Atmospheric and Oceanic Science Letters 14 (2021) 100019

Contents lists available at ScienceDirect

Atmospheric and Oceanic Science Letters

journal homepage:
http://www.keaipublishing.com/en/journals/atmospheric-and-oceanic-science-letters/

Improvement of the global prediction system of the COVID-19 pandemic based on the ensemble empirical mode decomposition (EEMD) and autoregressive moving average (ARMA) model in a hybrid approach
基于集合经验模态分解和自回归-移动平均模型的 COVID-19 流行病全球预测系统预测结果改进

Chuwei Liu, Jianping Huang*, Fei Ji, Li Zhang, Xiaoyue Liu, Yun Wei, Xinbo Lian

Collaborative Innovation Center for Western Ecological Safety, Lanzhou University, Lanzhou, China

ARTICLE INFO

Keywords:
COVID-19
prediction
hybrid EEMDARMA method
historical data
关键词:
COVID-19
预测
EEMD–ARMA混合方法
历史数据

ABSTRACT

In 2020, the COVID-19 pandemic spreads rapidly around the world. To accurately predict the number of daily new cases in each country, Lanzhou University has established the Global Prediction System of the COVID-19 Pandemic (GPCP). In this article, the authors use the ensemble empirical mode decomposition (EEMD) model and autoregressive moving average (ARMA) model to improve the prediction results of GPCP. In addition, the authors also conduct direct predictions for those countries with a small number of confirmed cases or are in the early stage of the disease, whose development trends of the pandemic do not fully comply with the law of infectious diseases and cannot be predicted by the GPCP model. Judging from the results, the absolute values of the relative errors of predictions in countries such as Cuba have been reduced significantly and their prediction trends are closer to the real situations through the method mentioned above to revise the prediction results out of GPCP. For countries such as El Salvador with a small number of cases, the absolute values of the relative errors of prediction become smaller. Therefore, this article concludes that this method is more effective for improving prediction results and direct prediction.

摘要

2020年, 新型冠状病毒肺炎 (COVID-19) 在世界范围内迅速传播.为准确预测各国每日新增发病人数, 兰州大学开发了 COVID-19 流行病全球预测系统 (GPCP). 在本文的研究中, 我们使用集合经验模态分解 (EEMD) 模型和自回归-移动平均 (ARMA) 模型对 GPCP 的预测结果进行改进, 并对发病人数较少或处于发病初期, 不完全符合传染病规律, GPCP 模型无法预测的国家进行直接预测.从结果来看, 使用该方法修正预测结果, 古巴等国家预测误差均大幅下降, 且预测趋势更接近真实情况.对于萨尔瓦多等发病人数较少的国家直接进行预测, 相对误差较小, 预测结果较为准确.该方法对于改进预测结果和直接预测均较为有效.

1. Introduction

Coronavirus disease 2019 (COVID-19) is a novel infectious disease caused by a virus closely related to the SARS (severe acute respiratory syndrome) virus. COVID-19 has caused hundreds of thousands of deaths worldwide and was declared a global pandemic by the World Health Organization (WHO) on 11 March 2020 (WHO, 2020a). The COVID-19 pandemic has far-reaching consequences beyond the spread of the disease itself; it also has influence on quarantine measures, including political, cultural, and social implications. The WHO World Health Assembly made a global commitment to unite the world to fight COVID-19 (WHO, 2020b). The potential effects of COVID-19 have prompted extensive research to study the characteristics of the virus. Because the virus is new, it is challenging to predict when this disease will disappear. However, it has been found that about 60% of confirmed global COVID-19 cases have occurred in places with temperatures of 5 °C–15 °C. The pandemic spread to high latitudes in spring and summer, and countries located in mid-latitudes face the risk of a second wave of COVID-19 this autumn (Huang et al., 2020a). Therefore, short-term prediction is critical to better manage the societal, economical, cultural, and public health consequences of the pandemic (Petropoulos and Makridakis, 2020), especially in high-risk countries.

Researchers worldwide have been predicting the development of the outbreak by using existing mathematical and statistical methods, includ-

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

https://doi.org/10.1016/j.aosl.2020.100019
Received 24 June 2020; Revised 26 August 2020; Accepted 8 September 2020
Available online 9 December 2020

ing stochastic simulations, lognormal distribution (Linton et al., 2020), machine learning, and artificial intelligence (Tuli et al., 2020). The SEIR (susceptible, exposed, infectious, and removed) and SIR (susceptible, infectious, and removed) infectious disease models are the most widely used (Wu et al., 2020; Wang et al., 2020; Yang et al., 2020). A global prediction system (Global Prediction System of the COVID-19 Pandemic; GPCP) based on the SIR model was recently developed (Huang et al., 2020b). The system determines the parameters through historical data fitting, which allows it to make targeted predictions for various countries and obtain better prediction results. However, the development of the epidemic is complicated, and there are differences between the prediction results of the GPCP system model and the real data; thus, the results need further revision.

Various methods have been used to revise prediction results. For example, the analogue–dynamical approach is used to revise weather forecast models (Zheng et al., 2013; Yu et al., 2014). To modify the model results, it is necessary to analyze and predict the difference between the predicted results and the true values (forecast residuals). The residuals fitted by the model to historical data are nonstationary and nonlinear. In this study, we used the ensemble empirical mode decomposition (EEMD) method, which is an adaptive and temporal local data analysis method (Wu et al., 2007; Wu and Huang, 2009). EEMD is a time series analysis method based on the empirical mode decomposition (EMD) method (Huang et al., 1998; Huang and Wu, 2008), which decomposes complicated data series into finite quasi-periodic components at different frequencies and is suitable for adaptive analysis of nonlinear and nonstationary time series. The EMD/EEMD method has been used to analyze nonlinear and nonstationary data in climatic and oceanic analyses (Wu et al., 2011; Ji et al., 2014; Chen et al., 2017) and for biomedical signal processing (Colominas et al., 2014).

Methods to predict time series include support vector machines (Wang et al., 2010), artificial neural networks (Jiang et al., 2003), and genetic programming (Koutroumanidis et al., 2009). Box and Jenkins introduced a time series analysis approach called the autoregressive moving average (ARMA) method (Box and Jenkins, 1976), which combines the advantages of the autoregressive (AR) model and the moving average (MA) model. The AR model quantifies the relationship between current data and previous data, and the MA model solves the problem of stochastically changing terms. The ARMA model has been applied to forecasting meteorological elements (Torres et al., 2005) and macroeconomic evolution (Anghelache et al., 2016). The model only needs time series data, so the residuals' prediction of the infectious disease model can be better applied in it.

In this paper, we report upon work to improve the GPCP by applying the ARMA and EEMD methods to the results of the SIR model for the number of new cases in each country. We then use the method to predict the number of new cases in countries with fewer cases.

2. Data and methods

2.1. Data

We used the cumulative number of cases from the COVID-19 Data Repository published by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (https://github.com/CSSEGISandData/COVID-19). The number of new cases is the difference between the cumulative number of cases on the current day and the cumulative number of cases on the previous day. The fitting and prediction data are from the GPCP system, and temperature in the model was ignored.

2.2. EEMD method

Based on the EMD method, the EEMD method has various improvements (Wu and Huang, 2009). White noise is added to the original sequence, and the sequence is decomposed into a set of amplitude-frequency-modulated oscillatory components (intrinsic mode functions). These steps are repeated using a different white noise sequence each time, and the corresponding intrinsic mode functions are obtained as the final decomposition result. The detailed procedures can be found in previous studies (Huang et al., 1998; Huang and Wu, 2008; Wu and Huang, 2009).

We first performed seven-point smoothing on the original residual sequence and EEMD decomposition on the smoothed sequence. The ratio of the additional noise to the standard deviation of the original sequence was set to 0.1 and repeated 100 times.

2.3. ARMA method

For the p-order AR (p) model, the current value of the time series is expressed as follows (Box and Jenkins, 1976; Wang et al., 2015):

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t. \tag{1}$$

For the q-order MA (q) model, q previous values expressed as random errors, and the current value of the time series is expressed as follows:

$$y_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}. \tag{2}$$

From the above, the ARMA model is expressed as follows:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}, \tag{3}$$

where y_t is the predictive value, ϕ_i is the correlation coefficient with each previous value, θ_i is the correlation coefficient with the previous white noise, ε_t is the white noise process with zero mean and variance, and ε_{t-i} is the previous noise term.

2.4. Hybrid EEMD–ARMA method

From the perspective of fitting the real data, the residuals of the model fitting and the real data are nonstationary and nonlinear time series. The EEMD method can extract signals from such sequences and decompose them into different oscillatory components. The GPCP system parameters are obtained by fitting real data, and the parameters in the GPCP system are fitted from the real situation, so the initial model prediction mainly represents the trend which depends on human factors such as government policies. Therefore, the residuals basically reflect the oscillation of the infectious disease and are suitable for the EEMD method. The ARMA method's prediction only depends on the time series, and no other information is needed. In addition, the increase in cases in countries with a small number of cases tends to show an increase in oscillation, a decrease in oscillation, or a stable oscillation within a certain range without obvious peaks, so the SIR model is not applicable. Owing to the advantages of the EEMD and ARMA methods, we propose a hybrid EEMD–ARMA method, which is motivated by the idea of "decomposition and ensemble" (Yu et al., 2008; Guo et al., 2012). The seven-point smoothed original residual sequence is decomposed into several subsequences by the EEMD method. Each of subsequence is then predicted by the ARMA method, and the final predicted value can be obtained by summing the predicted values of each subsequences. The hybrid EEMD–ARMA method has a better effect on the high-frequency oscillations. Therefore, in our work, we used the hybrid EEMD–ARMA method to process sequences containing high-frequency oscillations. To improve the quality of the prediction, for the residual sequence, the first few days before the number of newly added cases reaches the peak are

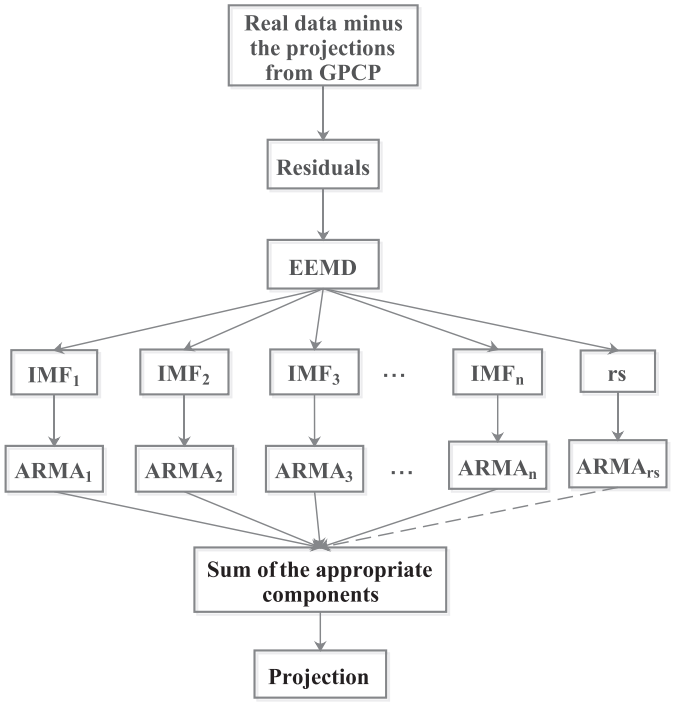

Fig. 1. Flowchart of the prediction of the residuals (add the dotted line when directly predicting the number of new people, and remove the dotted line when predicting the residuals).

selected as the starting time of the sequence. For the predicted number of new cases series, the days around which the number of cases starts to fluctuate are selected as the starting time of the series. The procedures for the hybrid EEMD–ARMA predicting method are shown in Fig. 1.

Seven-point smoothing is performed on the residual sequence of the fitting result. The smoothed sequence is decomposed by the EEMD method, and the residual difference component is removed. The first-order difference is calculated compared with other components, and then the ARMA model is used to predict each component. The prediction results of the appropriate components are selected for summation as the final residual prediction result.

For countries with a small number of cases, after the original sequence of newly added cases is decomposed by EEMD, the components are not removed and the ARMA predictions are made for each component directly, yielding the prediction results.

2.5. Calculation of relative error

The relative error is calculated as

$$\text{Relative error} = \frac{\text{ave}(\text{Result}_f) - \text{ave}(\text{Result}_r)}{\text{ave}(\text{Result}_r)} \times 100\%, \quad (4)$$

where $\text{ave}(\text{Result}_f)$ is the average of the prediction for several days, and $\text{ave}(\text{Result}_r)$ is the average of the real data for several days.

3. Results

3.1. Prediction improvements

Fig. 2 shows the improvement of the prediction effect from 6 to 15 May 2020 by some countries using the hybrid EEMD–ARMA method to correct residuals. Judging from the relative error of the 10-day prediction before and after the revision, Italy is the one with the most obvious improvement. The relative error has improved from 83.23% to −10.22%. Netherlands has the best prediction effect after correction; the relative error before correction is 35.65%, and the relative error after correction is reduced to −0.07%. Using the GPCP system for direct prediction, only 15 of the 34 countries listed in Fig. 2 have a relative error with an absolute value of less than 40%. After correction, the number of countries with an absolute value of less than 40% has increased to 24. This method offers great improvements for prediction, and has the potential to be effective for future predictions.

Fig. 3 compares the prediction results of the number of newly increased people before and after the 10-day (6–15 May) correction in six countries (Cuba, Romania, Italy, Spain, Netherlands, and Sweden), and gives the respective relative errors. The peak number of new cases in these countries has already appeared, and some countries are in a steady state (Spain and Sweden). Some countries are in the stage of decline in the number of new cases (Cuba, Romania, Italy, and Netherlands). These countries experience better prediction effect after correction. The absolute values of the relative errors before and after the six countries' corrections decreased by 32.44%, 3.46%, 73.01%, 23.55%, 35.58%, and 21.78%. Judging from the revised results, the relative error of the six countries has been reduced, and the new development trend is more in line with the real situation.

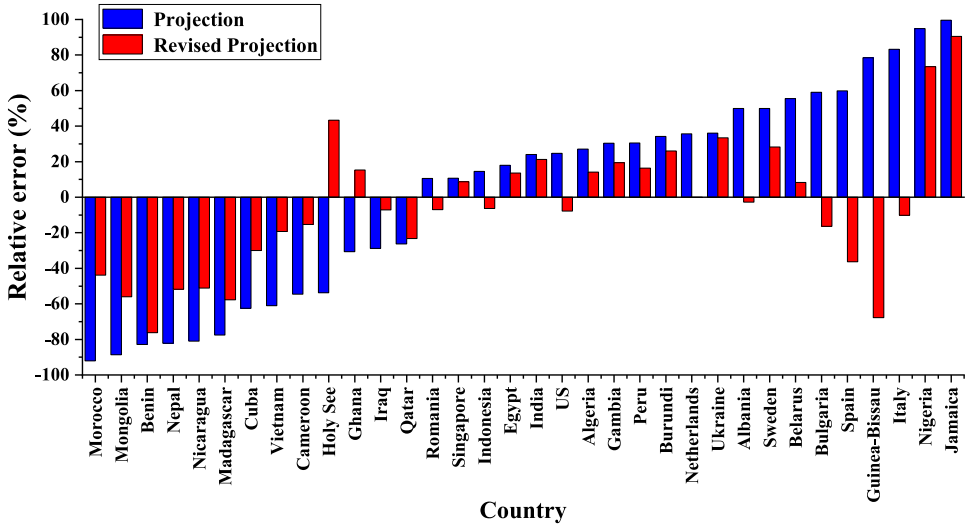

Fig. 2. The relative errors of the hindcast results before and after correction from 6 to 15 May 2020 in some countries.

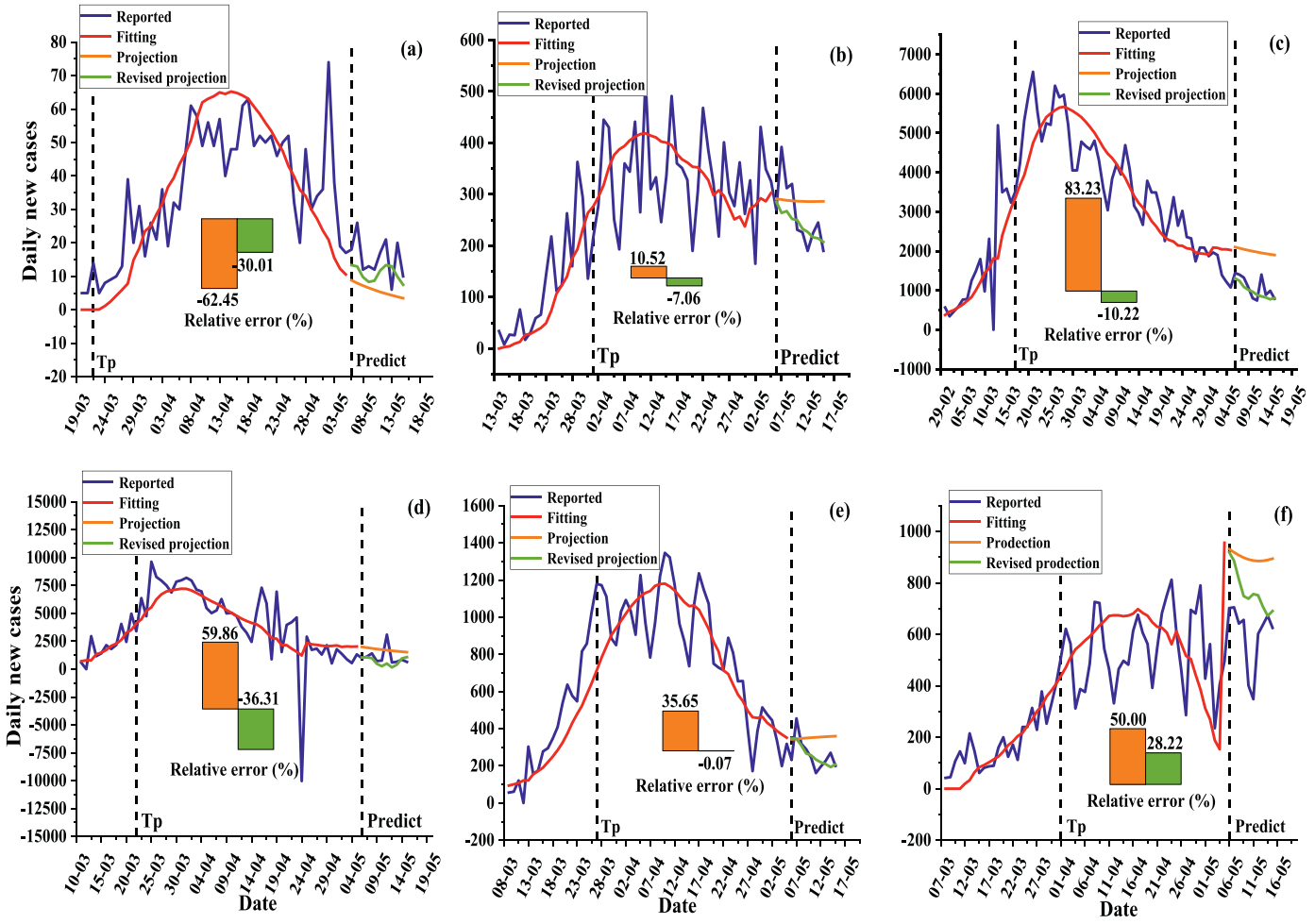

Fig. 3. Projections and relative errors before and after correction in six countries ((a) Cuba, (b) Romania, (c) Italy, (d) Spain, (e) Netherlands, and (f) Sweden) from the date of the emergence of confirmed cases to 15 May 2020. The reported confirmed cases (the date of the emergence of confirmed cases to 15 May) are shown as blue lines, while the historical simulated cases (the date of the emergence of confirmed cases to 5 May) are shown as red lines. The hindcast cases before revising (6–15 May) and after revising (6–15 May) are shown as orange and green lines, respectively. The two black dotted lines represent the time when the original residual sequence was used for correction (Tp) and the time when the prediction starts (Predict). The bar graphs show the relative errors of the results of projections before (orange bars) and after (green bars) correction.

3.2. EEMD–ARMA direct prediction

Fig. 4 shows the prediction results and relative errors of the new cases based on the EEMD–ARMA method for 10 days (6–15 May) in 6 countries. Some of these countries have a relatively small number of new cases (Sri Lanka), and the accuracy of the GPCP system prediction is low. Some countries are still in the stage of rapid increase in the number of new cases, and there has not been a peak (El Salvador, Kuwait, South Africa, Sri Lanka, and Bolivia). The EEMD–ARMA method is used directly for prediction, and the relative error of the 10-day prediction in these six countries is less than 40%. This method performs well in El Salvador (−9.84%), Kuwait (−6.93%), South Africa (−0.09%), Sri Lanka (0.97%), and Bolivia (3.57%), where the fluctuation amplitude is small, and the prediction effect of Sudan (−38.21%), with a sudden increase in amplitude, is slightly worse. Overall, this method has better predicted the change range and trend of the new population in these countries.

4. Conclusion

COVID-19 has spread rapidly and severely affects human health and economic development worldwide. Therefore, it is paramount to accurately predict the development of the epidemic in various countries to provide data for relevant organizations. Overall, the SIR model provides a good prediction, but it has some limitations. For example, there are errors in predictions for countries that enter a decline in new cases after the peak, and the predictions for countries that have not yet reached the peak are less accurate during the increase in cases. To improve our understanding of the global impact of COVID-19 and to better predict the number of COVID-19 cases in different countries, we developed a hybrid EEMD–ARMA method to correct the results of GPCP and make direct predictions for countries with small numbers of daily new cases. Our method provides more accurate and reliable predictions of the spread of COVID-19, and we hope the method will eventually inform strategic government responses.

Based on our results, for cases that used the hybrid EEMD–ARMA method to make corrections and predictions, within 10 days of the back-prediction, the changes and trends in the number of new cases were closer to the actual situation. The relative errors were lower, and the prediction was better. Fighting the epidemic requires a concerted international effort, which we believe will eventually control the disease.

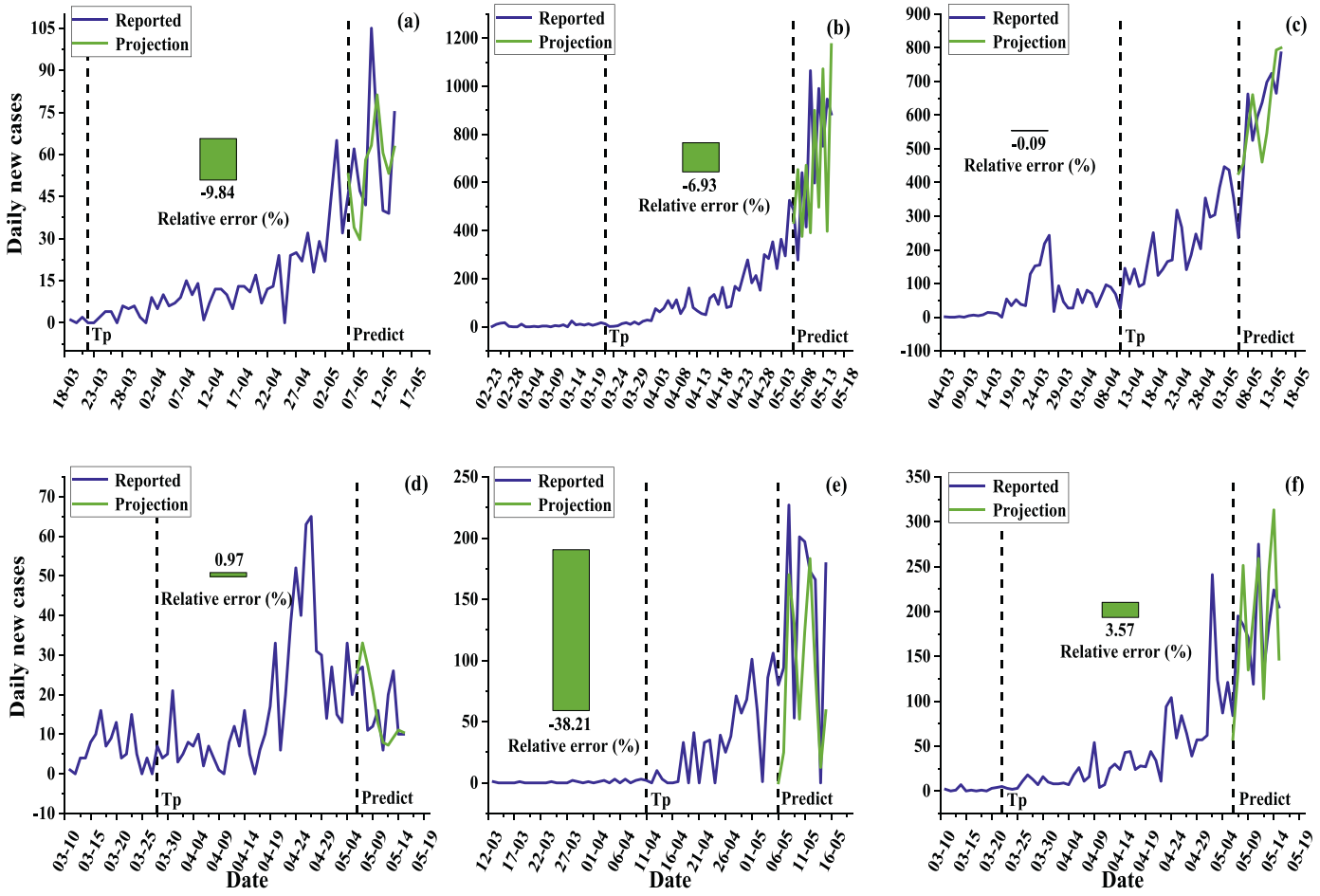

Fig. 4. Projections and relative errors in six countries ((a) El Salvador, (b) Kuwait, (c) South Africa, (d) Sri Lanka, (e) Sudan, and (f) Bolivia) from the date of the emergence of confirmed cases to 15 May 2020. The reported confirmed cases (the date of the emergence of confirmed cases to 15 May) are shown as blue lines, and the hindcast cases (6–15 May) are shown as green lines, respectively. The two black dotted lines represent the time when the original residual sequence was used for correction (Tp) and the time when the prediction starts (Predict). The bar graphs show the relative errors of the results of projections.

Declaration of Competing Interest

No potential conflict of interest was reported by the authors.

Funding

This work was jointly supported by the National Natural Science Foundation of China [grant numbers 41521004 and 41875083] and the Gansu Provincial Special Fund Project for Guiding Scientific and Technological Innovation and Development [grant number 2019ZX-06].

Acknowledgments

The authors acknowledge the CSSE at Johns Hopkins University for providing the COVID-19 data.

References

Anghelache, C., Grabara, J., Manole, A., 2016. Using the dynamic model ARMA to forecast the macroeconomic evolution. Rom. Stat. Rev. Suppl. 64 (1), 3–13.

Box, G.E.P., Jenkins, G.M., 1976. Times Series Analysis: Forecasting and Control. Prentice-Hall, Englewood Cliffs.

Chen, X.Y., Zhang, X.B., Church, J.A., Watson, C.S., King, M.A., Monselesan, D., Legresy, B., Harig, C., 2017. The increasing rate of global mean sea-level rise during 1993-2014. Nat. Clim. Change 7 (7), 492–495. doi:10.1038/nclimate3325.

Colominas, M.A., Schlotthauer, G., Torres, M.E., 2014. Improved complete ensemble EMD: a suitable tool for biomedical signal processing. Biomed. Signal Process. Control 14 (14), 19–29. doi:10.1016/j.bspc.2014.06.009.

Guo, Z.H., Zhao, W.G., Lu, H.Y., Wang, J.Z., 2012. Multi-step forecasting for wind speed using a modified EMD-based artificial neural network model. Renew. Energy 37 (1), 241–249. doi:10.1016/j.renene.2011.06.023.

Huang, J.P., Zhang, L., Liu, X.Y., Wei, Y., Liu, C.W., Lian, X.B., Huang, Z.W., et al., 2020b. Global prediction system for COVID-19 pandemic. Sci. Bull. 65 (22), 1884–1887. doi:10.1016/j.scib.2020.08.002.

Huang, N.E., Shen, Z., Long, S.R., Wu, M.C., Shih, H.H., Zheng, Q., Yen, N.-C., Tung, C.C., Liu, H.H., 1998. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. 454 (1971), 903–995. doi:10.1098/rspa.1998.0193.

Huang, N.E., Wu, Z.H., 2008. A review on Hilbert-Huang transform: method and its applications to geophysical studies. Rev. Geophys. 46 (2), RG2006. doi:10.1029/2007RG000228.

Huang, Z.W., Huang, J.P., Gu, Q.Q., Du, P.Y., Liang, H.B., Dong, Q., 2020a. Optimal temperature zone for the dispersal of COVID-19. Sci. Total Environ. 736, 139487. doi:10.1016/j.scitotenv.2020.139487.

Ji, F., Wu, Z.H., Huang, J.P., Chassignet, E.P., 2014. Evolution of land surface air temperature trend. Nat. Clim. Change 4, 462–466. doi:10.1038/nclimate2223.

Jiang, D.H., Zhang, Y., Hu, X., Zeng, Y., Tan, J.G., Shao, D.M., 2003. Progress in developing an ANN model for air pollution index forecast. Atmos. Environ. 38 (40), 7055–7064. doi:10.1016/j.atmosenv.2003.10.066.

Koutroumanidis, T., Sylaios, G., Zafeiriou, E., Tsihrintzis, V.A., 2009. Genetic modeling for the optimal forecasting of hydrologic time-series. J. Hydrol. Amst. 368, 156–164. doi:10.1016/j.jhydrol.2009.01.041.

Linton, N.M., Kobayashi, T., Yang, Y.C., Hayashi, K., Akhmetzhanov, A.R., Jung, S-m., Yuan, B.Y., Kinoshita, R., Nishiura, H., 2020. 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), 538. doi:10.3390/jcm9020538.

Petropoulos, F., Makridakis, S., 2020. Forecasting the novel coronavirus COVID-19. PLoS ONE 15 (3), e0231236. doi:10.1371/journal.pone.0231236.

Torres, J.L., García, A., Blas, M.D., Francisco, A.D., 2005. Forecast of hourly average wind speed with ARMA models in Navarre (Spain). Sol. Energy 79 (1), 65–77. doi:10.1016/j.solener.2004.09.013.

Tuli, S., Tuli, S., Tuli, R., Gill, S.S., 2020. Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing. Internet Things 11, 100222. doi:10.1016/j.iot.2020.100222.

Wang, G., Qiu, Y.F., Li, H.X., 2010. Temperature forecast based on SVM optimized by PSO algorithm. In: Proceedings of the International Conference on Intelligent Computing and Cognitive Informatics (ICICCI), pp. 259–262. doi:10.1109/ICICCI.2010.24.

Wang, H.W., Wang, Z.Z., Dong, Y.Q., Chang, R.G., Chen, X., Xu, X.Y., Zhang, S.X., et al., 2020. Phase-adjusted estimation of the number of coronavirus disease 2019 cases in Wuhan, China. Cell Discov. 6 (1), 10. doi:10.1038/s41421-020-0148-0.

Wang, W.C., Chau, K.W., Xu, D.M., Chen, X.Y., 2015. Improving forecasting accuracy of annual runoff time series using ARIMA based on EEMD decomposition. Water Resour. Manag. 29 (8), 2655–2675. doi:10.1007/s11269-015-0962-6.

WHO, 2020a. Coronavirus disease (COVID-19) pandemic. Available online: https://www.who.int/zh/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—11-march-2020. (accessed 28/09/2020).

WHO, 2020b. Coronavirus disease 2019 (COVID-19): situation report. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports.

Wu, J.T., Leung, K., Leung, G.M., 2020. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. Lancet 395, 689–697. doi:10.1016/S0140-6736(20)30260-9.

Wu, Z., Huang, N.E., Long, S.R., P., C.-.K., 2007. On the trend, detrending, and variability of nonlinear and nonstationary time series. Proc. Natl. Acad. Sci. 104 (38), 14889–14894.

Wu, Z.H., Huang, N.E., 2009. Ensemble empirical mode decomposition: a noise-assisted data analysis method. Adv. Adapt. Data Anal. 1 (1), 1–41. doi:10.1142/S1793536909000047.

Wu, Z.H., Huang, N.E., Wallace, J.M., Smoliak, B.V., Chen, X.Y., 2011. On the time-varying trend in global-mean surface temperature. Clim. Dyn. 37, 3–4. doi:10.1007/s00382-011-1128-8.

Yang, Z.F., Zeng, Z.Q., Wang, K., Wong, S.-.S., Liang, W.H., Zanin, M., Liu, P., 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. doi:10.21037/jtd.2020.02.64.

Yu, H.P., Huang, J.P., Li, W.J., Feng, G.L., 2014. Development of the analogue-dynamical method for error correction of numerical forecasts. J. Meteorol. Res. 28 (5), 934–947. doi:10.1007/s13351-014-4077-4.

Yu, L., Wang, S.Y., Lai, K.K., 2008. Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm. Energy Econ. 30 (5), 2623–2635. doi:10.1016/j.eneco.2008.05.003.

Zheng, Z.H., Huang, J.P., Feng, G.L., Chou, J.F., 2013. Forecast scheme and strategy for extended-range predictable components. Sci. China Earth Sci. 56 (5), 878–889. doi:10.1007/s11430-012-4513-1.