Bulletin of Mathematical Biology (2021) 83:1
https://doi.org/10.1007/s11538-020-00834-8

Society for Mathematical Biology

METHODS

Sequential Data Assimilation of the Stochastic SEIR Epidemic Model for Regional COVID-19 Dynamics

Ralf Engbert[1] · Maximilian M. Rabe[1] · Reinhold Kliegl[2] · Sebastian Reich[3]

Received: 13 July 2020 / Accepted: 5 November 2020 / Published online: 8 December 2020

Abstract

Newly emerging pandemics like COVID-19 call for predictive models to implement precisely tuned responses to limit their deep impact on society. Standard epidemic models provide a theoretically well-founded dynamical description of disease incidence. For COVID-19 with infectiousness peaking before and at symptom onset, the SEIR model explains the hidden build-up of exposed individuals which creates challenges for containment strategies. However, spatial heterogeneity raises questions about the adequacy of modeling epidemic outbreaks on the level of a whole country. Here, we show that by applying sequential data assimilation to the stochastic SEIR epidemic model, we can capture the dynamic behavior of outbreaks on a regional level. Regional modeling, with relatively low numbers of infected and demographic noise, accounts for both spatial heterogeneity and stochasticity. Based on adapted models, short-term predictions can be achieved. Thus, with the help of these sequential data assimilation methods, more realistic epidemic models are within reach.

Keywords Stochastic epidemic model · Sequential data assimilation · Ensemble Kalman filter · COVID-19

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11538-020-00834-8) contains supplementary material, which is available to authorized users.

✉ Ralf Engbert
ralf.engbert@uni-potsdam.de

Maximilian M. Rabe
maximilian.rabe@uni-potsdam.de

Reinhold Kliegl
reinhold.kliegl@uni-potsdam.de

Sebastian Reich
sebastian.reich@uni-potsdam.de

[1] Department of Psychology, University of Potsdam, Potsdam, Germany

[2] Division of Training and Movement Sciences, University of Potsdam, Potsdam, Germany

[3] Institute of Mathematics, University of Potsdam, Potsdam, Germany

1 Introduction

The initial spread of the novel coronavirus in Germany (RKI 2020) resulted in containment measures based on reduced traveling and social distancing (Anderson et al. 2020). In epidemic standard models (Anderson et al. 1992; Kucharski et al. 2020), which provide a dynamical description of epidemic outbreaks (Bolker and Grenfell 1995; Schwartz and Smith 1983), containment measures aim at a reduction of the contact parameter. Since the contact parameter is one of the critical parameters that determine the speed of increase of the number of infectious individuals, estimating the contact parameter is a key basis for epidemic modeling (Lourenço et al. 2020).

From early on, the situation of COVID-19 has been characterized by extreme spatial heterogeneity (RKI 2020). In the initial phase of the outbreak, spatial heterogeneity was caused by random travel-based imports of infectious cases and enhanced by local events with increased contacts; after introduction of non-pharmaceutical interventions, spatial heterogeneity was sustained. Therefore, over the full observation period, the assumption of homogeneous mixing must be relaxed (Grenfell et al. 1995), and coupled dynamics of regional models seem to be a more adequate description (Li et al. 2020). However, when modeling a relatively small region with a population size of $N = 10^5$, compared to the country level with populations of $N = 10^7$ to 10^9, one must address the problem of stochasticity (Engbert and Drepper 1994; Grenfell et al. 1995) (Sect. 2.2). The combination of dynamical modeling and substantial fluctuations calls for sequential data assimilation methods for parameter inference (Law et al. 2015; Reich and Cotter 2015) as widely used, for example, in numerical weather prediction (Bauer et al. 2015).

We investigate how the stochastic SEIR epidemic model (Anderson et al. 1992) applies to regional data of COVID-19 incidence under non-pharmaceutical interventions, i.e., where epidemic dynamics were confined to regions and coupling between them could be neglected. The model assumes S, E, I, and R compartments representing susceptible, exposed, infectious, and recovered individuals (Fig. 1). This model is particularly important for the description of the spread of COVID-19, since infectiousness seems to peak on or before symptom onset (He et al. 2020b) and models without the exposed compartment cannot adequately address the time delay between the build-up of exposed and infectious individuals.

Since we are interested in short-term modeling (weeks to months), we neglect birth and death processes as a first-order approximation for the dynamics of the model. The disease-related model parameters are the rate parameters $a = 1/Z$ (with an average latency period Z) and $g = 1/D$ (with a mean infectious period D), which can be estimated independently from the analysis of infected cases (He et al. 2020b; Li et al. 2020). Therefore, the time-dependent contact parameter β is the most critical parameter that needs to be determined via data assimilation (Reich and Cotter 2015). It is directly related to the basic reproductive number r in a SEIR-type model (Sect. 2.1), which is average number of secondary cases by each infected case in a population consisting of susceptible individuals only (Dietz 1993; He et al. 2020a). Therefore, non-pharmaceutical interventions that aim at $r < 1$ translate into the relation $\beta < g$ in the model.

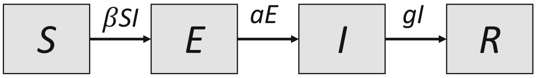

Fig. 1 The SEIR model. The population is divided into four compartments that represent susceptible, exposed, infectious, and recovered individuals. The contact parameter β is critical for disease transmission, and $1/a$ and $1/g$ are the average durations of exposed and infectious periods, respectively. Unlike in the standard model, the birth and death processes are neglected in short-term simulations discussed throughout the current study

In the following, we will use a combination of sequential data assimilation and stochastic modeling on the regional level to estimate the spatial heterogeneity in the spread of epidemics and to show how to use such a combined approach for epidemic predictions and uncertainty quantifications.

2 Mathematical Model and Statistical Inference

2.1 SEIR Model and Basic Reproductive Number

The SEIR epidemic model is a four-compartment model with susceptibles (individuals who are able to contract the disease), exposed (individuals who are infected but not yet infectious), infectious, and recovered (individuals who are immune). The model is typically formulated as a system of ordinary differential equations (ODE), i.e.,

$$\dot{S} = m - (m + \beta I)S \tag{1}$$
$$\dot{E} = \beta SI - (m + a)E \tag{2}$$
$$\dot{I} = aE - (m + g)I \tag{3}$$
$$\dot{R} = gI - mR\ , \tag{4}$$

where the total number of individuals $N = S + E + I + R$ is constant under temporal evolution due to $\dot{N} = 0$. The ODE system, Eqs. (1–4), has a non-trivial equilibrium point, denoted as epidemic equilibrium ($S^\star$, $E^\star$, $I^\star$, $R^\star$), where the number of susceptibles $S^\star$ at equilibrium is related to the basic reproductive number r by $r \cdot S^\star = 1$. Since we aim at a short-term description of the system, we neglect birth and death processes here; equivalent to the limit $m \to 0$, we obtain

$$r = \frac{1}{S^\star} = \frac{a\beta}{(m + a)(m + g)} \to \frac{\beta}{g} \quad \text{for} \quad m \to 0\ . \tag{5}$$

We use a numerical values of $g = 1/3$ per day, equivalent to an average infectious period of $D = 3$ days (Li et al. 2020). The critical condition for disease containment $r < 1$ is obtained for $\beta < \beta_{\text{crit}} = 1/3$ per day in our model. The median latency period has been estimated as 5.2 days (He et al. 2020b). Here, we used a numerical value of $a = 1/5.2$ per day for the rate parameter of the exposed individuals.

Table 1 Transitions and transition probabilities in the stochastic SEIR model. The transition are from state $X = (S, E, I, R)^{\mathrm{T}}$ to X' with probability $W_{X \to X'}$

X'				$W_{X \to X'}$
$S-1$	$E+1$	I	R	$\beta SI/N$
S	$E-1$	$I+1$	R	aE
S	E	$I-1$	$R+1$	gI

2.2 The Stochastic SEIR Model

While the classical model is formulated as a system of ordinary differential equations, we are exploring its application to a relatively low number of cases in the early phase of the current epidemics on the regional level with population sizes from 10^5 to 10^6. Therefore, we use the stochastic SEIR model in the form of a master equation (Engbert and Drepper 1994), which is particularly useful for modeling small numbers of infected individuals occurring in smaller regions or at the beginning of an epidemic.

The demographic SEIR model contains four variables denoted by $X = (S, E, I, R)^{\mathrm{T}} \in \mathbb{N}^4$, representing the number of individuals in each of the four classes with a constant population size $N = S + E + I + R$. The transition rate of the ODE compartmental model translates into transition probabilities in the master equation formulation for the evolution of the model's conditional probability density, that is,

$$\frac{\mathrm{d}}{\mathrm{d}t} p(X|X_0, t) = \sum_{X' \neq X} \left\{ W_{X' \to X}\, p(X'|X_0, t) - W_{X \to X'}\, p(X|X_0, t) \right\} \tag{6}$$

with transition probabilities given in Table 1 and initial condition X_0. The individual trajectories for the model's temporal evolution can be simulated exactly and numerically efficiently (Engbert and Drepper 1994) using Gillespie's algorithm (Gillespie 1976).

2.3 Sequential Data Assimilation

Publicly available data on the cumulative number of infected individuals are used to infer the model states $X = (S, E, I, R)^{\mathrm{T}}$ and the contact parameter β of the stochastic SEIR model. Note that the cumulative number of infected individuals corresponds to $Y = I + R$ in the SEIR model.

In the present study, we combine sequential data assimilation for the model states with an approximate log-likelihood function for the contact parameter (Reich and Cotter 2015). The basic algorithmic idea is to propagate an ensemble of M model forecasts using Gillespie's algorithm up to the next available observation point t_k. The forecast ensemble is denoted by $X_{\mathrm{f}}^{(n)}(t_k)$ with $n \in \{1, \ldots, M\}$. We used an ensemble size of $M = 100$ in this study. The reported cumulative number of infected individuals $y_{\mathrm{obs}}(t_k)$ is then used via a linear regression approach to obtain the adjusted model states $X_{\mathrm{a}}^{(n)}(t_k)$. This step is implemented via the ensemble Kalman filter in its formulation of (Sakov and Oke 2008; Reich and Cotter 2015). While the forecast ensemble is used to compute the temporary negative log-likelihood $L(t_k, \beta)$ of the model's contact parameter β at time t_k, the adjusted model states serve as starting values for the next Gillespie prediction cycle.

The above algorithm is run with a fixed range of contact parameters $\beta \in [\beta_{\min}, \beta_{\max}]$ and for a fixed time window $[t_{\text{initial}}, t_{\text{final}}]$ of available data points $y_{\text{obs}}(t_k)$. The best fit contact parameter $\beta_*(t_k)$ at any time t_k is the one that minimizes the temporary negative log-likelihood function, that is,

$$\beta_*(t_k) = \arg\min_{\beta} L(t_k, \beta) \tag{7}$$

with $L(t_k, \beta)$ defined by (13) below.

2.3.1 Ensemble Kalman Filter

The reported cumulative number of infected individuals $y_{\text{obs}}(t_k)$ is linked to the model states $X = (S, E, I, R)^{\text{T}}$ via

$$Y(t_k) := I(t_k) + R(t_k) = HX(t_k) \ , \tag{8}$$

i.e., $H = (0, 0, 1, 1)$. As an initial condition, we set $I(t_0)$ equal to the reported number of infected cases and $R(t_0) = 0$ so that $y_{\text{obs}}(t_0) = I(t_0)+R(t_0)$, and $E(t_0) = g/a\cdot I(t_0)$ with additive noise. We assume that the errors in the observed $y_{\text{obs}}(t_k)$ are additive Gaussian with mean zero and variance ρ. We set $\rho = 10$ in our experiments.

The ensemble Kalman filter provides a computationally robust algorithm for updating a model-based forecast ensemble $X_{\text{f}}^{(n)}(t_k)$ at time t_k into posterior model states $X_{\text{a}}^{(n)}(t_k)$ utilizing the observed $y_{\text{obs}}(t_k)$ and its forward model (8). The assimilation step is based on a Gaussian or linear regression approximation of the underlying Bayesian inference problem with the forecast ensemble constituting a Monte Carlo approximation to the prior distribution over the model states $X(t_k)$ (Evensen 2006; Reich and Cotter 2015). More precisely, the ensemble Kalman filter is based on the empirical mean

$$m_{\text{f}}(t_k) := \frac{1}{M}\sum_{n=1}^{M} X_{\text{f}}^{(n)}(t_k) \in \mathbb{R}^4 \tag{9}$$

and the empirical covariance matrix

$$P_{\text{f}}(t_k) := \frac{1}{M}\sum_{n=1}^{M} \left(X_{\text{f}}^{(n)}(t_k) - m_{\text{f}}(t_k)\right)\left(X_{\text{f}}^{(n)}(t_k) - m_{\text{f}}(t_k)\right)^{\text{T}} \in \mathbb{R}^{4\times 4} \tag{10}$$

of the forecast ensemble. These two quantities are used to quantify the forecast uncertainty. Combining the implied Gaussian approximation of the forecast uncertainty with the assumed linear forward model (8) leads to the ensemble Kalman filter update formula (Sakov and Oke 2008; Reich and Cotter 2015)

$$X_{\mathrm{a}}^{(n)}(t_k) := X_{\mathrm{f}}^{(n)}(t_k) - \frac{1}{2}K(t_k)\left\{HX_{\mathrm{f}}^{(n)}(t_k) + Hm_{\mathrm{f}}(t_k) - 2y_{\mathrm{obs}}(t_k)\right\} \tag{11}$$

with the Kalman gain defined by

$$K(t_k) := P_{\mathrm{f}}(t_k)H^{\mathrm{T}}\left\{HP_{\mathrm{f}}(t_k)H^{\mathrm{T}} + \rho\right\}^{-1} \in \mathbb{R}^{4\times 1}. \tag{12}$$

One notes that the Kalman gain $HK(t_k)$ increases as the forecast variance $HP_{\mathrm{f}}(t_k)H^{\mathrm{T}}$ in the observed quantity (8) increases while being bounded by one from above. This in turn implies that the forecast is more strongly affected by the data and vice versa in case the forecast variance is reduced. We note that the resulting analysis ensemble $X_{\mathrm{a}}^{(n)}(t_k) \in \mathbb{R}^4$, $n \in \{1, \ldots, N\}$, can be mapped back onto the integers $\mathbb{N}^4$ if needed.

2.3.2 Model evidence

The model's negative log-likelihood at an observation time t_k is approximated by

$$L(t_k, \beta) := \frac{1}{2}\frac{|Hm_{\mathrm{f}}(t_k) - y_{\mathrm{obs}}(t_k)|^2}{HP_{\mathrm{f}}(t_k)H^{\mathrm{T}} + \rho} + \frac{1}{2}\log(HP_{\mathrm{f}}(t_k)H^{\mathrm{T}} + \rho). \tag{13}$$

Note that the first contribution penalizes the data misfit, while the second penalizes model uncertainty. The smaller the negative log-likelihood, the better the chosen model parameter β fits the data $y_{\mathrm{obs}}(t_k)$ at time t_k. The best parameter fit over a time window $[t_{\min}, t_{\max}]$ is defined as the value of β which minimizes the cumulative negative log-likelihood

$$L_{\mathrm{cum}}(\beta) = \sum_{t_k=t_{\min}}^{t_{\max}} L(t_k, \beta), \tag{14}$$

that is,

$$\beta_* := \arg\min_{\beta} L_{\mathrm{cum}}(\beta). \tag{15}$$

3 Methods

3.1 Parameter Recovery from Simulated Data

To test the inference scheme, we simulated data for 20 days. In Fig. 2a, the black line indicates the evolution of the SEIR model's predicted cumulative numbers of infected individuals, $Y(t_k) = HX(t_k) = I(t_k) + R(t_k)$. As in the real data, red dots represent the daily number of reported cases. In the simulation, the contact rate was chosen as $\beta_{\mathrm{true}} = 0.6$. In the following, we analyzed whether this true value could be recovered using the inference procedures described above.

We varied the contact rate β and determined the cumulative negative log-likelihood values $L_{\mathrm{cum}}(\beta)$, Eq. (14). The position of the minimum of $L_{\mathrm{cum}}(\beta)$ indicates the best estimate for the numerical value of the underlying contact rate β_*, Eq. (15). The position of the minimum turns out to be close to the true value, $\beta_* \approx \beta_{\mathrm{true}} = 0.6$

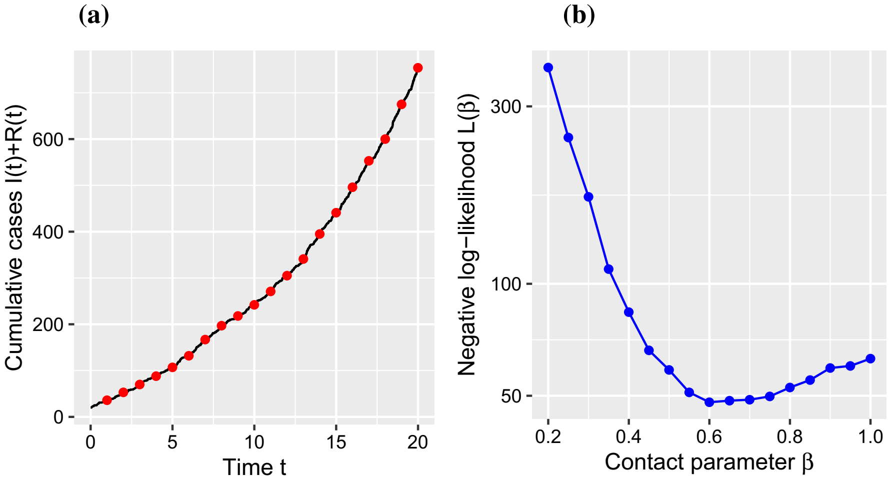

Fig. 2 (Color figure online)Parameter recovery analysis. **a** Simulated data with $b = 0.6$. **b** Negative log-likelihood $L_{\text{cum}}(\beta)$ indicates a minimum at about the true parameter value

(Fig. 2b). Thus, parameter recovery can be demonstrated for a relatively short time series of 10 observations, which represents a typical dataset in the early phase of newly emerging epidemics. Next, we apply our inference scheme to real data.

3.2 Application to Empirical Data

Since the parameter inference was successful with simulated data, the next step was an application to empirical observations. We applied the inference framework to two regional data sets from the RKI data base. As an example, we selected the COVID-19 time series for Köln (RKI data, population size $N = 1,085,664$), which includes 27 days of observations with more than 30 cases and is plotted in Fig. 3a. The parameter estimation yields an estimate for the contact rate of $\beta_* \approx 0.7$ (Fig. 3b). Thus, an analysis of the negative log-likelihood function produced qualitatively similar results for the simulated SEIR time series and the empirical data for a representative region. In the main text, we carry out an estimation of the time-resolved instantaneous optimal parameter values $\beta_*(t_k)$, Eq. (7), using the instantaneous negative log-likelihood function $L(t_k, \beta)$, Eq. (13).

We found that our results were relatively insensitive to the choice of the measurement error variance ρ appearing in (12) and (13). At the same time, we emphasize that the errors in the reported cumulative numbers of infected individuals are complex, may vary over time, and will certainly impact on the inferred parameters. The same applies to the unknown initial model states $X(t_0) = (S(t_0), E(t_0), I(t_0), R(t_0))^{\text{T}}$ and their uncertainties.

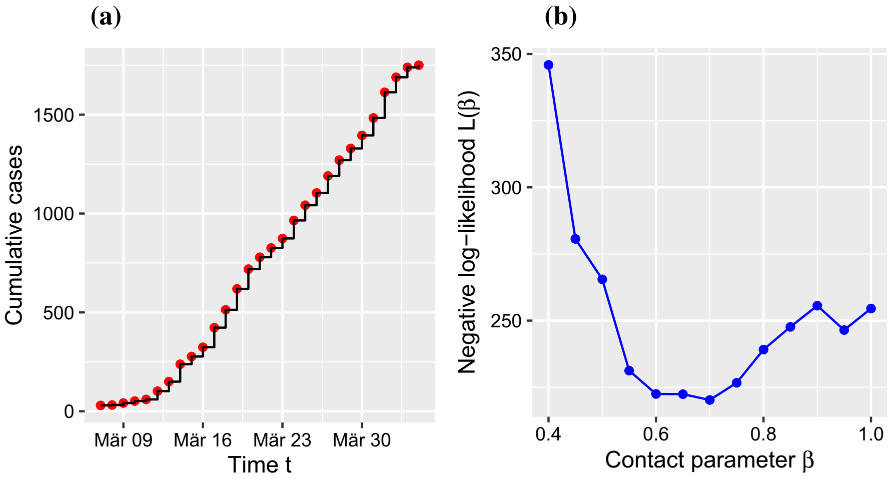

Fig. 3 (Color figure online)Contact parameter estimates for real data. **a** Data for Köln; date refers to report of case at RKI. **b** Negative log-likelihood $L_{\text{cum}}(\beta)$ for Köln give a minimum at $\beta_* \approx 0.7$

3.3 RKI Data on COVID-19 in Germany

The Robert Koch Institute (RKI), the central scientific institution in the field of biomedicine within the portfolio of the Federal Ministry of Health, provides daily access to the number of confirmed cases, deaths, and recovered patients, broken down into 412 counties, six age groups, and by gender. As they are official records, only cases certified by doctors or laboratories according to a strict medical protocol in accordance with the Infection Protection Act are entered into the data base. The exact time of an infection is usually not known. The associated time stamp refers to the date on which the local health authority became aware of the case and recorded it electronically. As records are passed from the physician or lab via local and state health authorities to the RKI, there is a delay of several days before cases are reported on the website. Thus, the statistics relating to the most recent three or four days are incomplete and cannot be interpreted; retrospective updates and corrections are made for all days of the pandemic spell as they become available. Also the date at which the case is reported at the RKI (i.e., the Erkrankungsdatum/date of illness) is not more recent than the date at which the infection occurred. We use data up to and including April 30, as reported on June 22, 2020; they are included as part of the supplement.

4 Results

The key motivation of the current study was to apply sequential data assimilation to the stochastic SEIR model to estimate the contact parameter. We successfully applied an ensemble Kalman filter (Evensen 2006; Law et al. 2015; Reich and Cotter 2015) to recover the contact parameter from simulated data (Sect. 3.1). When applied to empirical data on a regional level, the estimation of the contact parameter produces a comparable evidence profile (Sect. 3.2).

In the early phase of the COVID-19 outbreak in Germany, the reported cumulative numbers of cases increased rapidly (Fig. 4a,b); however, the epidemic dynamics vary from region to region. This spatial heterogeneity is due to the different onset times of the disease in different regions, but it is also enhanced by variations in the local contact parameters β. In response to the containment measures, we expect β to change over time.

4.1 Time Dependence of the Contact Parameter

An estimation of the time dependence of the contact parameter is achieved via the model's best fit. An approximative instantaneous negative log-likelihood $L(t_k, \beta)$ of the contact parameter β at observation time t_k is obtained from the ensemble Kalman filter (Sect. 2.3). Thus, by determining the minimum of $L(t_k, \beta)$ with respect to β at time t_k, we estimate the time dependence of the best fit $\beta_*(t_k)$ (Fig. 4c). The black line represents the average time dependence for all 320 regions included in the analysis; standard deviations are indicated by the gray area. The results for the two example regions are given in their corresponding colors.

The non-pharmaceutical interventions to counter the spread of COVID-19 were implemented at slightly different points in time across Germany. In the majority of the regions, closings of schools and other educational institutions started on March 16, while large-scale contact bans were implemented on March 22. As a result of these social distancing measures will have an impact on the contact parameter, we expected to observe a related drop in the contact parameter over time. Before we present a corresponding analysis, it should be made clear that none of these measures can produce an immediate effect on the observed cases of infected individuals because of the latency period. Since sequential data assimilation will need several data points to adapt the model to the data, the related interval should be as long as possible to achieve a reliable estimate of the contact parameter. Therefore, we selected the average value of $\beta_*(t_k)$ over the three days from March 17 to March 19 as a pre-intervention value. The average over March 31 to April 2 is taken as an estimate of the post-intervention value. To analyze the effect across regions, we computed average values β_{pre} (March 17-19) and β_{post} (March 31-April 2) of the relevant $\beta_*(t_k)$ for all regions. The resulting scatter plot indicates a clear reduction of the numerical value of the contact parameter from β_{pre} to β_{post} (Fig. 4d). The reduction is statistically significant (Wilcoxon test, $p < 0.01$).

4.2 Simulations with Time-Dependent Contact Parameter

The contact parameter β is the most critical parameter determining the dynamics of the stochastic SEIR model. After a time-resolved estimation of the best fit $\beta_*(t_k)$, we are able to generate simulations from an initial state to predict the future trajectory (Fig. 5). Simulations I begin with the first epidemic day in the corresponding region with greater than or equal to 30 cases. The initial number of infected I_0 is set to

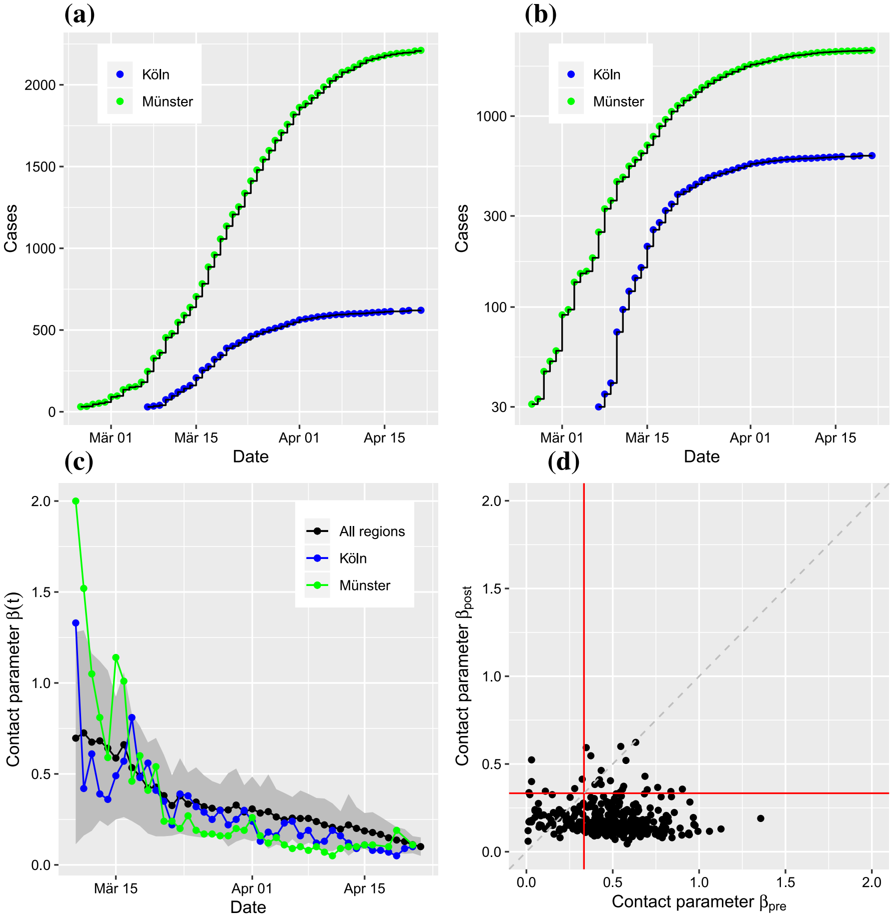

Fig. 4 (Color figure online) Analysis of best fit time-dependent contact parameters $\beta_*(t_k)$; date refers to report of case at RKI. **a** For two regions (LK Köln and LK Münster), the cumulative numbers show a strong increase after different disease onset times. **b** Semi-logarithmic scaling suggests approximate exponential growth in early as well as later regimes. **c** The time-dependent contact parameter $\beta_*(t_k)$ indicates a small decrease over time due to social distancing interventions (black: average for 320 regions; red, blue: contact parameter for the examples above; gray shading: standard deviation across regions. **d** Scatter plot of the time averaged contact parameter β_{pre} before intervention and β_{post} after intervention. Note that the critical value for disease containment is $\beta_{\text{crit}} = 1/3$ per day in our model (red lines)

the observed number of cases $y_{\text{obs}}(t_0)$, while the initial number of exposed is set to $E_0 = g/a \cdot I_0$, which would hold at epidemic equilibrium for $m > 0$ (for $m = 0$ both E and I tend to zero with $E^\star/I^\star \approx g/a$). The initial number of infected was disturbed by noise representing uncertainties in the initial model states. The initial number of susceptibles was set to N, which must be replaced by an estimate as the epidemics in unfolding more strongly, of course. Forward iterations with the estimated time-varying contact parameter show that the slope of the epidemic curve is approximately

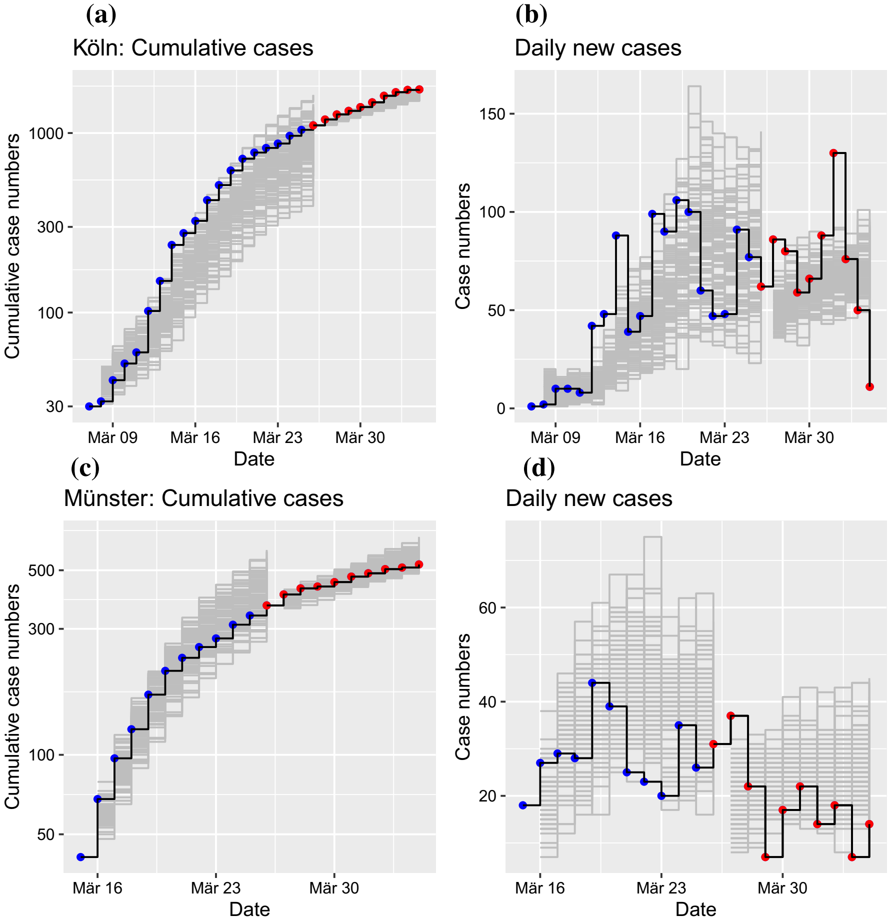

Fig. 5 (Color figure online) Simulations of the stochastic SEIR model for two example regions. Simulations I indicate an ensemble of 100 runs of the model with initial conditions from the first epidemic day with number of cases greater than or equal to 30 (gray: ensemble of trajectories; blue: observations). Simulations II start at March 26, using an ensemble size of 100 after data assimilation (gray: ensemble of trajectories; red: observations). **a** Cumulative cases of infected individuals over time for LK Köln. **b** Daily reported new cases for Köln. **c** Cumulative cases for LK Münster. **d** Daily new cases for Münster. Date refers to report of case at RKI

reproduced by the model (Fig. 5a,c; gray lines indicate the ensemble of simulated trajectories; blue points are observed data).

Simulation II starts at March 26 and exploits the full potential of sequential data assimilation. The sequential data assimilation approach via the ensemble Kalman filter (Sect. 2.3.1) is based on the forward modeling of an ensemble of trajectories. After each time step (1 day), the ensemble of trajectories is compared to the next observation and adjusted via a linear regression step. Thus, we obtained an adapted ensemble of internal model states for each epidemic day. Here, we exploit this fact to run a forward simulation with initial conditions from the assimilated ensemble of internal model states. The corresponding forward simulations are close to the real time-evolution of the epidemics in the two example regions (Fig. 5a,c; gray lines indicate the ensemble of simulated trajectories; blue points are observed data). A related plot of the daily reported new cases indicates an approximately constant level of numbers of new cases for Köln (Fig. 5b) and slowly decreasing daily new cases for Münster (Fig. 5d); both predictions are in agreement with empirical observations.

4.3 Predictions for Two Different Scenarios

The forward simulations discussed in the previous section demonstrated the predictive power of the SEIR model when using sequential data assimilation. In the next step, we generated simulations under two different scenarios. In scenario I, we started with the adapted ensemble of internal model states after data assimilation (April 16) and iterated the model forward with the mean contact parameter estimated from the period of April 14 to April 16, that is well after interventions were implemented (Fig. 6, green area). The simulations continue to match the time course of infected cases for both example regions (Fig. 6a,b). Daily reported case numbers show a decline for both regions (Fig. 6c,d).

In scenario II, we assumed that all governmental intervention measures had been terminated. Therefore, we used the estimated contact parameters from the period of March 17 to March 19. Again, we started simulations with the adapted ensemble of internal states after sequential data assimilation (Fig. 6, red area). For both example regions, we observe a strong increase in infected cases under scenario II (Fig. 6c,d). This dramatic increase can be seen most clearly in the plot of daily numbers of new cases (for more examples, see section A).

5 Discussion

The ongoing worldwide spread of the new coronavirus exerts enormous pressure on healthcare systems, societies and governments. Therefore, predicting the epidemic dynamics under the influence of non-pharmaceutical interventions (NPI) is an important problem from a data science and mathematical modeling perspective (Maier and Brockmann 2020). The motivation of the current work was to explore the potential of sequential data assimilation (Law et al. 2015; Reich and Cotter 2015) to create a regional epidemic model as a forecasting tool.

The standard epidemic SEIR-type models implement a compartmental description under the assumption of homogeneous mixing of individuals (Anderson et al. 1992). More realistic modeling approaches must account for spatial heterogeneity due to time-varying disease onset times, regionally different contact rates, and the time dependence of the contact rates due to the implementation of containment strategies. However, these regional descriptions require models that include the effects of demographic stochasticity due to the limited size of populations and the low number of cases in the region considered (Bittihn and Golestanian 2020). The effects of such statistical

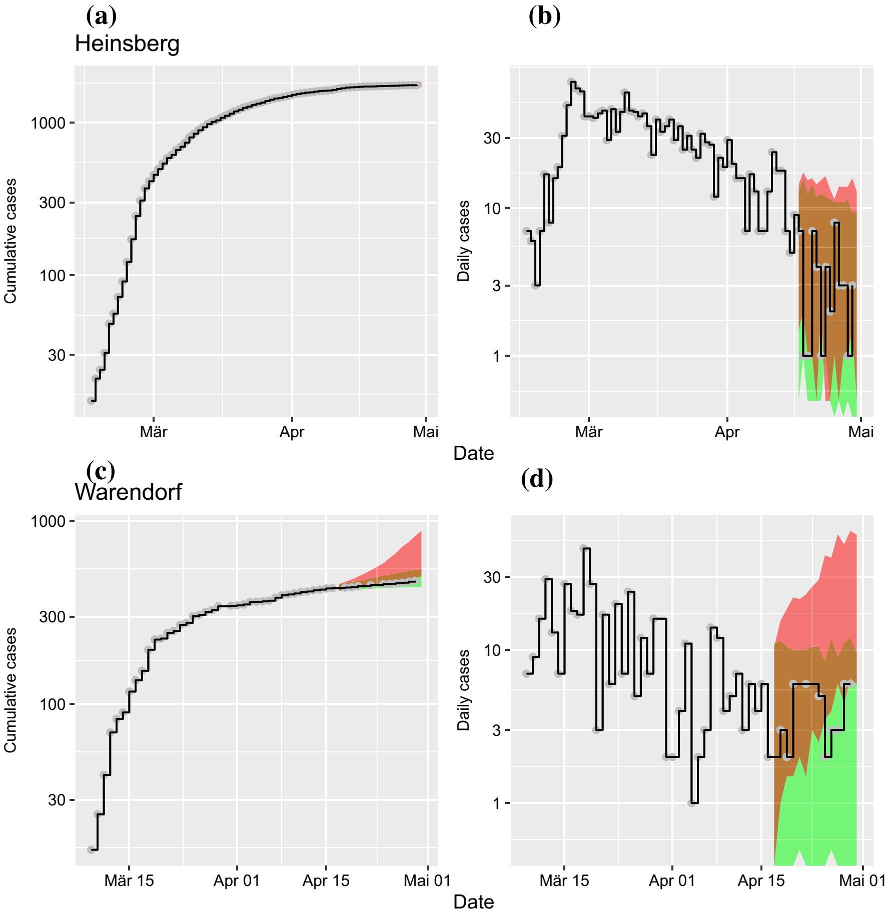

Fig. 6 (Color figure online) Model predictions for COVID-19 after data assimilation in comparison to data (black lines). In scenario I (green area), an assimilated ensemble of internal model states starts the forecast with contact parameter β_{post} (continuation of social distancing interventions). In scenario II (red area), the equivalent forecast is generated with contact parameter β_{pre} (termination of interventions). **a** Predictions for cumulative case numbers in Heinsberg. **b** Predictions of daily new cases in Heinsberg. **c** Predictions of cumulative cases for Warendorf. **d** Predictions of daily new cases for Warendorf. Date refers to report of case at RKI

fluctuations are inherently reproduced via stochastic versions of the standard epidemic models (Engbert and Drepper 1994; Grenfell et al. 1995).

We have demonstrated the potential of sequential data assimilation to reproduce COVID-19 dynamics at the level of a regional, stochastic model. With the help of the ensemble Kalman filter (Evensen 2006), we successfully recovered the contact parameter from the simulated data and obtained reliable estimates from the empirical data. The contact parameter is the most critical free parameter in the stochastic SEIR model, since the other parameters (mean exposed and infectious duration) can be estimated independently from observed time series (He et al. 2020b; Li et al. 2020).

Moreover, the contact parameter of the SEIR model is directly related to the basic reproductive number r (Liu et al. 2020). Therefore, our approach could also be framed as a model-based method for statistical inference of the basic reproductive number.

Next, we ran a time-resolved data assimilation that generated estimates of the time dependence of the contact parameter. The drop in mean contact rates from an early (β_{pre}, March 17 to 19) to a later period (β_{post}, March 31 to April 2) indicates the effect of non-pharmaceutical interventions. We also generated model prediction for two different scenarios. In scenario I, simulations produce forecasts with start date April 16. The previously assimilated ensemble provide the initial conditions and the contact parameter is set to the value estimated for the post-intervention period from April 14 to April 16. In scenario II, we replaced the post-intervention contact parameter with its pre-intervention value, estimated from the data for the period March 17 to March 19. As a result, the two scenarios predict rather different temporal developments (decline of daily new cases for scenario I and strong increase for scenario II). Therefore, our model predictions suggest that lifting of the non-pharmaceutical interventions could potentially turn the epidemic dynamics back to the exponential increase from before their implementation. Such predictions can easily be scaled up to the federal state level (Bundesländer) or to the country level; a corresponding predictive model will be potentially quite robust due to its explicit modeling of spatial and temporal heterogeneities, captured by a separate time course of the contact parameter for each region.

A recent simulation study by Li et al. (2020) used a similar approach of sequential data assimilation for dynamic epidemic models. However, they implemented a deterministic SEIR model and extended it with additional noise assumptions. We proposed the usage of the stochastic SEIR model in the formulation of a master equation (Engbert and Drepper 1994) which can be simulated exactly and numerically efficiently using Gillespie's algorithm (Gillespie 1976). A more complex spatiotemporal stochastic model has been considered in Arenas et al. (2020).

Furthermore, the state-parameter estimation in Li et al. (2020) utilizes the ensemble Kalman filter directly on an augmented state space (Reich and Cotter 2015). Contrary to that study, we found a direct application of the ensemble Kalman filter to the augmented state space (X, β) not suitable because of the strongly nonlinear interaction between the model states X and the contact parameter β. This led us to consider a two stage approach which combines the ensemble Kalman filter for state estimation with a likelihood-based inference of the contact parameter β (Reich and Cotter 2015). The proposed two-stage approach can be extended to the estimation of multiple model parameters including the generally unknown initial states of the stochastic SEIR model. However, the computational complexity will increase exponentially with the number of parameters to be estimated and more refined Monte Carlo methods for combined state and parameter estimation will be required if the total number of parameters exceeds three or four (Reich and Cotter 2015).

Our current study was mainly motivated by the methodological problem of a possible contribution from data assimilation to epidemics modeling based on a stochastic SEIR model. There are obvious limitations within our current modeling framework, which we did not address because of our methodological focus. Longer-term predictions ($\sim$ months) are important, but they critically depend on an estimation of undocumented infections (see Li et al. 2020). Such hidden infections create, after recovery, an unknown reduction in the number of susceptibles, which slows down the epidemic dynamics; such an effect is currently not included in our current model. However, it seems compatible with our framework to extend the SEIR model by an additional class of undocumented infected individuals (Li et al. 2020).

Another important limitation of these results comes from the simplifying assumption that there is no coupling to neighboring regions. As a consequence, the regional differences in the contact parameter could be at least partly due to differences in the contacts between the regions. Couplings between the regions (Li et al. 2020) could also be integrated into our modeling framework. However, the non-coupling approximation might be realistic in the situation of social distancing and travel bans during the period investigated here.

Acknowledgements We thank Klaus Dietz, Tübingen, for comments on the manuscript. This work was supported by a grant from Deutsche Forschungsgemeinschaft, Germany (SFB 1294, Project No. 318763901).

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability Statement Data and source code for simulations, analyses, and figures are available via Open Science Framework (OSF) at https://osf.io/7dshm/

References

Anderson RM, Anderson B, May RM (1992) Infectious diseases of humans: dynamics and control. Oxford University Press, Oxford

Anderson RM, Heesterbeek H, Klinkenberg D, Hollingsworth TD (2020) How will country-based mitigation measures influence the course of the COVID-19 epidemic? Lancet 395(10228):931–934

Arenas A, Cota W, Gomez-Gardenes J, Gómez S, Granell C, Matamalas JT, Soriano-Panos D, Steinegger B (2020) A mathematical model for the spatiotemporal epidemic spreading of COVID19. medRxiv:2020.03.21.20040022

Bauer P, Thorpe A, Brunet G (2015) The quiet revolution of numerical weather prediction. Nature 525(7567):47–55

Bittihn P, Golestanian R (2020) Containment strategy for an epidemic based on fluctuations in the sir model. *preprint* arXiv:2003.08784

Bolker B, Grenfell BT (1995) Space, persistence and dynamics of measles epidemics. Philos Trans R Soc Lond B Biol Sci 348(1325):309–320

Dietz K (1993) The estimation of the basic reproduction number for infectious diseases. Stat Methods Med Res 2(1):23–41

Engbert R, Drepper F (1994) Chance and chaos in population biology-models of recurrent epidemics and food chain dynamics. Chaos Solitons Fract 4(7):1147–1169

Evensen G (2006) Data assimilation. Springer, New York

Gillespie DT (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. J Comput Phys 22(4):403–434

Grenfell BT, Kleczkowski A, Gilligan C, Bolker B (1995) Spatial heterogeneity, nonlinear dynamics and chaos in infectious diseases. Stat Methods Med Res 4(2):160–183

He D, Zhao S, Li Y, Cao P, Gao D, Lou Y, Yang L (2020a) Comparing COVID-19 and the 1918–1919 influenza pandemics in the United Kingdom. Int J Infect Dis 98:67–70

He X, Lau EH, Wu P, Deng X, Wang J, Hao X, Lau YC, Wong JY, Guan Y, Tan X (2020b) Temporal dynamics in viral shedding and transmissibility of COVID-19. Nat Med 26(5):672–675

Kucharski AJ, Russell TW, Diamond C, Liu Y, Edmunds J, Funk S, Eggo RM, Sun F, Jit M, Munday JD et al (2020) Early dynamics of transmission and control of COVID-19: a mathematical modelling study. Lancet Infect Dis 20(5):553–558

Law K, Stuart A, Zygalakis K (2015) Data assimilation: a mathematical introduction. Springer, New York

Li R, Pei S, Chen B, Song Y, Zhang T, Yang W, Shaman J (2020) Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2). Science 368(6490):489–493

Liu Y, Gayle AA, Wilder-Smith A, Rocklöv J (2020) The reproductive number of covid-19 is higher compared to sars coronavirus. J Travel Med, 27

Lourenço J, Paton R, Ghafari M, Kraemer M, Thompson C, Simmonds P, Klenerman P, Gupta S (2020) Fundamental principles of epidemic spread highlight the immediate need for large-scale serological surveys to assess the stage of the sars-cov-2 epidemic. medRxiv:2020.03.24.20042291

Maier BF, Brockmann D (2020) Effective containment explains subexponential growth in recent confirmed COVID-19 cases in China. Science 368(6492):742–746

Reich S, Cotter C (2015) Probabilistic forecasting and Bayesian data assimilation. Cambridge University Press, Cambridge

RKI (2020) Robert Koch Institute (RKI): COVID-19 Data for Germany. https://npgeo-corona-npgeo-de.hub.arcgis.com/. Accessed 08 April 2020

Sakov P, Oke P (2008) A deterministic formulation of the ensemble Kalman filter: an alternative to ensemble square root filters. Tellus 60A:361–371

Schwartz IB, Smith H (1983) Infinite subharmonic bifurcation in an SEIR epidemic model. J Math Biol 18(3):233–253