Nonlinear Dyn (2023) 111:15573–15589
https://doi.org/10.1007/s11071-023-08632-2

ORIGINAL PAPER

Data assimilation method for improving the global spatiotemporal predictions of epidemic dynamics yielded by an ensemble Kalman filter and Metropolis–Hastings sampling

Feng Liu · Xiaowei Nie · Adan Wu · Zebin Zhao · Chunfeng Ma · Lijin Ning · Yajie Zhu · Liangxu Wang · Xuejun Guo · Xin Li

Received: 17 January 2023 / Accepted: 27 May 2023 / Published online: 18 June 2023

Abstract Assimilating the latest epidemic data can improve the predictions of epidemic dynamics compared with those using only dynamic models. However, capturing the nonlinear spatiotemporal heterogeneity remains challenging. We propose a data assimilation method to simultaneously update the parameters and states with respect to their spatiotemporal variation intervals by (1) developing a susceptible-infected-removed-vaccinated model by considering vaccination strategy and quarantine periods and (2) assimilating real-time epidemic data using an ensemble Kalman filter for daily updates of the state variables and Metropolis–Hastings sampling for weekly parameter estimation. Synthetic experiments and a WebGIS-based global prediction system demonstrate the sufficient nowcasting accuracy of this method. An analysis of the system outcomes shows that modeling vaccination details, embedding reasonable model and observation errors, using up-to-date parameters, and avoiding the prediction of sporadic cases can increase the correlation coefficient and coefficient of determination by more than 31.35% and 161.19%, respectively, and decrease the root mean square error by more than 54.17%. Our prediction system has been working well for more than 700 days. Its worldwide nowcasting accuracies have been continuously improved, where the overall correlation coefficients, coefficient of determination, and threat percent score exceed 0.7, 0.5 and 65%, respectively. The proposed method lays promising groundwork for the real-time spatiotemporal prediction of infectious diseases.

Keywords Information fusion · COVID-19 · Parameter estimation · Forecasting and nowcasting · Epidemic model

F. Liu · A. Wu · Z. Zhao · C. Ma
Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou 730000, China

X. Nie (✉) · X. Guo · X. Li
State Key Laboratory of Tibetan Plateau Earth System, Environment and Resources (TPESER), Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing 100101, China
e-mail: xwnie@anso.org.cn

X. Nie
The Alliance of International Science Organizations, Beijing 100101, China

L. Ning
College of Earth and Environmental Sciences, Lanzhou University, Lanzhou 730000, China

Y. Zhu
College of Physics and Electric Engineering, Northwest Normal University, Lanzhou 730070, China

L. Wang
School of Environmental and Geographical Sciences, Shanghai Normal University, Shanghai 200234, China

1 Introduction

Studies concerning epidemic dynamics based on mathematical transmission models have been implemented for over 200 years and have laid the groundwork for epidemic prediction and control. The first contribution was provided by Daniel Bernoulli [1], and the canonical susceptible-infected-recovered (SIR) model was proposed by Kermack and McKendrick [2]. Subsequent studies used either complex networks [3, 4] or fractional differential equations [5–8] to extend the SIR model by incorporating more heterogeneities and correlations into the transmission processes [9]. However, due to the remarkable spatiotemporal variations exhibited by related factors [10], such as non-pharmaceutical interventions [11], population movements [12], and social contacts [13], as well as the uncertainties caused by virus mutations and vaccinations [14], mathematical models have long faced challenges in capturing epidemic dynamics in terms of multiple factors. Furthermore, small differences in the quantities, distributions and availability levels of the initial data and parameters have significant impacts on epidemic curves, implying that predictions of epidemic dynamics are highly sensitive to prior knowledge and have poor reliability. Recently, some studies switched to machine learning approaches [15] to find equivalent but faster methods for computing epidemic dynamics [16]. However, machine learning approaches are blind without background etiology knowledge [17]. Therefore, predicting epidemics with mathematical approaches still requires further research.

One way to improve the predictions of epidemic dynamics is to introduce data assimilation [18]. Data assimilation is an information fusion technique that incorporates observation data with numeric dynamic models [19, 20]. The best unbiased estimates of model trajectories are obtained by weighting the errors in both the models and observations [21], and the errors in the whole system are accordingly reduced [22]. In data assimilation, since imperfect models can be revised based on the available observation information, studies with limited dynamic process knowledge can benefit from more realistic scenarios. In the field of epidemiology, a general lack of COVID-19 knowledge was combined with the influences of socioeconomic factors such as population mobility [23] and medical cost burdens [24]. These dynamics vary for different socioeconomic statuses and may be heterogeneous in a population [25]. Therefore, the introduction of data assimilation has the potential to greatly improve epidemic modeling. Examples of early data assimilation studies in epidemiology include influenza forecasting by assimilating Google Flu Trends data into an SIR model [26, 27] and cholera forecasting in cooperation with the rainfall model [28]. During the COVID-19 pandemic, data assimilation techniques have also been applied to parameter estimation through the use of Kalman filters [29, 30] or calculus of variations [31] in epidemic models. As the first step required for generating real-time forecasts [26], epidemiological data assimilation is an option whose potential was suggested by the above studies.

However, previous epidemic prediction studies exhibit shortcomings, as data assimilation has generally been used to either estimate parameters or update states. These strategies lead to a limited understanding of epidemic dynamics. On the one hand, epidemic models, e.g., SIR-type models, are nonlinear and sensitive to the model parameters and the initial values of the state variables [32], implying that tiny variances in the initial state variables and parameters of the models can lead to considerable prediction errors. On the other hand, the spread of epidemics is highly variable from spatial and temporal perspectives [33], so during the lifecycle of the assimilation windows, updating the spatiotemporally variable parameters and model states at regular temporal intervals is essential for capturing epidemic dynamics. However, the temporal interval for updating parameters is much different from that for states. Generally, model parameters are constant and represent invariant control measures or the transmissibility levels of viruses over a long time, such as contact rates or vaccine effectiveness levels; however, model states are more random and may continue to change over short intervals. Therefore, determining a better assimilation strategy that addresses the problem of simultaneously updating the parameters and states with respect to their temporal variation intervals can improve the understanding of epidemic dynamics.

In this study, we propose a data assimilation method for simultaneously updating the parameters and states of a prediction model with respect to their spatiotemporal variations and further develop an epidemic data assimilation system. The method integrates weekly parameter estimation steps and daily

state updates to address their temporal variations during all stages of the disease spread lifecycle. The system is designed to improve the real-time and operational predictions concerning the COVID-19 pandemic. Considering that the ongoing COVID-19 pandemic has inevitably entered a more stable phase, this system provides epidemic dynamic nowcasting to inform the public and further provides long-term forecasts for control measure validation, thereby satisfying the primary goal of establishing a global COVID-19 prediction system and providing timely information services. The proposed data assimilation method and corresponding prediction system both exhibit real-time epidemic forecasting.

This paper is organized as follows. Section 2 introduces the framework of the study, including the major materials (e.g., the adopted epidemic data and model), the utilized methods (e.g., parameter estimation and data assimilation algorithms and error metrics), and the procedure used to integrate these materials and methods. Section 3 presents the main detailed results obtained in synthetic and real-world experiments based on the proposed method. The performance of the proposed epidemic data assimilation system is also estimated. Section 4 provides an overall discussion on the results and limitations of this study, as well as the improvements yielded. Conclusions are drawn in Sect. 5.

2 Materials and methods

The proposed data assimilation method is composed of a susceptible-infected-removed-vaccinated (SIRV) model, a Metropolis–Hastings (M-H) sampling approach for estimating the parameters of the SIRV model, and an ensemble Kalman filter (EnKF) for assimilating the daily confirmed cases into the SIRV model. The above components are integrated using a common data assimilation toolkit (ComDA) [34]. Figure 1 illustrates how these components are integrated.

2.1 Epidemic model

We extend the SIR model by introducing vaccinated and quarantined populations, i.e., an SIRV model [35], considering a two-dose vaccination strategy:

$$\begin{cases} S(t) = (1-\alpha) \times [N_a - I(t-1) - R(t-1) - V(t-1)] \\ \quad \dfrac{dI(t)}{dt} = \beta \times \dfrac{S(t)}{N} \times I(t-1) - \gamma \times I(t-1) \\ \quad \dfrac{dR(t)}{dt} = \gamma \times I(t-1) \\ V(t) = V_1(t-1) \times r_1 + V_2(t-1) \times r_2 \\ Q(t) = \alpha \times [N_a - I(t-1) - R(t-1) - V(t-1)] \end{cases}, \tag{1}$$

where $S(t)$, $I(t)$, $R(t)$, $V(t)$, $Q(t)$, N and N_a represent the susceptible, infected, removed, vaccinated, quarantined, total and active populations, respectively. The α, β, and γ parameters denote the protection proportion and the rates of contact and removal, respectively. $V_1(t)$ and r_1 are the number of people who are administered one dose and the corresponding vaccine effectiveness level, respectively, and $V_2(t)$ and r_2 are their counterparts with two vaccine doses.

Note that in (1), the active population N_a is defined as the population that has the ability and is willing to contact others during all stages of the pandemic lifecycle. Therefore, individuals who stay at home voluntarily are not included in the active population and are not considered in the epidemic dynamics. Additionally, $S(t)$ and $Q(t)$ are not determined by differential equations since they do not steadily increase and instead vary depending on the current public health policy. In contrast to other groups, a person can be quarantined at one time point, finish their quarantine at the next time point, and then belong to the susceptible population. This implies that $S(t)$ and $Q(t)$ are random and complementary. To estimate them, we first use a strategy for estimating the active population on each day and determining the sizes of the quarantined and susceptible populations in terms of the time-dependent protection proportion α.

This model should be flexible so that it can accommodate interventions in different scenarios, as vaccination programs and non-pharmaceutical interventions (NPIs) vary during all stages of the pandemic lifecycle with location and time. Another advantage of the SIRV model is that, unlike canonical compartmental models, it uses an estimated active population N_a to replace the total population at each simulation time step. N_a represents the true number of populations facing the disease, which generates a random quarantined population at each timepoint. Therefore, it

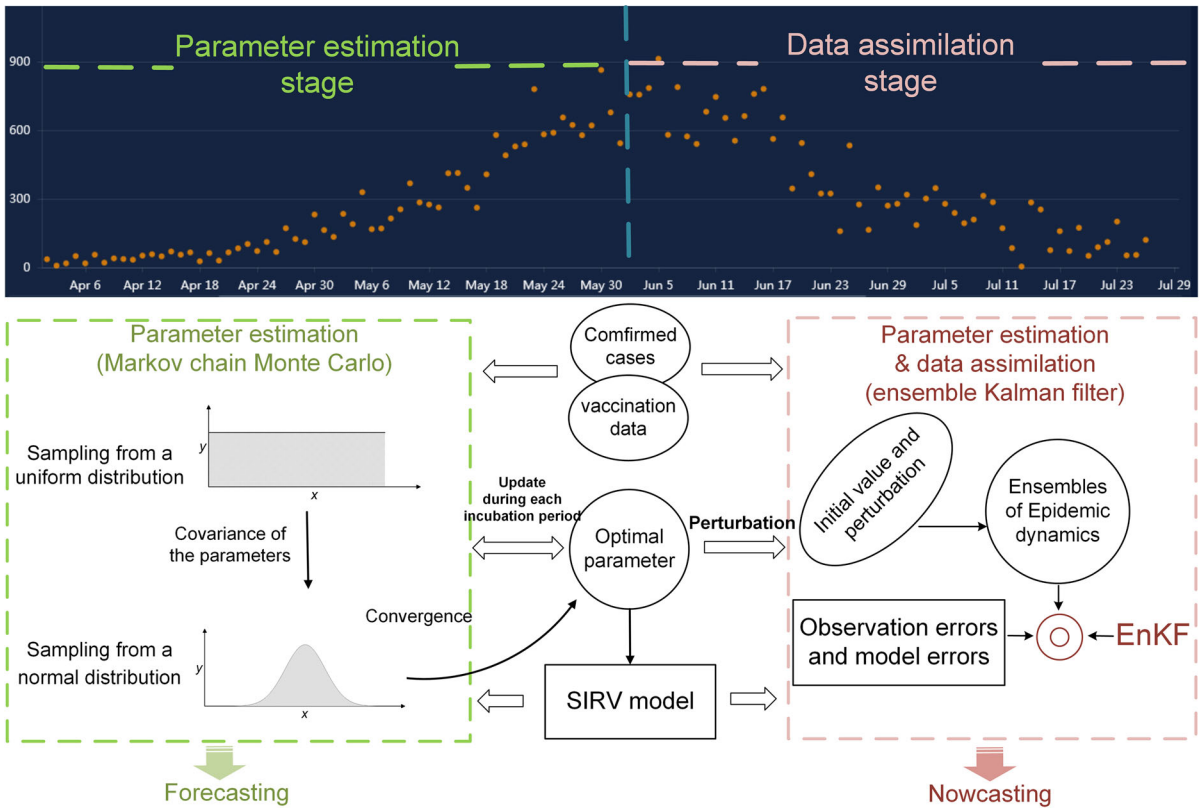

Fig. 1 The framework of the epidemic data assimilation method for simultaneously updating the model parameters and states. The epidemic curve is divided into a parameter estimation stage and a data assimilation stage, where the segmentation point occurs 7 days before the initial data assimilation time point

provides an alternative for modeling the quarantine period in the epidemic model, as some individuals may be quarantined while others are at the end of their quarantines. The total number of quarantined populations at each time is determined by N_a and the protection proportion.

The vaccine expiration date is not included in (1) because the relevant data remain in the research phase [36]. Instead, we investigate a similar issue, i.e., how vaccine effectiveness qualitatively affects the model outcomes, in the following experiments.

2.2 Parameter estimation

To adjust the model's parameters and reach the steady state for each special case, parameter estimation is performed before executing data assimilation. Here, we use a typical Markov chain-Monte Carlo (MCMC) algorithm, i.e., M-H sampling [37, 38]. By learning the prior knowledge contained in observations, M-H sampling discovers stationary estimates for the model parameters in the multidimensional probability space. In M-H sampling, the posterior distribution of the unknown parameters θ conditioned on the reference observations Ω is $P(\theta|\Omega) \propto P(\Omega|\theta)P(\theta)$, where $P(\theta)$ is the prior of the parameters and $P(\Omega|\theta)$ is the likelihood function.

As shown in Fig. 1, M-H sampling mainly contains two steps. In the first step, candidate samples are generated for the parameters according to the proposed uniform distribution, and the accepted samples are determined by comparing the model's outputs with Ω. In the second step, the same procedure is performed, but the proposed distribution follows $N(\theta, cov)$, where cov is a covariance matrix. The diagonal of the covariance matrix is set to the variances of the accepted samples in the first step, and the other elements are set to zero. The final estimated parameters are the modes of the accepted samples in the second step. The pseudocode of M-H sampling is provided in "Appendix A1".

2.3 EnKF

A nonlinear system of time-dependent states x according to a deterministic dynamic $\boldsymbol{M}$ and extra observations (real-world data) y is written as [39]

$$\begin{cases} x_{t+1} = \boldsymbol{M}(x_t) + \varepsilon \\ y = \boldsymbol{H}(x_{t+1}) + \epsilon \end{cases} \tag{2}$$

Here, $\boldsymbol{H}$ is the observation mapping operator. $\boldsymbol{H}$ is used to convert the state vector x to specific variables that can be directly compared with the available observations y. The subscript t denotes the calendar day or runtime. The independent variables $\varepsilon \sim N(0, Q)$ and $\epsilon \sim N(0, R)$ are the model error and observation error, respectively, where Q and R are the covariance matrices of the model error and observation error, respectively.

The EnKF [40, 41] is one of the most commonly used assimilation algorithms in geosciences, and it resolves the nonlinear estimation problem by employing the ensemble forecasting scheme. In the EnKF, the final estimate is

$$x^a = x_{t+1} + \boldsymbol{K}(y - \boldsymbol{H}(x_{t+1})), \tag{3}$$

where $\boldsymbol{K} = P^f \boldsymbol{H}^{\boldsymbol{T}} \left(\boldsymbol{H} P^f \boldsymbol{H}^{\boldsymbol{T}} + R\right)^{-1}$ is the Kalman gain, $P^f = \frac{1}{N-1}(x_{t+1} - \overline{x_{t+1}})(x_{t+1} - \overline{x_{t+1}})^T$ is the variance of the ensemble simulation of $\boldsymbol{M}(x_t)$, and $\overline{x_{t+1}}$ is the average value of the state vector. Figure 2 illustrates the concept of ensemble data assimilation.

In the epidemiology data assimilation workflow (Fig. 1), $\boldsymbol{M}$ and $\boldsymbol{H}$ are the epidemic model and identity matrix, respectively. The state x denotes the number of infected people. The pseudocode of the EnKF algorithm applied during the data assimilation stages is provided in Appendix A2.

2.4 Configurations

The model's parameters are calibrated by M-H sampling during the parameter estimation stage. M-H can find the optimal model parameters by repeatedly sampling from the multidimensional probability space and then adjusting the model structure according to the observation series over a long time interval. The sampling space comprises three parameters (the protection proportion α, contact rate β, and removal rate γ) in the SIRV model and one state variable (the active population N_a). Here, the sampling ranges for α, β, γ and N_a are $[0.01, 1]$, $[0.0001, 0.5]$, $[0.000001, 0.1]$ and $[1, N]$, respectively, where N denotes the total population. The sampling ranges are experiential since they must be widely applicable. The parameter estimation stage is further divided into several timespans over 7 days, and M-H sampling is performed in each of the timespans. The numbers of sampling iterations for the uniform and normal distributions are 3000 and 1500, respectively. The corresponding optimal parameters (i.e., modes) are used to reconstruct the past epidemic curves and to set up the SIRV model.

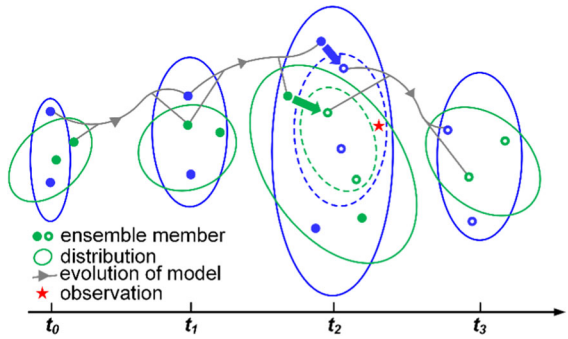

Fig. 2 Schematic of the data assimilation process using an EnKF. Two model states (colored in blue and green) consist of their distributions and samples (ensemble members) and evolve in accordance with the model's dynamics or rules at each time step. At t_2, after the ensemble members (in either the blue or green state) are updated by assimilating observation (red star), the uncertainty levels of the model states are reduced (the distributions turn from solid ellipses into dotted ellipses), and the ensemble members converge in distributions with smaller buffer zones. The averages of the updated ensemble members are regarded as the best unbiased estimations for the model states. Generally, the numbers of model states and ensemble members are much larger than 2. (Color figure online)

The model's trajectory is continuously adjusted during the data assimilation stage. Here, the EnKF is used to assimilate the daily confirmed cases at a specific discrete time. The optimal parameters obtained during the parameter estimation process of the last time span are input into the SIRV model. The initial field consists of the above parameters and the infected population (with 10% perturbation). Since we assume that the system state (i.e., the infected population) equates with the daily confirmed cases, which have been introduced as the observations to be assimilated, the observation operator is the corresponding identity matrix. The ratio of the observation error to the model error is 1:10, implying that the observations (daily confirmed cases) are assigned higher weights when generating the final

state variable values. The EnKF follows the standard workflow to update the states by calculating the ensemble averages of 100 ensemble members.

The forecasting results highlight three possible values for disease spread in the next 60 days and further provide convincing predictions when the factors that affect the epidemic change. The forecasting effect is enhanced by performing a scenario analysis after the parameter estimation stage; for example, it is assumed that typical NPIs may result in a variety of control measures [42, 43]: mitigating (using the protection proportion α in the SIRV model, which is the regular value of the intervention factor), mild (using 0.5α, indicating relaxed policies) and suppressive (using 1.5α, indicating severe interventions). We adjust the dynamics and then update the future epidemic curve according to the three intervention factors.

Nowcasting predicts the next 7 days of daily confirmed cases during the data assimilation stage. To generate real-time nowcasting results, the parameter estimation and data assimilation stages both drift over time, which means that the last time span also moves forward to update the optimal parameters. This strategy ensures that the SIRV model is up-to-date based on the latest epidemic curve.

2.5 Data

We collect the following open access datasets: (1) the infection data from the COVID-19 Data Repository [44]; (2) the vaccination data from the official website of Our World in Data [45]; (3) the population density data from the Population Estimation Service (https://doi.org/10.7927/H4DR2SK5); (4) the Map server from Tianditu (http://lbs.tianditu.gov.cn/server/MapService.html); and 5) the gross national income per capita data from the World Bank's list of economies published in June 2020. The first two datasets are automatically downloaded at 14:00 (UTC + 8) each day by retrieving files software (GNU Wget). We use a 7-day piecewise polynomial fitting method to obtain smoother infection data and to avoid the potential overfitting problem in the subsequent work.

2.6 Error metrics

To calculate the explainability, bias and correlation levels of the outcomes, the coefficient of determination (R^2), root mean square error (RMSE), and correlation coefficient (r) [46] are introduced as follows.

$$R^2 = 1 - \frac{\sum_{i=1}^{T}\left(x_t^a - \mathrm{y}_t^o\right)^2}{\sum_{i=1}^{T}\left(x_t^a - \overline{\mathrm{y}_t^o}\right)^2}, \tag{4}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(x_t^a - \mathrm{y}_t^o\right)^2}, \tag{5}$$

$$r = \frac{\sum_{t=1}^{T}\left(x_t^a - \overline{x_t^a}\right)\left(y_t^o - \overline{\mathrm{y}_t^o}\right)}{\sqrt{\sum_{t=1}^{T}\left(x_t^a - \overline{x_t^a}\right)^2}\sqrt{\sum_{t=1}^{T}\left(y_t^o - \overline{\mathrm{y}_t^o}\right)^2}}, \tag{6}$$

where T represents the full period of data assimilation and x_t^a and y_t^o represent the assimilation and observation time series at time t, respectively. $\overline{x_t^a}$ and $\overline{\mathrm{y}_t^o}$ represent the average assimilation and observation values, respectively.

Additionally, we use the threat percent score (TPS) to measure the proportion of successful predictions. The TPS is similar to the threat score [47] but is defined as the distance between an estimation and an observation rather than that between dichotomous (true or false) forecasts:

$$\mathrm{TPS} = \begin{cases} 1 - |x^a - y^o|/y^o, \text{if } \mathrm{y^o} \\ \quad \text{is located in the uncertainty range} \ ., \\ 0, \text{otherwise} \end{cases} \tag{7}$$

where x^a and y^o are the assimilation results and true confirmed cases, respectively. The uncertainty ranges consist of the ensemble averages and their $\pm$ 25% buffer zones.

The nowcasting accuracy is higher when the TPS, R^2 and r are closer to 1 or when the RMSE is closer to 0. Regarding the TPS, we store all the nowcasting data and the true number of confirmed cases over the last 7 days and then compare them by using (7) in chronological order. With respect to the other error metrics, we see the assimilation data and the true number of confirmed cases as the assimilation results and the observation time series, namely, x_t^a and y_t^o in (4) $\sim$ (6), respectively.

3 Results

3.1 Synthetic experiments

The data assimilation method was tested in advance with synthetic experiments. The first step was to generate observations using Model (1). The configurations were as follows: the parameters α, β, and γ were 0.05, 0.05, and 0.01, respectively. The total population size N and initial infected population size were 1 million and 1, respectively. It was assumed that 0.15% of the total population was vaccinated each day. A 500-day time loop was then executed based on Model (1), and 40% perturbations were added to the corresponding factitious epidemic curve to generate the observations. The number of ensembles was 100. The model error and observation error were time-dependent and were set to 10% of the ensemble averages and observations. In the second step, we randomly selected 100 30-day subperiods as the parameter estimation stages (Fig. 3) and performed 7-day nowcasting with only parameter estimation (as the control method) or with both parameter estimation and data assimilation (the proposed method, as shown in Fig. 1). Finally, we calculated error metrics (R^2, RMSE, and r) between the model outputs (or observations) and the nowcasting results for each subperiod, as well as their averages.

The error metric results demonstrate that our proposed method outperformed the control method. Although the correlations between the two nowcasting methods and model outputs (r=0.99 for each) are indistinguishable, the R^2 (0.60 for the control method and 0.98 for the proposed method) and RMSE values (226.32 for the control method and 38.83 for the proposed method) show that the proposed method captured the true epidemic curve with higher accuracy. The error metrics between the nowcasting results and observations demonstrates consistent performance; i.e., for the control method, r=1.00, R^2=0.62 and RMSE = 221.29, and for the proposed method, r=1.00, R^2=0.98 and RMSE = 30.57. This experiment illustrates that learning the observations with data assimilation during the nowcasting stage can improve the predictability of epidemic dynamics.

3.2 WebGIS-based global system implementation

The data assimilation method was applied in a COVID-19 data assimilation (COVDA) system to generate daily COVID-19 predictions for 193 countries or regions. The source data (see the 'Data' section) were automatically downloaded each day, and all system configurations were adaptive without any manual intervention. For example, the parameters were customized for a specific country or region during a specific period, and the observations and model errors were self-amplified or diminished according to the model states. The system was used to estimate the numbers of confirmed cases for the upcoming 7 days in the nowcasting output and to propose 3 possible scenarios for the next 60 days in the forecasting output. Both the nowcasting and forecasting results were visualized using a WebGIS-based interface on personal computers and mobile devices (Fig. 4). The interactive webpage allowed users to select countries or regions, epidemic periods, and other features (e.g., confirmed data, prediction data and error ranges) in the plots.

3.3 The 60-day forecasting module and its performance

The 60-day forecasting module is an independent dashboard in the system web interface (Fig. 4d). Here, we present examples concerning three countries to explain how the system can provide different scenarios regarding the spread of the disease. These instances are all based on epidemic phases on specific dates, i.e., March 23. For each example, three scenarios (the yellow, red, and blue curves) corresponding to the mitigating, mild and suppressive policies were considered with different disease spread dynamics from April 18 to June 16. For example, compared to the mild scenario, the obvious rise of the epidemic curve based on real-world data (documented cases) after March 23 in Fig. 5a indicates that milder NPIs may have been implemented. Similarly, the flat curve and falling curve obtained based on real-world data (Fig. 5b and c, respectively) may have been caused by the mitigation and suppression of NPIs, respectively. Therefore, comparisons between the curves produced based on real-world data and simulated scenarios allow the public to estimate whether and how to change the current NPIs to control the spread of the disease.

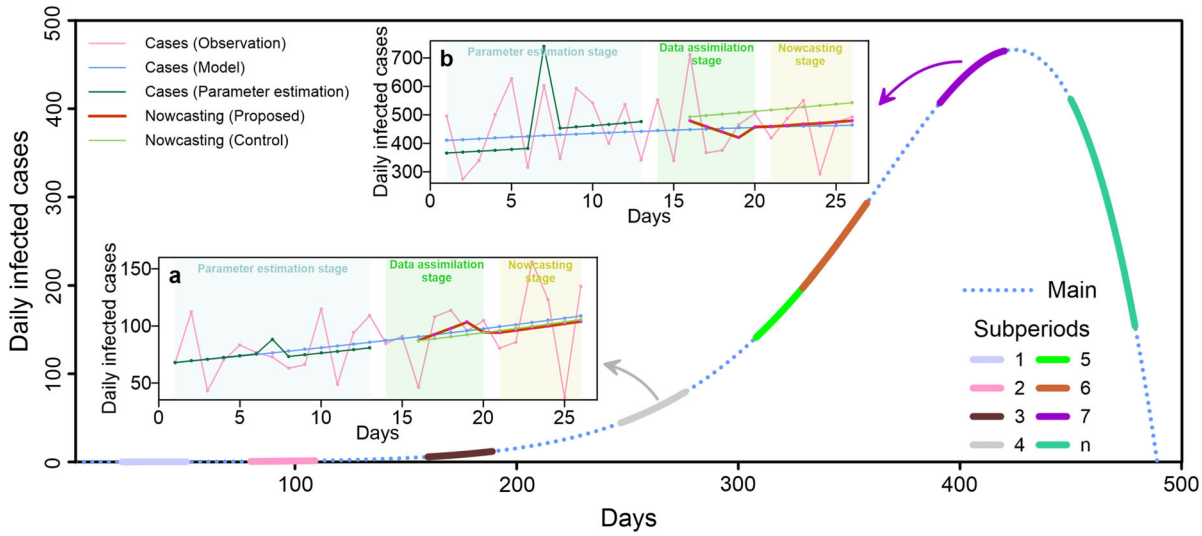

Fig. 3 Synthetic epidemic nowcasting experiments conducted based on the developed strategy with parameter estimation (Metropolis–Hastings sampling) and data assimilation (ensemble Kalman filter). The main epidemic curve (dotted line) was generated using an SIRV model, and 100 30-day subperiods (bold short lines) were randomly selected to perform the parameter estimation process, followed by 7-day nowcasting with the control and proposed strategies. Subplots a and b are synthetic experiments implemented during two random subperiods. Observation cases (pink fine line) included 40% perturbations of the model outputs (blue fine lines), and both the control (green fine line) and proposed strategies (red line) were based on the adjusted model with parameter estimation. (Color figure online)

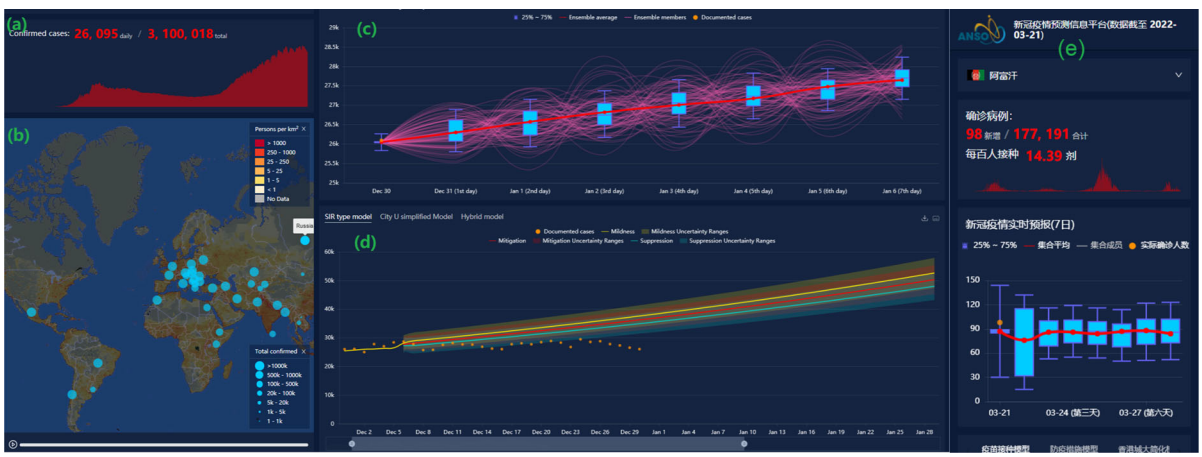

Fig. 4 The web interface of the real-time global COVID-19 data assimilation prediction system comprises 4 web dashboards for **a** selecting a country and displaying its epidemic curve (193 countries/regions in total), **b** determining the number of country-level cumulative confirmed cases using geographic information, **c** performing nowcasting using data assimilation and (d) forecasting using parameter estimation and scenario analysis. **e** is the corresponding mobile dashboard (in Chinese). All figures were captured from the web interface

3.4 Nowcasting performance

To evaluate the nowcasting performance of the proposed approach, the whole running period of the COVDA system was divided into three stages. The system was first released in early June 2020, and its formal version was launched on 31 August 2020. The system was upgraded on 1 January 2021 to include ongoing vaccination. Thus, Stage 1 was from June 2020 to December 2020, in which we employed a

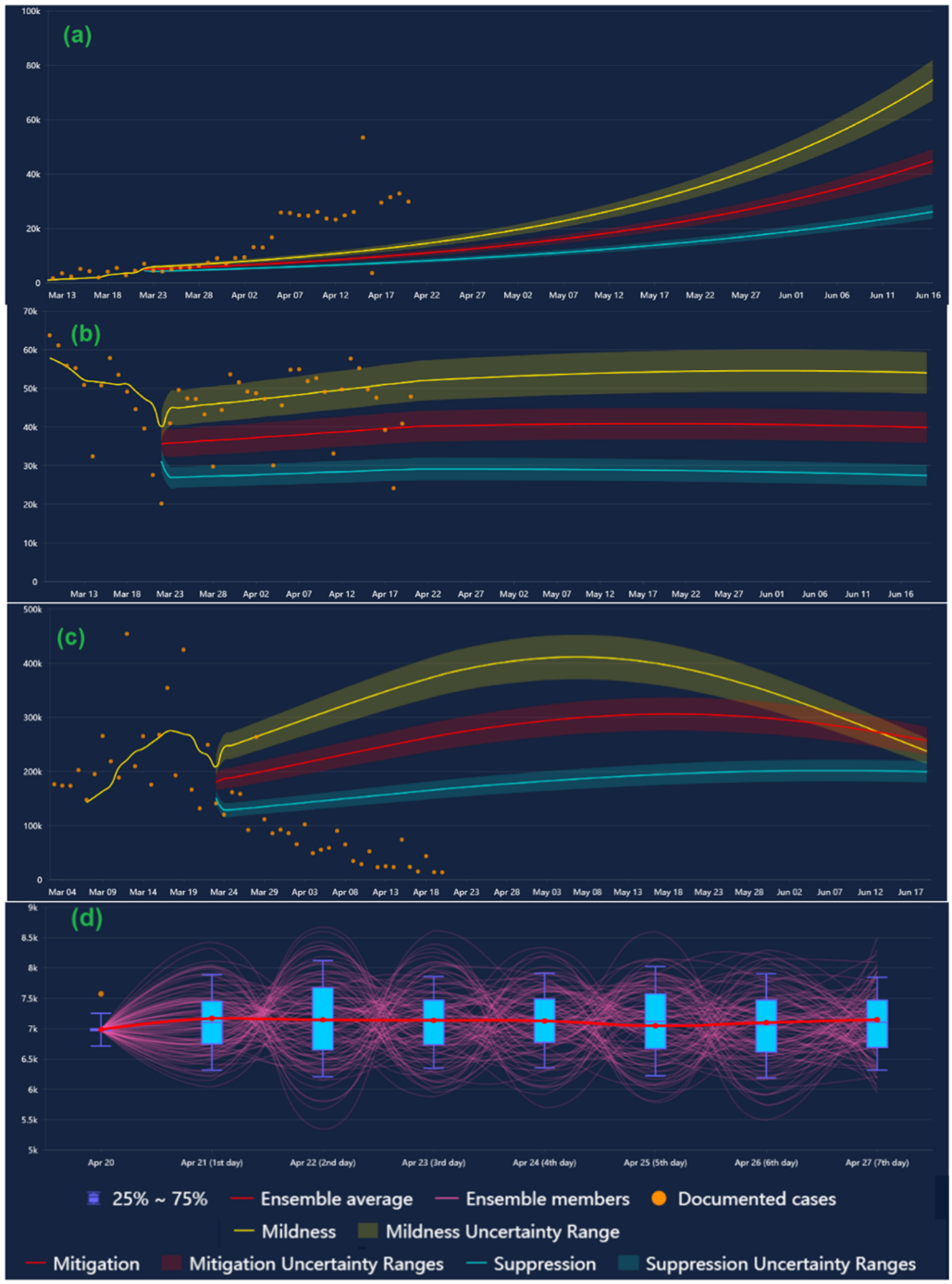

Fig. 5 Sixty-day forecasting results obtained in scenarios under **a** mild, **b** mitigating and **c** suppressive non-pharmaceutical interventions (NPIs). The excessive rise or fall of the epidemic curve based on real-world data indicates milder or more suppressive NPIs, respectively. **d** Presents the 7-day nowcasting results obtained with 100 ensemble members and the corresponding ensemble average. All figures were captured from the web interface

simple SIR model. Stage 2 was from January 2021 to December 2021, during which vaccinations were included in the system, and the vaccine effectiveness levels ranged from 85 to 95% (customized for SARS-CoV-2). Stage 3 was from January 2022 to March 2022, in which the vaccine effectiveness levels were updated for the variants (ranging from 30 to 60%) in accordance with the effectiveness of the vaccines against the SARS-CoV-2 variants [48, 49].

The nowcasting results are validated by the measures listed in the "Error metrics" section. Figure 6 shows the 7-day average error metrics for the system. The results display three patterns. First, the nowcasting accuracies increased stage by stage as the dynamic model and parameters were improved. For example, during Stage 1, the correlation coefficient (Fig. 6a, blue line) fell from 0.4747 to 0.1961. The system performed poorly due to (1) the absence of a vaccinated population in the SIR model, (2) the unreasonable ratio of the observation error to the model error (1:1), and (3) the system instability exhibited during the initial launch stage. During Stage 2, we developed the SIRV model to address the global rollout of the COVID-19 vaccine and updated the ratio of the observation error to the model error (1:10). Then, the corresponding values increased (ranging from 0.6235 to 0.3125). During Stage 3, we updated the vaccine effectiveness levels to match the global spread of the SARS-CoV-2 variants, and the correlation coefficients further increased (ranging from 0.6943 to 0.3401). Compared to that achieved in other stages, the prediction performance attained in Stage 1 was poor, indicating that the accuracy of pandemic dynamics modeling decreases when vaccination and reasonable observation and model errors are not considered. Second, the accuracies decreased daily during the nowcasting cycle (7 days). In particular, from the first day until the 7th day, the averaged r values were 0.5975, 0.3181, 0.3161, 0.3221, 0.3088, and 0.2832 (Fig. 6a, solid line); the averaged R^2 values were 0.3365, 0.0981, 0.0818, 0.0726, 0.0587, 0.0499, and 0.0305 (Fig. 6b, solid line); and the averaged RMSE and TPS values presented similar trends (Fig. 6c and d, respectively, solid lines). Compared with the first day, the performances of the error metrics worsened and were basically consistent during the 2nd and 7th days, implying that the long-term prediction results were inferior to the short-term prediction results. Regarding the RMSEs, although the performances were slightly better on the 6th and 7th days for some of the subplots, we believe that the insufficiency of the samples for producing a large infected population contributed to this phenomenon. Third, we classified the countries and regions as having outbreaks if their average numbers of daily confirmed cases were larger than 1000. The corresponding curves (Fig. 6, dotted line) are easily identifiable by their larger r (37.35% more than global average, similarly hereinafter), R^2 (49.83%), and TPS (6.94%) values and their smaller RMSEs (93.40% less than the global average). These results indicate that the predictions were more accurate for countries/regions with outbreaks. This phenomenon was more significant in the RMSEs, where the global RMSEs continued to rise after the 2nd day.

A further interpretation of Fig. 6 is as follows. The results confirm that one-day prediction performed better than long-term prediction. During the first and second nowcasting days, the correlation coefficients were reduced by almost half, the coefficient of determination decreased by more than two-thirds, and a large RMSE increase was observed. Distinguished from the other metrics, the TPS fell gradually (by only approximately 5%) from the first to the second nowcasting day. This phenomenon was mainly caused by the buffer zones (see the 'Error metrics' section) used for comparison with the ground truths; i.e., the smaller the buffer zones were, the more obvious the decrease became. For example, if $\pm$ 2.5% buffer zones were used, the TPS would decrease by approximately 15%. Additionally, the system performances achieved for the countries and regions with outbreaks surpassed the worldwide performances in terms of every error metric, meaning that the predictions in these countries yielded greater accuracy. These results further support the conclusion that, in the real world, the SIRV model is more appropriate for a population with a sufficiently high infection rate, exhibiting better performance than that achieved in applications with sporadic cases.

We further evaluated the one-day nowcasting error metrics (Fig. 7). The lifecycle of the data assimilation system was also divided into the same three stages. Clearly, the system performance improved after the model operation and error evaluation were improved (Stage 2) or the parameters were closer to reality (Stage 3). During Stage 1, the nowcasting accuracies experienced a lasting upgrade. This is because the assimilation method continued to be revised, including

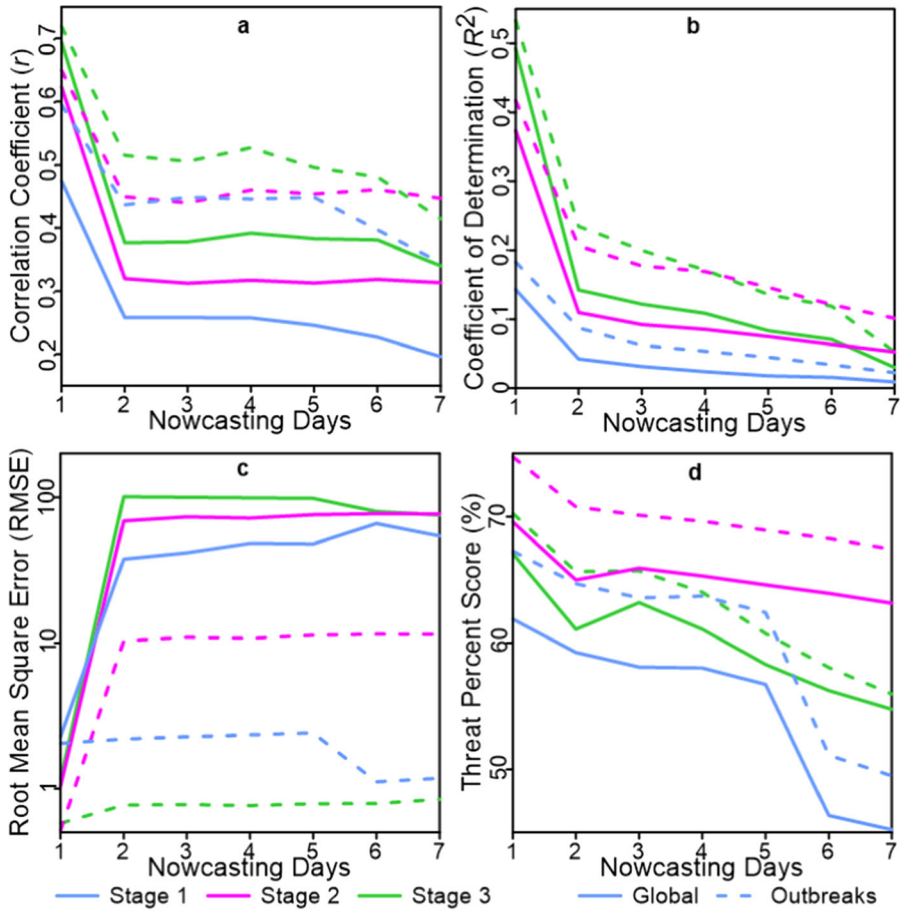

Fig. 6 The 7-day average nowcasting error metrics induced by the epidemic data assimilation system during Stage 1 (from June 2020 to December 2020), Stage 2 (from January 2021 to December 2021), and Stage 3 (from January 2022 to March 2022). Countries/regions with outbreaks (dotted line) were defined as having daily average numbers of confirmed cases that were larger than 1000. Note that the Y-axis in the RMSE plot is logarithmic

determining the appropriate sampling ranges, timespans, and loop counts during parameter estimation or the number of ensemble members and the correct observation preprocessing strategy (data smoothing) during assimilation. During Stage 2, due to the introduction of the vaccination module, the system operation was stable (compared with those of Stage 1, r and R^2 increased by 31.35% and 161.19%, respectively, and the RMSE decreased by 54.17%) but faced the issue of virus variants that affected the epidemic dynamics; therefore, the nowcasting accuracies slightly declined at the end of Stage 2. During Stage 3, we used an updated vaccine effectiveness mechanism that was customized for the SARS-CoV-2 variants, which helped improve the performance of the system (compared with those of Stage 2, r and R^2 increased by 11.36% and 31.87%, respectively, and the RMSE decreased by 10.94%). The nowcasting accuracy curves produced for the countries and regions with outbreaks were higher than the world average (r and R^2 increased by 13.27% and 16.50% on average, respectively, and the RMSE decreased by 50.87% on average), implying that sporadic cases may result in poor performance for the SIRV model.

4 Discussion

Since it was released in early June 2020, COVDA has been working well for more than 700 days. The nowcasting accuracies of COVDA have been continuously improved for 193 countries and regions (the overall correlation coefficients, coefficients of determination, and TPSs exceed 0.7, 0.5, and 65%, respectively).

Compared with the existing methods, the main advantage of our approach is that it finds a solution to the nonlinear spatiotemporal heterogeneity estimation problem concerning epidemic dynamics. By introducing the EnKF and M-H sampling, the parameters and states of the epidemic model are simultaneously

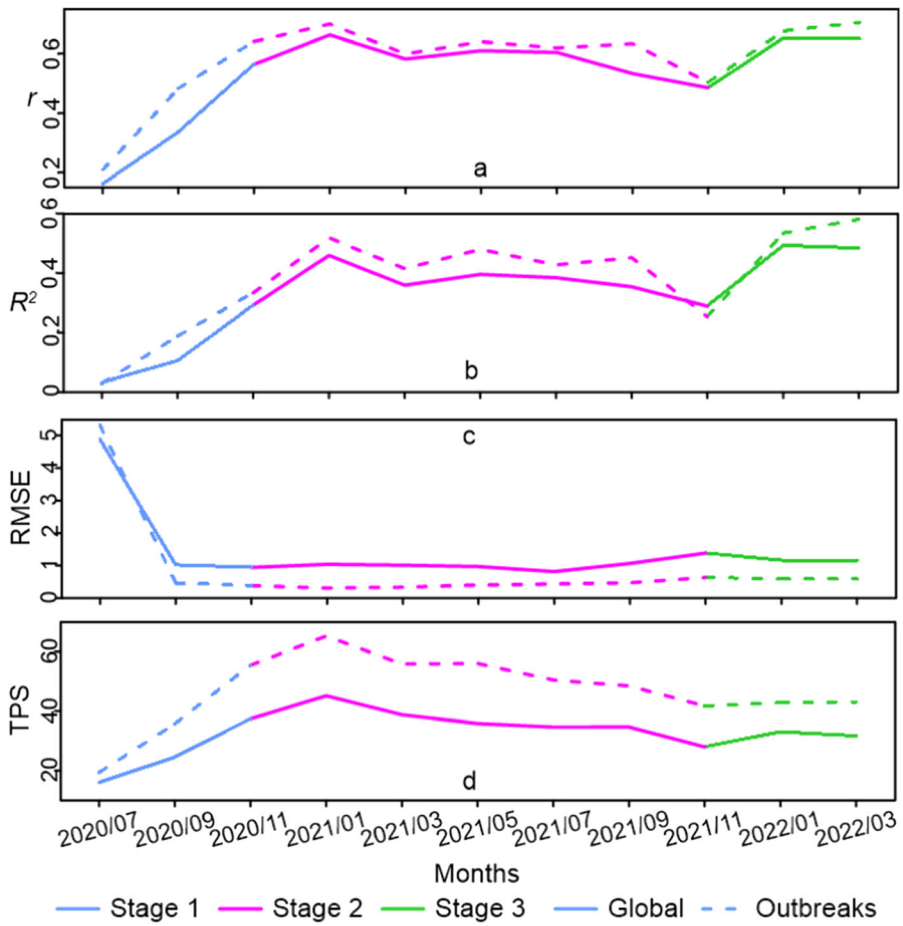

Fig. 7 The one-day nowcasting error metrics, i.e., the correlation coefficients (r), coefficients of determination (R^2), RMSEs, and TPSs, produced during Stage 1 (from June 2020 to December 2020), Stage 2 (from January 2021 to December 2021), and Stage 3 (from January 2022 to March 2022). The countries and regions with outbreaks (dotted line) were defined as having daily average numbers of confirmed cases that were larger than 1000

updated according to their spatiotemporal variations. The results of synthetic experiments show that these solutions can improve the accuracy of epidemic dynamics prediction. Another advantage is that we developed a real-time spatiotemporal data assimilation system, where the proposed method had been put into practice. This system has accumulated a large amount of prediction data, permitting us to complete comprehensive evaluations and post-analyses of the effectiveness of the proposed method.

Our study demonstrated that the development of a general epidemic assimilation method and a corresponding prediction system for real-time and operational use is promising. Utilizing the EnKF to directly assimilate the number of confirmed cases into the SIRV model can result in an epidemic curve similar to that of the real world. Although the SIRV model is too simple to capture epidemic tendencies, an evaluation of the assimilation results showed that the total TPS produced for one day was more than 85% and that obtained for 7 days was approximately 77.6%. Note that this evaluation was global, and the parameters for each country were self-adaptive without any specific settings. However, the temporal variations in the error metrics of the assimilation results indicate decreases in predictability with increasing time scale. The error metrics in Fig. 6 also illustrate that the system can be further improved, especially for forecasting times exceeding 2 days.

The limitations of our study mainly originate from this simplified data assimilation framework. The model and observation errors require further study in both epidemic model comparisons and data validations, and they should be quantified and input into the assimilation system. In our study, the configurations of these errors were empirical such that the assimilation results were closer to the observations. This empirical configuration was also applied to the number of ensembles, which was set to 100. For the assimilation algorithm, the effectiveness of the EnKF in the epidemic prediction system was demonstrated. However, other algorithms, such as

canonical particle filters, calculus of variations, or emergent algorithms customized for epidemiology, should be further tested to better capture the nonlinearity, non-Gaussianity, heterogeneity, and randomness of the spread of epidemics. Regarding the observations, identifying the true infected population is challenging due to the virus incubation period and the timespan between infection and confirmed, which follows a Weibull distribution. Therefore, we directly assimilated the number of confirmed cases into the system, implying that the number of daily confirmed cases equaled the number of daily infections; however, these two variables are not the same, so they may have negative impacts on the prediction results.

Considering the proposed data assimilation framework, we recommend the following potential improvements in future work.

(1) Complex but more real epidemic models, such as contact networks [50], are also suitable for this assimilation framework. Compared with SIR-type models, a contact network model performs better in terms of capturing the heterogeneity and randomness in transmission processes. However, the use of more real models may require a high-performance computing architecture.
(2) Regarding the investigation of compound epidemic patterns for combining pathophysiological transmission processes and data-driven models, data assimilation can provide an alternative strategy based on information fusion, such as Bayesian model averaging.
(3) The development of observation operators can enable the good use of big data related to epidemiology. As a primary component of data assimilation, observation operators can characterize the relationships between states and observations. For example, if the infected population is regarded as a model state, then introducing conversions from the available observations (i.e., observation operators) to the infected population can provide comprehensive state calculations and improve the predictability of epidemic assimilation. By means of observation operators, the pathogenesis of the factors, which are so indefinite or complex that they cannot be included in epidemic models, can be incorporated into the assimilation system to update the infection population and influence the epidemic curve.

5 Conclusions

To tackle the nonlinear spatiotemporal heterogeneity problem of epidemic dynamics estimation, a new method is proposed and can be used to simultaneously estimate the parameters and update states with respect to their temporal variation intervals. This method uses information fusion techniques (including the EnKF and Metropolis–Hastings sampling) to harmonize the available health-related data and epidemic model and provides better predictions. A synthetic test showed that the proposed method benefits the capture of the epidemic curve and yields improved prediction accuracy. This method has been further applied in COVDA, which is a real-time and operational epidemic data assimilation system. COVDA has been operationally used for more than 700 days and has produced a large amount of prediction data. This permits us to conduct a comprehensive post-analysis. Some improvements of COVDA, such as vaccination modeling, the use of reasonable model and observation errors, and avoidance of sporadic case prediction, can increase the correlation coefficient and coefficient of determination by more than 31.35% and 161.19%, respectively, and decrease the RMSE by more than 54.17%. Future studies on epidemic data assimilation should focus on integrating more real and data-driven epidemic model operators with information fusion techniques and developing observation operators to make good use of big data related to epidemiology.

Author contributions Feng Liu, Xiaowei Nie, and Xin Li contributed to the study conception and design. Feng Liu, Xiaowei Nie, Adan Wu, Liangxu Wang, and Xin Li contributed to the methodology and visualization. Feng Liu, Adan Wu, Zebin Zhao, Chunfeng Ma, Lijin Ning, Yajie Zhu, and Xuejun Guo implemented the framework and system. Feng Liu, Zebin Zhao, Lijin Ning, and Yajie Zhu collated and analyzed the data. Feng Liu, Adan Wu, Zebin Zhao, and Xin Li contributed to the initial draft of the manuscript. Xiaowei Nie, Xuejun Guo, and Xin Li acquired the funding. All authors revised the manuscript and approved the final version.

Funding This work was supported in part by Alliance of International Science Organizations under (Grant numbers ANSO-SBA-2021-05 and ANSO-SBA-2020-13), and National Natural Science Foundation of China (Grant numbers 42271442).

Data availability Data that support the findings of this study included pandemic diagnosis data, vaccination data and

prevention and control measures. Verified COVID-19 data were obtained from Johns Hopkins University (https://github.com/CSSEGISandData/COVID-19). Vaccine data were obtained from Our World in Data (https://github.com/owid/covid-19-data). The time series data of global pandemic prevention and control measures (detailed rules) were obtained from the Oxford COVID-19 government response tracking system (https://github.com/OxCGRT/covid-policy-tracker). The components of the source code and instructions for implementation pertaining to this research can be accessed via the GitHub repository (https://github.com/uniliufeng/ComDA).

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Appendix A1: The pseudocode of parameter estimation

Algorithm 1. Parameter estimation (Metropolis–Hastings sampling)

Inputs: $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{N_a}$ denote the ranges of the protection proportion and the rates of contact and removal in the epidemic model (S1), respectively. $\boldsymbol{\Omega}$ denotes the real-world data for confirmed cases. $\boldsymbol{m_1}$ and $\boldsymbol{m_2}$ denote the maximum loop counts of the sampling steps. $\boldsymbol{r_1}$ and $\boldsymbol{r_2}$ denote the desired accepted times of the sampling steps.

Formulate the sampling space $\boldsymbol{\Theta} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{N_a}\}$.

k=0.

While ($\mathbf{k} \leq \boldsymbol{m_1}$ && $\mathbf{accepted\ time} < \boldsymbol{r_1}$)

Choose a parameter vector $\boldsymbol{\theta}^{(k)}$ according to uniform distributions.

Generate a result from the epidemic model using $\boldsymbol{\theta}^{(k)}$.

Generate the acceptance rate from a uniform distribution $\boldsymbol{U}(\mathbf{0}, \mathbf{1})$.

Compare the result with $\boldsymbol{\Omega}$ and determine whether the sample can be adopted according to the acceptance rate. If yes, accepted time++.

k++.

End while

Formulate a normal proposal distribution $\boldsymbol{N}(\boldsymbol{\theta}, \boldsymbol{cov})$, where $\boldsymbol{cov}$ is the covariance matrix of the samples adopted in the previous loop.

k=0.

While ($\mathbf{k} \leq \boldsymbol{m_2}$ && $\mathbf{accepted\ time} < \boldsymbol{r_2}$)

Choose an initial parameter vector $\boldsymbol{\theta}^{(k)}$ according to $\boldsymbol{N}(\boldsymbol{\theta}, \boldsymbol{cov})$.

Generate a result from the epidemic model using $\boldsymbol{\theta}^{(k)}$.

Generate the acceptance rate from a uniform distribution $\boldsymbol{U}(\mathbf{0}, \mathbf{1})$.

Compare the result with $\boldsymbol{\Omega}$ and determine whether the sample can be adopted according to the acceptance rate. If yes, accepted time++.

k++.

End while

Outputs: the posterior probability distribution of the parameters based on the samples in the last loop.

Appendix A2: The pseudocode of the EnKF algorithm

Algorithm 2. EnKF during the data assimilation stage

Inputs: the assimilation period; n denotes the number of ensemble members; p denotes the value of the ensemble perturbation; R and Q denote the values of the observation and model errors, respectively; $\boldsymbol{\Omega}$ denotes the real-world data for confirmed cases; the dates with the updated parameters.

While (date i is in the assimilation period)
If (date i is the date of the updated parameters)
  Update the parameters
End if
Use p to perturb $\boldsymbol{\Omega}$ on date i
Input the perturbed data into the epidemic model
Run the model for one step and generate the ensemble results
If (date i is the date of assimilation)
  Derive R, Q, and the Kalman gain from the ensemble results
  Obtain the new ensemble results with the Kalman gain
End if
Calculate the ensemble average as the nowcasting data
End while

Outputs: the nowcasting time series data.

References

1. Dietz, K., Heesterbeek, J.A.P.: Daniel Bernoulli's epidemiological model revisited. Math. Biosci. **180**, 1–21 (2002)
2. Kermack, W.O., McKendrick, A.G.: A contribution to the mathematical theory of epidemics. Proc. Royal Soc. London Ser. A Contain. Papers Math. Phys. Character **115**, 700–721 (1927)
3. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature **393**, 440–442 (1998)
4. Barabási, A.-L., Albert, R.: Emergence of scaling in random networks. Science **286**, 509–512 (1999)
5. Kumar, A., Prakash, A., Baskonus, H.M.: The epidemic COVID-19 model via Caputo–Fabrizio fractional operator. Waves Random Complex Media (2022). https://doi.org/10.1080/17455030.2022.2075954
6. Kumar, A., Prasad, R. S.: On dynamical behavior for approximate solutions sustained by nonlinear fractional damped Burger and Sharma–Tasso–Olver equation. Int. J. Modern Phys. B 2350228.
7. Gao, W., Baskonus, H.M.: Deeper investigation of modified epidemiological computer virus model containing the Caputo operator. Chaos, Solit. Fract. **158**, 112050 (2022)
8. Ciancio, A., Yel, G., Kumar, A., Baskonus, H.M., Ilhan, E.: On the complex mixed dark-bright wave distributions to some conformable nonlinear integrable models. Fractals **30**, 2240018 (2022)
9. Newman, M.E.: Spread of epidemic disease on networks. Phys. Rev. E **66**, 016128 (2002)
10. Snoeijer, B.T., Burger, M., Sun, S., Dobson, R.J.B., Folarin, A.A.: Measuring the effect of Non-Pharmaceutical Interventions (NPIs) on mobility during the COVID-19 pandemic using global mobility data. Npj Digital Med. **4**, 81 (2021)
11. Ge, Y., et al.: Untangling the changing impact of non-pharmaceutical interventions and vaccination on European COVID-19 trajectories. Nat. Commun. **13**, 3106 (2022)
12. Chen, X., Zhang, A., Wang, H., Gallaher, A., Zhu, X.: Compliance and containment in social distancing: mathematical modeling of COVID-19 across townships. Int. J. Geogr. Inf. Sci. **35**, 446–465 (2021)
13. Hoang, T., et al.: A systematic review of social contact surveys to inform transmission models of close-contact infections. Epidemiology **30**, e74 (2019)
14. Huang, B., et al.: Integrated vaccination and physical distancing interventions to prevent future COVID-19 waves in Chinese cities. Nat. Human Behav. **5**, 695–705 (2021)

15. Han, B.A., Schmidt, J.P., Bowden, S.E., Drake, J.M.: Rodent reservoirs of future zoonotic diseases. PNAS **112**, 7039–7044 (2015)
16. Wiemken, T.L., Kelley, R.R.: Machine learning in epidemiology and health outcomes research. Annu. Rev. Public Health **41**, 21–36 (2020)
17. Robins, J.M.: Data, design, and background knowledge in etiologic inference. Epidemiology **12**, 313–320 (2001)
18. Li, X., Zhao, Z., Liu, F.: Big data assimilation to improve the predictability of COVID-19. Geography Sustain. **1**, 317–320 (2020)
19. Li, X., Liu, F., Fang, M.: Harmonizing models and observations: Data assimilation in Earth system science. Sci. China Earth Sci. **63**, 1059–1068 (2020)
20. Bocquet, M., Brajard, J., Carrassi, A., Bertino, L.: Data assimilation as a learning tool to infer ordinary differential equation representations of dynamical models. Nonlin. Processes Geophys. **26**, 143–162 (2019)
21. Moore, A.M., et al.: Synthesis of ocean observations using data assimilation for operational, real-time and reanalysis systems: a more complete picture of the state of the Ocean. Front. Mar. Sci. **6**, 90 (2019)
22. Talagrand, O.: Assimilation of observations, an introduction (gtspecial issueltdata assimilation in meteology and oceanography: Theory and practice). J. Meteorol. Soc. Jpn. Ser II **75**(1B), 191–209 (1997)
23. Tian, H., et al.: An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. Science **368**, 638 (2020)
24. Kyu, H.H., et al.: Global, regional, and national burden of tuberculosis, 1990–2016: results from the global burden of diseases, injuries, and risk factors 2016 study. Lancet. Infect. Dis **18**, 1329–1349 (2018)
25. Oliver, N., et al.: Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle. Sci. Adv. **6**, eabc0764 (2020)
26. Shaman, J., Karspeck, A.: Forecasting seasonal outbreaks of influenza. PNAS **109**, 20425–20430 (2012)
27. Shaman, J., Karspeck, A., Yang, W., Tamerius, J., Lipsitch, M.: Real-time influenza forecasts during the 2012–2013 season. Nat. Commun. **4**, 2837 (2013)
28. Pasetto, D., Finger, F., Rinaldo, A., Bertuzzo, E.: Real-time projections of cholera outbreaks through data assimilation and rainfall forecasting. Adv. Water Resour. **108**, 345–356 (2017)
29. Li, R., et al.: Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). Science **368**, 489 (2020)
30. Evensen, G. et al.: An international assessment of the COVID-19 pandemic using ensemble data assimilation. medRxiv, 2020.2006.2011.20128777 (2020)
31. Nadler, P., Wang, S., Arcucci, R., Yang, X., Guo, Y.: An epidemiological modelling approach for COVID-19 via data assimilation. Eur. J. Epidemiol. **35**, 749–761 (2020)
32. Ma, C., et al.: Understanding dynamics of pandemic models to support predictions of COVID-19 transmission: parameter sensitivity analysis of SIR-Type models. IEEE J. Biomed. Health Inform. **26**, 2458–2468 (2022)
33. Castro, M.C., et al.: Spatiotemporal pattern of COVID-19 spread in Brazil. Science **372**, 821–826 (2021)
34. Liu, F., Wang, L., Li, X., Huang, C.: ComDA: a common software for nonlinear and non-gaussian land data assimilation. Environ. Model. Softw. **127**, 104638 (2020)
35. Liu, F., et al.: Return to normal pre-COVID-19 life is delayed by inequitable vaccine allocation and SARS-CoV-2 variants. Epidemiol. Infect. **150**, e46 (2022)
36. Baraniuk, C.: How long does covid-19 immunity last? BMJ **373**, n1605 (2021)
37. Hastings, W.K.: Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**, 97–109 (1970)
38. Zhu, G.F., et al.: 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 (2014)
39. van Leeuwen, P.J., Künsch, H.R., Nerger, L., Potthast, R., Reich, S.: Particle filters for high-dimensional geoscience applications: a review. Q. J. R. Meteorol. Soc. **145**, 2335–2365 (2019)
40. Evensen, G.: The ensemble Kalman filter for combined state and parameter estimation. Control Syst. IEEE **29**, 83–104 (2009)
41. Loos, S., et al.: Ensemble data assimilation methods for improving river water quality forecasting accuracy. Water Res. **171**, 115343 (2020)
42. Zhao, Z., et al.: Prediction of the COVID-19 spread in African countries and implications for prevention and controls: a case study in South Africa, Egypt, Algeria, Nigeria, Senegal and Kenya. Sci. Total Environ. **729**, 138959 (2020)
43. Zhao, Z., et al.: Stringent nonpharmaceutical interventions are crucial for curbing COVID-19 transmission in the course of vaccination: a case study of south and southeast Asian countries. Healthcare **9**, 1292 (2021)
44. Dong, E., Du, H., Gardner, L.: An interactive web-based dashboard to track COVID-19 in real time. Lancet. Infect. Dis **20**, 533–534 (2020)
45. Mathieu, E., et al.: A global database of COVID-19 vaccinations. Nat. Human Behav. **27**, 205 (2021)
46. Chicco, D., Warrens, M.J., Jurman, G.: The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comput. Sci. **7**, e623 (2021)
47. Hall, T., Brooks, H.E., Doswell, C.A.: Precipitation forecasting using a neural network. Weather Forecast. **14**, 338–345 (1999)
48. Bernal, J.L., et al.: Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant. New Engl. J. Med. **385**(7), 585–594 (2021). https://doi.org/10.1056/NEJMoa2108891
49. Fowlkes, A., et al.: Effectiveness of COVID-19 vaccines in preventing SARS-CoV-2 infection among frontline workers before and during B.1.617.2 (Delta) variant predominance -

Eight U.S. Locations, December 2020-August 2021. MMWR Morb. Mortal. Wkly. Rep. **70**, 1167–1169 (2021)
50. Liu, F., Li, X., Zhu, G.: Using the contact network model and Metropolis-Hastings sampling to reconstruct the COVID-19 spread on the "Diamond Princess." Sci. Bullet. **65**, 1297–1305 (2020)