE USP BEFORE PREDICTION

Example: In the typhoon clouds' covering area, we can use *MBR* to describe the cloud area. Our model defines two pairs of coordinate values to describe the cloud area at a certain time point, which are at the lower left corner (x_l, y_l) and at the upper right corner (x_r, y_r). In the predicting model, XML is an irregular multiway tree. Querying multiway tree is rather complicated, so we transform the original data model into a binary tree before querying. The data model is transformed into binary tree for practical use so that two pairs of the coordinates are divided into four groups. Considering the actual operation, the data model is converted into a binary tree when uncertain spatiotemporal data is processed, and two groups of coordinates of the delineated position are divided into four groups based on the time (x_l, t), (y_l, t), (x_r, t), (y_r, t), respectively. After combining digital form for prediction with the grey dynamic model, the results show the accuracy is able to meet the requirements of the next prediction. A set of data is obtained during data extracting process each time. The time is defined as t_m, where m represents any value from 1 to n. Each query corresponds to a time t_m, this time node is the value of the left sub-tree of VAL's father node. The transforming processes are: take the part of the XML storage structure which stores the coordinates of the *MBR*; keep the leftmost line; delete the rest of the right lines (for example, delete the dotted lines in the figure as shown in Fig. 5(a).); connect the sibling node from left to right; and finally converted into a binary tree. As a result, when querying for the coordinates of the position nodes (x_l, y_l, x_r, y_r), the information is within the binary tree at the root of *VAL*, the subtree of whose left child's right child has the value, as shown in Fig. 5(b). We create stack *VAL1*, *VAL2*, *VAL3* and *VAL4* to store the value (x_l, y_l, x_r, y_r) and their corresponding



(a)

(b)

FIGURE 5. (a). Transform multi-tree into binary tree. (b) Data structure of binary tree.

time information of each query. After the results of grey prediction are calculated, the predicted value of possibility is returned to the XML database to be stored. Each time the information is updated, the grey dynamic model will predict the next time point information and update the XML database in the real time.

The problem of predicting the position of uncertain spatiotemporal objects is the core issue to be solved in this paper. Based on the above definitions and descriptions, the structure of uncertain spatiotemporal XML data model integrated with grey dynamic model is shown in Fig. 6.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<Forecast>
<patt>Conjunctive</patt>
<model>Grey Model</model>
  <original data>23.1, 23.5, 23.8, 24, 24.5</original data>
  <forecast data>24.5</forecast data>
</Forecast>
```

FIGURE 6. Structure of forecast node with grey dynamic model.

In Fig. 6, *Conjunctive* indicates that the predicted type of the spatiotemporal objects is coincident. Besides, a set of *original data* is involved to get the *forecast data*.

Lemma 1: The bottom left corner (X_l, Y_l, f_x, t_1) and the top right corner (X_r, Y_r, f_x, t_1) can be operated as (X_l, f_x, t_1), (X_r, f_x, t_1) and (Y_l, f_x, t_1), (Y_r, f_x, t_1).

Proof: Because the grey model is to deal with a single sequence, we need to deal with (x, y) separately in this model. Since X_l, X_r, Y_l, Y_r are relatively independent, we process them with grey dynamic model on the time sequence when predicting the positions of uncertain spatiotemporal objects by the proposed model. Now the tuples are (X_l, f_x, t_1), (X_r, f_x, t_1), (Y_l, f_x, t_1), (Y_r, f_x, t_1) where f_x represents the function of processing uncertain spatiotemporal data using grey dynamic model.

Suppose $X_l = 0$, $Y_l \neq 0$, then (X_l, Y_l, f_x, t_1) is equal to (Y_l, f_x, t_1); suppose $X_r \neq 0$, $Y_r = 0$, then (X_r, Y_r, f_x, t_1) is equal to (X_r, f_x, t_1); suppose $X_r \neq 0$, $Y_l \neq 0$, vector $\vec{X}$ and vector $\vec{Y}$ are represented as the current node, which is actually one of the coordinates, with grey model in Fig.7, and vector $\vec{Z}$ is represented as the node of the next time point. Since

$$\vec{Z} = \vec{X} + \vec{Y} \tag{2}$$

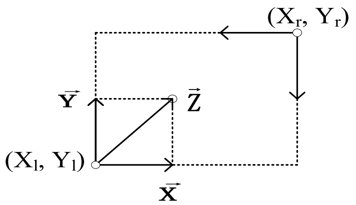


FIGURE 7. The corner of MBR.

thus

$$(\vec{Z}, f_x, t_l) = (\vec{X}_l, f_l, t_l) + (\vec{Y}, f_x, t_l) \tag{3}$$

In the grey dynamic model, we use the latest data to predict the trajectory of uncertain spatiotemporal objects in real-time, so we can get accurate results.

Definition 8: Merge n time points into a group, the time sequence from t_{i-n+1} to t_i is following:

$$\begin{array}{cccc}
(X_{li-n+1}, t_{i-n+1}) & (X_{li-n+2}, t_{i-n+2}) & \ldots & (X_{li}, t_i) \\
(Y_{li-n+1}, t_{i-n+1}) & (Y_{li-n+2}, t_{i-n+2}) & \ldots & (Y_{li}, t_i) \\
(X_{ri-n+1}, t_{i-n+1}) & (X_{ri-n+2}, t_{i-n+2}) & \ldots & (X_{ri}, t_i) \\
(Y_{ri-n+1}, t_{i-n+1}) & (Y_{ri-n+2}, t_{i-n+2}) & \ldots & (Y_{ri}, t_i)
\end{array}$$

In Definition 8, each row represents a set of predicted sequences, the subscripts represent ordinal numbers, l represents the coordinates of lower left corner, and r represents the coordinates of upper right corner. There are n pairs of points, $(X_l, Y_l,)$ and (X_r, Y_r), representing position coordinates, and each line of the data is predicted by the grey dynamic model. Suppose the current time is t_i, then the next time is t_{i+1}. Therefore the grey dynamic model chooses n data from t_{i-n+1} to t_i to predict the uncertain spatiotemporal data. The predicted results correspond to (X_l, X_r, Y_l, Y_r) as position coordinates.

Only when the structure of the prediction and the actual value match with the requirement of accuracy, will we consider if the prediction is formula valid. At the end of this round of prediction, the oldest data is discarded, namely the data at t_{i-n+1}. Then update n sets of data from t_{i-n+2} to t_{i+1} to predict again, and update the original sequence and the predicted results in the real time. If one group of time changes too shortly, the updated uncertain spatiotemporal data cannot generate the grey equation. In that case, we can use the first prediction to obtain the predicted formula. After collecting m sets of data, we can update the formula with the time length of m. In this situation, the data keep changing.

When predicting the properties of an uncertain spatiotemporal object, we make predictions based on the types of the uncertain spatiotemporal objects. When the properties of uncertain spatiotemporal objects can be described in terms of data such as predicting the annual rainfall in a region, we can develop formula to predict. In particular, according to h_i, which represents the annual rainfall of one of the past few years, the sequence equations of (h_i, f_x, t_i) can be developed to predict the rainfall h_{i+1} of the next year. In this equation, f_x represents the function of *GM(1, 1)*.In this paper, we only focus on *GM(1, 1)* used in predicting the spatial information of the uncertain spatiotemporal objects.

Because the predictions of uncertain spatiotemporal data may enhance the uncertainty, the similarity of the single result with the actual value cannot meet the requirement of prediction. Therefore, we expand the predicting range, taking the original uncertain spatiotemporal data into consideration.

Definition 9: A model called *REPM*. Define R as the predicted result of each group, $\Delta\varepsilon$ represents the mean residual error of m groups which is from (h_i, f_{xi}, t_i, to (h_{i+m-1}, f_{xi+m-1}, t_{i+m-1}), and then the residual error in the range of ($h_i \sim h_{i+m-1}$) called uncertain range by prediction of grey models is:

$$R_p(t) = R_o(t) \pm \Delta\varepsilon \tag{4}$$

where $R_p(t)$ is the function which predicts the range of uncertain spatiotemporal data, and $R_o(t)$ is the function in grey dynamic model calculates a sequence o f n numbers within one group, as shown in Fig. 8.

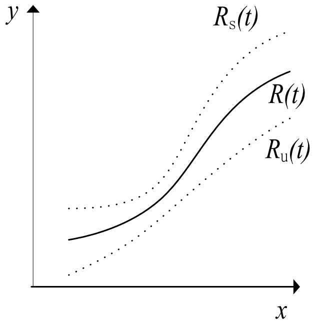


FIGURE 8. Predicted range of $f_{(t)}$.

Definition 10: A model called *URGM*, which varies over time T:

- $URGM = MBR\{(min\text{-}x, max\text{-}x) \cup (min\text{-}y, max\text{-}y)\}$, when $T = t_i$.
- $URGM = \mathrm{U}_{\mathrm{i}}^{\mathrm{i+n}}\ MBR_i\{(min\text{-}x, max\text{-}x) \cup (min\text{-}y, max\text{-}y)\}$, when $\mathrm{T} = t_i, t_{i+1}, \ldots, t_{i+n}$.

In Definition 10, when T is a time point, the grey predicted region is a *MBR* (denoted as M) whose coordinates are represented as (*min-x*, *min-y*) and (*max-x*, *max-y*). When T is a time series which is an assemble of n time points $t_0, t_1, t_2, \ldots, t_n$, $M_0, M_1, M_2, \ldots, M_n$ represent the central positions of the real typhoon.

First of all, we consider (*min-x*, *max-x*) when y is constant as shown in Fig. 9 (a). As described in Definition 9, five groups of x is predicted by the sequence when t is between t_{i-n+1} and t_i, and then the predicted coordinate of x is x_i when t is t_{i+1}. Similarly when x is constant, the (*min-y*, *max-y*) is shown in Fig. 9 (b). Lastly, the region of prediction is a series of *MBRs* when x and y is variable at the same time, as shown in Fig. 9(c).

As we can see in Fig. 9, the sequence of t_0, t_1, t_2, t_3, t_4, $\ldots$, t_n is represented as the time series. The black spots are predicted based on the previous five groups of data as shown in Fig. 9(a) and Fig. 9(b). In Fig. 9(c), the yellow spots represent the coordinates of the *MBR* for each time moment, and the red spots M_0, M_1, M_2, M_3, M_4, $\ldots$, M_n represent the true positions of spatiotemporal objects.

3) PREDICTED STEPS

On the basis of the above definitions and investigations, we have the steps of predicting uncertain spatiotemporal XML data using the grey dynamic model in the following:

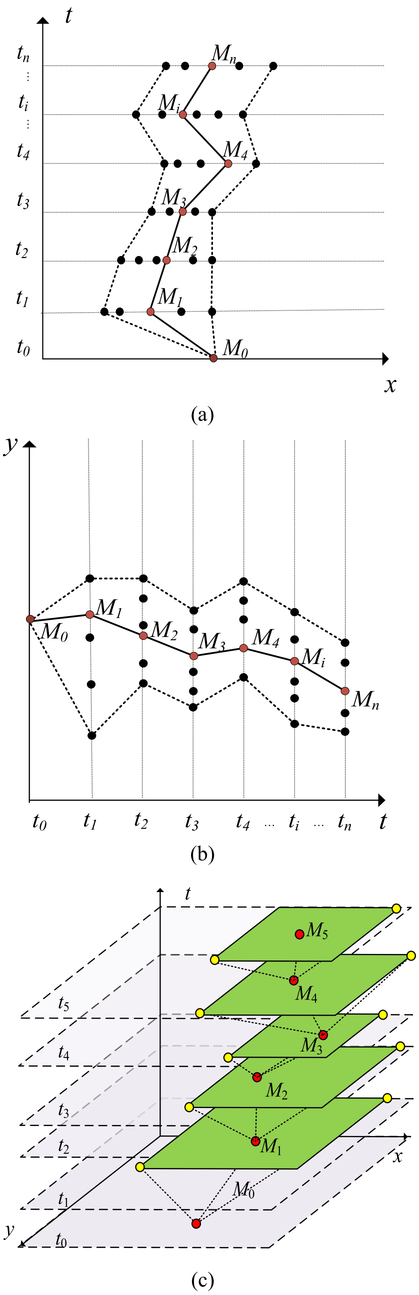


FIGURE 9. (a) The series of (*min-x*, *max-x*). (b) The series of (*min-y*, *max-y*). (c) The series of *MBRs* when *x* and *y* is variable at the same time.

Step 1: Extract data information. First of all, query the predicted information and use it to make prediction.

Step 2: Process the information. Transform the structure of the data model from multiway tree into binary tree.

Step 3: Use grey dynamic model (f_x) to process data where f_x represents the function in grey dynamic model. The original sequence $x^{(0)}$ is extracted from each group (X, f_x, t) or (h_i, f_x, t_i). *GM(1, 1)* model of the winterization equation is developed by accumulated sequence $x^{(1)}$ in the grey dynamic model. List the undetermined coefficients. The matrix B

and the data vector Y_n are transformed by the accumulated sequence $x^{(1)}$. After calculating the coefficients, the time responding equation can be obtained.

Step 4: Analyze the accuracy and make the next prediction. The residual $E_{(k)}$ and the relative error $e_{(k)}$ are calculated at each time, and the post-test error ratio is calculated based on these two values. After the prediction is assessed, if the evaluation turns out to be ideal, take the coefficient of the solution as the coefficient of the time response formulation equation and predict the next unknown time moment based on the corresponding time response formulation.

Step 5: Return the predicting results. The grey dynamic model returns the calculated probability of the prediction, and stores it in the XML database under the *forecast* node. Each time the data is updated, the grey dynamic model will predict the next time point and updates the XML database in real time.

4) ALGORITHM OF PREDICTION WITH GREY MODEL BASED ON XML (PGX)

A new algorithm of prediction with grey model based on XML (called PGX) has been represented. It is used to process spatiotemporal data and return predicted results. Grey model algorithm is redefined and applies to spatiotemporal datasets in PGX. Additionally, in order to compare Markov model with PGX, some modification is done as shown in the following.

The algorithmic PGX is used to predict attributes of spatiotemporal objects. It will return the predicted results and updated XML database when the input is USP. PGX processes USP with five steps.

- Step 1. Initialize USP, an uncertain spatiotemporal object. And the initial values will be given in this step. This step is shown in line 01.
- Step 2. Access data of USP with XML scheme, where P, T, *OID*, *ATTR* are the basic attributes of USP according to Definition 5. In this step, PGX can access different attributes of USP or ATTR. And the selected attributes will be stored as XML scheme, a multi tree which is represented as *Mtree_XML* in line 05. For example, selecting spatial of ATTR can work well when predicting the trajectory of USP. This step is shown from line 02 to 06.
- Step 3. Scheme transformation. In order to read data easily, the *Mtree_XML* needs to be transformed to binary tree, which is represented as *Btree_XML* in line 08. And this step is shown from line 07 to 10.
- Step 4. Data process. In this step, algorithm such as grey model or Markov model will be selected to predict the attributes of USP. This step is shown from line 11 to 15.
- Step 5. Update XML database. The experimental results from step 4 will be input to update XML database. This step is shown from line 16 to 18.

Algorithm 1 Prediction With Grey Model Based on XML (PGX)

```
Input: Uncertain spatiotemporal object- - -USP
Output: Predicting results and updated XML database
 01. initialize_USP();
 //access data of USP with XML scheme.
 02. define access_data(OriginalData){
 03.     input P, T, OID;
 04.     access ATTR.property as Access_data;
 05.     return Access_data as Mtree_XML;
 06. }
 07. define schema_transformation(Mtree_XML) {
 08.    Btree_XML <– Mtree_XML;
 09.     return Btree_XML;
 10. }
     //prediction with grey model or markov model.
 11. define data_process (Btree_XML) {
 12.     doubel result = Grey_Model();
 13.     double result = Markov_Model();
 14.     return predicting_results;
 15. }
 16. define Update_Forecast(predicting_results){
 17.    return update XML database;
 18. }
```

The algorithm, grey model, is used to process data in PGX. In line 01, the grey model function is defined; From line 02 to 05, the *Btree_XML* is divided into different arrays, the consequences of X_r, X_l, Y_r, Y_l is used to make prediction by grey model; From line 06 to 11, do grey model method for predicting. Each *array* denotes *Array_*X_r*[], Array_*X_l*[], Array_*Y_r*[], Array_*Y_l*[]*, and n is the length of each predicted results. And in line 9, f_x is the grey model's predicting function which is used to process data; From line 12 to 13, results are returned to PGX.

The algorithm, Markov, is also used to process data in PGX. In line 01, define the Markov model function; From line 02 to 05, divide the *Btree_XML* into different arrays, the results of X_r, X_l, Y_r, Y_l, which is used to make prediction by Markov model; From line 06 to 17, the state_transition matrix of *ATTR.attributes* is initialized. Here, we choose spatial position to be discussed. And in line 10, *total* represents the amount of transition from Array[i] to Array[$i+1$]; From line 18 to 20, make prediction with Markov model method. And in line 19, *m[k,j]* is needed in *Mx* which is the predicting function of Markov model.

IV. DISCUSSION

Markov Model is a statistical model. In order to compare it with PGX proposed in this paper, we define the state variable of Markov, and give the definition of Markov hit rate.

Definition 11: Define the state variable and the initial state in the Markov model as increment and (x_0, y_0) respectively, then the increments of x_0 or y_0 are:

$$\begin{aligned} \Delta x &= x_i - x_{i-1} \\ \Delta y &= y_i - y_{i-1} \end{aligned} \tag{5}$$

where each Δx or Δy is set as an increment.

Algorithm 2 Grey Model Algorithm

```
Input: Btree_XML
Output: Prediction by grey model
define Grey_Model() {
   group(Btree_XML){
     get Array_$X_r$[], Array_$X_l$[], Array_$Y_r$[], Array_$Y_l$[];
     return $X_r$[], $X_l$[], $Y_r$[], $Y_l$[];
   }
   for(i=0; i<Array[].length-n; i++) {
     from i to i+n−1{
       for (each array in Array_$X_r$[], Array_$X_l$[], Array_$Y_r$[], Array_$Y_l$[]) {
         do(Array, $f_x$, $t_j$);
       }
     }
   return results;
}
```

Algorithm 3 Markov Model Algorithm

```
Input:Btree_XML
Output: Prediction by Markov_Model
define Markov_Model() {
   group(Btree_XML){
     get Array_Xr[], Array_Xl[], Array_Yr[], Array_Yl[];
     return Xr[], Xl[], Yr[], Yl[];
   }
   initialize IMatrix() {
     set threshold;
     group m[0,j] and m[k,0] states by threshold;
     int $g = 0$;
     int sum=total;
     for (each array in Array_Xr[], Array_Xl[], Array_Yr[],
      Array_Yl[]) {
       when Array[i+1]-Array[i] is in m[0,j] and Array[i]-Array[i-1]
       is in m[k,0] {
           g++;
       }
       return m[k,j]= (g/sum)*100%;
     }
   }
   for (each array in Array_Xr[], Array_Xl[], Array_Yr[], Array_Yl[]) {
     do( Array,Mx,tj);
   }
}
```

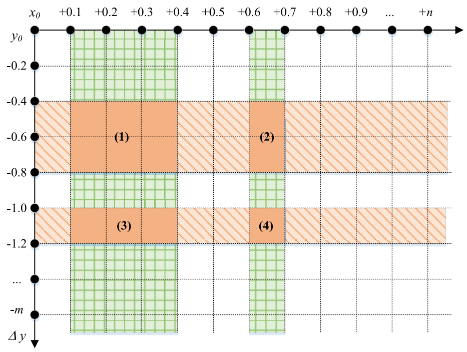


FIGURE 10. State Partition and Prediction of Markov Model.

For example, suppose $\Delta x = 0.1$, and $\Delta y = -0.2$, the state partition and prediction of Markov model is shown in Fig. 10. The state vector of Δx is (0, 1, 1, 1, 0, 0, 1, 0, 0, …, n) and the state vector of Δy is (0, 0, 1, 1, 0, 1, …, m), and then the checkered range represents the Δx range, the diagonal range represents the Δy range, and the shaded range (*DPR*) represents the junction of Δx and Δy, shown as (1), (2), (3), (4) in the Fig. 10. It is noted that the state vector in Fig.10 is the increments vector. The initial state value should add Δx and Δy, and then we can get the next predictive position.

Markov model divides the predicting range according to the state partition, which affected by the selection of the threshold. As a result, the predicting range results using Markov model cannot indicate the actual useful performance because it contains extra range by the selection of the threshold. Accordingly, we further define unit area hit rate in the following.

Definition 12: Overall Hit Rate of Markov Model (*OHRM*) is represented as:

$$OHRM = \frac{\min s}{mouts + \min s} \times 100\% \tag{6}$$

where *mins* and *mouts* are the times when predictive value is in the *DPR* or not, respectively.

Definition 13: Unit Area Hit Rate of Markov Model (*UHRM*) is represented as:

$$UHRM = \frac{\min s}{S \times (mouts + \min s)} \times 100\% \tag{7}$$

where S is the predicting range.

V. EXPERIMENT

A. EXPERIMENTAL SETUP

All the evaluations have been implemented in MATLAB2016b, Dev C++, SQL Server2016, and performed on a system with 2.4 GHz Intel Core i5 processor with 8 GB RAM running on a Windows 7 system.

To evaluate our method, we use typhoon data to make prediction [55]. The most common features of a typhoon are locations (latitude and longitude) at a particular time. So we measured these three elements in the experiments, including time, latitude and longitude. In order to simplify the experiments, we used the typhoon center's position to replace the *MBR* (X_l, Y_l), (X_r, Y_r) in our model. The experimental data is shown in Table 5 and explanation on symbols is shown in Table 6.

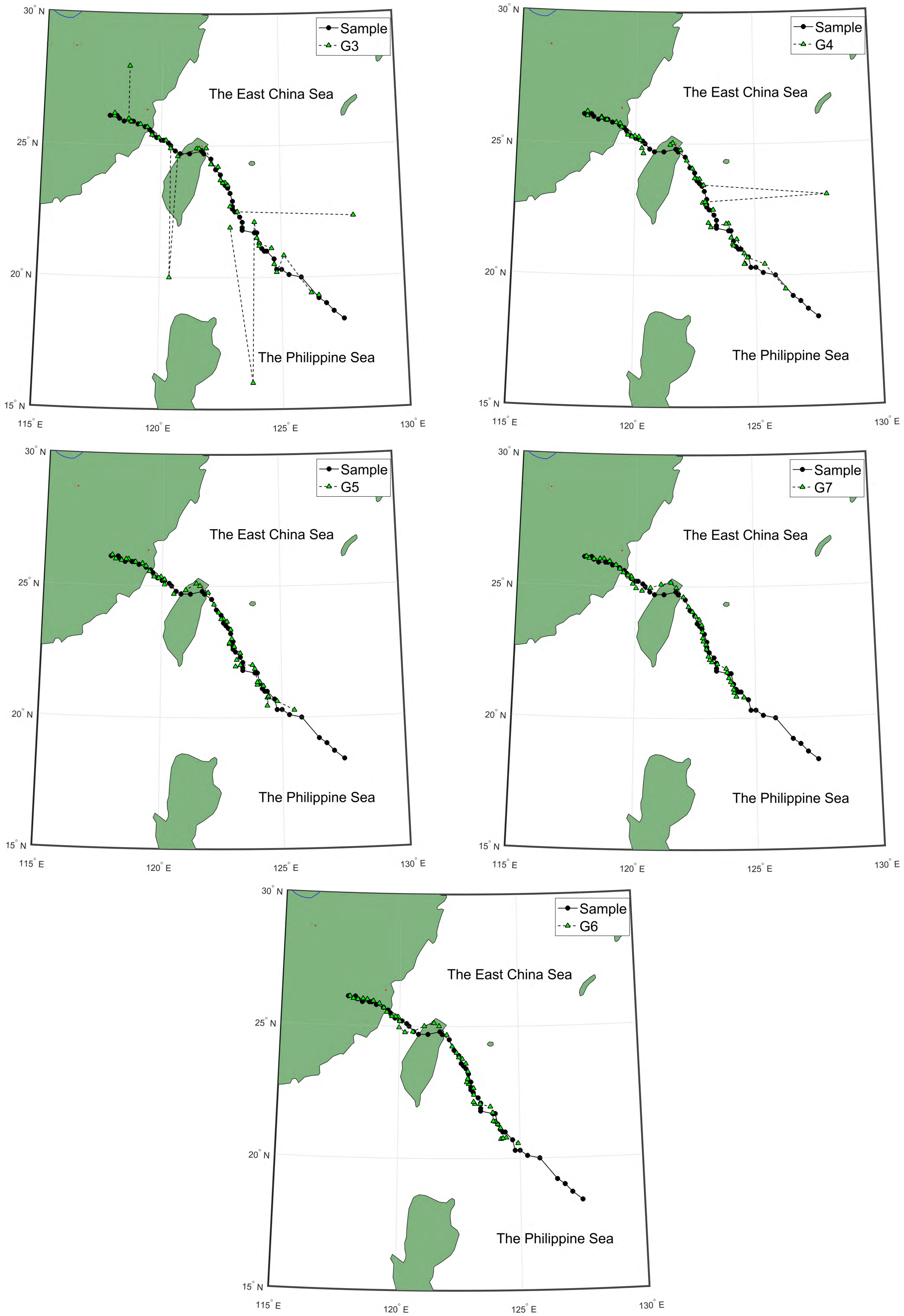


FIGURE 11. The comparisons between the prediction and actual data when using *G3, G4, G5, G6,G7*.

TABLE 5. Typhoon Nesat on July 2017.

Y/M/D	Time	Serial number	Latitude N	Longitude E	Y/M/D	Time	Serial number	Latitude N	Longitude E
2017/7/27	17:00	1	18.4	127.5	2017/7/29	15:00	26	23.5	122.7
2017/7/27	20:00	2	18.7	127.1	2017/7/29	16:00	27	23.6	122.6
2017/7/27	23:00	3	19	126.8	2017/7/29	17:00	28	23.9	122.5
2017/7/28	2:00	4	19.2	126.5	2017/7/29	18:00	29	24.1	122.3
2017/7/28	5:00	5	20	125.8	2017/7/29	19:00	30	24.5	122.1
2017/7/28	8:00	6	20.1	125.3	2017/7/29	20:00	31	24.7	121.8
2017/7/28	11:00	7	20.3	125	2017/7/29	21:00	32	24.8	121.7
2017/7/28	14:00	8	20.3	124.8	2017/7/29	22:00	33	24.7	121.2
2017/7/28	17:00	9	20.7	124.7	2017/7/29	23:00	34	24.7	120.8
2017/7/28	19:00	10	21	124.4	2017/7/29	0:00	35	24.8	120.6
2017/7/28	20:00	11	21	124.3	2017/7/30	1:00	36	25	120.4
2017/7/28	21:00	12	21.1	124.2	2017/7/30	2:00	37	25.1	120.3
2017/7/28	22:00	13	21.3	124.1	2017/7/30	3:00	38	25.2	120.1
2017/7/28	23:00	14	21.7	124	2017/7/30	4:00	39	25.2	120
2017/7/29	1:00	15	21.7	123.9	2017/7/30	5:00	40	25.3	119.8
2017/7/29	4:00	16	21.8	123.4	2017/7/30	6:00	41	25.5	119.6
2017/7/29	6:00	17	21.9	123.4	2017/7/30	7:00	42	25.6	119.5
2017/7/29	7:00	18	22.1	123.4	2017/7/30	8:00	43	25.7	119.3
2017/7/29	8:00	19	22.3	123.3	2017/7/30	9:00	44	25.8	119
2017/7/29	9:00	20	22.5	123.1	2017/7/30	10:00	45	25.9	118.8
2017/7/29	10:00	21	22.6	123	2017/7/30	11:00	46	25.9	118.7
2017/7/29	11:00	22	22.7	123	2017/7/30	12:00	47	25.9	118.4
2017/7/29	12:00	23	22.9	123	2017/7/30	13:00	48	26	118.2
2017/7/29	13:00	24	23.2	122.9	2017/7/30	14:00	49	26.1	118.1
2017/7/29	14:00	25	23.4	122.8	2017/7/30	17:00	50	26.1	117.8

B. EXPERIMENTAL SETUP

In the experiments, we took longitude and latitude data as X and Y respectively. In order to facilitate the experiments, we replaced *MBR* with the center position of typhoon. The predicted results are the moving curve of the typhoon's center that corresponds to time. In order to obtain the appropriate predicted method by which the predicted result is the most consistent with the actual data, n pairs of the data in $t_{i-n+1}-t_i$ are used as original data and the results are compared with the actual data (the value n ranges from 3 to 7). We compared them in five aspects (accuracy, relative error, posterior difference, Data fitting and response time).

1) ACCURACY

The accuracy of the results can be drawn from the comparison between the predicted results and the actual values as shown in Fig. 11, where the black thin solid line is actual data and the dotted line represents the predicted data. Since each time we take n positions of typhoon for prediction, each result of the prediction is a subsequent of the $(n+1)^{\text{th}}$ data. When data *G3* is taken, the results are shown in Fig. 11 (a), *G4* in Fig. 11 (b), *G5* in Fig. 11 (c), *G6* in Fig. 11 (d), and *G7* in Fig. 11 (e).

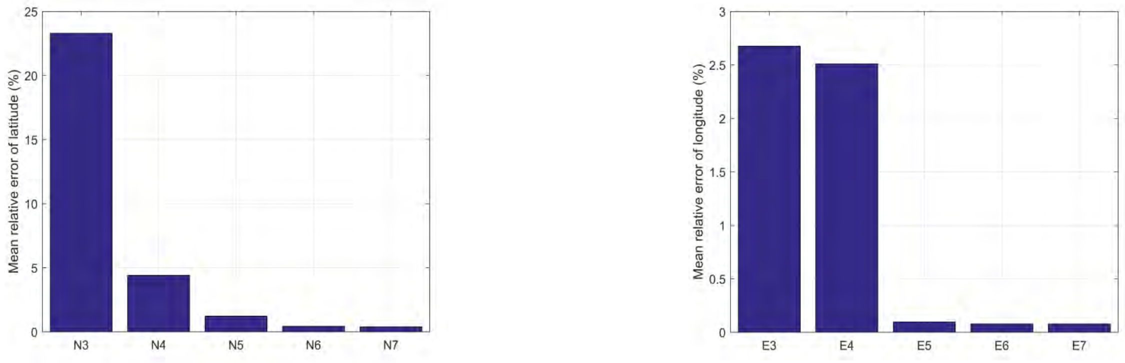
We can observe that on the aspect of accuracy from Fig. 11, when *G3* and *G4* are taken into the experiments, there are serious distortions in the prediction.

When the experiment data is *G5*, there is a serious distortion of a single value in the prediction. The prediction is more volatile. The accuracy of the results is greatly improved when *G6* and *G7* are used, and the predicted results change more smoothly when a large number of reference samples are consulted.

The process of grey dynamic prediction is to predict the data value of group $n+1$ by the pattern of the first n data sets, assuming that the data (data $n+1$) changes by the same pattern. However, unexpected situations occur sometimes in reality, and it is difficult to avoid fluctuations in the data trend. In that kind of situation, the predicted results obtained from the first n sets of data may be biased.

2) RELATIVE ERROR

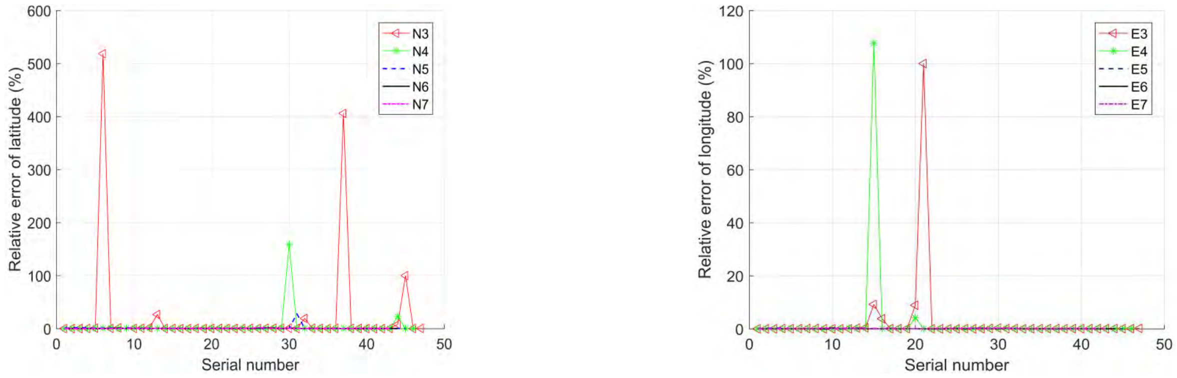
In this group of experiments, relative error is used to measure the deviation degree of predicted data from the actual data. We calculate the relative error in each prediction. When n equals to 4, average relative error of latitudes longitudes are



(a)

(b)

FIGURE 12. **(a) Comparisons on average relative error of latitudes. (b) Comparisons on average relative error of longitudes.**



(a)

(b)

FIGURE 13. **(a) Diagram of relative error of latitudes. (b) Diagram of relative errors of longitude.**

TABLE 6. **Explanation on symbols.**

Symbols	Explanations
N3	3 serial numbers of latitude as a group, so n = 3
N4	4 serial numbers of latitude as a group, so n = 4
N5	5 serial numbers of latitude as a group, so n = 5
N6	6 serial numbers of latitude as a group, so n = 6
N7	7 serial numbers of latitude as a group, so n = 7
E3	3 serial numbers of longitude as a group, so n =3
E4	4 serial numbers of longitude as a group, so n =4
E5	5 serial numbers of longitude as a group, so n =5
E6	6 serial numbers of longitude as a group, so n =6
E7	7 serial numbers of longitude as a group, so n =7
G3	Use 3 typhoon positions in the model
G4	Use 4 typhoon positions in the model
G5	Use 5 typhoon positions in the model
G6	Use 6 typhoon positions in the model
G7	Use 7 typhoon positions in the model

shown in Fig. 12(a) and Fig. 12(b). The relative error of each group's latitude and longitudes are shown in Fig. 13(a) and Fig. 13(b).

It can be observed that average relative errors of *G5*, *G6* and *G7* are all less than 0.02 from Fig. 15. Generally speaking, the predicted errors of these three groups are small. From Fig. 14, it can be concluded that the relative errors of *N6* and *N7* are stable within the range of 0.05, while the relative error of *N5* group is less than 0.05 except for one data having a large deviation. The relative errors of *E5*, *E6* and *E7* are stable within the range of 0.0035.

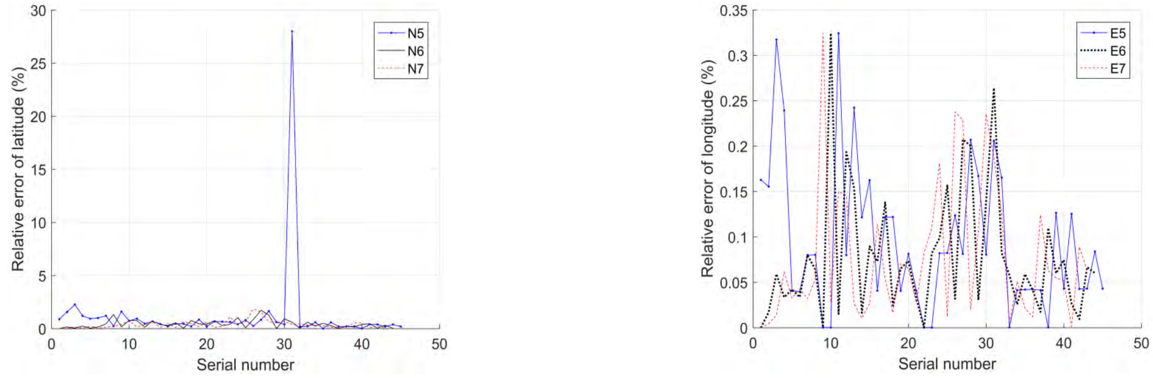
3) POSTERIOR DIFFERENCE

The posterior difference *C* represents the degree of reliability of the $(n+1)^{\text{th}}$ value calculated by the predictive equation.

$$C = \frac{S_2}{S_1} \times 100\% \tag{8}$$

$$\mathrm{S}_1 = \sqrt{\frac{1}{N}\sum_{K=1}^{N}[x_{(k)}^{(0)} - \bar{x}]^2} \tag{9}$$

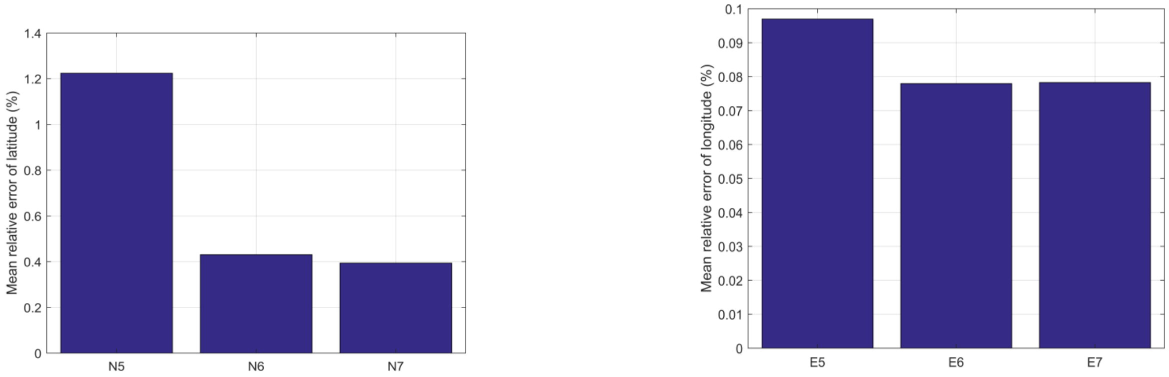
$$\mathrm{S}_2 = \sqrt{\frac{1}{N-1}\sum_{K=2}^{N}[E_{(k)} - \overline{E}]^2} \tag{10}$$



(a)

(b)

FIGURE 14. (a) Diagram of relative error of latitudes. (b) Diagram of relative error of longitudes.



(a)

(b)

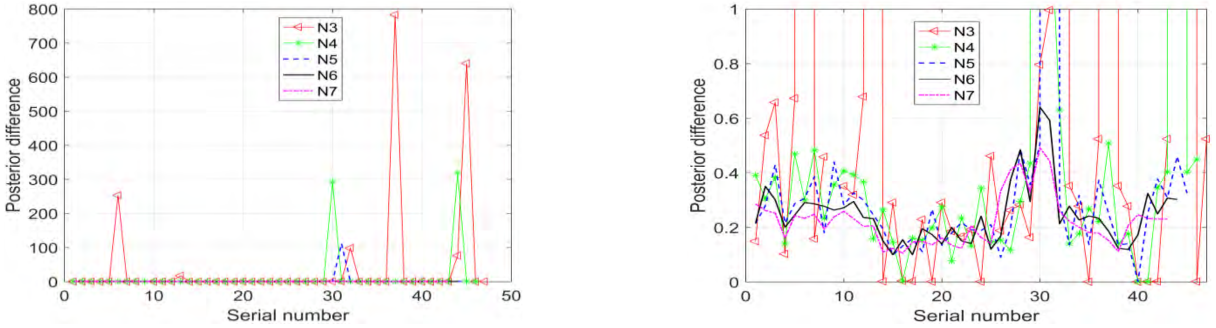
FIGURE 15. (a) Comparisons on average relative error of latitudes. (b) Comparisons on average relative error of longitudes.

where S_2 is the variance of the standard deviation of the residual error, S_1 denotes the standard deviation of the original data, and $x_{(k)}^{(0)}$ is the data of sequence 0 and k^{th}, and $E_{(k)}$ is the residual error.

We recalculate n sets of predicted data based on the equation compared with the actual data. When the data distribution is relatively stable, the posterior variance can be used as a very important measurement of predictive reliability. When the posterior variance is less than 0.35, it is considered that the predicted data is quite accurate. When the data is not stable enough, it may appear that the $(n+1)^{\text{th}}$ predicted data is still close to the actual data, even if the posterior variance is large. This is also a normal phenomenon because the $(n+1)^{\text{th}}$ data is not taken into account when calculating the posterior variance. So in either case, the posterior variance is an important criterion to measure the reliability of the prediction.

As the posterior difference is calculated by comparing the predicted data of the first n groups with the actual data respectively, which is shown in Fig. 16(a). According to the predicted equation, when the posterior variance is less than 0.35, the prediction is considered accurate. Therefore, those values under 1 should be only concern. By showing only this kind of predicted results, we have Fig. 16(b). It can be seen that posterior differences of *N6* and *N7* is the smallest ones. The *N3* and *N4* are the two largest ones, followed by *N5* group. In that case, we can draw a figure to reflect the number of each group's posterior difference elements and the proportion of elements above 0.35, as shown in Fig. 16. We can see that with the increase of the number of groups, the proportion of the posterior difference greater than 0.35 decreases gradually, indicating that the reliability of data prediction increases gradually. After removing group *N3*, *N4* and redrawing the figure, we have Fig. 17.

In Fig. 18, the blue histogram represents the number of elements included in each groups' latitude data matrix, and the yellow line represents the proportion of elements whose posterior difference exceeds 0.35 in each set of data. We can see that the proportion of posterior difference above 0.35 decreases as the amounts of groups increases, thus indicating that the reliability of data forecasting grows When we remove group *N3*, *N4* and redraw the figure as shown in Fig. 17, the posterior difference trends basically the same although the grouping among N5, N6 and N7 differs from each other. The posterior difference is relatively large from serial 1 to 14, and the posterior difference from groups



(a)

(b)

FIGURE 16. (a) Distribution on posterior difference of latitude. (b) Distribution on posterior difference of latitude under 1.

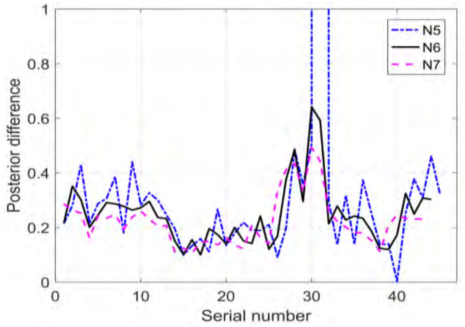


FIGURE 17. Distribution of latitudes' posterior difference on *N5, N6, N7*.

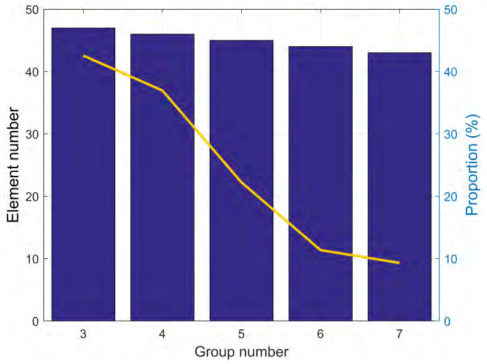


FIGURE 18. Posterior difference on latitudes of each group and the proportion of those above 0.35.

15 to 25 is smaller, followed by two peaks, which are more obvious in *N5* and *N6*, whereas the trend of *N7* is steadier, while it has a smaller value. We can draw a conclusion that as the grouping (sequence number) increases, changes in the posterior difference also tend to be stable. These characteristics are due to the distribution of the original data, as detailed data is shown in Table 7, Table 8, and Table 9.

It can be found that a number of a posteriori difference peaks occur when the original data has the same value, such as group 31 of *N5*, group 30 of *N6*, group 30 of *N7*, where 24.7 and 24.8 occurred multiple times. These data tend to become both larger and smaller, which is a great disadvantage

TABLE 7. N=5 and posterior difference of latitude > 0.35.

Group i	Posterior difference	Predicted sequence
3	0.4301	19, 19.2, 20, 20.1, 20.3
7	0.3855	20.3, 20.3, 20.7, 21, 21
9	0.4401	20.7, 21, 21, 21.1, 21.3
28	0.4863	23.9, 24.1, 24.5, 24.7, 24.8
29	0.3569	24.1, 24.5, 24.7, 24.8, 24.7
30	0.5054	24.5, 24.7, 24.8, 24.7, 24.7
31	111.4036	24.7, 24.8, 24.7, 24.7, 24.8
36	0.3732	25, 25.1, 25.2, 25.2, 25.3
42	0.3776	25.6, 25.7, 25.8, 25.9, 25.9
44	0.4604	25.8, 25.9, 25.9, 25.9, 26

TABLE 8. N=6 and posterior difference of latitude > 0.35.

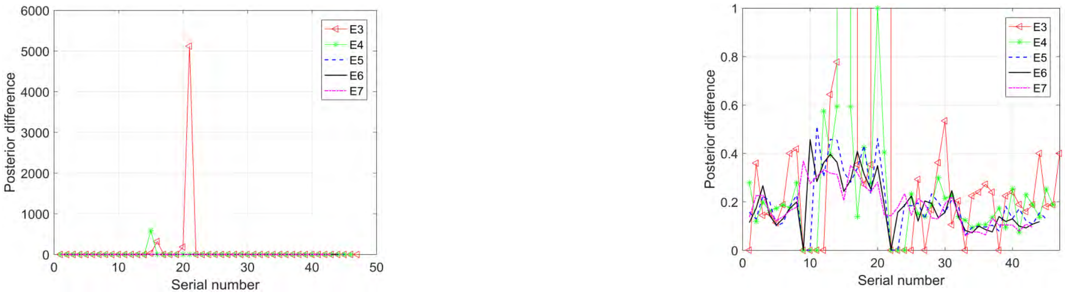
Group i	Posterior ifference	Predicted sequence
2	0.3505	18.7, 19, 19.2, 20, 20.1, 20.3
27	0.3742	23.6, 23.9, 24.1, 24.5, 24.7, 24.8
28	0.4845	23.9, 24.1, 24.5, 24.7, 24.8, 24.7
30	0.6409	24.5, 24.7, 24.8, 24.7, 24.7, 24.8
31	0.5917	24.7, 24.8, 24.7, 24.7, 24.8, 25

TABLE 9. N=7 and posterior difference of latitude > 0.35.

Groupi	Posterior difference	Predicted sequence
27	0.4110	23.6, 23.9, 24.1, 24.5, 24.7, 24.8, 24.7
28	0.4352	23.9, 24.1, 24.5, 24.7, 24.8, 24.7, 24.7
30	0.4921	24.5, 24.7, 24.8, 24.7, 24.7, 24.8, 25

for the prediction. But at the same time, according to the figure and table above, it can be concluded that larger group number *n* can be well compatible with these kinds of data distribution. Therefore, we can conclude that in order to improve the accuracy of prediction, we should obtain differently valued original data. If this is unavoidable, it's a good way to maximize *n* groups.

The posterior difference is calculated by comparing the predicted data of the first *n* groups with the real data, distribution on posterior difference of longitude in shown in Fig. 19(a). When the posterior variance is less than 0.35, it is considered that the prediction is quite accurate, so we only care about those values under 1. By only showing this kind of predicted results, we have Fig. 19(b).



(a)

(b)

FIGURE 19. (a) Distribution on posterior difference of longitude. (b) Distribution on posterior difference under 1.

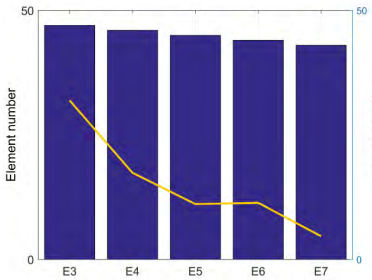


FIGURE 20. Posterior difference on longitude of each group and the proportion of those above 0.35.

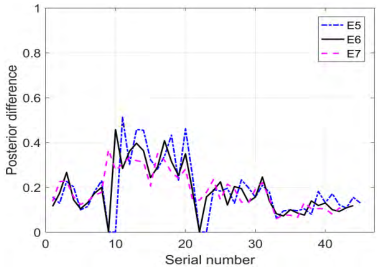


FIGURE 21. Distribution of latitudes' posterior difference on *E5*, *E6*, *E7*.

Similarly, it can be seen that posterior differences in group *E3* and *E4* are the two largest ones, *E5* follows the posterior difference of E6 and E7 are the smallest ones, so we can draw a figure that reflects the number of each group's posterior difference elements and the proportion of elements above 0.35 as shown in Fig. 20

We can see that the proportion of posterior difference above 0.35 decreases as the amount of groups increases, thus indicating that the reliability of data forecasting grows. After removing group *E3*, *E4* and redrawing the figure, we have Fig. 21.

We can use similar analysis on longitudes which have been done on latitudes. Detailed data is shown in Table 10, Table 11, and Table 12.

Similar conclusions can be drawn from Table 10, Table 11, and Table 12 that when the data in a sequence have the similar or same values, the posterior difference of the prediction would be too large. In addition, the larger portion the same data has, the less satisfactory the results are. For example, in group 13 of *E6*, the data 123.4 appears three times and the posterior difference is 0.3965. In the group *E6* of *E7* the data 123.4 appears three times, resulting in the posterior difference of 0.3504.

Fig. 22 shows the average values of posterior difference. The prediction made by grey dynamic model is only reliable when the posterior difference is no greater than 0.35. We can see from Fig. 22(a) and 22(b) that group *G3* and *G4* have not met with this requirement. So we redraw the figure only on *G5*, *G6* and *G7*. The value of a posteriori difference on the group 5 is far above 0.35, which means that taking five groups of sample data as the original data for prediction

TABLE 10. E=5 and posterior difference of latitude > 0.35.

Group i	Posterior difference	Predicted sequence
11	0.5124	124.3, 124.2, 124.1, 124, 123.9
13	0.4580	124.1, 124, 123.9, 123.4, 123.4
14	0.4544	124, 123.9, 123.4, 123.4, 123.4
18	0.4327	123.4, 123.3, 123.1, 123, 123
20	0.4611	123.1, 123, 123, 123, 122.9

TABLE 11. E=6 and posterior difference of latitude > 0.35.

Group i	Posterior difference	Predicted sequence
10	0.4572	124.4, 124.3, 124.2, 124.1, 124, 123.9
12	0.3611	124.2, 124.1, 124, 123.9, 123.4, 123.4
13	0.3965	124.1, 124, 123.9, 123.4, 123.4, 123.4
14	0.3628	124, 123.9, 123.4, 123.4, 123.4, 123.3
17	0.4080	123.4, 123.4, 123.3, 123.1, 123, 123

TABLE 12. E=7 and posterior difference of latitude > 0.35.

Group i	Posterior difference	Predicted sequence
9	0.3681	124.7, 124.4, 124.3, 124.2, 124.1, 124, 123.9
16	0.3504	123.4, 123.4, 123.4, 123.3, 123.1, 123, 123
30	0.4921	24.5, 24.7, 24.8, 24.7, 24.7, 24.8, 25

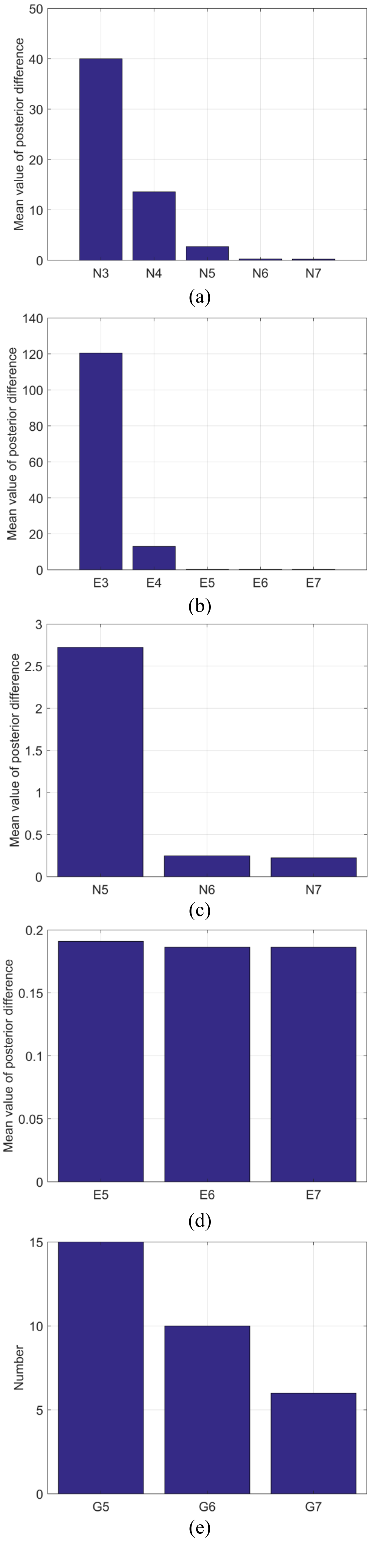


FIGURE 22. **(a) Average values of latitude's posterior difference. (b) Average values of longitude's posterior difference. (c) Mean values of latitude's posterior difference on *N5 N6 N7*. (d) Mean values of longitude's posterior difference on *N5 N6 N7*. (e) Numbers of data that doesn't meet the requirement of accuracy.**

is less reliable. While values of the group E6 and group E7 are both less than 0.35, the predictions have a stronger reliability. Here we have the number of elements that don't meet the requirement. We can see from the Fig. 22(e) that the prediction is more reliable when group number is G7.

4) DATA FITTING

Fig. 11(a), (b), (c), (d) and (e) show the actual trajectory of the typhoon on the map and the trajectory predicted by the grey dynamic model. As the amount of reference groups for prediction increases, the predicted value of the trajectory becomes closer to the actual value and the fitting degree becomes higher. But no matter how well they fit, there are still some gaps between the predicted points and the actual points. When considering an actual situation, suppose that a typhoon is about to occur in a place, the relevant departments need to predict the future location of the typhoon quickly, so that evacuation and protection can be organized in advance. Therefore, this prediction should be as comprehensive as possible and should be able to predict a regional scope, in order to maximize the safety of the residents' lives and property.

In order to evaluate the degree of fitting, our model uses the predicted value and the sample to calculate the relative error, which means that the sample should be located in the area whose center is the predicted data and the radius is the relative error. We calculate the average residual errors of the first n groups and then combine them with the predicted data. So we have the range calculated by Definition 10 as the upper bound and the lower bound of the predicted latitude and longitude. Theoretically, this range of prediction should include the trajectory of the moving typhoon. The experimental results are shown in Fig. 23

The red dotted line in Fig. 23 is the actual value of the sample data, that is, the path of the moving typhoon, and the gray area is the closed area enclosed by the predicted upper and lower bounds of the latitude and longitude. As seen from Fig. 23(a), due to the small number of reference sample, the extreme values in the predicted data are excessive, making it too dispersed in the map and unable to form a closed area. For Fig. 23(b) and 23(c), there are still large gaps between the predicted area and the sample value because of extreme values. In Fig. 23(d) and 23(e), there is no extreme value and the inclusion of actual data in the predicted area has improved greatly compared with the first three images, but the results are still not satisfactory. The trajectory of the typhoon is affected by uncertainty in the process of movement, especially some of these factors only happen at a specific time. They should not be included in our model as the accuracy of the predicted range may be interfered.

Based on the analysis above, we propose a new way to predict the range denoted as Largest Region Predicted Method or LRPM. According to the characteristics of grey dynamic predicted model, we can select different values of sample as prediction reference group which may increase the uncertainty in the predictive model. Therefore, we remove the extreme values in the predicted results of *G4*, *G5*, *G6*, *G7*, and select the maximum and minimum values from the rest of the predicted data as the border of the region, shown in Fig. 24.

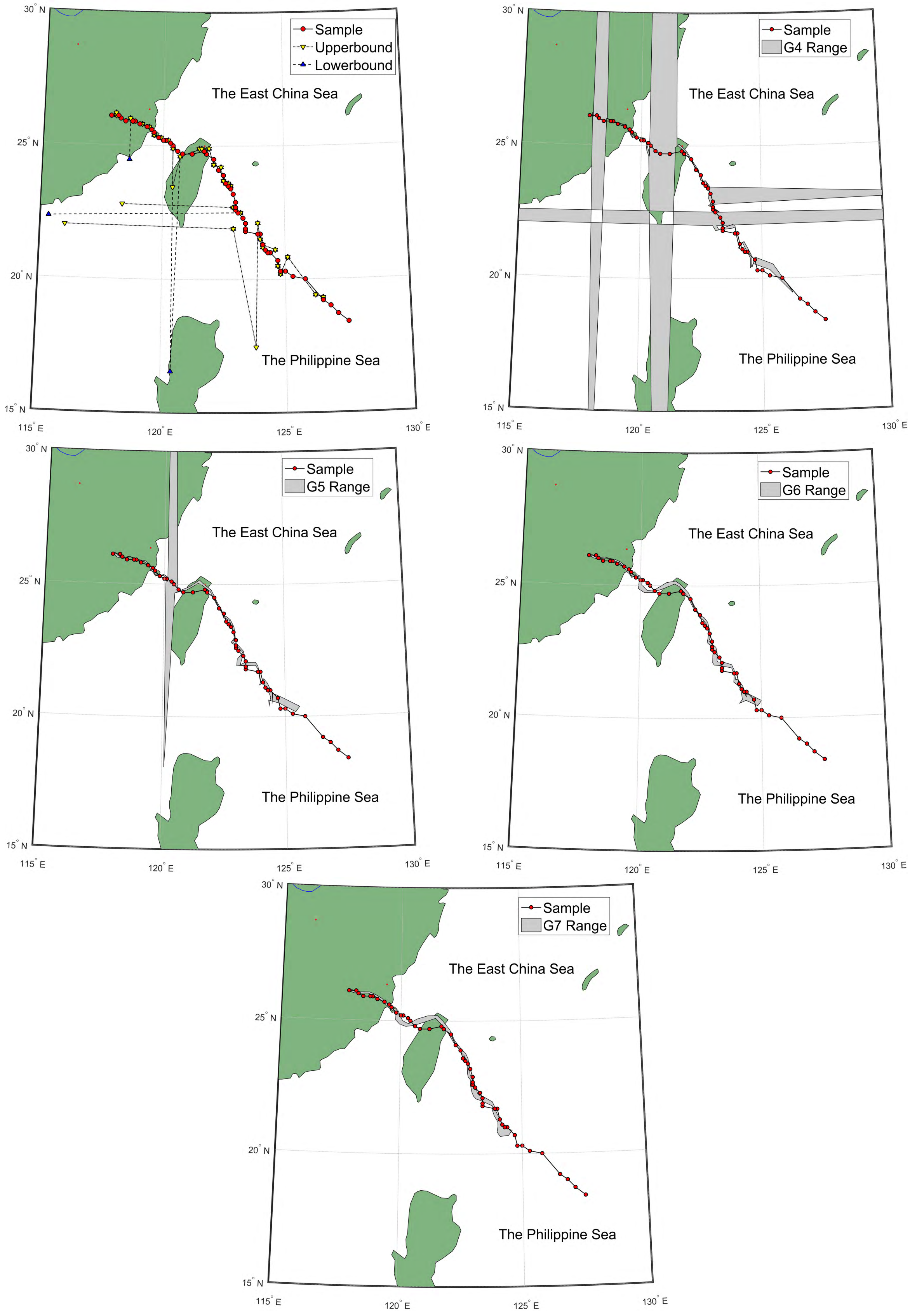


FIGURE 23. Uncertain regions of G3, G4, G5, G6, G7 predicted by REPM.

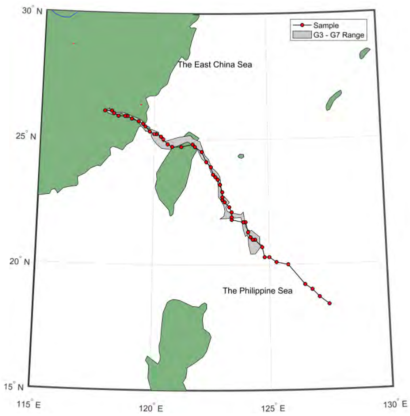


FIGURE 24. **Uncertain regions predicted by LRPM.**

We use *G3*, *G4*, *G5*, *G6*, *G7* to obtain the maximum and minimum values, and use these values as the upper and lower bounds of the range. As seen from Fig. 24, we have greatly increased the actual sample value inclusion in this way.

5) RESPONSE TIME

For data storage and query, the traditional way is to use a relational database, and in this paper, we use improved XML to do this. The comparison between these two methods on time efficiency is shown in Fig. 25. As seen from the histogram, XML is far superior to the relational database in terms of the efficiency of the storage and query, and it is proved that XML is a better way to store and query predictive data. In order to get the effect of the grouping number on predicted time, we present the combined data in longitude and latitude to show the average predicted time of *G3*, *G4*, *G5*, *G6* and *G7*, as shown in Fig. 26. It can be seen that the predicted time of these five groups are at most the degree of 10^{-3}, which also ensure the efficiency of the actual predicted application. By comparison, it can be concluded that it will cost more time querying data than predicting the results in XML structures.

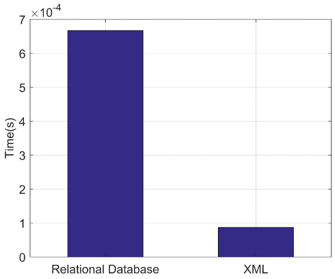


FIGURE 25. **Query times of relational database and XML.**

The experimental data is totally 50 couples which is divided into several groups. Therefore, the total number is certain, and the predicted time decreases as the number of groups increases. Theoretically, as the number of groups increases, the predicted time is usually shorter.

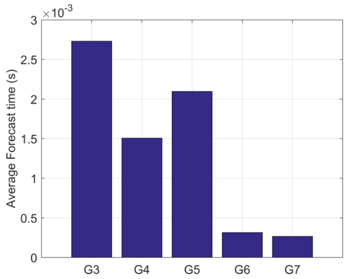


FIGURE 26. **Average sums of time used to predict latitude and longitude.**

In the following, we compare XML with relational database model in query time. When querying predicted data, it is necessary to convert the multi-tree data model built by XML into a binary tree data model, and it is necessary to convert and establish the binary tree data model. The total conversion time and the total query time in the XModel add up to 0.00008718s. Compared with the storage structure of spatiotemporal data in relational database [56], EModel queries the same set of experimental data and for an average time of 0.0006674s. From the comparison, we can see the advantages of storage based on XML. The query time of each group by XML and relational database is shown in Fig. 27.

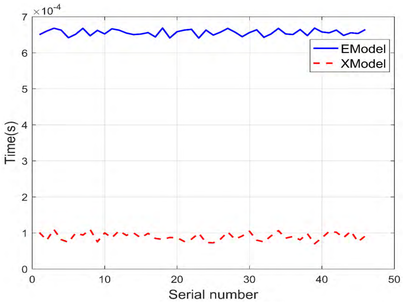


FIGURE 27. **Query time of EModel and Xmodel.**

Compared with XML, using traditional relational database is better in generalization, data storage, data persistence and query service. But the reading speed is low in relational database. In our experiment, the only thing needed is to transform the multi tree into binary tree by storing uncertain spatiotemporal data integrated XML. The experiment demonstrates that the transform time is less than that of querying database. The time of querying XML database is only 0.001 percent of transformation, which is negligible. Hence, XML is the better choice in the storage and querying of uncertain spatiotemporal data. At the same time, we are studying how to combine relational database with XML to store uncertain spatiotemporal data effectively.

C. EVALUATION ON MARKOV MODEL

In combination with *State* and *Probability* of data in the last 20 groups predicted in experiment, the results show that minimum value was 0.1 and maximum value was 0.5, as shown in Table 13 and Table 14.

TABLE 13. The state partition of latitude.

Group num.	State 1	State 2	State 3	State 4	State 5	State 6
31	0.5	0	0	0.5	0	0
32	0.1	0.2	0.2	0.1	0.2	0.1
33	0	0.375	0.5	0.125	0	0
34	0	0.5	0	0.25	0.25	0
35	0	0.5	0	0.25	0.25	0
36	0	0.375	0.5	0.125	0	0
37	0.1	0.2	0.2	0.1	0.2	0.1
38	0	0.375	0.5	0.125	0	0
39	0	0.375	0.5	0.125	0	0
40	0	0.5	0	0.25	0.25	0
41	0	0.375	0.5	0.125	0	0
42	0.1	0.2	0.2	0.1	0.2	0.1
43	0	0.375	0.5	0.125	0	0
44	0	0.375	0.5	0.125	0	0
45	0	0.375	0.5	0.125	0	0
46	0	0.375	0.5	0.125	0	0
47	0	0.5	0	0.25	0.25	0
48	0	0.5	0	0.25	0.25	0
49	0	0.375	0.5	0.125	0	0
50	0	0.375	0.5	0.125	0	0

TABLE 14. The state partition of longitude.

Group num.	State 1	State 2	State 3	State 4	State 5
31	0.14	0	0.43	0.43	0
32	0.14	0	0.43	0.43	0
33	0	0.1	0.17	0.72	0
34	0	0	0.67	0.3	0
35	0	0	0.67	0.3	0
36	0.14	0	0.43	0.43	0
37	0.14	0	0.43	0.43	0
38	0	0.1	0.17	0.72	0
39	0.14	0	0.43	0.43	0
40	0	0.1	0.17	0.72	0
41	0.14	0	0.43	0.43	0
42	0.14	0	0.43	0.43	0
43	0	0.1	0.17	0.72	0
44	0.14	0	0.43	0.43	0
45	0.14	0	0.43	0.43	0
46	0.14	0	0.43	0.43	0
47	0	0.1	0.17	0.72	0
48	0.14	0	0.43	0.43	0
49	0.14	0	0.43	0.43	0
50	0	0.1	0.17	0.72	0

And the *States* in Table 13 and Table 14 is the *States* divided by *threshold* according to Definition 11. It is used to confirm the state-transition matrix which is an essential parameter in Markov Model. In combination with other probabilities, the values could be divided into eight segments: [0, 0.1], (0.1, 0.12], (0.12, 0.125], (0.125, 0.23], (0.23, 0.25], (0.25, 0.375], (0.35, 0.5] and (0.5, 1]. In addition, when the *Probability* value is not less than 0.5, it is considered that this state is more likely to occur, so the threshold value should be less than 0.5. Therefore, the threshold values of 0.1, 0.12, 0.125, 0.23, 0.25 and 0.375 can be selected. Similarly, in the longitude predicting results, the minimum value is 0.1 and the maximum value is 0.72. According to the same principle above, the *Probability* values can be divided into six segments, which are [0, 0.1], (0.1, 0.14], (0.14, 0.17], (0.17, 0.3], (0.3, 0.43], (0.43, 0.67] and (0.67, 0.72]. Then the thresholds are 0.1, 0.12, 0.15, 0.17, 0.35 and 0.43 in the range less than 0.5.

According to the *Probability* range of latitude and longitude, 10 sets of common thresholds are selected, which are 0.1, 0.12, 0.125, 0.15, 0.17, 0.23, 0.25, 0.35, 0.375 and 0.43 following the discussion above. And these thresholds will be used to select the group when the probability is equal or greater than corresponding threshold. Experimental results will be displayed in the following discussion according to different thresholds.

1) PREDICTED TIME

We test the predicted time under different thresholds. When the threshold is equal to 0.1, the predicted time is about 0.025s, while the predicted time under other thresholds is close to 0.02s. The consumption time in the first experiment is obviously more than that in the other groups, because matrixs need to be created and data need to be loaded. As a result, we select the average time of the last nine sets of predicted time as the final prediction of Markov model, as shown in Fig. 28.

2) MEMORY CONSUMPTION

Memory consumption of Markov Model is shown in Fig. 29. The experimental results show that there is no significant relationship between memory consumption and the selected

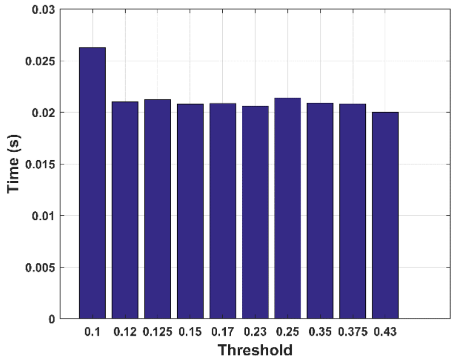


FIGURE 28. Predicted time of Markov Model.

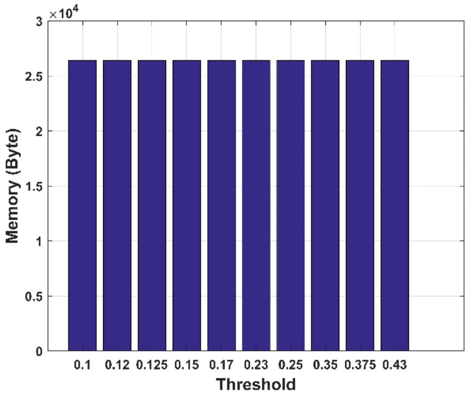


FIGURE 29. Memory consumption of Markov Model.

threshold. Since we employ the same experimental method, and the data we used was independent in the prediction, the detailed memory usage maintained about 26416 Byte fluctuate up and down shown in Table 15.

TABLE 15. The data of memory consumption using Markov model.

Memory(Byte)	Threshold	Memory(Byte)	Threshold
26430	0.1	26412	0.23
26390	0.12	26421	0.25
26420	0.125	26418	0.35
26423	0.15	26415	0.375
26416	0.17	26416	0.43

3) UNIT AREA HIT RATE

The variables used in this experiment are latitude and longitude, so the predicting *Area* is $S = \Delta\text{latitude} \times \Delta\text{longtitude}$. Therefore, the concept of Square Degree (*SD*) is defined as the unit of S. A unit *SD* is defined as the area of a rectangle with two sides' length of a unit latitude and a unit longitude.

In Fig. 30, the histogram represents the *Area* of the range predicted under each threshold. From Fig. 30, we can see that the maximum *Area* is more than 700 *SD* and the minimum is less than 10 *SD*. The results are reasonable, because when the selected thresholds increase, the *State* range may decrease, leading to the reduction of the A*rea*.

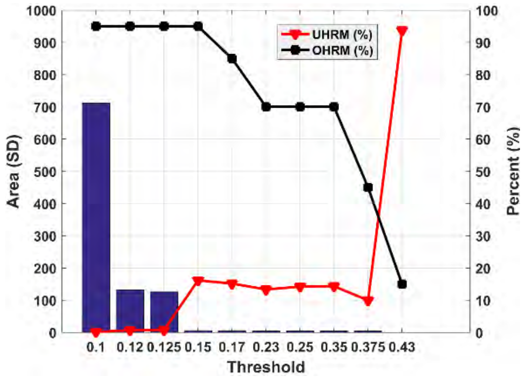


FIGURE 30. Area and hit rate.

According to Definition 13, Overall Hit Rate of Markov Model (*OHRM*) is the percentage of the predicted points as shown in Fig. 30, the maximum *OHRM* is 95%, and the minimum is 15%. With the increase of thresholds, the *OHRM* shows a downward trend.

Specifically, when the threshold is less than 0.15, *OHRM* remains at 95%. Analyzing the reason, on the one hand, *OHRM* is stable because the increments of thresholds are small. On the other hand, *OHRM* is high because the smaller threshold of this segment, the more *states* is selected. So the predicted range has been extended, which contains more actual values. *OHRM* decreases when the thresholds are selected at [0.15, 0.23], because the threshold gradient is higher than before (0.02, 0.06 to 0.02, 0.05, 0.025). What's more, some data points are filtered out with the threshold increasing in this range. When the thresholds are selected at [0.23, 0.35], *OHRM* remains the same. Within this range, the forecast range of latitude has a certain impact, while the forecast range of longitude is the same as the threshold value of 0.15. When the thresholds are selected at [0.35, 0.43], *OHRM* always drops, and the falling speed is accelerated. The reason is that there are more data points of latitude and longitude filtered out by high threshold value. It is concluded that when the selection of thresholds can affect both latitude and longitude predicted range, *OHRM* will change dramatically.

According to the analysis of the *Area* and *OHRM* of Markov predicted range above, it is necessary to reduce the threshold and expand the predicted range in order to improve *OHRM* of Markov model. In actual situation, the typhoon forecast needs the request of accurate and moderate scope, which ensures that people can carry on the protection more quickly, accurately and effectively before typhoon comes. Therefore, considering the size of predicted range on the effect is important. So we defined the concept of *unit area hit rate* (*UHRM*) according to Definition 14, that is, the ratio of the *OHRM* to the *Area* of predicted range.

As shown in Fig. 30, when the thresholds are in [0.1, 0.125], the *Area* of prediction is large and *OHRM* is remained at a high level of 95%, so *UHRM* remains below 1%. When the threshold is 0.15, *UHRM* is improved to 16% due to deduction of *Area*, while *OHRM* is still 95%. When the thresholds are in [0.15, 0.375], *UHRM* fluctuates down because *Area* doesn't change significantly, and *OHRM* decreases. When the threshold increases to 0.43, *OHRM* also reduces, but *UHRM* still improves greatly, reaching more than 90%. From the analysis above, we can see that *UHRM* is a more favorable indicator to measure the advantages and disadvantages of the methods.

According to Fig. 30, *OHRM* is 95% when the thresholds are 0.12 and 0.125, and *Area* is only 1/5 comparing with area when threshold is 0.1. *OHRM* is more than 70% when the thresholds are 0.15, 0.17, 0.23, 0.25, and 0.35. *Area* is less than 10, and *UHRM* is about 15%. So we select these eight sets of thresholds to count again, as shown in Fig. 31, which

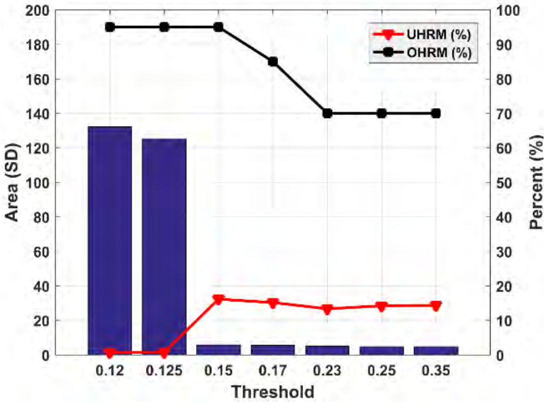


FIGURE 31. Area and hit rate.

shows that when the threshold is 0.15, the area is the smallest, and *OHRM* and *UHRM* reach the maximum of these groups. So the threshold of 0.15 is the most appropriate choice for this experiment.

D. COMPARISONS

1) PREDICTED TIME

In Grey model and Markov's multi-group experiments, the predicted time is calculated as shown in Fig. 32. The predicted time of the Grey model is less than 0.015s, and that of Markov model is more than 0.02s. Therefore, we can conclude that the Grey model is superior to the Markov model in terms of memory consumption.

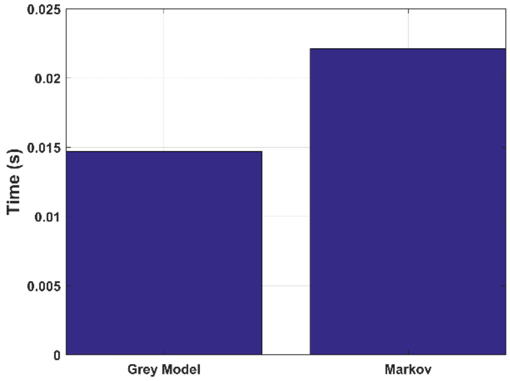


FIGURE 32. **Predicted time of Grey model and Markov model.**

2) MEMORY CONSUMPTION

By comparing the memory usage in Matlab between Grey model and Markov model, their memory consumptions are shown in Fig. 33. The memory consumption of the Grey model is about 8,000 byte, while that of the Markov model is about 26,000 byte, nearly 3.25 times as much as that of the Grey model. Thus, the performance of Grey model is obviously better than that of Markov model in memory consumption.

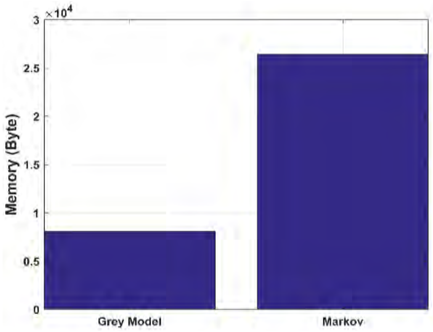


FIGURE 33. **Memory consumption of Grey model and Markov model.**

The memory consumption seems negligible due to its only MB degree. But it is worth mentioning that our dataset is small with only 50 series of data. Small static datasets are not enough to estimate their memory consumption such as our dataset. But dynamic datasets will be pervasive in the long run. In this situation, estimating the memory consumption makes sense.

3) UNIT AREA HIT RATE (UHRM)

Comparisons on *UHRM* are between the G7 group with the best performance in Grey model and the best performed group with threshold equals to 0.15 in Markov model, shown in Fig. 34. The results show that *UHRM* in Markov model is better about 1.5% higher than that in Grey model.

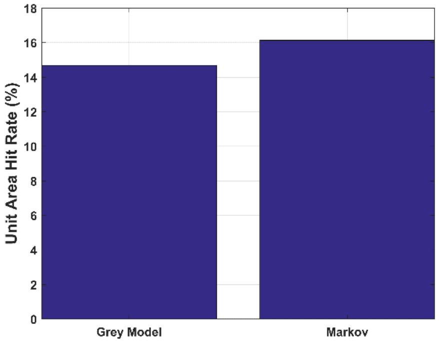


FIGURE 34. **Unit area hit rate of Grey model and Markov model.**

Analyzing the reason, the location of the typhoon at the next time should be predicted firstly when Grey model is used to predict range, and then the scope should be extended outwards as the center of the circle. The scope of Markov model is determined by *State*, and probabilities of different *States* are generated by prediction. As a result, the scope of the prediction is wider, and then the *UHRM* is also higher. However, it is worth noting that the *State* matrix of Markov model needs to be changed dynamically according to different data sets, and Grey model doesn't need these transformations

E. EXPERIMENTAL CONCLUSION

1) QUERY PERFORMANCE

Query performance is mainly measured by response time. Through the statistical analysis of data, we can see that most time is taken up by the storage and query in the actual situation, and the calculation of prediction takes much less time than the former two. At the same time, by comparing the traditional relational database and the proposed approach in this paper, it is proved that querying uncertain space-time data in XML has higher efficiency and greater advantages.

2) PREDICTED PERFORMANCE

a: GREY MODEL WITH XML (PGX)

Accuracy and Relative Error: From the comparisons between predicted results with the actual data and numerical analysis, we can conclude that bizarre error would occur when the data are divided into 3 or 4 groups, and when the group number is 5, some data have greater errors, but when the group number is 6 or 7, the predicted results are satisfactory.

Posterior Difference: In the evaluation of the predicted reliability, we conclude that posterior difference is not only related to the group number, but also relies on the original data's smoothness. When the data do not change significantly, all original data should be treated with caution. And such data may be deleted if necessary. When the data distribution is suitable for prediction, the posterior difference will increase if the number of groups is too small. So selecting the appropriate grouping number is the key factor. In this paper, group 7 is the most appropriate choice.

b: MARKOV MODEL

Accuracy: With the increase of threshold, the predicted range of Markov model method is shrinking. At the same time, *OHRM* is decreasing due to the reduction of the forecast range. In particular, *OHRM* will greatly change when the selection of threshold can affect the predicting range of latitude and longitude at the same time. *UHRM* is the ratio of *OHRM* to the area of the predicted range, which is also a key measure of the appropriate threshold selection. *UHRM* increases with the increasing of thresholds. The smaller the threshold is, the greater the impact of predicted area on *UHRM* is. When the threshold is larger, the impact of *OHRM* on *UHRM* is stronger. The results show that *OHRM* or *UHRM* is the most appropriate one when the threshold is 0.15.

c: THE COMPARISONS BETWEEN PGX AND MARKOV MODEL

Through the comprehensive experiments, Grey model and Markov model are used to predict the trajectory of typhoon, respectively. The results show that Grey model is superior obviously to Markov model in both time consumption and the memory consumption. What's more, Markov model is slightly better than Grey model in *UHRM*. However, it's necessary to consider the division of *States* and threshold selection according to different situations when applying the Markov model into practice. As a consequence, it's more complicated than the Grey model. Therefore, Grey model is better than Markov model on the prediction of uncertain spatiotemporal data.

To sum up, we can draw the following conclusions:

i) PGX, the more predicted groups, the smaller consistency and the relative errors are;

ii) PGX, the distribution of the original data has a great impact on the posterior difference. When selecting the original data, we should pay attention to the unregulated data and select the appropriate group number;

iii) In Markov model, the selection of the thresholds has a great impact on *UHRM* which is dataset oriented. The thresholds should not be too large or too small when setting a threshold.

iv) Markov model is slightly better than Grey model in *UHRM*, while Grey model is superior to Markov model significantly in terms of time consumption and memory usage. Therefore, Grey model method has excellent performance in the prediction of uncertain spatiotemporal data.

v) PGX can ensure the predicted accuracy and at the same time improve the storage and query performance of uncertain spatiotemporal data effectively.

VI. CONCLUSION

This paper proposes an approach called PGX which predicts uncertain spatiotemporal data by integrating grey dynamic model with XML. Considering the multilayered information of uncertain spatiotemporal data as well as its hierarchical representation, we establish several groups of experiments to show its advantages. Firstly, uncertain spatiotemporal XML model grows fast and has a large storage capacity, which makes it adaptable to acclimatize to the long-term expansion of database. Secondly, experimental results show that PGX achieves a minimum mean accuracy of 0.5% in a short time. Markov' *UHRM* is better about 1.0% than that in PGX, but the *State* of matrix of Markov needs to be changed dynamically as memory consumption increases, while PGX doesn't need any transformation. Finally, compared with the traditional database storage, PGX has the advantages in querying time over relational database. The experimental results show that PGX can effectively improve the efficiency of information storage and retrieval when the experimental predicted accuracy (with a relative error between 5% and 0.5%) is guaranteed. What's more, the query time based on XML is 86.94% shorter than SQL Server.

In conclusion, PGX performed well in the prediction of uncertain spatiotemporal data in our study.

VII. FUTURE WORK

In this paper, we discussed the prediction of typhoon trajectory by using its previous spatial position. We did a lot of work to discuss the precision with different parameters. But the built object USP has a series of attributions, such as OID and ATTR.trend. Besides, geographical and meteorological factors were not taken into account. This work is promised to be done in the future.

REFERENCES

[1] M. Santos, C. Bateira, C. Hermenegildo, and L. Soares, "Hydrogeomorphologic GIS database in Northern Portugal, between 1865 and 2010: Temporal and spatial analysis," *Int. J. Disaster Risk Reduction*, vol. 10, pp. 143–152, Dec. 2014.

[2] G. Jeon, M. Anisetti, D. Kim, V. Bellandi, E. Damiani, and J. Jeong, "Fuzzy rough sets hybrid scheme for motion and scene complexity adaptive deinterlacing," *Image Vis. Comput.*, vol. 27, no. 4, pp. 425–436, 2009.

[3] Y. Tao, G. Kollios, J. Considine, F. Li, and D. Papadias, "Spatio-temporal aggregation using sketches," in *Proc. 20th Int. Conf. Data Eng.*, 2004, pp. 214–225.

[4] H. Jeung, Q. Liu, X. Zhou, and H. T. Shen, "A hybrid prediction model for moving objects," in *Proc. ICDE*, 2008, pp. 70–79.

[5] T. Cheng and J. Wang, "Integrated spatio-temporal data mining for forest fire prediction," *Trans. GIS*, vol. 12, no. 5, pp. 591–611, 2008.

[6] H. Jeung, M. L. Yiu, C. S. Jensen, and X. Zhou, "Path prediction and predictive range querying in road network databases," *VLDB J.*, vol. 19, no. 4, pp. 585–602, 2010.

[7] W. Boulila, I. R. Farah, B. Solaiman, H. Ben Ghézala, and K. S. Ettabaa, "A data mining based approach to predict spatiotemporal changes in satellite images," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 13, no. 3, pp. 386–395, 2011.

[8] L. Bai, L. Yan, and Z. M. Ma, "Determining topological relationship of fuzzy spatiotemporal data integrated with XML twig pattern," *Appl. Intell.*, vol. 39, no. 1, pp. 75–100, 2013.

[9] Z. Xu *et al.*, "From Latency, through Outbreak, to decline: Detecting different states of emergency events using Web resources," *IEEE Trans. Big Data*, vol. 4, no. 2, pp. 245–257, Jun. 2016.

[10] J. Le Coz, B. Renard, F. Branger, R. Le Boursicaud, and L. Bonnifait, "Combining hydraulic knowledge and uncertain gaugings in the estimation of hydrometric rating curves: A Bayesian approach," *J. Hydrol.*, vol. 509, pp. 573–587, Feb. 2014.

[11] H. McMillan, J. Freer, T. Krueger, M. Clark, and F. Pappenberger, "Impacts of uncertain river flow data on rainfall-runoff model calibration and discharge predictions," *Hydrol. Processes*, vol. 24, no. 10, pp. 1270–1284, 2010.

[12] S. Liu, J. Forrest, and Y. Yang, "A brief introduction to grey systems theory," *Grey Syst., Theory Appl.*, vol. 2, no. 2, pp. 89–104, 2012.

[13] Y. Yang, S. Liu, and R. John, "Uncertainty representation of grey numbers and grey sets," *IEEE Trans. Cybern.*, vol. 44, no. 9, pp. 1508–1517, Sep. 2014.

[14] C. Hamzacebi and H. A. Es, "Forecasting the annual electricity consumption of Turkey using an optimized grey model," *Energy*, vol. 70, pp. 165–171, Jun. 2014.

[15] G. Sun, X. Guan, Z. Zhou, and X. Yi, "Grey relational analysis between hesitant fuzzy sets with applications to pattern recognition," *Expert Syst. Appl.*, vol. 92, pp. 521–532, Feb. 2018.

[16] X. Zhang, F. Jin, and P. D. Liu, "A grey relational projection method for multi-attribute decision making based on intuitionistic trapezoidal fuzzy number," *Appl. Math. Model.*, vol. 37, no. 5, pp. 3467–3477, 2013.

[17] H. Kuang, M. A. Bashar, K. W. Hipel, and D. M. Kilgour, "Grey-based preference in a graph model for conflict resolution with multiple decision makers," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 45, no. 9, pp. 1254–1267, Sep. 2015.

[18] L. Bai, L. Yan, and Z. Ma, "Interpolation and prediction of spatiotemporal data based on XML integrated with grey dynamic model," *Int. J. Geo-Inf.*, vol. 6, no. 4, p. 113, 2017.

[19] L. Bai, L. Yan, and Z. M. Ma, "Fuzzy spatiotemporal data modeling and operations in XML," *Appl. Artif. Intell.*, vol. 29, no. 3, pp. 259–282, 2015.

[20] Z. M. Ma and L. Yan, "Modeling fuzzy data with XML: A survey," *Fuzzy Sets Syst.*, vol. 301, pp. 146–159, Oct. 2016.

[21] Z. Ma, L. Bai, Y. Ishikawa, and L. Yan, "Consistencies of fuzzy spatiotemporal data in XML documents," *Fuzzy Sets Syst.*, vol. 343, pp. 97–125, Jul. 2017, doi: 10.1016/j.fss.2017.03.009.

[22] T. Li and Z. M. Ma, "Object-stack: An object-oriented approach for top-*k* keyword querying over fuzzy XML," *Inf. Syst. Frontiers*, vol. 19, no. 3, pp. 669–697, 2017.

[23] S. Clark, "Traffic prediction using multivariate nonparametric regression," *J. Transp. Eng.*, vol. 129, no. 2, pp. 161–168, Mar. 2003.

[24] J. Haworth and T. Cheng, "Non-parametric regression for space–time forecasting under missing data," *Comput., Environ. Urban Syst.*, vol. 36, no. 6, pp. 538–550, 2012.

[25] W. Mathew, R. Raposo, and B. Martins, "Predicting future locations with hidden Markov models," in *Proc. ACM Conf. Ubiquitous Comput.*, 2012, pp. 911–918.

[26] J. Yang, J. Xu, N. Zheng, Y. Chen, and M. Xu, "Predicting next location using a variable order Markov model," in *Proc. 5th ACM SIGSPATIAL Int. Workshop GeoStreaming*, 2014, pp. 37–42.

[27] S. Qiao, D. Shen, X. Wang, N. Han, and W. Zhu, "A self-adaptive parameter selection trajectory prediction approach via hidden Markov models," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 1, pp. 284–296, Feb. 2015.

[28] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 802–810.

[29] X. Ma, Z. Dai, Z. He, J. Ma, Y. Wang, and Y. Wang, "Learning traffic as images: A deep convolutional neural network for large-scale transportation network speed prediction," *Sensors*, vol. 17, no. 4, p. 818, 2017.

[30] Y. Chen, L. Shu, and L. Wang, "Poster abstract: Traffic flow prediction with big data: A deep learning based time series model," in *Proc. IEEE Conf. Comput. Commun.*, May 2017, pp. 1–2.

[31] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: A deep learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 865–873, Apr. 2015.

[32] M. Morzy, "Mining frequent trajectories of moving objects for location prediction," in *Proc. Int. Workshop Mach. Learn. Data Mining Pattern Recognit.*, 2007, pp. 667–680.

[33] P.-R. Lei, T.-J. Shen, I.-J. Su, and W.-C. Peng, "Exploring spatial-temporal trajectory model for location prediction," in *Proc. 12th IEEE Int. Conf. Mobile Data Manage. (MDM)*, Jun. 2011, pp. 58–67.

[34] M. Zolotukhin, E. Ivannikova, and T. Hämäläinen, "Novel method for the prediction of mobile location based on temporal-spatial behavioral patterns," in *Proc. IEEE Int. Conf. Inf. Sci. Technol. (ICIST)*, Mar. 2013, pp. 761–766.

[35] X. Ma and Z.-B. Liu, "The kernel-based nonlinear multivariate grey model," *Appl. Math. Model.*, vol. 56, pp. 217–238, Apr. 2018.

[36] B. Zeng, W. Meng, and M. Tong, "A self-adaptive intelligence grey predictive model with alterable structure and its application," *Eng. Appl. Artif. Intell.*, vol. 50, pp. 236–244, Apr. 2016.

[37] Y. Wei and D.-H. Hu, "Deficiency of the smoothness condition and its remedy," *Syst. Eng.-Theory Pract.*, vol. 29, no. 8, pp. 165–170, 2009.

[38] J. Liu, S. Liu, N. Zhang, and Z. Fang, "New strengthening buffer operators based on adjustable intensity and their applications," *J. Grey Syst.*, vol. 26, no. 3, pp. 117–126, 2014.

[39] L.-F. Wu, S.-F. Liu, D.-L. Liu, T.-X. Yao, and W. Cui, "Non-homogenous discrete grey model with fractional-order accumulation," *Neural Comput. Appl.*, vol. 25, no. 5, pp. 1215–1221, 2014.

[40] Y. H. Wang, Q. Liu, W. Cao, X. Li, and J. Tang, "Optimization approach of background value and initial item for improving prediction precision of GM(1,1) model," *J. Syst. Eng. Electron.*, vol. 25, no. 1, pp. 77–82, 2014.

[41] Z.-X. Wang and P. Hao, "An improved grey multivariable model for predicting industrial energy consumption in China," *Appl. Math. Model.*, vol. 40, no. 11, pp. 5745–5758, 2016.

[42] C.-J. Chang, D.-C. Li, C.-C. Chen, and Y.-H. Huang, "A novel gray forecasting model based on the box plot for small manufacturing data sets," *Appl. Math. Comput.*, vol. 265, pp. 400–408, Aug. 2015.

[43] Y. Wang, Y. Dang, S. Liu, and Y. Li, "An approach to increase prediction precision of GM(1,1) model based on optimization of the initial condition," *Expert Syst. Appl.*, vol. 37, no. 8, pp. 5640–5644, 2010.

[44] Y.-C. Hu and P. Jiang, "Forecasting energy demand using neural-network-based grey residual modification models," *J. Oper. Res. Soc.*, vol. 68, no. 5, pp. 556–565, 2017.

[45] T. X. Yao and S. F. Liu, "Characteristics and optimization of discrete GM (1, 1) model," *Syst. Eng., Theory Pract.*, vol. 29, no. 3, pp. 142–148, 2009.

[46] L. Wu, S. Liu, and Y. Wang, "Grey Lotka–Volterra model and its application," *Technol. Forecasting Social Change*, vol. 79, no. 9, pp. 1720–1730, 2012.

[47] L. Wu, S. Liu, S. Yan, D. Liu, and L. Yao, "Grey system model with the fractional order accumulation," *Commun. Nonlinear Sci. Numer. Simul.*, vol. 18, no. 7, pp. 1775–1785, 2013.

[48] N.-M. Xie and S.-F. Liu, "Discrete grey forecasting model and its optimization," *Appl. Math. Model.*, vol. 33, no. 2, pp. 1173–1186, 2009.

[49] J. Ye, Y. Dang, and B. Li, "Grey-Markov prediction model based on background value optimization and central-point triangular whitenization weight function," *Commun. Nonlinear Sci. Numer. Simul.*, vol. 54, pp. 320–330, Jan. 2018.

[50] U. Kumar and V. K. Jain, "Time series models (Grey-Markov, Grey Model with rolling mechanism and singular spectrum analysis) to forecast energy consumption in India," *Energy*, vol. 35, no. 4, pp. 1709–1716, 2010.

[51] H.-L. Wong and J.-M. Shiu, "Comparisons of fuzzy time series and hybrid Grey model for non-stationary data forecasting," *Appl. Math. Inf. Sci.*, vol. 4, pp. 409–416, 2012.

[52] W. Froelich and J. L. Salmeron, "Evolutionary learning of fuzzy grey cognitive maps for the forecasting of multivariate, interval-valued time series," *Int. J. Approx. Reasoning*, vol. 55, no. 6, pp. 1319–1335, 2014.

[53] Y.-H. Wang, "Using neural network to forecast stock index option price: A new hybrid GARCH approach," *Qual. Quantity*, vol. 43, no. 5, pp. 833–843, 2009.

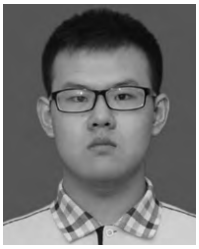
[54] X.-Y. Zeng, L. Shu, J. Jiang, and G.-M. Huang, "Triangular fuzzy series forecasting based on grey model and neural network," *Appl. Math. Model.*, vol. 40, no. 3, pp. 1717–1727, 2016.

[55] *Typhoon Data, Typhoon Nesat*. Accessed: Jul. 2017. [Online]. Available: http://typhoon.weather.com.cn/

[56] L. Bai, Z. Jia, and J. Liu, "Transformation of fuzzy spatiotemporal data from XML to object-oriented database," *Earth Sci. Inform.*, vol. 11, no. 3, pp. 449–461, 2018.

CHENGJIA SUN was born in Anshan, Liaoning, China, in 1996. She is currently pursuing the Degree with the School of Computer and Communication Engineering, Northeastern University, Qinhuangdao Campus. Her main studies include uncertain spatiotemporal data and network security.

LUYI BAI received the Ph.D. degree from Northeastern University, China. He is currently an Associate Professor with Northeastern University at Qinhuangdao, China. He has authored papers in several journals such as *Integrated Computer-Aided Engineering*, *Fuzzy Sets and Systems*, *Applied Intelligence*, and *Applied Artificial Intelligence*. His current research interests include uncertain databases, fuzzy spatiotemporal extensible markup language data management, and knowledge graph. He is a member of CCF.



LEI KANG is currently pursuing the Degree with the School of Computer and Communication Engineering, Northeastern University, Qinhuangdao Campus. He participated in relevant research on uncertain spatiotemporal data.

SHANHAO LI is currently pursuing the bachelor's degree in computer science with Northeast University at Qinhuangdao, China. His major interest is in artificial intelligence.

NAN LI is currently pursuing the bachelor's degree with the School of Computer and Communication Engineering, Northeastern University at Qinhuangdao, China. Her main research interests include fuzzy spatiotemporal database.