



Evolutionary modeling-based approach for model errors correction

S. Q. Wan^{1,2}, W. P. He³, L. Wang⁴, W. Jiang⁵, and W. Zhang⁶

¹Yangzhou Meteorological Office, Yangzhou, China

²Department of Physics, Yangzhou University, Yangzhou, China

³National Climate Center, China Meteorological Administration, Beijing, China

⁴School of Computer Science, Beijing Institute of Technology, Beijing, China

⁵Jiangsu Meteorological Bureau, Nanjing, China

⁶Department of Atmospheric Sciences, Lanzhou University, Lanzhou, China

Correspondence to: W. P. He (wenping_he@163.com)

Received: 3 December 2011 – Revised: 20 July 2012 – Accepted: 24 July 2012 – Published: 14 August 2012

Abstract. The inverse problem of using the information of historical data to estimate model errors is one of the science frontier research topics. In this study, we investigate such a problem using the classic Lorenz (1963) equation as a prediction model and the Lorenz equation with a periodic evolutionary function as an accurate representation of reality to generate “observational data.”

On the basis of the intelligent features of evolutionary modeling (EM), including self-organization, self-adaptive and self-learning, the dynamic information contained in the historical data can be identified and extracted by computer automatically. Thereby, a new approach is proposed to estimate model errors based on EM in the present paper. Numerical tests demonstrate the ability of the new approach to correct model structural errors. In fact, it can actualize the combination of the statistics and dynamics to certain extent.

1 Introduction

Monthly, seasonal, annual and inter-annual climatic predictions became the next targets of the frontier research of atmospheric sciences since the successful implementation of short-term weather forecasts. However, with the discovery of chaos, it is well known that the climate system is a complex nonlinear system, and climate prediction is a great challenge to researchers. Moreover, inherent defects of statistical methods and dynamical methods result in significant uncertainties of their applications in climate prediction. Statistical methods are mainly used to search prediction clues from historical observational data, based on the historical behavior

and statistical regularities of the climate system. The results of statistical model are sometimes satisfactory, but such statistical regularities and the prediction models derived from such statistical methods are often unstable due to the effects of nonlinearity of the climate system (Gu, 1958).

The prediction problem based on dynamical methods can be regarded as a problem of initial value of differential equations. It implies that it is necessary to ensure sufficiently precise initial values and boundary conditions as well as the accuracy of the prediction model in order to obtain a reliable prediction. In fact, these conditions cannot be satisfied completely. Despite the continuous improvement and optimization of numerical model and data assimilation systems, it is difficult to improve the time limit of weather forecasts beyond two weeks (Lorenz, 1965, 1969) due to the complexity of the atmospheric motion. Therefore, it is crucial to develop alternative ways to improve the capability of the climate prediction model (Schubert, 1985; Vannitsem and Toth, 2002; Chou, 2003a, b; Li and Ding, 2011).

The dynamics-statistics approach is an important step for the improvement of climate prediction models. Extensive research has been carried out using a combination of statistics and dynamics. Qiu and Chou (1988) noted that the observational data could be regarded as sufficiently precise solutions for atmospheric models, and could be used to correct the errors of model parameters by solving an inverse problem. Cao (1993) and Feng (2004) suggested that the atmospheric motion is an irreversible process, and they introduced a memory function that can make use of historical observational data to deduce the self-memorization equation for the atmospheric motion including multi-time observational data.

Chou (2007) argued that past daily weather changes, especially the recent evolutionary status of the atmosphere, contain the information of numerical model errors. It should be used to correct the model errors. Some related mathematical and numerical issues in the geophysical fluid dynamics and climate dynamics have been discussed (Li and Wang, 2008). These methods have shown to be useful techniques for climate prediction in numerical experiments and applications. It is an evidence of the effectiveness of the dynamics-statistics approach.

In fact, the idea of the correction of the model errors is old. Many studies provided some objective methods of estimating model errors, but their use for improving model performance, is comparatively small. Examples are given by Schubert (1985), Klinker and Sardeshmukh (1992). D'Andrea and Vautard (2000) proposed a methodology for the correction of systematic errors in a simplified atmospheric general circulation model, and confirmed that this improvement actually stems from the flow dependence of the model error. Such flow dependence was found in the Euro-Atlantic sector, while similar attempts to establish this relation in other sectors of the globe failed. Vannitsem and Toth (2002) investigated the short-term dynamics of model errors by means of numerical analysis of the Lorenz (1984) low-order atmospheric system, and found that the short-term mean square error evolution is mainly characterized by an initial quadratic or linear behavior, depending on the dynamical properties of the model error source terms. Also, there are studies of the impact of model errors associated with parameter errors. Duan and Zhang (2010) investigated the effect of initial errors and model parameter errors on a significant "spring predictability barrier" (SPB) for El Niño events. They inferred that initial errors, rather than model parameter errors, may be the dominant source of uncertainties that cause a significant SPB for El Niño events. Mu et al. (2010) suggested that conditional nonlinear optimal perturbation approach (CNOP) can be applied to estimate model parameter errors for El Niño events, but did not consider other kinds of model errors at present. Despite this, it is expected that CNOP will play an important role in the studies of atmospheric and oceanic sciences.

However, most of the works mentioned above remain in the theoretical stage, and many issues need to be addressed for their application in weather forecasts (Chou et al., 2007). For example, traditional methods for solving inverse problems of differential equations face one essential problem of ill-posed characteristic, such as the instability of approximate solutions. The models describing the complex climate system can only approximately describe the major dynamic processes of atmospheric motion. Therefore, not all the details are described in these prediction models. In fact, climate change can be viewed as a long-range evolutionary process with self-adaptation. So, it is unnecessary to describe all the characteristics of the problem in detail, and all it requires is

a solution in accordance with the evolutionary laws of the nature.

In the case that only limited information is known on a dynamic system, a possibility is to replace human intelligence with computational intelligence in some steps of the traditional modeling, including the development of assumptions, the construction and the calculation of the model. This is the idea of so-called evolutionary modeling (EM; Cao et al., 2000). Evolutionary computation is a mathematical modeling approach based on the evolutionary laws of nature (Back et al., 1997; Cao et al., 2000). In EM, simple coding techniques can be used to represent a variety of complex mathematical structures. The approximate solutions of inverse problem can be automatically searched by computer programs. This step is iterated for some number of generations until the termination criterion of the run has been satisfied. The characteristics of evolutionary algorithms are self-organizing, self-adaptive and self-learning. Natural selection, namely survival of the fittest, and evolution strategies provide many conveniences and advantages for solving inverse problems of complex differential equations.

In the present work, the problem of how to use combination of the dynamics and statistics to correct model errors has been studied. On the basis of the EM method, a new approach is proposed to correct model errors. Using the new approach, dynamic information of model errors can be extracted from historical observational data. The results of numerical experiments based on the Lorenz (1963) model have been preliminarily validated in terms of their ability to correct model errors. It must be noted that the model error studied here mainly refers to the dynamic structural error in a prediction model.

2 Algorithm of EM

EM is a calculation strategy and method on the basis of the principles of biological natural selection and genetic inheritance (Back et al., 1997; Cao et al., 2000). Complex problems can be solved by means of computer simulations of the natural evolutionary process in EM. In accordance with rules of natural evolution, such as the "survival of the fittest", the complexity of the computing and solution can be controlled in EM. Evolutionary computation provides an effective approach to deal with those problems (Holland, 1975), which are complex and intractable problems by traditional mathematics. Generally, differential equations can be expressed as binary trees in EM (Fig. 1). In these binary trees, each tree structure uniquely corresponds to a complex or simple function within the solution space of the problem. These functions generally consist of arithmetic operations, logic operations, variables, constants, and elementary functions, such as sine, cosine, and exponential function. The tree structure can be changed dynamically by using genetic operators, such as crossover and mutation. The evolution of these tree structures continues from one generation to another generation until the

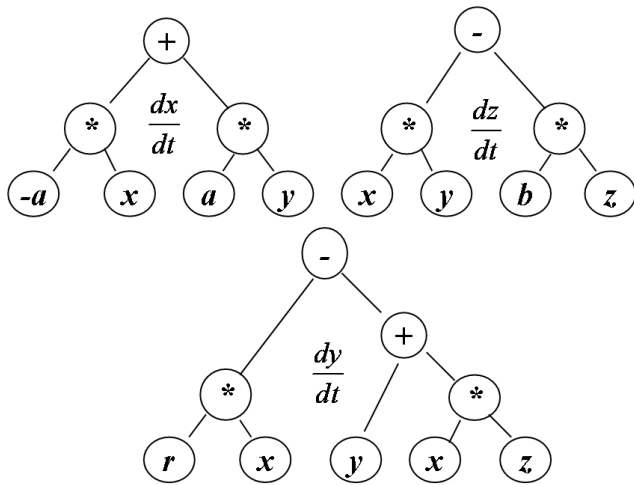


Fig. 1. Hierarchical structure diagram of binary tree of the Lorenz model.

termination criterion of the run has been satisfied. The complexity of a function can be controlled by setting the depths of the layers of a binary tree. The detailed algorithm of EM can be found in Cao et al. (2000). As an example, the classic three-dimensional Lorenz (1963) model can be written as the form:

$$\begin{cases} \frac{dx}{dt} = ax + ay = L_x(x, y) \\ \frac{dy}{dt} = rx - (y + xz) = L_y(x, y, z) \\ \frac{dz}{dt} = xy - bz = L_z(x, y, z). \end{cases} \quad (1)$$

Here, the three parameters, a , r and b , are positive, and are called the Prandtl number, the Rayleigh number and a physical proportion related to convective scale, respectively. In the present study, the three parameter values are 10.0, 28.0 and $8/3$, respectively. When $r > 24.74$, the Lorenz system exhibits chaotic behavior. The binary tree form of the Lorenz model is shown in Fig. 1.

In evolutionary algorithms, an individual mathematical model structure corresponds to a unique binary tree (Fig. 1). Three popular methods can be used to implement evolutionary computation, namely, genetic algorithm (GA), gene expression programming (GEP; Ferreira, 2001), and genetic programming (GP; Saiedian, 1997). The GP and GEP methods will be selected to implement EM program in the present study.

3 Correction scheme for model errors

Numerical prediction models are the dynamic equations constructed according to the basic laws of atmospheric and climate motion as currently understood. Undoubtedly, these models use simplified and approximate equations. It can only describe main features of atmospheric motion. In fact, the

problem is that the disturbance terms have to be overlooked, or certain important dynamic elements are not included due to the limitation of the existing knowledge. One of the current measures for correcting errors is a post-processing strategy for model prediction results, i.e. statistically correct on the prediction results. However, the omission of dynamic characteristics of the model error affects the quality of the results of the statistical correction.

3.1 Description of the inverse problem

An inverse problem is a general framework that is used to convert observed measurements into information about a physical object or system that we are interested in. For example, if we have measurements of the Earth’s gravity field, then we might ask the question: “given the data that we have available, what can we say about the density distribution of the Earth in that area?” The solution to this problem (i.e. the density distribution that best matches the data) is useful because it generally tells us something about a physical parameter that we cannot directly observe (Keller, 1976; http://en.wikipedia.org/wiki/Inverse_problem).

In contrast, with a forward problem, an inverse problem of differential equations is based on existing results to deduce the cause of the results. In other words, on the basis of the solution or partial solution of one differential equation, the unknown components of the equation can be deduced. In terms of differential equations, Leverentiev and his colleagues argued that the inverse problem can be defined as follows: an inverse problem for a partial differential equation is any problem involving the determination of the coefficients or right-hand side of a partial differential equation on the basis of certain functionals of the solution of the equation (Leverentiev et al., 2003). Without loss of generality, the general form of differential equations can be written as

$$L \cdot u(x, y, t) = f(x, y, t), (x, y) \in \Omega, t \in (0, \infty). \quad (2)$$

Here, L is the differential operator, and $u(x, y, t)$ is the solution of the differential equation. The function $f(x, y, t)$ is the right-hand side source term of the equation. When L is unknown, this is called an inverse problem of operator identification; furthermore, when the right-hand source term $f(x, y, t)$ is unknown, this is called a source-term inverse problem. The ill-posedness and nonlinearity of an inverse problem make the related theories and solutions more complex and difficult than those of a forward problem.

In order to investigate the inverse problem based on the use of the information of historical data to estimate model errors, Eq. (1) is regarded as an approximate model derived from practice, while the corresponding true model is represented by

$$\begin{cases} \frac{dx}{dt} = L_x(x, y) \wedge E_x(x, y, z) \\ \frac{dy}{dt} = L_y(x, y, z) \wedge E_y(x, y, z) \\ \frac{dz}{dt} = L_z(x, y, z) \wedge E_z(x, y, z). \end{cases} \quad (3)$$

Here, $E(x, y, z)$ is the potential unknown error term, which is a composite function, such as trigonometric, exponential and power functions. The symbol “ \wedge ” represents connecting nodes. Generally, the symbol can only be basic arithmetic operators, such as “+”, “-”, “ \times ”, “ \div ”. If $L(x, y, z)$ is the main dynamic structure of the function, “ \wedge ” can be determined. In this case, the symbol “ \wedge ” usually is “+”. Even for a simple dynamic equation, such as Eq. (1), it is difficult to determine the specific mathematical expression of $E(x, y, z)$ using traditional methods. On the basis of the definition of the inverse problem, determining the error term $E(x, y, z)$ in Eq. (3) is a classic inverse problem, namely, solving the source term on the right-hand side.

The limitation of those traditional methods is that model structure must be determined in advance, and then model parameters can be estimated. However, it generally depends on personal experience to construct model structure. It is a complicated problem, especially when the data space takes on the form of hyper-surfaces with multi-parameters and multi-variables.

3.2 Error correction scheme of evolutionary algorithm-based prediction model

The main steps for correcting prediction model error terms through historical data are as follows: an approximate prediction model and an accurate prediction model are constructed, respectively. The solutions of the approximate model are regarded as the prediction results of a specific system, and the solutions of the accurate model are regarded as “observational data”. Subsequently, the EM algorithm is used to automatically search for the error term. During modeling, a large ensemble of mathematical model structures is randomly formed by means of the search in the basic function library using the evolution program. Simultaneously, the parameters of the individual model structures will be continuously estimated and optimized by the parameter optimization module. According to certain criteria, poor individual model structures will be eliminated, and good model structures will be retained, which can be continuously used to EM of the next generation. Finally, several mathematical models with relatively minimal errors will be obtained.

Firstly, in Eq. (3), for example, the function $L(x, y, z)$ must be unchanged in the EM. Then, the function will be expressed as the binary tree hierarchy (Fig. 1), which can be represented in a certain program code, and then can be placed into the evolutionary process. Obviously, obtaining the expression of the error function $E(x, y, z)$ is an inverse problem

that solves the right-hand side of the differential equation. Here, $L(x, y, z)$ is the approximate model, and the correction scheme of the prediction equation is as follows.

1. Constructing Gene(L), the binary tree of the main function $L(x, y, z)$.
2. Specifying randomly the node symbols and the constants contained in the error terms $E(x, y, z)$, and the nodes include {+, -, \times , \div , sin, cos, ln, exp} in the present study. The populations Pop with a certain size will be generated randomly.
3. Forming several binary tree structures Gene_{*i*}(E) of the error functions using the evolutionary algorithm GP and GEP. Meanwhile, Gene_{*i*}(E) will concurrently evolve with the main binary tree Gene(L) by applying evolutionary operations, including selection, hybridization and mutation. By controlling the depths of the binary tree layers of Gene_{*i*}(E), several complex-controllable composite binary trees (i.e. the revised function) can be obtained.
4. Computing and evaluating the adaptive values of these composite binary trees in combination with “observational data”. Retaining the superior individuals Gene_{*i*}(E) in the generation, and then these superior individuals will be used as EM of the next generation.
5. Repeating the above procedures until a new function body meets the predetermined conditions (i.e. the termination condition).

3.3 Numerical experiments

In order to test the performance of the error correction scheme presented in this study, it is necessary to generate “observational data”. To deal with this issue, a classic three-dimensional Lorenz model is regarded as an inaccurate prediction equation $L(x, y, z)$. The Lorenz model with an error term $E_x = 5 \sin(\sin(x))$ is regarded as an accurate prediction equation, and connecting node $\wedge = “+”$, as follows:

$$\begin{cases} \wedge = “+” \\ E_x(x, y, z, t) = 5 \sin(\sin(t)) \\ E_y(x, y, z, t) = 0 \\ E_z(x, y, z, t) = 0. \end{cases}$$

Thus, Eq. (3) can be written as

$$\begin{cases} \frac{dx}{dt} = L_x(x, y) + 5 \sin(\sin(t)) \\ \frac{dy}{dt} = L_y(x, y, z) \\ \frac{dz}{dt} = L_z(x, y, z). \end{cases} \quad (4)$$

In Eq. (4), the integral initial values of x_0 , y_0 and z_0 are 1.19, 3.08 and 21.65, respectively. The integral step is 0.1. The first

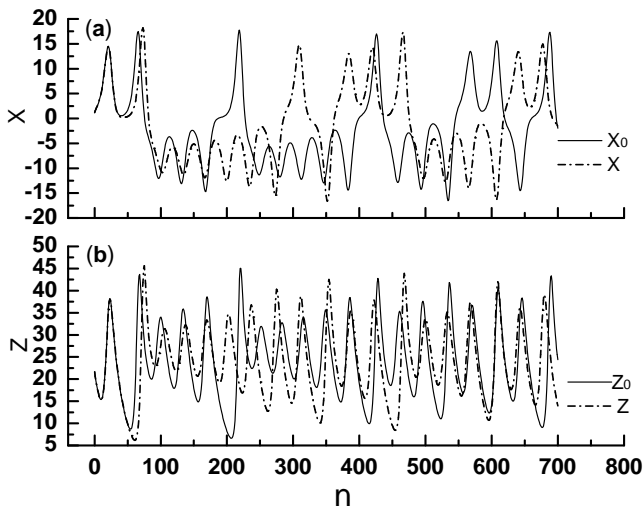


Fig. 2. The “observed value” (x_0, z_0) and the corresponding prediction data (x, z) of the prediction model, where the sample size is 700. (a) Variable x in the Lorenz model; and (b) variable z in the Lorenz model.

700 integral steps of Eq. (4) are regarded as “observed data” (solid line in Fig. 2). The corresponding prediction data can be generated by the classic Lorenz model (Eq. 1), and the solution of the approximation model has been shown in Fig. 2 (dash line). Comparing the prediction data with the observational data, it is clear that there is some small difference between the prediction data and the observational data when the integral time is less than about 180 steps. Thereafter, the differences gradually become larger, and even the phases of evolution of the prediction data and the observational data are sometimes in opposition.

In accordance with the correction scheme, the GP and GEP are used in the modeling approach. The details of the related parameters are already listed in Table 1, and a relative error function is used as the fitness function,

$$\text{fitness}(f(x)) = \frac{100.0}{\sum_{i=1}^N (x_i - x_{i'})^2}. \quad (5)$$

Equation (5) is called the evaluation function, which is a quantitative indicator of the quality of model, and also provides the termination condition of the EM program. Here, x_i is the “observational data”, and $x_{i'}$ is the solutions of the revised model by EM. It is the most crucial element of the evolutionary algorithm for the problem to be solved. The selection of the fitness function has a significant impact on the convergence speed and results of the evolutionary algorithm. The evaluation function (Eq. 5) is comparatively consistent with the law of evolution, the “survival of the fittest”. It retains a suitable adaptive range and can create a state space that can be addressed by the evolutionary algorithm. A suitable evaluation function can improve the convergence

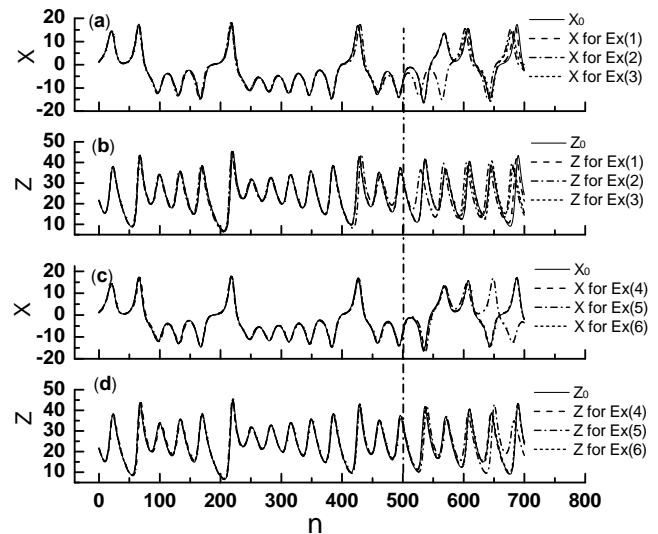


Fig. 3. Error correction based on the first component $L_x(x, y)$ using EM. (a) and (b) are the fitting and prediction data of the corrected models using the error function $E_x^{(1)}, E_x^{(2)}$ and $E_x^{(3)}$, respectively. (c) and (d) are the same as (a) and (b), but for $E_x^{(4)}, E_x^{(5)}, E_x^{(6)}$, respectively.

velocity of the evolutionary algorithm and avoid excessively rapid or slow evolution of populations. Thus, it can ensure to obtain several reliable solutions with high quality. A few prediction equations have been listed in Table 2 obtained by the error correction scheme.

Generally, it is difficult to have the knowledge on the forms and positions of the error terms. The error corrections have been implemented for the three functions $L_x(x, y), L_y(x, y, z), L_z(x, y, z)$, respectively. Based on the results of the EM, seven representative error functions are shown in Table 2 for the error corrections on the first function $L_x(x, y)$ in the Lorenz model (Eq. 1). The error function $E_x^{(0)}$ obtained by EM is identical to the real error term because the function $E_x^{(0)}$ can be transformed into the original error function. Indeed, it is no doubt that the probability for obtaining an identical error function is very low in climate models based on the existing technology. The reason is that the variables and physical mechanisms of the atmospheric motion are more complicated than those of the current example. The other six error functions can be divided into two categories. The first three error functions only contain time variable t , and the style of these functions is similar to that of the real error function. In contrast, there are more than one variable contained in the other three error functions. It is easy to find that most of these error functions obtained by EM are similar to the given error term E_x , namely, trigonometric functions are the main form in these error functions.

The fitting curves and prediction curves are shown in Fig. 3. The sample size of the training data is 500, and that of the prediction data is 200. Obviously, they are all good

Table 1. Parameters in the EM procedure.

Setting	GP	GEP
Population size	100	100
Generation	100–1000	500–1000
Selection method	Elite	Elite
Functions	+, −, *, /, sin, cos, ln, exp	+, −, *, /, sin, cos, ln, exp
Random constant range	0–100	0–100
Crossover prob.	75 %	75 %
Mutation prob.	4.4 %	4.4 %
Minimum level of binary tree	3	/
Maximum level of binary tree	5	/
Head length	/	20
Observations size	700	700

Table 2. Error functions and errors of the equations obtained by EM (for variable x).

Object revised	Error function		Expression of error function	Errors (based on x)	
	function name	Including variables		Fitness	Prediction
dx/dt	$E_x^{(0)}$	t	$10 \sin(0.5 \sin(t)) \cos(0.5 \sin(t)) + 5 \sin(\sin(t))$	0	0
	$E_x^{(1)}$	t	$2 \sin(t) + 1$	0.26	1.37
	$E_x^{(2)}$	t	$3.248 \sin(\cos(t))$	1.50	8.41
	$E_x^{(3)}$	t	$\frac{4.584}{t+2}$	0.56	3.14
	$E_x^{(4)}$	x, y	$5 \cos(x) \cos^2\left(\frac{\cos(y)-0.5}{y}\right)$	0.038	0.02
	$E_x^{(5)}$	x, t	$\sin(\sin(\sin(y+t))) + 1.554$	0.02	1.68
$E_x^{(6)}$	x, z	$\ln(\cos(e^x) + \cos(\cos(z+0.5))) + 5$	0.03	0.06	

fitting results for different error functions by EM, and the prediction errors are largely reduced by the correction before reaching the first 150 steps. Even though the prediction error becomes slightly larger in the final 50 steps, the overall varying trend remains unaffected. It is also indicated that there exists a time limit in the correction efficiency of these error functions. In Fig. 3a and b, the adaptive value calculated by Eq. (5) is less than 2, and the maximum of the prediction error is only 8.41. Obviously, it is a successful error correction. Figure 3c and d show the results of the revised equation corrected by three relatively more complex error items: $E_x^{(4)}$, $E_x^{(5)}$ and $E_x^{(6)}$. They contain more than one variable, including components (x, y, z) and t . Even though they are more complicated than the original error function, the effects of the error correction are still satisfactory. Except for the prediction result of the revised equation by $E_x^{(5)}$, the average prediction errors are very small. The corrected prediction results are similar for the prediction of the variables y and z by error correction on the first function $L_x(x, y)$ in Eq. (4). It

shows the ability of EM to reduce model error using historical observational data.

The results of the error corrections on the second function $L_y(x, y, z)$ of the Lorenz model are listed in Table 3. The original error term E_x is in the first function $L_x(x, y)$ in Eq. (4), but the error corrections on the second function $L_y(x, y, z)$ can achieve the effect of reducing prediction error by the automatic searching process of the EM and the correction effect is almost the same as with the function $L_x(x, y)$. Apart from the error functions $E_y^{(1)}$ and $E_y^{(2)}$, the fitting values of other four revised functions are less than 1.0. The fitting value of $E_y^{(1)}$ is the largest in the six cases; however, it can be seen that the prediction error also can be largely reduced (Fig. 4). For the other three error functions $E_y^{(4)}$, $E_y^{(5)}$, $E_y^{(6)}$, the average prediction errors of the revised function are very small. The prediction errors of the revised model by the error function $E_y^{(4)}$, $E_y^{(5)}$, $E_y^{(6)}$ become relatively larger when the integration time is greater than about 150 units. The results further indicate that there exist time limits in the correction efficiency by these error functions. It

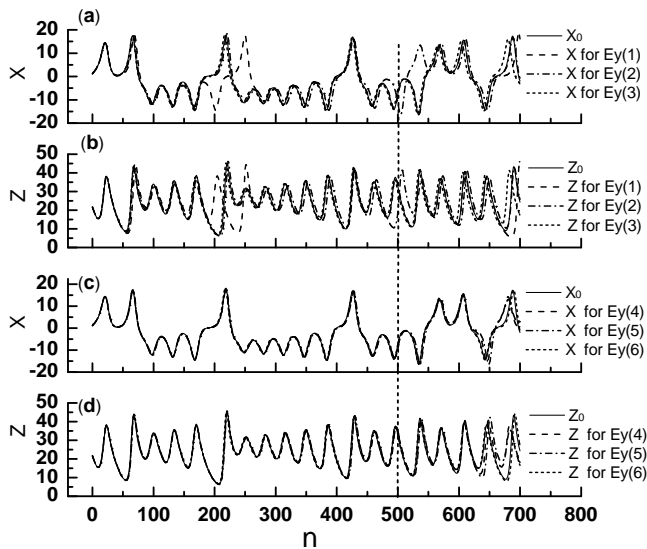


Fig. 4. Error correction based on the second component $L_y(x, y, z)$ using EM. (a) and (b) are the fitting and prediction data of the corrected models using the error function $E_y^{(1)}, E_y^{(2)}, E_y^{(3)}$, respectively. (c) and (d) are the same as (a) and (b), but for $E_y^{(4)}, E_y^{(5)}, E_y^{(6)}$, respectively.

shows that the structural errors of a prediction model can be adjusted by correcting its different components of the model. Similar error corrections have been implemented for the third function $L_z(x, y, z)$. The results are listed in Table 3 and shown in Fig. 5. The effects of the corrections are similar with those of the other two components in the Lorenz model.

Based on these results, it can be concluded that various forms of error function can be obtained by the EM algorithm. Some of them are simple error function forms, which may be almost identical to the original error function in the inaccurate equation; however, there are also complex forms, which contain various kinds of compound functions. In certain time range, they provide similar dynamic characters. In summary, the information of the prediction error contained in the observation data can be effectively extracted by EM. EM provides a variety of possible forms of error function, which can be used to largely reduce the prediction error in a certain time limit.

4 Conclusions

The development of a numerical model largely depends on the continued improvement of the model and the quality of observational data. The dynamical methods have already become an important research direction in climate prediction. The dynamics-statistics approach is a significant and effective way to improve prediction capabilities under the conditions of existing models and data. Different from the simple

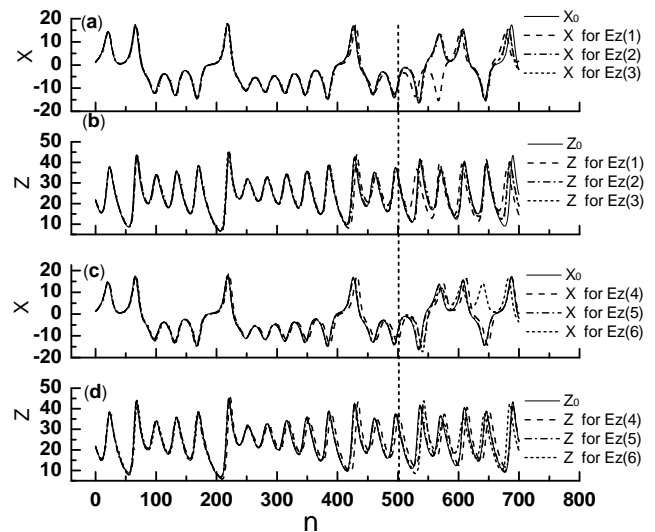


Fig. 5. Error correction based on the third component $L_z(x, y, z)$ using EM. (a) and (b) are the fitting and prediction data of the corrected models using the error function $E_z^{(1)}, E_z^{(2)}, E_z^{(3)}$, respectively. (c) and (d) are the same as (a) and (b), but for $E_z^{(4)}, E_z^{(5)}, E_z^{(6)}$, respectively.

statistical correction strategy, this paper proposes a direct correction method of model error based on the EM algorithm, which can make use of dynamic regularities inherent in historical data.

In comparison with the original prediction equation, the numerical results indicate that prediction errors are significantly reduced after the correction by the EM algorithm and the feasibility and ability of the correction scheme proposed in this paper. The present work demonstrates that, using the EM algorithm, the abundance of information present in historical data can be used to correct model errors. The present method can make full use of the advantages of dynamics and statistics.

Various forms of error function can be obtained by EM algorithm, and it will be a difficult task to choose an appropriate equation in application. An ensemble forecasting method can be used to deal with this problem. Based on the prediction errors of the model corrected by EM, the proportion of the prediction results of the different revised model can be quantitatively allocated. It can reduce the uncertain of the prediction results by the revised model as well as avoiding randomness of a single revised model.

Only a simple three-dimensional Lorenz equation is selected to study the scientific problem for model error correction in the present paper, but our work has undoubtedly paved a new way for the implementation of a combined dynamics-statistics. The results of the numerical experiment will be a great encouragement for researchers to continually study in this field, and it further demonstrates that the combination of dynamics and statistics is an important and feasible approach

Table 3. Error functions and errors of the equations obtained by EM (for variable y).

Object reviewed	Error function		Expression of error function	Errors (based on x)	
	Function name	Including variables		Fitness	Prediction
dy/dt	$E_y^{(1)}$	t	$\sin(\sin(t) + \sin(t + 2)) + 0.5$	4.58	2.45
	$E_y^{(2)}$	t	$\sqrt[4]{10t^2}$	2.28	7.60
	$E_y^{(3)}$	t	$\cos(0.5t + t \sin(t)) + 0.642$	0.96	3.74
	$E_y^{(4)}$	y, z	$e^{\frac{\sqrt{5}}{y - \sqrt{z} - \sqrt{yt}/1.5}}$	0.03	0.22
	$E_y^{(5)}$	x, y, t	$(\sin(\cos(x)) + \sqrt{yt})/5$	0.04	0.25
	$E_y^{(6)}$	x, y, z	$\cos\left(xy - \frac{1}{2\ln(2z) + y - 5}\right)$	0.02	0.08

Table 4. Error functions and errors of the equations obtained by EM (for variable z).

Object reviewed	Error function		Expression of error function	Errors (based on x)	
	Function name	Including variables		Fitness	Prediction
dz/dt	$E_z^{(1)}$	t	$\frac{e^{0.5-t}}{\ln(0.5t^2+2)}$	1.44	7.9
	$E_z^{(2)}$	t	$\frac{\sqrt{5-(0.5t-e^t)}}{e^t}$	0.87	2.29
	$E_z^{(3)}$	t	$\sin(e^{2t+0.4}) + (\sin(e^t) + 0.4)/5$	0.42	2.33
	$E_z^{(4)}$	x, y, z	$\frac{\sin(x+5/x)}{(\sin(y)+\ln(z)+5)} - t$	0.12	0.32
	$E_z^{(5)}$	x, y, z	$\frac{2}{x} \cos\left(\frac{2z \cos(0.5y)}{y}\right)$	0.02	0.04
	$E_z^{(6)}$	z	$e^{\sin(e^z)-z}$	0.03	0.22

for the improvement of prediction precision. In future work, the idea presented in this paper will be gradually applied to higher dimensional ordinary differential and partial differential equations. The ultimate goal is to apply the present method to a variety of numerical prediction models.

Currently, EM can be optimized by means of parallel computing or cloud computing. It can be realized by implementing multithread parallel computing not only for the correction of model errors of single dynamic equation but also for that of multi-equations. Based on this, the computational efficiency of EM can be improved greatly. In fact, the numerical calculations used in this paper all are based on multi-core multi-threaded algorithms, which obviously improves the computational efficiency of EM. So we believe that the present study could be applied to more complex systems for the correction of model errors including the model errors of large-scale numerical prediction models.

Acknowledgements. The authors thank the anonymous reviewers and the editor for helpful suggestions, and thank Zuojun Yu in Hawaii University for improving the English in our manuscript. This research was jointly supported by the National Natural Science Foundation of China (Grant No. 41005041), National Basic Research Program of China (973 Program) (2012CB955902), the National Natural Science Foundation of China (Grant No. 40905034), and the Special Scientific Research Fund of Meteorological Public Welfare Profession of China (Grant Nos. GYHY201106015 and GYHY201106016).

Edited by: S. Vannitsem

Reviewed by: two anonymous referees

References

- Bäck, T., Hammel, U., and Schwefel, H. P.: Evolutionary computation: Comments on the history and current state, *Evolutionary Computation*, IEEE Trans. Evolut. Comput., 1, 3–17, 1997.
- Cao, H. X.: Self-memorization Equation in Atmospheric Motion, *Sci. China*, 23, 104–112, 1993.
- Cao, H. X., Kang, L., Chen, Y., and Yu, J.: EM of systems of ordinary differential equations with genetic programming, *Genet. Program. Evol. M.*, 1, 309–337, 2000.
- Chou, J. F.: Short Term Climatic Forecast: Present Condition, Problems and Way Out (I), *Bimonthly of Xinjiang Meteorology*, 26, 1–4, 2003a (in Chinese).
- Chou, J. F.: Short Term Climatic Forecast: Present Condition, Problems and Way Out (II), *Bimonthly of Xinjiang Meteorology*, 26, 1–4, 2003b.
- Chou, J. F.: An Innovative Road To Numerical Weather Prediction – From Initial Value Problem To Inverse Problem, *Acta Meteorol. Sin.*, 65, 673–682, 2007 (in Chinese).
- D’Andrea, F. and Vautard R.: Reducing systematic errors by empirically correcting model errors, *Tellus*, 52A, 21–41, 2000.
- Duan, W. and Zhang, R.: Is model parameter error related to a significant spring predictability barrier for El-Nino event?, *Result from Theoretical Model*, *Adv. Atmos. Sci.*, 27, 1003–1013, 2010.
- Feng, G. L., Dong, W. J., Li, J. P., and Chou, J. F.: On the stability of the difference scheme in the self-memorization model, *Acta Phys. Sin.*, 53, 2389–2395, 2004 (in Chinese).
- Ferreira, C.: Gene expression programming: A new adaptive algorithm for solving problems, *Complex Syst.*, 13, 87–129, 2001.
- Gu, Z. C.: On the utilization of past data in numerical weather forecasting, *Acta Meteorol. Sin.*, 29, 176–186, 1958 (in Chinese).
- Holland, J. H.: *Adaptation in natural and artificial systems*, Ann Arbor MI, University of Michigan Press, 1975.
- Keller, J. B.: Inverse problems, *Amer. Math. Monthly*, 83, 107–118, 1976.
- Klinker, E. and Sardeshmukh, P. D.: The diagnosis of mechanical dissipation in the atmosphere from deterrnialarge-scale balance requirements, *J. Atmos. Sci.*, 49, 608–627, 1992.
- Lavrentiev, M. M., Avdeev, A. V., and Priimenko, V. I.: *Inverse Problems of Mathematical Physics*, Berlin, Walter de Gruyter, 2003.
- Li, J. P. and Chou, J. F.: Existence of atmosphere attractor, *Sci. China*, 40, 215–224, 1997.
- Li, J. P. and Ding, R. Q.: Temporal-spatial distribution of atmospheric predictability limit by local dynamical analogues, *Mon. Weather Rev.*, 139, 3265–3283, 2011.
- Li, J. P. and Wang, S. H.: Some mathematical and numerical issues in geophysical fluid dynamics and climate dynamics, *Commun. Comput. Phys.* 3, 759–793, 2008.
- Lorenz, E. N.: Deterministic nonperiodic flow, *J. Atmos. Sci.*, 20, 130–141, 1963.
- Lorenz, E. N.: A study of the predictability of a 28 variable atmospheric model, *Tellus*, 17, 321–333, 1965.
- Lorenz, E. N.: Atmospheric predictability as revealed by naturally occurring analogues, *J. Atmos. Sci.*, 26, 636–646, 1969.
- Mu, M., Duan, W., Wang, Q., and Zhang, R.: An extension of conditional nonlinear optimal perturbation approach and its applications, *Nonlin. Processes Geophys.*, 17, 211–220, doi:10.5194/npg-17-211-2010, 2010.
- Qiu, C. J. and Chou, J. F.: A new approach to improve the numerical weather prediction, *Sci. China*, 31, 1132–1142, 1988.
- Saiedian, H.: An evaluation of extended entity-relationship model, *Inform. Software Tech.*, 39, 449–462, 1997.
- Schubert, S. D.: A statistical–dynamical study of empirically determined modes of atmospheric variability, *J. Atmos. Sci.*, 42, 3–17, 1985.
- Vannitsem, S. and Toth, Z.: Short-term dynamics of model errors, *J. Atmos. Sci.*, 59, 2594–2604, 2002.
- Verhulst, P.: *Recherches mathématiques sur la loi d’accroissement de la population*, *Nouveaux Mémoires de l’Académie Royale des Sciences et des Belles-Lettres de Bruxelles*, 18, 1–32, 1848.