A Universal Model for Prediction of COVID-19 Pandemic Based on Machine Learning

Lu Wang
Tsinghua University, The Future Laboratory https://orcid.org/0000-0002-3913-8862
De-wei Han
BMI Technologies Co. Ltd.
Kang Li
BMI Technologies Co. Ltd.
Xu Yang
BMI Technologies Co. Ltd.
Xiao-lei Yin
Tsinghua University, The Future Laboratory
Jing Qiu
Tsinghua University, The Future Laboratory
Dong-xue Liang (liang.laurel@hotmail.com)
Tsinghua University, The Future Laboratory
Zhao-yuan Ma
Tsinghua University, The Future Laboratory

Research Article

Keywords: COVID-19, pandemic trend prediction, machine learning

Posted Date: May 26th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-31164/v1

A Universal Model for Prediction of COVID-19 Pandemic Based on Machine Learning

Lu Wang (王路)[1]，De-wei Han (韩德伟)[2]，Kang Li (李康)[2]，Xu Yang (杨勖)[2]，Xiao-lei Yin (殷小雷)[1]，Jing Qiu (邱婧)[1]， Dong-xue Liang (梁冬雪)[1*]，Zhao-yuan Ma (马兆远)[1]

1, The Future Laboratory, Tsinghua University, Beijing 100086, China

2, BMI Technologies Co., Ltd., A111, 44 Third North Ring Road, Haidian District, Beijing 100088, China

* Corresponding author: Dong-xue Liang (liang_laurel@mail.tsinghua.edu.cn)

Abstract: **Background:** With the current worldwide spreading of the coronary virus (COVID-19) pandemic, accurately predicting the rate of spread of the virus has become an urgent need. **Methods:** In this article we propose a universal COVID-19 prediction model that is independent of country-specific factors in this paper. By analyzing the pandemic data in China, we combined the advantages of Gaussian function with that of chi-square distribution function, to render an innovative mathematical model named the H-Gaussian with five parameters to be learned, and solved the parameters by a gradient descent algorithm. **Results:** We trained the model with partial historical pandemic data to predict subsequent pandemic trends in several regions, and validated the predictions with real data. The H-Gaussian model was experimentally shown to correctly predict the pandemic trends, and the parameters had good interpretability. **Conclusions:** On this basis, the global trends of the pandemic are given based on the data currently available, as well as suggestions for subsequent prevention strategies.

Keywords: COVID-19; pandemic trend prediction; machine learning.

1. Background

The accurate prediction of the pandemic trend of the novel coronavirus (COVID-19) not only can help us understand the development of the pandemic in advance, but also have an important guiding role in the deployment of medical resources, accurate assessment of the implementation time, intensity of the control measures, and the

final effect of measures implemented. At present, the pandemic situation of the new COVID-19 [1][2] in China has been controlled, but the world pandemic situation is getting worse. It is particularly important to propose a universal pandemic trend prediction model that can adapt to the situations of various countries. However, the significant differences in the actual conditions of various countries make the study of this model challenging. In the research of infectious disease trend prediction, SIR model [3] and SEIR model [3] are the frequently used models, which have strong theoretical backgrounds and good interpretability. However, such models generally contain a large number of unknown parameters that need to be estimated, and it is difficult to take into account the control measures gradually implemented during the development of the pandemic. Researchers also use other models, such as exponential and polynomial growth rate models, to predict epidemic trends in the early stages of the pandemic. Empirical data such as HIV / AIDS [4] and Ebola [5] all display a sub-exponential growth pattern. These models also should take a variety of factors into consideration, for instance, spatial heterogeneity, cluster infection, and temporal heterogeneity of infection parameters, which increase the model complexity.

In this paper, by analyzing the characteristics of the domestic pandemic development in China, we designed an innovative mathematical model based on the Gaussian function and the advantages of chi-square distribution. The relevant parameters in the new model were automatically learned through the gradient descent algorithm. In this article, we used some real pandemic data in several countries to train several models, and used the trained models to give predictions of the pandemic trends in these countries. A comparative analysis was made with the real data, and the effectiveness of the model was proven. Afterwards, we used the current data from several countries to train the models separately and give the predictions of future data. We hope that the predictions of the pandemic trends in many countries would provide a certain scientific basis for the prevention and control of the worldwide pandemic.

This article is organized as follows: Section 2 conducts a descriptive analysis of the detailed case data sets of the COVID-19 infections, and establishes the mathematical model; Section 3 qualitatively and quantitatively analyzes the validity of the model through experimental results, and analyzes the future of several countries. The pandemic situation trends are predicted; Section 4 and Section 5 summarize the full text and give relevant pandemic prevention and control recommendations.

2. Methods

2.1. Data sources and analysis

The data used in this article refers to the latest pandemic data released by the National Health Commission of the People's Republic of China [6], Health Commission

of Hubei Province [7], World Health Organization (WHO) [8], and Johns Hopkins Coronavirus Resource Center [9].

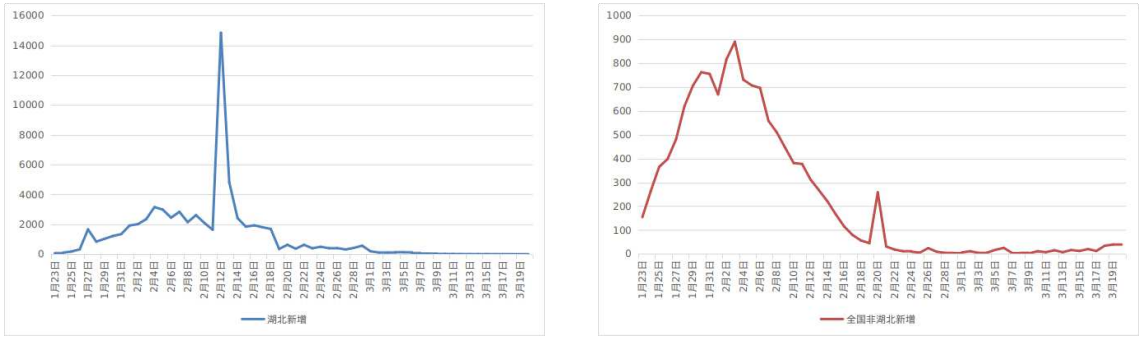


Figure 1, Daily new infections in Hubei Province (left), and daily new infections in non-Hubei parts of China (right).

By analyzing the historical daily new infection curves in Hubei and non-Hubei parts of China, we can find that the whole process of the pandemic can be divided into four stages: in the initial stage, because people have little knowledge about the virus, and the detection methods are underdeveloped, the number of the reported infection is very low. The second is the outbreak period. During this period the detection methods are continuously improved, and people realize the significance of the situation and they immediately go to hospital when symptoms appear. So the number of reported infections begin to increase exponentially. However, with the increasing awareness of everyone, the gradual improvement of protection and response measures, people actively or passively began to take quarantine orders, and go to hospital in time for symptoms. Governments promptly build hospitals that can accommodate more infected people and suspected infected people. As a result, the growth rate of the new number of infected people will decline until the inflection point appears. Then there comes the recovery period. With people's understanding of the dangers of the virus, the protection measures will be further strengthened, and the number of new infections will be gradually reduced due to the further reduction of travel and personal contact. However, because there has been no specific vaccines, the number of new infections per day will decrease more slowly. During the closing period, the number of new infections per day will not drop to zero soon, but will stay at a low level for a period of time. This is because although protection has not been relaxed during this period, factors such as the beginning of movement of personnel, the influx of overseas personnel, and asymptomatic infected people make it difficult to reduce the number of daily new infections to zero.

In this article, we proposed a novel model to fit the number of daily new infections across the four periods mentioned above. The model was used to fit the pandemic situation in several countries and predict the situation for a period of time.

2.2. The model

In the deep learning model for processing time series data, for instance, RNN and LSTM, the supervised learning method requires the data of the complete disease cycle as a training sample before the model can be trained. In addition, because of the high complexity of the deep models and the huge amount of parameters, a lot of data training is required to make the model better generalization ability. Due to the particularity of the COVID-19 pandemic, there is no large amount of data of the complete disease cycle for model training. Deep learning is not a good choice in this application. The traditional machine learning models play the opposite way. Only a small number of data can be used for training an efficient model. Considering that the four stages of infectious disease development have an intrinsic semi-symmetry, we first thought of using the Gaussian model to fit the number of daily new infections. The Gaussian model is defined as follows,

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(\frac{-(x-u)^2}{2\sigma^2}\right) \quad (1)$$

The Gaussian model has many advantages as a pandemic prediction model: (1) very few parameters, (2) strong interpretability, and (3) easy to solve and adjust the parameters. But on the other hand, because the Gaussian model is too simple and the it is strictly symmetrical, it cannot well express the semi-symmetric trend of the number of daily new infected patients given control measures. Therefore, we used the chi-square (χ^2-) distribution model to fit the number of new infections per day. The definition of the χ^2-distribution model is as follows:

$$f(x) = \begin{cases} \frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} exp\left(-\frac{x}{2}\right), x < 0 \\ 0, otherwise \end{cases} \quad (2)$$

This model is a semi-symmetric model, so it can well fit the non-identical increase and decrease rates in the number of daily new infections during the outbreak period and the recovery period. However, this model is a piecewise function, so it cannot fit the growth of the number of infectious patients in the initial stage well. Combining the advantages of the Gaussian model and the chi-square model, the new model should not only fit the slow increase of the number of new infections in the initial stage of the outbreak, but also the rapid increase of the number of new infections in the outbreak period, and not only fit the slow decrease in the new infection in the recovery period, but also fit the very slow decline in the number of newly infected people in the closing period. In other words, the new model is required to fit all the characteristics of the four stages of the pandemic. This is realized by adding a correction term to the Gaussian function to modify the symmetric curve into asymmetric. Finally, a novel mathematical model is proposed by combining the advantages of the Gaussian and the chi-square models, and we name it the H-Gaussian model, defined as follows:

$$f(x) = \frac{a}{\sqrt{2\pi}\sigma} exp\left(\frac{-(x-u-|x-u|^m+1)^2}{2\sigma^2}\right) + C \quad (3)$$

in which, $|x-u|^m + 1$ is the correction term that breaks the symmetry of the Gaussian function and enables stronger fitting capabilities, and the value of the order parameter m is in the range $0 \leq m \leq 1$. When $m = 0$, the H-Gaussian model degrades to the original Gaussian model。The effect of different values of m is illustrated in figure 2,

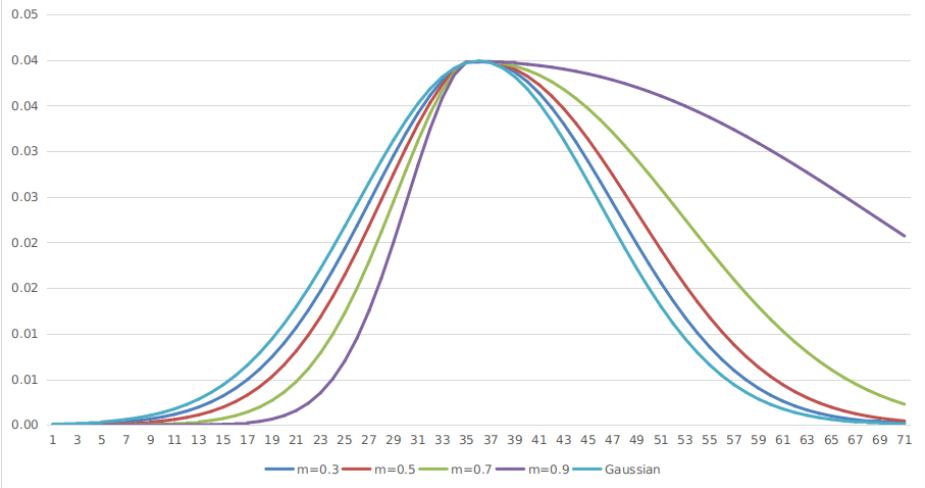


Figure 2, Illustration of H-Gaussian curves with different values of m. When $m = 0$, the H-Gaussian model is equivalent to Gaussian model. Other parameters of the H-Gaussian model are set as $a = 1, \sigma = 1, \mu = 0, C = 0$.

It can be seen from the curves that the H-Gaussian model can well fit the four stages of the pandemic: the situation of slow growth in the initial period, rapid growth in the outbreak period, slow decline in the recovery period, and a very long closing period to return to zero. There are five parameters in the H-Gaussian model, a, σ, μ, m, C, respectively, and these parameters need to be trained using the pandemic data. By integrating the prevention and control situations of each country in the five parameters, the model can adapt to and fit the pandemic trends of different countries.

2.3. Loss function

The mean square error (MSE) has a very good geometric meaning. It corresponds to the commonly used Euclidean distance. The method of solving a model based on the minimum mean square error is called "least squares". Its objective is to find a curve that makes the sum of Euclidean distances between the curve and the samples smallest. In order to make the estimated value closer to the observed value, the least square method is used to construct the loss function. The MSE loss function is defined as follows,

$$J(\theta) = \frac{1}{2n}\sum_{i=1}^{n}[f_\theta(x_i) - y_i]^2 \quad (4)$$

in which x_i denotes the sample， y_i denotes the ground truth value， θ denotes the parameters in the model.

2.4. Gradient descent algorithm

When the model is very simple, the closed form solution of the model parameters can be directly obtained using the least square method. It is more convenient to use the gradient descent algorithm when the model is more complex. Using gradient descent algorithm, all parameters are updated according to the following formula:

$$\theta = \theta - \eta\Delta\theta \quad (5)$$

where η is the learning rate, and $\Delta\theta$ is the gradient of the parameter θ.

For simplicity, let $\boldsymbol{P} = x - \mu - |x - \mu|^m + 1$, the gradients of each parameters are calculated as,

$$\frac{\partial J}{\partial a} = \frac{1}{n}\sum_{i=1}^{n}[f(x_i) - y_i]\frac{1}{\sqrt{2\pi}\sigma}exp\left(-\frac{\boldsymbol{P}^2}{2\sigma^2}\right) \quad (6)$$

$$\frac{\partial J}{\partial \sigma} = \frac{1}{n}\sum_{i=1}^{n}[f(x_i) - y_i]\,exp\left(-\frac{\boldsymbol{P}^2}{2\sigma^2}\right)\left(\frac{\boldsymbol{P}^2}{\sigma^3} - \frac{a}{\sqrt{2\pi}\sigma^2}\right) \quad (7)$$

$$\frac{\partial J}{\partial \mu} = \frac{1}{n}\sum_{i=1}^{n}[f(x_i) - y_i]\frac{a}{\sqrt{2\pi}\sigma}exp\left(-\frac{\boldsymbol{P}^2}{2\sigma^2}\right)\frac{\boldsymbol{P}}{\sigma^2}[1 + m(x-\mu)|x-\mu|^{m-2}] \quad (8)$$

$$\frac{\partial J}{\partial m} = \frac{1}{n}\sum_{i=1}^{n}[f(x_i) - y_i]\frac{a}{\sqrt{2\pi}\sigma}exp\left(-\frac{\boldsymbol{P}^2}{2\sigma^2}\right)\frac{\boldsymbol{P}}{\sigma^2}(-|x-\mu|^m\, ln\,|x-\mu|) \quad (9)$$

$$\frac{\partial J}{\partial C} = \frac{1}{n}\sum_{i=1}^{n}[f(x_i) - y_i] \quad (10)$$

3. Results

3.1 Training processes

In preparation of the data, we use 0, 1, 2 ... n to denote the date since the beginning of the pandemic. For a certain country or area, the dataset is organized as an ordered set of n+1 sample pairs Ω ~ ((0, y0), (1, y1), (2, y2) ... (n, yn)), where yi denotes the number of new infectious on the i-th day. Because gradient descent is sensitive to the initialization of the parameters, we use {a = 100000, σ = 15, m = 0.7, C = 0} for initialization. The optimizer is ADAM [10] ($\beta 1$ = 0.9, $\beta 2$ = 0.999, eps = 1e-8), learning rate = 1e-3, weight decay = 1e-5. The pseudocode of training process is as follows:

Algorithm 1: training steps of H-Gaussian model

Input：training set Ω， learning rate, weight decay

Process：

Initializing a, σ, μ, m, C

for epoch = 1,...,K **do**

  for batch = 1,...,B **do**

    $\boldsymbol{x}, \boldsymbol{y} \leftarrow$ randomly select n pairs of training samples from Ω

    $\boldsymbol{z} \leftarrow f_{a,\sigma,\mu,m,C}(\boldsymbol{x})$

    $L \leftarrow \frac{1}{2n}\sum_{i=1}^{n}(\boldsymbol{z}-\boldsymbol{y})^2$

  //Compute the gradients of parameters a, σ, μ, m, C, according to equations (6) (7) (8) (9) (10), and upgrade the parameters

  $a \overset{+}{\leftarrow} -\nabla_a L,\ \sigma \overset{+}{\leftarrow} -\nabla_\sigma L,\ \mu \overset{+}{\leftarrow} -\nabla_\mu L,\ m \overset{+}{\leftarrow} -\nabla_m L,\ C \overset{+}{\leftarrow} -\nabla_C L$

  end for

end for

Output：trained parameters a, σ, μ, m, C

3.2 Test results

First, the Gaussian model and the proposed H-Gaussian model are used to fit the historical pandemic data in non-Hubei parts of China, and the effects of the two models on the data fitting are compared.

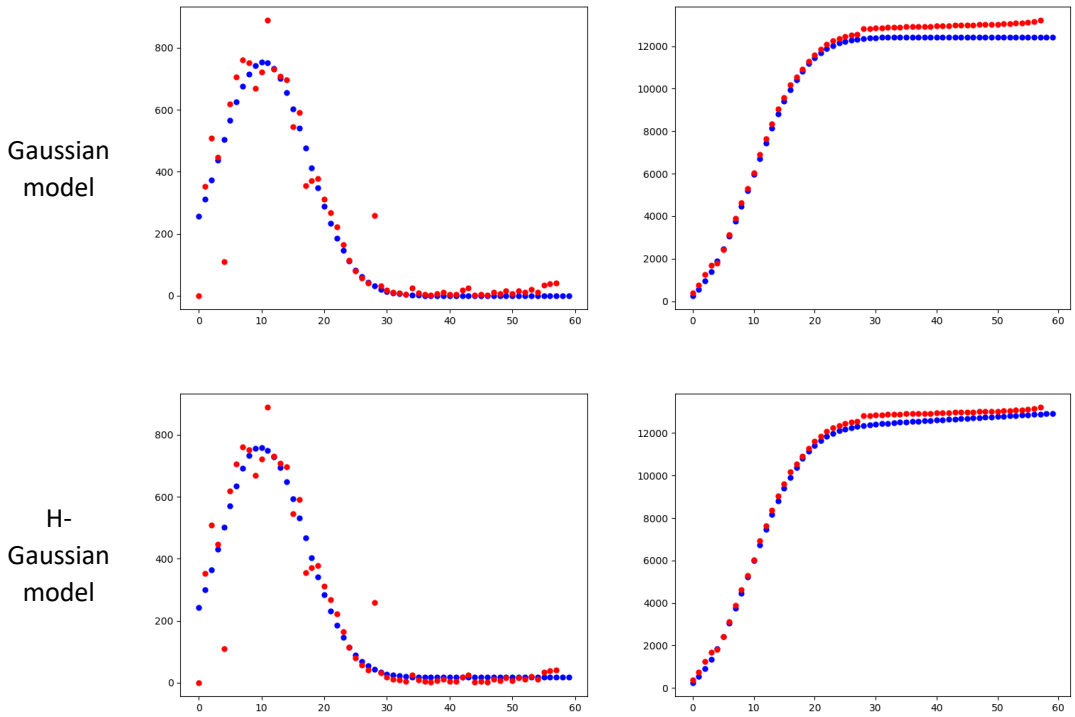
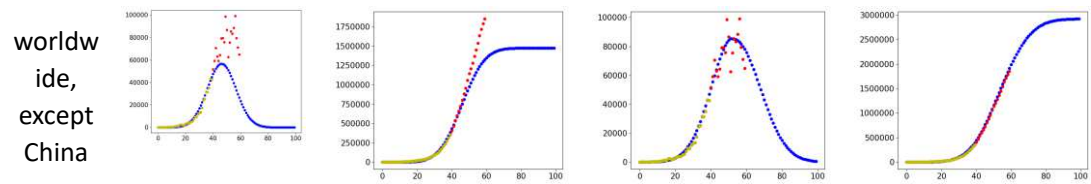


Figure 3, Comparison of the fitting by the two models on the pandemic data in non-Hubei parts of China. *Left column*: number of daily new infectious people versus time. *Right column*: number of cumulative infectious people versus time.

In figure 3, the data from 23 January 2020 to 10 March 2020 are used for training the models. The blue dots represent the predicted values, and the red dots represent the actual data. It can be seen from the figure that the relevant parameters of the respective models can be learned using the training dataset. In this case, both the Gaussian and the H-Gaussian models have a good fitting effect. Since a correction term is added, the H-Gaussian model performs better.

Then, we select several sets of data in several countries or areas to verify and analyze the H-Gaussian model. The worldwide data except China, the United States of America (USA) and Italy are taken as three examples. We use the data from 15 February to 25 March (41 days in total) as the training data to train three separate models, and use the data from March 26 to 15 April (20 days in total) to test and verify the models. The results are illustrated in the following figure,



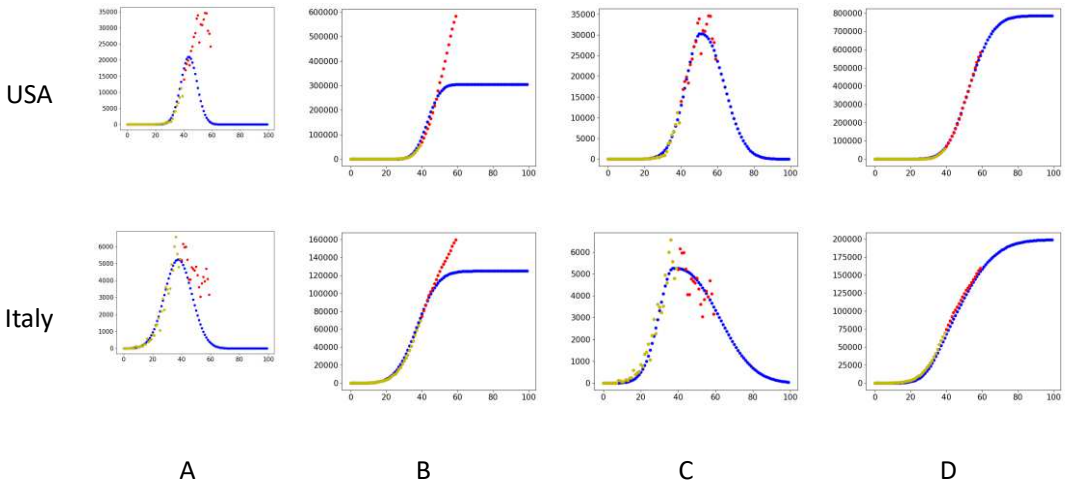


Figure 4, Comparison of the fitting effects of the Gaussian model and H-Gaussian model on the pandemic data worldwide (except China), the USA, and Italy. The yellow dots represent the training data of 41 days, the red dots represent the test data of the subsequent 20 days, and the blue dots represent the predicted values. Column A: number of daily new infections given by the Gaussian models. Column B: number of daily cumulative infections given by the Gaussian models. Column C: number of daily new infections given by the H-Gaussian models. Column D: number of daily cumulative infections given by the H-Gaussian models.

The H-Gaussian models give predictions for the inflection points, whereas the Gaussian models fail, resulting in a great deviation from the test data. The mean square errors (MSE) of the two models on the training set and the test set are listed in Table 1,

Table 1, comparison of the fitting effects of the Gaussian model and H-Gaussian model. The MSE of the two models are computed.

	worldwide, except China		USA		Italy	
error (MSE)	training error	test error	training error	test error	training error	test error
Gaussian	**3645652**	1042539291	1672497	379591331	1672497	379591331
H-Gaussian	3858970	**85215538**	**329815**	**7549344**	**329815**	**7549344**

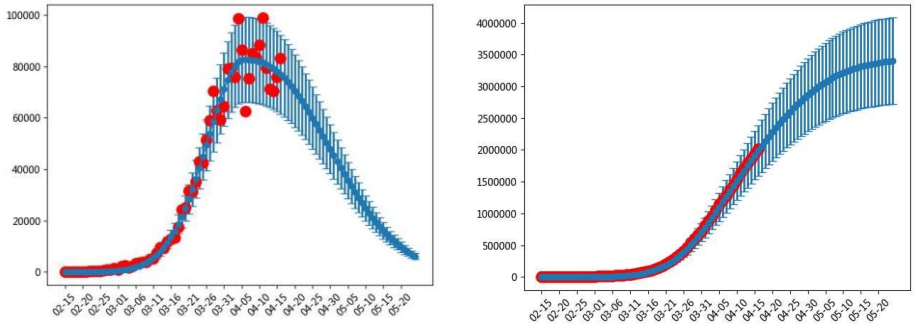
Combining the results of Figure 4 and Table 1, it can be concluded that when very few real data are used to predict the pandemic, the mean square error of the Gaussian

model is generally much greater than that of the H-Gaussian model, indicating that the H-Gaussian model has a higher capacity in fitting complicated data. Moreover, the Gaussian model cannot accurately predict the inflection point, nor can it predict the approximate daily number of newly infected people at the inflection point, resulting in a large error between the subsequent prediction and the real data. It is very impressive to see that the H-Gaussian model uses the very limited training data to accurately predict the approximate date of the occurrence of the inflection point in different countries, and the approximate number of new infections per day around the inflection point, without any prior knowledges of the inflection point. As the training sample size increases, the prediction results will be more accurate.

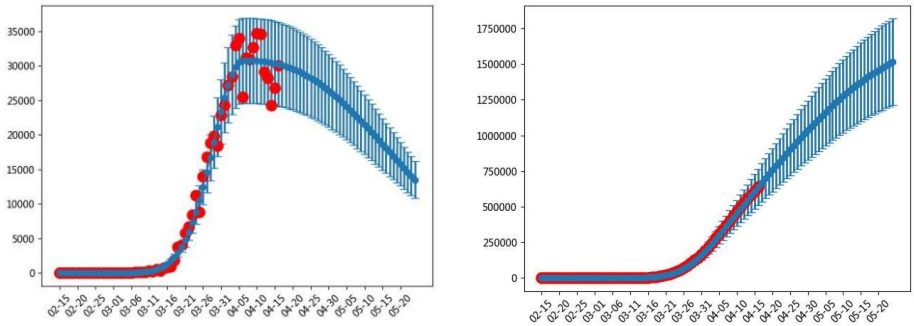
3.3 Future prediction

Next, we use all the data from 15 February to 15 April (61 days in total) of several regions to train separate H-Gaussian models and give predictions of the future development of the COVID-19 pandemic. The results are illustrated in Figure 5,

worldwide, except China



USA



Italy

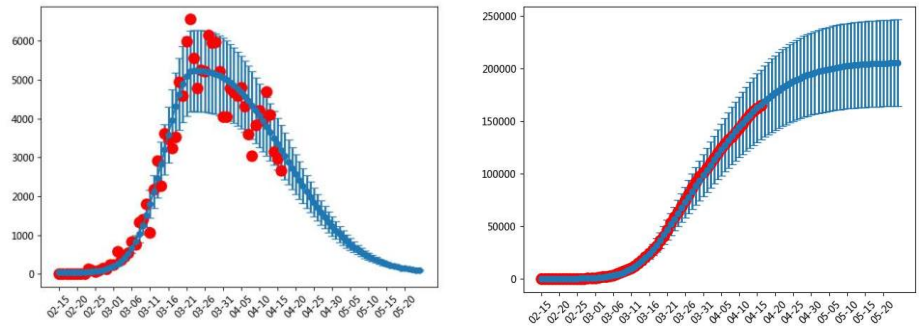


Figure 5, the fitting and prediction of H-Gaussian models trained on the data of 61days in several regions. The red dots represent the training data, the blue dots represent the fitting and predictions given by the H-Gaussian model with the optimized parameters, and the error bars represent the fitting and predictions given with extreme parameters. *Left column*: daily number of new infections. *Right column*: daily number of accumulative infection.

It should be noted that the optimized parameters are trained with all the 61 days of data, whereas the extreme parameters are trained with a subset of the training data containing only the upper bound samples or the lower bound samples. The optimized parameters trained on the 41-days data set and the 61-days data set are listed in Table 2.

Table 2, comparison of parameters of the H-Gaussian model after training with different data sets

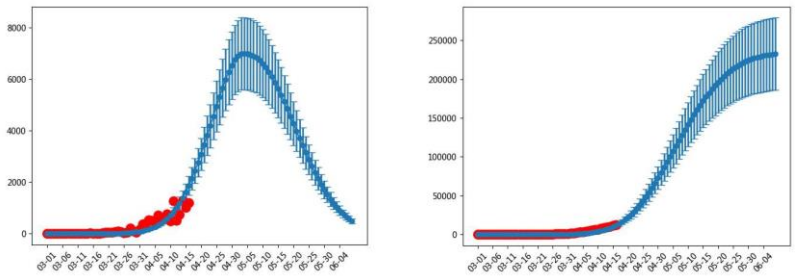
	worldwide, except China		USA		Italy	
training set size	41	61	41	61	41	61
a	85479	82815	30374	30788	5274	5206
σ	13.47	14.8	10	11	12	12.4
μ	51	50	50	50.73	37	36.5
m	0.4	0.7	0.47	0.88	0.78	0.78
C	48	42	4	0	0	35

It is easy to see from table 2 that the parameters trained using the 41-day dataset are not much different from the parameters trained using the 61-day dataset, further validating the stability and effectiveness of the H-Gaussian model. Among the five

parameters, σ, m modify the shape of the curve, which reflects the speed of virus transmission. The smaller σ is and the larger m is, the larger the virus transmission rate is, and the slower the number of daily new infections decreases after the inflection point. The parameter a influences the height of the curve, which reflects the number of infections on the day of the inflection point. μ affects the phase of the curve, and further reflects the date when the inflection point appears. In the end, C is a bias factor. These observations are roughly consistent with the actual situation, which once again proves the effectiveness of the model.

Finally, we use the H-Gaussian model to predict the outbreak in India. The inflection point of India has not yet appeared, so the model is used to fit and predict the data from March 1 to April 17 in India.

prediction 1, inflection point on 2 May



prediction 2, inflection point on 25 April

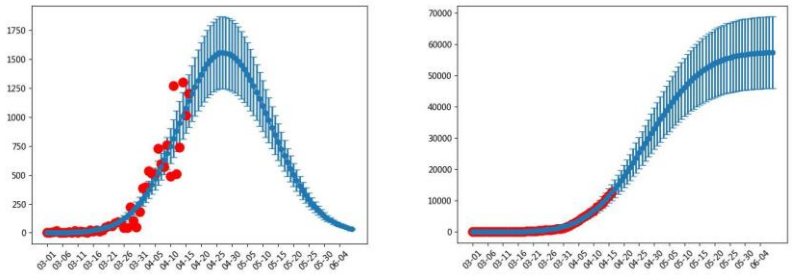


Figure 6, H-Gaussian model predictions of India's pandemic with different inflection points. *Left column*: daily number of new infections. *Right column*: daily number of accumulative infection.

The top row in Figure 6 illustrates the pandemic trend fitted by the model based on the data from 1 March to 17 April. It can be predicted that the subsequent Indian pandemic situation is still serious, and the final number of infected people may be as high as 240,000. The bottom row illustrates the pandemic trend if more effective prevention and control measures are immediately taken to deal with the virus spread, and the inflection point is supposed to arrive seven days earlier. Then the final number of infected people will be reduced by approximately 180,000 to 60,000.

4. Discussion

The data used in this article ends on 17 April 2020. Because the pandemic situation will be affected by various factors, such as policy, whether there is a strong quarantine measure, and when the vaccine appears, this article is limited to using historical data as a basis. Therefore, in actual applications, it is required to constantly modify the model parameters according to the actual data to give more accurate prediction results.

5. Conclusions

In this article, we combine the advantages of Gaussian and chi-square distribution functions, and propose an innovative mathematical model to fit and predict the daily number of COVID-19 infections. The proposed H-Gaussian model has five parameters which can be learned with gradient descent algorithm. We use the infection data from several regions to train and test the model, and prove its effectiveness and stability. We further give interpretations and explanations for each parameter.

The pandemic situation is the same for every country, and it will experience an outbreak and gradually disappear. Judging from the experiences of China, South Korea, Italy, and Spain, the sooner people pay attention, the sooner they will take corresponding measures, and the stricter the measures taken, the more beneficial it will be for the prevention and control of the virus. At present, India is in an outbreak period. It is necessary to strengthen response and control measures and reduce the number of final infections in order to minimize losses.

Declarations

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Availability of data and materials: The authors used publicly accessible data published by National Health Commissions of China (http://www.nhc.gov.cn/xcs/yqtb/list_gzbd.shtml), Hubei Health Commission (http://wjw.hubei.gov.cn/bmdt/ztzl/fkxxgzbdgrfyyq/xxfb/), World Health Organization (https://www.who.int/docs/default-source/coronaviruse/situation-reports), and Johns Hopkins University (https://coronavirus.jhu.edu/map.html).

Competing interests: Not applicable.

Funding: Not applicable.

Author's contributions: LW put forward the model formula, wrote the initial version of the computer program, and was a major contributor in writing the manuscript. DWH designed the parameters in the model formula, co-wrote the initial version of the computer program. KL did the mathematical derivations, and optimized the computer program. XY collected the data, and optimized the computer program. XLY analyzed the model formula. JQ analyzed the data. DXL visualized the data, drew the figures, and reviewed the computer program. Prof. ZYM initiated and organized the research, supervised the results, and reviewed the manuscript. All authors read and approved the final manuscript.

Acknowledgements: Not required.

Reference

1. Chen N. S., Zhou M., Dong X., et al., Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet, 2020, 395: 507-13. https://doi.org/10.1016/S0140-6736(20)30211-7

2. Lu R. J., Zhao X., Li J., et al., Genomic characterization and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet, 2020, 395:565-74. https://doi.org/10.1016/S0140-6736(20)30251-8

3. Kuznetsov, Y. A., Piccardi, C., Bifurcation analysis of periodic *SEIR* and *SIR* epidemic models. *J. Math. Biol.* 32, 109–121 (1994). https://doi.org/10.1007/BF00163027.

4. May R. M., Anderson R. M., Transmission dynamics of HIV infection. Nature, 1987, 326: 137-142.

5. Chowell G., Viboud C., Hyman J. M., et al. The Western Africa Ebola virus disease epidemic exhibits both global exponential and local polynomial growth rates. PLoS Currents, 2015, 7: 1-10. doi: 10.1371/currents.outbreaks.8b55f4bad99ac5c5db3663e916803261

6. http://www.nhc.gov.cn/xcs/yqtb/list_gzbd.shtml.

7. http://wjw.hubei.gov.cn/bmdt/ztzl/fkxxgzbdgrfyyq/xxfb/

8. WHO. Coronavirus disease (COVID-2019) situation reports [EB/OL]. [2020-04-17] https://www.who.int/docs/default-source/coronaviruse/situation-reports

9. COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU), https://coronavirus.jhu.edu/map.html

10. Diederik P. K., Jimmy B., Adam: A Method for Stochastic Optimization, 3rd International Conference on Learning Representations, ICLR 2015

Figures

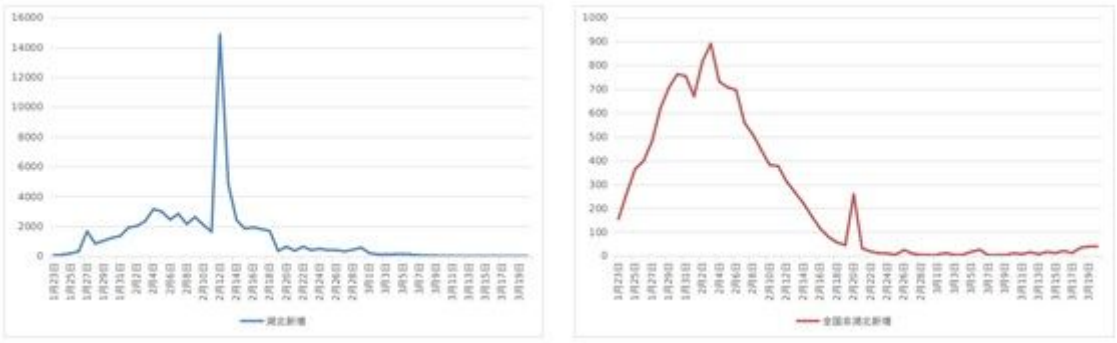


Figure 1

Daily new infections in Hubei Province (left), and daily new infections in non-Hubei parts of China (right).

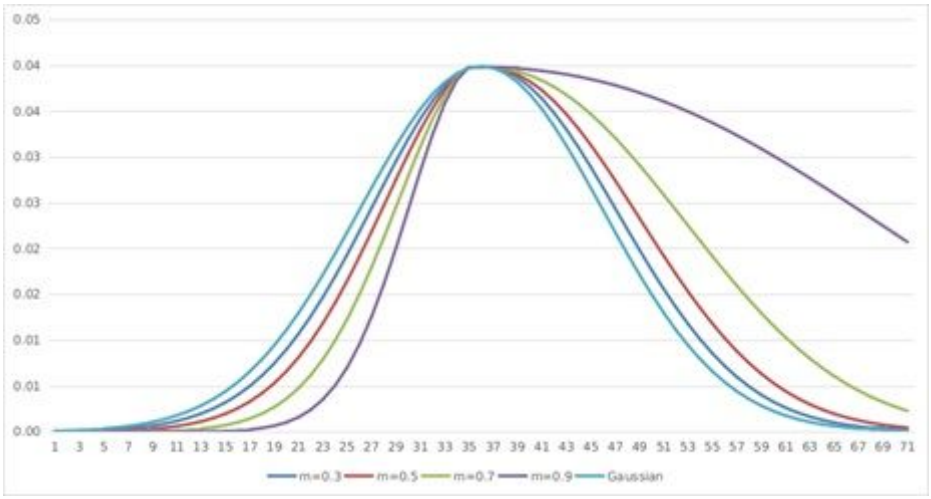


Figure 2

Illustration of H-Gaussian curves with different values of m. When m=0, the H-Gaussian model is equivalent to Gaussian model. Other parameters of the H-Gaussian model are set as a=1,σ=1,μ=0,C=0.

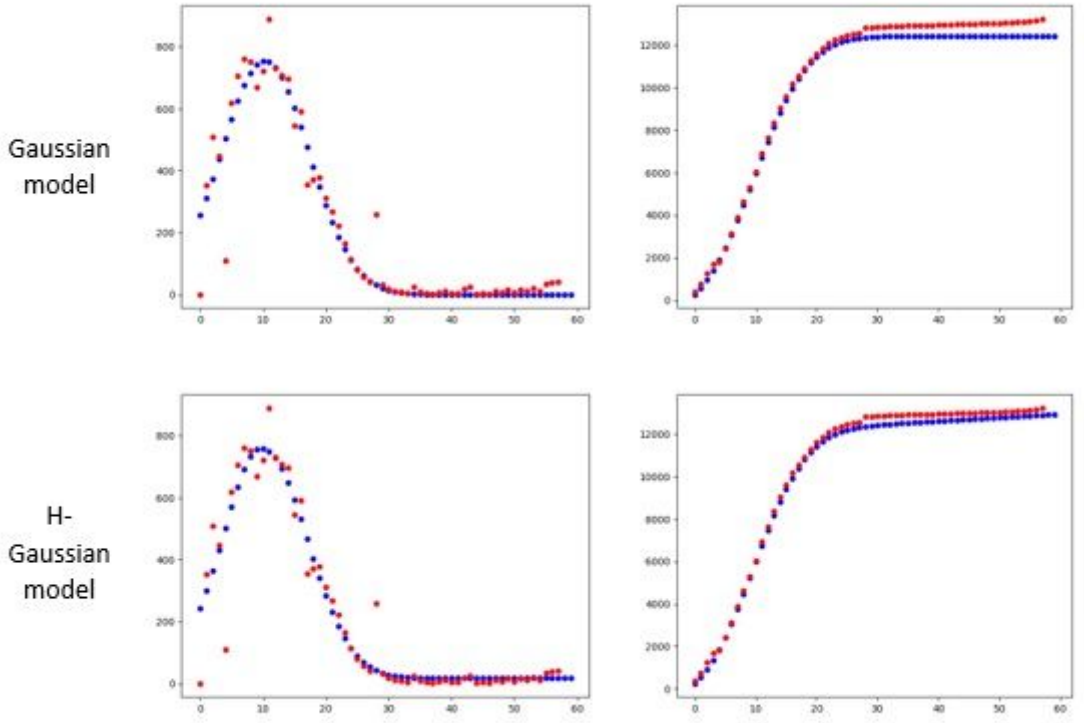


Figure 3

Comparison of the fitting by the two models on the pandemic data in non-Hubei parts of China. Left column: number of daily new infectious people versus time. Right column: number of cumulative infectious people versus time.

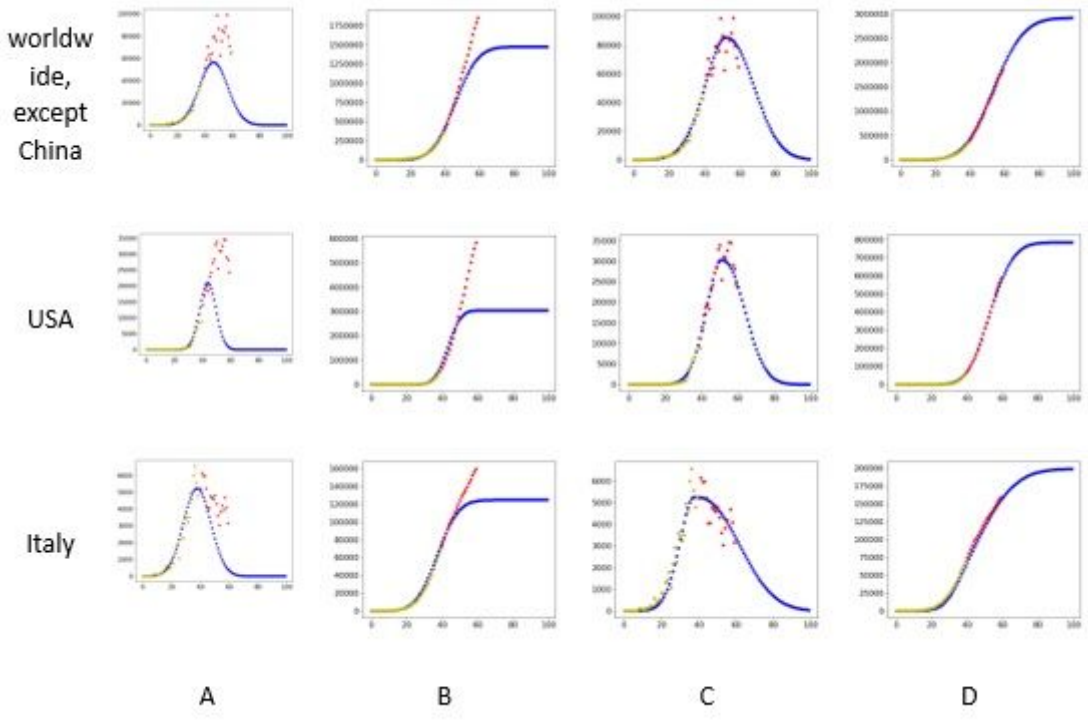


Figure 4

Comparison of the fitting effects of the Gaussian model and H-Gaussian model on the pandemic data worldwide (except China), the USA, and Italy. The yellow dots represent the training data of 41 days, the red dots represent the test data of the subsequent 20 days, and the blue dots represent the predicted values. Column A: number of daily new infections given by the Gaussian models. Column B: number of daily cumulative infections given by the Gaussian models. Column C: number of daily new infections given by the H-Gaussian models. Column D: number of daily cumulative infections given by the H-Gaussian models.

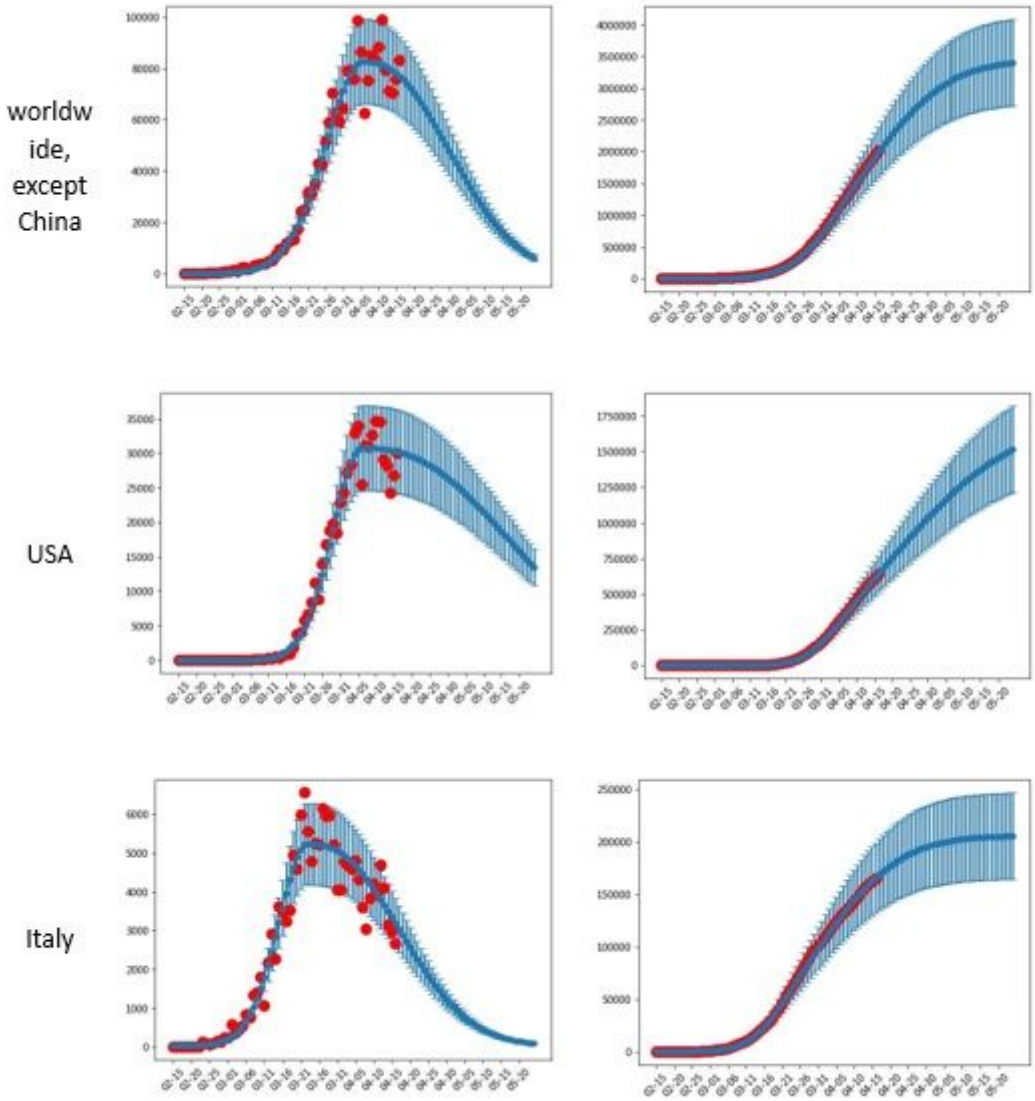


Figure 5

the fitting and prediction of H-Gaussian models trained on the data of 61days in several regions. The red dots represent the training data, the blue dots represent the fitting and predictions given by the H-Gaussian model with the optimized parameters, and the error bars represent the fitting and predictions

given with extreme parameters. Left column: daily number of new infections. Right column: daily number of accumulative infection.

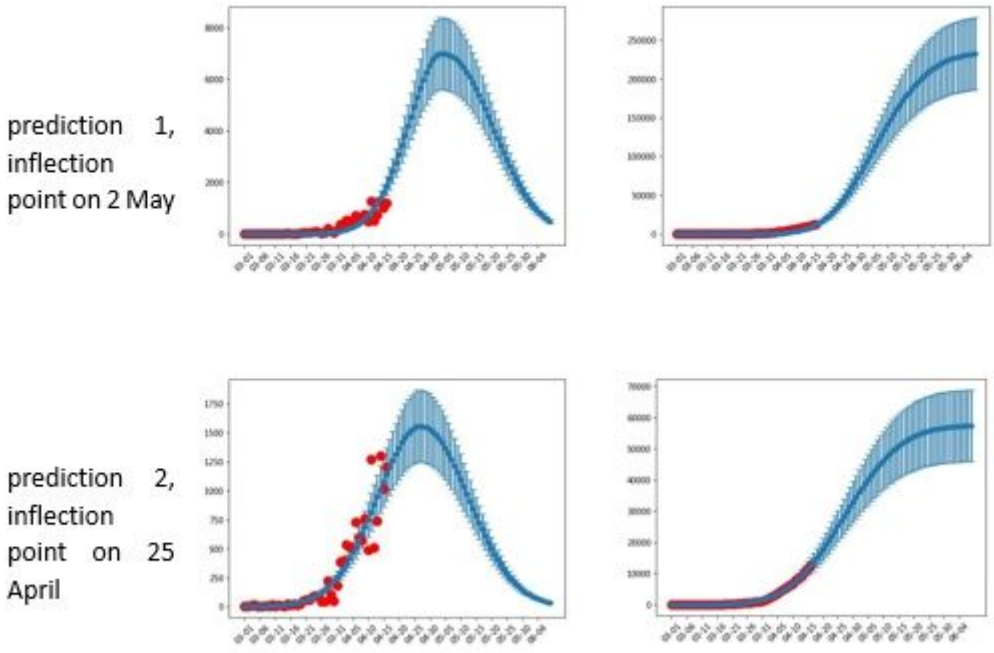


Figure 6

Gaussian model predictions of India's pandemic with different inflection points. Left column: daily number of new infections. Right column: daily number of accumulative infection.