Under approval at medRxiv (manuscript ID 2020/063768)

Sequential data assimilation of the stochastic SEIR epidemic model for regional COVID-19 dynamics

Ralf Engbert*$^{1-4}$, Maximilian M. Rabe^{4}, Reinhold Kliegl^{2,5}, and Sebastian Reich$^{1-3,6}$
[1] *Research Focus Data-Centric Sciences,* [2] *Research Focus Cognitive Science,* [3] *DFG Collaborative Research Center 1294,*
[4] *Department of Psychology,* [5] *Division of Training and Movement Sciences,* [6] *Institute of Mathematics*
University of Potsdam, Germany

April 14, 2020

Abstract

Newly emerging pandemics like COVID-19 call for better predictive models to implement early and precisely tuned responses to their deep impact on society. Standard epidemic models provide a theoretically well-founded description of dynamics 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 challenges containment strategies, in particular, due to delayed epidemic responses to non-pharmaceutical interventions. However, spatial heterogeneity questions the adequacy of modeling epidemic outbreaks on the level of a whole country. Here we show that sequential data assimilation of a stochastic version of the standard SEIR epidemic model captures dynamical behavior of outbreaks on the regional level. Such regional modeling of epidemics with relatively low numbers of infected and realistic demographic noise accounts for both spatial heterogeneity and stochasticity. Based on adapted regional models, population level short-term predictions can be achieved. More realistic epidemic models that include spatial heterogeneity are within reach via sequential data assimilation methods.

The evolving spread of the novel coronavirus in Germany [1] resulted in containment measures based on reduced traveling and social distancing [3]. In epidemic standard models [2, 14], which provide a dynamical description of epidemic outbreaks [7, 20], 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, estimation of the contact parameter is a key basis of epidemic modeling [17].

The current situation of COVID-19 is characterized by extreme spatial heterogeneity [1]. In the initial phase of the outbreak, this heterogeneity is caused by random travel-based imports of infectious cases and enhanced by local events with increased contacts. As a consequence, the assumption of homogeneous mixing must be relaxed [12] and coupled dynamics of regional models seem to be a more adequate description [15]. However, when modeling a relatively small region with population size $N = 10^5$ compared to the country level with $N = 10^7$ to 10^9, demographic stochasticity [9, 12] must be addressed (see The stochastic SEIR model). The combination of dynamical modeling with substantial fluctuations calls for sequential data assimilation methods for parameter inference [5, 19].

We investigate the stochastic SEIR epidemic model [2] for application to regional data of COVID-19 incidence. The model assumes S, E, I, and R compartments representing susceptible, exposed,

[*] Address correspondence to: ralf.engbert@uni-potsdam.de

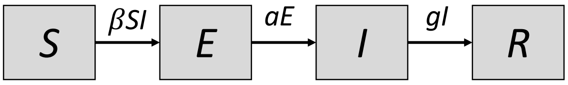

Figure 1: The SEIR model. The population is composed of four compartments that represent susceptible, exposed, infectious, and recovered individuals. The contact parameter β is critical for disease transmission, $1/a$ and $1/g$ are the average duration of exposed and infectious periods, resp. Different from the standard model, birth and death processes are neglected for the short-term simulations discussed throughout the current study.

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 [13], such that models without the exposed compartment cannot adequately address the time-delay between 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. Disease-related model parameters are the rate parameters $a = 1/Z$ (with average latency period Z) and $g = 1/D$ (with mean infectious period D), which can be estimated independently from analysis of infected cases [13, 15]. Therefore, the time-dependent contact parameter β is the most critical parameter that needs to be determined via data assimilation [19]. The contact parameter β is directly related to the basic reproductive rate R_0 in a SEIR-type model (see SEIR model and basic reproductive rate). Therefore, non-pharmaceutical interventions that aim at $R_0 < 1$ translate into the relation $\beta < g$ in the model.

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

Results

The key motivation of the current study was to apply sequential data assimilation of the stochastic SEIR model to estimate the contact parameter. Using simulated data, we successfully applied an ensemble Kalman filter [10, 19] for recovery of the contact parameter from data (see Parameter recovery from simulated data). When applied to empirical data on the level of a region, the estimation of the contact parameter produces a comparable evidence profile (see Application to empirical data).

In the early phase of the outbreak of COVID-19 in Germany, the reported cumulative numbers of cases strongly increase (Fig. 2a,b), however, epidemic dynamics vary on the regional level. Such spatial heterogeneity is due to different onset times of the disease in different regions, but is also enhanced by variations in the local contact parameters β. In response to containment measures, we expect β to change over time.

Time-dependence of the contact parameter

Estimation of the time-dependence of the contact parameter is done 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 (see Model inference based on sequential data assimilation). 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. 2c). The black line reports the average time dependence for all 320 regions included in the analysis; standard deviations are indicated by the grey area. Results for the two example regions are given by their corresponding colors.

The non-pharmaceutical interventions in the spread of COVID-19 were implemented at slightly varying points in time across Germany. In the majority of regions, closings of schools and other educational institutions started on March 16th, while large-scale contact bans was implemented on March 22th. Since 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 any of these measures cannot produce an immediate effect on the observed cases of infected individuals because of the latency period. To use a reliable estimate of the contact parameter, the related interval should be as long as possible, since sequential data assimilation will need several data points to adapt the model to the data. Therefore, we selected the average value of $\beta_*(t_k)$ over the three days from March 17th to March 19th as a pre-intervention value. The average over March 31st to April 2nd 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. A scatter plot indicated a clear reduction of the numerical value of the contact parameter from β_{pre} to β_{post} (Fig. 2d). The reduction is statistical significant (Wilcoxon test, $p < 0.01$).

Finally, the overall time-dependence $\beta_*(t_k)$ shows a decreasing trend, however, Figure 2c suggests that there is an additional weekly cycle with local minima at weekends (March 22 and 29). Both of the example regions show this effect as well. For the RKI data, we do not expect that seemingly reduced contact parameters are a simple consequence of increased reporting delay over the weekend, since this database is continuously updating the reported cases back in time (see RKI data on COVID-19 in Germany).

Simulations with time-varying contact parameter

The contact parameter β is the most critical parameter determining the dynamics of the stochastic SEIR model. After 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. 3). Simulations I are started from the first epidemic day in the corresponding region with greater than or equal to 30 cases. The initial numbers of infected I_0 were set to the observed number of cases $y_{\text{obs}}(t_0)$, while the initial numbers of exposed were set to $E_0 = g/a \cdot I_0$, which would hold at epidemic equilibrium. The initial number of infected people was disturbed by noise representing uncertainties in the initial model states. Forward iteration with the estimated time-varying contact parameter show that the slope of the epidemic curve is approximately reproduced by the model (Fig. 3a,c; grey lines indicate the ensemble of simulated trajectories; blue points are observed data).

At March 26th, we started simulations II which exploits the full potential of sequential data assimilation. The sequential data assimilation approach via the ensemble Kalman filter (see Ensemble Kalman filter) is based on the forward modeling of an ensemble of trajectories. After each time step (1 day), the ensemble of trajectory is compared to the next observation and adjusted via a linear regression step. Therefore, we obtained an adapted ensemble of internal model states at each epidemic day. Here we exploit this fact in 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. 3a,c; grey lines indicate the ensemble of simulated trajectories; blue points are observed data). A related plot of the daily reported new cases indicates approximately constant level of numbers of new cases for Köln (Fig. 3b) and slowly decreasing daily new cases for Münster (Fig. 3d); both predictions are in agreement with empirical observations.

Predictions for two different scenarios

The forward simulations discussed in the previous section demonstrated the predictive power of the SEIR model after 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 4th) and iterated the model forward with the mean contact parameter estimated in the week March 29th to April 4th after implementation of interventions (Fig. 4, green area). The simulations smoothly continue the time-course of infected cases for both example regions (Fig. 4a,b). Daily reported case numbers show a decline for both regions (Fig. 4c,d).

In scenario II, we assumed that all governmental intervention measures were terminated. Therefore, we used the estimated in the week March 15th to March 21st. Again we started simulations with the adapted ensemble of internal states after sequential data assimilation (Fig. 4, red area). For both example regions, we observe a strong increase in infected cases under scenario II (Fig. 4c,d). The

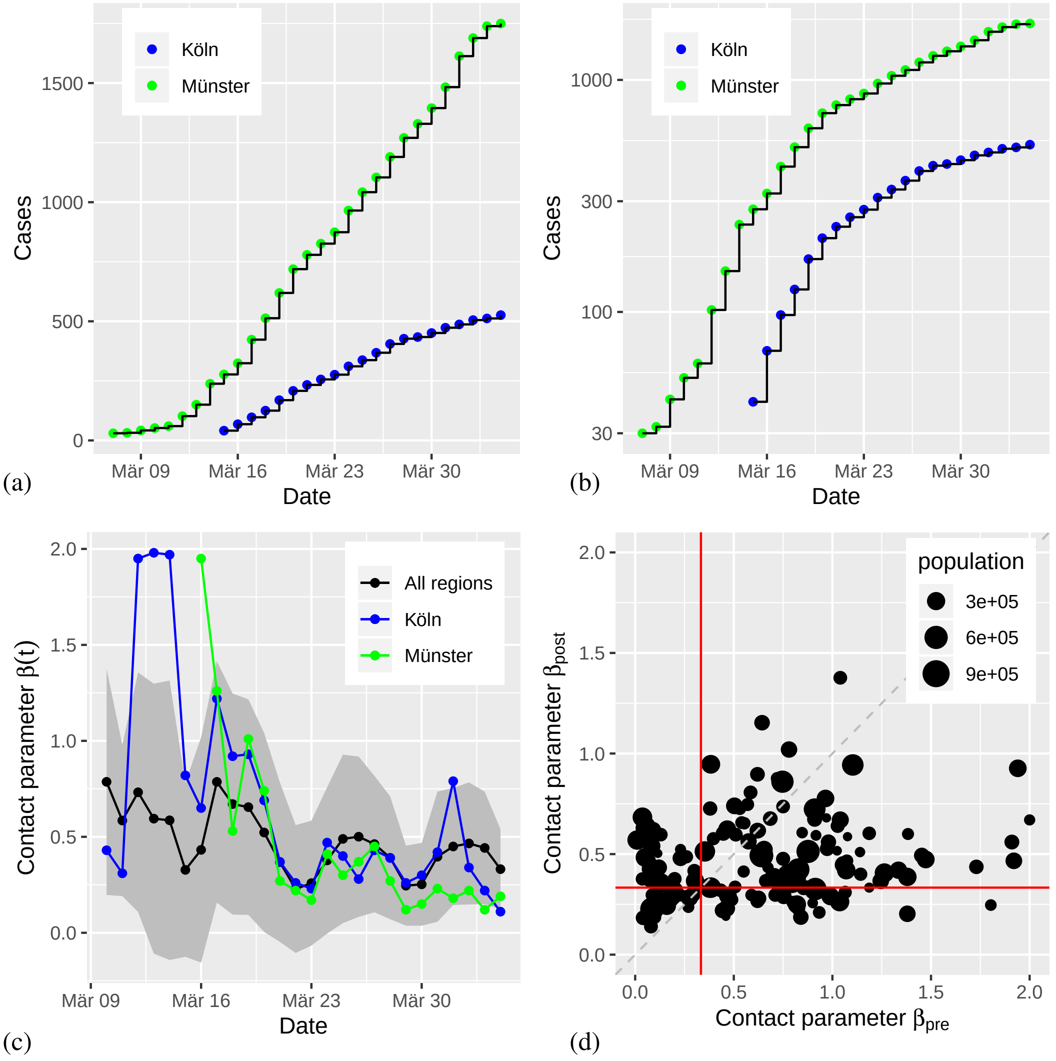

Figure 2: Analysis of time-dependent contact best fit contact parameters $\beta_*(t_k)$. (a) For two regions (LK Köln and LK Münster), the cumulative numbers show strong increase after different disease onset times. (b) Semi-logarithmic scaling suggests approximate exponential growth in early as well 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; grey shading: standard deviation across regions. (c) 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 $\beta_{\text{crit}} = 1/3$ in our model (red lines).

dramatic increase can be seen most clearly in the plot of daily numbers of new cases (for more examples see Supplementary Information Appendix).

Discussion

The ongoing worldwide spread of the new coronavirus exerts enormous pressure on health 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 [18]. The motivation of the current work was to explore the potential of sequential data assimilation [19] for a regional epidemic model as a forecasting tool.

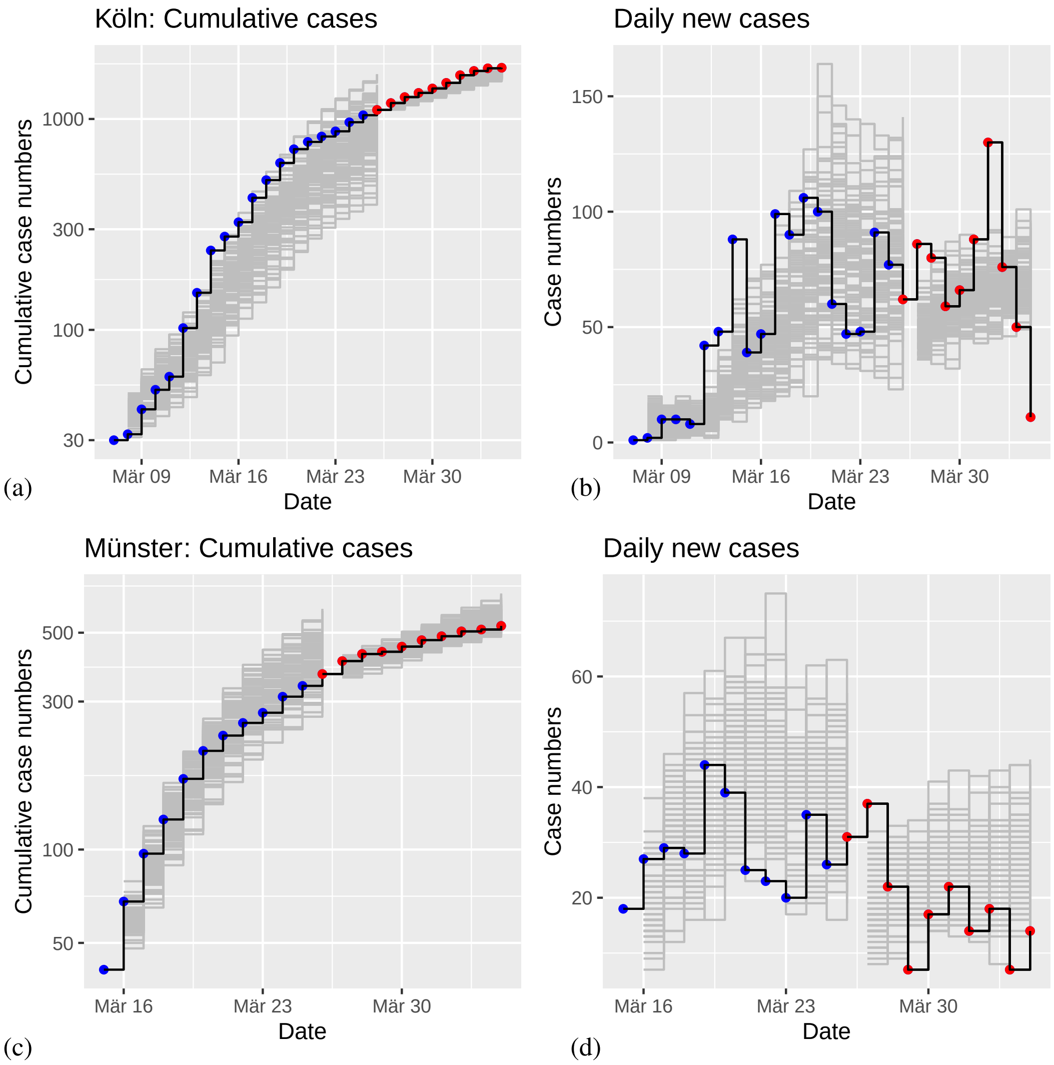

Figure 3: Simulations of the stochastic SEIR model for two example regions. Simulations I indicate an ensemble of 100 runs of the model with initial conditions from the first epidemic day with number of cases greater than or equal to 30 (grey: ensemble of trajectories; blue: observations). Simulations II start on March 26th, using an ensemble size of 100 after data assimilation (grey: 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.

Standard epidemic SEIR-type models implement a compartmental description under the assumption of homogeneous mixing of individuals [2]. More realistic modeling approaches require spatial heterogeneity due to time-varying disease onset times, regionally different contact rates, and the time-dependence of the contact rates due to implementation of containment strategies. However, regional descriptions require models that include effects of demographic stochasticity due to limited population sizes and cases numbers in the region considered [6]. Effects of such statistical fluctuations are inherently reproduced via stochastic versions of the standard epidemic models [9, 12].

We demonstrated the potential of sequential data assimilation for COVID-19 dynamics at the level of a regional, stochastic model. Based an the ensemble Kalman filter [10], we successfully recovered the contact parameter from simulated data and obtained reliable estimates from empirical data. The contact parameter is the most critical free parameter in the stochastic SEIR model, since other

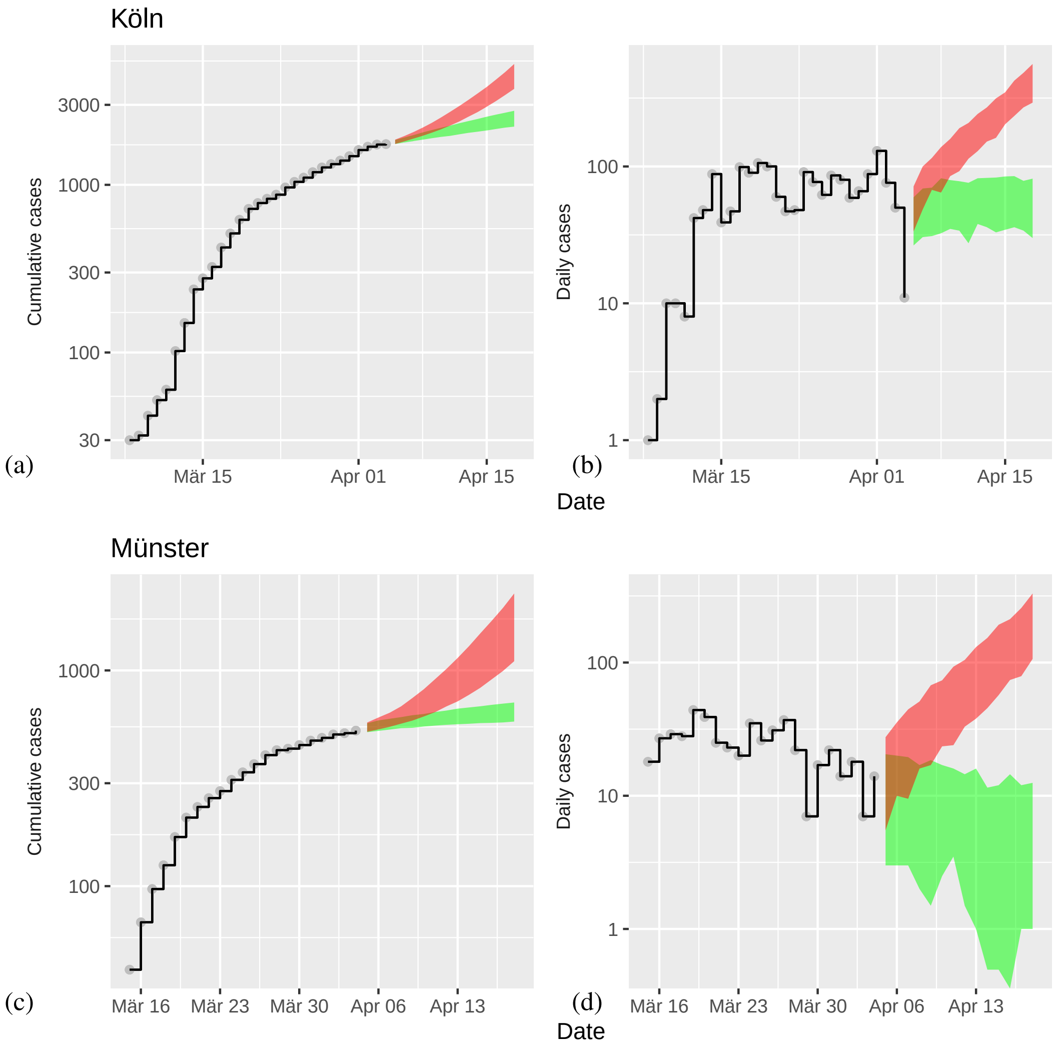

Figure 4: Model predictions for COVID-19 after data assimilation. In scenario I (green area), an assimilated ensemble of internal model states starts the forecast with contact parameter β_{post} (continuation of social distancing interventions). In scenario II (red area), the equivalent forcase is generated with contact parameter β_{pre} (termination of interventions). (a) Predictions for cumulative case numbers in Köln. (b) Predictions of daily new cases in Köln. (c) Predictions of cumulative cases for Münster. (d) Predictions of daily new cases for Münster.

parameters (mean exposed and infectious duration) can be estimated independently from observed time series [13, 15]. Moreover, the contact parameter of the SEIR model is directly related to the basic reproductive rate R_0 [16]. Therefore, our approach could also be framed as model-based method for statistical inference of the basic reproductive rate.

Next, we ran time-resolved analyses that generated estimates of the time-dependence of the contact parameter. The drop in mean contact rates from an early (β_{pre}, March 17-19) to a later period (β_{post}, March 31-April 2) indicated the effect of non-pharmaceutical interventions. We also generated model prediction under two different scenarios. In scenario I, started simulations from April 4th with sampling from the assimilated ensemble as initials conditions and the contact parameter estimated for the post-intervention period (March 29-April 4). In scenario II, we replaced the post-intervention contact parameter by its pre-intervention value (March 15-21). As a results, 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 off the current interventions would clear switch the epidemic dynamics to the exponential increase before implementation of non-pharmaceutical interventions. 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 because of explicit modeling of spatial and temporal heterogeneities, captured by a separate time-course of the contact parameter for each region.

The recent simulation study Li et al. [15] used a similar approach of sequential data assimilation for dynamic epidemic models. However, the deterministic SEIR model was implemented and extended by additional noise assumptions. We proposed the usage of the stochastic SEIR model in the formulation of a master equation [9] which can be simulated exactly and numerically efficiently using Gillespie's algorithm [11]. A more complex spatiotemporal stochastic model has been considered in [4].

Furthermore, the state-parameter estimation in [15] utilises the ensemble Kalman filter directly on an augmented state space [19]. 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 the two stage approach, as presented in this study, combining the ensemble Kalman filter for state estimation with a likelihood based inference of the contact parameter β.

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 in our current modeling framework, which we did not address because of the methodological focus. Longer-term predictions ($\sim$ months) are important, but clearly dependent critically on the estimation of undocumented infections (see Li et al. [15]). Such hidden infections create, after recovery, an unknown reduction in the number of susceptibles that slows down 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 [15].

Another important limitation of these results comes from the simplifying assumption that there is no coupling to neighboring regions. As a consequence, regional differences in the contact parameter might in fact be due to differences in contacts between regions. Introducing couplings between regions [15] could also be integrated in our modeling framework. However, the no-coupling approximation might be realistic in the current situation of social distancing and contact ban.

Linking back to the potential of NPI to control COVID19 until vaccination or medicine is available we close with an observation relating to a recent study [8]. Our analyses covered the same time span as Dehning et al. [8] and we also picked up a systematic pattern of contact reduction which we interpret as a weekly cycle (see Statistical modeling of trends and weekly oscillations). Indeed, the dates of the interventions analyzed by Dehning et al. are confounded with day of the week; the three dates refer to three Sundays in March. It is reasonable to assume that in Germany there is much less contact to persons outside the family context on weekends than during working days. Of course, there are also other reasons why certified reports of COVID19 infections are less frequent for weekends. For example, people are more likely to go to their GP on Monday than to the emergency room on Sunday, especially as long as the symptoms are mild. No reason, in our opinion, is a lack of recording at the RKI; the cases reported for the weekend are added on Monday and Tuesday. One advantage of the method proposed here is that we recover effects with comparatively little data at the level of regions, not an entire country. In the absence of a vaccine or of medication, having such an epidemic forcasting tool seems almost like a necessary precondition for selective and optimal timing of tightening and loosening of NPI-based containment measures.

Materials and Methods

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 by 412 counties, six age groups, and sex. As they are official records, only cases certified by doctors or labs 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 available on the website. Statistics relating to the most recent three or four days are incomplete and cannot be interpreted; retrospective updates and corrections are possible for all days of the pandemic spell as they become available. We use data inclusive April 4, as reported on April 8, 2020; they are included as part of the supplement.

SEIR model and basic reproductive rate

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

$$\begin{aligned} \dot{S} &= m-(m+\lambda)S && (1)\\ \dot{E} &= \lambda S-(m+a)E && (2)\\ \dot{I} &= aE-(m+g)I && (3)\\ \dot{R} &= gI-mR\,, && (4) \end{aligned}$$

where the total number of individual $N = S + E + I + R$ is constant under temporal evolution due to $\dot{N} = 0$. The ODE system, Eq. (1-4), has a non-trivial equilibrium point, denoted as epidemic equilibrium (S_0, E_0, I_0, R_0), where the number of susceptibles S_0 at equilibrium is related to basic reproductive rate. 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

$$R_0 = \frac{1}{S_0} = \frac{a\beta}{(m+a)(m+g)} \to \frac{\beta}{g} \quad \text{for} \quad m \to 0\,. \qquad (5)$$

We use a numerical values of $g = 1/3$, equivalent to an average infectious period of $D = 3$ days, and $a = 0.192$, or an average latency period $Z = 5.2$ [13]. As a consequence, the critical condition for disease containment $R_0 < 1$ is obtained for $\beta < \beta_{\text{crit}} = 1/3$ in our model.

The stochastic SEIR model

While the classical model is formulated as a system of ordinary differential equations, we are exploring the application to relatively low numbers 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 [9], which is particularly useful for modeling small numbers of infected individuals occuring in smaller regions or in the beginning of epidemics.

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 constant population size $N = S + E + I + R$. The transition rate of the ODE compartmental model translate into transition probabilities in the master equation formulation for the evolution of the model's conditional probability density, that is,

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

with transition probabilities given in Table 1 and initial condition X_0. Single trajectories for the model's temporal evolution can be simulated exactly and numerically efficiently [9] using Gillespie's algorithm [11].

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

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

Model inference based on sequential data assimilation

Publicly available data on the cumulative number of infected individuals is 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 [19]. 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 adjusted model states $X_{\mathrm{a}}^{(n)}(t_k)$. This step is implemented via the ensemble Kalman filter [19]. 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 over a fixed range of contact parameters $\beta \in [\beta_{\min}, \beta_{\max}]$ and over a fixed time window $[t_{\mathrm{initial}}, t_{\mathrm{final}}]$ of available data points $y_{\mathrm{obs}}(t_k)$. The best fit contact parameter $\beta_*(t_k)$ at any time any t_k is found as the one that minimises 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.

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

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

i.e., $H = (0, 0, 1, 1)$. As initial condition, we set I_1 as the number of infected cases, $R_1 = 0$, so that $y_{\mathrm{obs}}(t_0) = I_1 + R_1$, and $E_1 = g/a \cdot I_1$ with additive noise. We assume that the errors in the observed $y_{\mathrm{obs}}(t_k)$ is additive Gaussian with mean zero and variance ρ. We set $\rho = 10$ in our experiments. The analysis step of the ensemble Kalman filter is now based on the empirical mean

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

and the empirical covariance matrix

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

of the forecast ensemble. These two quantities are used to quantify the forecast uncertainty. Combining the forecast uncertainty with the assumed data error model leads to the linear regression formula

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

with the Kalman gain defined by

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

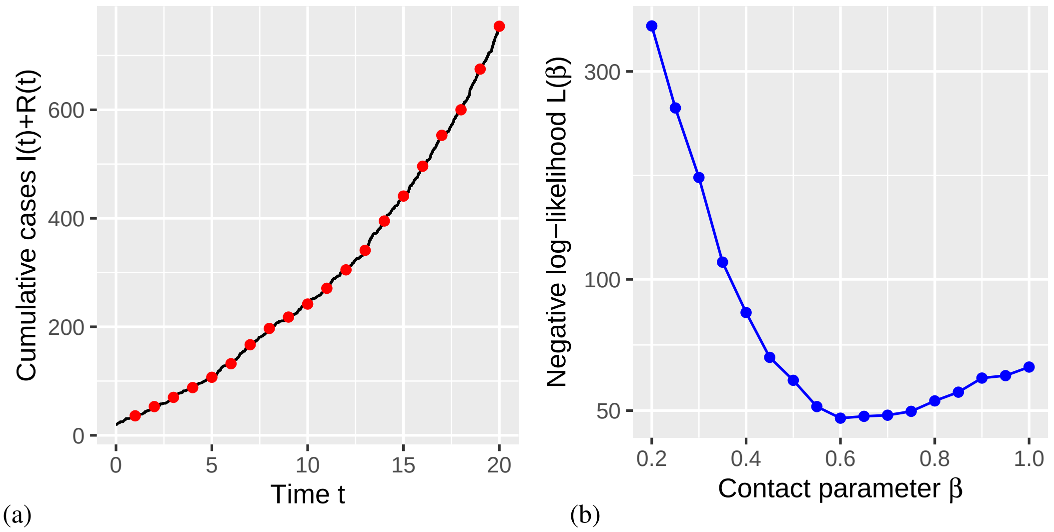

Figure 5: Parameter recovery analysis. (a) Simulated data with $b = 0.6$. (b) Negative log-likelihood $L_{\text{cum}}(\beta)$ indicates a minimum at about the true parameter value.

The resulting analysis ensemble $X_{\text{a}}^{(n)}(t_k) \in \mathbb{R}^4$, $n \in \{1, \dots, N\}$, can be mapped back onto the integers $\mathbb{N}^4$ if needed.

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

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

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

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

that is,

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

Parameter recovery from simulated data To test the inference scheme, we simulated data for 20 days. In Figure 6, the black line indicates the evolution of the SEIR model's predicted cumulative numbers of infected individuals, $Y(t_k) = HX(t_k) = I(t_k) + R(t_k)$. Red dots represent the daily number of reported cases as in real data. In the simulation, the contact rate was chosen as $\beta_{\text{true}} = 0.6$. 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 $L_{\text{cum}}(\beta)$, Eq. (14). The position of the minimum of $L_{\text{cum}}(\beta)$ 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, $\beta_* \approx \beta_{\text{true}} = 0.6$ (Fig. 6b). Thus, parameter recovery can be demonstrated for a relatively short time series of 10 observations, which represents a typical data-set in the early phase of newly emerging epidemics. Next, we apply our inference scheme to real data.

Application to empirical data Since parameter inference was successful for 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

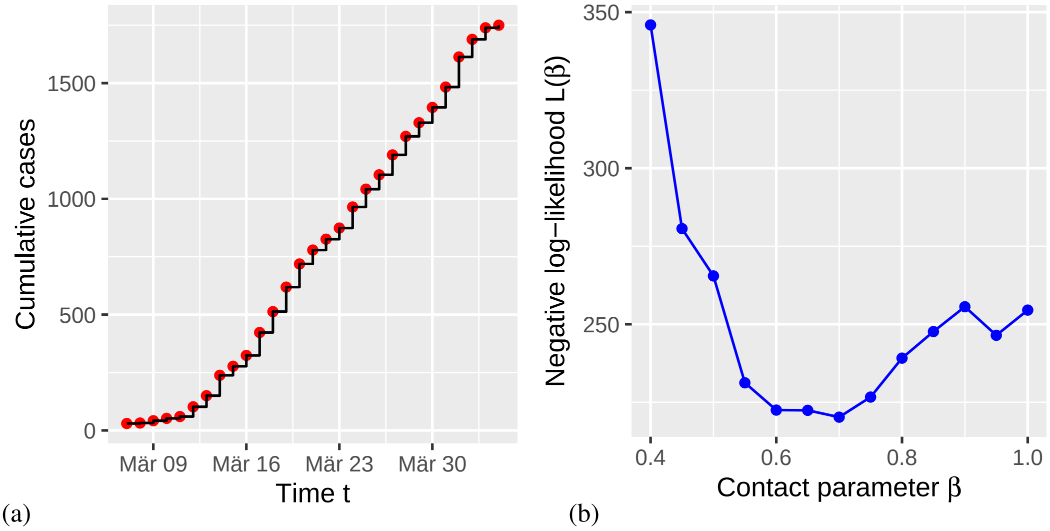

Figure 6: Contact parameter estimates for real data. (a) Data for Köln. (b) Negative log-likelihood $L_{\text{total}}(\beta)$ for Köln give a minimum at $\beta_* \approx 0.7$.

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 Figure 7. Parameter estimation yields an estimate for the contact rate of $\beta_* \approx 0.7$ (Fig. 7b). Thus, analysis of the negative log-likelihood function produced qualitatively similar results for simulated SEIR time series and empirical data for a representative region. In the main text, we carry out an estimation of the time-resolved instantaneous optimal parameter values $\beta_*(t_k)$ 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 emphasise 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.

Statistical modeling of trends and weekly oscillations Mean certified cases, computed across regions per day, reveal a strong daily profile with local minima always falling on Sundays. The corresponding mean of log contact parameters reveals the same oscillation. The negative slopes are compatible with the expectation that a decrease in contact rates causes a decrease in daily cases.

These illustrative analyses imply two considerations. If evaluated against statistics of daily cases, obviously containment measures must take the weekly cycle into account and avoid confounds in the design. There are many potential sources for this cycle, some of them possibly quite trivial (e.g., the number of tests carried out). However, if experimental research and statistical modeling can establish that part of these fluctuations are indeed due to reduced contact rates beyond the family context on weekends, then dynamical models may help with the prediction of the timing of tightening and loosening decisions in local contexts. For example, one may consider moving to a three- or four-workday week for some time, in well-defined contexts, and in targeted regions to facilitate this dynamic.

Acknowledgments

This work was supported by a grant from Deutsche Forschungsgemeinschaft, Germany (SFB 1294, project no. 318763901).

Data and Code Availability

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

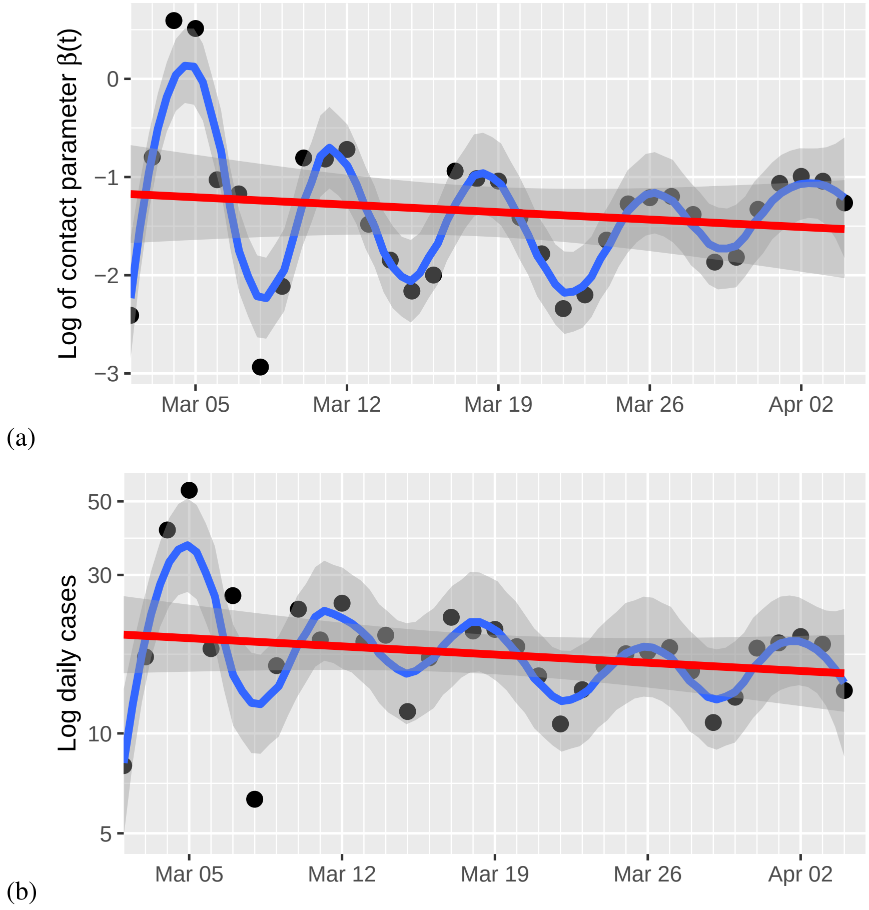

Figure 7: Weekly cycles in contact parameter estimates and daily reported cases. (a) Log contact parameter estimates; (b) Confirmed daily cases. Black: means across regions; red: simple regression of means on date; blue: smoothing with span 0.28 days;grey bands are 95% confidence intervals. RKI data from 10 April 2020.

References

[1] Robert Koch Institute (RKI): COVID-19 Data for Germany. https://npgeo-corona-npgeo-de.hub.arcgis.com/, accessed: 2020-04-08.

[2] Roy M Anderson, B Anderson, and Robert M May. *Infectious diseases of humans: dynamics and control*. Oxford University Press, 1992.

[3] Roy M Anderson, Hans Heesterbeek, Don Klinkenberg, and T Déirdre Hollingsworth. How will country-based mitigation measures influence the course of the COVID-19 epidemic? *The Lancet*, 395(10228):931–934, 2020.

[4] Alex Arenas, Wesley Cota, Jesus Gomez-Gardenes, Sergio Gómez, Clara Granell, Joan T Matamalas, David Soriano-Panos, and Benjamin Steinegger. A mathematical model for the spatiotemporal epidemic spreading of COVID19. *medRxiv:2020.03.21.20040022*, 2020.

[5] Peter Bauer, Alan Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55, 2015.

[6] Philip Bittihn and Ramin Golestanian. Containment strategy for an epidemic based on fluctuations in the sir model. *preprint arXiv:2003.08784*, 2020.

[7] Benjamin Bolker and Bryan Thomas Grenfell. Space, persistence and dynamics of measles epidemics. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 348(1325):309–320, 1995.

[8] Jonas Dehning, Johannes Zierenberg, F Paul Spitzner, Michael Wibral, Joao Pinheiro Neto, Michael Wilczek, and Viola Priesemann. Inferring covid-19 spreading rates and potential change points for case number forecasts. *preprint arXiv:2004.01105*, 2020.

[9] R Engbert and FR Drepper. Chance and chaos in population biology-models of recurrent epidemics and food chain dynamics. *Chaos, Solitons & Fractals*, 4(7):1147–1169, 1994.

[10] Geir Evensen. *Data Assimilation. The Ensemble Kalman Filter*. Springer-Verlag, New York, 2006.

[11] Daniel T Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434, 1976.

[12] Bryan T Grenfell, A Kleczkowski, CA Gilligan, and BM Bolker. Spatial heterogeneity, nonlinear dynamics and chaos in infectious diseases. *Statistical Methods in Medical Research*, 4(2):160–183, 1995.

[13] Xi He, Eric HY Lau, Peng Wu, Xilong Deng, Jian Wang, Xinxin Hao, Yiu Chung Lau, Jessica Y Wong, Yujuan Guan, Xinghua Tan, et al. Temporal dynamics in viral shedding and transmissibility of covid-19. *preprint medRxiv:2020.03.15.20036707*, 2020.

[14] Adam J Kucharski, Timothy W Russell, Charlie Diamond, Yang Liu, John Edmunds, Sebastian Funk, Rosalind M Eggo, Fiona Sun, Mark Jit, James D Munday, et al. Early dynamics of transmission and control of covid-19: a mathematical modelling study. *The Lancet Infectious Diseases*, 2020.

[15] Ruiyun Li, Sen Pei, Bin Chen, Yimeng Song, Tao Zhang, Wan Yang, and Jeffrey Shaman. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2). *Science*, 2020.

[16] Ying Liu, Albert A Gayle, Annelies Wilder-Smith, and Joacim Rocklöv. The reproductive number of covid-19 is higher compared to sars coronavirus. *Journal of Travel Medicine*, 2020.

[17] José Lourenço, Robert Paton, Mahan Ghafari, Moritz Kraemer, Craig Thompson, Peter Simmonds, Paul Klenerman, and Sunetra Gupta. 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*, 2020.

[18] Benjamin F Maier and Dirk Brockmann. Effective containment explains sub-exponential growth in confirmed cases of recent covid-19 outbreak in mainland china. *preprint arXiv:2002.07572*, 2020.

[19] Sebastian Reich and Colin Cotter. *Probabilistic forecasting and Bayesian data assimilation*. Cambridge University Press, 2015.

[20] Ira B Schwartz and HL Smith. Infinite subharmonic bifurcation in an SEIR epidemic model. *Journal of Mathematical Biology*, 18(3):233–253, 1983.

Supplementary Information Appendix

This PDF file includes additional examples for regional modeling.

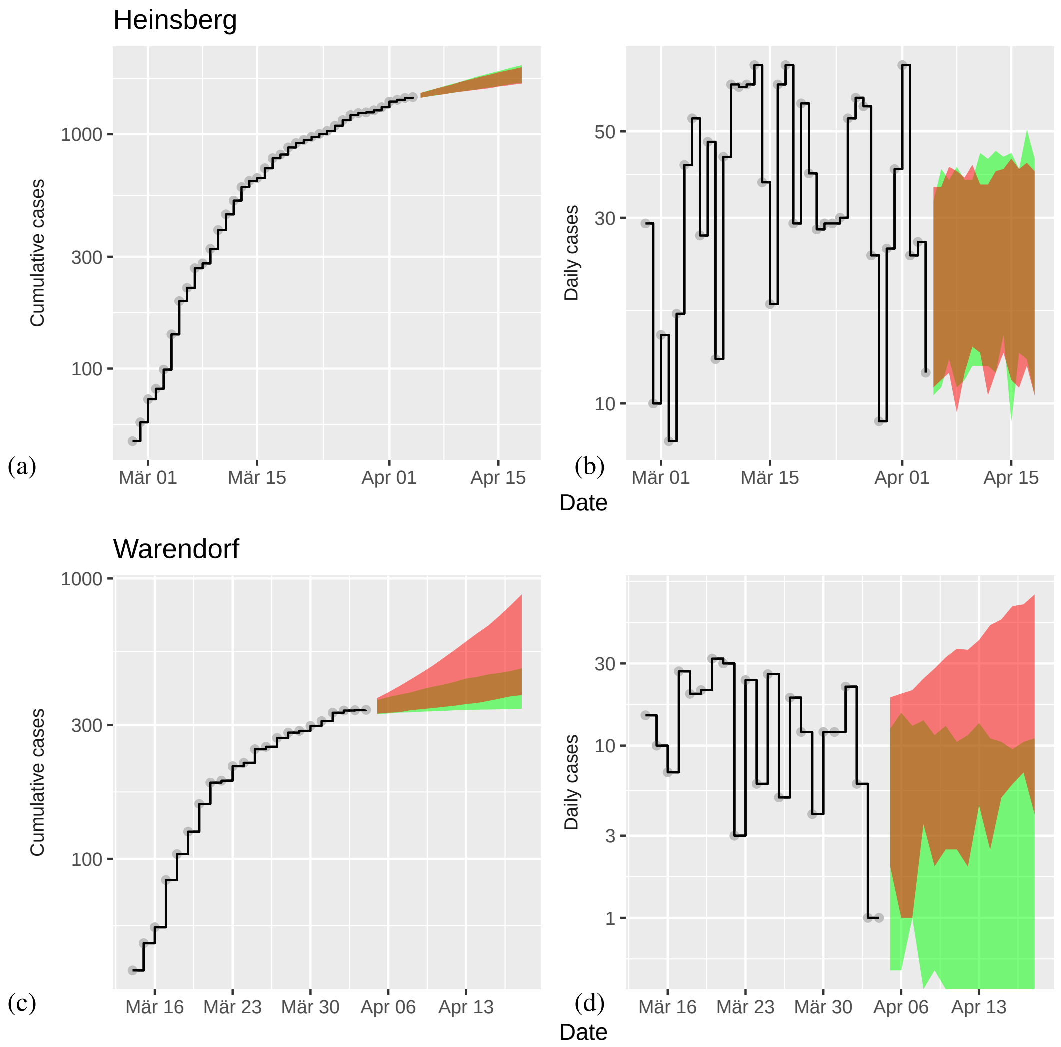

Figure S1: Model predictions for COVID-19 after data assimilation. For details of modeling scenarios I and II see main text and Figure 4.

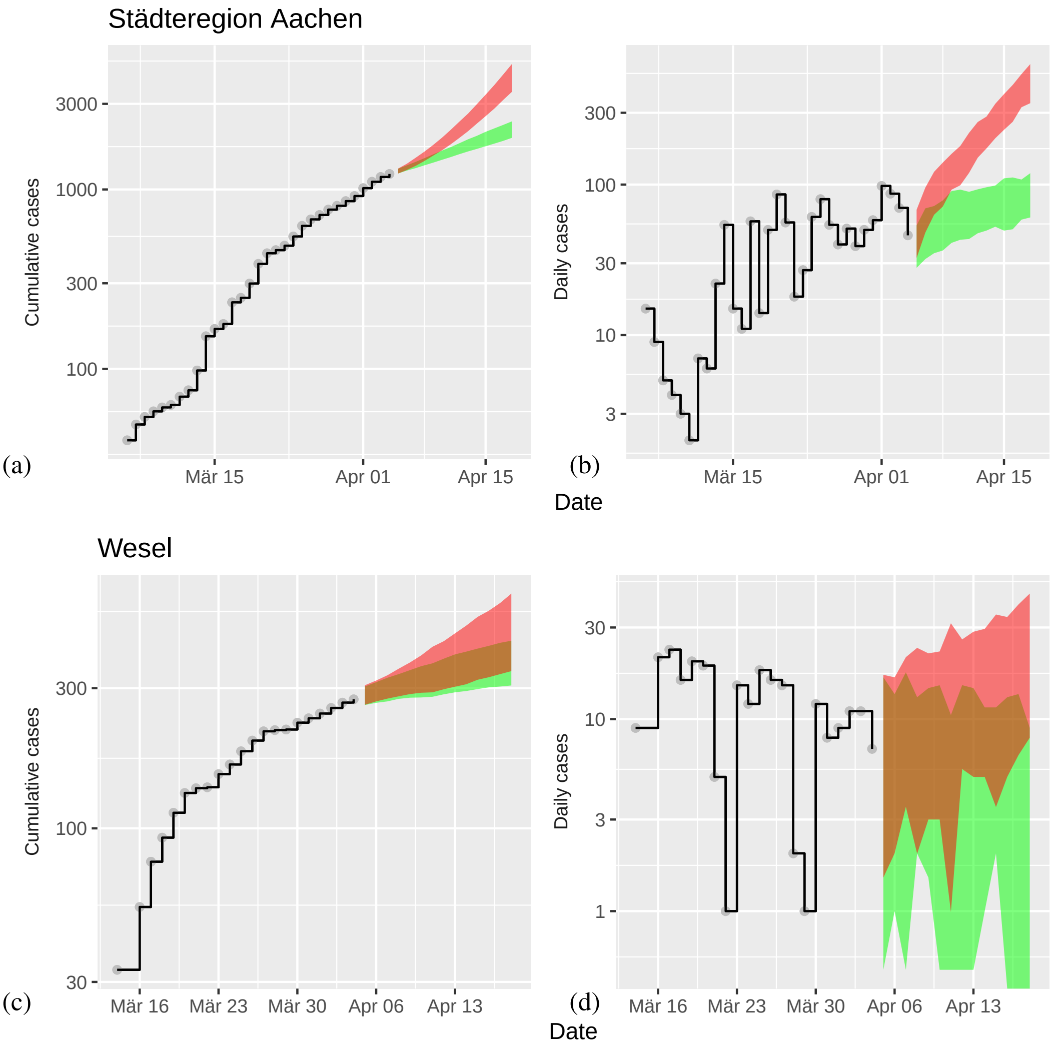

Figure S2: Model predictions for COVID-19 after data assimilation. For details of modeling scenarios I and II see main text and Figure 4.

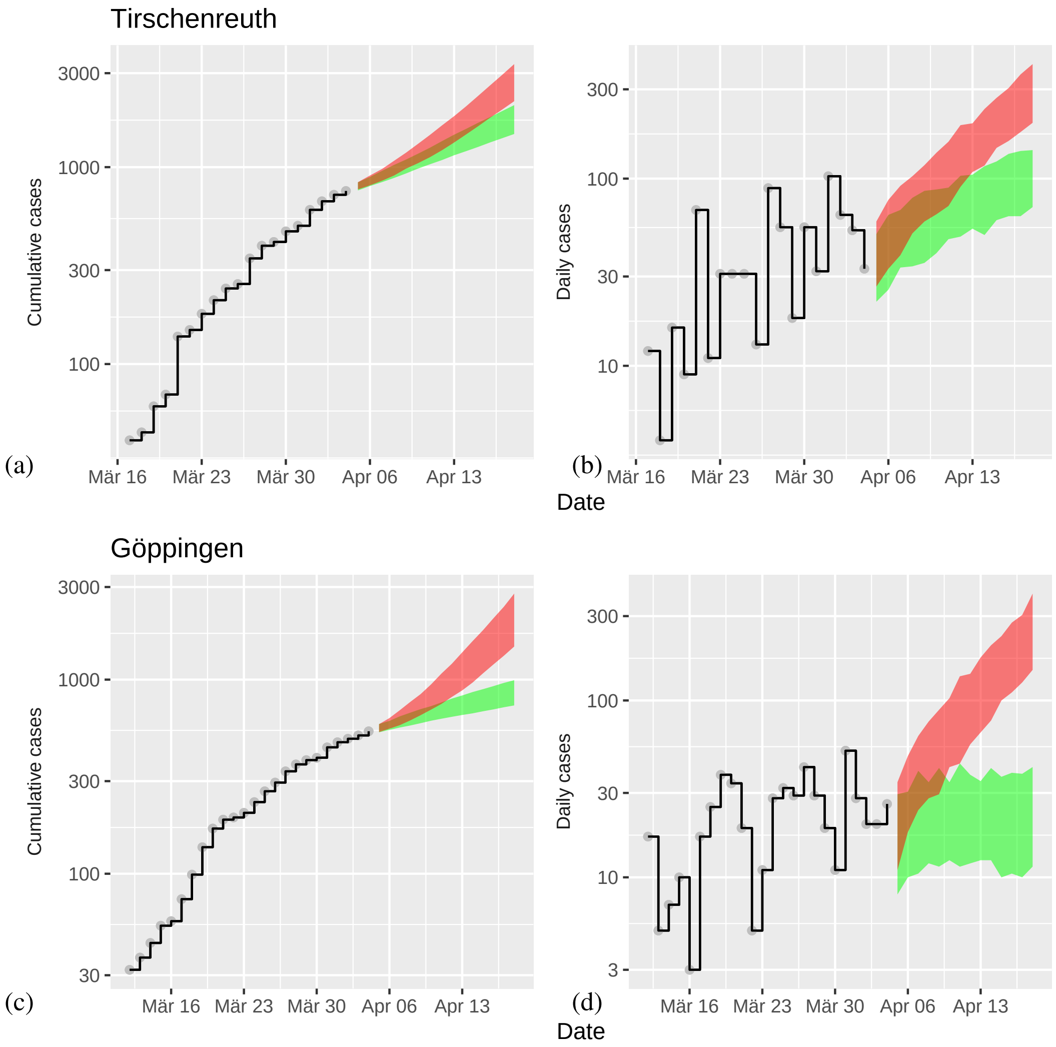

Figure S3: Model predictions for COVID-19 after data assimilation. For details of modeling scenarios I and II see main text and Figure 4.

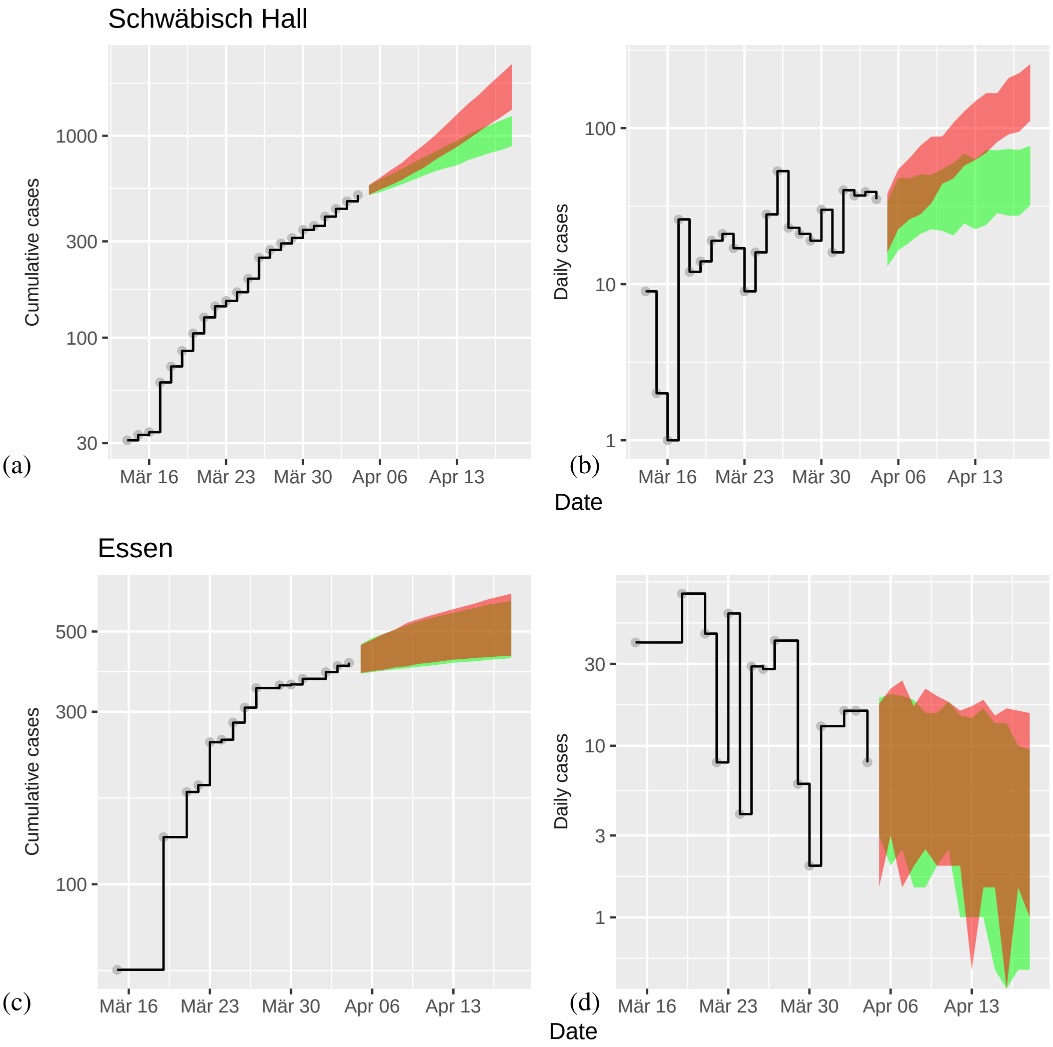

Figure S4: Model predictions for COVID-19 after data assimilation. For details of modeling scenarios I and II see main text and Figure 4.

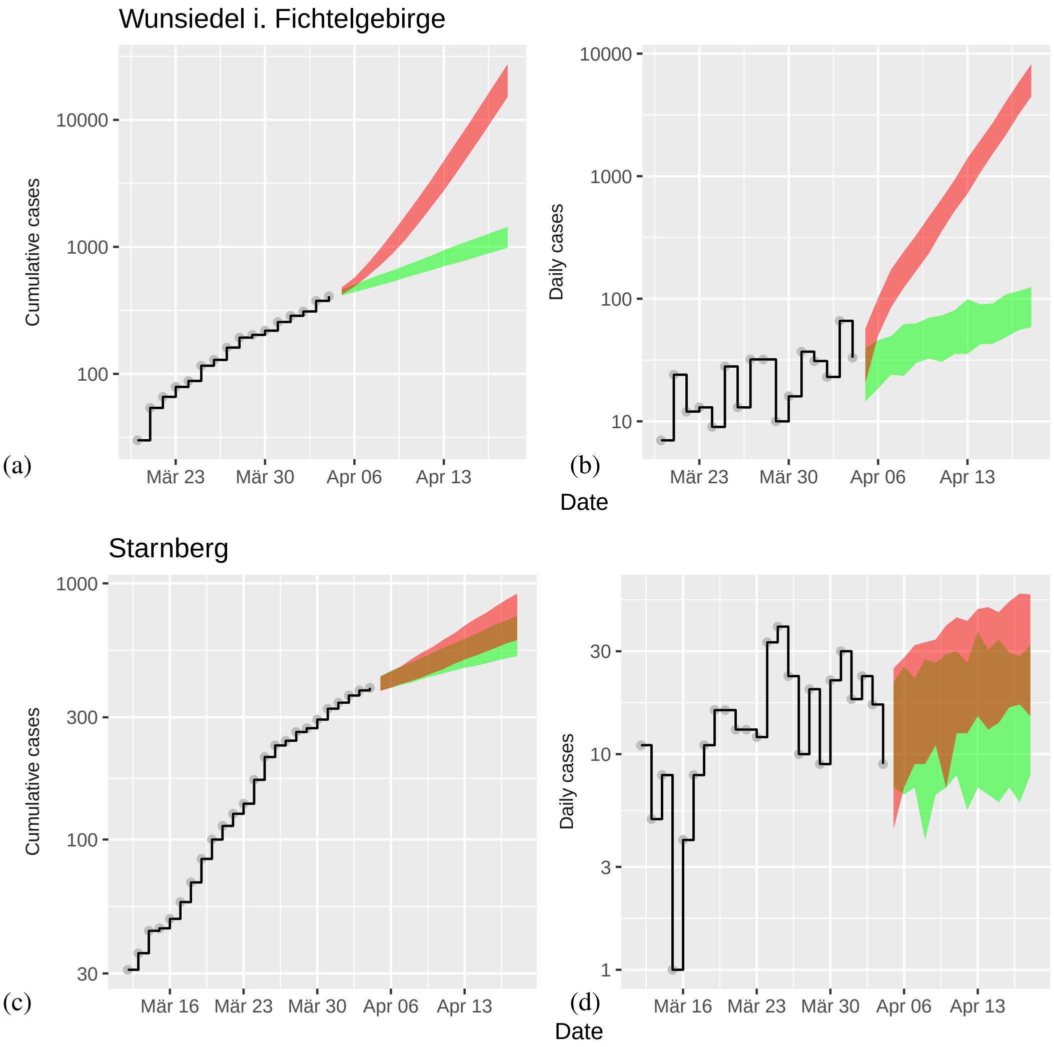

Figure S5: Model predictions for COVID-19 after data assimilation. For details of modeling scenarios I and II see main text and Figure 4.

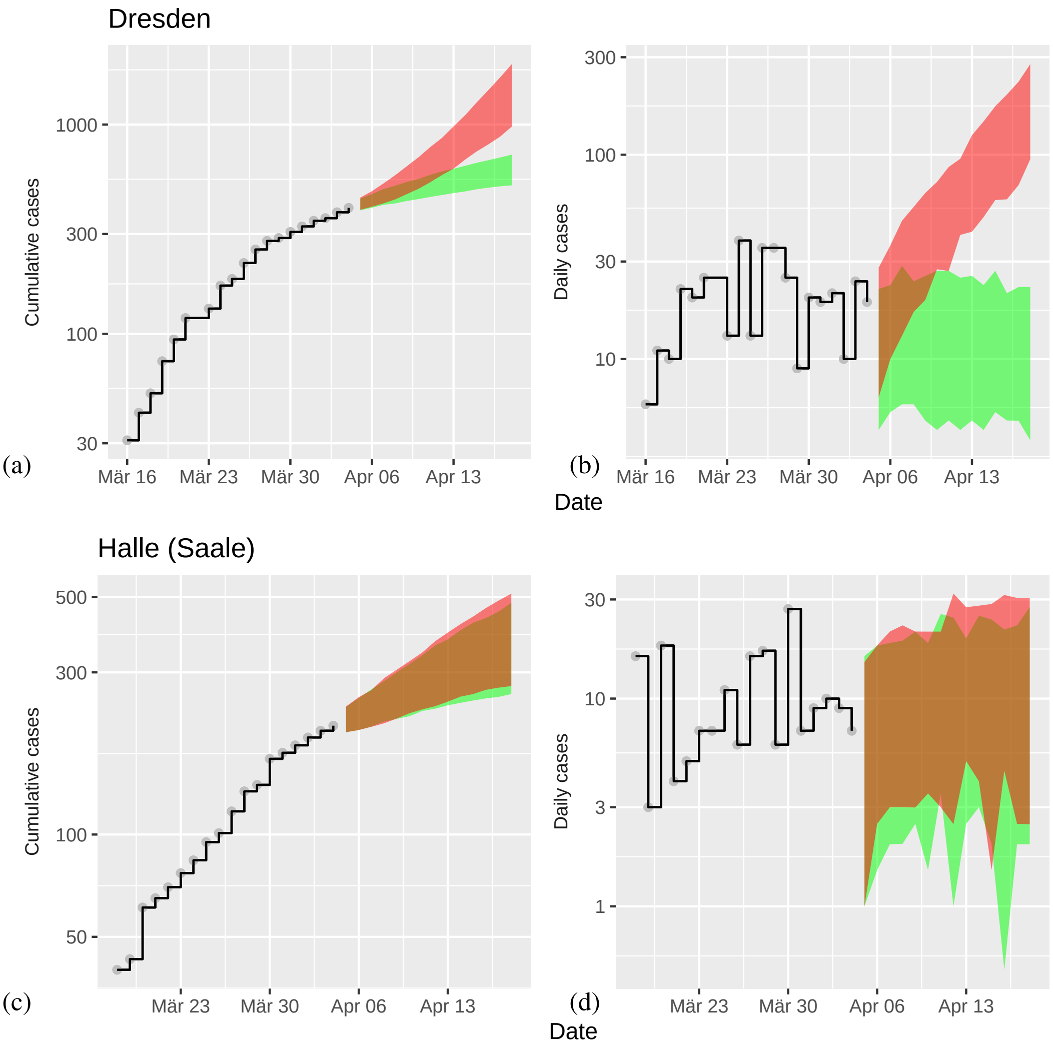

Figure S6: Model predictions for COVID-19 after data assimilation. For details of modeling scenarios I and II see main text and Figure 4.

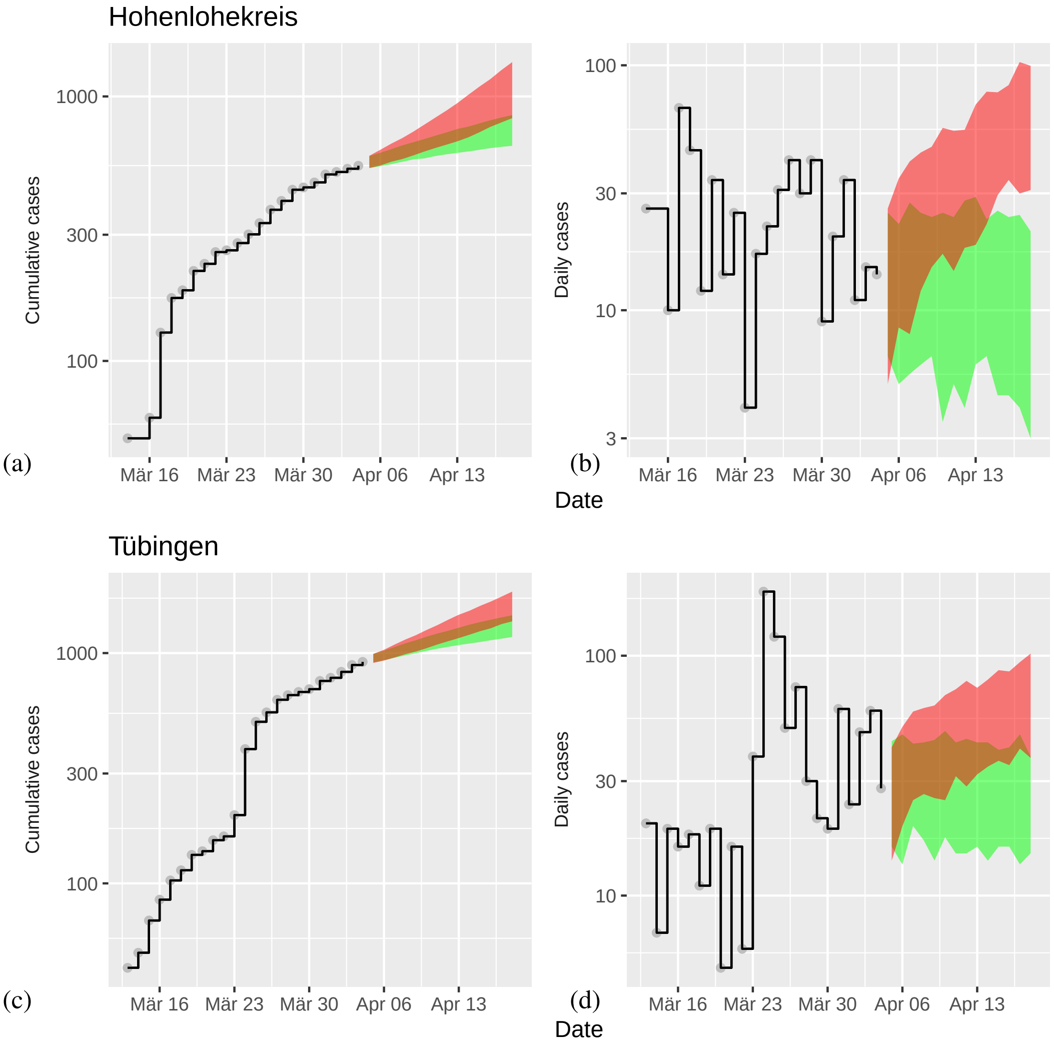

Figure S7: Model predictions for COVID-19 after data assimilation. For details of modeling scenarios I and II see main text and Figure 4.

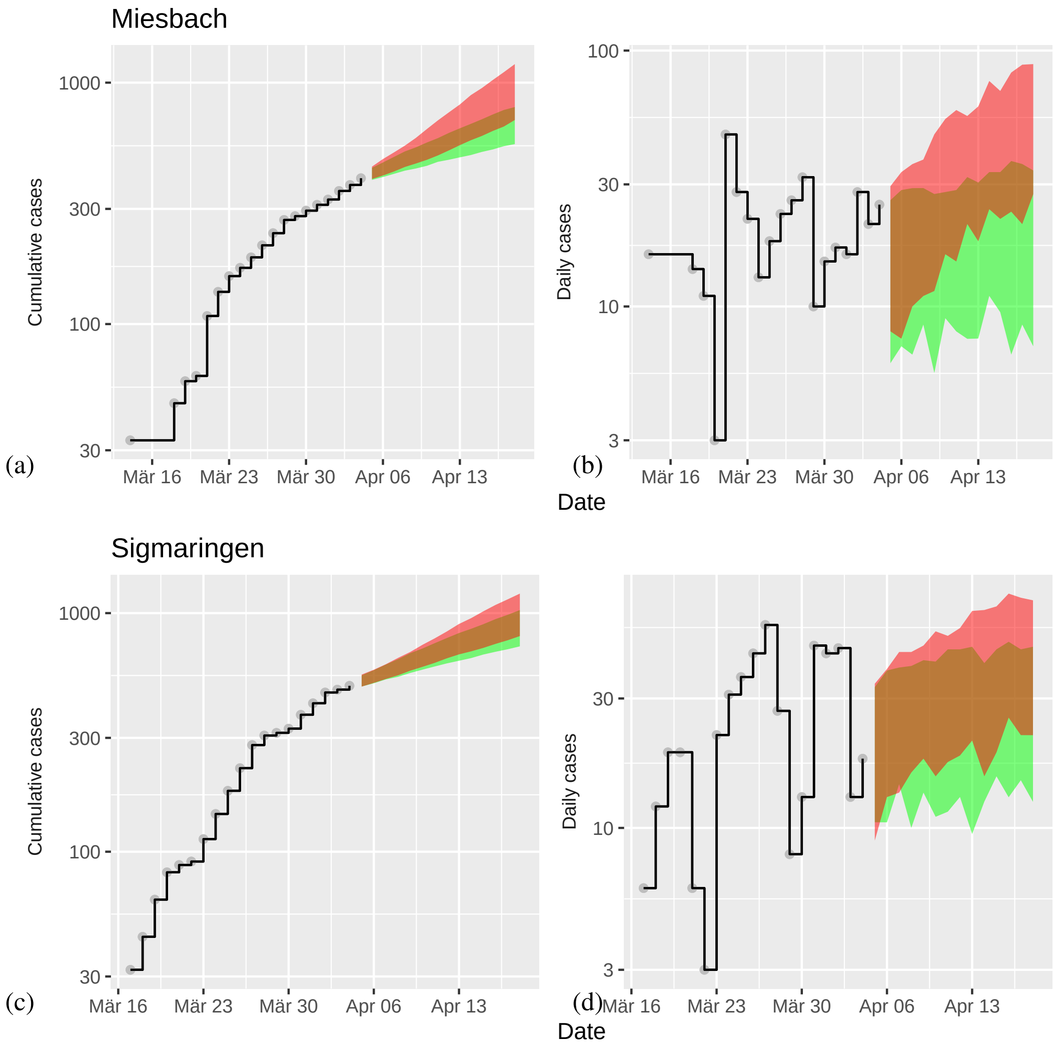

Figure S8: Model predictions for COVID-19 after data assimilation. For details of modeling scenarios I and II see main text and Figure 4.