Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Perspective

Big data assimilation to improve the predictability of COVID-19

Xin Li a,b,c,x, Zebin Zhao c,d, Feng Liu d

a National Tibetan Plateau Data Center, Key Laboratory of Tibetan Environmental Changes and Land Surface Processes, Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing 100101, China
b Center for Excellence in Tibetan Plateau Earth Sciences, Chinese Academy of Sciences, Beijing 100101, China
c University of Chinese Academy of Sciences, Beijing 100049, China
d Key Laboratory of Remote Sensing of Gansu Province, Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou 730000, China

HIGHLIGHTS

• Accurate prediction of COVID-19 is essential for achieving the United Nations SDGs.
• Predictability of COVID-19 is improved by data assimilation and parameter estimation.
• A data assimilation based COVID-19 forecasting system is developed.

ABSTRACT

The global outbreak of COVID-19 requires us to accurately predict the spread of disease and decide how adopting corresponding strategies to ensure the sustainable development. Most of the existing infectious disease forecasting methods are based on the classical Susceptible-Infectious-Removed (SIR) model. However, due to the highly nonlinearity, nonstationarity, sensitivities to initial values and parameters, SIR type models would produce large deviations in the forecast results. Here, we propose a framework of using the Markov Chain Monte Carlo method to estimate the model parameters, and then the data assimilation based on the Ensemble Kalman Filter to update model trajectory by cooperating with the real time confirmed cases, so as to improve the predictability of the pandemic. Based on this framework, we have developed a global COVID-19 real time forecasting system. Moreover, we suggest that big data associated with the spatiotemporally heterogeneous pathological characteristics, and social environment in different countries should be assimilated to further improve the COVID-19 predictability. It is hoped that the accurate prediction of COVID-19 will contribute to the adjustments of prevention and control strategies to contain the pandemic, and help achieve the SDG goal of “Good Health and Well-Being”.

1. Global cooperation to fight COVID-19

The global outbreak of the coronavirus disease 2019 (COVID-19) requires us to collaboratively fight against it and ensure the achievement of the United Nations Sustainable Development Goals (SDGs) (United Nations, 2020). Since December 2019, the global outbreak of COVID-19 has affected the health of tens of millions of people around the world, and is likely to affect hundreds of millions in the near future. It will profoundly change the global socioeconomic structure and have political impacts (Jiao et al., 2020; Zhao et al., 2020a). COVID-19 has become a truly global, systemic and cascading emergency, posing a real threat to global sustainable development. In retrospect, the global response to this challenge has not received the attention it deserves. From a long-term perspective, it is necessary to formulate a new special target based on COVID-19 within the framework of the United Nations SDGs to respond to large-scale health problem. Promoting healthy societies can

---

x Corresponding author: Prof. Dr. Xin Li, National Tibetan Plateau Data Center, Key Laboratory of Tibetan Environmental Changes and Land Surface Processes, Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing 100101, China.

E-mail address: xinli@itpcas.ac.cn (X. Li).

https://doi.org/10.1016/j.geosus.2020.11.005
2666-6839/©2020 The Authors. Published by Elsevier B.V. and Beijing Normal University Press (Group) Co., LTD. on behalf of Beijing Normal University. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
expand the scope of SDG 3 (Good Health and Well-being) (Guo, 2020). The normal development of the current pandemic requires that we adopt global cooperation and solidarity, scientific and technological innovation, decision-making based on scientific knowledge, strengthening and combining our efforts (United Nations, 2020), and "coordination, classification, and collaboration" to tackle this challenge (Fu et al., 2020; Zhao et al., 2020a).

2. Data assimilation to improve the predictability of COVID-19

Reasonable epidemic prediction can provide the total number of infections, the lifecycle of the epidemic, the arrival time of the epidemic peak, and the assessment of the epidemic severity. In addition, model forecasting provides the scientific foundation for decision-making and intervention strategy adjustment. Since the outbreak of COVID-19, various models have been proposed to simulate, analyze, and predict the pandemic (Wu et al., 2020; Chinazzi et al., 2020). With the increasingly severe COVID-19 spread, infectious disease models have played key roles in the prediction of pandemic trends, scientific prevention and control, and outbreak assessment. Studies of infectious disease models have a long history and can be traced back to the 1760s, with the famous Susceptible-Infectious-Removed (SIR) model established in the 1920s (Kermack and McKendrick, 1927). The SIR model divided the population under natural conditions into three categories, namely, susceptible individuals, infected cases and removed cases.

Most of the existing infectious disease forecasting methods are based on the classical SIR model. The SIR-type model can be used to accurately predict uncontrolled disease spread at the early stage since it has a clear pathological dynamic mechanism. However, for medium- and long-term epidemic spread (such as 60 days or more), SIR-type models have poor predictability, which can be defined as the qualitative or quantitative correctness of the system predictions. The reasons are that the epidemic characteristics and transmission patterns vary by region and season (Zhan et al., 2019). The parameters of each SIR-type model are spatial and time variant at different stages, i.e., affected by changes in pathological characteristics, the social environment, the human intervention for prevention and control measures (Tian et al., 2020), medical conditions, and self-protection awareness (Scarpino and Petri, 2019). In addition, the stochasticity and nonstationarity of the models indicate that the prediction of SIR-type models is sensitive to the initial value of the model and the total population in the region. Therefore, due to the nonlinearity, heterogeneity, and randomness of the epidemic dynamics, and the significant impact of prevention and control policies, a better epidemic forecasting method to improve predictability is required, both for volatile and normal epidemic spreads.

We propose that data assimilation combined with parameter estimation can improve the predictability of COVID-19: the parameters are optimally estimated based on real-time epidemic data before simulation; then, the trajectory of the infectious disease model can be adjusted dynamically by assimilating real-time observations such as confirmed infected cases, therefore reducing the accumulated errors of the prediction system. This approach is similar to real-time numerical weather predictions. Data assimilation is rooted in and developed based on estimation theory, cybernetics, and chaos theory, fusing model simulation and true observations, and then enhancing the system’s predictability and observability (Li et al., 2020b).

Data assimilation has been used to predict the spread of infectious diseases and provide a theoretical reference for prevention based on science. A main advance of the data assimilation technique is that it combines dynamics with real-time observation data to recursively optimize both state variables and parameters of the epidemic model to improve disease spread predictability (Shaman et al., 2013; Zhan et al., 2019). A successful application of data assimilation to improve the prediction of COVID-19 is manifested in a recent work, which combined the Susceptible Exposed Infectious Recovered (SEIR) model and the Ensemble Adjustment Kalman Filter with a Bayesian inference framework to explain the impact of undiagnosed cases in Wuhan on the rapid spread of COVID-19 (Li et al., 2020a).

Additionally, there are significant uncertainties in the cognition of dynamic infectious disease model parameters that are affected by the pathological characteristics of the virus and human activities, such as nonpharmaceutical interventions; therefore, reasonable parameter estimation is crucial for improving model accuracy (King et al., 2008). Parameter estimation for infectious disease models is usually conducted in a Bayesian framework or by heuristic optimization. For example, the Bayesian inference and Markov Chain Monte Carlo (MCMC) algorithm have been successfully introduced in SIR and SEIR models to predict COVID-19 spread in Wuhan, China (Rodà et al., 2020). Furthermore, the particle swarm optimization method was utilized to estimate SEIR model parameters (He et al., 2020), and the iterative filtering method using the maximum likelihood function was applied in the prediction of COVID-19 spread (Fox et al., 2020).

Overall, compared with the previous pandemic cases studies (Tian et al., 2020; Yang et al., 2020), we aim to integrate the advantages of both data assimilation and parameter estimation to improve the predictability of COVID-19 and reduce forecast uncertainty, and further enable real-time COVID-19 prediction.

3. Real time COVID-19 prediction

We integrated data assimilation, parameter estimation, and infectious disease models to predict COVID-19 spread. We retrospectively forecasted the COVID-19 outbreak in Wuhan by using Common software for Data Assimilation (ComDA) (Liu et al., 2020b), which not only integrates the classic algorithms of data assimilation and parameter estimation, i.e., Metropolis-Hasting (MH) (Zhu et al., 2014) and Ensemble Kalman Filter (EnKF), but also employs the contact network model and SEIR model as the model operators. In addition, we successfully reproduced the "Diamond Princess" epidemic by employing Bayesian inference and MH parameter estimation methods combined with an infectious disease model. The MH algorithm repeatedly learns the daily confirmed cases by sampling and iterating in the multidimensional parameter space to obtain the optimal estimation of the parameters (Liu et al., 2020a). At the same time, the MH algorithm has been used in the improved SEIR model to predict the pandemic in six African countries in different intervention scenarios, and the validations are consistent with the real pandemic within 20 days of the initial stage (Zhao et al., 2020b). These results showed that our strategies can improve the predictability of pandemic and reduce forecast uncertainty.

We then developed a COVID-19 prediction system based on real-time operational data assimilation and parameter estimation. We used MCMC sampling to estimate the model parameters and EnKF to continuously update the model trajectory by assimilating the latest pandemic data into the dynamic framework for seven-day nowcasting and two-month forecasting (Fig. 1). The forecasting was further facilitated by scenario analysis with three assumed non-pharmaceutical interventions. Moreover, a pandemic forecast information platform (http://covda.tpdc.ac.cn/#all) was successfully launched and is capable of predicting the spread of COVID-19 in 202 countries/regions all over the world. The epidemiological data (i.e., the population densities, daily and cumulative confirmed cases) were tracked by country/region both in the plots and in a geographic information dashboard. The forecasting has achieved ideal results with an average prediction accuracy of approximately 77% for the next 7 days. Therefore, this forecasting system can provide references for the adjustment of relevant prevention and control measures. This system is different with existing real-time visualization systems, such as the Johns Hopkins University visualization system, but our study is more inclined to forecast the spread of COVID-19 for a period of time in the future, with a simple and clear interface that makes it easy to view the trend of COVID-19 by country.
4. Summary and prospect

Big data innovation is particularly important for combating COVID-19 to ensure sustainable development of human society (Murdoch and Detzky, 2013; Pu, 2020). Big data from social media, mobile data and other sources, contains more sophisticated information than daily confirmed cases, particularly on the spatiotemporal variation of social distancing, population mobility, and pathological characteristics in different countries. These data provide timely and useful information for a more reasonable parameterization of the COVID-19 prediction model (Wu et al., 2020; Zhou et al., 2020), constrain data assimilation and parameter estimation, and enable the design of more realistic scenarios for long-term prediction.

In summary, data assimilation, and parameter estimation can help tackle COVID-19 by enhancing the predictability of the pandemic. Meanwhile, the adjustment of epidemic trajectory by introducing real time big data of pathological characteristics and social environment poses a promising way toward big data assimilation. We hope that the reliable prediction of COVID-19 will contribute to the adjustment of prevention and control strategies to combat the pandemic, ensuring human health, and achieving the goal of “Good Health and Well-Being” established by the United Nations SDG.

Declaration of Competing Interest

The authors declare that they have no conflict of interest.

Acknowledgements

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDA20100104), the Science-based Advisory Program of the Alliance of International Science Organizations (Grant No. ANSO-SBA-2020-07), the National Natural Science Foundation of China (Grant No. 41801270), and the Foundation for Excellent Youth Scholars of NIEER, CAS.

References

Chinazzi, M., Davis, J., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S., Piontti, A., Mu, K., Rossi, L., Sun, K., Viboud, C., Xiong, X., Yu, H., Halloran, M., Jr., I., Vespi-gnani, A., 2020. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science 368 (6489), 395–400.
Fox, S., Pasco, R., Tec, M., Du, Z., Lachmann, M., Scott, J., Meyers, L., 2020. The impact of asymptomatic COVID-19 infections on future pandemic waves. medRxiv. https://doi.org/10.1101/2020.06.22.20137489.
Pu, B., 2020. Promoting geography for sustainability. Geogr. Sustain. 1 (1), 1–7.
Pu, B., Zhang, J., Wang, S., Zhao, W., 2020. Classification—coordination—collaboration: A systems approach for advancing sustainable development goals. Natl. Sci. Rev. 7 (5), 626–840.
Gao, H., 2020. Big data platforms for a global pandemic. https://sustainabledevelopment.un.org/?page=view&nr=1656&type=230&menu=2059. (accessed 2 Apr 2020).
He, S., Peng, Y., Sun, K., 2020. SEIR modeling of the COVID-19 and its dynamics. Nonlinear Dyn. 101 (3), 1667–1680.
Jiao, N., Chen, F., Hou, Z., 2020. Combating climate change in a post-COVID-19 era. Sci. Bull. 65 (23), 1958–1960.
Kermack, W., McKendrick, A., 1927. A contribution to the mathematical theory of epidemics. In: Proceedings of The Royal Society of London, 115. Royal Society, UK, pp. 700–721.
King, A., Ionides, E., Pascual, M., Bouma, M., 2008. Inapparent infections and cholera dynamics. Nature 454 (7206), 877–880.
Li, R., Pet, S., Chen, B., Song, Y., Zhang, T., Yang, W., Shaman, J., 2020a. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). Science 368 (6490), 489–493.
Li, X., Liu, F., Fang, M., 2020b. Harmonizing models and observations: Data assimilation in Earth system science. Sci. China Earth Sci. 63 (8), 1059–1068.
Liu, F., Li, X., Zhu, G., 2020a. Using the contact network model and Metropolis-Hastings sampling to reconstruct the COVID-19 spread on the “Diamond Princess”. Sci. Bull. 65 (15), 1297–1305.
Liu, F., Wang, L., Li, X., Huang, C., 2020b. ComDA: A common software for nonlinear and non-Gaussian land data assimilation. Environ. Modell. Softw. 127, 104638.

Murdoch, T., Detsky, A., 2013. The inevitable application of big data to health care. Jama 309 (13), 1351–1352.

Roda, W., Varughese, M., Han, D., Li, M., 2020. Why is it difficult to accurately predict the COVID-19 epidemic? Infect. Dis. Modell. 5, 271–281.

Scarpino, S., Petri, G., 2019. On the predictability of infectious disease outbreaks. Nat. Commun. 10 (1), 1–8.

Tian, H., Liu, Y., Li, Y., Wu, C., Chen, B., Kraemer, M., Li, B., Cai, J., Xu, B., Yang, Q., Wang, B., Yang, P., Cui, Y., Song, Y., Zheng, P., Wang, Q., Bjornstad, O., Yang, R., Grenfell, B., Pybus, O., Dye, C., 2020. An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. Science 368 (6491), 638–642.

Shaman, J., Karspeck, A., Yang, W., Tamerius, J., Lipsitch, M., 2013. Real-time influenza forecasts during the 2012–2013 season. Nat. Commun. 4, 2837.

United Nations, 2020. The Sustainable Development Goals Report 2020. https://unstats.un.org/sdgs/report/2020/ (accessed 7 July 2020).

Wu, J., Leung, K., Leung, G., 2020. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: A modelling study. The Lancet 395 (10225), 689–697.

Yang, Z., Zeng, Z., Wang, K., Wong, S., Liang, W., Zanin, M., Liu, P., Cao, X., Gao, Z., Mai, Z., Liang, J., Liu, X., Li, S., Li, Y., Ye, F., Guan, W., Yang, Y., Li, F., Luo, S., Xie, Y., Liu, B., Wang, Z., Zhang, S., Wang, Y., Zhong, N., He, J., 2020. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. J. Thorac. Dis. 12 (3), 165–174.

Zhan, Z., Dong, W., Lu, Y., Yang, F., Wang, Q., Jia, P., 2019. Real-Time Forecasting of Hand-Foot-and-Mouth Disease Outbreaks using the Integrating Compartment Model and Assimilation Filtering. Sci. Rep. 9 (1), 1–9.

Zhao, W., Zhang, J., Meadows, M., Liu, Y., Hua, T., Fu, B., 2020a. A systematic approach is needed to contain COVID-19 globally. Sci. Bull. 65 (11), 876–878.

Zhao, Z., Li, X., Liu, F., Zhu, G., Ma, C., Wang, L., 2020b. Prediction of the COVID-19 spread in African countries and implications for prevention and control: A case study in South Africa, Egypt, Algeria, Nigeria, Senegal and Kenya. Sci. Total Environ. 729, 138959.

Zhou, C., Su, F., Pei, T., Zhang, A., Du, Y., Luo, B., Cao, Z., Wang, J., Yuan, W., Zhu, Y., Song, C., Chen, J., Xu, J., Li, F., Ma, T., Jiang, L., Yan, F., Yi, J., Hu, Y., Liao, Y., Xiao, H., 2020. COVID-19: Challenges to GIS with big data. Geogr. Sustain. 1 (1), 77–87.

Zhu, G., Li, X., Su, Y., Zhang, K., Bai, Y., Ma, J., Li, C., Hu, X., He, J., 2014. Simultaneous parameterization of the two-source evapotranspiration model by Bayesian approach: Application to spring maize in an arid region of northwest China. Geosci. Model Dev. 7, 1467–1482.