European Journal of Epidemiology (2020) 35:749–761
https://doi.org/10.1007/s10654-020-00676-7

COVID-19

An epidemiological modelling approach for COVID-19 via data assimilation

Philip Nadler[1] · Shuo Wang[1] · Rossella Arcucci[1] · Xian Yang[1] · Yike Guo[1]

Received: 2 May 2020 / Accepted: 10 August 2020 / Published online: 4 September 2020

Abstract
The global pandemic of the 2019-nCov requires the evaluation of policy interventions to mitigate future social and economic costs of quarantine measures worldwide. We propose an epidemiological model for forecasting and policy evaluation which incorporates new data in real-time through variational data assimilation. We analyze and discuss infection rates in the UK, US and Italy. We furthermore develop a custom compartmental SIR model fit to variables related to the available data of the pandemic, named SITR model, which allows for more granular inference on infection numbers. We compare and discuss model results which conducts updates as new observations become available. A hybrid data assimilation approach is applied to make results robust to initial conditions and measurement errors in the data. We use the model to conduct inference on infection numbers as well as parameters such as the disease transmissibility rate or the rate of recovery. The parameterisation of the model is parsimonious and extendable, allowing for the incorporation of additional data and parameters of interest. This allows for scalability and the extension of the model to other locations or the adaption of novel data sources.

Keywords Data assimilation · 2019-nCov · Inference · Bayesian updating · Compartmental model

Introduction

The global outbreak of n-Cov2019 and the possibility of severe social and economic costs worldwide requires immediate action on suppresion measures. In order to evaluate the efficacy of past and future policy measures to fight and contain the spread of n-Cov2019, a robust and quantifiable analysis system is required. We propose a methodology for forecasting the spread of n-Cov2019 and show how to estimate latent infection rates, accounting for high uncertainty in observation and model specification, which is done by combining real-time Bayesian updating with epidemiological models.

To show the generalisability of our updating approach we first embed a standard SIR model in our framework and then develop a custom compartmental SIR model which is fit to data related to the spread of the coronavirus worldwide which we name SITR. The SITR model adds an additional compartment for patients under treatment T and allows for more granular inference on the underlying dynamics of the epidemic, separating confirmed cases under treatment with latent unconfirmed cases of Covid19. The models are embedded in a data assimilation framework, a form of recursive Bayesian estimation [1], which conducts model updates when new observations become available. The assimilation scheme lends itself naturally to the problem because the procedure allows the model to dynamically adjust infection rates in real time, while taking into account the uncertainty in the data via the specification of covariance matrices.

The uncertainty quantification and choice of covarainace matrices is being analyzed using a hybrid data assimilation approach, which is applied to make results robust to initial conditions. We use the model to infer the amount of infected people and both, the disease transmissibility rate, as well as the rate of recovery. The time-varying parameter structure of

✉ Philip Nadler
p.nadler@imperial.ac.uk

✉ Yike Guo
y.guo@imperial.ac.uk

Shuo Wang
shuo.wang@imperial.ac.uk

Rossella Arcucci
r.arcucci@imperial.ac.uk

Xian Yang
xian.yang08@imperial.ac.uk

[1] Data Science Institute, Imperial College London, London SW7 2AZ, UK

the model allows for the incorporation and analysis of policy action, such as if the shutdown of transportation or closure of schools affect transmissibility.

In line with other researchers, our model estimates indicate that the number of infected people is a number of magnitudes higher than the actual reported number of confirmed reported infections. We also find that compared to static models, updating the parameters in a dynamic fashion leads to an upward correction of the true number of infected people as well as reducing forecasting errors.

We estimate both short term and long term dynamics in Italy, the United States and the United Kingdom, finding a peak of infections in the middle of March and a flattened but sustained number of infected cases in the United States and the United Kingdom. We furthermore analyze the transmissibility rate and find that they decreased after imposed initial lockdown measures, but increased again after a loosening of restrictions end of May. The rest of the paper is structured as follows: Section two discusses related work. Section three and four introduce the dynamic model as well as the SITR compartmental model. Section five discusses how the uncertainty in the data is incorporated in the model. Section six states discusses empirical results and section seven concludes.

Related work

The spread of the novel type of respiratory virus as well as the dramatic economic consequences trying to contain it has led to a rapid engagement of the scientific community, with many different areas of research being explored.

Using compartmental models in epidemiology, authors such as [2, 3] and [4] have done the first studies on the size of the outbreak in China. They applied standard SIR models with static parameters to estimate the basic reproductive number and analyze the exponential growth of the virus in Wuhan. The work of [4] in particular combines standard SEIR models with travel data obtained from Tencent and found that epidemic dynamics show exponential patterns in multiple major cities with a lag behind the Wuhan outbreak of about one to two weeks.

First studies using data assimilation for epidemiological modelling have been conducted by other authors such as [5] and [6], which studied the techniques on different cases such as influenza using standard SIR models, although none has considered the issue of the robust covariance estimates as discussed in [7] or [8].

Further studies such as [9] and [10] study time varying parameters in more detail, although only for standard SIR models with no relationship to the current corona virus outbreak. We are the first to conduct a study of the current spread of 2019-nCoV using data assimilation. We furthermore contribute by providing a novel framework which enables the prior computation of covariance matrices, adding robustness to epidemiological assimilation models. Although many compartmental models are available, we base our initial studies on SIR models, since it allows us to verify the dynamics of the assimilation scheme as a robust benchmark and compare it to extensions of the model later on. We will specify the exact model choice and specification in the next section.

The adaptive epidemiological model

We introduce an adaptive epidemiological modelling framework which combines a SIR model whose model parameters are time-varying with data assimilation techniques. We base our model on the most basic compartmental model, which is the SIR. Describing and implementing the assimilation scheme in the basic structure of a SIR model allows us to analyze its initial performance and derive additional modifications later on. Further complexities will be introduced when more granular data is available. The current confirmed cases are mostly symptomatic cases with many asymptomatic cases being unconfirmed due to limited testing capacities.

For Covid19, a SEIR model adding an compartment for exposure is a possible candidate for further extensions. But given the lack of more granular data with all confirmed infection cases being symptomatic and no other data on exposure levels, adding additional parameters with insufficient data can lead to a bias of parameter estimates in the assimilation procedure, and also increases model complexity without adding understanding of the underlying mechanics of the Covid19 infections [11, 12]. The lack of meaningful and accurate data which fits the model assumptions for more complex models strongly affects the performance of data assimilation and thus makes a sparse and parsimonious model preferable [1].

The standard SIR model

We start our analysis with a standard SIR model [13], which is a system of three interrelated non-linear ordinary differential equations without an explicit analytical solution. The dynamics of the model are given by:

$$\frac{dS}{dt} = -\beta \frac{IS}{N} \tag{1}$$

$$\frac{dI}{dt} = \beta \frac{IS}{N} - \gamma I \tag{2}$$

$$\frac{dR}{dt} = \gamma I \tag{3}$$

where S denotes the susceptible population size, I the infected people who are not isolated from the population and R the recovered population. The total population is given by N. The parameters β and γ denote the transmission and recover rate of the virus infection. Note that for the outbreak, the susceptible number S is observable, which we label as the population of the country under analysis. The recovered population R denotes the population not infectious anymore and being removed from the population, which for the studied examples is the number of confirmed cases, since confirmed cases are hospitalized and isolated and not infecting the general population anymore.

The adaptive DA-SIR model

Data Assimilation (DA) is a technique to incorporate observations into a theoretical model where uncertainty is quantified [1]. It allows for problems with uneven spatial and temporal data distribution and redundancy to be addressed such that models can ingest information. DA is a vital step in numerical modeling and has become a main component in the development and validation of mathematical models in meteorology, climatology, geophysics, geology and hydrology [14]. Recently, DA is also applied to numerical simulations of geophysical applications, medicine, biological science and finance [15]. Data assimilation can be applied to a variety of problems where an uncertainty quantification has to be included [16] or where latent parameters need to be computed taking into account new observations.

The Adaptive DA-SIR model is a model which incorporates data assimilation with a compartmental SIR model. We use DA as an adaptive modelling approach which integrates new observations into our compartmental model to enhance the accuracy of forecasts as well as computing model parameters of interest, in our case β and γ in the SIR model.

The SIR model in equations (1)–(3) can be discretized with respect to the time variable, giving the following equations:

$$S_{t+1} = S_t - \beta \frac{I_t S_t}{N} \tag{4}$$

$$I_{t+1} = I_t + \beta \frac{I_t S_t}{N} - \gamma I_t \tag{5}$$

$$R_{t+1} = R_t + \gamma I_t \tag{6}$$

For a given time step t and assuming to have observations of the variable R_t we denote here with R_t^{obs}, the DA problem consists in computing the minimum of the cost function

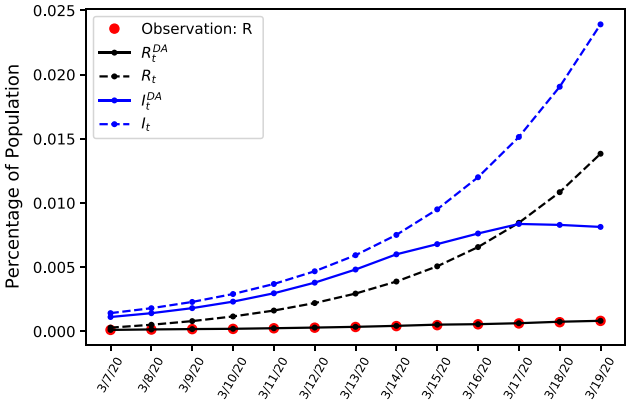

Fig. 1 Comparing estimates during early stages of the outbreak of confirmed cases R and unobservable amount of infected people I in Italy, using static and updating parameters

$$J(I) = \sum_{i=t+1}^{t+\tau} ||R_i^{obs} - R_i^{pred}(I, \beta, \gamma)||_{\mathbf{Q}_t^{-1}} + ||I - I_t^{pred}||_{\mathbf{P}_t^{-1}} \tag{7}$$

and

$$I_t^{DA} = \underset{I}{\operatorname{argmin}}\, J(I) \tag{8}$$

where R^{pred} is a predicted value generated by the SIR model, and where $\mathbf{Q}$ and $\mathbf{P}$ denote the the background and the observation covariance matrices, representing an estimation of the errors in the data. To estimate the parameter, we minimize

$$\beta_t, \gamma_t = \underset{\beta,\gamma}{\operatorname{argmin}} \sum_{i=t+1}^{t+\tau} \|R_i^{obs} - R_i^{pred}(I_t^{DA}, \beta, \gamma)\|_{\mathbf{Q}_t^{-1}} \tag{9}$$

Data assimilation is very sensitive to initial conditions and the choices of the covariance matrices, since they quantify uncertainty and determine how much weight is assigned to new observations which are assimilated into the model. Thus their calibration needs to be properly chosen, which we outline in detail in Sect. 5.

The data we use representing S_t, I_t and R_t is given by the official government numbers and is available at [17, 18]. The solution of the DA problem in (7) leads to a modified extended Kalman filtering algorithm where an SIR model is used to compute the forward steps, e.g. in the time window $[t, t+M]$. Where I_t^{DA} are the values of I_t computed after the assimilation of R_t^{obs} as in Eq. 8. To illustrate and put results into perspective, we compare results of our adaptive DA-SIR model with the common SIR model for an example case.

Both models use the same initial conditions given by the observed data. In Fig. 1 we compare model performance of the standard SIR model and show how assimilation of new observations generates updated model dynamics in the DA-SIR model that do differ from standard SIR model predictions

Table 1 Selected data points for predicted number of infected and treated patients for a dynamic model and a static ODE model

Date	3/7/20	3/9/20	3/11/20	3/13/20	3/15/20	3/17/20	3/19/20
I_t^{DA}	67,050	108,540	178,152	289,418	422,682	525,755	534,073
I_t	85,190	137,428	221,401	355,915	570,180	908,422	1,434,817
R_t^{DA}	17,311	47,899	97,212	176,569	303,921	507,379	830,126
R_t	6705	10,582	14,595	21,030	30,945	37,853	49,265

by a wide margin, as is illustrated in the figure. Selected values of the graph are available in Table 1 where we compare estimated confirmed cases and latent infection rates for both the static SIR, and dynamic DA-SIR model.

The dynamic model given by the solid lines fit the observed values of confirmed cases R_t and interpolates the number of infected people I_t. The dashed lines represent the standard SIR model and show how not updating the model from the initial conditions leads to overestimation of infectious cases if interpolating according to simple exponential SIR dynamics, with a inferior model fit when comparing the observed confirmed cases denoted in red. This illustrates how without updating the parameters the number of infected people is overestimated and the assimilation of new observations helps to adjust the trajectory of likely infections in the future. Having shown the large difference between static and dynamic SIR models we next introduce a further refined extension of the dynamic assimilation model.

The extended epidemiological assimilation scheme

The SITR model

Having illustrated the benefits of embedding the SIR model in a DA framework, we aim to further exploit the available data to do more fine tuned inference. In the previous case of the simple SIR model, both recovered and isolated patients were categorized as R. We revise the SIR model by introducing an intermediate compartment T. Here, T represents the number of people being treated, given by the difference between accumulated confirmed cases and recovered or deceased patients R. Instead of just observing one variable, the number of confirmed cases, we are now observing two variables: the currently confirmed cases being treated T and removed infectious population due to recovery or being deceased R. The model is given by

$$\frac{dS}{dt} = -\beta^e I \tag{10}$$

$$\frac{dI}{dt} = \beta^e I - \alpha I \tag{11}$$

$$\frac{dT}{dt} = \alpha I - \gamma T \tag{12}$$

$$\frac{dR}{dt} = \gamma T \tag{13}$$

The parameter $\beta_t^e = \beta \frac{S_t}{N_t}$ is the real transmission rate over time, taking into account the total population size N as in the SIR model. Assuming all the parameters $\boldsymbol{\theta} = [\beta^e, \alpha, \gamma]$ time dependent, the SITR model in equations (10)-(13) can be discretized with respect to the time variable, giving the following equations:

$$S_{t+1} = S_t - \beta_t^e I_t \tag{14}$$

$$I_{t+1} = I_t + \beta_t^e I_t - \alpha_t I_t \tag{15}$$

$$T_{t+1} = T_t + \alpha_t I_t - \gamma_t T_t \tag{16}$$

$$R_{t+1} = R_t + \gamma_t T_t \tag{17}$$

which is a linearized approximation of the original SIR model with the additional compartment T. This provides the model prediction of the compartment states $\mathbf{X} = [S, I, T, R]^T$ given all parameters including β^e. The other variables are the same as in the SIR model, where S denotes the susceptible population, I the infected people who are not isolated from the population. The parameters γ and α denote the recovery and transition rate given by total of incubation and admission days. To extend the model and incorporate information not just of the last timestep, we introduce a model extension which bases model predictions on a sliding window of length τ, similar to a 4D-VAR approach [1]. For a given time window $[t+1, t+\tau]$ and assuming to have observations of the variable T_t which we denote here with T_t^{obs}, the resulting assimilation scheme is given by

$$J(I) = \sum_{i=t+1}^{t+\tau} ||T_i^{obs} - T_i^{pred}||_{\mathbf{Q}_i^{-1}} + ||I - I_t^{pred}||_{\mathbf{P}_\tau^{-1}} \tag{18}$$

and

$$\mathbf{X}_{t+1}^{pred} = F(\mathbf{X}_t, \boldsymbol{\theta}_t) \tag{19}$$

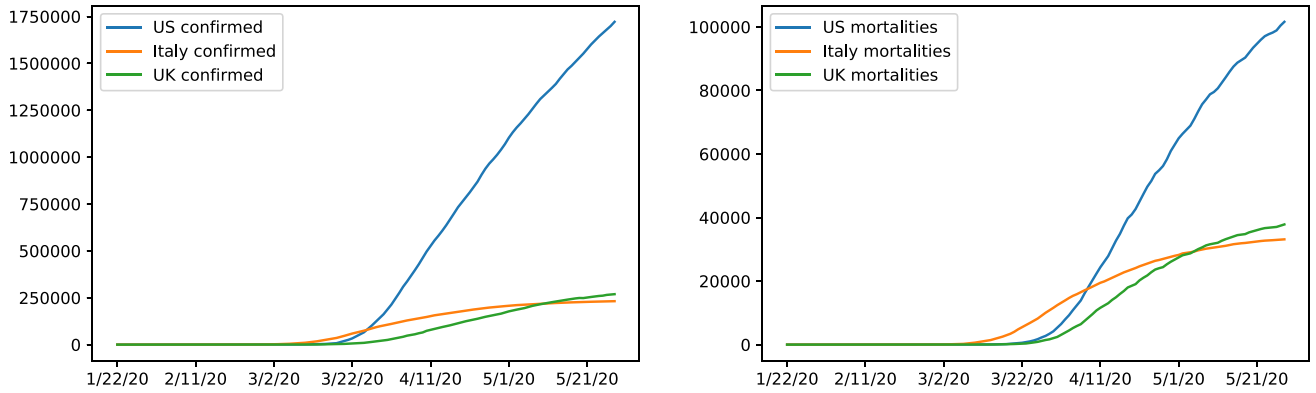

Fig. 2 Number of mortalities and confirmed cases for the United States, Italy and the United Kingdom

$$I_t^{DA} = \underset{I}{\operatorname{argmin}}\, J(I) \tag{20}$$

which in a first step infers the number of infected people I. To estimate the infection rate, in a second step we minimize

$$\beta_t^e = \underset{\beta^e}{\operatorname{argmin}} \sum_{i=t+1}^{t+\tau} ||T_i^{obs} - T_i^{pred}||_{\mathbf{Q}_i^{-1}} \tag{21}$$

which updates β conditioned on assimilated values of I. The resulting algorithm implements a 4D-VAR assimilation scheme in cost function (18), where forecasts and parameter estimates are based on a sliding window over time. Without preconditioning, the algorithm updates the model parameter values with the noise and observation matrices $\mathbf{Q}$ and $\mathbf{P}$ being fixed hyperparameters. In order to present results which may have major policy implications, correct and robust estimation of initial conditions and hyperparameters is of high importance, we therefore introduce a formalization and preconditioning of the covariance matrices $\mathbf{Q}$ and $\mathbf{P}$ before applying the assimilation scheme, which is named hybrid data assimilation.

Uncertainty in infection rates

The data that is being observed is highly aggregated and suffers from uncertainty firstly due to human measurement error and secondly due to number of confirmed cases being a noisy subset of the true number of infections. Furthermore different definitions for confirmed cases or the cause of mortalities due to Covid19 in various countries adds noise and uncertainty to the data. The figures in Fig. 2 show example data of confirmed cases and mortalities.

This uncertainty in the data mandates a methodology that incorporates the uncertainty into the model. The DA-SIR model takes this uncertainty into account via the values of the covariance matrices $\mathbf{Q}$ and $\mathbf{P}$.

The state and observation covariance matrices $\mathbf{Q}$ and $\mathbf{P}$ determine the weight of new observations when updating the parameters of the model. As is stipulated in the cost function in Eq. 7, the inverse of the covariance matrices are used to weight the terms of the observation and model operator. Thus very noisy data yielding large error covariance matrices will result in less weight of the term containing the model operator, i.e. less certainty is put on the data.

Therefore the next section studies different covariance matrix setups to take data uncertainty into account. We include detailed steps for the computation of the covariance matrix in the appendix, where we outline the ensemble variational approach applied to generate robust covariance matrix estimates.

Sensitivity analysis

As we mentioned in Sect. 4.1, the choice of the covariance matrices strongly affect the efficiency and the accuracy of the assimilation approach. As the available data is not accurate enough, in order to justify our estimations, we run a sensitivity analysis to study the impact of our estimated parameters and covariance matrices into the model predictions, using a subset of the data as illustration. To illustrate the hyperparameter sensitivity we compare the number of estimated infected people and we apply a mean root squared forecasting error (MRSFE) metric:

$$MRSFE = \sum_{n=0}^{N} \left(\frac{\sum_{\tau=\tau_0}^{T-h} \sqrt{(y_{t,n}^r - \hat{y}_{t,n})^2}}{T-h-\tau_0+1} \right) \tag{22}$$

where $\hat{y}_{t,n}$ represents the model prediction, $y_{t,n}^r$ the real observation with forecast horizons defined by $h = 1$, and $\tau_0 = 1$

Table 2 Sensitivity analysis for different values of observation and model error covariance matrices. The first two rows show number of latent infected patients and patients under treatment predicted for the 28.05.2020, the last day in the sample. The last two rows show the mean forecasting errors for treated patients and confirmed cases over the full sample for each covariance matrix configuration. The table exemplifies a bad fit with a high amount of forecasting errors when using a naive unit covariance setup

P Value	0.1	1	1	1	100	100
Q Value	0.5	100	1	10	1	10
Treatment	230,760	231,336	230,792	231,517	230,782	230,934
Infections	210,311	2,102,126	2,102,549	2,102,431	2,102,576	2,102,603
MRSFE T	888	863	901	807	829	836
MRSFE R	39,476	39,291	39,431	39,381	39,459	39,469

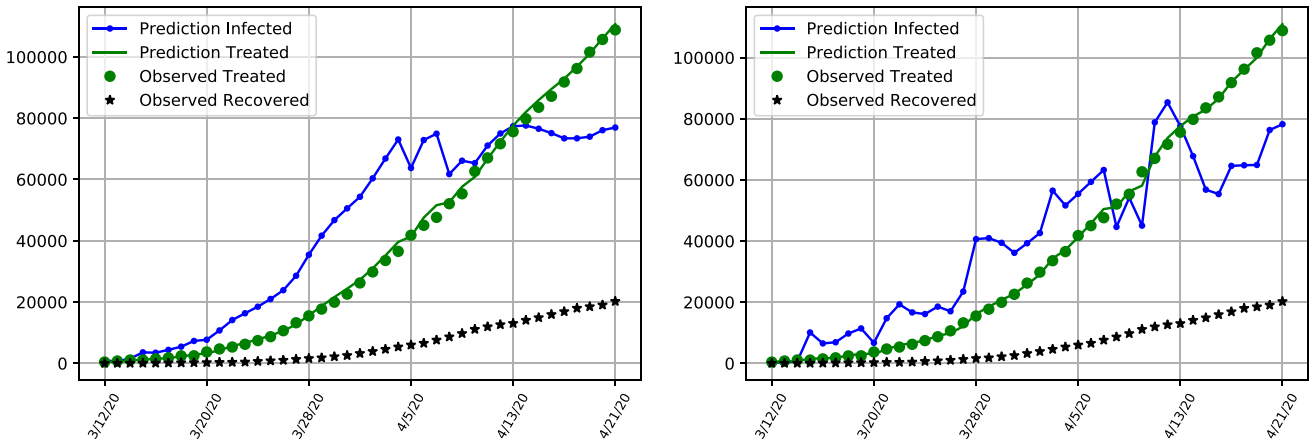

Fig. 3 Number of infected and treated cases for the United Kingdom, depicting the difference between infection numbers for robust (left) and naive unit covariance matrices (right)

the starting period of the forecast for n variables. The results are given in Table 2.

The results in the table show that a naive setup using unit covariance matrices leads to detrimental fit of the model with relatively high forecasting errors compared to other covariance combinations, with generally better performance for high values for observation covariance matrix **P**, meaning that prediction accuracy increases when less weight is put on the model dynamics and more weight on the observations. The lowest forecasting error for T is given with a unit observation matrix **P** and an error covariance matrix **Q** of 10, showing that the optimal combination of covariance matrix values is non-linear and requires a carefully considered estimation algorithm, as our proposed ensemble approach.

The computation of the covariance matrices is performed via a hybrid-assimilation approach where the covariance matrices are estimated using an ensemble approach. In this approach, the values of the covariance matrices are generated through sampling multiple synthetic trajectories using the SITR which is initialized with different parameters for each draw as well as calculating the empirical residual covariance matrix for the data. The details of the procedure are outlined in the appendix.

Figure 3 depicts different infection curves given the naive unit covariance setup on an example time period of the data in Table 2 and show that the dynamics are affected by the choice of the covariance matrices. The updated model assimilates new observations of infected patients and people recovered from the virus. The long run dynamics predict a recent spike in the number of infected people in the United Kingdom. The total number of people being treated in hospitals follows with a small lag and is still growing towards the end of the example set.

The trajectories for patients under treatment are similar but the dynamics for latent infections contain some discrepancies. The left hand side depicts the trajectory of estimated latent infections and patients over time using the ensemble approach with robust covariance matrices, whereas the right hand figure depicts results using a simple unit covariance setting. The left hand side depicts a much smoother and more well-behaved series of treated cases as well as infections, approximating a more reasonable stable growth path, whereas the unit covariance case would yield much more erratic, unrealistic fluctuations in infection numbers over time, with many sudden drops in infection numbers.

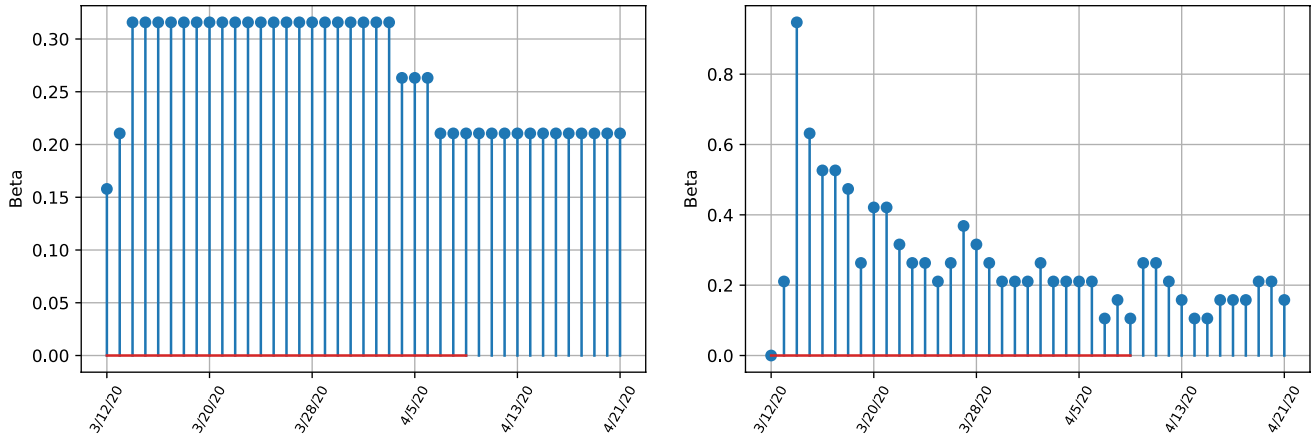

Fig. 4 Estimates for transmissibility rates beta for the United Kingdom, depicting the difference between infection numbers for robust (left) and unit covariance matrices (right)

Table 3 Selected data points for predicted number of infected and treated patients, as well as the MRSFE. Results are obtained using a naive unit covariance matrix

Date	03–09	03–16	04–12	04–19	05–21	05–28
Infected Patients	21	85	12,629	18,492	36,042	37,837
Treated Patients	300	1458	71,650	101,575	214,866	231,290
RSFE Treated	261	110	2049	1592	1308	498
RSFE Infections	43	114	4439	13,950	133,162	172,417

Table 4 Selected data points for predicted number of infected and treated patients, as well as the MRSFE. Results are obtained using the hybrid assimilation covariance matrix

Date	03–09	03–16	04–12	04–19	05–21	05–28
Infected Patients	31	154	16,755	32,023	168,831	209,999
Treated Patients	290	1405	72,363	101,051	214,629	232,074
RSFE Treated	10	53	713	524	237	784
RSFE Infections	10	69	4126	1353	132,789	172,162

Since the number of treated patients is observable, we use generated forecasts of treated persons as a forecasting metric to evaluate the model fit. The Tables 3 and 4 give excerpts from the forecasts values of infected and treated patients as well as the corresponding MRSFE for treated patients in hospitals. For the unit covariance Table 3 it is observable how the unrealistic dip in forecasted infections which is visible in the right side plot in Fig. 3 causes a large spike in forecasting errors for treated people beginning of April onwards. Comparing it with a hybrid assimilation approach in table 4 reveals an overall lower number of forecasting error and better fit. The sensitivity of results confirm the need for a more rigorous algorithm of covariance estimates given the few and noisy datapoints.

Comparing the left and right bottom figures of Fig. 4, the transmissibility rate β shows less variation over time, which differs from the model without ensembles where strong variation is visible. Both estimates depict the downward trend of transmissibility. The high variability of the unit covariance matrix estimate implies that the transmissibility is affected more easily by external factors which change the dynamics of new infections. Thus the robust model estimates imply that, within the sample period, the transmission rate is more stable and unaffected by changes in observations because a uncertainty weight is given by the covariance matrix estimates. We next proceed to apply out methodology to the full dataset analysing the US, the UK and Italy.

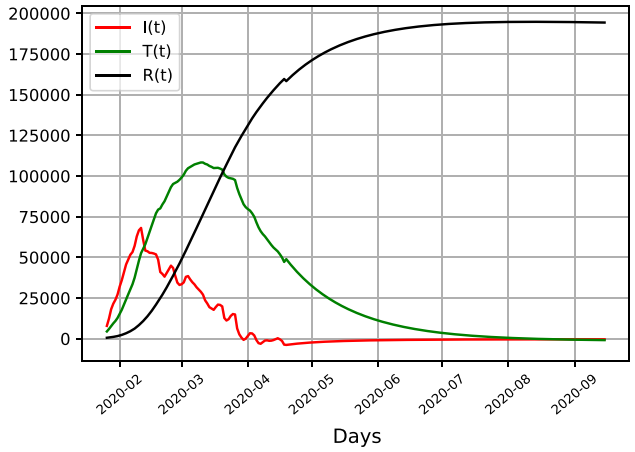

Fig. 5 SITR results for Italy, showing estimates of latent infections (red), recoveries (black) and hospitalizations (green)

Empirical results

Trend analysis of international data

To illustrate the flexibility of our approach we apply our analysis to an international comparison with additional results for the United States, the United Kingdom as well as Italy.

For the models we focus on the number of infected people extrapolated from the number of confirmed cases and recoveries. The data was obtained from the John Hopkins University Coronavirus Resource Center[1]. We compare data and show test results for different forecast horizons and parameter estimates. Given the confirmed coronavirus cases we infer the amount of infected people and do forecasts to estimate the approximate development of the epidemic. We estimate the model on a sample of daily data until the 28/05/20 and evaluate models based on their fit and infection curves.

Although long-term forecasts are of limited use in fast changing scenarios such as the current pandemic, they nevertheless can provide rough guidance on a potential peak of infection rates. Comparing cases for all three countries Fig. 5 depicts the long-run dynamics of the epidemic in Italy, where according to the model the peak of infections has already occurred in march, with a gradual decrease in infection numbers afterwards. The absolute number of patients under treatment decrease throughout April and May.

Comparing Italian results with the United States and the United Kingdom in Fig. 6 shows that the trajectories of Italy differs from both countries, with the UK and the US showing similar patterns. When infection numbers in Italy already peaked, the number of latent infections depicted in red has not decreased, but has merely stabilized at a high level, whereas the hospitalized cases under treatment are still growing, with a forecasted peak by end of April, which contrasts to the Italian peak in March. The number of infections are increasing within sample, as is the forecasted number of infections. The peak of latent infections stabilses at slightly below half a million cases in the US and 100,000 cases in the UK. For both countries the total number of infections tappers off by the end of September, whereas this already happens by July for the case of Italy.

The different levels of infections are likely due to different inception dates of the pandemic, having started earlier in Italy than the United States, with an eventual peak of the United States not visible yet given the current data sample. The results indicate that the pandemic has reached a peak in Italy recently, the dynamics for the United Kingdom and especially the United States indicate that no plateau has been reached yet and that the number of infections is likely to increase.

Short term dynamics

The results given by the dynamic SITR model highlight the different phases of development in Italy and the United Kingdom and the United States, as is depicted in Figs. 7, 8 and 9 respectively. The figures depict the predicted latent infection numbers, observed and predicted hospital treatment numbers as well as the number of recoveries. We also provide accompanying tables where we report the MRSFE fit of observed confirmed and treated cases. We first discuss the trajectories of infections and follow up with a discussion on the parameter estimates in the next section.

Results align with the previous analysis where in Fig. 7 it is observable that in Italy the number of latent asymptomatic cases has decreased with the majority of patients being hospitalized with a decreasing trend for both latent infections and hospitalized patients. The pattern of latent infections follows a shaped curve, with the majority of infections already peaked in march and now being on a trajectory exhibiting signs that the pandemic is under control. Table 5 provides additional details for the model performance, showing the number of latent infections and treated patients in the model. We show the forecasting errors when fitting the model in order to be able to compare model fit between all three countries. The forecasting errors depict how the model fits the data, with the estimates of treated patients performing particularly well compared to the UK and US (Tables 6 and 7). At the end of the sample on the 28th of May the number of latent infected patients is slightly above 180,000 and the number of patients under treatment is below 48,000, with the values exemplifying the decrease from the previous month.

[1] https://coronavirus.jhu.edu/map.html

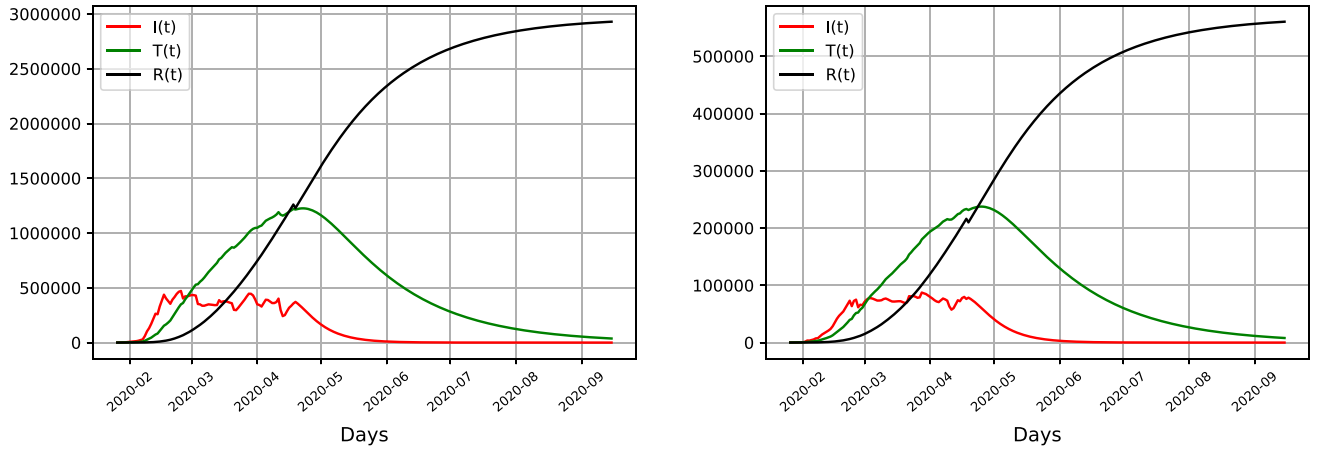

Fig. 6 SITR results for the US (left) and UK (right), showing estimates of latent infections (red), recoveries (black) and hospitalizations (green)

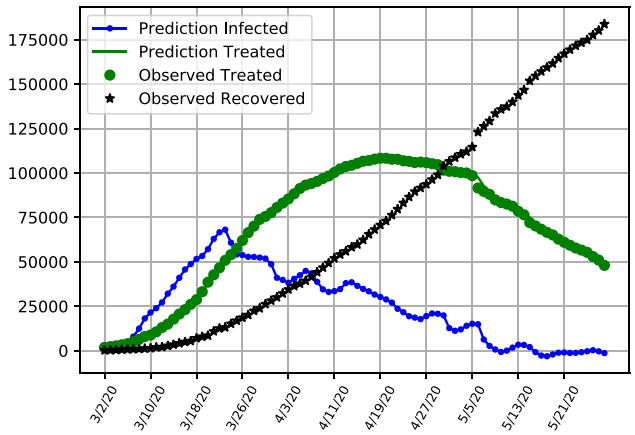

Fig. 7 SITR short run dynamics Italy, showing recoveries (black), latent infections (blue) as well as observed and predicted numbers of treated patients (green)

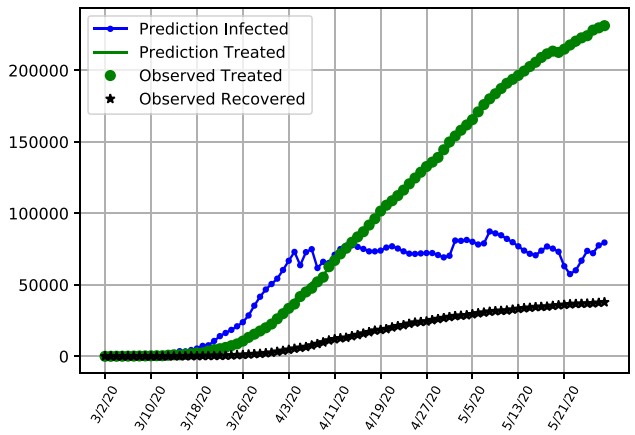

Fig. 8 SITR short run dynamics in the United Kingom, showing recoveries (black), latent infections (blue) as well as observed and predicted numbers of treated patients (green)

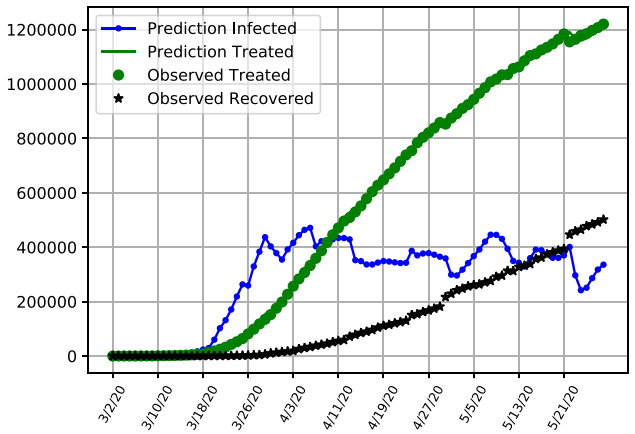

Fig. 9 SITR short run dynamics in the United States, showing recoveries (black), latent infections (blue) as well as observed and predicted numbers of treated patients (green)

Overall number of infected but not hospitalized patients increases initially and are starting to trend downward after a peak at the end of March.

The hospitalization numbers are displaying a curve like behaviour, following the infections with a lag. The maximum number of patients under treatment are reached in the middle of April, with a further flattening curve towards the end of the sample.

This is in contrasts to the United States and the United Kingdom as is visible in Figs. 9 and 8 where the number of latent infections has been relatively constant and exhibits less of an downward trending behaviour. With the number of hospitalized infections increasing, this is likely due to the early relaxation and less stringent quarantine restriction in the United States compared to Italy. When inspecting both infected and hospitalized compartments, results of the SITR model aligns with policy choices, with a flattening tendency

Table 5 Selected data points for predicted number of infected and treated patients, as well as the MRSFE. Results are obtained using the hybrid assimilation covariance matrix for Italy

Date	03–09	03–16	04–12	04–19	05–21	05–28
Infected Patients	1187	4907	54,110	70,715	167,046	183,746
Treated Patients	7985	23,073	102,253	108,257	60,960	47,986
RSFE Treated	226	1156	1405	224	195	1958
RSFE Confirmed	292	1562	2619	199	18,980	25,316

Table 6 Selected data points for predicted number of infected and treated patients, as well as the MRSFE. Results are obtained using the hybrid assimilation covariance matrix for the UK

Date	03–09	03–16	04–12	04–19	05–21	05–28
Infected Patients	31	154	16,755	32,023	168,831	209,999
Treated Patients	290	1405	72,363	101,051	214,629	232,074
RSFE Treated	10	53	713	524	237	784
RSFE Confirmed	10	69	4126	1353	132,789	172,162

Table 7 Selected data points for predicted number of infected and treated patients, as well as the MRSFE. Results are obtained using the hybrid assimilation covariance matrix for the US

Date	03–09	03–16	04–12	04–19	05–21	05–28
Infected Patients	29	117	59,074	111,282	393,120	501,607
Treated Patients	559	4544	496,239	647,527	1,184,027	1,220,146
RSFE Treated	32	152	7806	3911	13667	15347
RSFE Confirmed	31	281	630,550	113,096	616,649	726,317

for Italy, contrasting with results for the United States and the United Kingdom which followed later or with less rigorous quarantine restrictions. Table 6 depicts selected entries and forecasting errors for the United Kingdom, where at the end of the sample in May the amount of latent infections is estimated to be around 209,000.

Table 7 shows selected values for the United States. Even taking account the higher amount of infection numbers in the United States, the high amount of prediction errors is noticeable, indicating that the data is noisy ad the model fit is not performing as well as in the other two cases.

Given these results the number of infections in the United States and the United Kingdom are likely to keep on growing, especially if the government is not considering the implementation of tighter regulations and quarantine measures. The results furthermore demonstrate how the assimilation framework can be extended to multiple countries and provide robust results given the large uncertainty in infection estimates.

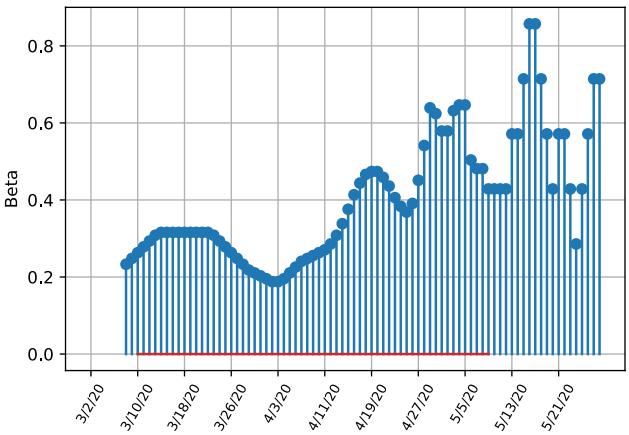

Fig. 10 SITR short run transmissibility dynamics for Italy

Lockdown effects on transmissibility

To compare the predicted dynamics of our model estimates and to evaluate policies, we extend the analysis and discuss the dynamics of the model parameters. Following the same framework, we analyze the estimated transmissibility rate

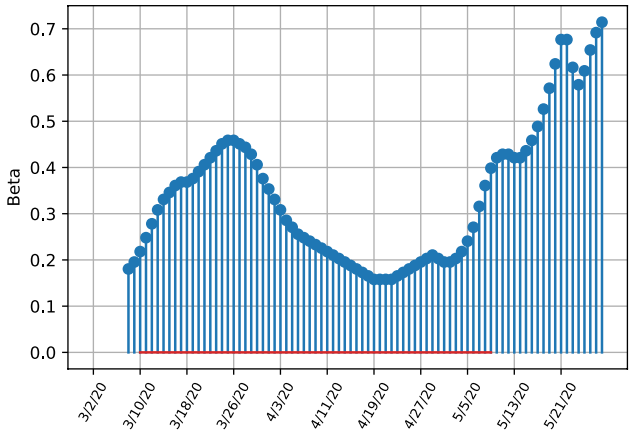

Fig. 11 SITR short run transmissibility dynamics for the United States

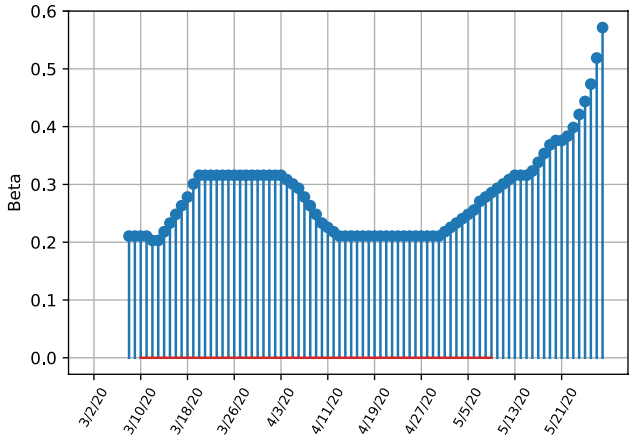

Fig. 12 SITR short run transmissibility dynamics for the United Kingdom

over time in all three countries for which we depict weekly averaged results. Italy has seen the earliest surge of Covid19 cases but kept its quarantine measures intact. The US and UK were followed by rising surges with a delay, but have had a looser approach to quarantine measures.

As given in Fig. 10, the infection rate β reached a high level of 0.3 from March 10th to to around March 20th, followed by a gradual decreasing period with an infection rate bottoming out at a value of 0.2, showing initial successes in lockdown measures, which were enforced from the 9th of March onwards. This shows how the lockdown order has first stabilized the transmissibility rate and later led to a decrease. Towards the end of the sample with increasing relaxation of the lockdown measures an increase in β is observable. Variation of the parameters is very high towards the end of the sample, especially compared to examples of the UK and US, being evidence that the model parameter estimates are less clear on a strong increase in infection rates as in the US or the UK.

Overall,the Italian dynamics in Figs. 7 and 10 clearly indicate that the number of undetected latent infections is decreasing, with also the number of known and hospitalized cases having crossed their peak value as well and approaching a very low value in the sample. The infection rates indicate a decrease after the enforcement of quarantine measures, with a increase in transmissibility towards the end of the sample, although the model parameters exhibit strong variation in parameter estimates.

The United States have not reached a similar level as Italy. The trajectories in Fig. 11 illustrate that after a rapid growth and high peak value of transmissibility on the 26th of March, a strong decrease followed, showing the effectiveness of the lockdown measures that were enforced in many states from the 23th of March onwards. After an initial strong decrease of the transmissibility rate after imposing a lockdown, the infection rate has strongly increased again at the end of the sample, with the transmissibility values increasing from a trough of 0.18 on the 19th of April to 0.7 at the end of May. This, together with the trajectories in Fig. 9 show that restriction measures have only lead to initial successes with a stabilization of latent infection cases with the total number of cases still increasing. Especially towards the end of the sample at the end of May the amount of infections is increasing again.

The development of the United Kingdom is very similar to the United States, where initial lockdown measures managed to decrease the number of latent infections, with the initial growth in infection numbers in the middle of March is not as strong as in the Italy or the US. This is explainable when taking into account the parameter estimates given in Fig. 12. The parameter plot visualises this development clearly, where transmissibility has initially decreased from the 4th of April onward and stabilized at a level of 0.2 and steadily increased from the end of April onwards at the end of the sample. The parameter estimates unequivocally indicate a strong increase in infection rates, with the last estimated peaking at around 0.6, although the relative increase from it's lowest value throughout April is not as pronounced as in the US indicating that the increase in transmissibility has not been affected as much by changing policies or behaviour in the population as in the US.

Overall the results indicate that Italy is progressing well in containing the virus, the UK and especially the US are struggling to contain the virus in the medium-term. Especially the parameters indicate worsening developments, where in all three countries transmissibility has decrease initially due to quarantine measures but increased towards the end of the sample, although evidence is less clear and more uncertain for the Italian transmissibility rates.

Conclusion and future work

We introduced a novel epidemiological assimilation scheme to forecast and evaluate the current corona pandemic worldwide with a specific focus on the United States, the United Kingdom and Italy. We combined compartmental models in epidemiology with data assimilation schemes showing the advantage of real-time forecasting and parameter estimation in the current crisis. We discussed the benefits and differences in infection numbers when models are updated on a daily basis compared to static modelling. We then introduced a model extension allowing us to observe patients being treated, and patients being removed from the infectious population, which we labelled SITR. Since models are sensitive to estimates of the covariance matrices, we add a hybrid ensemble approach which allows for robust covariance matrix estimates. We find that in Italy the peak of infections has been reached already, with the number of patients being treated peaking middle of April. The trajectories of the US and UK are less clear, with a likely increase in the medium term, with both countries showing a strong increase in transmissibility rates after an initial decrease due to lockdown measures.

The generalisability of our model allows the addition of different compartments to the model, and also allows for the implementation in a variety of cases and countries, where in our experiments the model gives forecasts and parameter estimates for three different countries. Since this work focused mainly on the methodology of providing a robust recursive Bayesian estimation for the current nCov-2019 outbreak, we propose a further in depth-study of the parameter estimates and an extended comparative study across countries. Future work can add further complexities to the model, such as taking into account different mortality rates due to population age, cultural norms or quality of the healthcare system, providing applicability and robustness of the model for different datasets and scenarios.

We encourage both researchers and policymakers to run similar test results with data from other countries or on a more local level to estimate potential infection rates of outbreaks and the rate of transmission to implement the correct policy measures to contain and mitigate adverse effects of the pandemic.

Acknowledgements We are grateful for helpful discussions and feedback from Joseph Wu, Neil Ferguson and other participants at the Royal Society based DSI workshop "Scientists against CoViD-19 and beyond"

Appendix

Hybrid data assimilation

In this section we describe the computation of the covariance matrices as explained in Sect. 5. We estimate values for both the state and observation covariance matrices $\mathbf{Q}$ and $\mathbf{P}$ by using an ensemble approach [7, 19]. The values for $\mathbf{P}$ are based on an estimate of the residual covariance matrix of the stationary observed time series. Following the cost function give by Eq. 7, with $\mathbf{x}^b$ representing an individual background state vector, $\mathbf{x}^b = [S, I, T, R]$. The full ensemble of state vectors is given by

$$\mathbf{x}^b_{(1)}, \mathbf{x}^b_{(2)}, \dots, \mathbf{x}^b_{(N)} \tag{23}$$

If the ensemble mean is defined as $\overline{\mathbf{x}}^b$, then $\mathbf{V}_{ens}$, the background state perturbations are computed via

$$\mathbf{V}_{ens} = \mathbf{X}^b = \frac{1}{\sqrt{N-1}}(\mathbf{x}^b_{(1)} - \overline{\mathbf{x}}^b, \mathbf{x}^b_{(2)} - \overline{\mathbf{x}}^b, \dots, \mathbf{x}^b_{(N)} - \overline{\mathbf{x}}^b) \tag{24}$$

In this case, $\mathbf{V}_{ens}$ and $\mathbf{X}^b$ are a n x N matrix called the ensemble background perturbation matrix. The rank-deficient version of the background error covariance matrix is defined as $\mathbf{Q}^*$ with

$$\mathbf{Q}^* = \mathbf{X}^{bT}\mathbf{X}^b \tag{25}$$

The ensemble is static, meaning that it does not evolve dynamically with time, but it still incorporates flow-dependent information at the start time which is still beneficial for an extended Kalman filter or 4D analysis.

The way the ensembles are chosen and computed determines the accuracy of ensemble DA.

The ensemble needs to be computed in such a way that the time dependent variability of the background error covariance matrix, as well as the correlation of variables is captured by the sampling procedure.

The method we devise is to divide the collection of background states, $\underline{\mathbf{x}}^b$ based on the size of the ensemble into N equally sized groups with each group being denoted by $\underline{\mathbf{x}}^b_{(i)}$ meaning that ensemble members belong to the ith group. The mean and standard deviation of each group is

then estimated and used to sample the ensemble members from.

Algorithm 1 Build Ensemble

Inputs: $\underline{\mathbf{x}}^b$
$i = 0$, $N =$ ensemble size, $n = length(\mathbf{x}^b)$
for $\underline{\mathbf{x}}^b_{(i)}$ in $array\_split(\underline{\mathbf{x}}^b, N)$ **do**
  $\mu_{(i)} = mean(\underline{\mathbf{x}}^b_{(i)})$
  $\sigma_{(i)} = standard\_deviation(\underline{\mathbf{x}}^b_{(i)})$
  $ensemble[:, i] = normal\_distribution(\mu_{(i)}), \sigma^2_{(i)})$
  $i = i + 1$
end for
$ensemble\_mean = mean(ensemble)$
for $i = 0, 1, .., N$ **do**
  $\mathbf{V}_{ens}[:, i] = ensemble[:, i] - ensemble\_mean$
end for
return $\mathbf{V}_{ens}$

Algorithm 1 describes in detail how $\mathbf{V}_{ens}$ is computed and ensembles are formed. The full background state matrix, $\underline{\mathbf{x}}^b$ is split into N groups each of size $n \times \frac{n}{N}$. Both, the means as well as the standard deviations of the n rows are estimated and used to generate draws from a multivariate Gaussian distribution to form the ensemble. In order to form $\mathbf{V}_{ens}$, for each ensemble member the corresponding mean is estimated and then subtracted, computing the standard deviation. To put results into perspective we discuss the difference between standard assimilation and hybrid approaches by conducting a sensitivity analysis next.

References

1. Asch M, Bocquet M, Nodet M. Data assimilation: methods, algorithms, and applications. 2016;12
2. Imai N, Dorigatti I, Cori A, Riley S, Ferguson NM. Estimating the potential total number of novel coronavirus cases in wuhan city, china, 2020.
3. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, Ren R, Leung KSM, Lau EHY, Wong JY et al. Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia. New England J Med. 2020;382(13):1199–1207. https://doi.org/10.1056/NEJMoa2001316
4. Wu JT, Leung K, Leung GM. Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in wuhan, china: a modelling study. Lancet. 2020.
5. Rhodes CJ, Hollingsworth TD. Variational data assimilation with epidemic models. J Theor Biol. 2009;258(5):591–602.
6. Bettencourt LMA, Ribeiro RM, Chowell G, Lant T, Castillo-Chavez C. Towards real time epidemiology: data assimilation, modeling and anomaly detection of health surveillance data streams. NSF Workshop on Intelligence and Security Informatics. pp 79–90, 2007.
7. Wang X, Parrish D, Kleist D, Whitaker J. Gsi 3dvar-based ensemble-variational hybrid data assimilation for ncep global forecast system: single-resolution experiments. Monthly Weather Rev. 2013;141(11):4098–117.
8. Bonavita M, Hólm E, Isaksen L, Fisher M. The evolution of the ecmwf hybrid data assimilation system. Quart J R Meteorol Soc. 2016;142(694):287–303.
9. Bettencourt LMA, Ribeiro RM. Real time bayesian estimation of the epidemic potential of emerging infectious diseases. PLoS ONE. 2008;3(5):e2185.
10. Cobb L, Krishnamurthy A, Mandel J, Beezley JD. Bayesian tracking of emerging epidemics using ensemble optimal statistical interpolation. Sp Spatio Temp Epidemiol. 2014;10:39–48.
11. Li MY, Muldowney JS. Global stability for the seir model in epidemiology. Math Biosci. 1995;125(2):155–64.
12. Seibert J, Staudinger M, van Meerveld HJI. Validation and overparameterization—experiences from hydrological modeling. In: Computer Simulation Validation, Springer, 2019. pp. 811–834.
13. Anderson RM. Discussion: the kermack-mckendrick epidemic threshold theorem. Bull Math Biol. 1991;53(1–2):1.
14. Cuomo S, Galletti A, Giunta G, Marcellino L. Numerical effects of the gaussian recursive filters in solving linear systems in the 3dvar case study. Numer Math Theory Methods Appl. 2017;10(3):520–40.
15. Nadler P, Arcucci R, Guo Y-K. A scalable approach to econometric inference. In PARCO, 2019. pp. 59–68.
16. Arcucci R, D'Amore L, Pistoia J, Toumi R, Murli A. On the variational data assimilation problem solving and sensitivity analysis. J Comput Phys. 2017;335:311–26.
17. John hopkins university coronavirus resource center. https://coronavirus.jhu.edu/map.html, 2020. Accessed May 24, 2020.
18. National health commission of the people's republic of china. http://web.archive.org/web/20080207010024/. Accessed July 02, 2020.
19. Lim EM, Solana MM, Pain C, Guo Y-K, Arcucci R. Hybrid data assimilation: An ensemble-variational approach. In 2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), IEEE, 2019. pp. 633–640.