An international initiative of predicting the Sars-Cov-2 pandemic using ensemble data assimilation

Foundations of Data Science
©American Institute of Mathematical Sciences
Volume **3**, Number **3**, September **2021**
doi:10.3934/fods.2021001
pp. **413–477**

AN INTERNATIONAL INITIATIVE OF PREDICTING THE SARS-COV-2 PANDEMIC USING ENSEMBLE DATA ASSIMILATION

Geir Evensen*,[1], Javier Amezcua[2], Marc Bocquet[3], Alberto Carrassi[2,4], Alban Farchi[3], Alison Fowler[2], Pieter L. Houtekamer[5], Christopher K. Jones[6], Rafael J. de Moraes[7] Manuel Pulido[8], Christian Sampson[6], and Femke C. Vossepoel[7]

[1]NORCE and NERSC
Bergen, Norway

[2]Dept. of Meteorology
University of Reading and NCEO, UK

[3]CEREA, joint laboratory École des Ponts ParisTech and EDF R&D
Université Paris-Est, Champs-sur-Marne, France

[4]Mathematical Institute
University of Utrecht, Netherlands

[5]Environment and Climate Change Canada
Dorval, Québec, Canada

[6]Renaissance Computing Institute
University of North Carolina, Chapel Hill, USA

[7]Department of Geoscience and Engineering
Delft University of Technology, Delft, Netherlands

[8]FaCENA, UNNE and IMIT, CONICET
Corrientes, Argentina

Abstract. This work demonstrates the efficiency of using iterative ensemble smoothers to estimate the parameters of an SEIR model. We have extended a standard SEIR model with age-classes and compartments of sick, hospitalized, and dead. The data conditioned on are the daily numbers of accumulated deaths and the number of hospitalized. Also, it is possible to condition the model on the number of cases obtained from testing. We start from a wide prior distribution for the model parameters; then, the ensemble conditioning

2020 *Mathematics Subject Classification.* Primary: 65K10, 65K99.
Key words and phrases. SARS-CoV-2, ensemble data assimilation, ESMDA, parameter estimation, model calibration.
The first author is supported by NORCE.
*Corresponding author: Geir Evensen.

leads to a posterior ensemble of estimated parameters yielding model predictions in close agreement with the observations. The updated ensemble of model simulations has predictive capabilities and include uncertainty estimates. In particular, we estimate the effective reproductive number as a function of time, and we can assess the impact of different intervention measures. By starting from the updated set of model parameters, we can make accurate short-term predictions of the epidemic development assuming knowledge of the future effective reproductive number. Also, the model system allows for the computation of long-term scenarios of the epidemic under different assumptions. We have applied the model system on data sets from several countries, i.e., the four European countries Norway, England, The Netherlands, and France; the province of Quebec in Canada; the South American countries Argentina and Brazil; and the four US states Alabama, North Carolina, California, and New York. These countries and states all have vastly different developments of the epidemic, and we could accurately model the SARS-CoV-2 outbreak in all of them. We realize that more complex models, e.g., with regional compartments, may be desirable, and we suggest that the approach used here should be applicable also for these models.

1. **Introduction.** We have developed a methodology and software to study the evolution of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic in a country or region and used it to perform an international comparative study across six countries, Argentina, Brazil, England, France, Norway, The Netherlands, four states (New York, California, Alabama, and North Carolina) of the USA, and the province Québec of Canada.

Their diverse geographical locations, including seasonal phase opposition, demography, population densities, and social habits, led the epidemic to evolve differently and impact their very different healthcare systems differently. Moreover, the implemented counter-measures have been mixed in rigor, timing concerning the epidemic status, and effectiveness. Matters complicate further by the very diverse data collection protocols, data quality, and degree of accessibility. Accomplishing a comprehensive analysis of such a complicated situation to infer particular aspects of the SARS-CoV-2 pandemic and predicting its course requires a joint modeling and data analysis approach and a unified protocol. The scope of this work regards the development of such a shared system. It uses a state-of-art nonlinear ensemble data-assimilation method, the ensemble smoother with multiple data assimilation (ESMDA, [21]), typically used in geosciences or in petroleum reservoir modeling, to study the SARS-CoV-2 pandemic.

Our system combines existing data with a modular SEIR (Susceptible, Exposed, Infectious, Recovered) model with age classes and additional compartments for sick but quarantined, hospitalized, and dead. We have integrated the SEIR model into an ensemble-based data-assimilation framework where the setup resembles methods commonly used for parameter estimation in petroleum reservoir models. The system updates the model state and calibrates its parameters to fit a time series of indirect and noisy observations of deaths, hospitalization, and infected people. The current setup constitutes an efficient tool for real-time monitoring and prediction of the SARS-CoV-2 epidemic. As opposed to off-line parameter estimation, our system "corrects" the model as soon as new data become available. A key feature of data assimilation is that the model will always track the data to an extent proportional to their assumed accuracy, allowing for a straightforward treatment of incomplete and noisy data, such as those for SARS-CoV-2. The ensemble data assimilation provides for assessing the impact of the implemented measures on the magnitude of the effective reproductive number, $R(t)$, as a function of time. Moreover, with

knowledge of the fatality or hospitality rates, we can infer the number of undetected infectious and predict the number of fatalities, hospitalizations, and infected people, under prescribed social distancing scenarios.

Some previous works have used data assimilation for epidemiology, in the context of both variational and Kalman filter-like methods, [6, 57, 33]. Most of the studies use sequential filtering approaches, e.g., the iterative filter and the ensemble Kalman filter (EnKF) for Cholera [44, 53], or the Ensemble Adjustment Kalman Filter (EAKF) for influenza and Ebola [72, 73, 63].

Recent studies on SARS-CoV-2 still primarily use a filtering approach. [22], used the EnKF to update parameters in a stochastic SEIR model. In contrast, [45] used the EAKF combined with a network of SEIR models, simulating different connected "cities" but without age stratification. In the context of variational data assimilation, [62] performed parameter estimation and predictions using a SIR model. [4] proposed a modified SEIR model that distinguishes between symptomatic and asymptomatic and utilized data assimilation to identify appropriate observing strategies.

We believe that critical characteristics of COVID-19, such as its approximately two-week timescale from infection to death and the nonlinear observation-state-parameter relation, motivate the use of smoothers instead of filters. In a smoother, state and parameter estimation is done by constraining the parameters on all the data in a time window, not only on current data. Classical adjoint-based smoothers, however, can suffer from multiple minima, typical of nonlinear systems using sparse and noisy data. Iterative ensemble smoothers, such as ESMDA, can, in many cases, mitigate these issues. Notably, [69] also used ESMDA for parameter estimation in an SEIR model.

We have used the same model and ensemble data-assimilation method (the ESMDA), with similar experimental configurations, for all the countries (hereafter referred to as "cases"). However, we have performed different experiments to respond to specific country-dependent questions and their various data provision and characteristics, demonstrating our system's enormous versatility. We also present several sensitivity experiments to key data assimilation parameters to highlight our approach's properties and robustness. The model-system code and input files for the countries under consideration are available from Github: https://github.com/geirev/EnKF_seir.

The outline of the paper is as follows: Section 2 describes the SEIR model used, and in Section 3, we give a brief introduction to the use of ensemble methods for model calibration. In Section 4, we discuss general aspects of the data assimilation experiments. Sections 5–12 present individual results from the different countries. In contrast, in Section 13, we present an overall comparative assessment of the obtained results across the modeled countries and states.

2. **SEIR model with age classes.**

A ubiquitous but straightforward model for epidemic modeling is the SEIR (Susceptible, Exposed, Infectious, and Recovered) model [7]. However, for a realistic description of the SARS-CoV-2 epidemic, we need to use an extended SEIR-model variant. A better formulation is the one conceptualized in Figure 1 and given by the following set of equations.

$$\frac{\partial \mathbf{S}_i}{\partial t} = -\left(\sum_{j=1}^{n} \frac{R_{ij}(t)\mathbf{I}_j}{\tau_{\text{inf}}} \right) \mathbf{S}_i \tag{1}$$

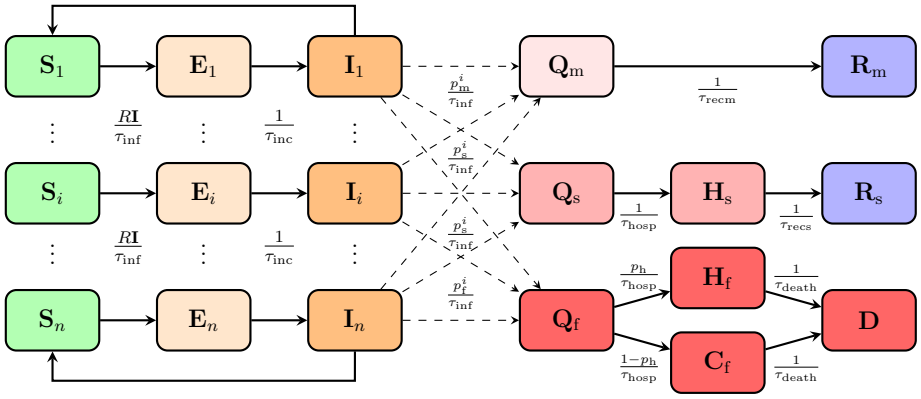

FIGURE 1. Flow diagram of the SEIR model.

$$\frac{\partial \mathbf{E}_i}{\partial t} = \left(\sum_{j=1}^{n} \frac{R_{ij}(t)\mathbf{I}_j}{\tau_\mathrm{inf}} \right) \mathbf{S}_i - \frac{1}{\tau_\mathrm{inc}} \mathbf{E}_i \tag{2}$$

$$\frac{\partial \mathbf{I}_i}{\partial t} = \frac{1}{\tau_\mathrm{inc}} \mathbf{E}_i - \frac{1}{\tau_\mathrm{inf}} \mathbf{I}_i \tag{3}$$

$$\frac{\partial \mathbf{Q}_\mathrm{m}}{\partial t} = \sum_{i=1}^{n} \frac{p_\mathrm{m}^i}{\tau_\mathrm{inf}} \mathbf{I}_i - \frac{1}{\tau_\mathrm{recm}} \mathbf{Q}_\mathrm{m} \tag{4}$$

$$\frac{\partial \mathbf{Q}_\mathrm{s}}{\partial t} = \sum_{i=1}^{n} \frac{p_\mathrm{s}^i}{\tau_\mathrm{inf}} \mathbf{I}_i - \frac{1}{\tau_\mathrm{hosp}} \mathbf{Q}_\mathrm{s} \tag{5}$$

$$\frac{\partial \mathbf{Q}_\mathrm{f}}{\partial t} = \sum_{i=1}^{n} \frac{p_\mathrm{f}^i}{\tau_\mathrm{inf}} \mathbf{I}_i - \frac{1}{\tau_\mathrm{hosp}} \mathbf{Q}_\mathrm{f} \tag{6}$$

$$\frac{\partial \mathbf{H}_\mathrm{s}}{\partial t} = \frac{1}{\tau_\mathrm{hosp}} \mathbf{Q}_\mathrm{s} - \frac{1}{\tau_\mathrm{recs}} \mathbf{H}_\mathrm{s} \tag{7}$$

$$\frac{\partial \mathbf{H}_\mathrm{f}}{\partial t} = \frac{p_\mathrm{h}}{\tau_\mathrm{hosp}} \mathbf{Q}_\mathrm{f} - \frac{1}{\tau_\mathrm{death}} \mathbf{H}_\mathrm{f} \tag{8}$$

$$\frac{\partial \mathbf{C}_\mathrm{f}}{\partial t} = \frac{(1-p_\mathrm{h})}{\tau_\mathrm{hosp}} \mathbf{Q}_\mathrm{f} - \frac{1}{\tau_\mathrm{death}} \mathbf{C}_\mathrm{f} \tag{9}$$

$$\frac{\partial \mathbf{R}_\mathrm{m}}{\partial t} = \frac{1}{\tau_\mathrm{recm}} \mathbf{Q}_\mathrm{m} \tag{10}$$

$$\frac{\partial \mathbf{R}_\mathrm{s}}{\partial t} = \frac{1}{\tau_\mathrm{recs}} \mathbf{H}_\mathrm{s} \tag{11}$$

$$\frac{\partial \mathbf{D}}{\partial t} = \frac{1}{\tau_\mathrm{death}} \mathbf{H}_\mathrm{f} + \frac{1}{\tau_\mathrm{death}} \mathbf{C}_\mathrm{f} \tag{12}$$

The total population number normalizes these equations. Thus the sum of all model variables, including the deceased $\mathbf{D}$, equals one and is time-invariant. For the whole population to stay constant in time, the equations' right-hand sides need to sum to zero.

We have stratified the populations of susceptible, exposed, and infectious into age groups $\mathbf{S}_i$, $\mathbf{E}_i$, and $\mathbf{I}_i$, see, e.g., [13]. As in the standard SEIR model, Eqs. (1)

Parameter	First guess	Description
τ_{inc}	5.5	Incubation period
τ_{inf}	3.8	Infection time
τ_{recm}	14.0	Recovery time mild cases
τ_{recs}	5.0	Recovery time severe cases
τ_{hosp}	6.0	Time until hospitalization
τ_{death}	16.0	Time until death
p_{f}	0.009	Case fatality rate
p_{s}	0.039	Hospitalization rate (severe cases)
p_{h}	0.4	Fraction of fatally ill going to hospital

TABLE 1. The table gives a set of first-guess model parameters. As we could not find scientific estimates of these parameters, we set their values based on available information from the internet and initial model-tuning experiments. We leave it to the data assimilation system to fine-tune the parameter values.

Age group	1	2	3	4	5	6	7	8	9	10	11
Age range	0–5	6–12	13–19	20–29	30–39	40–49	50–59	60–69	70–79	80–89	90–105
Population	351159	451246	446344	711752	730547	723663	703830	582495	435834	185480	45230
p–mild	1.0000	1.0000	0.9998	0.9913	0.9759	0.9686	0.9369	0.9008	0.8465	0.8183	0.8183
p–severe	0.0000	0.0000	0.0002	0.0078	0.0232	0.0295	0.0570	0.0823	0.1160	0.1160	0.1160
p–fatal	0.0000	0.0000	0.0000	0.0009	0.0009	0.0019	0.0061	0.0169	0.0375	0.0656	0.0656

TABLE 2. The p numbers indicate the fraction of sick people in an age group ending up with mild symptoms, severe symptoms (hospitalized), and fatal infection. The population-weighted averages (for the Norwegian population) of the case-fatality rate is 0.0090, and the rate of severe (hospitalized) cases is 0.039.

and (2) describe how the interaction between the infectious and the susceptible leads to the newly exposed. The effective reproductive numbers between different age groups R_{ij} together with the infection time scale τ_{inf} determine the rate of new infections. We will discuss the formulation used for R_{ij} in detail below. Note that the interaction between the susceptible and infectious constitutes the only source of nonlinearity in the model. Table 1 provides a set of default values of all the transition time scales in the model.

Equation (3) describes the exposed persons' transfer from $\mathbf{E}_i$ into the infectious group $\mathbf{I}_i$ at a rate given by the incubation time scale τ_{inc}.

The different age groups of infectious $\mathbf{I}_i$, transition into the various quarantined groups of sick, $\mathbf{Q}_{\text{m}}$, $\mathbf{Q}_{\text{s}}$, and $\mathbf{Q}_{\text{f}}$, based on the fractions p^i_{m}, p^i_{s}, p^i_{f}, and the infection time scale τ_{inf}, as modeled by Eqs. (4–6). The fractions refer to the portion of patients with mild symptoms, hospitalized patients with severe symptoms, and the fatally ill patients, and specify how the virus affects people of different age groups. The subscripts m, s, and f refer to mild, severe, and fatal symptoms. Thus, the model includes different probabilities for dying or being hospitalized dependent on the age group. The fractional coefficients sum to one for each age group. Table 2 provides an example set of fractions that illustrate how the SARS-CoV-2 virus affects older people more severely. We have assumed that a patient will not infect anyone while in a quarantined group.

Equation (10) describes how the patients with mild symptoms in $\mathbf{Q}_\mathrm{m}$ will recover and transition into the group of recovered with mild symptoms $\mathbf{R}_\mathrm{m}$, on a time scale τ_recm, without going to the hospital. Severely sick patients in $\mathbf{Q}_\mathrm{s}$ transfer to the hospital compartment $\mathbf{H}_\mathrm{s}$, on a time scale τ_hosp, as described by Eq. (7). After that, Eq. (11) models their recovery, which occurs on a time scale τ_recs, into the compartment of patients recovered from severe disease $\mathbf{R}_\mathrm{s}$.

Equation (8) models the fraction p_h of fatally-ill patients in $\mathbf{Q}_\mathrm{f}$ admitted to a hospital $\mathbf{H}_\mathrm{f}$ on the time scale τ_hosp. In Norway, the fraction p_h is around 0.4 since many fatalities were older people living in care homes, and they were usually not admitted to hospitals when they got infected by SARS-CoV-2. Thus, in Eq. (9) we also allow for a fraction, $1 - p_\mathrm{h}$, of fatally-ill patients that are not admitted to a hospital but rather transfer to $\mathbf{C}_\mathrm{f}$. The purpose of the $\mathbf{C}_\mathrm{f}$ variable is to include the fatally-ill patients not measured as hospitalized. Introducing $\mathbf{C}_\mathrm{f}$ allows us to use realistic fractions p_f of fatally-ill patients and still condition on the measured hospitalization numbers $\mathbf{H}_\mathrm{s} + \mathbf{H}_\mathrm{f}$. Within a few weeks of the pandemic, we had access to accurate estimates of the fraction dying within and outside hospitals for several countries. This partition of the fatally-ill patients turned out to be important for most of the cases discussed in this paper.

The fatally ill patients in $\mathbf{H}_\mathrm{f}$ and $\mathbf{C}_\mathrm{f}$ end up in the group of dead $\mathbf{D}$, on a time scale τ_death, as described by Equation (12).

We initialize the model with a country's total population divided among the age groups. We set an initial number of exposed and infectious reflecting the situation in a country, e.g., a sudden import of exposed or infectious split within some age groups. All other variables are set to zero initially. Tables 1 and 2 provide all the default model parameters. We can set the effective reproductive number in different ways, as discussed in Section 2.2.

2.1. **Some model aspects.** We chose to use an SEIR model for several reasons. First, the SEIR model concept is simple, yet it has provided realistic simulations of previous pandemics. Moreover, with the number of cases growing, the SEIR-model's aggregated variables provide more accurate statistical estimates. The model is nonlinear, but the nonlinearity does not pose substantial problems using ensemble data-assimilation methods, and a nonlinear model is required to model the pandemic over time realistically. An alternative could be to use machine learning and autoregressive models. Still, with the initial lack of observations to train networks or to compute statistics, the availability of well-tested dynamical SEIR models made our approach attractive. Moreover, machine-learning methods would not estimate the uncertainty that is an essential aspect of data assimilation. The additional compartments of hospitalized and deaths made it easy to condition on the observations. The starting point for our model was the online calculator: http://gabgoh.github.io/COVID/index.html.

In this study, we are modeling populations with vastly regional inequalities and different access to hospitals and proper care among various countries and population groups. Still, we have used the same model for diverse populations. This approach worked out by allowing the $\mathbf{C}_\mathrm{f}$ variable to include all fatally sick patients not admitted to a hospital. We can then estimate the parameters determining fatally and severely ill patients' fractions, p_f, and p_s (see Table 1), which differ in different countries and depend on the healthcare system's functioning and capacity. By doing so, we can model the pandemic in various countries.

It is possible to use different definitions of hospitalized, e.g., counting all patients hospitalized because of SARS-CoV-2 infections or counting only those in ICU by redefining the p_{s} parameter. Furthermore, it is possible to separate each of $\mathbf{H}_{\mathrm{s}}$ and $\mathbf{H}_{\mathrm{f}}$ into groups of intubated and non-intubated patients. With recent data, this addition would provide more detailed information regarding the need for ventilators, although it would not alter the total number of hospitalized and dead.

As a final note on the model, our data-assimilation approach allows us to determine the time scales and fractions defined in Table 1. However, results indicate that most of them are only marginally updated by assimilating data (i.e., their prior values were accurate).

2.2. **The effective reproductive number.** The most critical parameter in the model is the value of the effective reproductive number $R(t)$. When $R(t)$ is higher than one, this leads to an exponential growth of the epidemic, while the epidemic will die out when $R(t)$ is less than one. As long as $R(t) > 1$, the model simulates the development towards the population's immunity. If at some time, $R(t)$ becomes less than one and stays so for all future times, the epidemic dies out. In the experiments below, we will estimate $R(t)$ for the past using data assimilation. When running predictions, we set the mean and variance of $R(t)$ to reflect the anticipated effect of planned or expected policies and interventions.

It is convenient to use a square matrix $\mathbf{R}(t)$ with a size equal to the number of age groups, which allows for using different reproductive numbers in between the different age groups. It is then possible to include different transmission rates among children (who have fewer symptoms and maybe are less infectious) or elders (who got more easily infected in care homes). It also allows for simulating increased transmission among children from reopening schools or reduced transmissions among adults working from home.

Further, $\mathbf{R}(t)$ can change with time to reflect changes in implemented measures related to, e.g., social distancing. The equation describing $\mathbf{R}(t)$ in the model is

$$\mathbf{R}(t) = R(t)\hat{\mathbf{R}}, \tag{13}$$

where the scalar function $R(t)$ is estimated, while $\hat{\mathbf{R}}$ is a prescribed matrix that can differ for up to three different periods. In the model, $\hat{\mathbf{R}}$ is scaled such that for the age-weighted norm, we have $\mathbf{a}^{\mathrm{T}}\hat{\mathbf{R}}\mathbf{a} = 1$, with the vector $\mathbf{a}$ containing the fractions of the total population per age class. Thus, the scalar function $R(t)$ defines the effective reproductive number, while $\hat{\mathbf{R}}$ distributes prescribed relative transmissions among the age groups. The default matrix $\hat{\mathbf{R}}$ has all elements $\hat{R}_{ij} = 1.0$, and does not differentiate the transmissions between different age groups. We use the default $\hat{R_{ij}} = 1.0$ in all the experiments except for the Norwegian and England cases, where we present tables with the numbers used. In this work, we did not attempt to estimate the individual elements of $\hat{\mathbf{R}}$ since this would require much more detailed observations, e.g., an accurate number of cases per age group. Instead, we used the flexibility of setting different values for the elements in $\hat{\mathbf{R}}$ to model specific scenarios like the impact of sending children back to school in Norway. The overall model performance is most sensitive to the overall $\mathbf{R}(t)$, as the norm of $\hat{\mathbf{R}} = 1$, and then the scalar function $R(t)$, which we estimate, determines the growth or decay of the pandemic. The individual terms of $\hat{\mathbf{R}}$ act to differentiate the exposure among various age groups. Thus, using different values for the elements in $\hat{\mathbf{R}}$ in combination with conditioning on observed deaths and hospitalizations leads to an

adjustment of the total number of cases since the various age groups have different hospitalization and death rates. Initially, $R(0)$ equals the basic reproductive number R_0, and the initial transmission rate is $\beta = R_0/\tau_{\text{inf}}$ for an SEIR model without demography [7]. We use effective reproductive numbers $R(t)$, defined as the average number of secondary cases generated by each infectious person. In the absence of control measures, the effective reproductive number is the fraction of the number of susceptible times R_0. It is possible to reduce the effective reproductive number by introducing intervention measures such as social distancing. In contrast, the basic reproduction number remains unaffected, as it is a measure of the initial rate of infections when there are no interventions [2, page 347].

For the SARS-CoV-2 epidemic, it is essential to keep $R(t)$ below one to avoid exponential growth in the number of infected and thus hospitalized and dead. We can introduce various measures to reduce $R(t)$ to below one, e.g., immunization by vaccines, reducing social contacts, improving hygienic standards by washing hands, and wearing masks.

It is possible to specify different priors for $R(t)$. We can define $R(t)$ to be continuously or piecewise continuously varying in time, and we can add uncertainty to it, allowing us to update $R(t)$ using the ensemble methods described below.

2.3. **Intervention periods.** In the model, we have defined three main periods. The first initial period is from the start date of simulation until the introduction of interventions. In this period, we spin up the model from an uncertain initial condition, (size of the population in $\mathbf{I}_i$ and $\mathbf{E}_i$), and using a significant uncertainty around the prior reproductive number R_1. The second "lockdown" period is from the start of interventions until the "present" time. For this period, we can set a lower prior value of the effective reproductive number, R_2, to reflect the interventions' expected impact. The final period is for the prediction where we must assume the future distribution for $R(t)$ around the prior R_3. For the three different periods, we can give various structure matrices $\hat{\mathbf{R}}$.

3. **Ensemble data-assimilation methods for model calibration.** There is a vast literature on the use of ensemble Kalman filter (EnKF) type methods for sequential state and model parameter estimation, in high-dimensional and nonlinear inverse problems. Ensemble data assimilation is now standard and state of the art in a many operational prediction systems in the geosciences [14], including weather prediction [35], and petroleum applications [1]. The most popular ensemble-based data assimilation methods build on the EnKF [29, 12, 34, 3, 28, 25, 26].

Ensemble data-assimilation methods are popular for solving the state and parameter estimation problem in geosciences and petroleum applications. Being a highly nonlinear problem, it represents a formidable challenge whose solution led to a great stream of research and development [5]. We have used the Ensemble Smoother with Multiple Data Assimilation (ESMDA) [21] to calibrate model parameters in our SEIR model. The choice of ESDMA is motivated by the need to use a method that is efficient with large ensemble sizes and capable of handling non-linearities. We thus benefit from the extensive existing theory and methods initially flourished in geosciences and petroleum research. Here we study to what extent it can be useful to assess and predict the SARS-CoV-2 pandemic. ESMDA is an iterative ensemble smoother, meaning that it computes the posterior parameter estimates in one global computation using all data simultaneously. The smoother-approach differs

from the filter-approach in that the latter adds updates by sequentially introducing the measurements as they become available in time.

Let us write the model Eqs. (1–12) in compact form as

$$\mathbf{y} = \mathbf{g}(\mathbf{x}). \tag{14}$$

Here, $\mathbf{x}$ is a vector containing all the model's uncertain parameters, including initial conditions and the time-varying $R(t)$. The predicted measurements, $\mathbf{y}$, relate to the input parameters, $\mathbf{x}$, through the model, $\mathbf{g}$, which includes the model equations and a transformation of the model prediction onto the measurements. We have a time series of measurements, $\mathbf{d}$, of the number of deaths, the number of people hospitalized, and the number of cases,

$$\mathbf{d} \leftarrow \mathbf{y} + \boldsymbol{\epsilon}, \tag{15}$$

with stochastic errors, $\boldsymbol{\epsilon}$. The inverse problem solves for $\mathbf{x}$ given the predicted measurements, $\mathbf{y}$, and the observations, $\mathbf{d}$. It is possible to frame the inverse problem using Bayes' theorem as

$$f(\mathbf{x}|\mathbf{d}) \propto f(\mathbf{d}|\mathbf{g}(\mathbf{x}))f(\mathbf{x}), \tag{16}$$

where $f(\mathbf{x}|\mathbf{d})$ is the posterior probability density function of the parameters, $\mathbf{x}$, conditioned on the data $\mathbf{d}$. Equation (16) defines the so-called smoothing problem and is derived by, e.g., [24]. Our current approach is to use ensemble methods to approximately solve this equation [24, 5].

Developments originating from the petroleum applications have led to the use of new iterative ensemble smoothers such as the ESMDA by [21] and the ensemble randomized maximum likelihood (EnRML) by [18, 19]. Similar methods have been developed in the geosciences, albeit within a time-sequential context [9, 8], and have become popular for solving inverse problems of moderate nonlinearity. Recently, [55, 30] introduced a new efficient formulation of the EnRML that searches for the solution in the ensemble subspace, and this method is now operational in petroleum applications [30, 27]. Additionally, in applied mathematics, one has solved a broader class of inverse problems using the so-called ensemble Kalman inversion methods (EKI) [36, 17].

Without going into the mathematical details, the ensemble methods work as follows.

1. First sample a large ensemble of realizations of the prior uncertain parameters (e.g., the parameters listed in Table 1, the function $R(t)$, and the initial infectious $\mathbf{I}_i$, and exposed $\mathbf{E}_i$), given their prescribed first-guess values and standard deviations.
2. Integrate the ensemble of model realizations to produce a prior ensemble prediction characterizing the uncertainty.
3. Compute the posterior ensemble of parameters using the misfit between prediction and observations and the correlations between the input parameters and the predicted measurements.
4. Finally, compute the posterior ensemble prediction by a forward ensemble integration. The posterior ensemble is then the "optimal" model prediction with the ensemble spread representing the uncertainty.

The third step is essentially a linear-regression update. E.g., assuming that an increase in a parameter value yields an increase in a predicted measurement, then there is a positive correlation between the parameter and the prediction. It is then

possible to use this correlation to adjust the parameter value such that the posterior model prediction is closer to the observations.

For solving the SARS-CoV-2 data-assimilation problem, it is convenient to use an ensemble smoother. The main reason is that the time-varying and poorly known reproductive number, $R(t)$, at a particular time, determines the model predicted deaths and hospitalizations, two to three weeks later. A filtering data-assimilation method would update the predicted state variables but not be suitable for correcting a possible model bias caused by a poor estimate of $R(t)$ a couple of weeks earlier.

For a comprehensive explanation of using ESMDA to solve parameter-estimation problems, see [24]. The ESMDA algorithm solves the Bayesian parameter-estimation problem by gradually introducing the measurements' information to mitigate nonlinearity. The method handles strong nonlinearity, high model-state dimension, and a vast number of observations. In the Appendix, we provide a brief mathematical description of the ESMDA and comment on its sensitivity to the number of steps and ensemble size.

4. Introduction to assimilation experiments.

In this section, we discuss general aspects of our approach, such as the model's identifiability, the implications imposed by limitations in the observations, and how these limitations impact the way we can use the measurements.

4.1. Summary of uncertain model parameters.

In this study, we use a data-assimilation method to estimate the following uncertain parameters:

1. The six time scales, τ_{inc}, τ_{inf}, τ_{recm}, τ_{recs}, τ_{hosp}, and τ_{death} as listed, with their default prior values, in Table 1.
2. The case fatality rate, p_{f}, and the hospitalization rate for severe cases p_{s}, also listed with their default prior values in Table 1.
3. The initial number of exposed $\mathbf{E}_0$ and infectious $\mathbf{I}_0$, which were divided equally among the age groups.
4. The effective reproductive number $R(t)$ as a function of time.

In Table 1, we give default values of the time scales and rates. We used these values in most experiments, although in some cases, we modified them based on specific choices or additional available information.

We have assumed the time scales and rates to be constant in time. This assumption makes sense in this early study, where we only condition the parameters on observations available over an initial short period of about three months. At later times in the pandemic, we expect that some of these parameters may change, e.g., the case fatality rate, p_{f}, may decline if more effective medications and treatments become available. Conversely, significant overloading of the healthcare capacity can limit medical treatment availability and increase the value of p_{f}. In future studies, we can easily include new uncertain parameters to be estimated, or we can make them time-dependent as in the case of $R(t)$.

When using ensemble methods to condition the model parameters on observations, the definition of prior distributions for the parameters serves as an effective regularization of the inverse problem. The parameters that impact the predicted measurements will be constrained by the observations, while the rest of the parameters will retain their prior distributions. In the following sections, we discuss how we can identify the various model parameters and variables given the limited set of available observations.

When estimating the time-continuous effective reproductive number, $R(t)$, we impose that it is smooth in time by specifying a time decorrelation length in the ensemble perturbations added to the first guess of $R(t)$. The assumption of $R(t)$ to be smooth in time reduces the parameter estimation problem's effective dimension and is realistic.

4.2. **Overview of observed data.** We only had access to a somewhat limited selection of relevant observations. In most countries, we were able to obtain numbers of deaths and hospitalizations. However, these data are uncertain and collected differently in different countries. Additionally, most of the countries published the number of cases, i.e., positive tests, and sometimes also the number of people tested. Nevertheless, the procedures for selecting the population to be tested differed in various countries, and with time, and limited testing capacity induced large biases in these data.

Thus, in most experiments, we have conditioned the model only on the measured deaths and hospitalization numbers. In different countries, we had to process these data differently, as some death records would include deaths in care homes and others not. Also, the data did not always include deaths occurring outside the hospitals. The hospitalization data sometimes only included ICU patients or excluded severely ill patients in care homes.

It was also challenging to use the number of registered cases since these data are biased. However, we have used the case data in some experiments assuming a fraction of positive tests over the real number of cases and a significant measurement error. This approach sometimes helped to constrain the overall number of cases to "realistic" numbers.

One particular data set would be of great value for calibrating the model: i.e., unbiased random tests of positive cases in the population. Such data would allow for better estimation of the total number of infections in society and reduce some of the biases we likely have in our current results.

4.3. **Different countries' focus.** When running cases for different countries, we have examined several topics and properties of the system. While in all cases, we have run hindcast experiments where we calibrate the model system to the current observations followed by scenario predictions, there are also several additional topics studied that differ between countries and states. In Norway, we model scenarios of reopening schools and evaluate the impact of enhanced infections among children. For England, we assess the effect of conditioning on different data types in hindcast experiments. In the Québec case, we study the reliability of potential probabilistic short-range forecasts issued with the system. In The Netherlands, we evaluate the impact of conditioning on different data types (including ICU patients), and we examine the effect of using different priors for $R(t)$. In France, we run hindcast experiments with different priors for $R(t)$ while in Brazil, there is an additional focus on understanding the impact of varying observation errors. For Argentina, the main focus is on scenario predictions and assimilating only the accumulated deaths due to insufficient hospitalization data quality. For the US, we also run sensitivity analysis with different priors for $R(t)$ and use hindcast experiments for assessing the impact of interventions. We present a brief synthesis of the results from the different in Section 13.1.

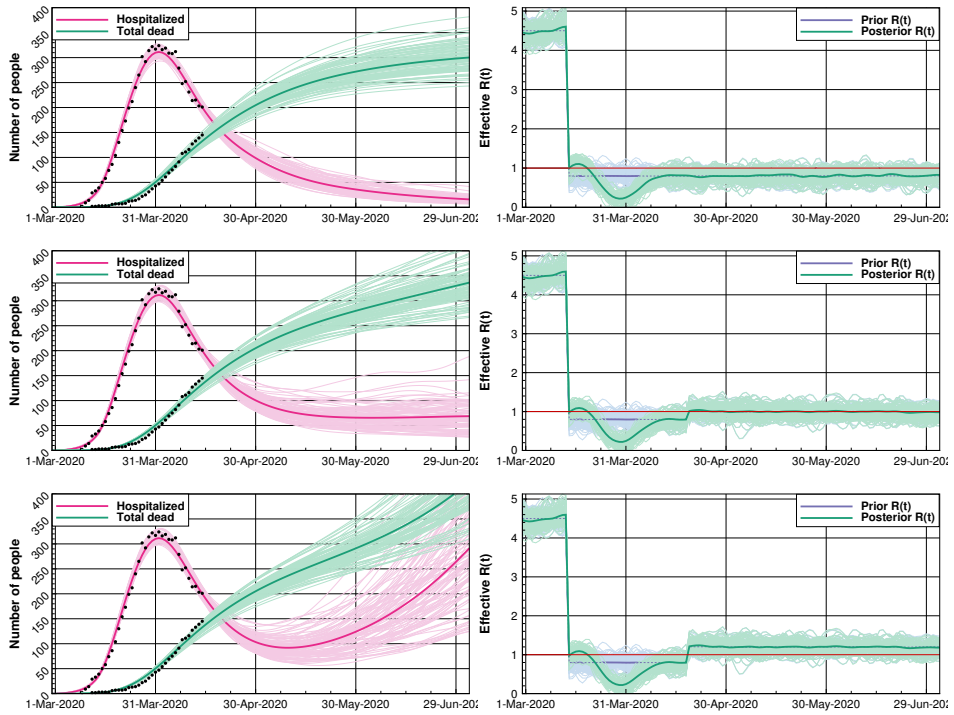

FIGURE 2. Norway: This figure summarizes scenarios related to opening up kindergartens and schools on the 20th of April. The left plots show the ensemble means and the 100 first ensemble realizations, for the number of hospitalized and the accumulated amount of deaths for different scenarios of future $R(t) = 0.8$, 1.0, and 1.2. The right plots show the prior and posterior ensembles of $R(t)$. The red thin line in the plots for $R(t)$ is an indication of the value $R(t) = 1$ for easier identification.

5. **A case study for Norway.** In Norway, the Government acted quickly and imposed a lockdown, closing all kindergartens, schools, universities, and all noncritical functions on March 15th. Several companies ordered their staff to work from home. All restaurants and bars closed. Still, society remained partly open. People were allowed to go hiking and training outside as long as there was no physical contact between them. The interventions did not confine Norwegians to their homes as in several other countries, and the psychological impact has probably been less severe. Considerable controversy surrounded the ban against going to cabins in the mountains or along the coast. However, this ban most likely reduced the spread of the virus across different regions, and it prevented potential overloading of the limited health-care systems in rural areas. Note that our model treats a country as a whole and does not segregate variables into regional compartments with exposure between regions. Thus, we could not model the impact of the "cabin ban".

The interventions came after about two to three weeks with sporadic imports of cases, mainly through people returning from skiing vacations in Austria and Italy. In the days before the close-down, there was an increasing number of cases where the

Age groups	1	2	3	4	5	6	7	8	9	10	11
1	**3.3**	1.8	1.8	1.3	1.3	1.0	0.9	0.9	0.9	0.9	0.9
2	1.8	**3.3**	1.8	1.3	1.3	1.3	0.9	0.9	0.9	0.9	0.9
3	1.8	1.8	**0.9**	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9
4	1.3	1.3	0.9	**0.9**	0.9	0.9	0.9	0.9	0.9	0.9	0.9
5	1.3	1.3	0.9	0.9	**0.9**	0.9	0.9	0.9	0.9	0.9	0.9
6	1.0	1.3	0.9	0.9	0.9	**0.9**	0.9	0.9	0.9	0.9	0.9
7	0.9	0.9	0.9	0.9	0.9	0.9	**0.9**	0.9	0.9	0.9	0.9
8	0.9	0.9	0.9	0.9	0.9	0.9	0.9	**0.9**	0.9	0.9	0.9
9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	**0.9**	0.9	0.9
10	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	**0.9**	0.9
11	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9	**0.9**

TABLE 3. Norway: This $\hat{\mathbf{R}}$-matrix increases transmissions among children after opening kindergartens and schools on April 20th. We chose the numbers *ad-hoc* to give a qualitative impact of opening kindergartens and schools. To estimate these transmissions' correct values, we will need access to additional data that are not yet available.

infection's origin was unknown, making clear that the virus was spreading through the population. The initial lockdown was active for five weeks, until April 20th, when the kindergartens reopened. The schools opened for the first four cohorts one week later on April 27th. During April, Norway had a steady reduction in hospitalized COVID-19 patients from a maximum of 322 patients on March 30th. The government, therefore, decided to reopen society gradually, starting with kindergartens and schools. During this period, the number of deaths and new cases per day did not increase. In the following weeks, some restaurants started reopening, although with a minimum distance between tables and no accumulation of people. As of June 6th, Norway had experienced only three deaths in the previous two weeks, only 24 COVID-19 patients were in the hospital, and the number of newly detected cases per day was around ten. At this point, Norway was able to test anyone with a symptom and run active contact tracing related to all positive cases. Thus, the general opinion was that Norway gained control of the epidemic.

The initial model development was motivated by the need for monitoring and predicting the epidemic in Norway. In particular, we believed that the introduction of ensemble-based data-assimilation methods for model calibration would lead to an ideal tool for making short-term predictions, evaluating scenarios, and estimating the impact of different interventions on the effective reproductive number. We communicated the model scenarios to Norwegian authorities to provide decision support related to managing the interventions. Early during the pandemic, we wished to awake awareness of the severity of the epidemic and the crucial importance of introducing interventions to halt the virus's spread. Furthermore, we wanted to explain how precarious the situation was concerning the reproductive number's value and the meaning of exponential growth. Later, the model provided useful short-term predictions to plan for additional hospital beds, ventilators' needs, and healthcare resources management. The model also allowed to assess the impact of interventions on the value of the reproductive number, which is essential information when managing different measures.

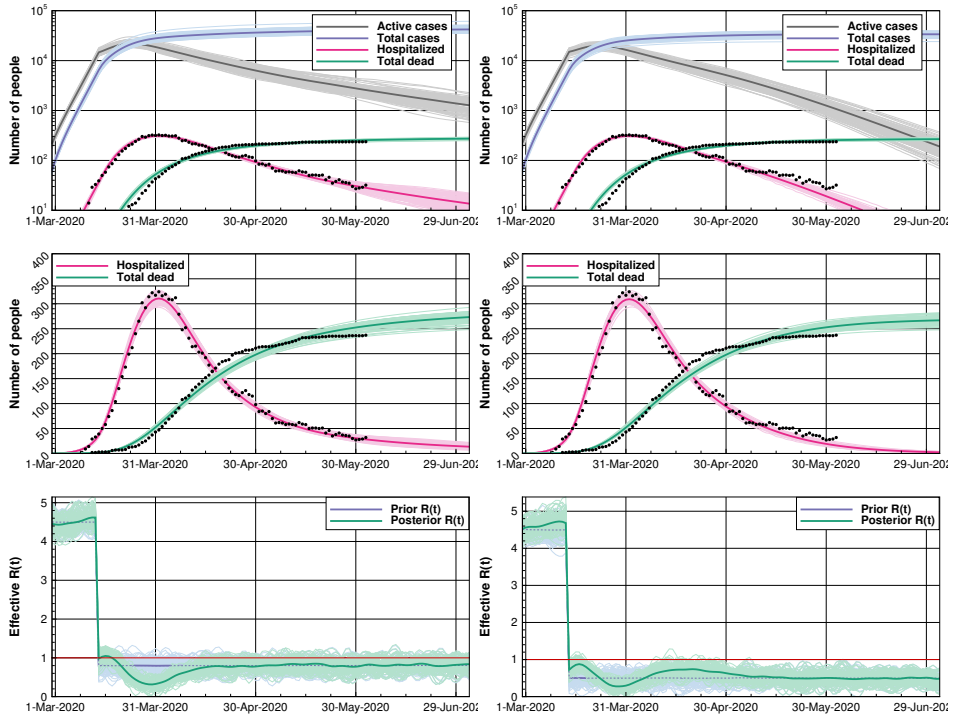

FIGURE 3. Norway (base case): For the two cases (left and right plots), the only difference is the prior-guess for $R(t)$ after starting the interventions. For the first two rows of plots, we show the posterior ensemble means and the 100 first realizations of the posterior solution from ESMDA. The blue lines are the total number of cases, while the gray lines give the number of active cases. The red curves denote the number of hospitalized, and green lines show the total number of deaths. The upper plot uses a log y-axis. The second row is a zoom of the upper plot using a linear y-axis. The lower plots show the corresponding prior and posterior estimates of $R(t)$ for the two cases. The red thin line in the plots for $R(t)$ is an indication of the value $R(t) = 1$ for easier identification.

5.1. **Observations and uncertain parameters.** With a relatively small population and a well-developed health-care service, we had access to reasonably accurate daily observations of the number of hospitalized and deaths in Norway. Also, we could access the daily number of positive cases. However, with the selected testing of prioritized groups, e.g., health-care workers, the cases data were strongly biased and were only used qualitatively to set some prior parameters of the model. The national newspapers published all the observed data, e.g.: https://www.vg.no/spesial/2020/corona/#norge.

The initial values of the parameters used in the model are the ones given in Table 1. For the time-scales and the rates, p_{f} and p_{s}, we set an uncertainty of 10%. Table 2 gives the age-specific distribution of the relative fractions of mild, severe, and fatal disease (we used these values for all the experiments in this paper). We

set the parameter $p_{\mathrm{h}} = 0.4$ for Norway based on published data of the fraction of deaths occurring in hospitals versus care homes.

5.2. **Exp. 1: Impact of gradually re-opening schools.** We initially added age groups to the model to simulate the impact of reopening schools and kindergartens. By increasing the transmissions among children who would be interacting at school, and introducing a slight increase among the children and their parents, we reasoned that this would lead to a higher effective reproductive number. If the effective reproductive number were to get above one, the spreading would be unstable. Thus, excessive transmissions among children would have to be balanced by less spread among the adult population.

On April 14th, we used the ensemble method presented in Section 3 with the model described in Section 2 to run sensitivity simulations for assessing the impact of gradually reopening children's schools and kindergartens from April 20th. The age-based model considered 11 age groups as defined in Table 2.

The values in $\hat{\mathbf{R}}$ were initially equal to 1.0 and the same between all age groups until April 20th when schools and kindergartens reopened. Then, we introduced the $\hat{\mathbf{R}}$ matrix from Table 3, used in the simulation after letting children back to kindergartens and schools. The matrix allowed for using different transmission factors between different age groups. On the diagonal, the value gives the transmission of disease within the same age group. The off-diagonal terms are the transmissions between age groups. We assumed that open kindergartens and schools would lead to significant spreading of the virus within these groups. We also included an increased transmission between parent groups and children. As explained in Section 2, this matrix's weight-averaged norm is one, so it does not contribute to the effective reproductive number. This choice of $\hat{\mathbf{R}}$ is specific for Norway. For countries with other demographic conditions, e.g., if multiple-generations families are typically living together in the same house, and we would need to consider these differences when specifying $\hat{\mathbf{R}}$.

A base run fitted model parameters to the total observed deaths and the current number of hospitalized using data until April 14th. We configured the model with a spinup period before the start of the interventions on March 15th. In this spinup period, we used a prior value of $R_1 = 4.5$ and a standard deviation of 0.2. From March 15th to April 14th, during the interventions, we used a prior value of $R_2 = 0.80$ and a standard deviation of 0.15. The model parameters, including $R(t)$, were then updated using the ESMDA method.

We communicated three prediction scenarios using the $\hat{\mathbf{R}}$ matrix in Table 3 from April 20th, and with different prior values of $R(t)$ in the prediction phase, $R_3 = 0.8$, $R_3 = 1.0$, and $R_3 = 1.2$, all with zero-mean Gaussian noise with standard deviation equal to 0.15. Figure 2 presents the results of these cases, with increasing value of R_3 for the prediction, from top to bottom.

It turned out that we were too pessimistic in the predictions. The lockdown in Norway was effective and nearly stopped further spreading the virus during the five weeks between March 15th and April 20th. In practice, we ran scenarios with values of R_3 that were too high and predicted a more severe development of the epidemic than what occurred.

More important than predicting the exact number of fatalities and hospitalized was to illustrate the qualitative evolution of the pandemic in the three cases when the effective reproductive number is less than one (stable), equal to one (neutral), or

larger than one (unstable). In particular, we wished to raise awareness among the population and decision-makers about the pandemic's functioning and severity. If everyone understands how epidemics develop under different conditions, it is easier to argue for and accept the implemented interventions. The three scenarios with varying values of R_3 result in vastly different future predictions, thus evidencing how it is essential to keep $R(t)$ less than one.

Small changes in the interventions can lead to an $R(t)$ larger than one, and two-three weeks later, a bloom of new cases will result. Therefore, at an early stage of the pandemic, it is essential to halt the virus spreading abruptly and obtain control of the active infections. After that, it is possible to open society gradually while monitoring the impact on the reproductive number. Finally, from the $\hat{\mathbf{R}}$ matrix in Table 3, we see that it is possible to balance increased spreading within some age groups by fewer transmissions in other age groups.

Based on these simulations, a conclusion was the following: *The opening of children's schools and kindergartens would likely yield growth in new cases for the young age groups. The virus would then spread to their parents, teachers, and the rest of the community. A continued effective lockdown in the reminder of society would be required to stabilize the growth. Still, the measures would have to be sufficient, or we might have experienced exponential growth again. With open schools, it is not likely that the rest of the society would manage to keep $R(t)$ below one due to the additional commuting and people going back to work. We suggested that the Coronavirus's extreme transmission rate would make the opening of society very risky. We would likely experience multiple new local exponential blooms of infected people.* We have later experienced a few local "blooms" and an increase in $R(t)$, but, as will be discussed below, the partial and gradual reopening strategy has worked out very well.

5.3. **Exp. 2: Recent scenarios for Norway.** A more recent scenario includes all the data until the end of May, and Figure 3 presents the results. The left panels show an optimized case and an online estimation of $R(t)$ during the lockdown period. After March 15th, we used a prior value of $R(t) = 0.8$ with a standard deviation of 0.15. The interventions implemented on March 15th led to an immediate reduction of $R(t)$ in the following week, and $R(t)$ reached a minimum of 0.3 to 0.4 on March 31st. During April, $R(t)$ gradually restored towards its initial value of 0.8. Using data up to the end of May, we would expect to see an impact on $R(t)$ through mid-May. Thus, to complement this discussion, we re-ran the experiment using a prior value of $R(t) = 0.6$. The panels in the right column of Figure 3 show the results, and we see that there was indeed a positive uptick of $R(t)$ until mid-May. It turned out that the prior value of $R(t)$ in the first case happened to be close to the correct value for May.

5.4. **Summary of results for Norway.** A key result from these experiments is that, by conditioning the model on the time-series of observed hospitalizations and deaths, it is possible to estimate the effective reproductive number, $R(t)$, as a function of time, until about two weeks before the last observation. This promising result should make it possible to compute the impact on $R(t)$ from introducing or removing specific interventions, as is exemplified in Exp. 2.

The experiments, Exp. 1 and Exp. 2, confirm that new measured deaths and hospitalizations impact the updates of $R(t)$ about two weeks earlier. Thus, infections must have happened about two weeks earlier since it takes some time before infected people get hospitalized or die.

These experiments also confirm that society's partial reopening led to an increase in $R(t)$ but, luckily, $R(t)$ stayed well below one. Also, from the current prediction, at the end of May, Norway had about 3,000 active cases. If the interventions would continue to retain $R(t) \sim 0.80$, then the number of active cases should be well below 1,000 at the end of July, and there should only be a few hospitalized patients.

6. **A case study for England.** At the time of writing, the UK has been the worst-hit country in Europe in terms of deaths. Some fundamental differences between the UK and Norway help explain the differences in the epidemic's evolution. These differences include the UK's higher population density, 273 people per square km compared to 14 in Norway, with most of the population concentrated within the South-East England and London, and different social habits and wealth distribution. The UK government's response to the epidemic was also much different, initially attempting to contain and delay the outbreak and only going into lockdown once approximately 700 deaths had already occurred in England and 9,000 people had tested positive.

Due to the different recording protocols across the devolved countries, data collection for the whole UK is not straightforward. To enable more consistent use of data, we focused on England only, which makes up approximately 85% of the UK population.

The following chronology is useful to understand the build-up of the epidemic in England.

- By February 19th, there were a total of 20 positive cases registered. On March 5th, the first two deaths due to COVID-19 occurred.
- On March 17th, the UK Prime Minister urged people not to go to public places (pubs, theatres, gyms) and work from home. Universities started closing that week.
- On March 20th, the government passed legislation stating the closing of schools and non-essential businesses [54]. The lockdown started effectively on March 23rd.
- On May 13th, lockdown measures were slightly relieved: workers who cannot work from home have been allowed to come back to work, although they have been discouraged from using public transport.
- On June 1st, a more substantial relaxation of the lockdown measures commenced. Nurseries and schools were allowed to reopen for children up to age six and those in their final year of primary school, while bars and restaurants remained closed, and public gatherings were not permitted. Outdoors meetings for up to six people were allowed.

6.1. **Observations and uncertain parameters.** We can obtain data on accumulated deaths for England from three different sources:

- The National Health Service (NHS) England provides data on deaths recorded against the actual date of death [49]. Deaths outside hospitals, such as those in care homes, are not included.
- Public Health England (PHE) provides figures for deaths that have had a diagnosis of COVID-19 confirmed by a PHE or NHS laboratory [65]. They

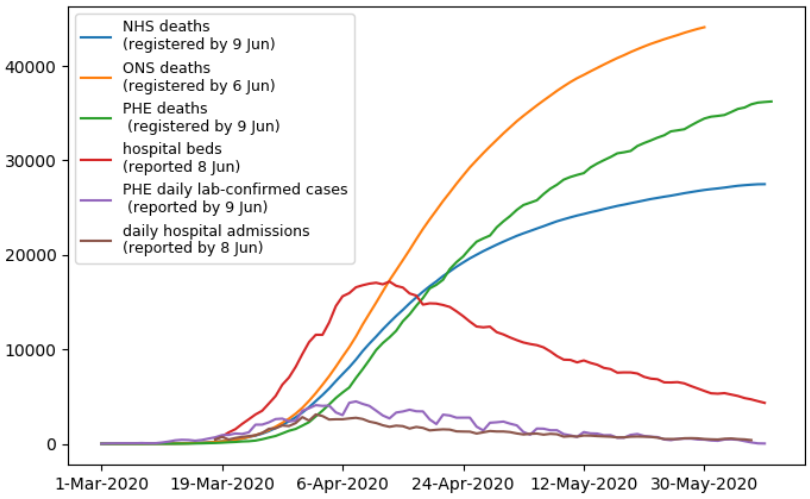

FIGURE 4. England: The plots show the data available for assimilation. Three different agencies report the number of deaths, the UK government press conference publishes the number of people in hospitals, CHESS reports the daily hospital admissions, and PHE presents the number of new cases.

offer death numbers according to the day they were reported, not the day they occurred.

- Office for national statistics (ONS) publishes deaths, where the death certificate refers to COVID-19, every Tuesday [68]. These deaths include cases outside the hospital and cases where COVID-19 is suspected, but no formal diagnostic test has occurred. They only report registered deaths up to 11 days before the date of publication.

Figure 4 compares the accumulated deaths reported from the three agencies. The end dates are June 9th for the PHE (green) and NHS (blue) data. However, only data from before June 4th can be considered complete. In contrast, the ONS (orange) delay reporting their data so that the figures will not change. The discrepancies between the three lines reflect different ways of attributing the death to COVID-19 and the date of occurrence. For instance, note how the PHE data, which reports by registration date, display a more rapid increase in deaths during the working week than at the weekends. We can estimate the proportion of deaths occurring outside of hospitals by taking the difference of the NHS data from the ONS data. This difference increases until saturating at approximately 39% on May 29th. The ONS data include some deaths that are not confirmed by a COVID-19 test. Thus, in the model, we set the fatally ill fraction that goes to the hospital to a slightly reduced value of $p_{\mathrm{h}} = 0.7$.

The ONS data is the most appropriate for our purposes, and we shall use it in our data assimilation experiments; the first available data are on March 5th. It is, however, worth noting that due to multiple causes of death, we can not necessarily

attribute the deaths solely to COVID-19 from the ONS data. ONS also provides data on the average number of deaths in the previous five years for comparison to try to disentangle this. For example, in the week ending April 17th, 2020, there were 22,351 deaths registered in England and Wales, of which 8,758 attributed to COVID-19. This number compares to an average of 10,497 deaths in the previous five years for the same week. The scale of these numbers brings confidence that the new presence of COVID-19 is responsible for the dramatic increase in deaths. Unfortunately, many excess deaths are not necessarily caused by the virus but by pressure on the NHS. Thus, more people are dying from causes other than COVID-19, that one would otherwise cure. Our model does not cover these pandemic consequences, which poses an essential factor when interpreting the data.

Data on the number of people in hospital with COVID-19 in England and Wales is being collected by PHE and reported by the daily UK government press conferences [67]. These are the data for hospitalization used in our experiments, and we plot them in Figure 4 as the red line; the first data is on March 20th.

For completeness, we also show the daily cases testing positive for SARS-CoV-2 reported by PHE (purple line; note that, as with the deaths, only data up to June 4th can be considered complete). It is possible to compare these cases with daily admission rates reported by the COVID-19 hospitalization in the England Surveillance System [66] (CHESS, brown line). Note that the number of cases testing positive is closely following the hospital admission rates as people hospitalized have been receiving priority in testing. Hence, the number of positive cases reported can be assumed to be a significant underestimate of the actual number. As an illustrative exercise, we will use ESMDA to estimate this underestimation approximately and correct the testing data before their use in the experiments.

We start our simulation on February 20th and assimilate data from March 5th to May 29th (recall that hospitalization data begins on March 20th). The model parameters were calibrated against data for England during several tuning experiments. To account for the under-sampled testing, we set the initial number of infected people to $I_0 = 60$, three times the number of accumulated positive tests provided by PHE by February 19th. The initial number of exposed is $E_0 = 240$. Similarly to what was done in the Norwegian experiments in Section 5.2 and 5.3, we also set $R(t)$ to be discontinuous in time. The prior value of the reproduction number before the intervention was $R_1 = 3.87$ [31]. The model is informed about the intervention date of March 23rd by reducing the first guess for the reproduction number to $R_2 = 0.9$, but with a relative standard deviation set to 50%. A large standard deviation accounts for a large uncertainty in $R(t)$ after introducing the intervention. As seen from the results, it provides the needed variability in the ensemble members for DA.

We have defined the pre-lockdown transmission matrix, $\hat{\mathbf{R}}_1$, for England based on results from [47], who estimated "contact" matrices in a European study in 2006. We have adopted a transmission, $\hat{\mathbf{R}}_2$, used during the lockdown, from [42], who estimated a similar "contact" matrix for the UK one day after the introduction of the lockdown. Note that the age bins used within these studies are slightly different from our model's ones. Therefore, we mapped the 0-4 age bin used in [42] to our 0-5 age bin. We further mapped the 5-17 age bin used in [42] into our 6-12 and 13-19 age bins (with enhanced mixing between the two groups), and we mapped the 18-29 age bin used in [42] to our 20-29 age bin. Following a normalization as described concerning Eq. (13), these two matrices provide us with the transmission

Age groups	1	2	3	4	5	6	7	8	9	10	11
1	**2.0**	1.5	1.5	1.0	1.5	0.5	0.5	0.5	0.4	0.4	0.4
2	0.5	**8.0**	6.0	2.0	2.5	2.5	1.5	1.4	0.9	0.9	0.9
3	0.5	6.0	**8.0**	2.0	2.5	2.5	1.5	1.4	0.9	0.9	0.9
4	0.5	2.5	2.5	**6.0**	2.0	2.0	1.9	1.5	0.9	0.9	0.9
5	1.2	2.5	2.5	2.0	**3.0**	2.0	1.9	1.8	0.5	0.5	0.5
6	0.5	2.3	2.3	2.0	2.0	**3.0**	1.9	1.5	1.4	1.4	1.4
7	0.5	2.0	2.0	1.5	1.5	1.5	**2.0**	1.5	0.9	0.9	0.9
8	0.5	1.9	1.9	1.0	1.2	1.2	1.9	**1.5**	0.9	0.9	0.9
9	0.5	1.5	1.5	0.9	0.9	1.2	1.0	1.5	**1.5**	1.5	1.5
10	0.4	1.0	1.0	0.9	0.7	1.2	1.0	1.0	1.5	**1.5**	1.5
11	0.4	0.9	0.9	0.9	0.7	1.2	1.0	1.0	1.5	1.5	**1.5**

TABLE 4. England: The contact matrix $\hat{\mathbf{R}}_1$ used to describe the transmission between different age groups in England before the enforced lockdown on March 23rd. The same contact matrix is used for the prediction from June 1st. See the right panel of Figure 2A in [42] for a heat-map representation of the original matrix.

matrices needed by our model to describe the inter-age group transmission before and during the lockdown. Table 4 and 5 show the two contact matrices $\hat{\mathbf{R}}_1$ and $\hat{\mathbf{R}}_2$. The entries of $\hat{\mathbf{R}}_1$ indicate that before lockdown, school-age children exhibit the largest number of contacts consistent with the school environment. Note that these matrices are not necessarily symmetric. For example, while most children will interact with adults under 50, either guardians or teachers, not all adults will necessarily have contact with children. The transmission matrix for the lockdown period, $\hat{\mathbf{R}}_2$, displays a large drop in the number of daily contacts. However, the reduction is not the same across all age groups, with the school-age children showing the most dramatic decrease in contacts within their age group.

The values of the remaining parameters used for the England case are the same as for Norway (see Table 1).

6.2. **Exp. 1: Hindcasts over the data period.** The first experiments aim to study the goodness of the fit to data and calibrate the model against them. We consider three DA configurations where we assimilate different combinations of accumulated deaths (D), the daily number of hospitalized (H), and the total number of positive cases (C). The data were available for the period from March 5th to May 29th.

The number of ensemble members (5,000), the number of ESMDA iterations (32), as well as the analysis types (stochastic EnKF) are the same as for the Norwegian experiment. Similarly, we assume the observation-error standard deviations to be 5% of the data value. However, in the Norwegian example, we set the assumed error standard deviation's maximum value to 6 people. Given the grander scale of deaths in England, we change this to be 5% of 8,000, i.e., 400. The default is that the same errors apply to the number of hospital beds occupied with COVID-19 patients as the accumulated deaths. In the experiments assimilating the number of cases testing positive, we have assumed an error standard deviation of 10% for the case observations with a maximum of 50,000. We have corrected the data to account for biases in reporting, as described later. As in the Norwegian example, a

Age groups	1	2	3	4	5	6	7	8	9	10	11
1	**1.0**	0.9	0.9	0.8	1.0	0.5	0.5	0.4	0.3	0.3	0.3
2	0.5	**2.0**	1.5	0.9	1.0	1.0	0.5	0.4	0.3	0.3	0.3
3	0.5	1.5	**2.0**	0.9	1.0	1.0	0.5	0.4	0.3	0.3	0.3
4	0.5	1.0	1.0	**1.2**	1.0	1.0	0.9	0.5	0.4	0.3	0.3
5	0.8	1.0	1.0	0.9	**1.1**	0.9	0.9	0.5	0.4	0.3	0.3
6	0.5	1.0	1.0	1.0	1.0	**1.1**	0.9	0.5	0.4	0.3	0.3
7	0.5	0.6	0.6	0.9	0.9	0.9	**1.0**	0.7	0.5	0.5	0.5
8	0.5	0.6	0.6	0.8	0.9	1.0	1.0	**1.0**	0.5	0.5	0.5
9	0.5	0.6	0.6	0.6	0.5	1.0	0.9	0.9	**1.1**	1.1	1.1
10	0.5	0.6	0.6	0.6	0.5	1.0	0.9	0.9	1.1	**1.1**	1.1
11	0.5	0.6	0.6	0.6	0.5	1.0	0.9	0.9	1.1	1.1	**1.1**

TABLE 5. England: The contact matrix $\hat{\mathbf{R}}_2$ used to describe the transmission between different age groups in England during the lockdown from March 23rd to May 31st. See left hand panel of Figure 2A in [42] for a heat-map representation.

de-correlation half-length scale of the observations is set to 10 days to account for systematic errors in the reporting.

Assimilating accumulated deaths only (top panels of Figure 5) brings a close fit to the corresponding data. However, the observed daily hospitalizations are substantially overestimated (cf the hospitalization data in the mid-left panel of Figure 5). The estimated effective reproductive number $R(t)$ until the intervention on March 23rd displays a vast, possibly nonphysical, fluctuation from above 4.5 for two weeks and then decrease quickly to about 3.0 afterward. The performance of ESMDA significantly improved when we assimilated daily hospitalizations in combination with observed deaths (the case DH, mid panels). The fit to the hospitalizations data is excellent, and the estimated $R(t)$ at the pre-interventions period undergoes much smaller oscillations. However, its estimate for the first week after March 5th is still above the assumed basic reproduction number $R_1 = 3.8$. After that, $R(t)$ is below or close to 0.9 for the first half of the lockdown period. It then stays above the critical unitary value from mid-April to mid-May, and finally, slightly below 1 toward the end of the lockdown. Given that we are conditioning on data until May 29th, and the time scale of the process is two weeks, further data is required to accurately describe the last part of the lockdown.

As anticipated in Section 6.1, the reported number of cases testing positive for SARS-CoV-2 underestimates significantly the actual number of people infected by the virus. We can estimate the magnitude of this underestimation by comparing the number of cases predicted by experiment DH to the reported number of cases (purple line in Fig. 4). There is uncertainty in the number of cases predicted by our experiments, as represented by the ensemble spread; this uncertainty is far smaller than the magnitude of the differences between the predicted values and those observed. From the results of this comparison (not shown), we find that as the number of people infected grew exponentially before the lockdown started (on March 23rd), the percentage of positive cases not being reported increased due to a severe lack of testing kits. On March 25th, this number peaked, with reporting of only 1% of cases. As more tests became available, and we simultaneously had a

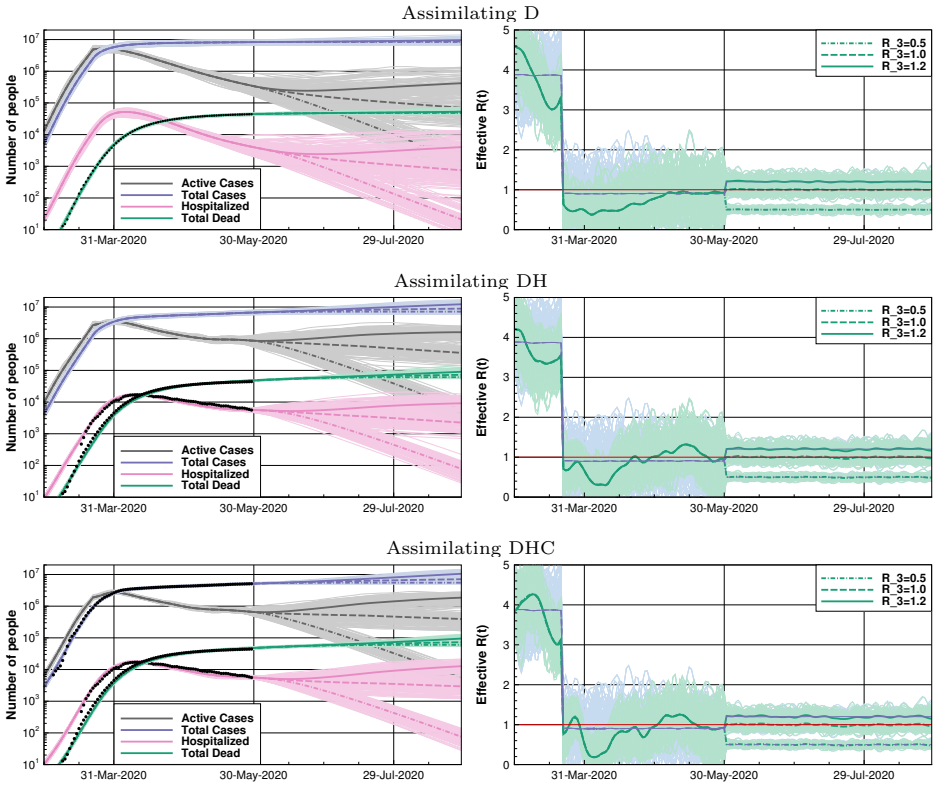

FIGURE 5. England. Left panels: ESMDA posterior estimates of the accumulated number of deaths (green), the daily number of hospitalizations (red), active (gray), and total cases (blue). Observations are displayed in black. Right panels: Prior and ESMDA posterior estimates of the effective reproduction number $R(t)$. Ensemble members (thin lines) and ensemble means (thick lines). Assimilation experiment D (top panels), DH (mid panels) and DHC (bottom panels). The red thin line in the plots for $R(t)$ is an indication of the value $R(t) = 1$ for easier identification.

reduction in the infection rate, the percentage of unreported cases decreased. From the beginning of May, about 2% of the accumulated cases are reported. The UK government achieved its self-imposed target of 100,000 tests per day for the first time on May 1st. One can compare the estimated percentage of reported cases to the percentage of asymptomatic cases. Estimates of the proportion of asymptomatic cases vary widely between studies. However, WHO suggests that 80% of infections are mild or asymptomatic [71], supported by [38]. In the UK, we were only testing people displaying symptoms. Therefore, we could conclude that, if the fraction of total accumulated cases reported is estimated to be about 2%, then approximately 10% of the accumulated symptomatic cases are being reported from the beginning of May in the UK.

Parameters	Prior	Posterior D	Posterior DH	Posterior DHC
I_0	59.97 (6.06)	61.32 (6.01)	61.95 (6.01)	60.08 (5.98)
E_0	240.64 (24.17)	246.68 (23.85)	251.56 (23.62)	239.43 (23.46)
τ_{inf}	3.80 (0.50)	2.73 (0.33)	3.13 (0.33)	2.83 (0.28)
τ_{inc}	5.50 (0.50)	4.62 (0.40)	4.79 (0.40)	5.14 (0.33)
τ_{recm}	13.98 (0.49)	13.99 (0.49)	13.94 (0.49)	13.94 (0.49)
τ_{recs}	4.99 (0.41)	4.98 (0.40)	4.13 (0.31)	3.57 (0.31)
τ_{hosp}	5.99 (0.51)	5.54 (0.49)	5.35 (0.48)	4.79 (0.39)
τ_{death}	15.99 (0.50)	15.64 (0.50)	15.13 (0.47)	14.43 (0.44)
p_{f}	0.009 (0.001)	0.009 (0.001)	0.014 (0.0009)	0.014 (0.0002)
p_{s}	0.039 (0.004)	0.039 (0.003)	0.011 (0.002)	0.015 (0.002)

TABLE 6. England: Prior and posterior mean and standard deviation for the time independent parameters estimated with ESDMA in the experiments D, DH and DHC.

To account for this change in the under-reporting of positive cases with time, we applied a piece-wise correction to PHE data. We estimated the correction from comparing the daily cases predicted by the DH experiment with the data (described above). This study resulted in the following correction of the data: between March 5th and March 20th, the inflation factor needed to correct the number of positive cases reported daily increased linearly from 25 to 120. The inflation factor then decreased linearly to 12 on April 7th, where it remained until April 29th, before increasing linearly to 60 on May 29th. As an exercise, we assimilated these "corrected" observations in the DHC experiment (shown as black points following the total cases in Figure 5).

Given the dramatic modifications to the original data, one should assimilate these corrected observations with caution, and we have increased their assumed standard deviation to 10%. As expected, the fit to the (corrected) observed cases are improved, and the results support that for about two weeks after March 5th (before the lockdown), $R(t)$ was stably larger than 3.8. Still, it decreases to about 3.0 and finally near 1 and below during lockdown. The increase in $R(t)$ above one from mid-April to mid-May, observed already in experiment DH, is also found here. However, as already pointed out, more recent data would be needed to constraint the solution in that period.

We present the values of the model parameters estimated by the ESMDA in the three experiments D, DH, and DHC, in Table 6, and we recall the prior values in the second column for reference. Numbers do not appear to be dramatically different across the different experiments and between the priors and posteriors. However, it is worth noting the differences in p_{s} and p_{f} when assimilating hospitalizations compared to not.

6.3. **Exp. 2: Prediction experiments.** From June 1st to September 1st, following the lockdown measures' relaxation, we run predictions under three different epidemic scenarios with the prior reproduction numbers $R_3 = 0.5$, 1.0, and 1.2. To the neutral scenario, $R_3 = 1.0$, we added a pessimistic, $R_3 = 1.2$, and an optimistic one, $R_3 = 0.5$, representing a situation without or with effective measures to contain the epidemic. We use the same transmission matrix for all scenarios for the pre-lockdown period: we set $\hat{\mathbf{R}}_3 = \hat{\mathbf{R}}_1$, and present the results in Fig. 5, commencing after the last data point (black point). For the optimistic case, $R_3 = 0.5$

(dot-dashed line), the differences in the forecasts between the experiments D, DH, and DHC, are minor. However, case D tends to differ more (in particular to predict substantially fewer deaths and hospitalizations), arguably due to not being constrained on hospitalization data. Even in this more favorable scenario where we assume the measures will keep $R(t)$ as small as 0.5, experiments DH and DHC predict more than 60,000 casualties by September 1st, as opposed to 48,000 from experiment D. Thus, even in this "favorable" scenario, the capability of simultaneously assimilating multivariate observations (deaths, hospitalization, and cases) allows for tracking 10,000 more deaths.

For the neutral case, $R_3 = 1.0$ (dashed line), DH, and DHC predict 70,000 deaths and 2,300 hospitalizations, while D predicts 50,000 casualties and 720 hospitalizations by September 1st. Further differences between the three experiments become evident when $R_3 = 1.2$ (solid line). The cases DHC and DH predict 90,000 deaths, against 50,000 for D, and 10,000 hospitalizations against less than 5,000 for experiment D. Note that DHC indicates systematically for all three values of R_3 considered, slightly more deaths and hospitalizations than DH. This example confirms the importance of access to a variety of useful data and their multivariate treatment that data assimilation offers.

6.4. **Summary of results for England.** Our results confirm the findings for the Norwegian case (see Section 5.4). In particular, we could also infer the time-dependent $R(t)$ up to two weeks before the last observation. The hindcasts, Exp. 1, have highlighted our approach's versatility, the ability to handle data of different types, such as deaths and hospitalizations, and significantly improve the estimate's fit to observations. We have also illustrated how to compare the predicted number of cases to the raw data. Thus, we provide an approximate quantification of the underestimation of the real number of cases.

A special consideration deserves the predictions, Exp. 2. As explained in Section 6.3, we prescribed scenarios by imposing prior values for the reproductive number, intended to reflect the impact of containment policies of different degrees. Nevertheless, other critical parameters of the model, notably the case fatality rate, p_f, or the transmission rates among age groups, were left unaltered to the values assigned to or estimated during the data period. Consequently, the model is not taking into account any improvement of, e.g., the medical care facilities or mitigating therapies that can reduce the case fatality rate or the implementation of a faster testing system capable of identifying infectious at an earlier stage of the illness. Our data assimilation framework will allow for running prediction experiments that incorporate scenarios for the case fatality and intra-age-groups transmission rates with appropriate modifications. We will consider these issues in a follow-up study. Before that, we must interpret the results of the predictions qualitatively.

7. **A case study for Québec, Canada.** This section evaluates options for real-time short-range prediction of the pandemic's evolution in Québec. In Section 7.1, we discuss the observations used, and we propose a processing step to account for reporting delays. In Section 7.2, we note that the SEIR modeling environment permits a relatively good match to the observations. In Section 7.3, we retro-actively perform short-range forecasts to study the reliability of potential probabilistic forecasts issued with the system. Limitations are discussed in Section 7.4.

7.1. **Observations and uncertain parameters.** Data for Québec are available from the INSPQ [41], which is the national public health institute for Québec. Data until May 28 have been used (as extracted on May 29). These include the accumulated number of deaths, the number of positive cases, and the current hospitalizations. The fatalities also include those occurring in care homes for the elderly and any situation where COVID-19 was a direct contributing factor. It does not include any deaths due to other causes such as delayed medical interventions in a stressed health care system or fear by sick persons of getting infected during a necessary hospital visit. We can attribute the deaths to the day of reporting or the day they occurred. In Québec, one had not designed the reporting system to provide rapid updates during an evolving epidemic, and counts for a specific day of occurrence only become relatively stable after about seven days. Thus, there are frequent changes made to the database. Accounting for the significant reporting delays is essential for the Québec case, as in Section 7.3, we focus on the quality of short-term forecasts.

In addition to reported hospitalizations and positive tests, we condition on reported deaths, d_{R}, to estimate the model parameters. To reflect the approximately four-day delay between an actual or modeled death, $d(t_i)$, at day t_i, and the reported death, we use:

$$d_{\mathrm{R}}(t_i) = \begin{cases} d(t_i = 0), & t_i < 4 \\ d(t_i - 4), & 4 \le t_i < 8\,. \\ 0.37d(t_i - 8) + 0.32d(t_i - 4) + 0.31d(t_i), & 8 \le t_i \end{cases} \tag{17}$$

Here, both $d(t)$ and $d_R(t)$ refer to cumulative total values at time t. The values of $d(t)$ only are updated as reports on deaths at time t come in with a delay. We used stabilized data to fit the coefficients in Eq. (17). Using more points in time did not substantially improve the agreement between the reported and transformed stabilized series. We applied this equation to the model predicted deaths, $d(t_i)$, whenever comparing the model values and the observations, $d_{\mathrm{R}}(t_i)$, such as for a figure.

We can estimate the excess mortality during the epidemic from weekly data published by the Québec Statistical Institute (ISQ) [39]. As of May 15th, the weekly totals were available until April 25th. The accumulated mortality from March 1st until April 25th, for the five years from 2015 until 2019, points to an average excess amount of 1919 deaths. The standard deviation in the five individual values is 218. For April 25th, the INSPQ reports a total of 1921 occurred deaths, which is remarkably close to the value provided by the ISQ. However, ISQ points out that the estimate of the most recent weekly total deaths only covers about 80% of the real number. Thus some evidence of additional excess mortality related to COVID-19 might appear in the future. Note also that for April 25th, we extracted the data from both sources on May 29th. The ISQ data indicated an excess of 2019 deaths, and the INSPQ total was 2001, which was still in good agreement with the value provided by the ISQ.

The number of hospitalizations is possibly accurate. But, we didn't know the hospitalized fraction of the fatally ill. Many deaths occurred at care homes for the elderly, so we used a fraction, $p_{\mathrm{h}} = 0.5$, in the experiments. On May 20th, the number of hospitalizations reduced by approximately 200 due to a different counting method. From May 20th onward, one removed the recovered patients still in the hospital, awaiting a transfer, from the number of hospitalized patients. We

Parameters	Prior(Std Dev)	DHC	DH	D
R_1	3.0(0.6)	-	-	-
R_2	1.0(0.5)	-	-	-
I_0	100.0(20.0)	67	98	96
E_0	240.0(48.0)	167	204	235
τ_{inc}	5.5(1.0)	5.2	3.4	3.8
τ_{inf}	3.8(0.6)	1.8	1.9	2.7
τ_{recm}	14.0(2.0)	14.1	12.8	14.8
τ_{recs}	5.0(1.0)	6.9	6.8	5.5
τ_{hosp}	6.0(1.2)	5.9	5.8	6.7
τ_{death}	10.0(2.0)	5.8	3.4	10.4
p_{f}	0.020(0.004)	0.020	0.021	0.023
p_{s}	0.039(0.006)	0.040	0.047	0.038
p_h	0.5(0)	-	-	-

TABLE 7. Québec: The set of prior model parameters and their standard deviations (a zero std dev denotes that the parameter is kept fixed). Columns DHC, DH and D show posterior values for, respectively, experiments DHC, DH and D. Note that p_h was supplied externally. The curves for R_1 and R_2 are shown in Figure 6.

distributed this shift of 200 over the preceding days as a sequence of 40 per day. This approach avoids a jump in the assimilated data.

The number of positive test cases includes likely associated cases, such as family members, where testing was deemed unnecessary. Initially, testing had to be limited to the subjects that were most likely to be positive, notably travelers. Later on, as the number of available tests improved, it became possible to test other groups, as well. Thus, over time, the testing became more comprehensive. Over the entire period, the assumption is that only 15% of positive cases show up in the reported statistic. Thus the reported number of cases was multiplied with a factor, 1.0/0.15, for use in the assimilation experiments. For the three data types, we use a relative error of five percent. We have neglected possible delays in the testing and reporting procedures for the number of positive cases. Finally, we obtained the population data from the Québec Statistical Institute for 2019 [40]. The total population is 8,484,965 and, as before for Norway, available per age range.

Table 7 gives the parameters used for the epidemic in Québec. The prior value of 0.02 for the case fatality rate is rather high, reflecting that the epidemic hit the most fragile part of the population very hard. The uncertainty in the parameter values is about twice the ones used for the other cases discussed in this manuscript. We choose these larger values to have an appropriate amount of uncertainty in the forecast ensembles. Contributing factors in the early development of the epidemic were the "Spring Break," which was from March 1st until March 6th, as well as close connections between the metropolitan areas of Montreal and New York. The SEIR model does not capture such events. Instead, the model is starting with relatively high estimates of the initially exposed and infectious. Even with the restrictive measures being introduced gradually over several days, we assume that they all became effective on March 23rd. For the initial period of free exponential growth, we use a prior estimate of $R_1 = 3.0 \pm 0.6$ for the effective reproductive number. To not bias our posterior estimate of the effective reproduction number to values

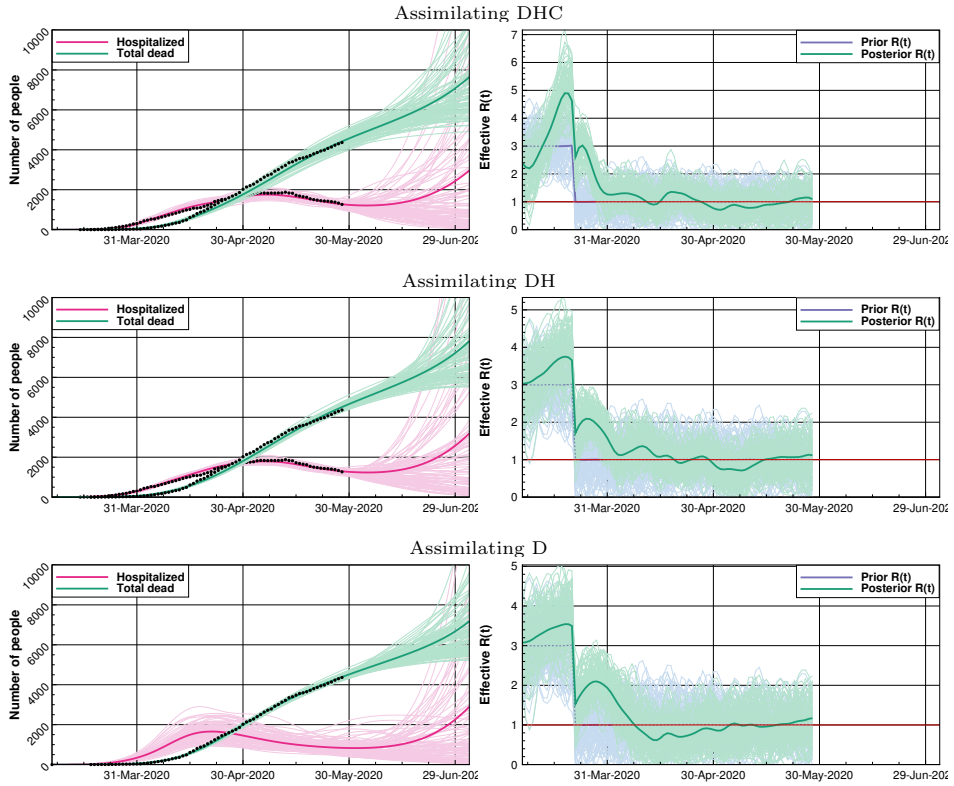

FIGURE 6. Québec: From top to bottom, the plots show the results from the three assimilation experiments DHC, DH, and D. The left column presents the accumulated number of deaths and the number of hospitalizations, and the right column shows corresponding the reproductive number $R(t)$ for the experiments. Time is since the start of the epidemic on March 8th. Observations are indicated with points when used to obtain the model fit. The solid lines are for the ensemble mean posterior estimates. After May 28th, we kept the realizations of $R(t)$ constant and equal to the latest values for the remainder of the simulation. The red thin line in the plots for $R(t)$ is an indication of the value $R(t) = 1$ for easier identification.

above or below unity, the prior estimate is $R_2 = 1.0 \pm 0.5$ for the second (and final) period. We estimate the effective reproductive number until May 28th and keep it constant after that. Note that the available data only weakly constrain the effective reproductive number in the final week.

7.2. **Exp. 1: Data assimilation experiments.** Like the England case, we have performed three types of experiments: DHC - assimilating deaths, hospitalizations, and positive cases; DH - assimilating fatalities and hospitalizations, and D assimilating deaths only. The upper plots of Figure 6 shows the results for experiment DHC. The actual and reported number of fatalities match rather closely from the

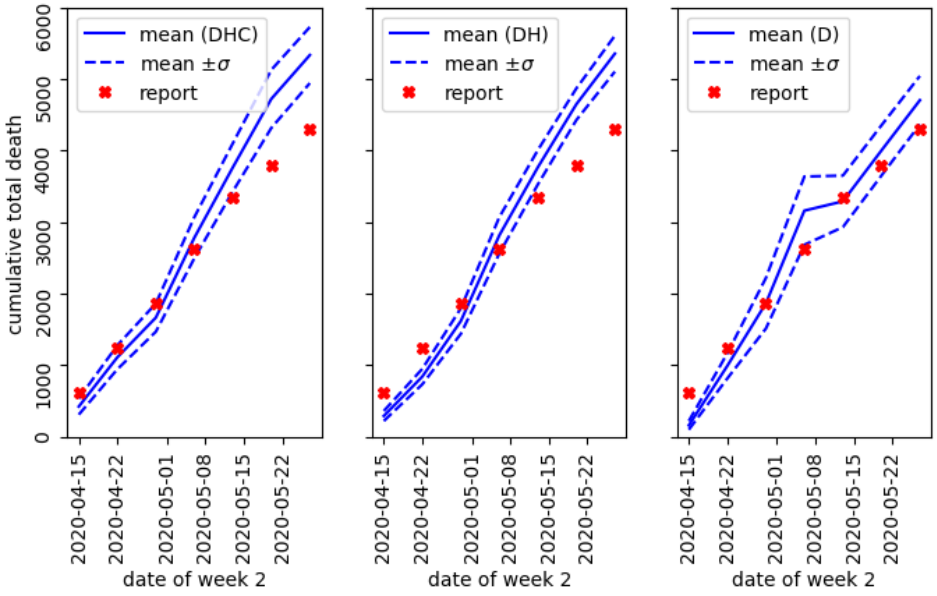

FIGURE 7. Québec: The plots show verification of retroactive week-two forecasts for experiments DHC, DH, and D. For the predictions, issued one week apart, we show the mean estimate with the full line, and the dashed lines give the mean value plus or minus one standard deviation. We indicate the reported values with crosses.

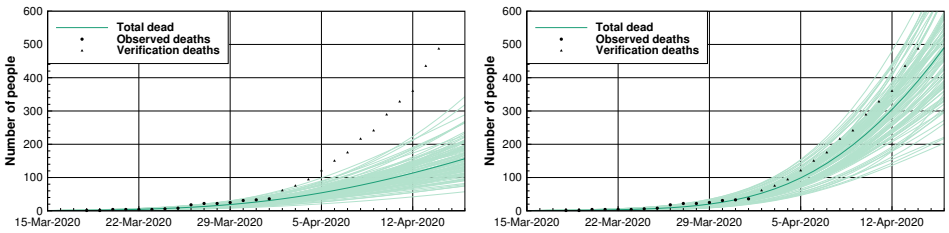

FIGURE 8. Québec: The plots present verification of the week-two forecasts (retroactively) issued on April 1st for experiment D (left) and the experiment DHC (right). The solid line denotes the mean prediction. The reported values used to fit the model parameters are indicated with circles, while the triangles are the values used for verification.

beginning of the epidemic. For the hospitalizations, the curves match after the initial week, and for the positive cases (not shown) after the initial two weeks.

We note that after the initial period of rapid growth in March, the epidemic gradually slowed down in April. As of May, the numbers of new cases, new deaths, and hospitalizations are relatively constant, or slowly decaying, resulting in near-unity estimates of the reproductive number. The gradual reduction in the reproductive

number may reflect both the gradual introduction of more restrictive measures for the general population and the progressive improvement of care homes' procedures to control the epidemic.

The middle plots of Figure 6 show the results for experiment DH. The numbers for the initial infectious and exposed are higher than in the experiment DHC (Table 7). However, note that the fraction of detected cases, which was assumed to be at 15% for experiment DHC, is essentially an unknown variable. Thus, additional information on the percentage of infected people would be necessary to calibrate the fraction and permit to connect the numbers of cases and the numbers of infectious and exposed.

Finally, the lower plots of Figure 6 shows experiment D, in which we only assimilated the observed number of deaths. All experiments feature a reasonably good fit between modeled and observed values, with generally minor differences. The reduction in hospitalizations in May suggests that the epidemic has peaked for all experiments. In the experiments DHC and DH, we do not fit the reduced rate of new deaths observed during May, as good as in experiment D that only assimilated observed deaths. This issue reflects the difficulty of having coherence between the SEIR model and all observed variables over the entire experimental period. Note that experiment D, which did not use the information on hospitalizations, features a broad ensemble spread for this variable. The increase of the mean-predicted number of hospital admissions in the last weeks of June results from having a few members with rapid exponential growth.

For the three experiments, the first wave of the epidemic likely peaked between mid-April and early May and gradually declined. In June, due to the exponential nature of the outbreak, the uncertainty multiplies during the forecast. Recent data on the epidemic's evolution would ideally support the gradual reopening up of society in June. It is thus essential that these data become available more quickly. Concerning predictions for a possible second wave of the outbreak and the eventual arrival at group immunity, it will be essential to perform additional random tests to estimate the fraction of infected persons.

7.3. **Exp. 2: Predictability experiment.** As a verification of the general methodology, we performed a sequence of retrospective 2-week forecasts for the configurations DHC, DH, and D. The first of these forecasts use data until April 1st and are valid on April 15th. After that, we made successive projections separated by one week, and we issued the final predictions on May 13th and valid for May 27th. The results in Figure 7 show that the size of the forecast errors corresponds reasonably well with the plus or minus one standard deviation range of the ensemble of model forecasts. Initially, the deaths are underpredicted, while during the last part, we overpredict them.

The unsuccessful forecast issued on April 1st, with configuration D, is shown in the left plot in Figure 8. Despite restrictive measures being in place, the virus started super spreading in houses for the elderly, leading to a rapid increase in the reported number of deaths as early as April 2nd. By April 15th, the actual number of fatalities is beyond the 1000 member forecast ensemble's members. We obtain a much better forecast with the DHC configuration (right plot of Figure 8). As positive cases and hospitalizations typically precede deaths by some days, the assimilation of these data helps predict the rising number of deaths. The experiments DHC and DH do, however, overpredict the number of reported deaths for the last two forecasts in May (Figure 7).

7.4. **Summary of results for Québec.** Whereas we generally obtained good agreement between the model and the observations, the model tends to overpredict the fatalities at the end of the period. A hypothesis is that the improving procedures in the health care system, notably in the care homes, made the SARS-CoV-2 epidemic less fatal. Additional testing and contact tracing likely increased the fraction of detected asymptomatic cases, which led to an excessive prediction of deaths. Finally, we may relate the apparent improvement to the reporting delays. On May 30th, the end-of-month adjustment reported an unusually large number of 202 deaths.

The need for setting large standard deviations for the initial parameter values may reflect that the SEIR model cannot describe the full complexity of the evolution of the epidemic in Québec. In numerical weather prediction, one has developed methods to ensure sufficient ensemble spread when using an imperfect forecast model [11, 70, 35]. Here, because of the general progression in response to the outbreak, it could also be appropriate to tune model parameters separately for the epidemic's different phases.

8. **A case study for The Netherlands .** We have applied an SEIR model with the ensemble data-assimilation method to study the effect of the prior assumptions on $R(t)$ on the estimates of the spreading of SARS-CoV-2 in The Netherlands. The first observed COVID-19 infection occurred on February 27th in the province of North Brabant. On March 1st, the government advised anyone with COVID-like symptoms in this province to stay home. On March 9th, the inhabitants of North Brabant were encouraged to work from home as much as possible. Three days later, the rest of the country followed this advice. When schools and daycare centers and restaurants and bars were closed down on March 16th, the contact between people decreased further. Children under the age of 12 years old were still allowed to play outside. Thus, the interaction among children of this age group has only slightly decreased due to government measures.

The spreading of the virus increased the pressure on the Dutch healthcare system, with a peak of 611 patients admitted to the hospital on a single day on March 27th [60]. The number of registered infections per day reached a maximum of 1398 on April 10th [60]. On April 21st, when the number of new infections and hospitalizations showed a consistent decrease, the Dutch government announced a slightly less stringent lockdown with the possibility for children until the age of 18 to practice outdoor sports as of April 28th, and the re-opening of daycare centers and elementary schools for children until the age of 12 as of May 11th. By this time, outdoor sports for groups of adults became possible, albeit at 1.5 m distance. A further opening-up took place on June 2nd, with restaurants and bars opening. In the same week, high schools re-opened while respecting the 1.5 m distance between students. Throughout Dutch society, the population adheres to the imposed 1.5 m distance between individuals, except for those using public transportation.

For The Netherlands, we consider a single experiment with four different cases to investigate the effect of assimilating different data types on the estimates of cases, hospitalized, and deaths. By choosing different priors for $R(t)$, we also investigate the sensitivity of the posterior estimate of the reproduction number to these priors.

8.1. **Observations and uncertain parameters.** For The Netherlands, data on registered infections, hospitalizations, and deaths resulting from the SARS-CoV-2 virus are shared daily by the Dutch RIVM [60] and in the Dutch media. The number

Parameter	First guess	Std. Dev.	Description
E_0	500.0	50.0	Initially exposed
I_0	400.0	40.0	Initially infectious
R_1	3.8	0.05	Reproduction number before interventions (Case 1DH and Case 1DI)
R_1	0.8	0.01	Reproduction number after first nation-wide intervention (Case 1DH and Case 1DI)
R_1	1.0	0.75	Reproduction number (Case 2DH)
p_{s}	0.010	0.0001	Hospitalization rate (Case 1DI)
p_{s}	0.039	0.0039	Hospitalization rate (Case 1DH, Case 2DH and Case 3DH)
p_{h}	0.5		Fraction of fatally ill going to hospital (Case 1DI)
p_{h}	0.6		Fraction of fatally ill going to hospital (Case 1DH, Case 2DH and Case 3DH)

TABLE 8. The Netherlands: The table gives values of the parameters used in Case 1, for the parameters that are different from those indicated in Table 1. The starting date of the simulations is February 20th, 2020.

Case	Assimilated data	Description
1DI	deaths, ICU patients	prior $R(t)$ equals 3.8 before intervention, 0.8 after
1DH	deaths, hospitalized	prior $R(t)$ equals 3.8 before intervention, 0.8 after
2DH	deaths, hospitalized	prior $R(t)$ equals 1.0
3DH	deaths, hospitalized	prior $R(t)$ equals 1.8 at start of simulation and gradually ramps down to 0.8

TABLE 9. The Netherlands: overview of cases.

of COVID-19 patients present in intensive care units (ICU) has been monitored intensively by the Dutch foundation of national intensive care evaluation (NICE), which also publishes numbers on the total number of COVID-19 patients in hospitals [51, 52]. For this manuscript, we use the hospitalized and ICU patient numbers from NICE and the death count from RIVM, all downloaded on June 3rd, 2020. For the partitioning of the population between the different age groups, we use the data of Statistics Netherlands (CBS) [15]. The reproduction number itself is assumed to be independent of age, so the distribution of p numbers across age groups is a direct function of the populations of the different age groups, similar to the Norwegian p numbers presented in Table 2.

We have performed a large number of sensitivity-runs to evaluate the choice of prior parameters for The Netherlands (not shown). Based on this, we have decided to use the first-guess values given in Table 1. In Table 8, we provide an overview of the country-dependent parameters used in the simulations.

We observe a strong linear correlation between the number of ICU patients with COVID-19 and the number of hospitalized COVID-19 patients during the onset of the epidemic. This relationship led us to conduct experiments in which we assimilated the numbers of ICU patients rather than the more uncertain data on hospitalized patients. The correlation changes after March 27th, shortly before the

number of ICU patients reaches a maximum of 1313 on April 7th. When the number of ICU patients is consistently decreasing, we observe a different correlation. Of course, there is a distinct difference between the admittance of ICU patients and patients that receive regular hospital care. When the pressure on the medical systems increases, this pressure will likely be observed more strongly in the standard care units than in intensive care units (provided that intensive care units' capacity is sufficient). This fundamental difference between the two cohorts may have consequences for the suitability of the observed numbers of ICU patients in a model configured to assimilate the total number of hospitalized.

8.2. **Exp. 1: Sensitivity to assimilated data type and prior reproduction number .** Exp. 1 for The Netherlands investigates the influence of the conditioning on different sets of data on the estimate of deaths and hospitalized. It also addresses how the prior for $R(t)$ affects the resulting posterior estimate of the reproduction number. In Exp. 1, we consider four different cases. In all cases, we assimilate the number of deaths, the number of ICU patients (Case 1DI), or the number of hospitalized (Cases 1DH, 2DH, and 3DH); see Table 9. In Cases 1DH and 1DI, the prior reproductive number $R_1 = 3.8$ before any country-wide intervention. After that, we reduced it to $R_2 = 0.8$. In Case 2DH, $R(t) = 1.0$, while in Case 3DH, $R(t)$ is a function of time that starts at 1.8 and gradually reduces to 0.8. The prior $R(t)$ of Case 3DH resembles more the slowly changing $R(t)$ shared by the RIVM, and we consider it to be the most knowledge-based prior out of the four different priors.

Experiments that assimilate the number of infected struggle to fit all three data types. Given that the number of registered infected may not represent the total number of positive cases, we decide not to use these data.

In Case 1DI, in which we assimilate ICU patients rather than hospitalized patients, the use of a lower value of p_s reflects the difference in hospitalization rate between these two groups of patients. The numbers set for I_0 and E_0 for the initial infected and exposed are relatively high compared to the number of registered positive cases. This choice accounts for the fact that the number of registered infected is likely an order smaller than the actual number of infected.

The number of deaths and hospitalized or ICU patients for Case 1DH or Case 1DI, respectively, are shown in the two top panels on the left-hand side of Figure 9. In Case 1DI, the estimated number of ICU patients has a reasonably good fit to the data. The fit to the observed number of hospitalized in Case 1DH is poorer, especially during the onset of the epidemic. The uncertainty of these numbers possibly caused this more significant misfit. In all simulations, the estimated number of deaths tends to be higher than the observed number of deaths in the epidemic's onset. The estimated number of fatalities tends to have a slightly flatter course in the period that follows.

In both the Cases 1DI and 1DH, the posterior estimate of $R(t)$, as depicted in Figure 9, appears to fit the prior value of 3.8 relatively well. We interpret the dip in $R(t)$ after the first intervention as an adaptation effect to the sudden change in the prior for $R(t)$. The differences between estimates of the number of deaths and $R(t)$ between the Cases 1DI and 1DH are relatively minor. For consistency with the other experiments in this paper, the additional cases shown assimilate the numbers of hospitalized patients.

To investigate the evolution of $R(t)$ without relying on the prior estimate, Case 2DH assumes a constant first-guess value of 1.0 and a relatively high standard deviation of 0.75. The resulting fit to the data is comparable to Cases 1DI and

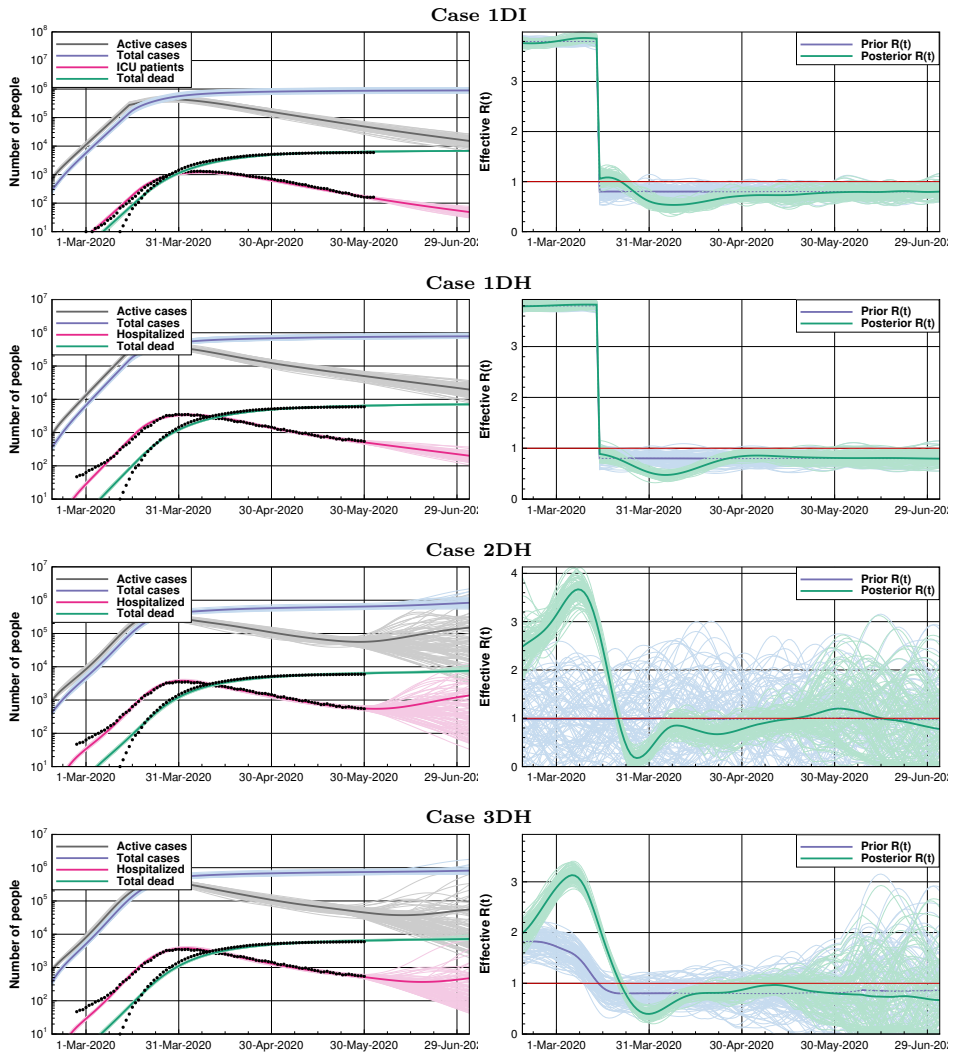

FIGURE 9. The Netherlands: The plots show the results from the Cases 1I, 1H, 2H, and 3H, from top to bottom. The left plots include model estimates of the number of hospitalized patients and dead, in addition to the total number of cases as well as the number of active cases. The right plots show the corresponding estimates of $R(t)$. The red thin line in the plots for $R(t)$ is an indication of the value $R(t) = 1$ for easier identification.

1DH. The evolution of $R(t)$ for Case 2DH, illustrated in Figure 9, is different from Cases 1DI and 1DH. Because of the constant prior $R(t)$, the posterior $R(t)$ tends to change gradually. When the actual $R(t)$ changes more rapidly, the model estimate will respond more slowly and exhibits a compensation of an $R(t)$ that drops below 0.5. The conditioning of the data in Case 2DH results in a peak of $R(t)$ around 3.8, roughly corresponding to the first guess for the initial R in Cases 1DI and 1DH.

The effectiveness of the interventions in reducing $R(t)$ is visible in the $R(t)$ staying below or close to the value of 1 from late March onward.

Case 3DH uses a prior $R(t)$ that mimics the estimate of $R(t)$ published by the RIVM[58]. Rather than having a step-wise change of $R(t)$, it assumes a gradual decrease from 1.8 to 0.8 over the first couple of weeks of the epidemic. Thus, it reflects an incremental impact of the measures. In this run, we allow the $R(t)$ to be influenced by the data assimilation by imposing a more significant uncertainty, although smaller than in Case 2DH. The resulting estimates of dead and intensive care patients are relatively accurate. The posterior $R(t)$ increases rapidly in the first weeks, suggesting that the prior estimate of an initial $R(t)$ was possibly too optimistic. See Figure 9 for the results.

8.3. **Summary of results for The Netherlands.** The results with the different prior assumptions for $R(t)$ illustrate that we can use different priors for $R(t)$ and still fit the data. Estimates of $R(t)$ tend towards values around three in the onset and just below one after the epidemic's peak. Slight changes of short duration in $R(t)$ do not appear to have much effect on the pandemic's evolution. The data-assimilation system seems to favor a gradual development of $R(t)$, and our results suggest that compensation effects occur when the actual $R(t)$ is less smooth.

The observations suggest a stabilization of COVID-19-related deaths of slightly over 6000, while our simulations indicate a total of approximately 8000 deaths. Because of the logarithmic scale, this difference, which is most vital for the second half of May, is not clear in Figure 9. We can think of two possible explanations for this overestimation of deaths by the model. The first possibility is that under-reporting causes the model to overestimate the number of deaths. Statistics Netherlands (CBS) reported a substantial number of excess deaths in late March and early April [16]. The total number of excess deaths, based on weekly estimates, adds up to 8765 over March 9th to May 10th. Based on available data for 2015-2019, we estimate the standard deviation of reported deaths in this period to be around 670. The number of reported COVID-19-related deaths over this period is 5597. RIVM suggests a correlation between the epidemic and the higher number of excess deaths in early April and states that the number of people dying is higher because not all people who die in the Netherlands have been tested for SARS-CoV-2 [59]. Hence, the under-reporting of the number of COVID-19 related deaths could explain why our estimates for the total number of deaths are higher than what the number of reported deaths would suggest.

Another possible explanation could be that over time, the Dutch health-care system is becoming more effective in its treatment of COVID-19. As the model parameters are assumed to be constant over time, the simulations cannot account for this gradual improvement. In the Québec Section 7.2, we presented a similar hypothesis to explain these simulation results.

With these possible causes of uncertainty impacting the estimated number of deaths, we consider our relatively high estimates as rough approximations of the actual number.

We have simulated additional cases where $R(t)$ increases to 1.0 when the schools re-opened in May. The number of COVID-19 deaths continues to grow in these simulations and does not stabilize as it would do in Cases 1DI and 1DH (Figure not shown). We can conclude from this that continued vigilance is of crucial importance for the case of The Netherlands. These results also clarify that further intensification

of contact between the Dutch population can result in an unstable growth of the number of COVID-19-related deaths.

9. **A case study for France.** We applied the SEIR-based ESMDA data assimilation method with data from the pandemics in France to assess the lockdown effect and run possible scenarios following the intervention.

The first occurrences of confirmed cases of COVID-19 in France were reported at the end of January 2020. However, biological a posteriori sample testing showed that there were isolated cases as early as November 2019. The first clusters of infectious patients were identified at the end of February in the Oise district (north of Paris area) and in the city of Mulhouse (east of France) at the beginning of March. A critical stage of public intervention, which recognizes the pandemics, started March 14th with the closure of schools and universities, and the enforcement of the first systematic restrictions on public meetings. A full lockdown was enacted on March 17th. This lockdown's primary goal was to avoid the saturation of intensive care units (ICUs) in hospitals (whose theoretical capacity was 4000 beds). Occupied ICU beds rose to 7,148 on April 8th before decreasing. Nonessential business and, to some extent, schools progressively re-opened on May 11th while enforcing physical distancing. Continuing with the lockdown-ending, restaurants, bars, and indoor sports facilities re-opened on June 2nd, together with lifting the restrictions on traveling beyond 100km. A few exceptions for the Paris region included hot spots (such as in the Val d'Oise district) that might still have a reproduction number higher than one.

9.1. **Observations and uncertain parameters.** Santé Publique France provides the data for France and makes them available on their daily updated dashboard [32]. These data have been officially published since March 2nd (see Figure 10). The published data are the cumulative deaths occurring at the hospitals, the current number of hospitalized patients for confirmed COVID-19 cases, and the current number of (COVID-19 related) patients in intensive care units. These numbers apply to the whole country. Although regional data are also available, we do not exploit them here. Also, one can access the daily number of patients released from the hospitals and the total number of confirmed cases. The dashboard also offered the number of deaths in care homes and other medical institutions (except hospitals) and confirmed cases in these institutions from mid-March. Given the lack of reactive compounds, PCR (polymerase chain reaction) testing was only applied to suspected patients and was initially systematically carried out in the hospitals. Hence the number of confirmed cases is by far not representative of the actual number of infected people.

Inaccuracies and delays in reporting deaths in care homes to Santé Publique France profoundly impact the accuracy of the total deaths observations. Moreover, we believe that, at least for France, we cannot model the deaths in care homes with the same SEIR-like dynamics as we do for the rest of the country. Care homes comprise confined spaces with an aged, weakened population with a very different $\hat{\mathbf{R}}$-matrix than the rest of the country. Thus we do not use these data in the conditioning. By contrast, the numbers of deaths at hospitals are likely more reliable. That is why, as opposed to the Québec case, the death number in care homes is not added to the death toll here.

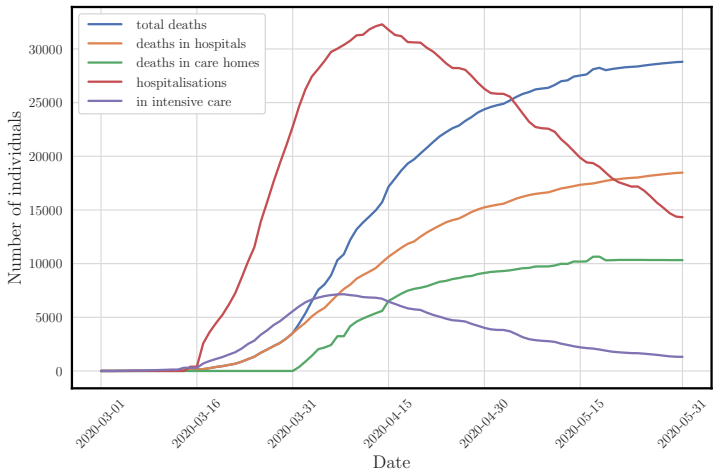

FIGURE 10. France: The plot shows official data curves for France from Santé Publique France up to May, 31.

9.2. **Exp. 1: Calibrated reference.** The first experiment (Exp. 1) serves as a reference for the application of ESMDA on the data for France. We have chosen the model parameters based on several sensitivity simulations, which yield a calibrated reference analysis. We focused on (i) the hospitalized patients and (ii) the accumulated deaths at hospitals. We have also experimented with the total deaths, which include the deaths in care homes. This approach yielded a significant misfit with the data, which is consistent with the above explanation that, for the French data, one cannot easily model the sum of hospitals and care homes deaths by using a global

Parameter	First guess	Std. Dev.	Description
t_0	February 16th	-	Start date of simulation
t_1	March 17th	-	Start date of intervention
t_2	May 11th	-	End of lockdown
R_1	3.5	0.20	$R(t)$ prior before intervention
R_2	0.65	0.20	$R(t)$ prior during lockdown
R_3	0.85	0.20	$R(t)$ prior after full lockdown
E_0	500	500	Initial Exposed
I_0	200	200	Initial Infectious
τ_{recs}	20	2	Recovery time severe cases
τ_{hosp}	6	0.5	Time until hospitalization
τ_{death}	7	1	Time until death
p_{f}	0.02	0.02	Case fatality rate
p_{s}	0.039	0.03	Hospitalization rate for severe cases
p_{h}	1	-	Fraction of $\mathbf{Q}_{\text{f}}$ that go to hospital

TABLE 10. France: The table gives a set of "calibrated" first-guess (i.e., prior) model parameters and their standard deviations used for France. p_{h} is set ot 1 to inform the model that care homes deaths are excluded from the death numbers. All other parameter settings are unchanged as compared to the ones given in Table 1.

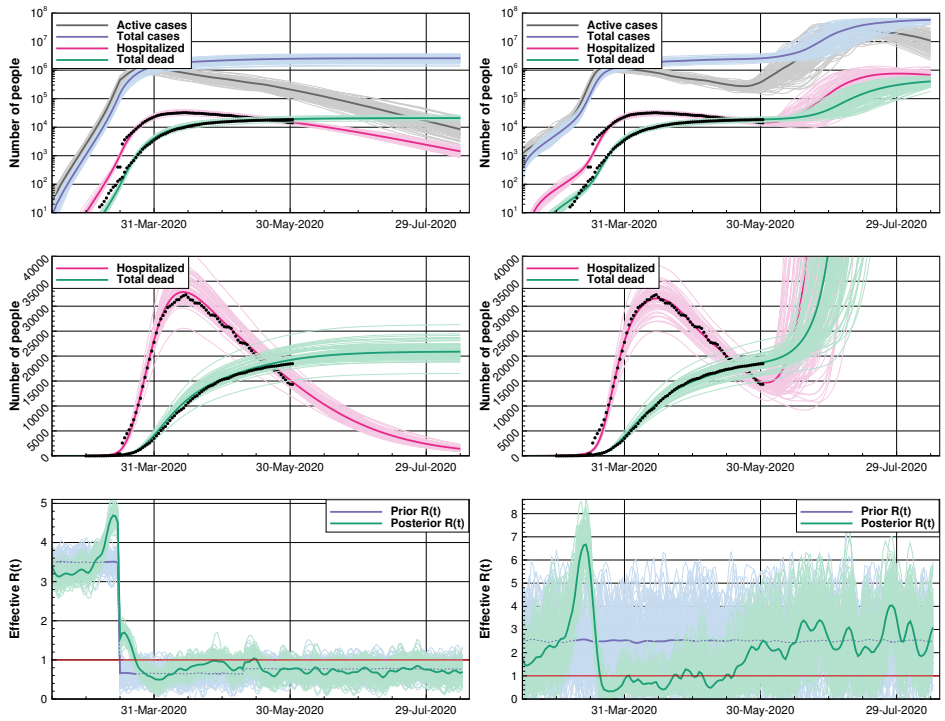

FIGURE 11. France: The figure shows the reference case (left) and case with an unknown intervention (right). The upper plots show the number of deaths at hospital, hospitalized patients, the total number of cases, and the currently active cases. The lower plots show the ensemble of effective reproduction number $R(t)$. The red thin line in the plots for $R(t)$ is an indication of the value $R(t) = 1$ for easier identification.

SEIR model. By contrast, we expect the reported deaths at hospitals and the hospitalized to be reliable and representative so that we choose a standard observation deviation of 1%, with a maximum error of 50 individuals. The experiments start on February 16th and end by late August.

Before the intervention and during the lockdown, the reproduction numbers follow INSERM (Institut national de la santé et de la recherche médicale), and Institut Pasteur studies [61]: $R_1 = 3.5$ with a standard deviation of 0.2 and $R_2 = 0.65$ with a standard deviation of 0.2. We also choose $R_3 = 0.77$ with a standard deviation of 0.2 for the reproduction number after the second intervention (re-opening). This latter reproduction number corresponds to a preliminary estimate from the French government on May 28th. Moreover we chose $\tau_{\text{recs}} = 20$ days, $\tau_{\text{death}} = 7$ days and $p_{\text{f}} = 0.02$ to reflect the situation in French hospitals. We set the initially exposed and infectious set to somewhat arbitrary values but with a huge standard deviation to account for the large uncertainty. The scaled age matrix $\hat{\mathbf{R}}$ is the default one, as explained in Section 2. Other key parameters have been left unchanged compared to the Norway case, except for a higher standard deviation for the case fatality rate

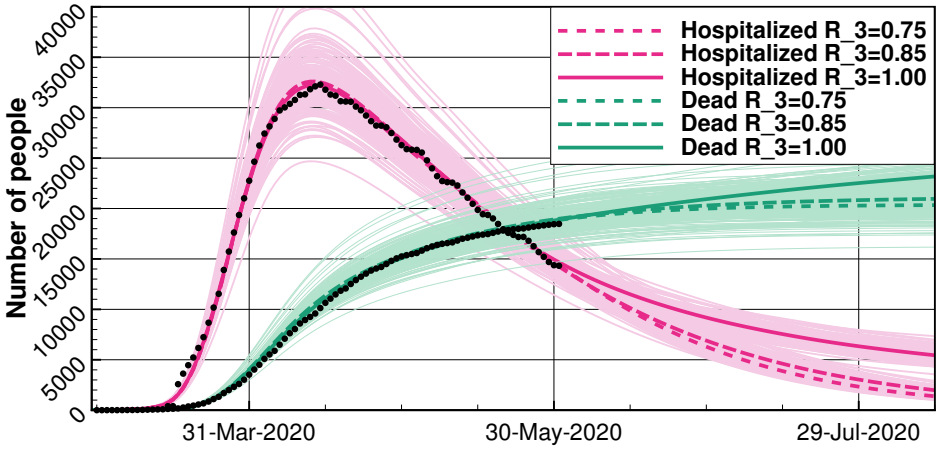

FIGURE 12. France: The plot presents forecasts of the reference case with three distinct scenarios after intervention setting the reproduction number to $R_3 = 0.75, 0.85, 1.00$. We have plotted the posterior values for the number of deaths at hospitals (green lines), and the hospitalizations (red lines), together with the first hundred realizations for each case.

and the hospitalization rate for severe cases. We present these numbers and their standard deviations in Table 10.

Figure 11 presents the assimilation results for the deaths at hospitals, hospitalized, and the reproduction number, as functions of time. The fit of our estimates to both sets of data is very good with thin posterior uncertainty. We could not obtain similar performance when using the total death numbers (i.e., including the deaths in care homes). However, the estimated reproduction number before intervention struggles to remain close to the consensus value of about 2.5 to 3.5 [61].

The mean number of cases as of May 31st is estimated to be 3.66 ± 0.57 million people, which is about $5.47\% \pm 0.8\%$ of the population. The fraction of the observed cases is estimated to be only 4.15% of the total number of cases.

9.3. **Exp. 2: No information on intervention.** In this second experiment (Exp. 2), we check if the data assimilation system can infer the lockdown period using the information from the observations. The prior reproduction number $R(t)$ is set constant in time at a fairly large first-guess value of 2.50 with a large standard deviation of 1.25, to make it as uninformative as possible. The other parameters are left unchanged, except for an adjustment of the initial Exposed and Infectious. The results for the time-dependent quantities are shown in Figure 11 (right panels). The fit to the observations is reasonably good for both the accumulated deaths and the hospitalized, although it is not as good as in Exp. 1. Furthermore, the algorithm detects a very significant change in reproduction number: $R(t)$ between March 15th and 20th and maintains it below one during the actual lockdown period. It shows that the lockdown was very effective in controlling the pandemic. The death toll explosion at the beginning of July (middle-right panel) is due to the high R_2 prior and is not realistic.

9.4. **Exp. 3: Forecast after lockdown.** In this last experiment, we are interested in projections after the lockdown. The exit from lockdown happened on May 11th, but it was very progressive, with a second stage on June 2nd. After this date, the evolution of the situation mainly depends on the scenario for the reproduction number, $R(t)$. We have investigated three scenarios:

1. In the first scenario, the reproduction number is kept at a low value of $R_3 = 0.75$. It is barely above the estimated value at the end of the lockdown, and it is compatible with the French government's preliminary estimate on May 28th.
2. In the second scenario, the reproduction number is raised to $R_3 = 0.85$, a somewhat ambitious targeted value.
3. In the third scenario, the reproduction number is raised to $R_3 = 1$, the threshold case.

For this experiment, the setup is the same as the one used in the previous experiments. The results are shown in Figure 12. The two stable cases ($R_3 = 0.75$ and $R_3 = 0.85$), where the virus spread remains under control, yield very similar results, with about 21,000 deaths at hospitals by the end of August. By the end of August, it is possible to estimate the total number of deaths using the empirical ratio of deaths at hospitals to the total deaths computed from the French data as of May 31st. This ratio is about 0.64, so we predict that the total number of deaths at the end of August to be $33,000$. Finally, the $R = 1$ scenario departs significantly from the stable scenarios with a significant resurgence in the death growth (25,000 deaths at hospitals by the end of August 2020). These scenarios plead for maintaining a discipline of social distancing for a few additional months. This result echoes the analog experiments performed for Norway (Fig. 2) and England (Fig. 5).

9.5. **Summary of results for France.** The health crisis in France is somehow different from other countries, so the first step in the series of experiments was to modify the model and data-assimilation parameters to reflect the french specifics better. In particular, we have chosen to exclude the deaths in care homes as these numbers are not as reliable and timely as the deaths in hospitals. With this calibrated reference setup, we have shown in Exp. 1 that it is possible to estimate the effective reproduction number $R(t)$, which confirms the other countries' conclusions. We relate the variation in $R(t)$ to the evolution of the situation and the government interventions. In particular, the lockdown effect is visible in the posterior, even when it is intentionally missing from the prior, as shown by Exp. 2.

Finally, as explained when summarizing the England case, one must interpret the results of Exp. 3 (forecast after the lockdown) qualitatively. Indeed, in the absence of observations, the forecast scenarios are mainly driven by the sole prior value of $R(t)$ and suffice to give a rough idea of the disease evolution. Better modeling would require updates of (i) the transmission rates between age groups as the population behavior is likely to change and of (ii) the case fatality ratio as the medical situation and the profile of sick individuals evolve.

10. **A case study for Brazil.** We have employed the compartmental model and the ensemble data assimilation described here to assess the possible future impacts on the relaxation of isolation policies and to illustrate the challenges involved in the prediction of SARS-CoV-2 spread in a vast country, mainly in the scenario of severe under-notification.

Many different sources (National Government, States Governors, and Cities Mayors) provide policy guidelines in a diverse country like Brazil. Therefore, it is a challenge to estimate the efficiency of, for instance, the distancing measures because of the different preventive policies adopted in each State [20]. For illustration, we have selected five different Brazilian states, one from each geopolitical region. In Figure 13, we show the evolution of SARS-CoV-2 in terms of the number of confirmed cases, deaths, and the mortality rate (CFR). The number of cases and deaths is still increasing in all the states, although at different growth rates and CFRs. For instance, there is a relatively slower increase of confirmed cases in the Rio Grande do Sul (green curve). But not in Pará (purple curve), where the confirmed cases have increased at approximately the same rate since the first death. It is also remarkable that the CFR tends to grow in Pará (purple curve) while it tends to decrease in time in Goiás (red curve).

Furthermore, there is a severe under-notification of both confirmed cases and deaths. The former is due to the lack of sufficient tests. The latter is related to a potential association of deaths, which might have been caused by COVID-19, to other respiratory diseases. Even though substantially more hospitalizations due to respiratory infections have been reported than in other years, it is unclear what the actual number of admissions is due to COVID-19 [56].

Therefore, due to the challenge of accounting for such diverse dynamics in a country-wise fashion, we have modeled only one Brazilian State in this work, the State of São Paulo. This State concentrates the most significant number of confirmed cases and deaths in Brazil and presents an increasing trend in the disease's evolution. We focus on the spread of SARS-CoV-2 under the quarantine level and forecast how to improve the situation by applying more restrictive policies. Also, due to the reported data reliability issue, we investigate the uncertainty in the predictions and their intrinsic impact in the policy-making.

For our data assimilation experiments, we make the following assumptions:

- Even after adopting a specific policy, its efficiency is uncertain. We assume that the policy results in a lower reproductive number, $R(t)$, but it can still be higher than one (meaning the disease will still spread). As of mid-June, the epidemic is still growing, as is supported by a recent study [37, 46].
- The observation error on the number of deaths is primarily due to the under-notification. Thus, we will consider observation errors up to 20%, although they might be even more substantial.
- Due to under-notification, we will not assimilate the number of confirmed cases, where there is an even higher uncertainty around the reports [10].
- There are no observations of the number of hospitalizations due to the lack of reliable sources.

10.1. **Observations and uncertain parameters.** The first death due to COVID-19 in Brazil happened on March 20, 2020, in the State of São Paulo. Immediately after, on March 21, 2020, the State Government declared state-wide isolation. All commerce and non-essential services were to close on March 24, 2020. The intervention included all age classes, without any distinction, from closing kindergartens to any other establishment or social event that could result in large gatherings. The isolation policies were mostly followed and were still in effect when we ran the experiments.

Parameter	Initial value	Std dev	
t_0	March 10th	-	Start date of simulation
t_1	March 23rd	-	Start date of interventions
t_2	June 1st	-	Start date of prediction
E_0	656.0	65.6	Initial Exposed
I_0	164.0	16.4	Initial Infectious
R_1	4.0	0.4	Prior $R(t)$ during spinup
R_2	1.0	0.1	Prior $R(t)$ during interventions
R_3	1.0	0.1	Prior $R(t)$ in prediction phase
p_h	0.2	-	Ratio of fatally sick hospitalised
p_f	0.065	0.0065	Case fatality rate (CFR)

TABLE 11. Brazil: The table gives the parameter values used in the SARS-CoV-2 simulations for São Paulo, Brazil. All other parameters were kept unchanged as compared to the ones given in Table 1.

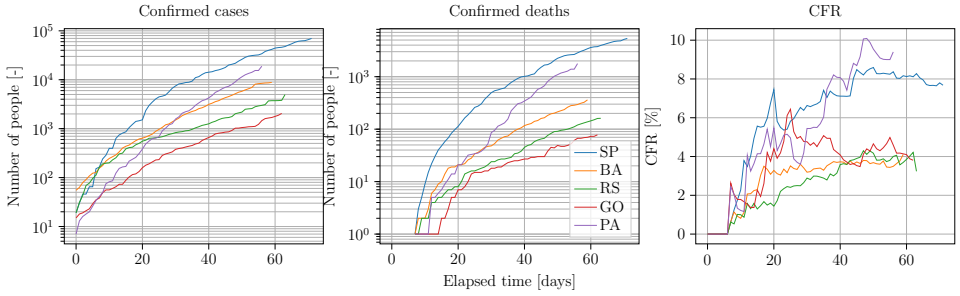

FIGURE 13. Brazil: The plots illustrate the evolution of SARS-CoV-2 in terms of the number of confirmed cases, deaths, and mortality rate (CFR), starting from the day of the first reported death due to COVID-19. The reported states are São Paulo (SP), from the Southeast geopolitical region; Bahia (BA), from the Northeast; Para (PA), North; Rio Grande do Sul (RS), South; and Goiás (GO), Mid West. Each state represents a different evolution of the disease. We have shifted the curves in time to correspond to the first confirmed case.

Table 11 presents the parameters used for the São Paulo State simulation. The small fraction of the fatally ill hospitalized is due to the limited ICU capacity and the high number of deaths associated with severe respiratory diseases happening outside hospitals. The reproductive number before the first intervention is around four, consistent with the numbers reported by WHO. The scaled age matrix $\hat{\mathbf{R}}$ is the default one, as explained in Section 2. Ideally, like the other parameters that were kept unchanged, we should have considered a more representative set of parameters. The issues with data gathering, as already mentioned, prevent a more accurate scenario representation.

10.2. **Exp. 1. Assessment of distancing measures.** In the first experiment, we assess the impact on the disease evolution, considering two different scenarios in terms of the efficiency of distancing measures. The first scenario, an unstable

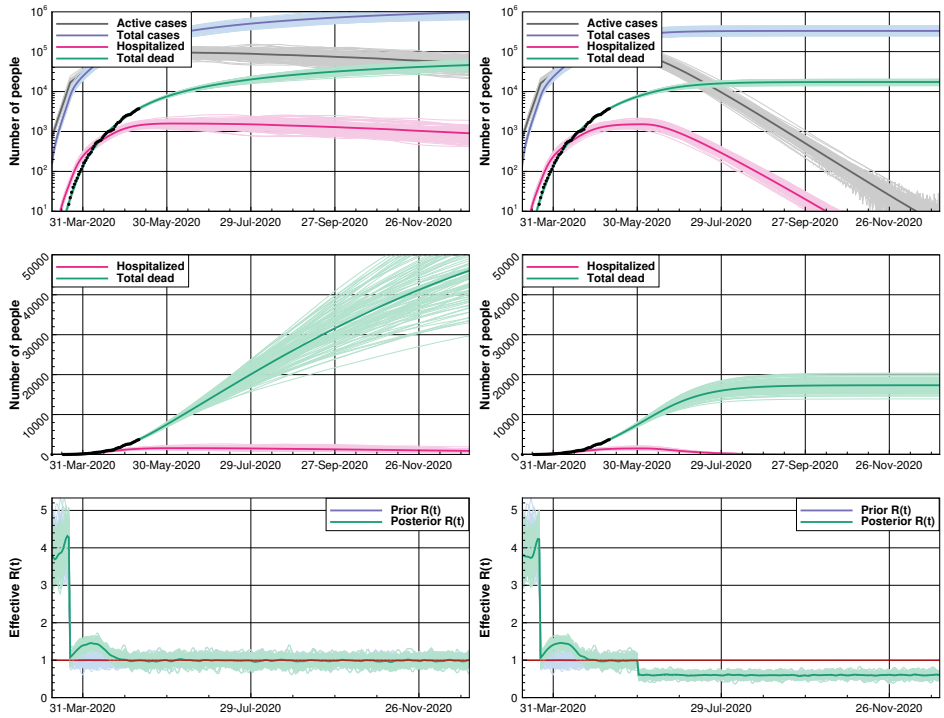

FIGURE 14. Brazil: The figure shows the simulation of the SARS-CoV-2 evolution in the São Paulo Brazilian State. In the left plots, we show a neutral case where the reproductive number $R(t) \sim 1.0$ with a standard deviation of 0.1 for the prediction. The right plots present a stable situation where $R(t) \sim 0.6$ with a standard deviation of 0.06, after a second intervention. The red thin line in the plots for $R(t)$ is an indication of the value $R(t) = 1$ for easier identification.

scenario, follows exactly the parameters depicted in Table 11. Figure 14 illustrates the results of this scenario in the left plots. Since we allow $R(t)$ to be higher than one, the number of deaths continues to grow until the end of 2020. In the second scenario, we keep the reproduction number at a low value of 0.6 after June 1st (simulating more efficient distancing measures, resulting in a stable scenario). We show the results of these data assimilation experiments in the right plots of Figure 14. As seen from the plots, this measure could result in the ending of COVID-19 deaths by 2020. The results illustrated in the left plots of Figure 14 show that, despite the isolation measures, we were not capable of reducing $R(t)$ below one. Thus, the mitigation measures were probably not as restrictive as, e.g., in Europe. However, this situation can change in the future with the implementation of stricter rules. Thus, it might be possible to reach a situation similar to the one illustrated in the right plots Figure 14.

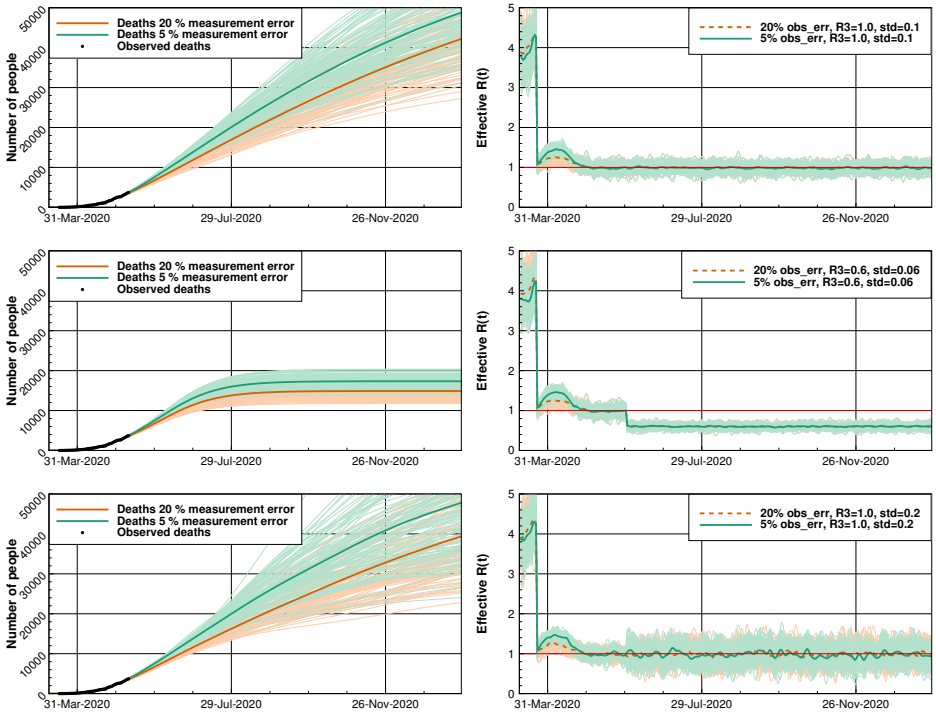

FIGURE 15. Brazil: The plots compare simulations considering relative observation errors of 5% and 20%. The upper and lower rows represent a neutral scenario with $R_3 = 1.0$. They differ in the standard deviations or $R(t)$ (0.1 and 0.2). The middle plots show a stable situation with $R_3 = 0.6$ and a standard deviation of 0.06, following a second assumed intervention. The left plots show the deaths, and to the right, we present $R(t)$ using different measurement errors.

10.3. **Exp 2. Impact of uncertainty in the policy-making process.** Next, we assess how uncertainty related to the effectiveness of interventions and under-reporting can impair policy-making. By setting different values and standard deviations on R_3 (the reproductive number used in the prediction period), we can simulate the impact of intervention uncertainty. We can examine the errors due to under-reporting by evaluating the prediction uncertainty due to observation errors.

We repeat Exp. 1, both stable and unstable scenarios, but changing the observation error from 5% to 20%. The first and second rows of Figure 15 illustrate the results. Note that 20% errors can even be considered low according to some investigations that indicate a possible association of deaths reported to be due to other severe respiratory diseases to COVID-19 [56]. Also, we apply a standard deviation of 0.2 after June 1st for the neutral case, shown in the third row of Figure 15. This example illustrates the uncertainty around the actual measures assuming no changes in the current policies.

It is incredibly challenging to construct, for instance, a plan to loosen up the isolation in the mid-long term, if the uncertainty around $R(t)$ is high. We support this conclusion by the broader spread of the posterior ensemble in the neutral scenario ($R(t) \sim 1.0$) where we assume a 0.2 standard deviation on $R(t)$, and compared with the situation with a standard deviation of 0.1 (first and third rows in Figure 15). The wide range of plausible predictions represents a higher risk in any decision-making process. Note also that the stable scenario results in a narrower spread of the posterior ensemble (center plots). While this can lead to less risk, the more substantial observation errors still imply a difference of thousands of deaths, as shown in the center panel's expected mean curves.

10.4. **Summary of results for Brazil.** The evolution of SARS-CoV-2 can vary considerably depending on the geopolitical region or State of large countries, mainly if vastly different policies are followed by each region, like in Brazil and unlike the European countries investigated in this paper. Additionally, even studies on individual states may not be entirely representative if we consider separate cities. Nevertheless, the significant uncertainty in the predictions, as in any other data assimilation or uncertainty quantification study, should be remedied by acquiring additional and more accurate data. More tests and a proper reporting of deaths, hospitalizations, and other vital data are of utmost importance to reduce the parameters' uncertainty, mainly on the reproduction number $R(t)$, hence in the predictions.

The unstable cases represent likely scenarios if the government does not enforce policies to control the spread of the disease. Together with the stable examples, the results illustrate how the data assimilation system can support decisions to mitigate catastrophic outcomes by raising awareness and providing a tool for decision support.

11. **A case study for Argentina.** The SARS-CoV-2 virus arrived in Argentina around late February/early March after several citizens returned from the summer holidays, mainly from Italy, Spain, and the United States. Argentina detected the first positive case on March 3rd. Considering the large spread in Europe and the extreme situation in Italy and Spain, a group of epidemiologists advised the President on the need for social-distancing measures to slow down the virus's spread. On March 20th, an extensive early lockdown in the whole country was established by the national government, at a time when the number of tested positive cases was relatively small, i.e., about 100. This lockdown included strict control of the external country boundaries and between states, closing all social activities and commercial businesses except shopping for food and accessing the health service. Individuals were not allowed to leave their houses except for buying food. After about 40 days, in late April, the government relaxed the lockdown, permitting more work activities every week.

11.1. **Observations and uncertain parameters.** During March, the tests for SARS-CoV-2 were collected around the country and evaluated at a single place, Instituto Malbran, in Buenos Aires city. Because of this, there was a significant delay in the detection of cases of about five days. Furthermore, they were somewhat limited at first but increased later, which may change the fraction of undocumented cases and affect the statistics. In late April, several authorized laboratories spread around the country started to analyze the tests. The Health Ministry produces a

daily official report with the number of positive cases and the number of accumulated deaths. These reports are available on the Health Ministry web page. We used these data to conduct the data assimilation experiments of the virus transmission in Argentina. There is still some capacity available in the health system up to the end of May (Health Ministry web page), and we consider the number of deaths due to COVID-19 to be rather precise. The standard deviation of the observed deaths was four at the start of the pandemic and increased up to six the last examined day, which contains an accumulated number of deaths of 539. Regrettably, the information on hospitalized people is not well updated on the available official databases. Therefore we have been unable to use this information in the data assimilation experiments.

Since the first detected cases were around early March and because of the early lockdown, we conducted the data assimilation experiments for Argentina with a smaller number of initial exposed and infected individuals than the other countries, 40 and 10 respectively, and standard deviation of 8 and 2. We kept the rest of the parameters to be the same as for the other experiments (Table 1). We set the prior values for the effective reproductive number, $R_1 = 3.7$, between March 1st and March 20th, $R_2 = 1.0$ during the lockdown between March 21st and April 27th, and $R_3 = 1.5$ after the relaxation of the lockdown.

11.2. **Exp 1. Hindcasts over the data period.** We conducted two experiments. One, Case DC, in which we conditioned both on the accumulated deaths and the reported number of cases, and Case D, in which we only conditioned on the deaths observations. Figure 16 presents the results from Case DC in the left plots and Case D in the right plots, respectively. In the upper plots, we show the reported number of accumulated cases and deaths as dots. For the number of cases, we assumed that this number represents only 15% of the actual number of cases. Since there is no observational evidence in Argentina about the reported total cases rate, we assumed the estimate from Chinese data by [45]. This undetected rate is in coherence with the rest of the experiments for other countries in this work. The thick lines correspond to the posterior density mean of the total and active cases, the number of hospitalized individuals, and the number of accumulated deaths. The estimated mean parameters exhibit only slight deviations from the proposed prior values, except for the case fatality rate, which we compute to be $p_f = 0.011$.

In the lower plots of Figure 16, we present the mean (line) and uncertainty, represented by the ensemble, of the effective reproductive number $R(t)$ as inferred with ESMDA. Overall, the estimated $R(t)$ in the pre-lockdown period is 3.7 and relatively close to the assumed prior mean. After the lockdown, $R(t)$ descends abruptly to about 1.1. After that, we see a slight increase in $R(t)$ to over 1.5 following the lockdown's relaxation. Note that the estimated $R(t)$ in Case DC, corresponds to a minimal number of reported cases and deaths. The early lockdown effectively diminished the increase in the number of new infections and the number of hospitalized and fatalities. Two months after the first reported COVID-19 case, the number of reported cases was still relatively small, 4783.

The early lockdown did reduce $R(t)$ below one for a short period, but the quarantine's relaxation led to values over one. Thus, the number of cases is still increasing as of the end of May, and the outbreak's peak is ahead. We partly attribute the lower reduction of $R(t)$ compared to other countries to South America's social inequality distribution. The lockdown was not effective in vulnerable neighborhoods

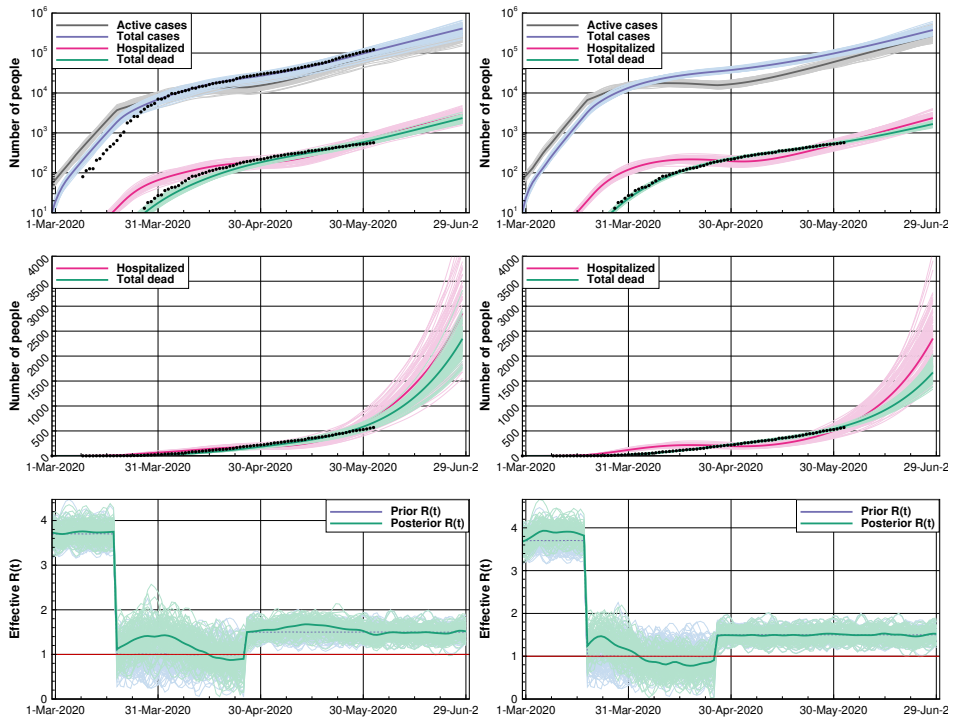

FIGURE 16. Argentina: The dots are the observations. In Case DC(left plots) both the accumulated deaths and the estimated number of cases were conditioned on, while in Case D(right plots) we only conditioned on the total number of deaths. The red thin line in the plots for $R(t)$ is an indication of the value $R(t) = 1$ for easier identification.

because of the living conditions. In those disadvantaged neighborhoods, many individuals live in the same small house, and most of the daily activities are conducted outside the house so that strict social distancing measures are not viable. The number of persons detected with COVID-19 in these neighborhoods has increased dramatically with time. Meta-population or agent-based models focusing on the neighborhood scale are required to model these patterns, particularly the spread's spatial distribution.

Since the reported number of cases by the health system may be underestimated, we conducted an experiment (Case D) in which we use only the number of deaths as observations. In this case, we obtain a closer match to the observed number of deaths. On the other hand, the inferred number of reported cases is larger than the reported cases. This result may indicate that the number of tested cases was lower than 15% of the total cases. Still, it may also be the result of the delay in the evaluation of the tests. As shown in the middle plots of Figure 16, the posterior mean of the accumulated deaths shows a closer match to the data. The effective reproductive number resulting from conditioning only on the number of accumulated deaths shows significant differences to the results from Case DC. The estimated mean exhibits a lower level of variability. It descends to values below one

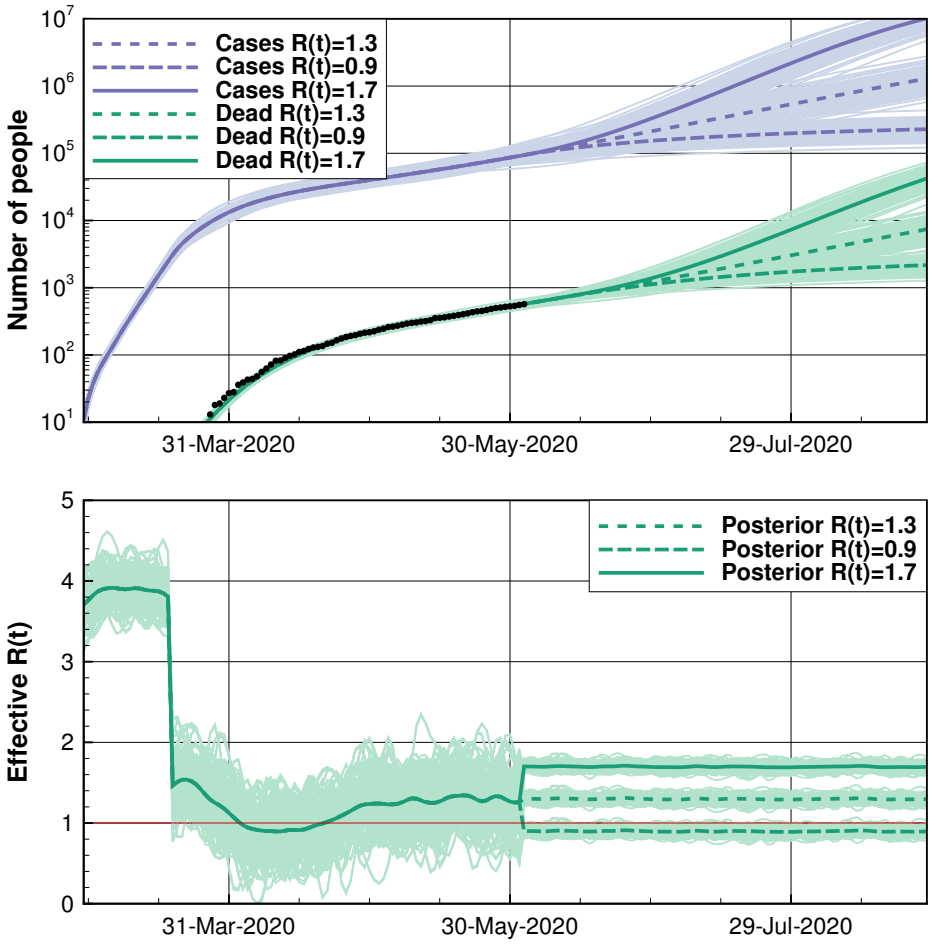

FIGURE 17. Argentina: Same as Figure 16 but for an experiment (Exp. 2) focused on a probabilistic prediction in which three scenarios, with different effective reproductive numbers imposed from June 1st, $R(t) = 1.7$, 1.3, and 0.9. The red thin line in the plots for $R(t)$ is an indication of the value $R(t) = 1$ for easier identification.

during the lockdown period, and then it increases to 1.4-1.5 after the relaxation of the lockdown. This behavior produces a flattening of the hospitalized number compared to Case DC.

11.3. **Exp 2. Prediction experiments.** To evaluate the system as a tool for decision making, we conducted a third experiment focused on probabilistic forecasts. As in Case DC and D, we started with the prior of $R(t) = 3.7$ until the introduction of interventions on March 20th, where the prior for $R(t)$ is reduced to $R(t) = 1.3$ until June 2nd. We then modeled three potential scenarios with $R(t) = 1.7$, 1.3, and 0.9, representing a further relaxation of the lockdown, a continuation of the current partial lockdown and social distancing recommendations, and a more

restrictive scenario with a return to the original lockdown measures. Because these are probabilistic scenarios, we reduced the standard deviation of R(t) to 0.05. This prediction experiment is in a similar vein to the one conducted in other countries. On the date (June 1st), when the prediction starts, Argentina has the highest reproductive number of the examined countries in this work. The upper plot of Figure 17 shows the evolution of the number of deaths and reported cases while the lower plot presents the effective reproductive number. In these scenarios, we only condition the model on the observed numbers of deaths, which are more reliable than the cases. The probabilistic predictions differ substantially in the three scenarios. On August 1st, the predicted mean number of deaths is 6800, 2800, and 1600, respectively, while the forecasted estimates of reported cases are 302000, 73000, and 27000 for the three scenarios.

11.4. **Summary of results for Argentina.** Because of the long time lag of death number compared to the reported cases, the estimated $R(t)$ using the reported cases has more recent information about the infections. Indeed, Case DC captures a change in the trend of the last infections with an increase in $R(t)$ to values of 1.6-1.7, which is still not detected in Case D (Figure 16). In this sense, reported cases may have some valuable information for producing forecasts. However, we must decide on the correct trade-off between different observation types since the reported case numbers are less reliable than death numbers. The top left panel of Figure 16 shows that ESMDA underestimates the number of deaths at the beginning of the epidemic, and then it overestimates them. This result is coherent to what we find in other countries like Quebec and The Netherlands. One interpretation is that the healthcare system improved the treatment of the COVID-19 disease. A given number of cases then translates into a smaller number of deaths, and the case fatality rate reduces. Another possibility is that older people's death rate was high because they were mainly infected in hospitals, while later on, the virus spread to younger persons. Hence, the mean age of the infected cases diminished with time, and so does the case fatality rate. If more data become available, future studies may evaluate these hypotheses.

The predicted number of accumulated cases and deaths are only indicative values. They are scenarios with a fixed effective reproductive number. In Argentina, we expect a continued increase in the effective reproductive number because of the relaxation of the lockdown measures. Hence, we anticipate approaching the pessimistic scenario. Given that the first COVID-19 outbreaks were in a few cities and then the virus spread to other cities, a metapopulation model with regional compartments and considering mobility between regions would probably provide a more precise long-range prediction of the pandemic evolution in Argentina. The country frontiers are closed (flights and car entrance to Argentina are prohibited). Thus, in the second stage, the virus should spread internally between cities due to all the interconnections. In contrast, during the first stage, the virus's spread was mainly produced by imported cases, i.e., passengers of international flights.

12. **A case study for The US.** Retrospective analysis carried out by the United States Center for Disease Control and Prevention (CDC) suggests sustained community spread of the novel SARS-CoV-2 virus occurred before detecting the first non-travel related case on February 26th in California. An analysis of RNA sequences from early patients suggested that a single lineage of the virus was imported directly or indirectly from China and began circulating between January 18th and

February 19th, 2020, followed by several importations of SARS-CoV-2 from Europe [43]. The largest number of flights between the United States and Europe typically go through airports in the country's northeast region, particularly in and near New York City. A lack of testing led to some uncontrolled spread early on as individual states began their first mitigation measures.

Individual states began implementing initial mitigation measures in mid-March to varying degrees, with additional measures typically being implemented in serial progression through early April. The timing and severity of restrictions varied significantly among states with no national standard enforced by the Federal Government. We can see an impact of these measures in Figure 18 by the initial drop in baseline mobility, measured from cell phone data and compiled by the Institute for Health Metrics and Evaluation (IHME) [48]. This data gives the percent decrease in mobility from the specific region's baseline, based on pre-pandemic travel behavior, and therefore accounts for the differences in the usual travel distances required for essential or typical trips in a more rural v.s. a more urban location. However, no distinction can be made between relatively safe, driving to a trailhead for a hike, or somewhat unsafe, grocery or restaurant, related travel. On March 16th, the Center for Disease Control (CDC) and the newly formed White House Corona Virus Task Force announced voluntary guidelines dubbed "15 days to slow the spread." These guidelines were later extended through April. They advised taking only essential trips combined with some social distancing measures but focused on following individual state governments' specific instructions. As a result, each state took its path with different degrees of restrictions and enforcement. The response of the people of each state varied as well. Protests against stay-at-home orders sprang up, people sometimes filled up closed beaches, and some churches defied closure orders.

With this backdrop of heterogeneity across the US, we carried out a case study of four states representing a cross-section of the country: New York, California, Alabama, and North Carolina. New York was chosen for two reasons. First, it was the hardest-hit state due to the population density in New York City and its strong connection to Europe through air travel. Secondly, based on mobility data, New York achieved a 66% reduction from baseline mobility, one of the most substantial decreases observed in the United States. California was chosen for similar reasons, with an achievement of a 55% reduction in mobility. It was also one of the quickest states to have locked down at least parts of it. On the other hand, it also experienced massive protests and defiance of beach closures. Alabama and North Carolina are chosen because of lower reductions of baseline mobility of 38% and 47%, respectively, and their contrasting approaches to government-mandated measures despite similar demographics. North Carolina implemented mitigation measures earlier than Alabama, with schools closed and gathering restrictions implemented on March 14th, compared to March 19th. Stay at home orders began on March 30th in North Carolina and April 4th in Alabama, with North Carolina keeping restrictions in place until May 8th as compared to April 30th in Alabama. The measures differed in severity to some degree, with essential businesses reduced to 20% capacity in North Carolina v.s. 50% in Alabama. At the end of May, the percent of tests returning positive in North Carolina was about 6.7% compared to approximately 11.3% in Alabama.

12.1. **Observations and uncertain parameters.** The data we use comes from the IHME group at the University of Washington. They have pre-compiled the historical data in a large CSV file, which we use to extract data for the 50 states:

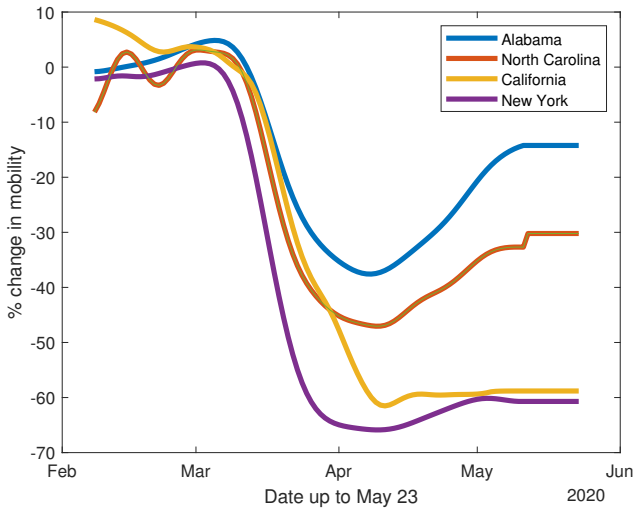

FIGURE 18. US: Mobilites for each of the states considered

Param	Prior	Std. Dev.	Description
T_s	20/2-2020	-	Start Date
E_0	50(NY,CA),10(AL),20(NC)	10(NY,CA,NC),2.0(AL)	Initial Exposed
I_0	10(NY,CA),1(AL), 2(NC)	5(NY,CA,NC),1.0(AL)	Initial infectious
p_f	0.18(NY), 0.009(CA,NC,AL)	0.001	CFR

TABLE 12. US: The parameters used in our experiments, the different values for the reproductive number at intervention steps are described in the sections for each case. Values without state indications are the same for all states. Any parameters not listed are the same as in Tab 1.

http://www.healthdata.org/covid/data-downloads. Their primary data source is John's Hopkins University. We use the accumulated number of deaths, the current number of COVID hospitalizations, and the accumulated number of positive cases. However, we do not condition our ESMDA steps on the number of cases due to the lack of early testing in the United States as well as considerable uncertainty in the percentage of cases it represents. Our population data for each state comes from the 2018 U.S. Census Bureau estimates. We use an age grouping similar to the Norwegian case with 11 age groups.

We assimilated the number of accumulated deaths and current hospitalizations using three different scenarios. First, we set the prior for the reproduction number to be $R(t) = 1$ with a large uncertainty to see how the ESMDA estimates $R(t)$. In all of the cases, the prior of $R(t) = 1$ does not appear to be representative of the actual value until around late March, which is consistent with the mobility data approaching their respective minimums around that time for all four states.

Secondly, we assume that every citizen followed the CDC's guidance issued officially on March 16th, causing an immediate drop in the reproduction number to $R(t) = 1$. And thirdly, we implemented a step-down decrease from a maximum $R(t)$

value to an intermediate value and then to $R(t) = 1$. We choose the date for setting the prior $R(t)$ to its intermediate value, to be the first date when the mobility data reaches half of its maximum reduction-value, typically around March 17th. After that, it follows a transition to $R(t) = 1$ when the mobility is ten percentage points away from the maximum reduction value, typically around March 23rd.

We have chosen the prior R values at the first two steps so that the forecast ensemble mean closely follows the number of hospitalizations and deaths. We set the final prior of $R(t)$ to one to gauge the success of the mitigation efforts for the given state. The time-varying value for $R(t)$ is updated in the assimilation step and implies that when $R(t) > 1$, after the final step down, the state has not met its mitigation goals. Similarly, when $R(t) < 1$, the state may have exceeded its mitigation goals.

12.2. **Exp. 1: Sensitivity analysis on $R(t)$.** In this case, we take a prior of $R(t) = 1$ and increase the uncertainty in $R(t)$ to have a standard deviation of 1 to see how the data may inform the reproduction number with the results shown in Figure 19. In this case, we condition on deaths and hospitalizations and observe how $R(t)$ differs from a base prior value of $R(t) = 1$ over time. Both New York and California show a sharp rise and then decrease with New York having the largest values for $R(t)$ before the implementation of measures. Both states issued similar rules at similar times but have very different urban structure. People in New York City often use crowded public transit, while in Los Angeles and other cities in California, personal vehicles tend to be the primary mode of transportation. North Carolina has initially slower spread but is comparable to Alabama in the early days with larger oscillations around $R(t) = 1$ after late March than California or New York. While this analysis is informative to give us an idea of how the virus was spreading in time, we also want to condition on priors that represent specific scenarios *intended* by lockdown measures. We consider these types of priors in the next two cases.

12.3. **Exp. 2: Immediate response to federal guidelines.** In this case, we impose a drop in the prior in $R(t)$ to take place when the first federal guidelines were issued on March 16th with the results shown in Figure 20. Clearly, from the mobility data in Figure 18, this drop is unrealistic as mobility decreased slowly from mid-March to mid-April. However, it is essential to establish that the assimilation of deaths and hospitalizations detects this decrease, as, for all cases, $R(t)$ remains well above one until the mobility indeed decreases enough. New York shows a sharp decrease in $R(t)$ in late March while the other three states lag until the beginning of April. Interestingly, in North Carolina, the value for $R(t)$ remains relatively high until around April 15th, peaking on March 31st despite that the mobility has decreased significantly during the weeks before. A reason might have been an influx of people from New York and the northeast with second homes in North Carolina. In all cases except New York, we see oscillatory behavior in posterior for $R(t)$ during the assimilation window. These correspond to increases and Decreases in hospitalizations and deaths weeks later. In the US, borders between states cannot be closed, and enforcement of restrictions can be difficult, leading to time-varying spreading rates. The sharp drop in mobility in the state of New York is apparent in this analysis. What is clear is that New York had the most consistent and rapid response in terms of controlling the reproduction rate.

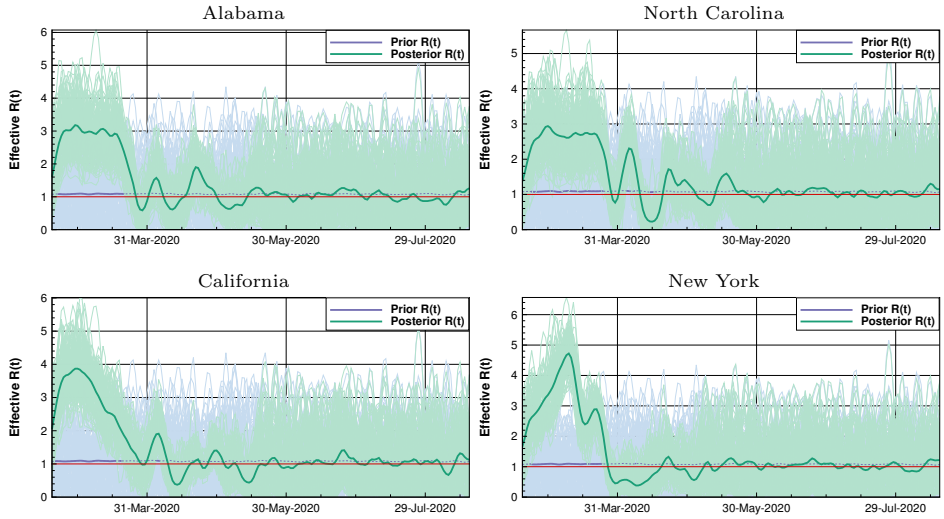

FIGURE 19. US Case 1: Large uncertainty in $R(t)$ The red thin line in the plots for $R(t)$ is an indication of the value $R(t) = 1$ for easier identification.

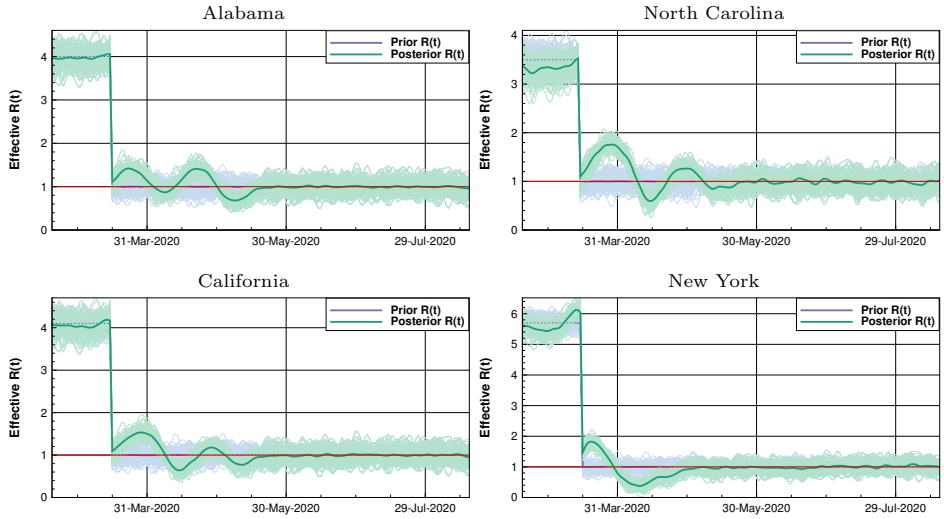

FIGURE 20. US Case 2: Assuming voluntary federal guidance was immediately observed by the citizens on March 16th. The red thin line in the plots for $R(t)$ is an indication of the value $R(t) = 1$ for easier identification.

12.4. **Exp. 3: Intermediate step down to $R = 1$.** In this case, we impose two drops in the value for R to better account for the slow decrease shown in the mobility data with the results shown in Figure 21. We choose the first two values for $R(t)$ so that the initial forecast mean closely follows the data on deaths

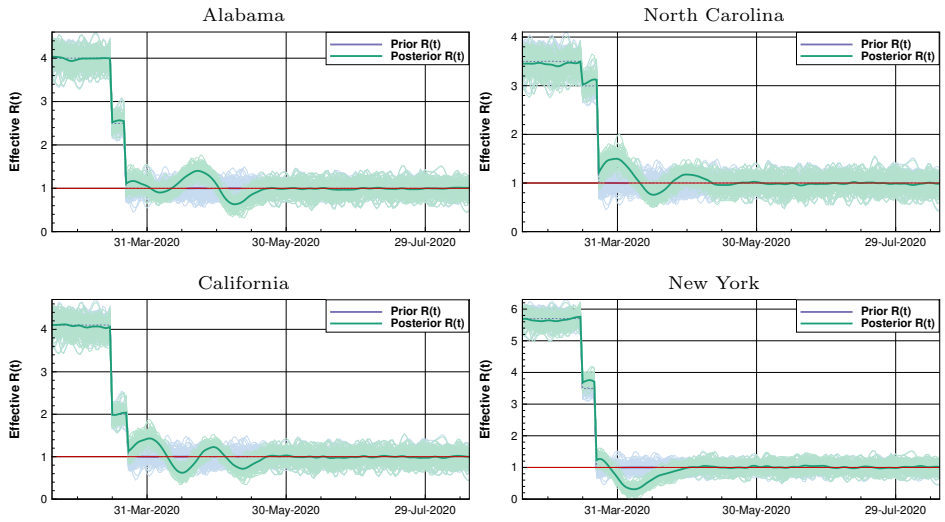

FIGURE 21. US Case 3: A gradual step down in $R(t)$ with the first and intermediate values chosen so that the prior mean closely follows the data until the time for which $R(t)$ is guessed to be one. The red thin line in the plots for $R(t)$ is an indication of the value $R(t) = 1$ for easier identification.

and hospitalizations until the step down to $R(t) = 1$. In this way, we can gauge how well a state has managed to control the spread of the virus, compare the projection of the "goal spread" (achieving $R = 1$), and the projection influenced by the data. It is immediately clear from the effective $R(t)$ ensemble plots that neither Alabama, California, or North Carolina was able to achieve constant control of spread. Alabama has some of the least reduction in mobility with the longest lag time in reaction to the federal guidance. It is notable that before the intervention, the R value for Alabama fits reasonably well at $R(t) = 4$ while for North Carolina, $R(t)$ is lower at just below 3.5. This result is despite the two states having a similar geographical distribution of the population. Further, Alabama typically had $R(t)$ above one with a dip below one shortly after the minimum mobility. The difference in $R(t)$ values between California and New York before intervention is also striking. New York fits well with values as high as $R(t) = 5.7$ while California hovers around $R(t) = 4$ in the early days. Both states have large Urban areas with large populations; however, the population density in New York urban centers, particularly NYC, is typically much higher. There are differences in travel methods as well, i.e., public versus private transportation, as discussed in Case 1. With all states but New York, we see oscillations around $R(t) = 1$, which we may link to specific spreading events. California and North Carolina experienced a series of anti-lockdown protests. The most massive protest in California, on May 1st, included more than 1000 people. It was one out of a series of protests numbering in the 100s, in the days before and after May 1st. There were two protests in North Carolina, 100 people on April 14th, and 300 people on April 21st. New York also experienced two demonstrations, one low-risk protest in Albany on April 22nd, where a parade of cars drove through Albany's capitol park. Another on May 1st, where hundreds

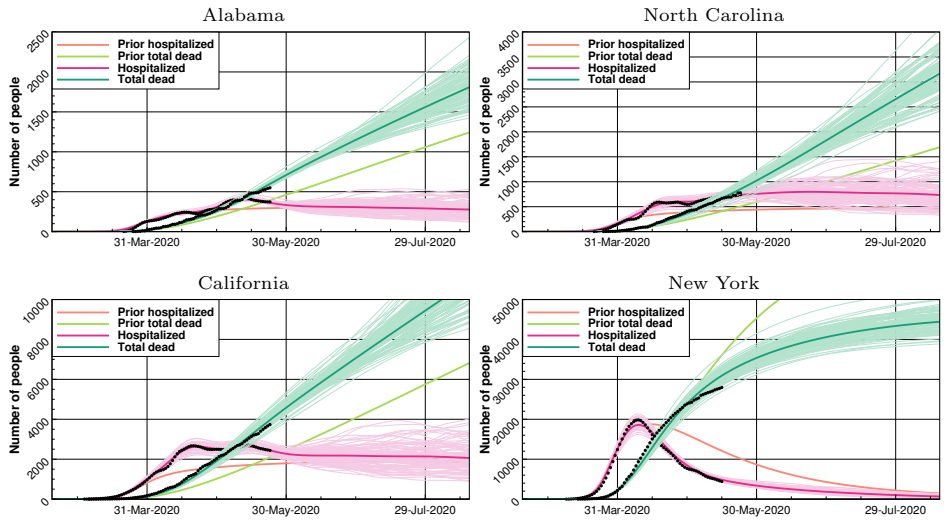

FIGURE 22. US: Forecasts for Case 3: Here we also show the prior forecast mean to highlight the difference between the idealized scenario and after analysis. In this case, only New York was able to achieve a trajectory that predicts fewer deaths than the idealized scenario. Furthermore, New York also is the only state in this study to make an end of the outbreak by late August.

of protesters converged on Commack, Long Island. However, New York achieved sustained $R(t)$ below $R = 1$ and then stabilized around $R = 1$ as in the last half of May with the data going up to May 23rd. By and large, New York citizens were far more cooperative than in many other states, but this could also be an artifact of the sheer number of infections and deaths New York suffered in the early days of its outbreak.

12.5. **Summary of results for The US.** We consider the step-down runs to be the most realistic in terms of forecasting for each state. In Figure 22, we show both the mean of the forecast, chosen to closely match the data on hospitalization and deaths, as well as the analysis projections before transitioning to $R(t) = 1$ as our prior. The projections past May 23rd, the last day of available data, in this case, rely on $R(t) = 1$ for the rest of the year. It seems that California, North Carolina, and Alabama will not do better than the situation represented by the forecast mean in terms of accumulated deaths. However, New York is on track to doing better than the $R(t) = 1$ prognosis after the last intervention due to its sustained $R(t) < 1$ through mid-April. If the other three states manage to maintain some $R(t) \approx 1$, the variations in the early days of the lockdowns do lead to more deaths than otherwise would have occurred. Furthermore, because $R(t)$ remained above or around $R = 1$, one did not achieve a rapid decrease in cases. The time to reach zero hospitalizations has increased significantly, and projects to continue through Summer 2020. On the other hand, New York is currently on track to gain control of the outbreak.

13. **Discussion.** We have demonstrated an approach for parameter estimation and model calibration in an age-stratified SEIR model to model and predict the SARS-CoV-2 epidemics. By assigning uncertainties to all poorly known parameters and characterizing them by specifying appropriate prior distributions, we can sample large ensembles of parameter realizations. The model predictions originating from the prior ensemble of parameters constitute our previous knowledge of the epidemic evolution, before data are taken into account, and have typically significant errors. An ensemble data-assimilation method allows for estimating both static model parameters and the time-varying effective reproductive number. Our ensemble data-assimilation approach, conditioning the model on observations, implies constraining the prior ensemble and compute an updated ensemble of parameter estimates that constitute the posterior solution. The posterior solution agrees well with the measurements and has a significantly lower uncertainty. As such, the use of ensemble data assimilation allows for efficient model calibration, and we propose it as an ideal tool for calibrating any epidemic model to observations.

13.1. **A brief synthesis of the results per cases.** By modeling the SARS-CoV-2 epidemics in several countries with vastly different predicaments and resources to handle the virus's spread, we have learned a lot about the modeling and data-assimilation system's capabilities predicting the epidemic. We restate here succinctly the experiments and the key results obtained in each of the cases. Otherwise we refer to the extended case-specific summaries in their respective sections and the overall discussion in Sections 13.2–13.7 below.

For Norway, we proved that it is possible to estimate $R(t)$ until about two weeks before the last observed value when conditioning on fatalities and hospitalization data. Gradually re-opening the schools and kindergartens led to only a small increase in $R(t)$, which remained less than one since the interventions were effective among the adult population. In June, the epidemic had nearly ended in Norway, thanks to effective implementation and public support of the interventions.

The results from England confirmed our system's capability to estimate $R(t)$ until two weeks before the last observation. Projections under three different scenarios of weak, mild, and strict containment policies predicted the pandemic's status in England three months ahead. We imposed the different scenarios by prescribing $R(t)$, but we left other influential parameters, such as the fatality rate, unchanged, resulting in a possible overestimate of projected deaths.

For Quèbec, we obtained an overall good fit to observed data, although there is a tendency to overestimate fatalities towards the end of the observation period. Like for England, we relate this overestimation to improvements in the health care systems' response to the pandemic situation, which we do not model using time-invariant model parameters. We also demonstrated a significant short-term prediction skill using the latest estimated $R(t)$-ensemble during the prediction.

For The Netherlands the system proved robust against the choice of vastly different priors for $R(t)$, with estimates converging towards values around three during the onset of the pandemic and slightly below one in the intervention period.

In France the reported deaths in care homes are not reliable, and we have not conditioned on them in the experiments. The results for France confirms our system's capabilities to estimate $R(t)$ to timely represent the implementation of measures imposed by the French authorities.

For Brazil the study focuses on the state of Sãu Paulo. Similarly to the previous cases, we estimate realistic values of $R(t) \simeq 1.0$ over the data period, and we

prescribed different future $R(t)$ scenarios and run predictions. We assessed the impact on the policies decision making in face of uncertainty on the interventions' effectiveness and data under-reporting.

For Argentina the fit to data during the data period is excellent initially but tends toward an overestimate at the end, as already seen in the some previous cases. Additional experiments assimilating infectious cases, albeit little reliable, help catch sudden changes in $R(t)$ undetected when only incorporating reliable death observations.

We studied four states in the USA, and examined different ways of defining the prior $R(t)$ to account for the implemented interventions, e.g., using a sudden change in the prescribed $R(t)$ or using in a step-wise decrease reflecting a gradual introduction of measures. This latter case more closely matched the mobility data and resulted in more realistic predictions. Results indicated that only the state of New York is on track to achieve controlling the pandemic. These experiments also showed that we could estimate an $R(t)$ that reflects time variability in the interventions' effectiveness.

13.2. **ESMDA convergence and ensemble size.** It is possible to interpret the ESMDA method as a sequence of tempered steps in the parameter space, starting from the prior and ending up at the posterior solution. The technique applies a linearization at the beginning of each pseudo-time step to determine the update direction. This linearization implies that, by reducing the step size, one expect the accuracy to improve. Therefore, irrespective of the ensemble size, we need to use sufficiently short steps to ensure the convergence of the ESMDA method. Previous studies using ESMDA have indicated that the technique requires about 16–32 steps to converge with reasonable accuracy. In sensitivity experiments (see link below), we could not visually observe any difference in the results using 32 and 64 MDA steps. Thus, we have used 32 MDA steps in all the experiments presented in this paper.

An ensemble size of 100 realizations will provide a reasonable solution, but there will be significant sampling errors. A sensitivity study indicated that using more than 5,000 ensemble members will not lead to visible changes in the results. Thus, we have used 5,000 realizations for all the final cases presented in the paper.

We present a more detailed discussion regarding the convergence properties of ESMDA in Sections A.1 and A.2 of the Appendix. A document with supplementary plots, illustrating the convergence of ESMDA, is available from the Github repository: https://github.com/geirev/EnKF_seir/blob/master/doc/sensitivity.pdf.

13.3. **Model initialization.** Given a prior set of model parameters, e.g., the defaults in Table 1, we can initialize the model by setting the start date, the initial number of infectious I_0, and exposed E_0. These numbers interact with the reproduction number $R(t)$ for the initial period between the start date and the interventions' introduction. As the basic reproductive number's value is somewhat uncertain, we have allowed for uncertainty in $R(t)$ in the initial period. Given the start date, we have some freedom to balance I_0, E_0, and $R(t)$ in the initial period, such that we match the observed onset and numbers of deaths and hospitalized. There are different combinations of these parameters that will give an equally good fit to the observations. However, the critical point is to have a realistic model estimate at the

onset of interventions. The initial period serves mostly as a spinup for the model and the ESMDA. The data will, after that, constrain the system.

In several of the cases presented here, we used ESMDA to estimate these initial parameters in a recursive process to find a good initial model state. After that, we initialized the data-assimilation experiments with these prior values using smaller uncertainties but added uncertainty to the remaining model parameters.

13.4. **Estimation of $R(t)$.** A strength of the method is that it is also possible to estimate the effective reproductive number as a function of time. In the cases in which we choose a constant prior $R(t)$ (for The Netherlands, France, and The US) and a high standard deviation to reflect the uncertainty, the observations steer the posterior estimate of $R(t)$ to more realistic values that reflect measures put in place to contain the pandemic. In this manner, it is possible to use data assimilation in a hindcast mode to analyze $R(t)$ and test the validity of different scenarios for its temporal evolution.

All experiments that suddenly lower the prior value of $R(t)$ exhibit a compensation effect shortly after this change. By implementing this reduction over two smaller steps, as in Case 3 for The US, this effect is less but still present. Using a more gradual $R(t)$ as in Case 3DH for The Netherlands results in a smoother $R(t)$ but does not eliminate the tendency of the system to produce a low value of $R(t)$ following a peak.

Since it takes about two weeks before new cases evolve into hospitalizations or deaths, it is possible to constrain or estimate $R(t)$ until about two weeks before the latest observation. For the last two weeks and into the future, $R(t)$ will stay close to its prior value. Consequently, it is possible to fit the model to long-term epidemics experiencing multiple waves and peaks — some of the US cases, which have more erratic behavior, illustrate this property.

13.5. **Predicting the future.** The driving parameter of an SEIR model is the effective reproductive number. Thus we need to know the future value of $R(t)$ to make reliable predictions. The model predictions depend exponentially on $R(t)$. As such, it is maybe more useful to run scenarios with values of $R(t)$ leading to either unstable $R(t) > 1$, neutral $R(t) = 1$, and stable evolution $R(t) < 1$ of the epidemics. These scenarios allow for predicting the severity of the pandemics under different regimes of interventions in different countries. We expect that societal and cultural diversity between countries also affect $R(t)$: some countries may need to implement stricter interventions than others, to keep $R(t)$ below one.

Assuming a country has implemented sufficient measures to reduce $R(t)$ below one, the system can realistically simulate the future decline of the epidemics under the current interventions. With time and more data, we also expect to use the system for quantifying the impact of various opening measures on $R(t)$. We already know that the reopening of schools and kindergartens in Norway only led to a slight increase in $R(t)$.

Many countries do not test all deaths for the presence of SARS-CoV-2. Consequently, the number of reported deaths is highly uncertain and tends to underestimate the actual number of COVID-19 deaths. In France, the COVID-19 deaths in care homes are reported with delay and are much less accurate than the number of deaths in the hospitals. Consequently, we multiply these numbers by an adjustment factor in the assimilation to correct for this bias. For some other countries, e.g., The Netherlands, Québec, England, and The US, the estimated number of deaths

is larger than the reported number. Some explanations can be, e.g., the health care system may be improving with time, the virus may be getting less lethal, or the age distribution of patients may shift to younger patients. This overestimation also influences the forecasts. The various projections of the number of fatalities reflect this uncertainty. Thus, we need to interpret the predicted COVID-19 deaths as rough estimates.

13.6. **Short term predictability.** The application for Québec also assessed the quality of making short-term predictions (two weeks ahead), using the current estimates of $R(t)$. We found a significant forecast skill, suggesting that it may be beneficial to use model predictions for evaluating the need and impact of implementing new interventions. Also, the short-term forecasts can be useful for planning the ICU capacity.

13.7. **Which data to condition on?** In this work, we have mostly conditioned the SEIR model on the accumulated number of deaths and the current number of hospitalized. In some examples, we have also included derived or adjusted data of the accumulated number of cases (e.g., for England).

An issue is that the basic SEIR model does not include hospitalized and deaths (see, e.g., the flow diagram in Figure 1). Thus, these observations are not straightforward to use, unless we assume to know (as we did here) the fraction of cases that die or go to the hospital. The registration of COVID-19 patients in the hospital may not be very accurate and using the number of ICU patients, as is done for The Netherlands, the time until hospitalization τ_{hosp}, and the hospitalization rate p_{s} must be adjusted to reflect the difference between the two groups of patients.

The accuracy of the reporting of hospitalized and deaths varies in time and from country to country. In The Netherlands, they don't always report the COVID-19-related deaths in care homes, and the same holds for deaths when the cause is unknown. In France, they report the care homes' deaths with significant and erratic delays. The inaccuracy in the available data may have caused the Netherlands' fit to be less good than for some of the other countries. In France, the data fit is very good, but not so much in the early weeks of the pandemics since the French reporting system was not fully operational until late March. The fit to the data in England is relatively good, but there are no data for hospitalized COVID-19 patients at the onset of the pandemic. These differences between countries form a likely reason for the difference in the fit of the data with the model, which is especially prominent in the early weeks of the pandemic.

Also, many patients in care homes for the elderly become severely ill and die without going to the hospital. When we assign uncertainties to the fraction of patients going to the hospital, we can obtain an excellent match to the observations of the number of deaths. If the case fatality rate p_{f} and the hospitalization rate p_{s} are unknown, it is impossible to constrain the SEIR model with the observed numbers of deaths and hospitalized. However, we can condition the SEIR model on the observations by using prior estimates of these parameters.

The observed number of cases would be useful if they were unbiased observations since they relate directly to the SEIR model variables. In theory, we could obtain reliable case data by testing the whole or an unbiased sample of the population. With such data, we would not only much better constrain the model, but we could even accurately estimate the case fatality rate p_{f} and the hospitalization rate p_{s}. When health institutes provide an official estimate of the number of cases, this is

typically a rough estimate of the total infected population that is always larger than the number of positive tests and hence biased. Selective and too little testing inevitably creates a bias in these data. Nevertheless, our data-assimilation approach permits us to estimate the total number of infected as part of the assimilation output when we assimilate death and hospitalized data. This posterior estimate of the number of infected, and its uncertainty, may then be used to evaluate an eventual bias in the raw data (see e.g., the England case).

We do not know the total number of infected, so we do not know the basic reproduction number. We see the number of deaths and the number of hospitalized, but since we do not know the actual number of infected, the severely and fatally ill fractions are both unknown. Thus, to obtain unbiased data, we should test extensively across the entire population.

With the above in mind, in some countries, we included the case data in the conditioning to illustrate our approach's capabilities. By doing so, we assumed that the case data only detect a specified fraction of the actual number of cases, and we appended to them a considerable uncertainty. The effect is that we might obtain a more realistic figure of infections (e.g., in England). In some cases, e.g., for Norway and The Netherlands, we instead used the case data to constrain the initial conditions to provide predictions in agreement with the assumed number of infections.

14. **Summary.** We have demonstrated the use of an ensemble-based data assimilation system for modeling and predicting the development of the SARS-CoV-2 outbreak in 11 countries and states. We introduced a variant of the SEIR type model and obtained realistic predictions of the epidemics, with uncertainty estimates, by conditioning the model ensemble on observed deaths and hospitalization numbers. The system clearly illustrates the inherent instability in infectious illnesses where the disease's development primarily depends on the effective reproductive number, $R(t)$. The data assimilation system allows for online and automated estimation and monitoring of $R(t)$. Unfortunately, there is a lag of two-three weeks between infections and registered hospitalization and deaths. However, we handled this challenge by using a smoother approach when assimilating data to constrain the model simultaneously on all data present in a time window. This lag causes that, when one observes a too high value of $R(t)$, it may already be too late to avoid another round of significant restrictions on society. Therefore, it is mandatory to reopen society slowly, step by step, and allow enough time to assess each step's impact before initiating the next one. In this respect, the system will provide an online estimation of any changes to $R(t)$, although with the delay related to the disease's development time. As a natural follow-on of this study, the authors currently include a spatial dimension in the model to turn it into a meta-population model composed of connected nodes, each represented by an SEIR. This modification will allow us to resolve and analyze the geographical distribution of the virus.

Acknowledgments. G. Evensen was supported by internal funding from NORCE (and his family was generous enough to allow and encourage him to work long days and nights given the urgency of this work). J. Amezcua, A. Carrassi, and A. Fowler were supported by the NERC award NCEO02004. F.C. Vossepoel was supported by a Delft Technology Fellowship. M. Pulido was supported by ANPCYT under grant CORR01 COVID-19 Federal. Both C. Jones and C. Sampson were supported by the US Office of Naval Research under grant N00014-18-1-2204. CEREA is a

member of the Institute Pierre-Simon Laplace (IPSL). We are grateful for having received extensive and constructive reviews that have led to a significant manuscript improvement.

Appendix A. **ESMDA algorithm.** As explained in [24, 23], ESMDA solves the traditional ensemble smoother (ES) update equations using a predefined number of steps. In each recursive update, ESMDA inflates the measurement errors to reduce the impact of the measurements. With correctly chosen inflation factors, the ESMDA update precisely replicates the standard Ensemble Smoother (ES) in the linear case. When the model or observation operators are nonlinear, it turns out that the use of multiple short update steps reduces the errors and improves the solution as compared to using one long update step in ES.

It is easiest to derive ESMDA by using a tempering of the likelihood function in (16) [50], which leads to a recursive minimization of a sequence of $n = 1, N_{\text{mda}}$ cost functions, [64, 24],

$$\begin{aligned}\mathcal{J}(\mathbf{x}_j^{n+1}) &= \left(\mathbf{x}_j^{n+1} - \mathbf{x}_j^n\right)^{\mathrm{T}} \left(\mathbf{C}_{xx}^n\right)^{-1} \left(\mathbf{x}_j^{n+1} - \mathbf{x}_j^n\right) \\ &+ \left(\mathbf{g}(\mathbf{x}_j^{n+1}) - \mathbf{d} - \sqrt{\alpha^{n+1}}\mathbf{e}_j^n\right)^{\mathrm{T}} \left(\alpha^{n+1}\mathbf{C}_{dd}\right)^{-1} \left(\mathbf{g}(\mathbf{x}_j^{n+1}) - \mathbf{d} - \sqrt{\alpha^{n+1}}\mathbf{e}_j^n\right),\end{aligned} \tag{18}$$

where we evaluate $\mathbf{C}_{xx}^n$ at the nth iterate $\mathbf{x}^n$, and we must have the following condition on the sequence of measurement-error inflation factors α^n:

$$\sum_{i=1}^{N_{\text{mda}}} \frac{1}{\alpha^n} = 1. \tag{19}$$

In each ESMDA step we must minimize $j = 1, \ldots, N$ cost functions, one for each of the N members of the ensemble.

Similarly to the derivation of the standard EnKF analysis update, we obtain the recursive update equations for ESMDA given by Eqs. (24) in the algorithm below. The update direction is computed based on a linearization around the prior realizations of each update step.

To compute the ESMDA solution, we start by sampling the initial ensembles from

$$\mathbf{x}_{j,0} \sim \mathcal{N}(\mathbf{x}^{\mathrm{f}}, \mathbf{C}_{xx}). \tag{20}$$

Then we integrate the model according to Eq. (14) to obtain the prior ensemble prediction for the first ESMDA step,

$$\mathbf{y}_{j,0} = \mathbf{g}(\mathbf{x}_{j,0}), \tag{21}$$

and we compute recursively the following for each iteration $n = 0, \ldots, N_{\text{mda}} - 1$:

1. Construct the sample covariances $\overline{\mathbf{C}}_{yy}^n$ and $\overline{\mathbf{C}}_{xy}^n$, from the ensemble realizations $\mathbf{y}_j^n$ and $\mathbf{x}_j^n$.
2. Sample the measurement perturbations
$$\mathbf{e}_j^n \sim \mathcal{N}(\mathbf{0}, \alpha_{n+1}\mathbf{C}_{dd}). \tag{22}$$
3. Generate the perturbed measurements
$$\mathbf{d}_j^n = \mathbf{d} + \mathbf{e}_j^n. \tag{23}$$
4. Compute the update
$$\mathbf{x}_j^{n+1} = \mathbf{x}_j^n + \overline{\mathbf{C}}_{xy}^n \left(\overline{\mathbf{C}}_{yy}^n + \alpha^{n+1}\mathbf{C}_{dd}\right)^{-1} \left(\mathbf{d}_j^n - \mathbf{y}_j^n\right). \tag{24}$$

5. Rerun the model to obtain the updated prediction

$$\mathbf{y}_j^{n+1} = \mathbf{g}(\mathbf{x}_j^{n+1}). \tag{25}$$

Repeat this procedure until $n = N_{\text{mda}} - 1$, which results in the ESMDA solution for $\mathbf{x}_j$ and $\mathbf{y}_j$.

As in ES, the update direction is computed based on linearization around each update step's current estimate. Thus, we can interpret the ES update as taking one long Euler step of length $\Delta\tau = 1$ in pseudo time τ, while in ESMDA we take a predefined number of shorter Euler steps of step length $\Delta\tau_i = 1/\alpha_i$ that satisfy Eq. (19) [24]. By updating the linearizations around each recursive estimate, ESMDA reduces the impact of nonlinearity and leads to improved solutions.

The ensemble size determines the statistical convergence of ensemble methods, which estimates the mean and standard deviation from the posterior ensemble. The central limit theorem tells us that the estimated ensemble-mean is a normal distribution sample with a standard deviation proportional to $\sigma/\sqrt{N}$. Here σ is the true standard deviation of the posterior estimate, and N is the ensemble size. Thus, assuming $\sigma = 1.0$, then using 5000 realizations gives $1/\sqrt{5000} \approx 1.4$. Thus, running multiple experiments with different random seeds would lead to estimates sampled from a normal distribution with a standard deviation of approximately 0.014 (or about 1.4 % of the true standard deviation of the posterior distribution).

The number of MDA steps determines how well we are resolving the nonlinearity of the problem. In a linear case, one step will give the exact solution (with infinite ensemble size). In the nonlinear case, we can interpret the method as computing a sequence of Euler steps that reduce the impact of nonlinearity. Using ESMDA, we will likely improve the estimate indefinitely by increasing the number of steps. With given convergence criteria, the required number of steps will depend on the problem's nonlinearity. Typical applications run 4–16 steps since an additional number of steps does not "significantly" improve the results to the extent that can justify the additional computations. Here we ran 32 steps because we could afford it with the simple model, and we could not visually see any difference in the results if we increased the number of steps further.

A.1. **Sensitivity of ESMDA to the number of steps.** We have performed sensitivity experiments to examine the convergence properties of the ESMDA algorithm. We expect that the solution's accuracy will improve with the number of steps until a certain level where there is nothing more to gain. [24] examined the convergence of ESMDA for a simple nonlinear scalar case and obtained minimal improvement after 16-32 steps. Similar results were found here when examining the posterior solutions for deaths and hospitalized using ESMDA with 1, 2, 4, 8, 16, 32, 64, and 128 steps, and similarly for the corresponding estimates of $R(t)$. From visual inspection of the results, it is hard to justify more than 16 steps. We decided to use 32 steps in all the simulations presented in this manuscript to ensure convergence with a margin. A document with supplementary material, including the results of the sensitivity tests, is available from the Github repository: https://github.com/geirev/EnKF_seir/blob/master/doc/sensitivity.pdf.

A.2. **Sensitivity of ESMDA to the size of the ensemble.** We have examined the convergence of ESMDA concerning the ensemble size. ESMDA, being based on a Monte Carlo algorithm, means that we can always improve the solution by increasing the ensemble size. However, we need to decide on a tradeoff between

ensemble size, N, and the number of ESMDA steps, N_{mda}, due to limitations on computing power. While the number of ESMDA steps impacts the algorithm's actual convergence towards the correct solution, the ensemble size impacts the precision of the statistical estimate (i.e., how accurate we describe the error) of the final solution. We find it most important to first converge to the correct physical solution and, after that, run with an as large as possible ensemble to reduce the sampling errors. We examined the results using different ensemble sizes and two different random seeds. Even using $N = 100$ ensemble members, the posterior predictions are consistent with the data and very similar to the cases with larger ensemble sizes. There is a visual difference in the results from changing the random seed, which is clearer from the estimated $R(t)$. When using 1000 or 5000 realizations, there is still a significant difference in the estimated $R(t)$. Between 5000 and 10000 realizations, we could hardly see any difference in the parameter estimates or predictions. Thus, we use 5000 realizations in all the simulations in this paper. We have included the results in the supplementary material found at: https://github.com/geirev/EnKF_seir/blob/master/doc/sensitivity.pdf

A.3. **Uniqueness of the solution.** Finally, we will comment on the degrees of freedom in the model parametrization versus the amount of information in the data. There are dependencies between several of the uncertain parameters. E.g., suppose we specify too few initial infected. In that case, this may be compensated for by a higher value of $\mathbf{R}(t)$ in the first period. If the model predicts too many infectious, it is still possible to match the accumulated deaths and hospitalizations by reducing the fractions p_{s} and p_{f}. Thus the system appears to be under-determined. However, by specifying prior distributions for the uncertain parameters, we ensure one unique posterior solution, and we obtain it by running the model with the ensemble of estimated parameters as input.

REFERENCES

[1] S. I. Aanonsen, G. Nævdal, D. S. Oliver, A. C. Reynolds and B. Vallès, Ensemble Kalman filter in reservoir engineering – A review, *SPE Journal*, **14** (2009), 393–412,

[2] S. Abrams, The analysis of multivariate serological data, in *Handbook of Infectious Disease Data Analysis*, CRC Press, 2019.

[3] J. L. Anderson and S. L. Anderson, A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts, *Mon. Weather Rev.*, **127** (1999), 2741–2758.

[4] E. Armstrong, M. Runge and J. Gerardin, Identifying the measurements required to estimate rates of COVID-19 transmission, infection, and detection, using variational data assimilation, *Infectious Disease Modelling*, to appear.

[5] M. Asch, M. Bocquet and M. Nodet, *Data Assimilation. Methods, Algorithms, and Applications*, Fundamentals of Algorithms, 11, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2016.

[6] L. M. A. Bettencourt, R. M. Ribeiro, G. Chowell, T. Lant and C. Castillo-Chavez, Towards real time epidemiology: Data assimilation, modeling and anomaly detection of health surveillance data streams, in *Intelligence and Security Informatics: Biosurveillance*, Lecture Notes in Computer Science, 4506, Springer, 2007, 79–90.

[7] J. C. Blackwood and L. M. Childs, An introduction to compartmental modeling for the budding infectious disease modeler, *Lett. Biomath.*, **5** (2018), 195–221.

[8] M. Bocquet and P. Sakov, An iterative ensemble Kalman smoother, *Q. J. R. Meteorol. Soc.*, **140** (2014), 1521–1535.

[9] M. Bocquet and P. Sakov, Joint state and parameter estimation with an iterative ensemble Kalman smoother, *Nonlin. Processes Geophys.*, **20** (2013), 803–818.

[10] C. Brasil, *Estimativa de Casos de COVID-19*, 2020. Available from: https://ciis.fmrp.usp.br/covid19-subnotificacao/.

[11] R. Buizza, M. Milleer and T. N. Palmer, Stochastic representation of model uncertainties in the ECMWF ensemble prediction system, *Q. J. R. Meteorol. Soc.*, **125** (1999), 2887–2908.
[12] G. Burgers, P. J. van Leeuwen and G. Evensen, Analysis scheme in the ensemble Kalman filter, *Mon. Weather Rev.*, **126** (1998), 1719–1724.
[13] H. Cao and Y. Zhou, The discrete age-structured SEIT model with application to tuberculosis transmission in China, *Math. Comput. Modelling*, **55** (2012), 385–395.
[14] A. Carrassi, M. Bocquet, L. Bertino and G. Evensen, Data assimilation in the Geosciences: An overview on methods, issues and perspectives, *WIREs Climate Change*, **9** (2018), 50pp.
[15] CBS, *Bevolkingspyramide*, Statistics Netherlands (CBS), 2020. Available from: `https://www.cbs.nl/nl-nl/visualisaties/bevolkingspiramide`.
[16] CBS, *Nearly 9 Thousand More Deaths in First 9 Weeks of COVID-19*, Statistics Netherlands (CBS), 2020. Available from: `https://www.cbs.nl/en-gb/news/2020/20/nearly-9-thousand-more-deaths-in-first-9-weeks-of-covid-19`.
[17] N. K. Chada, M. A. Iglesias, L. Roininen and A. M. Stuart, Parameterizations for ensemble Kalman inversion, *Inverse Problems*, **34** (2018), 31pp.
[18] Y. Chen and D. S. Oliver, Ensemble randomized maximum likelihood method as an iterative ensemble smoother, *Math. Geosci.*, **44** (2012), 1–26.
[19] Y. Chen and D. S. Oliver, Levenberg-Marquardt forms of the iterative ensemble smoother for efficient history matching and uncertainty quantification, *Comput. Geosci.*, **17** (2013), 689–703.
[20] COVID-19 in Brazil: "So what?", *The Lancet*, **395** (2020).
[21] A. A. Emerick and A. C. Reynolds, Ensemble smoother with multiple data assimilation, *Comput. Geosci.*, **55** (2013), 3–15.
[22] R. Engbert, M. M. Rabe, R. Kliegl and S. Reich, Sequential data assimilation of the stochastic SEIR epidemic model for regional COVID-19 dynamics, *Bull. Math. Biol.*, **83** (2021).
[23] G. Evensen, Accounting for model errors in iterative ensemble smoothers, *Comput. Geosci.*, **23** (2019), 761–775.
[24] G. Evensen, Analysis of iterative ensemble smoothers for solving inverse problems, *Comput. Geosci.*, **22** (2018), 885–908.
[25] G. Evensen, *Data Assimilation. The Ensemble Kalman Filter*, Springer-Verlag, Berlin, 2009.
[26] G. Evensen, The ensemble Kalman filter for combined state and parameter estimation: Monte Carlo techniques for data assimilation in large systems, *IEEE Control Syst. Mag.*, **29** (2009), 83–104.
[27] G. Evensen, Formulating the history matching problem with consistent error statistics, *Comput. Geosci.*, to appear.
[28] G. Evensen, Sampling strategies and square root analysis schemes for the EnKF, *Ocean Dynamics*, **54** (2004), 539–560.
[29] G. Evensen, Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, *J. Geophys. Res.*, **99** (1994).
[30] G. Evensen, P. N. Raanes, A. S. Stordal and J. Hove, Efficient implementation of an iterative ensemble smoother for data assimilation and reservoir history matching, *Front. Appl. Math. Stat.*, **5** (2019), 47pp.
[31] S. Flaxman, S. Mishra, A. Gandy, H. Unwin and H. Coupland, et al., *Report 13: Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries*, 2020. Available from: `https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-13-europe-npi-impact/`.
[32] Gouvernement de la République Française, *COVID-19: Carte et Données*, 2020. Available from: `https://www.gouvernement.fr/info-coronavirus/carte-et-donnees`.
[33] H. Gupta, K. K. Verma and P. Sharma, Using data assimilation technique and epidemic model to predict TB epidemic, *Internat. J. Comput. Appl.*, **128** (2015), 5pp.
[34] P. L. Houtekamer and H. L. Mitchell, Data assimilation using an ensemble Kalman filter technique, *Mon. Weather Rev.*, **126** (1998), 796–811.
[35] P. L. Houtekamer and F. Zhang, Review of the ensemble Kalman filter for atmospheric data assimilation, *Mon. Weather Rev.*, **144** (2016), 4489–4532.
[36] M. A. Iglesias, K. J. Law and A. M. Stuart, Ensemble Kalman methods for inverse problems, *Inverse Problems*, **29** (2013), 20pp.
[37] Imperial College COVID-19 Response Team, *Short-term forecasts of COVID-19 deaths in multiple countries*, 2020. Available from: `https://mrc-ide.github.io/covid19-short-term-forecasts/index.html`.

[38] A. J. Ing, C. Cocks and J. P. Green, COVID-19: In the footsteps of Ernest Shackleton, *Thorax*, **75** (2020), 613–613.
[39] Institut de la Statistique Québec, 2020. Available from: https://www.stat.gouv.qc.ca/statistiques/population-demographie/deces-mortalite/nombre-hebdomadaire-deces.html.
[40] Institut de la Statistique Québec: Population Data, 2019. Available from: https://www.stat.gouv.qc.ca/statistiques/population-demographie/structure/population-quebec-age-sexe.html#tri_pop=20.
[41] Institut National de Santé Publique Québec, 2020. Available from: https://www.inspq.qc.ca/covid-19/donnees.
[42] C. Jarvis, K. Van Zandvoort and A. Gimma, et al., Quantifying the impact of physical distance measures on the transmission of COVID-19 in the UK, *BMC Med*, **18** (2020), 1416–1430.
[43] M. A. Jorden, S. L. Rudman, E. Villarino, S. Hoferka and M. T. Patel, et al., Evidence for limited early spread of COVID-19 within the United States, January-February 2020, *Morbid. Mortal. Weekly Rep. (MMWR)*, **69** (2020), 680–684,
[44] A. A. King, E. L. Ionides, M. Pascual and M. J. Bouma, Inapparent infections and cholera dynamics, *Nature*, **454** (2008), 877–880.
[45] R. Li, S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang and J. Shaman, Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2), *Science*, **368** (2020), 489–493.
[46] T. A. Mellan, H. H. Hoeltgebaum, S. Mishra, C. Whittaker and R. Schnekenberg, et al., Report 21: Estimating COVID-19 cases and reproduction number in Brazil, (2020).
[47] J. Mossong, N. Hens, M. Jit, P. Beutels and K. Auranen, et al., Social contacts and mixing patterns relevant to the spread of infectious diseases, *PLoS Med*, **5**.
[48] C. J. L. Murray, Forecasting the impact of the first wave of the COVID-19 pandemic on hospital demand and deaths for the USA and European economic area countries, preprint.
[49] National Health Service, *Covid-19 Daily Deaths*, 2020. Available from: https://www.england.nhs.uk/statistics/statistical-work-areas/covid-19-daily-deaths/.
[50] R. M. Neal, Sampling from multimodal distributions using tempered transitions, *Statist. Comput.*, **6** (1996), 353–366.
[51] NICE, COVID-19 Infecties op de IC's, Nationale Intensive Care Evaluatie, 2020. Accessed from: https://www.stichting-nice.nl/.
[52] NICE, COVID-19 Infecties op de Verpleegadeling, Nationale Intensive Care Evaluatie, 2020. Available from: https://www.stichting-nice.nl/covid-19-op-de-zkh.jsp/
[53] D. Pasetto, F. Finger, A. Rinaldo and E. Bertuzzo, Real-time projections of cholera outbreaks through data assimilation and rainfall forecasting, *Adv. Water Res.*, **108** (2017), 345–356.
[54] Public Health, England, The health protection (coronavirus, business closure) (England) regulations 2020, 2020. Available from: https://web.archive.org/web/20200323004800/http://www.legislation.gov.uk/uksi/2020/327/pdfs/uksi_20200327_en.pdf.
[55] P. N. Raanes, A. S. Stordal and G. Evensen, Revising the stochastic iterative ensemble smoother, *Nonlin. Processes Geophys*, **26** (2019), 325–338.
[56] Registro Civil, *Portal da Transparencia - Especial COVID-19*, 2020. Available from: https://transparencia.registrocivil.org.br/especial-covid.
[57] C. J. Rhodes and T. D. Hollingsworth, Variational data assimilation with epidemic models, *J. Theoret. Biol.*, **258** (2009), 591–602.
[58] RIVM, *Briefing Update Coronavirus Tweede Kamer 20 Mei 2020*, National Institute for Public Health and the Environment, 2020. Available from: https://www.tweedekamer.nl/sites/default/files/atoms/files/presentatie_jaap_van_dissel_-_technische_briefing_20_mei_2020.pdf.
[59] RIVM, *Excess Mortality Caused by the Novel Coronavirus (COVID-19)*, National Institute for Public Health and the Environment, 2020. Available from: https://www.rivm.nl/node/155011.
[60] RIVM, *Ontwikkeling COVID-19 in Grafieken*, National Institute for Public Health and the Environment, 2020. Available from: https://www.rivm.nl/coronavirus-covid-19/grafieken.
[61] H. Salje, C. Tran Kiem, N. Lefrancq, N. Courtejoie and P. Bosetti, et al., Estimating the burden of SARS-CoV-2 in France, *Science*, **369** (2020), 208–211.
[62] J. L. Sesterhenn, Adjoint-based data assimilation of an epidemiology model for the COVID-19 pandemic in 2020, preprint, arXiv:2003.13071.

[63] J. Shaman, A. Karspeck, W. Yang, J. Tamerius and M. Lipsitch, Real-time influenza forecasts during the 2012–2013 season, *Nature Commu.*, **4** (2013), 1–10.
[64] A. S. Stordal and A. H. Elsheikh, Iterative ensemble smoothers in the annealed importance sampling framework, *Adv. Water Res.*, **86** (2015), 231–239.
[65] UK Government, *Coronavirus (COVID-19) in the UK*, 2020. Available from: https://coronavirus.data.gov.uk.
[66] UK Government, *National COVID-19 Surveillance Reports*, 2020. Available from: https://www.gov.uk/government/publications/national-covid-19-surveillance-reports/.
[67] UK Government, *Slides, Datasets and Transcripts to Accompany Coronavirus Press Conferences*, 2020. Available from: https://www.gov.uk/government/collections/slides-and-datasets-to-accompany-coronavirus-press-conferences/.
[68] UK Office for National Statistics, *Dataset: Deaths Registered Weekly in England and Wales, Provisional*, 2020. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/datasets/weeklyprovisionalfiguresondeathsregisteredinenglandandwales.
[69] J. van Wees, S. Osinga, M. van der Kuip, M. Tanck and M. Hanegraaf, et al., Forecasting hospitalization and ICU rates of the COVID-19 outbreak: An efficient SEIR model, *Bull. World Health Org.*, (2020).
[70] J. S. Whitaker and T. M. Hamill, Evaluating methods to account for system errors in ensemble data assimilation, *Mon. Weather. Rev.*, **140** (2012), 3078–3089.
[71] WHO, *Coronavirus Disease (COVID-19): Similarities and Differences with Influenza*, 2020. Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-similarities-and-differences-covid-19-and-influenza.
[72] W. Yang, M. Lipsitch and J. Shaman, Inference of seasonal and pandemic influenza transmission dynamics, *PNAS*, **112** (2015), 2723–2728.
[73] W. Yang, W. Zhang, D. Kargbo, R. Yang and Y. Chen, et al., Transmission network of the 2014–2015 Ebola epidemic in Sierra Leone, *J. Roy. Soc. Interface*, **12** (2015).

Received June 2020; 1st revision October 2020; 2nd revision November 2020.

E-mail address: geev@norceresearch.no
E-mail address: j.amezcuaespinosa@reading.ac.uk
E-mail address: marc.bocquet@enpc.fr
E-mail address: n.a.carrassi@reading.ac.uk
E-mail address: alban.farchi@enpc.fr
E-mail address: a.m.fowler@reading.ac.uk
E-mail address: peter.houtekamer@canada.ca
E-mail address: ckrtj@email.unc.edu
E-mail address: rafaeljmoraes@gmail.com
E-mail address: manpulido@gmail.com
E-mail address: xian22@email.unc.edu
E-mail address: f.c.vossepoel@tudelft.nl