Fundamental Research 4 (2024) 516–526

Contents lists available at ScienceDirect

Fundamental Research

journal homepage: http://www.keaipublishing.com/en/journals/fundamental-research/

Article

Development of the second version of Global Prediction System for Epidemiological Pandemic

Jianping Huang[a,b,*], Li Zhang[b], Bin Chen[a,b], Xiaoyue Liu[b], Wei Yan[b], Yingjie Zhao[b], Siyu Chen[a,b], Xinbo Lian[b], Chuwei Liu[b], Rui Wang[b], Shuoyuan Gao[b], Danfeng Wang[a]

[a] Collaborative Innovation Center for Western Ecological Safety, Lanzhou University, Lanzhou 730000, China
[b] College of Atmospheric Sciences, Lanzhou University, Lanzhou 730000, China

ARTICLE INFO

Article history:
Received 28 August 2022
Received in revised form 18 February 2023
Accepted 21 February 2023
Available online 10 September 2023

Keywords:
COVID-19
Epidemiological model
Prediction
GPEP
SEIR
Statistical-dynamic

ABSTRACT

Coronavirus disease 2019 (COVID-19) is a severe global public health emergency that has caused a major crisis in the safety of human life, health, global economy, and social order. Moreover, COVID-19 poses significant challenges to healthcare systems worldwide. The prediction and early warning of infectious diseases on a global scale are the premise and basis for countries to jointly fight epidemics. However, because of the complexity of epidemics, predicting infectious diseases on a global scale faces significant challenges. In this study, we developed the second version of Global Prediction System for Epidemiological Pandemic (GPEP-2), which combines statistical methods with a modified epidemiological model. The GPEP-2 introduces various parameterization schemes for both impacts of natural factors (seasonal variations in weather and environmental impacts) and human social behaviors (government control and isolation, personnel gathered, indoor propagation, virus mutation, and vaccination). The GPEP-2 successfully predicted the COVID-19 pandemic in over 180 countries with an average accuracy rate of 82.7%. It also provided prediction and decision-making bases for several regional-scale COVID-19 pandemic outbreaks in China, with an average accuracy rate of 89.3%. Results showed that both anthropogenic and natural factors can affect virus spread and control measures in the early stages of an epidemic can effectively control the spread. The predicted results could serve as a reference for public health planning and policymaking.

1. Introduction

The World Health Organization (WHO) officially declared the outbreak of coronavirus disease 2019 as a global pandemic in March 2020 [1]. The COVID-19 outbreak has been the most serious global public health emergency for nearly a century. It has caused a major crisis in global health security as well as economic and social order, and still affects the survival of mankind [2–4]. Beyond its spread, the COVID-19 pandemic has posed a series of longstanding social problems [2,5]. To better respond to an epidemic and prepare for its development in different situations, quantitative mathematical models that can quantitatively express epidemic evolution are essential.

The epidemiological model is an important method for predicting the spread of infectious diseases and mainly predicts the transmission speed, space scope, transmission route, and dynamic mechanism of infectious diseases to guide the effective prevention and control of infectious diseases [6]. The susceptible-infectious-removed (SIR) and susceptible-exposed-infectious-removed (SEIR) epidemiological models are the most widely used numerical modeling [7–9]. However, these models are built under a series of idealized assumptions, which may limit the accuracy and reliability of the prediction. To obtain more credible prediction results, more complex models should be developed to more realistically reproduce actual situations [10].

Although establishing an accurate epidemiological model to describe the spread of a pandemic is difficult, reported global pandemic data contain solutions for epidemiological processes [5,6,11]. Theoretically, it is possible to remedy the defects in previous epidemiological models by introducing the latest pandemic data [2,6,12]. Moreover, lessons regarding forecasting methods can be obtained from disciplines other than epidemiology. For example, lessons can be drawn from weather and climate predictions in atmospheric science [13]. In recent decades, atmospheric science has made remarkable progress in weather and climate predictions. Although the Earth System Model can make relatively reliable multiscale weather and climate predictions, it still requires

* Corresponding author.
E-mail address: hjp@lzu.edu.cn (J. Huang).

https://doi.org/10.1016/j.fmre.2023.02.030

significant improvements in parameterization schemes, involving significant progress to truly address many challenges such as addressing aerosols and clouds [14]. Therefore, the idea of a parameterization scheme in the Earth System Model is also a useful reference for the prediction of epidemiological models.

The second version of Global Prediction System for Epidemiological Pandemics (GPEP-2) was further developed based on a modified SEIR model, called the susceptible-exposed-infectious-quarantined-recovered-death-protected (SEIQRDP) model [15–18]. Although this version inherited the statistical-dynamic climate prediction method of the first version, it significantly improved the accuracy of the system. This model was modified from a modified SIR model to a modified SEIR model. In general, the GPEP-2 can describe more epidemiological characteristics than GPEP-1 and provide longer term prediction results, which also simulates more epidemic scenarios. By considering the impact of policies on the evolution of epidemic situations, the model can provide a more detailed scientific basis for formulating appropriate policies. Specifically, compared to the first-generation model, the second-generation model builds a control parameterization scheme. The model can predict the development of an epidemic in countries with strict control measures, such as China, by providing effective prediction information to government departments and offering more help for anti-pandemic causes. In addition, the second version retained the temperature parametrization scheme from the first-generation model. Because the model stability is better than that of the previous generation, seasonal predictions can be made for a longer period and early warnings can be provided for countries worldwide. This study presents the basic composition of the GPEP-2 model as well as the prediction results and their evaluation.

2. Model description

2.1. Modified SEIR model

GPEP-1 uses a modified SIR model that defines three disease states: susceptible (S), infected (I), and recovered and dead (R) cases. Based on the first version of the GPEP, we developed a second version (GPEP-2). The GPEP-2 was built based on a modified SEIR model [18]. This modified SEIR model [15] defines seven disease states: susceptible (S), protected (P), potentially infected (E, infected cases in a latent period), infected (I, infected cases that have not been quarantined), quarantined (Q, confirmed and quarantined cases), recovered (R), and mortality (D). The modified SEIR model emulates the time curve of an outbreak, and consists of the following equations:

$$dS(t)/dt = -\beta I(t)S(t)/N - \alpha S(t) \tag{1}$$

$$dE(t)/dt = \beta I(t)S(t)/N - \gamma E(t) \tag{2}$$

$$dI(t)/dt = \gamma E(t) - \delta I(t) \tag{3}$$

$$dQ(t)/dt = \delta I(t) - \lambda(t)Q(t) - \kappa(t)Q(t) \tag{4}$$

$$dR(t)/dt = \lambda(t)Q(t) \tag{5}$$

$$dD(t)/dt = \kappa(t)Q(t) \tag{6}$$

$$dP(t)/dt = \alpha S(t) \tag{7}$$

The sum of the seven categories is equal to the total population (N) at any given time.

$$S + P + E + I + Q + R + D = N \tag{8}$$

The model makes the following two assumptions:

(1) As the population of a certain country or region changes little over a short period, the total population (N) remains unchanged.
(2) During the epidemic, the population is evenly mixed.

The coefficients α, β, γ, δ, λ, and κ represent the protection rate, infection rate, inverse of the average latent time, rate at which infected people enter quarantine, time-dependent recovery rate, and time-dependent mortality rate, respectively. The population N was assumed to be constant, which meant that the birth and death rates were not considered. Compared with the traditional SEIR model, the improved model introduces two new disease states: protected (P) and quarantined (Q). Owing to the increasing awareness of self-protection during the pandemic, a group of individuals exists who are much less likely to be infected with the virus than susceptible individuals (i.e., protected cases). Due to the complexity of reality, it is not possible to accurately detect and quarantine all confirmed cases. Therefore, Q represents only the isolated confirmed cases that are no longer able to spread the virus. In addition, the mortality and cure rates in the modified model were not constant but changed over time based on actual conditions.

2.2. Determination of model coefficients

Due to the limitations of the simulation reality of epidemiological models and the unavailability of real epidemiological model coefficients, it is almost impossible to obtain the coefficients of a model directly from the real world. To better fit the epidemic curve, we introduced an improved inverting coefficients method into the model to improve its goodness of fit [17]. This method is used in statistical-dynamic numerical forecasting for weather and climate.

Specifically, to predict epidemics in countries worldwide, we first provided an initial value for each coefficient in the model. Subsequently, coefficient optimization algorithms (such as least-squares) and the latest epidemic data were used in real-time to invert the various coefficients in the model. The coefficients of the different parameterization schemes below are included: Specifically, the initial values of the coefficients and parameters were used to integrate the equation system. The minimum variance sum of the obtained time-series parameters and the actual data were adjusted through iterative calculations such that the initial values of the coefficients were adjusted and kept close to the real value. The coefficients used in the different scenarios are listed in Table S1. Finally, the coefficients obtained were substituted into the model to predict the subsequent development of the epidemic. The initial values of the coefficients and the initial number of people in various disease states may affect the accuracy of the inverted coefficients. To obtain more accurate and stable results, we combined empirical assumptions and changed the initial value several times to perform an inversion.

We also used a statistical dynamic forecasting method to predict outbreaks in China. For example, when a regional-scale outbreak occurred, owing to the lack of epidemiological data, we used the coefficients of a similar regional-scale outbreak to make predictions in the early stage. When sufficient data were available, they were used to make predictions. Furthermore, to enhance the stability of the traditional least-squares method (Gaussian–Newton algorithm), we use an improved damped least-squares method called the Levenberg–Marquardt algorithm [19,20]. This method inserts a damping coefficient into the Gaussian–Newton method to calculate a Hessian matrix. The benefit of introducing this damping coefficient is that it can converge rapidly in the steepest direction, even when the initial solution is far from the ideal values. This makes coefficient determination more robust [21]. Additionally, for all damping coefficients > 0, the coefficient matrix is positive definite, which places the Hessian matrix in the descending direction.

3. Parameterization schemes

"To parameterize" by itself means "to express in terms of parameters" [22]. Parameterization is a mathematical process that expresses

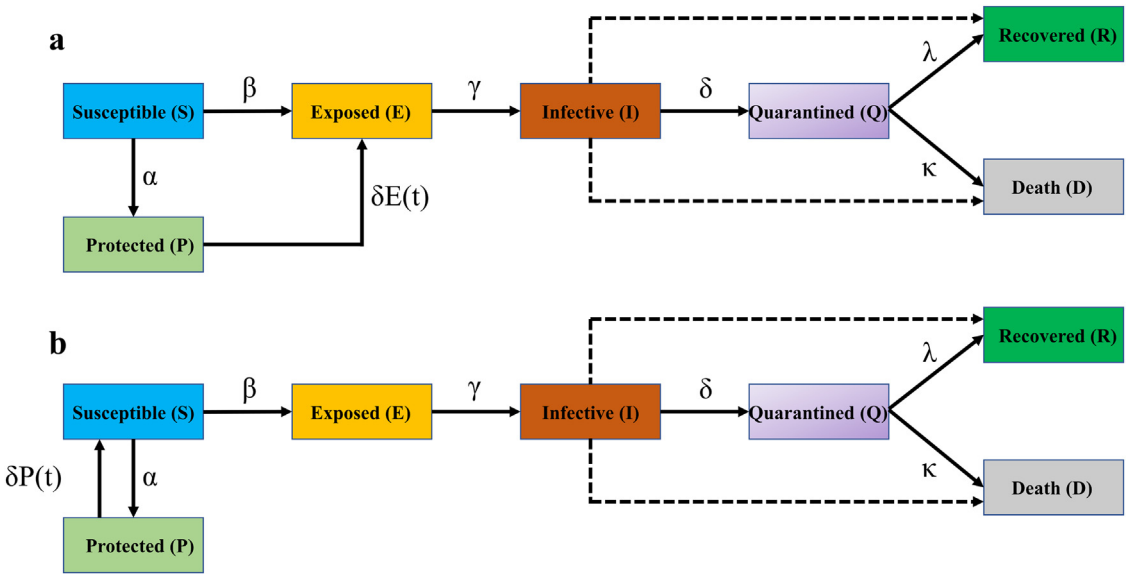

Fig. 1. (a) The schematic diagram of the parameterization scheme of massive gathering. (b) The schematic diagram of the parameterization scheme of unblock measures.

the state quantities of a system, process, or model as a function of independent variables, called parameters. The state of a system is typically determined using a finite set of coordinates. Therefore, the parameterization consisted of several real variables for each coordinate system. The number of parameters corresponds to the number of degrees of freedom of the system. Therefore, various processes in a model can be represented using several parameters or equations. Thus, the model can better describe the real world. By introducing parameterization concepts into epidemiological models, we constructed the following parameterization schemes and improved the model.

3.1. Parameterization of massive gatherings

Massive gatherings played an important role in virus spread. To simulate the impact of large-scale clustering on an epidemic quantitatively, we developed a scheme to switch between protected (P) and potentially infected (E) cases. A diagram of this is shown in Fig. 1a. In addition, we defined gatherings with > 200 and < 200 people as large- and small-scale gatherings, respectively. During massive gatherings, some protected cases are unprotected and exposed to COVID-19 [23,24]. Infected and asymptomatic cases among them will promote the epidemic spread [24]. We assumed that the proportion of infected and asymptomatic cases in the population could be calculated using the following equations:

$$\frac{dIa(t)}{dt} = M(t) \times \frac{Confirmed(t)}{Npop} \tag{10}$$

$$\frac{d\mathrm{P(t)}}{dt} = -M(\mathrm{t}) \tag{11}$$

$$\frac{d\mathrm{E(t)}}{dt} = \frac{dIa(t)}{dt} \tag{12}$$

where $\frac{dIa(t)}{dt}$ represents the number of undetected and asymptomatic infected cases in the gathering crowd, M(t) represents the number of gathered people, confirmed (t) represents the number of confirmed cases on day t, and Npop represents the population of the country or region. Protected cases ($\frac{d\mathrm{P(t)}}{dt}$) can reduce M(t), and potentially infected cases ($\frac{d\mathrm{E(t)}}{dt}$) can correspondingly increase $\frac{dIa(t)}{dt}$ during the gathering event in the scheme.

From the end of May to the beginning of June 2020, large-scale protests against violence broke out in the United States. We simulated this scenario using the aforementioned parameterization scheme for massive gatherings. During large-scale protests, large public gatherings, shouting, and shoulder-to-shoulder marching had already sown the seeds of a second outbreak in regions under initial control [23–25]. This makes it more difficult to contain epidemics in regions where the curve continues to rise. The use of tear gas and pepper spray against protesters may also have produced violent coughing and runny noses. These measures forced the protesters to remove their masks, making them more susceptible to the virus. A certain number of patients with latent diseases, including infected patients, may have participated in the protests. They may spread the disease to healthy protesters, police officers, and national guards who were not yet immune to the virus [25]. If close contacts of infected individuals are not fully tracked, they may spread the virus to other groups of people. All the above-mentioned scenarios could increase the risk of larger outbreaks.

The increase in the number of potentially infected individuals (δEt) in each city was estimated based on the ratio of the number of infected individuals (Qt) to the total population of the city (N). The timing of protests in each city was collected from local news reports. The number of protesters in each city was obtained using a modified model (Table S2). The daily increase in the number of potentially infected individuals (δEt) was used as the force input to the model. The model can then simulate the impact of protests on outbreaks. When the protests began, we forced group E to increase δEt. Fig. 2a shows the epidemic predictions for Miami and Florida. In Miami, the second outbreak peaked in mid-July 2020, with a maximum of 3000 new cases daily. The prediction results showed that protests played an important role and were the decisive factors in the second outbreak in these cities. With the protests, the second outbreaks were more severe and began earlier than the first.

3.2. Parameterization schemes of unblocking measures

In addition to the impact of large-scale gatherings on the pandemic, early unblocking due to economic pressure also had a significant impact. Once the control measures are lifted, the contact rate between people can increase significantly. This has provided a hotbed for the spread of COVID-19. To simulate the impact of lifting control measures, we assumed that after the relaxation of control measures, protected cases (P) were reduced by a certain percentage every day until society was

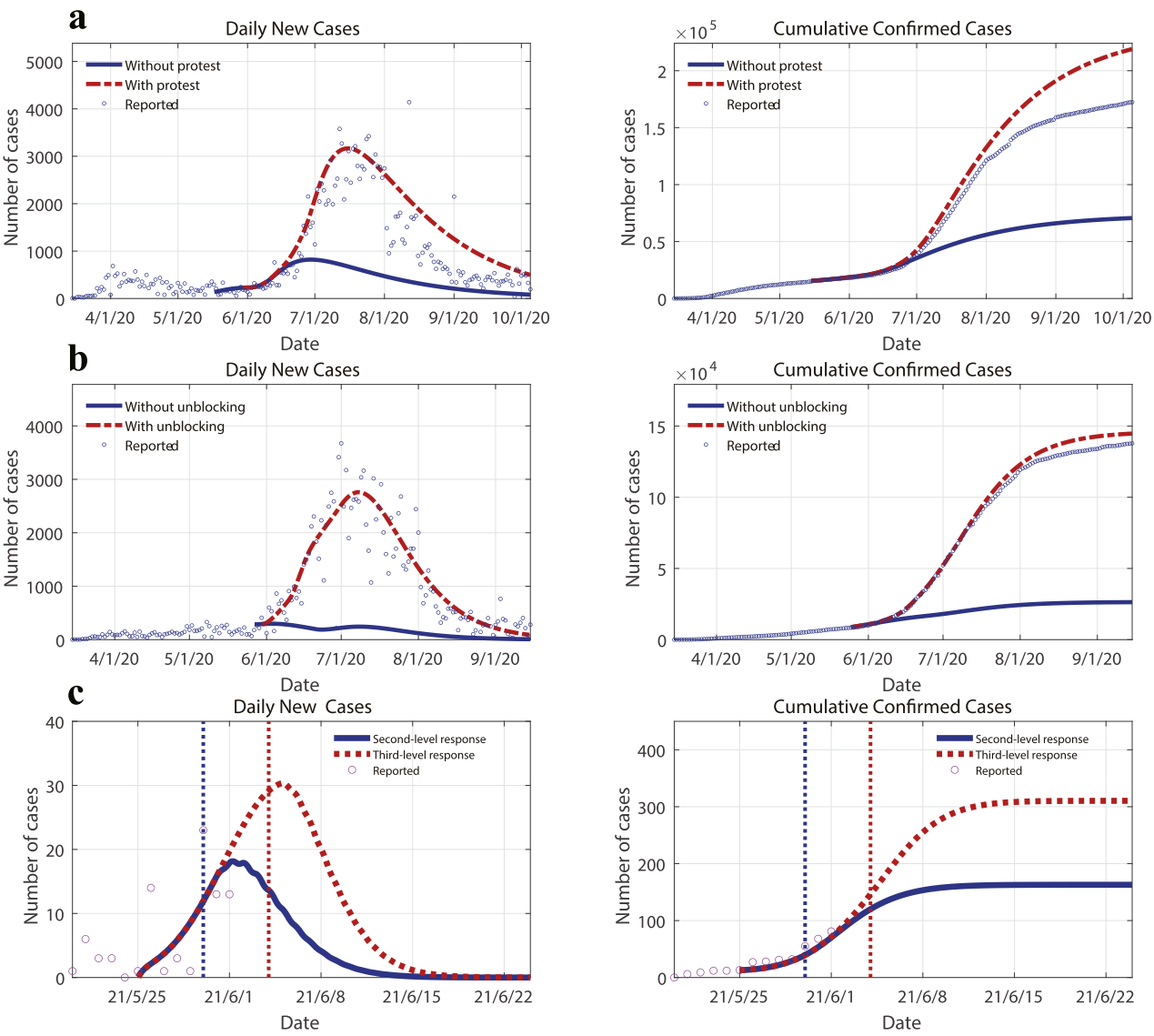

Fig. 2. (a) The impact of protests on the second outbreak in Miami, Florida. The blue dots denote the reported daily and cumulative cases of COVID-19. The solid blue line represents the simulation and prediction without protests, while the dashed red line denotes the simulation and prediction with protests. (b) The impact of unblocking on the second outbreak in Phoenix, Florida. The blue dots denote the reported daily and cumulative cases of COVID-19. The solid blue line represents the simulation and prediction without unblocking, while the dashed red line denotes the simulation and prediction with unblocking. (c) The simulation results for the sudden outbreak of COVID-19 in Guangzhou city on May 21, 2021 by the GPEP-2. The pink dots denote the reported daily and cumulative cases of COVID-19. The blue solid curve and red dashed curve denote the prediction of daily newly confirmed cases under the second- and third-level response, respectively. The red and blue vertical dashed lines represent the 5- and 10-day response times of the second- and third-level responses, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

unblocked to a certain extent after Days_lift. A diagram of this is shown in Fig. 1b. The equation is as follows:

$$P(t) = P(t) - P(t) \times P\_\text{lift\_rate(m)}\ m = 1, 2 \ldots \text{Days\_lift} \tag{13}$$

where $P(t)$ represents the protected cases at time t and P_lift_rate(m) represents the percentage reduction in protected cases on day m after lifting control measures. The P_lift_rate and Days_lift were retrieved from the model.

During the large-scale protests in the United States, the economy encountered huge difficulties. Therefore, government departments must implement control measures. To evaluate the contribution of unblocking measures to the promotion of the epidemic, we used the scheme described above to conduct simulation research. In addition, we assumed that protected cases were reduced by a certain percentage (P_lift_rate) daily after lifting the control measures. This process required a certain number of days (Days_lift). The timing of the lifting of the control measures was collected from local news reports. Using the optimization algorithm, the number of protesters (δEt) in each city, the reduction in percentage (P_lift_rate) and unblocking days (Days_lift) were obtained to calculate their impact on epidemic growth.

For example, the second outbreak peaked in early July 2020 in Phoenix, United States, with the maximum number of daily new cases reaching 2600 (Fig. 2b). The model also predicted enhanced secondary outbreaks in ten other cities in 2020, with massive gatherings and unblocking measure parameterization schemes. The results are summarized in Table S2. The predictions showed that the implementation of unblocking measures worsened the severity of the epidemic. Although unblocking measures temporarily alleviated the economic crisis, they brought about immeasurable long-term losses that may offset short-term benefits [26].

3.3. Parameterization of the control measures

Although many vaccines are currently being developed owing to the import and continuous mutation of the virus, it is almost impossible to avoid regional-scale local outbreaks in countries where the epidemic is under control. To accurately predict the scale of each local outbreak, a parameterized scheme for regional-scale outbreaks was developed based on the GPEP-2 model. This parameterization includes three coefficients: Days_con, E0, and Attenuation_rate. Because the first confirmed case is often not detected until several days after the infection, governments initiated control measures several days later. Days_con represents the government response time. There are several infected individuals at the beginning of the reported curves. Therefore, the reported epidemic curves differed from the actual curves, where E0 represented the number of initially exposed cases. Once the government initiated strong control measures, the epidemic infection rate decreased exponentially over time. The attenuation rate coefficient was used to represent the attenuation rate. β_0 represents the historical infection rate or base infection rate. Furthermore, if the local government took standard control measures in the corresponding version of the COVID-19 prevention and control protocol issued by the Chinese government, the response level was defined as a second-level response. If a local government adopted escalation measures, it was defined as a first-level response. If a local government adopted a downgrade measure, it was defined as a three-level measure.

$$\beta \begin{cases} \beta_0 & , t < Days\_con \\ \beta_0 * \text{Attenuation\_rate}^t, & t \geq Days\_con \end{cases} \tag{9}$$

The Chinese government has adopted strict epidemic control measures. We performed a simulation using the aforementioned parameterization scheme and achieved satisfactory results. As a large city with a population of approximately 20 million people, Guangzhou's epidemic control is a microcosm for the successful implementation of China's control policies. The outbreak in Guangzhou City started on May 25, 2021, and was caused by a delta-variant strain of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) that originated in India. This variant is one of the most virulent among the known variant strains of SARS-CoV-2. Therefore, the pressure on Guangzhou City during this round of the epidemic was greater than that during previous outbreaks. However, combined with the successful experience of several previous epidemic prevention and control measures and the contribution of large-scale vaccination, China once again successfully extinguished the epidemic in its early stages.

The prediction results of the GPEP-2 indicated that under second-level control, the number of newly confirmed cases in that round of the epidemic in Guangzhou City would peak at approximately 23 individuals on June 2, 2021 (Fig. 2c). The pandemic was expected to be effectively controlled by June 12, 2021. The estimated cumulative number of confirmed cases was approximately 163. After the end of the epidemic, the true cumulative number of infections in this round of the epidemic in Guangzhou was 153. The relative error of the system prediction was 6.5%. This demonstrated the reliable prediction ability of the GPEP-2.

3.4. Parameterization schemes of seasonal cycle

Apart from human factors that significantly impact epidemic development, many environmental factors change with the season, which can also affect or reflect the epidemic [27]. Among these, temperature is a major environmental factor [28].

Studies have confirmed that temperature affects the spread of SARS-COV-2. Temperature changes not only affect the activity of the virus itself but also affect human immunity and lifestyle [29]. The SARS-COV-2 can maintain high activity at lower temperatures, between 5 °C and 15 °C [28,30,31]. This is the temperature observed in autumn and winter at the mid-latitudes with the highest population densities. In addition, when the temperature is too high or low, people tend to gather indoors.

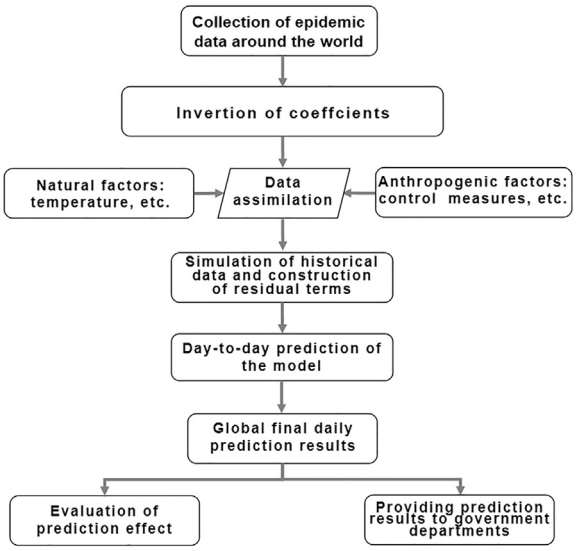

Fig. 3. The schematic diagram of system prediction process.

This increases the probability of viral transmission. To incorporate temperature into the model, we added a scheme to determine the effect of temperature on the infection rate.

$$\beta(t) = \beta_0(t) + F(t) \tag{14}$$

where $F(t)$ denotes the probability distribution function (PDF) obtained by Huang et al. [28]. They found that 60% of confirmed COVID-19 cases occurred in regions where the air temperature ranged from 5 °C to 15 °C. Using National Centers for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) reanalysis data, we calculated the global distribution of the annual PDF and included its influence on the infection rate. The results showed that high PDF values corresponded to ambient temperature, which is conducive to virus spread. In the Northern Hemisphere, the optimal band generally moved northward in summer (June, July, and August) and southward in winter (December, January, and February). In the Southern Hemisphere, the optimal band moves southward in summer (December, January, and February) and northward in winter (June, July, and August). Correspondingly, the probability of virus transmission in different temperature regions changed.

4. Process of system prediction

The entire prediction process included data collection, basic coefficient inversion and assimilation, simulation and prediction, model improvement, and policy formulation. Fig. 3 shows a schematic of the system prediction process.

Data collection: Data from countries worldwide were obtained from the COVID-19 Data Repository of the Center for Systems Science and Engineering at Johns Hopkins University (https://github.com/CSSEGISandData/COVID-19). When a small local outbreak in China was predicted, data were collected from the National Health Commission of China website. In addition to epidemic data, meteorological data, such as global 2-meter temperature and humidity grid data from ERA5 (the fifth-generation ECMWF atmospheric reanalysis of the global climate) were obtained. Reanalysis data from the ECMWF were also used for system predictions.

Basic coefficient inversion and assimilation: Historical data and a coefficient optimization algorithm were used to invert the basic model coefficients such as infection, isolation, mortality, and cure rates. A preliminary data assimilation module using data assimilation methods was constructed to integrate various parameterization schemes and data sources such as meteorological data.

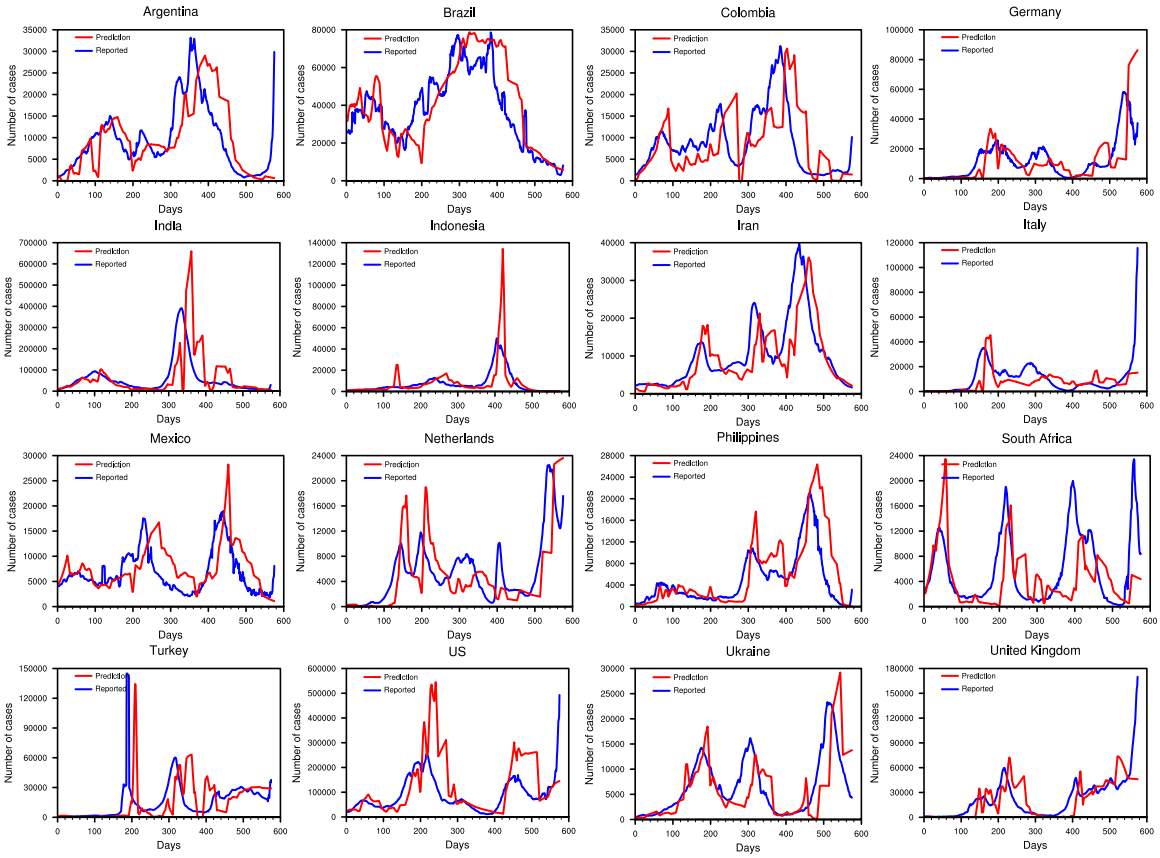

Fig. 4. The comparison of prediction results and reported data of 16 countries from 2020.6.1 to 2021.12.31. The solid blue lines represent the curve of the reported data of the epidemic. The solid red lines represent the curve of prediction of the epidemic. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The corresponding parameterization schemes were used for the different prediction scenarios. A temperature parameterization scheme was applied to all scenarios. Other parameterization schemes were applied according to different situations. For example, when predicting an epidemic in various countries, the model used unblocking and large-scale gathering parameterization schemes. The model used a control parameterization scheme to predict local outbreaks in China.

Simulation and prediction: After determining the basic coefficients of the model, historical data were fitted. The residual term was then constructed by comparing the fitted data with the reported data. By running the model with previously inverted basic coefficients, the daily prediction results for a single country or region were obtained. If an epidemic in a country or region has a specific form, corresponding parameterization schemes must be used. After determining the situation in different countries or regions, the corresponding model parameterization schemes were added to the model and adjusted individually. The prediction results for the corresponding scenarios were obtained. The final integration yields the global daily prediction result.

Model improvement and policy formulation: The prediction results were compared with reported data to test the prediction effect. In turn, improvements were made to the model, inversion algorithms, and parameterization schemes to improve the validity of the predictions. In addition to improving the model, a scientific basis is provided for the anti-epidemic causes in various countries, and the prediction results are provided to the WHO and government departments of various countries. In particular, when an outbreak occurs in China, regular and accurate predictions are provided to the National Health Commission of China and local governments. This has played a positive role in allocating medical resources to government departments in various countries, preparing isolation facilities, and understanding the development trends of the epidemic.

To verify the accuracy of the prediction system, we constructed the following scoring metrics to evaluate the prediction performance. This method is based on relative error. The scoring formulae for the epidemic predictions in various countries and China's domestic epidemic predictions were as follows:

$$PS_c = 1 - |MPE| \tag{15}$$

$$PS_d = 1 - |RE| \tag{16}$$

where PS_c and PS_d represent the prediction scores for epidemic predictions in various countries and China's domestic epidemic predictions, respectively. The MPE denotes the average relative error of each country's monthly cumulative prediction data. RE denotes the relative error of the cumulative prediction data for each outbreak.

5. Prediction results

5.1. Prediction of cases around the world

In 2020, COVID-19 had spread to most countries and regions worldwide. Moreover, the epidemics in various countries experienced several peaks and troughs. Fig. 4 shows a comparison between the reported curve of the pandemic and the predicted curve obtained from the average values of the GPEP-1 and GPEP-2 from June 2020 to December 2021 for the 16 most severely affected countries. The latest data were inputted into the model by rolling updates to predict the epidemic progress. To predict the developmental trend of the pandemic in a timely manner, the update frequency was set to approximately once every ten days. Predictions worldwide differ from refined predictions for cities in the United States, which use temperature, massive gathering, and unblocking parameterization schemes.

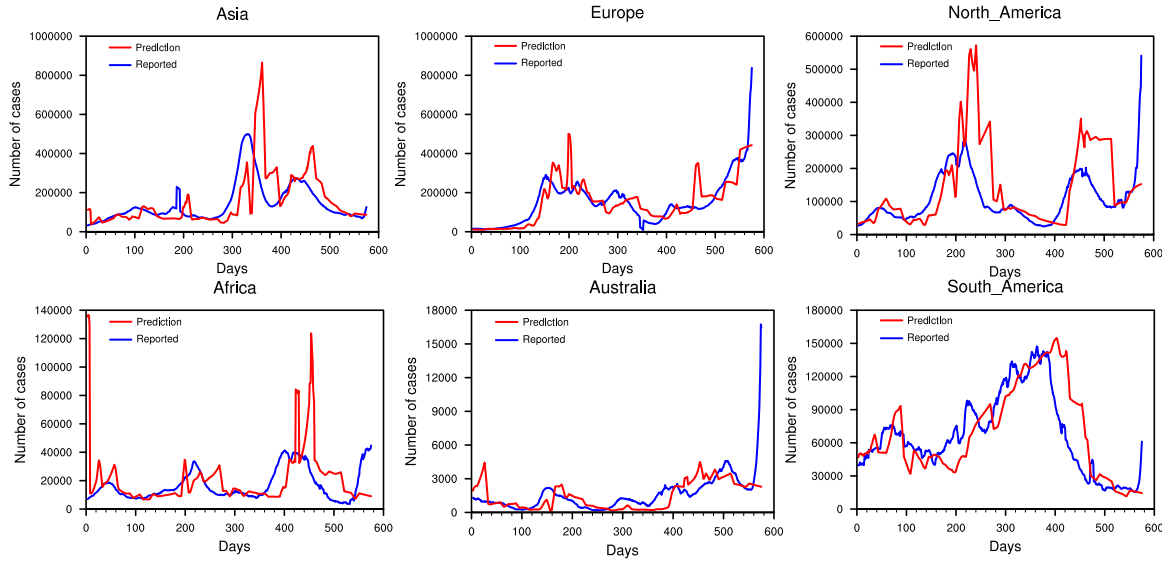

Fig. 5. The comparison of prediction results and reported data of 6 continents from 2020.6.1 to 2021.12.31. The solid blue lines represent the curve of the reported data of the epidemic. The solid red lines represent the curve of prediction of the epidemic. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

These 16 countries experienced at least two waves of epidemics in the last 3 years. Some countries such as the Netherlands and South Africa experienced four or five waves of epidemics. The epidemic peaks in each country showed a certain regularity and occurred at four time points. The first peak was observed in September 2020, corresponding to autumn in the Northern Hemisphere and the emergence of beta strains. The second peak corresponded to December 2020–January 2021. The third peak was detected in June 2021, which corresponded to the emergence of the delta strain. The fourth peak was detected from August to September 2021, corresponding to the start of autumn in the Northern Hemisphere. At the end of 2021, the emergence of the omicron variant caused an outburst of the epidemic in various countries, with a higher peak occurring in the subsequent period.

As shown in Fig. 4, the prediction results reflect the variation in the epidemic situation and serve as a reference. Except for the peaks in India and Indonesia, the peak predictions of the other countries were consistent in magnitude with the actual data. However, the model predictions lagged slightly behind the actual data. Although the prediction results are similar to the actual data in terms of trends, there is a gap between the two at the same time point. These limitations should be addressed in future studies.

In addition, we compiled the reported and predicted epidemic data for the six continents (Fig. 5). Asia, Europe, and North America experienced the worst outbreaks. The new daily peaks of the epidemic exceeded 400,000. With the spread of the omicron variant, the epidemic continued to deteriorate. Australia and South America have experienced relatively few outbreaks due to their smaller populations. However, Africa's medical system and other aspects of infrastructure are seriously underdeveloped; therefore, there is a problem of distortion in statistical data. This prediction captured trends in the development of the epidemic. Although there were deviations at certain times, the system can provide a basic reference for overall development trends. By updating the latest epidemic data in real time and inverting the latest model coefficients, the system can more accurately capture epidemic fluctuations.

The monthly cumulative summations of the predicted and reported epidemic data for all countries were performed and compared. The accuracy of the system was relatively high for countries with a small number of newly confirmed monthly cases (Fig. 6). This point had a regression line with a slope of 0.905 and probability of 1, indicating a high correlation between the predicted and actual outcomes. This indicated that the prediction results of the system exhibited a certain degree of credibility.

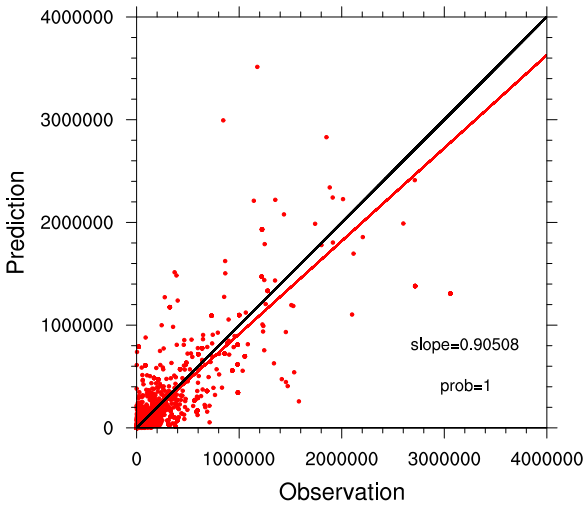

Fig. 6. Comparison of monthly predicted cumulative confirmed cases and reported cumulative confirmed cases from all countries. The horizontal and vertical axis represent the observations and model predictions, respectively.

Although they are somewhat discrete, the distribution of the points in the figure is relatively even on both sides of the 1:1 line. This is strongly correlated with a slight deviation from the peak. From a summation perspective, the deviations above and below the 1:1 line complement each other.

The predicted data for the 16 countries were summed monthly and compared with the reported data. As shown in Fig. 7a and Table S3, four statistics were calculated for the monthly summation: the mean (MEAN), prediction score (PS), root-mean-square error (RMSE), and correlation coefficient (CORR). From MEAN, it is easy to see that the epidemic data of the 16 countries are not a smooth transition, but three countries–the United States, Brazil, and India–are far ahead of other countries. This shows that the spatial distribution of the epidemic was not uniform, which may be related to population density, testing volume, etc. High PS values (> 60%) indicated the high accuracy and reliability of the prediction results. The magnitude of the RMSE was close to that of the MEAN. Thus, although the relative error of the system was small,

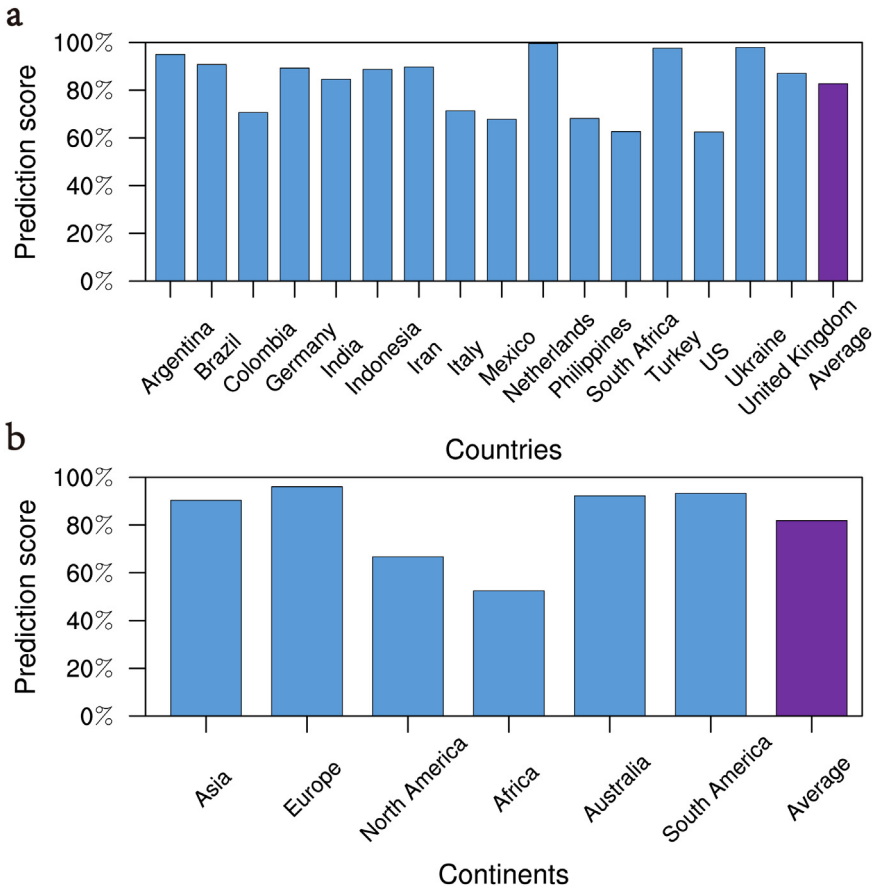

Fig. 7. Assessment of epidemic prediction for 16 countries and 6 continents. The blue bars represent the prediction scores of each country's and continent's epidemic prediction. The last purple bar represents the average prediction scores of all the country's and continent's epidemic prediction, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the predicted data were relatively discrete from the reported data. This may be due to large fluctuations in the reported data of these countries. Because the curve predicted by the system was relatively smooth, fluctuations in the reported data were not fully reflected. Sixteen countries show strong correlations, with an average of approximately 0.7.

Using the same algorithm, four metrics were calculated for six continents to evaluate the prediction accuracy of the system (Fig. 7b; Table S3). At the continental level, the data are uneven. The monthly average outbreak figures for Asia, Europe, and North America were much higher than those for the other three continents. This is similar to the differences between countries. The lower PS and RMSE values that are closer to the mean for these two continents are possibly due to the higher volatility of the data for North American and African countries. Moreover, the correlation was higher in all continents except Africa owing to poor detection and data collection capabilities.

5.2. Case prediction in China

In addition to predicting the epidemic situation in countries worldwide, another key task of this system is to predict the epidemic situation in China. Unlike large-scale outbreaks in other countries, domestic outbreaks in China are much smaller and shorter in duration, owing to strict control measures. After unblocking of control measures in Wuhan, China, on April 8, 2020, several regional-scale outbreaks occurred in China. To provide the government with a scientific basis for decision-making, we provided timely and accurate predictions of when the aforementioned outbreaks occurred. These predictions were based on the parameterization scheme of the regional-scale outbreak control measures described in Section 3. When the local epidemic first appeared, we obtained initial data from information released by the National Health Commission of China. The prediction of an epidemic is generally divided into two steps. First, due to the lack of data in the early stages of the epidemic, it is difficult to invert the coefficients. Therefore, the latest complete local epidemic data are used to invert the basic coefficients. Second, the basic coefficients are combined with the current epidemic data, and the control parameterization schemes are added. We then inverted the coefficients in the control parameterization scheme to predict the epidemic development trends.

Following these steps, we predicted 12 outbreaks in China (Fig. 8). As shown in Fig. 8, the system provided a very good prediction of various local epidemics in China. The developmental trends of the epidemics were predicted. The peak value and end time were in good agreement with the reported data. Because the scale of the local outbreaks in China was very small and the monitoring of the epidemic was relatively strict, as long as accurate data in the first few days can be provided, the system can predict and learn the general trend of the development of the epidemic. More than 30 outbreaks in China were effectively controlled during the initial stage; these data can be publicly accessed (http://covid-19.lzu.edu.cn/). Timely and effective control measures will remain the most effective means of dealing with the COVID-19 or other pandemics, such as the rapid detection of initial infections, timely and strict control measures by government departments, and large-scale epidemiological investigations using advanced technologies such as big data. Furthermore, vaccines play an important role in the fight against various mutant viruses.

Domestic outbreaks generally exhibited a unimodal pattern. During the epidemic period, except for a few cases, these outbreaks will end within 1 month. For policies, such as dynamic clearing, the system uses

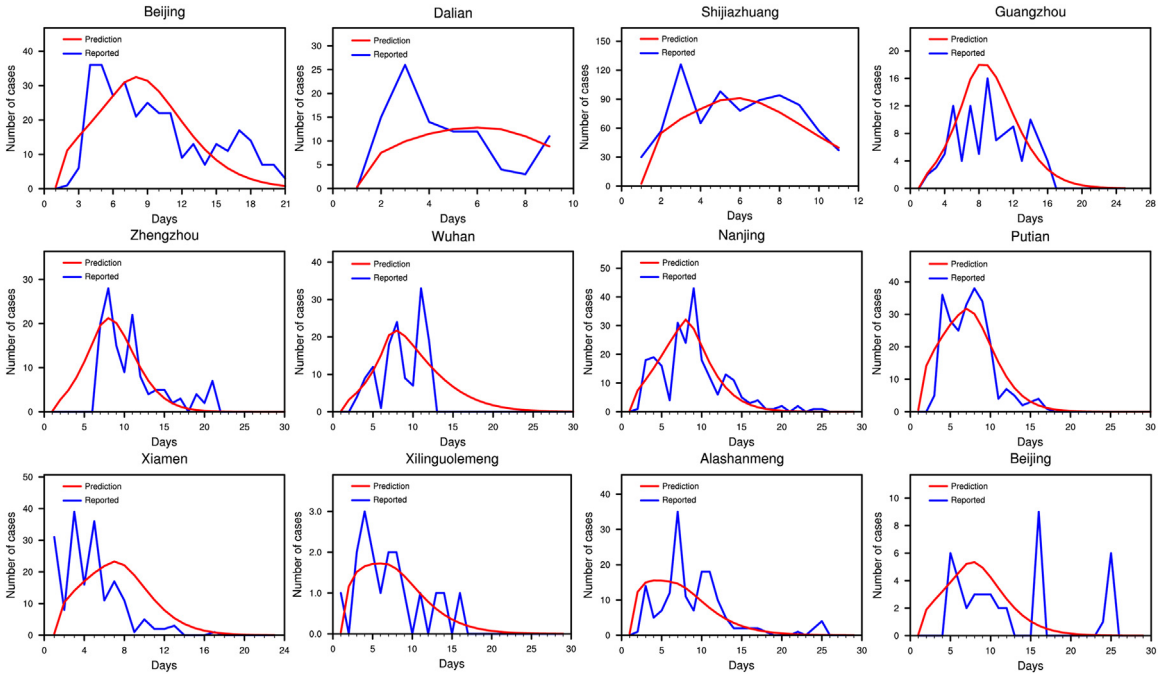

Fig. 8. The comparison of prediction results and reported data of 12 cities in China by the GPEP-2. The solid blue lines represent the curve of the reported data of the epidemic. The solid red lines represent the curve of prediction of the epidemic. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

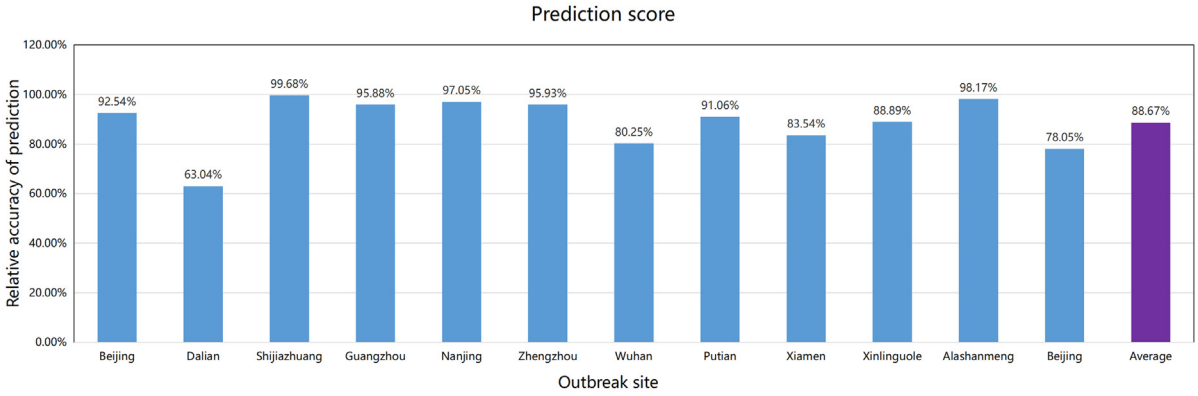

Fig. 9. Assessment of epidemic prediction for 12 outbreaks in China. The bars represent the prediction scores of each outbreak's epidemic prediction and their average (last purple bar). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

a control parameterization scheme to predict domestic epidemics. From a comparison of the predicted and reported data curves, the system can accurately capture the development trend of each epidemic and predict its end time. As shown in Fig. 9, the average prediction accuracy rate of the system for each epidemic reaches 89.3%, providing government departments with more accurate information on epidemic research and judgment. This can facilitate government departments to implement targeted deployment and response measures in advance.

Among the 12 local epidemics, except for Dalian, Wuhan, Xiamen, and Beijing (12th), the scores for the remaining cities were > 85%. Dalian and Beijing (12th) showed slightly larger deviations in epidemic predictions owing to a small upward trend in the final stage of the epidemic. In Wuhan and Xiamen, this was because the number of reported cases was too high or low at the beginning or end of the epidemic, respectively. This was closely related to the local testing capabilities. In several other epidemic predictions, the data curves conformed to unimodal characteristics and were relatively regular, without major changes. In the future, special optimization will be carried out for cases with fluctuating data to improve the prediction accuracy.

6. Discussion and conclusion

By improving the traditional SEIR model, this study developed the second-generation system GPEP-2. Based on the first-generation system (GPEP-1), the GPEP-2 adds protectors, quarantines, and deaths, and constructs various natural and anthropogenic factor parametrization schemes. The GPEP-2 improves the SIR model to the SEIQRDP model, which not only improves the stability of the model but also provides a basis for building more realistic parameterization. Unblocking and large-scale gathering parametrization schemes provide data support to simulate the relaxation or lifting of control policies. This can be conducive to the government's formulation of best-response strategies. Control parameterization schemes can be used in countries or regions that strictly control epidemics. Good simulation results are obtained. In

addition, for predictions in China under strict government and public intervention, if the epidemic trend does not match the unimodal trend predicted by the system, an early warning can be provided. An undiscovered chain or risk of infection may result in a new round of transmission. This provides valuable suggestions for the timely control of epidemic development trends.

The statistical-dynamic prediction method is effective for accurately inverting coefficients. It can also obtain model coefficients as close as possible to the real ones. Although compared with some generalized linear models (GLMs) and time-series models, the GPEP-2 has some deviations and lags in the rolling prediction of the epidemic situation in various countries worldwide, the dynamic system is included in the prediction. In addition, it can more effectively predict the inflection points, including the peaks. This is difficult to achieve using purely statistical models. These results proved that improving the epidemic model can provide good support for anti-epidemic causes.

However, the GPEP-2 has several limitations that must be addressed. Although the system counts asymptomatic infections as confirmed cases, certain errors may occur due to statistical problems. The current system builds only a single-point epidemic prediction model for each country and lacks a detailed description of the impact of population movement and socioeconomic factors on the epidemic. Although the system uses a rolling prediction method, and the data at different stages include information on vaccination and mutated viruses, these are not sufficient to accurately analyze the impact of vaccines and virus mutations [32,33]. A time parameterization scheme for infectivity, including vaccines and mutated viruses that change the viral pathogenicity, needs to be constructed. In addition, big data technology and the digitization level of contemporary society have greatly improved; however, large amounts of data are still rarely embedded in epidemiological models. It may be coarse and subjective to use the government's response level, such as control measures, as a quantitative standard. It will be more objective and quantitative to use big data or datasets to build indicators, such as NPI datasets [34]. Furthermore, the dynamic system of GPEP-2 must include a stochastic simulation to obtain a certain distribution interval and increase the predictive ability of the system. Moreover, data authenticity significantly affects prediction accuracy. For example, in most African countries, data quality is difficult to guarantee, and more coefficients may be provided rather than obtained by inversion. Furthermore, aerosol transmission, a major form of SARS-COV-2, has not received sufficient attention.

The limitations of this study include asymptomatic infections, the impact of external factors on the epidemic, the parameterization schemes of various processes, introduction of external data, and the validity of data in some regions. The prediction results confirmed that the core dynamic mechanism of the model reflected the changing trends of the epidemic. Therefore, improvements against the above limitations enable the model to provide more refined information and improve the prediction of small local fluctuations in time and space. Thus, the coverage of the model over time and space can be extended. Thus, the GPEP-2 requires further improvements to provide policymakers with a more refined and effective basis for decision-making.

Atmospheric science is one of the best fields for numerical simulation predictions [35]. A complex and sophisticated spatiotemporal four-dimensional numerical model can be used as a reference for epidemiological models. Furthermore, as environmental changes such as climate change become increasingly severe, studies have shown that the risk of > 58% of zoonotic diseases will increase [36]. Therefore, it would be beneficial to study the impact of environmental changes on infectious diseases by gridding an epidemiological model with climate and atmospheric chemistry models to formulate corresponding countermeasures in advance. Although epidemiologically gridded data are scarce, precedents exist for gridded epidemiological models [37]. Furthermore, because the climate and atmospheric chemical models are mature, we only need to build coupling modules. Therefore, the proposed scheme is feasible. Specifically, based on meteorological data, it was necessary to grid the collected epidemiological data. The interaction term between grid points was added to the mathematical equation system to grid the epidemiological model. Subsequently, this was converted into a meshed model. In addition to the gridding of the model, the improvement of the model itself, the update of the parameterization scheme, and the introduction of the traffic economy model were also the focus of subsequent improvements. Moreover, the model can be coupled with various economic, bioaerosol, and climate models.

Declaration of competing interest

The authors declare that they have no conflicts of interest in this work.

Acknowledgments

The authors acknowledge the Collaborative Research Project of the National Natural Science Foundation of China (L2224041) and the Chinese Academy of Sciences (XK2022DXC005): Frontier of Interdisciplinary Research on Monitoring and Prediction of Pathogenic Microorganisms in the Atmosphere, Self-supporting Program of Guangzhou Laboratory (SRPG22–007) and Gansu Province Intellectual Property Program (Oriented Organization) Project (22ZSCQD02). The authors acknowledge the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University for providing the COVID-19 data. We acknowledge E. Cheynet for providing the Modified SEIR Epidemic Model (fitting and computation).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.fmre.2023.02.030.

References

[1] X. Wu, Y.Q. Ye, A public health perspective on preventing and controlling the spread of Coronavirus Disease 2019, China CDC Wkly 2 (14) (2020) 237–240.
[2] M. Nicola, Z. Alsafi, C. Sohrabi, et al., The socio-economic implications of the coronavirus pandemic (COVID-19): A review, Int. J. Surg. 78 (2020) 185–193.
[3] COVID-19Data Repository By the Center for Systems Science and Engineering, CSSE) at Johns Hopkins University, 2020. https://github.com/CSSEGISandData/COVID-19.
[4] M. Cascella, M. Rajnik, A. Cuomo, et al., Features, Evaluation and Treatment Coronavirus (COVID-19), StatPearls (2022).
[5] World Health OrganizationCoronavirus Disease 2019 (COVID-19) Situation Report-51, 2020.
[6] F. Petropoulos, S. Makridakis, Forecasting the novel coronavirus COVID-19, PLoS ONE 15 (3) (2020) 1–8.
[7] J.T. Wu, K. Leung, G.M. Leung, Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: A modelling study, Lancet 395 (10225) (2020) 689–697.
[8] H.W. Wang, Z.Z. Wang, Y.Q. Dong, et al., Phase-adjusted estimation of the number of Coronavirus Disease 2019 cases in Wuhan, China, Cell Discov 6 (10) (2020) 1–8.
[9] Z.F. Yang, Z.Q. Zeng, K. Wang, et al., Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions, J. Thorac. Dis. 12 (3) (2020) 165–174.
[10] A. Godio, F. Pace, A. Vergnano, Seir modeling of the italian epidemic of sars-cov-2 using computational swarm intelligence, Int. J. Environ. Res. Public Health 17 (10) (2020) 1–19.
[11] N.M. Linton, T. Kobayashi, Y.C. Yang, et al., Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data, J. Clin. Med. 9 (2) (2020) 1–9.
[12] S. Tuli, S. Tuli, R. Tuli, et al., Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing, Internet Things 11 (2020) 1–16.
[13] F. Yang, S.J. Zhang, Q. Wang, et al., Analysis of the global situation of COVID-19 research based on bibliometrics, Heal. Inf. Sci. Syst. 8 (30) (2020) 1–10.
[14] G. Danabasoglu, J.-F. Lamarque, J. Bacmeister, et al., The Community Earth System Model Version 2 (CESM2), J. Adv. Model. Earth Syst. 12 (2) (2020) 1–35.
[15] L.R. Peng, W.Y. Yang, D.Y. Zhang, et al., Epidemic analysis of COVID-19 in China by dynamical modeling, medRxiv, 2020. https://doi.org/10.1101/2020.02.16.2002346.
[16] J.P. Huang, Y.H. Yi, Inversion of nonlinear dynamical model from the observation, Sci. China 34 (10) (1991) 1246–1251.
[17] J.P. Huang, Y.H. Yi, S. Wang, et al., An analogue-dynamical long-range numerical weather prediction system incorporating historical evolution, Q. J. R. Meteorol. Soc. 119 (33) (1993) 547–565.

[18] J.P. Huang, L. Zhang, X.Y. Liu, et al., Global prediction system for COVID-19 pandemic, Sci. Bull. 65 (22) (2020) 1884–1887.
[19] K. Madsen, H.B. Nielsen, O. Tingleff, Methods for Non-Linear Least Squares Problems, Society Industrial Applied Mathematics, 2004. https://doi.org/10.1155/2012/3129.
[20] L. Zhang, J.P. Huang, H.P. Yu, et al., Optimal parameterization of COVID-19 epidemic models, Atmos. Ocean. Sci. Lett. 14 (4) (2021) 1–5.
[21] A. Köházi-Kis, Relative effectiveness of the trust-region algorithm with precise second order derivatives, 6 (1) (2019) 1–7.
[22] D. Hughes-Hallet, W.G. McCallum, A.M. Gleason, Calculus: Single and Multivariable, Wiley, 2012.
[23] C. Kim, Images of police using violence against peaceful protesters are going viral, 2020. https://www.vox.com/2020/5/31/21275994/police-violence-peaceful-protesters-images.
[24] C. Rothenberg, S. Achanta, E.R. Svendsen, et al., Tear gas: An epidemiological and mechanistic reassessment, Ann. N. Y. Acad. Sci. 1378 (1) (2016) 96–107.
[25] S.M. Parodi, V.X. Liu, From containment to mitigation of COVID-19 in the US, JAMA 323 (15) (2020) 1441–1442.
[26] H. Brüssow, K. Timmis, COVID-19: Long covid and its societal consequences, Environ. Microbiol. 23 (8) (2021) 4077–4091.
[27] X.B. Lian, J.P. Huang, L. Zhang, et al., Environmental indicator for COVID-19 non-pharmaceutical interventions, Geophys. Res. Lett. 48 (2) (2021) 1–8.
[28] Z.W. Huang, J.P. Huang, Q.Q. Gu, et al., Optimal temperature zone for the dispersal of COVID-19, Sci. Total Environ. 736 (2020) 1–5.
[29] J.P. Huang, X.Y. Liu, L. Zhang, et al., The oscillation-outbreaks characteristic of the COVID-19 pandemic, Natl. Sci. Rev. 8 (8) (2021) 1–3.
[30] X.Y. Liu, J.P. Huang, C.Y. Li, et al., The role of seasonality in the spread of COVID-19 pandemic, Environ. Res. 195 (2021) 1–12.
[31] J. Omumbo, First Report of the WMO COVID-19 Task Team on Meteorological and Air Quality (MAQ) factors affecting the COVID-19 Pandemic, 2021.
[32] J. Li, R. Song, Z. Yuan, et al., Protective effect of inactivated COVID-19 vaccines against progression of SARS-CoV-2 omicron and delta variant infections to pneumonia in Beijing, China, in 2022, Vaccines 10 (8) (2022) 1–11.
[33] V. Thakur, R.K. Ratho, OMICRON (B.1.1.529): A new SARS-CoV-2 variant of concern mounting worldwide fear, J. Med. Virol. 94 (5) (2022) 1821–1824.
[34] COVID-19 Government Response Tracker, Blavatnik School Of Government At The University Of Oxford, (2020).
[35] W.J. Collins, N. Bellouin, M. Doutriaux-Boucher, et al., Development and evaluation of an Earth-System model - HadGEM2, Geosci. Model Dev. 4 (4) (2011) 1051–1075.
[36] C. Mora, T. McKenzie, I.M. Gaw, et al., Over half of known human pathogenic diseases can be aggravated by climate change, Nat. Clim. Chang. 12 (2022) 869–875.
[37] N.M. Ferguson, D.A.T. Cummings, C. Fraser, et al., Strategies for mitigating an influenza pandemic, Nature 442 (2006) 448–452.

Jianping Huang (BRID: 09265.00.65879) is a professor in College of Atmospheric Sciences and a director of Collaborative Innovation Center for Western Ecological Safety, Lanzhou University. He has long been dedicating to the study of long-term climate prediction, dust-cloud interaction and semi-arid climate change by combining field observations and theoretical study. Since the COVID-19 pandemic, he led his team to establish a global prediction system for COVID-19 pandemic by the combination of epidemic model and statistical-dynamic climate prediction methods.