Article ID:1000－1638(2025)01-0001-07 **DOI**:10.13484/j.nmgdxxbzk.20250101

Data-Driven Prediction in Complex Systems of Virus Evolution and Global Warming*

LUO Liaofu[1], LÜ Jun[2]

(1. *School of Physical Science and Technology, Inner Mongolia University, Hohhot* 010021, *China*;
2. *College of Science, Inner Mongolia University of Technology, Hohhot* 010051, *China*)

Abstract: A complex system is inherently high-dimensional. Recent studies indicate that, even without complete knowledge of its evolutionary dynamics, the future behavior of such a system can be predicted using time-series data (data-driven prediction). This suggests that the essential dynamics of a complex system can be captured through a low-dimensional representation. Virus evolution and climate change are two examples of complex, time-varying systems. In this article, we show that mutations in the spike protein provide valuable data for predicting SARS-CoV-2 variants, forecasting the possible emergence of the new macro-lineage Q in the near future. Our analysis also demonstrates that carbon dioxide concentration is a reliable indicator for predicting the evolution of the climate system, extending global surface air temperature (GSAT) forecasts through 2500.

Key words: data-driven prediction; complex system; virus evolution; global warming

CLC number: N94; O231.5 **Document code**: A

Virus evolution and global warming are typical examples of complex, time-varying systems. Even without complete knowledge of their evolutionary dynamics, future trends can be predicted using time-series data (data-driven prediction). Our approach relies on the idea that high-dimensional systems often contain redundant information, and that their essential dynamics can be captured through low-dimensional representations[1]. According to Takens' theorem on delay embedding, each time-series variable in a dynamical system can be used to reconstruct a low-dimensional representation. Delay embedding enables an isomorphic reconstruction of the original system from a single

* **Received date**: 2024-12-27

Foundation item: Natural science foundation of Inner Mongolia (2024LHMS06018); The basic scientific research funding for directly affiliated universities in the Inner Mongolia(JY20250094)

Biography: LUO Liaofu (1935－), male, a native of Shexian, Anhui, professor. Major in theoretical physics and theoretical biology. E-mail: lolfcm@imu.edu.cn

Corresponding author: LÜ Jun (1973－), male, a native of Bayannur, Inner Mongolia, professor, doctor. Major in theoretical physics and theoretical biology. E-mail: lujun@imut.edu.cn

time series. Thus, low-dimensional representations can serve as generalized predictors, allowing the identification of future dynamics in complex systems[2-4].

1 Evolution of SARS-CoV-2 variants

The continuous spread of the novel coronavirus over the past four years has been driven by successive waves of SARS-CoV-2 mutations. We previously introduced a mathematical model to analyze the dynamics of COVID-19 spread[5]. However, predicting the emergence of new strains and multiple macro-lineages remains a significant challenge. Given the crucial role of spike protein mutations in the rapid evolution of SARS-CoV-2, we proposed a model (the A-X model) to study the emergence of new strains on the phylogenetic tree and explained the patterns of SARS-CoV-2 macro-lineages[6-7]. Here, X represents a set of randomly generated sites in the spike protein. By expanding the stochastic sampling to a larger scale, we identified the statistical principles behind the emergence of new strains. Our findings show that the probability of a macro-lineage's emergence is linked to the number, x, of randomly generated sites within the X set. As x increases, the proportions of macro-lineages shift: lineage O surpasses lineage N, followed by lineage P surpassing lineage O, and eventually, lineage Q surpassing lineage P. We initially predicted the emergence of macro-lineage P, which has since been observed. Furthermore, we predicted the emergence of macro-lineage Q when x reaches a sufficiently large value[8]. These results provide a crucial theoretical framework for understanding SARS-CoV-2 evolution.

Fig. 1 shows the probability of macro-lineage emergence (PML) as a function of x, with the following demarcation values (99th percentile): New mutant ∈ N when $x \leqslant 20$; New mutant ∈ N or O when $21 \leqslant x \leqslant 30$; New mutant ∈ O when $31 \leqslant x \leqslant 37$; New mutant ∈ O or P when $38 \leqslant x \leqslant 62$; New mutant ∈ P when $63 \leqslant x \leqslant 78$; New mutant ∈ P or Q when $79 \leqslant x \leqslant 107$; New mutant ∈ Q when $x \geqslant 108$.

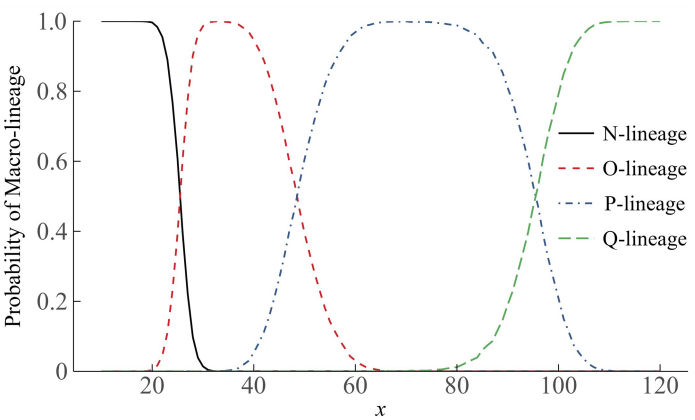

Fig. 1 PML(Probability of Macro-lineage) versus x for 61 mutants

Next, we analyze the relationship between the number of mutated sites x in a variant and its sample collection date. A linear regression of the number of mutated sites against the first sample collection date was performed, with results shown in Fig. 2[8]. The R-squared (R^2) and Residual Standard Error (RSE) are 0.91 and 6.51, respectively. These results suggest that the linear regres-

sion model provides a good approximation of the trend in the increasing number of mutated sites during viral evolution. The regression slope is $dx/dt = 1.268$ per month, with a 95% confidence interval of ± 0.103. If this linear relationship holds over a longer period, it will enable further predictions.

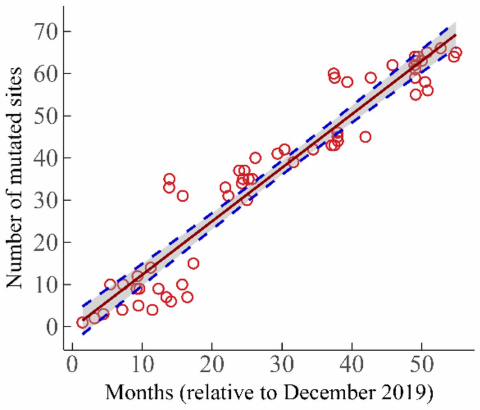

Fig. 2　Linear regression of the number of mutated sites versus the first sample collection date. Observational data are shown by cycles. The 95% confidence intervals are represented by the dotted lines around the straight line

Based on the results from Figs. 1 and 2, we predict that by the 62nd to 63rd month (starting from December 2019), the number of mutated sites will reach from $x = 78.25 \pm 3.76$ to 79.52 ± 3.85, and by the 86th month, it will reach $x = 108.69 \pm 6.07$. Using the demarcation values $x_1 = 79$ (initial emergence of macro-lineage Q) and $x_2 = 108$ (strong outbreak of macro-lineage Q) from Fig. 1, we forecast that macro-lineage Q will emerge around February 2025 (when $x \approx 79$) and, after approximately 23 months, will reach the strong outbreak stage (when $x \approx 108$)[8].

In conclusion, the spike protein mutations provide essential data for data-driven predictions. By integrating the novel strain emergence data from the A-X model and the assumption that the linear relationship between x and t holds, we predict that macro-lineage Q will emerge around February 2025 (when $x \approx 79$) and reach a strong outbreak stage about 23 months later (when $x \approx 108$).

Note: If other viral infections occur concurrently with SARS-CoV-2 in 2025 and 2026, the potential for competitive spread between two or more viruses in a region should be considered. To accurately simulate the cross-spread of two viruses, it is essential to account for the differences in the Cumulative Number of Infections (CNI) functions of each virus and their interactions. Detailed discussions can be found in ref. [5].

2　CO_2 concentration as an indicator of global surface air temperature

Global warming poses a significant challenge to earth's ecosystem[9-10]. Numerous studies on climate change, through observational data and attribution, consistently point to human activities, particularly CO_2 emissions since the industrial revolution, as the primary drivers[11-14]. The Intergovernmental Panel on Climate Change (IPCC) has synthesized this data, confirming that human emissions are a major cause of climate change. This consensus has driven global efforts by scientists and policymakers to develop mitigation strategies.

Traditionally, climate change research has focused on understanding how greenhouse gases

affect the atmospheric energy balance and the physical mechanisms of the global carbon cycle. However, due to the complexity of these systems, accounting for every detail remains challenging. We propose an alternative, data-driven approach. Our analysis reveals that historical CO_2 concentration data can be effectively modeled by exponential growth. Using the exponential function:

$$C = A \cdot \exp[B(t - T_0)] + C_0, (A = 55.408, B = 0.016581, C_0 = 258.61, T_0 = 1958) \quad (1)$$

we successfully simulated the annual variation of CO_2 concentrations from 1958 to 2023 at the Mauna Loa Observatory (Fig. 3)[15].

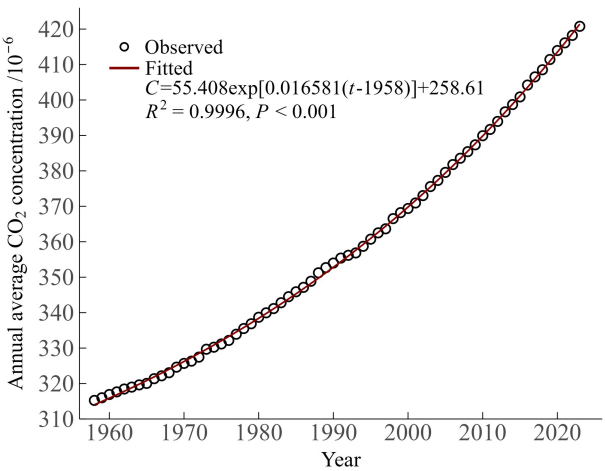

Fig. 3 Simulation of the variation of CO_2 concentration

Historical simulations play a key role in understanding the future climate dynamics. The IPCC introduced several Shared Socio-economic Pathways (SSPs) to project future greenhouse gas concentrations from 2015 onward[10]. According to Takens' theorem on delay embedding[2], historical data can be integrated with future projections within the framework of these five SSPs. By synthesizing historical and projected CO_2 concentration data from 2015 to 2500, we identified two distinct time-dependent patterns.

Before 2015 (denoted as y_0) and after a critical time point y (which varies between 2140 and 2235 across the five SSPs), the concentration follows pure exponential growth. Before 2015, the exponential model shows an increasing trend, while after y, it shows a decreasing trend. The second pattern captures the transition from exponential growth to exponential decay between 2015 and y. This transition is modeled by a modified exponential function (Eq. 2), which reflects the impact of socio-economic pathways and political conditions:

$$C = a \cdot \exp[\alpha(t - y_0) - \beta(t - y_0)^{\lambda}], t \in [y_0, y] \quad (2)$$

where parameters α, β, and λ are derived from simulations under different scenarios[15].

Global warming, defined as the long-term change in global average temperature from 1850 to 2020, does not follow a consistent annual increase. Using historical data, we established a linear relationship between temperature anomalies and CO_2 concentrations over extended periods (seven years or more). Based on this, using CO_2 concentration data we predicted global temperature changes up to 2500 under each SSP scenario[15]. The predicted temperature anomalies from 2015 to 2300 align with the projections in the IPCC Sixth Assessment Report (AR6)[10]. Furthermore, this approach

extends the forecast of global surface air temperature (GSAT) up to 2500, offering valuable insights into the trajectory of global temperature anomalies and informing proactive climate change mitigation efforts.

In summary, our analysis shows that the carbon dioxide concentration is an effective indicator for climate system evolution. Since high-dimensional systems often contain redundant information, and key features can be captured in low-dimensional representations, our data-driven predictions of global temperature anomalies suggest that low-dimensional embedding is a powerful method for studying the climate system, similar to techniques used in other complex systems[1-4].

3 Remarks

3.1 Fluctuation analysis

In the linear regression between the number of mutated sites and the first sample collection date (Fig. 2), the fluctuation is estimated by calculating the confidence interval of the predicted values in the regression model. Both CO_2 concentrations and temperature anomalies exhibit significant short-term fluctuations in climate warming. For example, carbon dioxide levels at the Mauna Loa Observatory fluctuate markedly from month to month. Similarly, natural temperature variations follow irregular cycles of 2 to 7 years, as seen in the El Niño/La Niña phenomena. To better understand these fluctuations in issues such as virus evolution and climate variation, a new analytical approach is necessary.

As is well known, chemistry and biology often study molecules, each of which has numerous internal degrees of freedom. Stochastic population kinetics offers a more accurate and realistic model of the biological world. Quaternions, discovered by the Irish scientist Hamilton in 1843, extended complex numbers by introducing three dimensions instead of one. Applying quaternions to describe stochastic population systems could provide more meaningful insights. A deeper understanding of quaternion phases and their relation to fluctuations was discussed in ref. [16] where, as an example, the quaternion-based model explained both the overall profile and fluctuation patterns of daily virus infection data. It is anticipated that the concepts and methods outlined in [16] can be applied to analyze the fluctuations observed in the evolution of both virus and climate systems.

3.2 CO_2 emission from land-use changes

Carbon dioxide (CO_2) emissions arise from two main sources: fossil fuel combustion and industrial activities, as well as land-use changes. To highlight the role of land-use change in CO_2 emissions, consider the case of Chilechuan Prairie in Inner Mongolia, northern China. Covering 10 km^2, this prairie has undergone restoration since 2012 and has become a prominent scenic spot by 2023 (Fig. 4).

Soil carbon sequestration plays a crucial role in the carbon balance of terrestrial ecosystems, storing over 90% of the carbon found in vegetation. Following its restoration, the prairie sequesters between 15000 tons and 45000 tons of carbon annually, equivalent to 55000 tons to 165000 tons of CO_2.

Given that global CO_2 emissions total approximately 40 Gt CO_2 per year, the restoration efforts

at Chilechuan Prairie highlight the potential for large-scale emission mitigation. Specifically, the carbon sequestration achieved by 2.42×10^5 to 7.26×10^5 grasslands of similar size to Chilechuan Prairie could offset current global CO_2 emissions.

However, when comparing this potential to the near-zero CO_2 emission targets for 2050, it becomes clear that scaling up such restoration projects is vital. The establishment of 100 grasslands of similar size to Chilechuan Prairie could significantly reduce CO_2 emissions from fossil fuel combustion and industrial activities, contributing to climate change mitigation.

This example emphasizes the critical role of land-use changes, such as the restoration of degraded ecosystems, in reducing carbon emissions and combating climate change. Projects like the restoration of Chilechuan Prairie offer practical solutions for achieving carbon neutrality and promoting sustainable environmental practices.

Fig. 4 Chilechuan Prairie in 2012 (left) and 2023 (right): Comparison before and after ecological restoration

References:

[1] WHITNEY H. Differentiable manifolds[J]. Annals of Mathematics, 1936, 37(3): 645-680.

[2] TAKENS F. Detecting strange attractors in turbulence[C]//Dynamical Systems and Turbulence. Berlin, Heidelberg: Springer, 1981: 366-381.

[3] SUGIHARA G, MAY R, YE H, et al. Detecting causality in complex ecosystems[J]. Science, 2012, 338(6106): 496-500.

[4] WU T, GAO X Y, AN F, et al. Predicting multiple observations in complex systems through low-dimensional embeddings[J]. Nature Communications, 2024, 15(1): 2242.

[5] LUO L F, LV J. Mathematical modelling of virus spreading in COVID-19[J]. Viruses, 2023, 15(9): 1788.

[6] LUO L F, LV J. An evolutionary theory on virus mutation in COVID-19[J]. Virus Research, 2024, 344: 199358.

[7] LUO L, LV J. Prediction on emergence of SARS-CoV-2 based on evolutionary theory of virus mutation[EB/OL]. (2024-08-29) [2024-12-25]. https://papers.ssrn.com/sol3/papers.cfm? abstract_id=4938698.

[8] LUO L F, LV J. Mutation sites increase over time in SARS-CoV-2 variants[EB/OL]. (2024-11-14) [2024-12-25]. https://www.preprints.org/manuscript/202411.0947/v1.

[9] MCCULLOCH M T, WINTER A, SHERMAN C E, et al. 300 years of sclerosponge thermometry shows global warming has exceeded 1.5℃[J]. Nature Climate Change, 2024, 14(2): 171-177.

[10] MASSON-DELMOTTE V P, ZHAI P, PIRANI S L, et al. IPCC, Summary for policymakers[M]// Climate change 2021-the physical science basis: Working group I contribution to the sixth assessment report of the intergovernmental panel on climate change. Cambridge: Cambridge University Press, 2023: 3-32.

[11] MANABE S, WETHERALD R T. Thermal equilibrium of the atmosphere with a given distribution of relative

humidity[J]. Journal of the Atmospheric Sciences, 1967, 24(3): 241-259.

[12] HASSELMANN K. On the signal-to-noise problem in atmospheric response studies[EB/OL]. Meteorology over the Tropical Oceans, 1979: 251-259 [2024-12-25]. https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_3030122.

[13] MANABE S, BROCCOLI A J. Beyond global warming: How numerical models revealed the secrets of climate change[M]. Princeton: Princeton University Press, 2020.

[14] HEGERL G, ZWIERS F. Use of models in detection and attribution of climate change [J]. Wiley Interdisciplinary Reviews Climate Change, 2011, 2(4): 570-591.

[15] LUO L F, LV J. Statistical theory on variation of carbon dioxide concentration in global warming[EB/OL]. Research Square, 2024: 3(2024-06-11) [2024-12-25]. https://doi.org/10.21203/rs.3.rs-4495753/v1.

[16] LUO L F, LV J. Quaternion in stochastic population dynamics[M]//BASWELL A R. Advances in Mathematics Research. New York: Nova Science Publishers, 2023, 33: 275-291.

复杂系统的数据驱动预测
——新冠病毒演化和全球气候变暖*

罗辽复[1]，吕　军[2]

(1. 内蒙古大学物理科学与技术学院，呼和浩特 010021；
2. 内蒙古工业大学理学院，呼和浩特 010051)

摘要：对于实际复杂系统，由于内在机制过于复杂，在无法获知其完整动力学方程情况下，可以通过演化产生的时间序列来进行预测。本文对两个复杂系统进行数据驱动预测。对新冠病毒演化问题，依据病毒刺突蛋白的氨基酸突变数据进行突变株的谱系预测，发现新的 Q 谱系可能在最近产生。对全球气候变暖问题，证明二氧化碳浓度是一个好的气候演化指标，由此出发将全球地表温度预测一直推广到 2500 年。

关键词：数据驱动预测；复杂系统；新冠病毒演化；全球气候变暖

中图分类号：N94；O231.5　　**文献标志码**：A

* **收稿日期**：2024-12-27

基金项目：内蒙古自然科学基金项目(2024LHMS06018)；内蒙古自治区直属高校基本科研业务费项目(JY20250094)

作者简介：罗辽复(1935－)，男，安徽歙县人，教授。主要从事理论物理和理论生物学研究。E-mail: lolfcm@imu.edu.cn

通信作者：吕　军(1973－)，男，内蒙古巴彦淖尔人，教授，博士。主要从事理论物理和理论生物学研究。E-mail: lujun@imut.edu.cn