One Health 21 (2025) 101266

Contents lists available at ScienceDirect

One Health

journal homepage: www.elsevier.com/locate/onehlt

Combining data assimilation of states and parameters for more precise infectious disease prediction☆

Zihan Hao [a], Shujuan Hu [a,b,*], Jianping Huang [a,b,*], Jiaxuan Hu [a], Zhen Zhang [a], Jingjing Zhang [a], Wei Yan [a,b], Han Li [a,b]

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

ARTICLE INFO

Keywords:
Infectious disease dynamic model
Prediction error
Data assimilation
Ensemble Kalman filter

ABSTRACT

Global escalation of infectious disease outbreak risks necessitates advanced predictive models. Despite methodological advances, errors in initial states and parameters of epidemiological dynamic models remain a key limitation to prediction reliability. To address this limitation, we propose an optimized data assimilation framework for combined state-parameter optimization based on Ensemble Kalman Filter. We design space transformations and adaptive covariance inflation driven by epidemic development and prediction errors, achieving a more stable update process and rapid response to epidemic changes. Through synthetic experiments and real-world case studies, the proposed scheme significantly reduces initial state and parameter errors, leading to a substantial improvement in prediction accuracy during the early stages of an epidemic. Compared with predictions without data assimilation, the average prediction error rate decreased by more than 50 % for 1-day predictions and by approximately 15 % for 7-day predictions. The prediction accuracy rate for the peak day of the epidemic and the peak number of infected cases reached more than 70 % in advance by 3 days. Critically, simple dynamical model integrated with our data assimilation framework outperform complex models without data assimilation. This study establishes data assimilation as an essential tool for epidemic forecasting and provides an extensible framework adaptable to multiple infectious diseases, offering critical support for public health decision making.

1. Introduction

Globalization and climate change have made infectious diseases emerge as a considerable public health threat. For instance, the novel coronavirus disease 2019 (COVID-19) has spread globally rapidly in recent years, resulting in more than 700 million infections and more than 7 million deaths. This number is continuing to increase according to the World Health Organization [1]. Additionally, climate change has created a more favorable environment for mosquitoes, increasing the risk of dengue fever. In 2024, Brazil experienced the worst dengue outbreak in its history. Devastating effects of infectious diseases on human health and socio-economic systems underscores the urgent need for timely and accurate forecasting [2,3]. All countries should establish a reliable and accurate infectious disease forecasting system for a better response to future public health crises.

At present, dynamic modeling and statistical modeling are the primary methods used for predicting infectious diseases. The compartmental model is the predominant framework for the dynamic modeling [4], and it has become the most commonly used tool for infectious diseases prediction and simulation [5–7]. It often divides the population into distinct states and describes their transition procedures by a comprehensive understanding of transmission mechanisms. Additionally, the rise of machine learning and deep learning has led to the widespread use of data-driven statistical models for predicting infectious diseases [8,9]. However, despite the availability of many infectious disease predicting models, accurately predicting the future trajectory of infectious diseases remains a significant challenge [10,11].

As we know, the description of transfer processes in compartmental models do not precisely align with real-world situations. This misalignment leads to incomplete dynamic modeling of epidemic spread, leaving certain physical processes unrepresented. Therefore, developing more complex dynamic models offers a natural approach to

☆ This article is part of a Special issue entitled: 'One Health framework for Inf Disease Modelling' published in One Health.

* Corresponding authors at: College of Atmospheric Sciences, Lanzhou University, Lanzhou 730000, China.
E-mail addresses: hushuju@lzu.edu.cn (S. Hu), hjp@lzu.edu.cn (J. Huang).

https://doi.org/10.1016/j.onehlt.2025.101266
Received 27 March 2025; Received in revised form 11 October 2025; Accepted 31 October 2025
Available online 4 November 2025

reducing errors in infectious disease forecasting. By incorporating a wider range of states and transmission processes, complex dynamic models can more accurately reflect real-world conditions, as far as expecting to improve the accuracy of predictions [2,12,13]. However, complex dynamic models introduce additional states and parameters, posing significant challenges in accurately estimating initial conditions. Since infectious disease dynamics are typically described by differential equations, precise initial values and parameters are essential for generating reliable predictions [14]. In reality, it is difficult to obtain accurate information about all states required for dynamic modeling. Government statistics usually contain only observable states, such as infections, recoveries, and deaths. Unobservable states, such as susceptible and exposed, are often unknown. Under such imprecise and incomplete observation, model parameters are also difficult to estimate accurately, especially in the early stages of epidemics [15]. Therefore, reducing errors in initial states and parameters is critical to improving the predictive performance and practical utility of epidemic models.

Data assimilation serves as a key approach to reducing initial value errors in dynamic models. It has long been regarded as one of the core methodologies in numerical weather prediction within meteorology [16]. By combining observations with model predictions, the data assimilation technique produces optimal estimates of model states. As early as 2012, a humidity-driven SIRS model was combined with the Ensemble Adjusted Kalman Filter (EAKF) for real-time forecasting of seasonal influenza outbreaks [17]. Subsequent studies compared various particle filters and ensemble filters in the context of influenza epidemics, demonstrating the effectiveness of Ensemble Kalman Filter (EnKF) methods in infectious disease modeling [18]. During the COVID-19 pandemic, data assimilation gained increased attention due to the urgent need for accurate epidemic forecasts. Many studies designed data assimilation algorithms to refine the states in COVID-19 prediction models [19–21]. Based on augmented state vector techniques, data assimilation enables simultaneous estimation of both model states and parameters [22]. Accordingly, a number of studies have applied such joint state-parameter estimation frameworks in COVID-19 dynamic modeling [23–25]. Combined optimization of states and parameters based on dynamic processes is a more reasonable and concise framework. However, applying data assimilation in epidemiology remains challenging due to the inherent nonlinearity and non-stationarity of epidemic processes. Covariance inflation and variable constraints are essential to maintain filter stability and avoid divergence in such settings. Commonly used covariance inflation techniques include correlated random walks [25,26] and fixed inflation factors applied separately to states and parameters [27]. To keep parameter estimates stable and within plausible ranges, parameter constraint techniques have also been developed in earth system modeling [28]. Building on these foundations, this study incorporates a state and parameter transformation scheme and an adaptive parameter covariance inflation strategy into the EnKF framework to enhance numerical stability and dynamic responsiveness in epidemic forecasting. Systematic simulation experiments and real-world case studies demonstrate how these enhancements significantly improve the predictive performance of infectious disease dynamics models.

2. Data and methodology

2.1. Infectious disease dynamic prediction model

The dynamic model of infectious disease employed by this study categorizes the population into five states: susceptible (S), exposed (E), infected (I), recovered (R), and dead (D). Assuming a total population of N, and $N = S + E + I + R + D$, then the following equations of dynamic model are satisfied:

$$\begin{cases} \dfrac{dS}{dt} = -\dfrac{\beta S(I+\theta E)}{N}, \\ \dfrac{dE}{dt} = \dfrac{\beta S(I+\theta E)}{N} - \alpha E, \\ \dfrac{dI}{dt} = \alpha E - \gamma I - \delta I, \\ \dfrac{dR}{dt} = \gamma I, \\ \dfrac{dD}{dt} = \delta I. \end{cases} \tag{1}$$

Here, the parameter β is interpreted as the probability that the susceptible people contact with infected individuals results in exposure to the virus. θ is the relative infectiousness of exposed individuals compared to infectious individuals. The parameter α is the conversion rate of exposed individuals to infected ones. The parameter γ denotes the recovery rate of infected individuals to recovered ones, and δ is the mortality rate from infected to dead individuals. Furthermore, we modify the current model to form dynamic prediction models of varying complexity to simulate different degrees of model error (Table 1). We label the SEIRD model that currently considers the infectivity of exposed individuals as SEIRD_e. The model that does not consider the infectivity of exposed individuals ($\theta = 0$) is labeled as SEIRD. After removing the death compartment, the SEIR model that considers the infectivity of exposed individuals is labeled as SEIR_e, and the model that does not consider the infectivity of exposed individuals is labeled as SEIR. Finally, after further removing the exposed compartment, we obtain the simplest prediction model SIR.

2.2. EnKF for combined state and parameter estimation

2.2.1. EnKF

By using the augmented state vector technique [22], we add the parameters of the infectious disease dynamic model to the state space and estimate the best combination of states and parameters. The main steps of the presented EnKF by this study can be divided into two parts: the prediction step and the analysis step. First, in the prediction step, each ensemble member predicts the states for the next step according to the epidemiological dynamic model:

$$x_t^f = M\left(x_{t-1}^a, \theta_{t-1}^a\right) + \varepsilon_t, \tag{2}$$

where $x_t^f \in R^{m\times n}$ is the ensemble of predicted state vector at time t, m is the number of ensemble samples, and n is the number of states. $x_{t-1}^a \in R^{m\times n}$ is the ensemble of analyzed value for the state vector at time $t-1$. $\theta_{t-1}^a \in R^{m\times l}$ is the ensemble of analyzed value for the parameter vector at time $t-1$, and l is the number of parameters.

Second, in the analysis step, we calculate the ensemble mean of predicted states $\overline{x}_t^f$ and parameters $\overline{\theta}_t$. Then, the prediction error covariance of the states P_t^x and parameters P_t^θ is obtained as follows:

Table 1
Information on infectious disease dynamics models of varying complexity.

Model name	Infectivity of exposed individuals	Parameters	Observable states	Unobservable states
SEIRD_e	Yes	$\beta,\theta,\alpha,\gamma,\delta$	$I,\ R,\ D$	$S,\ E$
SEIRD	No	$\beta,\alpha,\gamma,\delta$	$I,\ R,\ D$	$S,\ E$
SEIR_e	Yes	$\beta,\theta,\alpha,\gamma$	$I,\ R$	$S,\ E$
SEIR	No	β,α,γ	$I,\ R$	$S,\ E$
SIR	–	β,γ	$I,\ R$	S

$$\begin{pmatrix} \overline{x}_t^f \\ \overline{\theta}_t \end{pmatrix} = \begin{pmatrix} \frac{1}{m}\sum_i^m x_t^{f(i)} \\ \frac{1}{m}\sum_i^m \theta_t^i \end{pmatrix}, \tag{3}$$

$$\begin{pmatrix} P_t^x \\ P_t^\theta \end{pmatrix} = \begin{pmatrix} \frac{1}{m-1}\sum_{i=1}^m \left(x_t^{f(i)} - \overline{x}_t^f\right)\left(x_t^{f(i)} - \overline{x}_t^f\right)^T \\ \frac{1}{m-1}\sum_{i=1}^m \left(\theta_t^i - \overline{\theta}_t\right)\left(\theta_t^i - \overline{\theta}_t\right)^T \end{pmatrix}. \tag{4}$$

Since the parameters have no predicted values, here we let $\theta_t = \theta_{t-1}^a$. The parameters are updated only in the assimilation, and they remain unchanged in the prediction step. Next, update the state and parameters of ensemble members by calculating the Kalman gain matrix K_t:

$$\begin{pmatrix} x_t^{a(i)} \\ \theta_t^{a(i)} \end{pmatrix} = \begin{pmatrix} x_t^{f(i)} + K_t^x\left(y_t^i - Hx_t^{f(i)}\right) \\ \theta_t^i + K_t^\theta\left(y_t^i - Hx_t^{f(i)}\right) \end{pmatrix}, \tag{5}$$

$$K_t^x = P_t H^T\left(HP_tH^T + R\right)^{-1}, \tag{6}$$

where y_t is the observed value of the epidemic at time t, H is observation operator, R is the observation error covariance matrix, x_t^a is the analyzed value of states, and θ_t^a is the analyzed value of parameters. Specifically, for the observable state, the observation operator H is the identity matrix. For unobservable state, the Kalman gain matrix K_t^x is computed as follows:

$$K_t^x = \frac{\mathrm{cov}\left(x_t^{un}, x_t^o\right)}{\mathrm{var}\left(x_t^o\right) + R}, \tag{7}$$

where x_t^{un} is the model prediction for the unobservable state, and x_t^o is the model prediction for the observable state. The Kalman gain matrix used for parameter updating is calculated on the same principle as for unobservable variables:

$$K_t^\theta = \frac{\mathrm{cov}\left(\theta_t, x_t^o\right)}{\mathrm{var}\left(x_t^o\right) + R}. \tag{8}$$

The updated analysis value is the best estimate of the state at the current time step and serves as the initial value for the next time step's prediction.

2.2.2. State and parameter space transformation

To enforce physical constraints in the epidemiological dynamic model (non-negativity of state variables and $[0,\ 1]$ bounds for parameters), we implement state and parameter space transformation. During the update process of data assimilation, we apply the logarithmic transformation to the predicted results of states from the dynamic model, then perform the inverse transformation on the updated results to obtain the analyzed states. Similarly, we utilize logistic transformations for parameters, applying corresponding inverse transformations to the updated parameters to ensure they remain within their prescribed bounds.

For positivity-constrained states, the logarithmic transformation and its inverse transformation are given by:

$$\begin{cases} \widetilde{x} = \ln(x), \\ x = \exp(\widetilde{x}), \end{cases} \tag{9}$$

where, x represents the original state value, and $\widetilde{x}$ denotes the log-transformed value.

For bounded parameters, the logistic transformation and its inverse are given by:

$$\begin{cases} \widetilde{\theta} = \log\left(\frac{\theta}{1-\theta}\right), \\ \theta = \frac{1}{1+\exp\left(-\widetilde{\theta}\right)}, \end{cases} \tag{10}$$

where, θ is the original parameter value, and $\widetilde{\theta}$ is the logistic transformed value.

2.2.3. Adaptive covariance inflation

Observational noise often induces parameter oscillations that undermine prediction stability. Additionally, during shifts in control measures or environmental conditions, parameter response delays further compromise forecast accuracy. To address these challenges, we designed an adaptive covariance inflation mechanism. This mechanism adjusts the dispersion of updated posterior parameters, striking a balance between parameter stability and rapid response capability. Specifically, the covariance inflation factor $\lambda(t)$ decays gradually over time, with its decay rate $\tau(t)$ dynamically adjusted by prediction error: when prediction errors are large, the decay rate is reduced. This strategically reduces parameter oscillations during stable epidemic phases while preserving essential ensemble dispersion during abrupt changes. The formula for the covariance inflation mechanism is:

$$\lambda(t) = \lambda_0 \exp(-t/\tau(t)), \tag{11}$$

$$\tau(t) = \frac{\sqrt{\frac{1}{m}\sum_{i=1}^m \left(Hx_t^{f(i)} - y_t^i\right)^2}}{\sqrt{\frac{1}{m}\sum_{i=1}^m \left(y_t^i - \overline{y}_t\right)^2}}, \tag{12}$$

where, $\lambda_0 = 1.2$ is initial inflation factor.

2.3. Experimental design

2.3.1. Synthetic experiment

To fully assess the effectiveness of assimilated predictions, it is necessary to perform synthetic experiments under idealized conditions. Based on the SEIRD model considering the infectiousness of exposed individuals, we set the following parameters:$\beta = 0.5$, $\theta = 0.5$, $\alpha = 0.2$, $\gamma = 0.1$ and $\delta = 0.056$ and initial conditions:$S_0 = 999830$, $E_0 = 100$, $I_0 = 50$, $R_0 = 10$ and $D_0 = 10$. Then, we calculated the state sequence over 100 days as the "truth" of the epidemic. Further, 20 % random noise was introduced into the "truth" data, and 100 sets of noisy data were generated to simulate real observations. Noisy observable states (I, R, D) data is considered as "observations" and used as input for the dynamic prediction model and data assimilation model.

Since model parameters typically range from 0 to 1, we generate random parameters based on the beta distribution $Beta(2, 2)$ as initial parameters. During the experiments, we observed that the model rapidly forgets the initial parameter distribution and gradually converges toward the true parameters. We compared prediction errors between models with different structural errors under two scenarios: with and without data assimilation. Prediction models without data assimilation use the least squares method to update parameters in real time. The ensemble size of the Ensemble Kalman Filter was set to 1000. Additionally, we conducted sensitivity experiments with different ensemble sizes, as shown in Fig. S1.

2.3.2. Real experiment

To test the effectiveness of the assimilation algorithm in the real world, we selected outbreak report data of the Chinese provinces where the COVID-19 outbreak lasted more than 30 consecutive days in 2020 to conduct a realistic experiment. The outbreak report data are obtained from the Johns Hopkins University Center for Systems Science and En-

gineering's COVID-19 dataset, which is accessible via its GitHub repository (https://github.com/CSSEGISandData/COVID-19). Data variables include the cumulative number of confirmed cases, recoveries and deaths. Additionally, the number of existing confirmed cases was calculated by subtracting the number of recoveries and deaths from the cumulative number of confirmed cases. According to Eq. (13), the unknown states in the prediction model are initialized:

$$\begin{cases} E_0 = I_0 \ , \\ S_0 = N - E_0 - I_0 - R_0 - D_0 \ , \end{cases} \tag{13}$$

where, S_0 is the initial number of susceptible, E_0 is the initial number of exposed, I_0 is the initial number of infections, R_0 is the initial number of recovered, D_0 is the initial number of deaths, N is the number of population. In real experiments, the SEIRD model considering the infectiousness of exposed individuals was used for prediction.

2.3.3. Evaluation metric

We calculate the relative error metrics to assess the usefulness of the data assimilation algorithm. Since the true value of the epidemic states may be zero, we use the modified SMAPE indicator to measure the magnitude of the relative error. The formula for SMAPE is as the follows:

$$SMAPE = \frac{1}{n-1}\sum_{i=1}^{n}\frac{\left|\widehat{y}_i - y_i\right|}{\left(\left|\widehat{y}_i\right| + \left|y_i\right|\right)\Big/2} \ , \tag{14}$$

where $\widehat{y}_i$ is the evaluated value, y_i is the true value and n is the number of samples.

3. Results

3.1. Data assimilation enables more prompt and precise predictions

The joint assimilation of states and parameters significantly reduces state errors and parameter uncertainties in infectious disease dynamic model predictions. Synthetic experiment results demonstrate that the analyzed values of all states after data assimilation closely match the true values (Fig.S2), while model parameters progressively converge toward their true values (Fig.S3). Furthermore, comparative analysis of different covariance inflation methods (Supplementary Material S1) confirms that our adaptive approach achieves lower parameter estimation error. Data assimilation provides more reliable initial conditions for infectious disease dynamic model, enables more accurate and timely epidemic predictions.

Fig.1 presents the seven-day average prediction error rates across different predict starting days. Notably, during the early epidemic phase (less than 10 days), data assimilation substantially reduces prediction errors for both observable states (Fig.1(a)) and unobservable states (Fig.1(b)). This is crucial for rapid response in epidemic control. As observational data accumulates, prediction errors for observable states gradually decrease. However, specially for unobservable states, merely increasing data volume cannot fully eliminate errors caused by initial states inaccuracies and parameters biases, leading to evident error accumulation (Fig.1(b)). Data assimilation effectively suppresses the propagation of such errors, consistently yielding superior prediction accuracy for unobservable states compared to conventional methods. This finding confirms that data assimilation not only enhances the reliability of early-stage predictions but also mitigates error accumulation in later forecasts through continuous correction of system states and parameters, thereby providing enhanced robustness for infectious disease dynamic modeling.

3.2. Data assimilation enhances simple model beyond complex model

We simulate varying degrees of model error by intentionally modifying the model structure. Comparisons reveal that data assimilation consistently reduces prediction errors and mitigates the impact of model errors across all modified versions (Fig. 2(a)-(c)). Significantly, the simplest SIR model coupled with data assimilation outperforms SEIRD_e (a structurally complete model with no model error) without assimilation (Fig. 2(d)). This indicates that when coupled with data assimilation, simpler models can achieve superior predictive capability compared to complex models. While complex models with accurate structure yield optimal forecasting performance when integrated with data assimilation (Fig. 2(d)), employing simpler models with data assimilation proves more efficient in real-world forecasting contexts where infectious disease dynamics remain incompletely characterized.

In addition, we found that enhancing predictive performance by merely increasing model complexity is not always effective. Model complexity is not the sole determinant of predictive accuracy. Beyond model error, prediction errors in dynamical epidemic models also stem from errors in initial states and parameters estimates. As model complexity increases, so does the number of unknown state variables and parameters requiring estimation. The superior prediction performance of the SIR model over the SEIR model in our experiments is likely

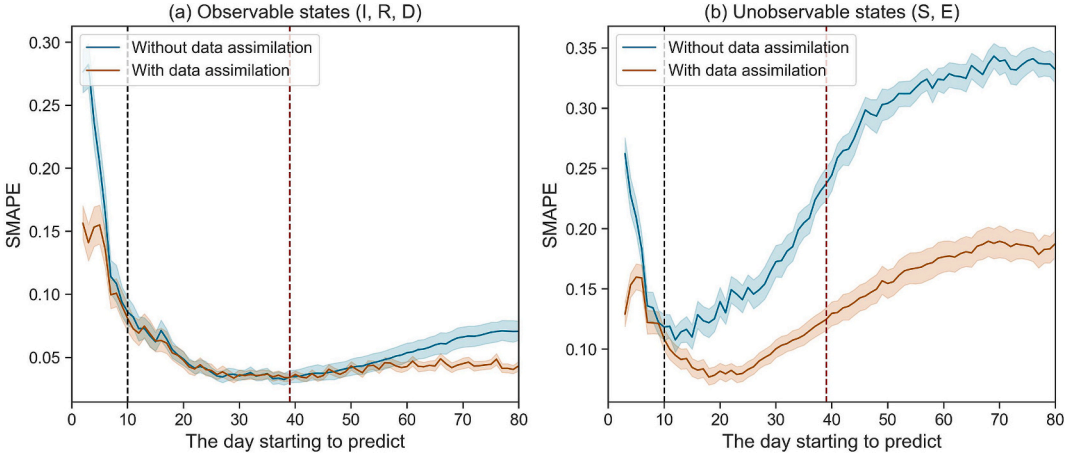

Fig. 1. Data assimilation effectively reduces prediction errors. The solid blue and orange lines depict the 7-day average SMAPE without and with data assimilation, respectively, with shaded regions indicating their 95 % confidence intervals. Critical epidemic timepoints are marked: day 10 (black dashed vertical line) and the epidemic peak day (red dashed vertical line). (a): displays the average SMAPE for predictions of observable states (I, R, D) across different forecast starting days during the epidemic; (b): presents corresponding results for unobservable states (S, E). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

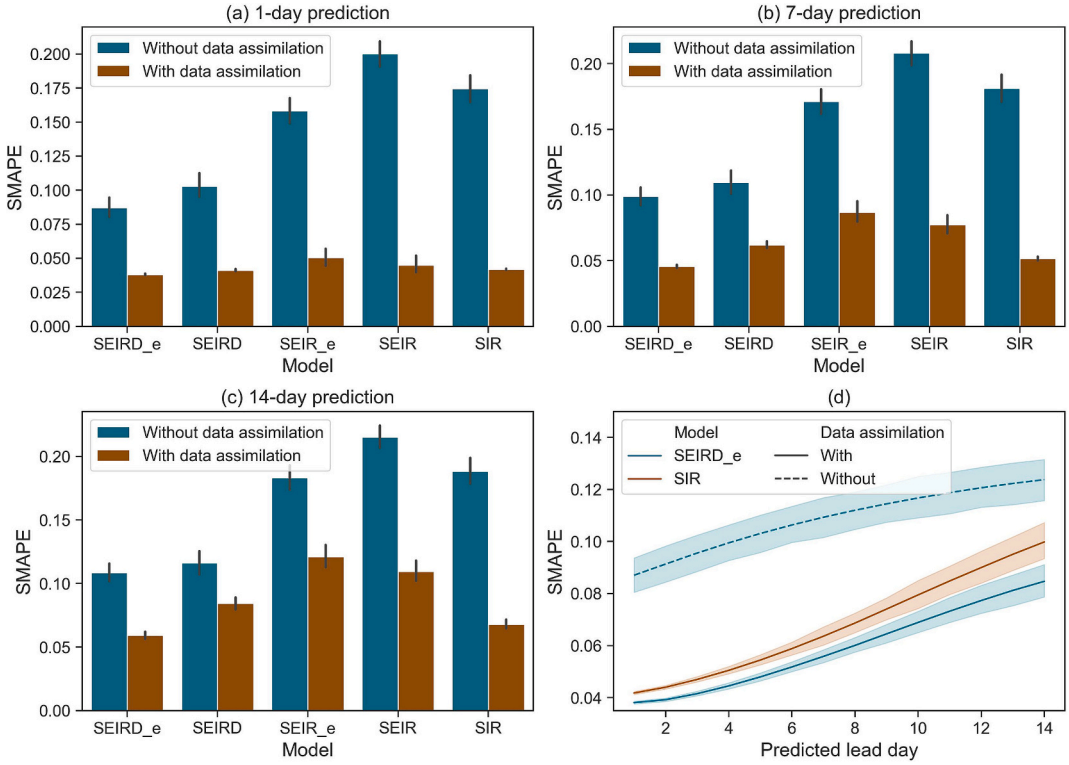

Fig. 2. Comparison of prediction results for models with varying complexity. (a)-(c): display SMAPE values for 1-day, 7-day, and 14-day predictions, respectively. Blue and orange lines indicate performance without and with data assimilation, with black lines showing 95 % confidence intervals. (d): compares prediction performance across three configurations: the simplest SIR model with data assimilation (orange solid line), the structurally complete SEIRD_e model without assimilation (blue dashed line), and the SEIRD_e model with assimilation (blue solid line). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

attributable to this reduced difficulty in states and parameters estimation.

3.3. Application to actual outbreaks of COVID-19 in China

In order to verify the validity of data assimilation in real-world scenarios, we conduct experiments by using the 2020 outbreak report data from 29 provincial administrative regions in China. As shown in Fig.3, data assimilation significantly enhances the predictive accuracy of the model. Notably, it heightens the model's sensitivity to epidemiological variations and fluctuations, enabling more responsive tracking of outbreak dynamics.

Fig.4(a) demonstrates consistent predictive improvement with data assimilation across all regions and all predicted lead days. Data assimilation has been shown to reduce 1-day prediction error rates by more than 50 % and 7-day error rates by approximately 15 % in comparison to predictions without data assimilation. Moreover, date and the number of infection cases at the epidemic peak are highly focused indicators in real-world infectious disease prediction. We define the peak day prediction accuracy as the proportion of regions for which the peak day is predicted correctly. The calculation formula is as $\frac{P_{acc}}{P}$, where P_{acc} is the number of regions with accurately predicted peak time, and P is the total number of regions. We consider the prediction to be accurate if the difference between the predicted and actual peak day is within three

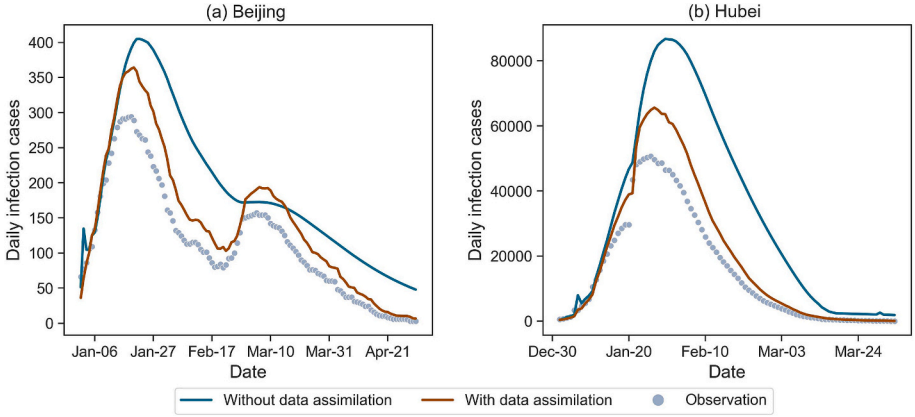

Fig. 3. Daily infection case predictions (1-day) for Beijing and Hubei Province. The orange and blue lines represent predictions with and without data assimilation respectively, while gray points indicate observed values. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

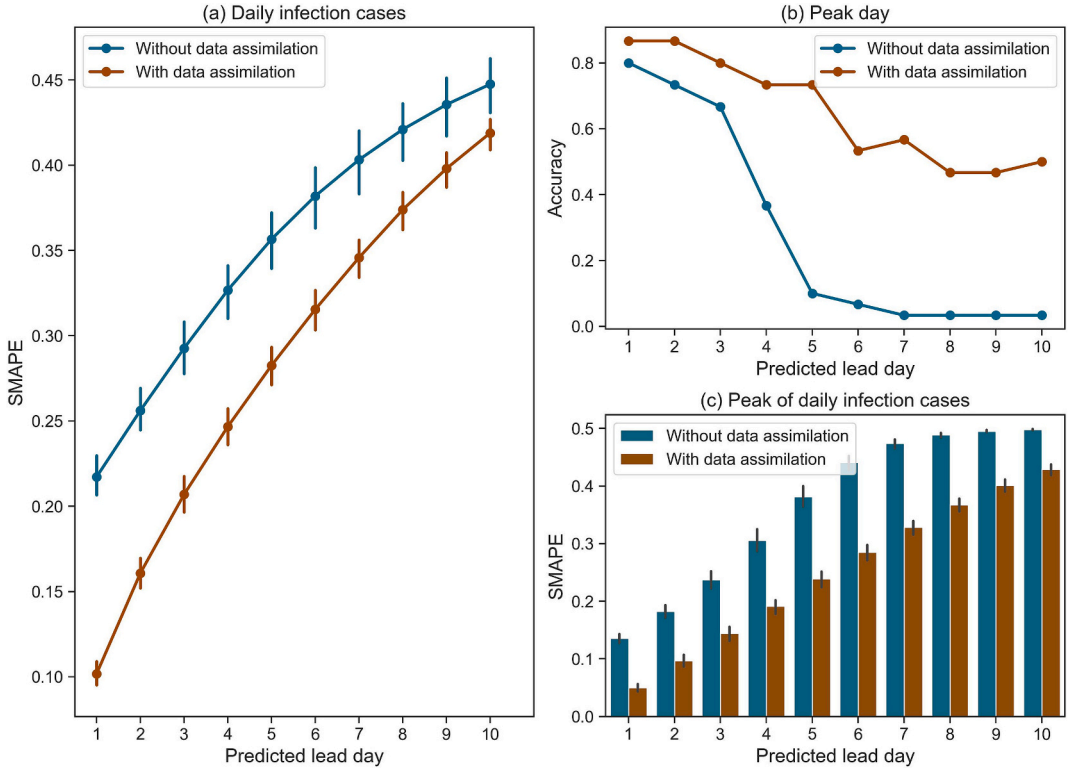

Fig. 4. Prediction performance for China's COVID-19 pandemic. Orange and blue lines indicate predictions with and without data assimilation respectively, while vertical bars denote 95 % confidence intervals. (a): shows prediction error across various predicted lead days; (b): shows peak day prediction accuracy at different predicted lead days; (c): shows peak infection cases predictions across various predicted lead days. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

days. As shown in Fig.4(b), data assimilation consistently improves accuracy in predicting peak day across all lead days. Without data assimilation, the accuracy of peak day prediction declines rapidly after a lead day of more than three days, while peak day prediction with data assimilation remains reliable even over longer lead days. Moreover, data assimilation achieves superior peak infection cases predictions across all regions (Fig.4(c)). The prediction accuracy rate for the peak day of the epidemic and the peak number of infected cases reached more than 70 % in advance by 3 days. Overall, data assimilation can markedly reduce the prediction error and enhance the prediction ability in real epidemics.

4. Discussion and conclusion

4.1. Contribution of state and parameter space transformation

EnKF relies fundamentally on the assumption that process and measurement noises are Gaussian. This assumption, however, is often violated in practical applications such as epidemiological modeling, where states and parameters can exhibit non-Gaussian distributions. We applied space transformations that projects state variables and parameters into a more Gaussian-compatible space. As shown in Fig. 5, which compares ensemble distributions of analysis and background (predictions from the infectious disease dynamic model) in different spaces during the data assimilation process, the transformation causes the skewness and kurtosis of the background distributions to more closely resemble those of a normal distribution. The updated analysis ensemble is expected to exhibit closer agreement with the observations (Fig. 5(a, c)). This transformation strategy helps preserve the physical interpretability of the updated states and parameters while enhancing their conformity with the Gaussian assumptions inherent in the EnKF framework, thereby improving numerical stability and estimation accuracy.

4.2. Real-time assimilation captures the dynamic evolution of the epidemic

During the early, data-scarce stages of an epidemic, data assimilation effectively reduces errors in initial values and parameters, enabling faster and more reliable outbreak predictions and enhancing the responsiveness of containment policies. In real-world epidemics, the dynamics of state transmission are not static, and transmission parameters change over time [10]. This necessitates constructing complex non-autonomous models. However, due to the greater number of states and parameters that are difficult to estimate, complex dynamic models face practical limitations. Balancing model complexity with accurate simulation of epidemic dynamics remains a critical technical challenge. We believe that combined optimization of both epidemic model parameters and states using data assimilation techniques can effectively address this issue. This technique achieves dynamic parameter evolution to accurately capture epidemic trends without increasing model complexity. It simultaneously mitigates prediction biases arising from insufficient modeling, enabling simple models to achieve superior predictive performance. For infectious disease prediction scenarios with unknown transmission mechanisms and high modeling difficulty, data assimilation facilitates the rapid deployment of predictive models. Undoubtedly, data assimilation holds significant practical application value in epidemic prediction.

4.3. Limitations of data assimilation in outbreak prediction

Although the method presented in this paper has achieved significant practical application value, it also has some limitations. Firstly, EnKF performance may be limited when dealing with highly nonlinear systems. In the future, data assimilation algorithms should be further optimized to meet the needs of more complex nonlinear infectious

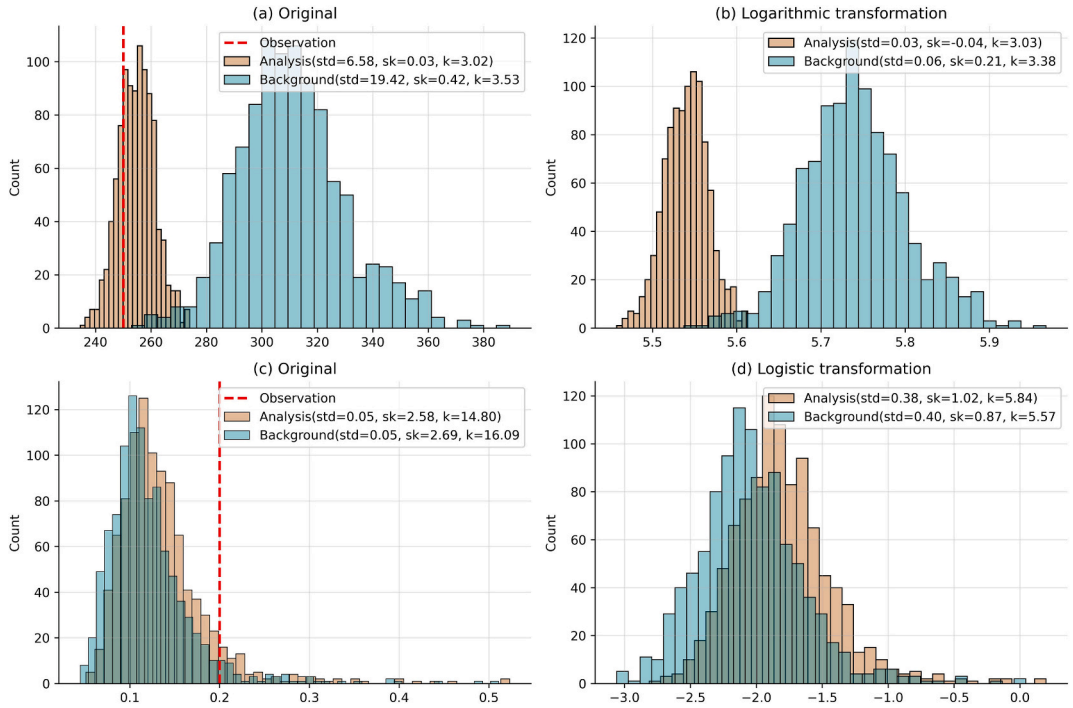

Fig. 5. Distributions of the analysis and background ensembles during data assimilation. The orange distributions represent the analysis ensembles, while the blue ones denote the background ensembles (forecast outputs from the infectious disease dynamic model). The red dashed vertical line indicates the observed value. Each distribution is annotated with its standard deviation (std), skewness (sk), and kurtosis (k). (a): shows the original forecast ensemble of the state variable I and the corresponding analysis ensemble after inverse transformation. Data assimilation brings the ensemble closer to the observation and reduces the standard deviation. (b): presents the dynamically forecast ensemble after logarithmic transformation and the updated analysis ensemble. Similarly, (c) and (d) depict the analysis and background ensemble distributions for the parameter α, before and after logistic transformation, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

disease dynamic systems. If sufficient data is available, recently developed data assimilation algorithms based on machine learning may offer more effective solutions, as they are better equipped to capture complex nonlinear relationships [29,30]. Secondly, due to the unknown errors in epidemic reporting data, accurately estimating the observation error covariance in data assimilation becomes challenging. This can influence the weighting of observations and model predictions during the assimilation process, thereby affecting the effectiveness of data assimilation [31]. Finally, this study developed an assimilation prediction framework based on the simple SEIRD model, demonstrating that more precise forecast outcomes can be achieved by refining the states and parameters. This framework can be flexibly applied to other infectious disease dynamic models. Future research could explore the data assimilation process of more complex dynamic models and focus on optimizing model errors arising from structural limitations.

4.4. Conclusion

With the ongoing challenges posed by global warming and increased ease of travel, environment is becoming more conducive to the survival and rapid spread of infectious diseases. As a result, accurate epidemic prediction has become increasingly critical for public health preparedness and response. However, inherent uncertainty in the initial states and parameters of infectious disease dynamic models remains a main obstacles to achieving reliable predictions. Our study proposes a generalized data assimilation framework for optimizing states and parameters in infectious disease dynamic models. Considering the characteristics of infectious disease prediction, we designed separate space transformation schemes for states and parameters to ensure that the data assimilation update process remains within reasonable bounds. Additionally, an adaptive covariance inflation strategy related to the epidemic development stage and prediction performance was employed to ensure stable parameter updates and a rapid response to changes in epidemic.

Synthetic experiments demonstrate that the proposed framework significantly reduces state and parameter errors, enabling more reliable early-stage predictions. It effectively estimates both observable and unobservable states, providing comprehensive insights into epidemic progression. Crucially, simple models coupled with data assimilation outperformed complex models without assimilation. Data assimilation reduces dependence on model complexity for infectious disease prediction and enhances the practicality of infectious disease dynamics models. Furthermore, validation using real-world COVID-19 data from China confirmed the framework's effectiveness and robustness. Compared to infectious disease dynamics models without data assimilation, prediction models using data assimilation demonstrate improved accuracy and stability of predictions. Data assimilation technology allows the dynamic model to adapt more efficiently to multi-peak epidemic outbreaks and irregular epidemic observation data. Additionally, data assimilation significantly improves the accuracy of predicting epidemic peak day and the peak number of infection cases.

In summary, this study emphasizes the significant role of data assimilation in improving epidemic prediction models. The combined assimilation of states and parameters offers a more accurate and adaptive approach, and it can be extended to the prediction of various infectious diseases. Data assimilation techniques are crucial for guiding timely public health responses, including resource allocation and intervention strategies.

CRediT authorship contribution statement

Zihan Hao: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Data curation. **Shujuan Hu:** Writing – review & editing, Supervision, Resources, Funding

acquisition, Conceptualization. **Jianping Huang:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **Jiaxuan Hu:** Writing – review & editing, Data curation. **Zhen Zhang:** Writing – review & editing, Data curation. **Jingjing Zhang:** Writing – review & editing. **Wei Yan:** Writing – review & editing. **Han Li:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was supported by the National Key Research and Development Program of China (2023YFC3503400), the Self-supporting Program of Guangzhou Laboratory (SRPG22-007), the Special Program of Guangzhou National Laboratory (GZNL2024A01004).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.onehlt.2025.101266.

Data availability

I have shared the link to research data in the manuscript

References

[1] WHO, Coronavirus Disease (COVID-19) Dashboard. https://covid19.who.int/, 2024 (Accessed 4 October 2024).
[2] J. Huang, L. Zhang, X. Liu, Y. Wei, C. Liu, X. Lian, Z. Huang, J. Chou, X. Liu, X. Li, K. Yang, J. Wang, H. Liang, Q. Gu, P. Du, T. Zhang, Global prediction system for COVID-19 pandemic, Sci. Bull. 65 (22) (2020) 1884–1887, https://doi.org/10.1016/j.scib.2020.08.002.
[3] C.C. John, V. Ponnusamy, S. Krishnan Chandrasekaran, R. Nandakumar, A survey on mathematical, machine learning and deep learning models for COVID-19 transmission and diagnosis, IEEE Rev. Biomed. Eng. 15 (2022) 325–340, https://doi.org/10.1109/RBME.2021.3069213.
[4] W.O. Kermack, A.G. McKendrick, G.T. Walker, A contribution to the mathematical theory of epidemics, Proc. R. Soc. Lond. A 115 (772) (1997) 700–721, https://doi.org/10.1098/rspa.1927.0118.
[5] O.N. Bjørnstad, K. Shea, M. Krzywinski, N. Altman, The SEIRS model for infectious disease dynamics, Nat. Methods 17 (6) (2020) 557–558, https://doi.org/10.1038/s41592-020-0856-2.
[6] S. He, Y. Peng, K. Sun, SEIR modeling of the COVID-19 and its dynamics, Nonlinear Dyn. 101 (3) (2020) 1667–1680, https://doi.org/10.1007/s11071-020-05743-y.
[7] Reiner, R. C., Barber, R. M., Collins, J. K., Zheng, P., Adolph, C., Albright, J., Antony, C. M., Aravkin, A. Y., Bachmeier, S. D., Bang-Jensen, B., Bannick, M. S., Bloom, S., Carter, A., Castro, E., Causey, K., Chakrabarti, S., Charlson, F. J., Cogen, R. M., Combs, E., … IHME COVID-19 Forecasting Team. (2021). Modeling COVID-19 scenarios for the United States. Nat. Med., 27(1), 1. doi: https://doi.org/10.1038/s41591-020-1132-9.
[8] M.O. Alassafi, M. Jarrah, R. Alotaibi, Time series predicting of COVID-19 based on deep learning, Neurocomputing 468 (2022) 335–344, https://doi.org/10.1016/j.neucom.2021.10.035.
[9] F. Shahid, A. Zameer, M. Muneeb, Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM, Chaos Solitons Fractals 140 (2020) 110212, https://doi.org/10.1016/j.chaos.2020.110212.
[10] S. Funk, S. Bansal, C.T. Bauch, K.T.D. Eames, W.J. Edmunds, A.P. Galvani, P. Klepac, Nine challenges in incorporating the dynamics of behaviour in infectious diseases models, Epidemics 10 (2015) 21–25, https://doi.org/10.1016/j.epidem.2014.09.005.
[11] W.C. Roda, M.B. Varughese, D. Han, M.Y. Li, Why is it difficult to accurately predict the COVID-19 epidemic? Infect. Dis. Model. 5 (2020) 271–281, https://doi.org/10.1016/j.idm.2020.03.001.
[12] S. Chang, E. Pierson, P.W. Koh, J. Gerardin, B. Redbird, D. Grusky, J. Leskovec, Mobility network models of COVID-19 explain inequities and inform reopening, Nature 589 (7840) (2021) 7840, https://doi.org/10.1038/s41586-020-2923-3.
[13] H. Li, J. Huang, X. Lian, Y. Zhao, W. Yan, L. Zhang, L. Li, Impact of human mobility on the epidemic spread during holidays, Infectious Dis. Model. 8 (4) (2023) 1108–1116, https://doi.org/10.1016/j.idm.2023.10.001.
[14] S. Pei, J. Shaman, Counteracting structural errors in ensemble forecast of influenza outbreaks, Nat. Commun. 8 (1) (2017) 925, https://doi.org/10.1038/s41467-017-01033-1.
[15] O. Melikechi, A.L. Young, T. Tang, T. Bowman, D. Dunson, J. Johndrow, Limits of epidemic prediction using SIR models, J. Math. Biol. 85 (4) (2022) 36, https://doi.org/10.1007/s00285-022-01804-5.
[16] A. Carrassi, M. Bocquet, L. Bertino, G. Evensen, Data assimilation in the geosciences: an overview of methods, issues, and perspectives, WIREs Clim. Change 9 (5) (2018) e535, https://doi.org/10.1002/wcc.535.
[17] J. Shaman, A. Karspeck, Forecasting seasonal outbreaks of influenza, Proc. Natl. Acad. Sci. U. S. A. 109 (50) (2012) 20425–20430, https://doi.org/10.1073/pnas.1208772109.
[18] W. Yang, A. Karspeck, J. Shaman, Comparison of filtering methods for the modeling and retrospective forecasting of influenza epidemics, PLoS Comput. Biol. 10 (4) (2014) e1003583, https://doi.org/10.1371/journal.pcbi.1003583.
[19] M.L. Daza-Torres, M.A. Capistrán, A. Capella, J.A. Christen, Bayesian sequential data assimilation for COVID-19 forecasting, Epidemics 39 (2022) 100564, https://doi.org/10.1016/j.epidem.2022.100564.
[20] P. Nadler, S. Wang, R. Arcucci, X. Yang, Y. Guo, An epidemiological modelling approach for COVID-19 via data assimilation, Eur. J. Epidemiol. 35 (8) (2020) 749–761, https://doi.org/10.1007/s10654-020-00676-7.
[21] A. Schaum, R. Bernal-Jaquez, L. Alarcon Ramos, Data-assimilation and state estimation for contact-based spreading processes using the ensemble kalman filter: application to COVID-19, Chaos, Solitons Fractals 157 (2022) 111887, https://doi.org/10.1016/j.chaos.2022.111887.
[22] G. Evensen, The ensemble Kalman filter for combined state and parameter estimation, in: IEEE Control Systems Magazine 29(3), 2009, pp. 83–104, https://doi.org/10.1109/MCS.2009.932223.
[23] G. Evensen, J. Amezcua, M. Bocquet, A. Carrassi, A. Farchi, A. Fowler, P. L. Houtekamer, C.K. Jones, R.J. de Moraes, M. Pulido, C. Sampson, An international initiative of predicting the SARS-CoV-2 pandemic using ensemble data assimilation, Found. Data Sci. 3 (3) (2021) 413–477, https://doi.org/10.3934/fods.2021001.
[24] R. Ghostine, M. Gharamti, S. Hassrouny, I. Hoteit, An extended SEIR model with vaccination for forecasting the COVID-19 pandemic in Saudi Arabia using an ensemble Kalman filter, Mathematics 9 (6) (2021) 636, https://doi.org/10.3390/math9060636.
[25] S. Rosa, M.A. Pulido, J.J. Ruiz, T.J. Cocucci, Transmission matrix parameter estimation of COVID-19 evolution with age compartments using ensemble-based data assimilation, PloS One 20 (4) (2025) e0318426, https://doi.org/10.1371/journal.pone.0318426.
[26] J. Liu, M. West, Combined parameter and state estimation in simulation-based filtering, in: A. Doucet, N. de Freitas, N. Gordon (Eds.), Sequential Monte Carlo Methods in Practice. Statistics for Engineering and Information Science, Springer, New York, 2001, pp. 197–223, https://doi.org/10.1007/978-1-4757-3437-9_10.
[27] J.J. Ruiz, M. Pulido, T. Miyoshi, Estimating model parameters with ensemble-based data assimilation: a review, J. Meteorol. Soc. Jpn. II 91 (2) (2013) 79–99, https://doi.org/10.2151/jmsj.2013-201.
[28] X.-M. Hu, F. Zhang, J.W.N. Gammon, Ensemble-based simultaneous state and parameter estimation for treatment of mesoscale model error: a real-data study, Geophys. Res. Lett. 37 (2010) L08802, https://doi.org/10.1029/2010GL043017.
[29] R. Arcucci, J. Zhu, S. Hu, Y.-K. Guo, Deep data assimilation: integrating deep learning with data assimilation, Appl. Sci. 11 (3) (2021) 3, https://doi.org/10.3390/app11031114.
[30] R. Cintra, H. de Campos Velho, S. Cocke, Tracking the model: data assimilation by artificial neural network, in: 2016 International Joint Conference on Neural Networks (IJCNN), 2016, pp. 403–410, https://doi.org/10.1109/IJCNN.2016.7727227.
[31] P. Tandeo, P. Ailliot, M. Bocquet, A. Carrassi, T. Miyoshi, M. Pulido, Y. Zhen, Joint Estimation of Model and Observation Error Covariance Matrices in Data Assimilation: A Review, 2018, https://doi.org/10.48550/arXiv.1807.11221.