AN EPIDEMIOLOGICAL MODELLING APPROACH FOR COVID19 VIA DATA ASSIMILATION

A PREPRINT

Philip Nadler*
Data Science Institute
Imperial College London
London SW7 2AZ, U.K.
p.nadler@imperial.ac.uk

Shuo Wang
Data Science Institute
Imperial College London
London SW7 2AZ, U.K.
shuo.wang@imperial.ac.uk

Rossella Arcucci
Data Science Institute
Imperial College London
London SW7 2AZ, U.K.
r.arcucci@imperial.ac.uk

Xian Yang
Data Science Institute
Imperial College London
London SW7 2AZ, U.K.
xian.yang08@imperial.ac.uk

Yike Guo*
Data Science Institute
Imperial College London
London SW7 2AZ, U.K.
y.guo@imperial.ac.uk

April 28, 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 China, the US and Italy. In particular, we develop a custom compartmental SIR model fit to variables related to the epidemic in Chinese cities, named SITR model. 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. We use the model to do 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 · Compartmental Model

1 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 the virus and show how to estimate latent infection rates, accounting for high uncertainty in observation and model specification.

This is done by combining real-time Bayesian updating with epidemiological models. We develop a custom compartmental SIR model which is fit to data related to the spread of the coronavirus in China which we name SITR. This is embedded in a data assimilation framework, a form of recursive Bayesian estimation [1], which conducts model updates when new observations become available. A hybrid data assimilation approach 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

*Corresponding Author

well as the rate of recovery. The time-varying parameter structure of 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 hospitalized people in China. 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 Wuhan and find a peak of infections in the beginning of March. We furthermore apply the model to data to a selection of developed and developing countries and find that infections peak in the middle of April. 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 the empirical results and section six concludes.

2 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.

3 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.

3.1 The Standard SIR Model

We start our analysis with a standard SIR model [11], 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 in cities such as Wuhan, the susceptible number S is observable, which we label as the population of Wuhan city. The recovered population R denotes the population not infectious anymore and being removed from the population, which for the Wuhan outbreak is the number of confirmed cases since confirmed cases are hospitalized and isolated and not infecting the general population anymore.

3.2 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 [12]. Recently, DA is also applied to numerical simulations of geophysical applications, medicine, biological science and finance. Data assimilation can be applied to a variety of problems where an uncertainty quantification has to be included [13] 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 discretised 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 minumum of the cost function

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

and

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

where $H : I \to R$ is a linear transformation function usually called observation function [1] which is here represented 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 into the data. Data assimilation is very sensitive to initial conditions and the choices of the covariance matrices. Their calibration needs to be properly chosen.

The data we use representing S_t, I_t and R_t is given by the official government numbers for Wuhan and is available at [14]. 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 Wuhan. 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 by a wide margin, as is illustrated in the figure. Selected values of the graph are available in Table 1 where we compare estimated cases and 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 underestimation of infectious cases.

date	01-15	01-19	01-23	01-27	01-31	02-03	02-07
I_t^{DA}	131	286	1182	4778	16130	30744	49801
I_t	166	342	893	2333	6093	12514	32656
R_t^{DA}	74	177	498	1338	3532	7279	19040
R_t	54	132	635	2700	7615	14491	26767

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

This illustrates how without updating the parameters the number of infected people is underestimated 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.

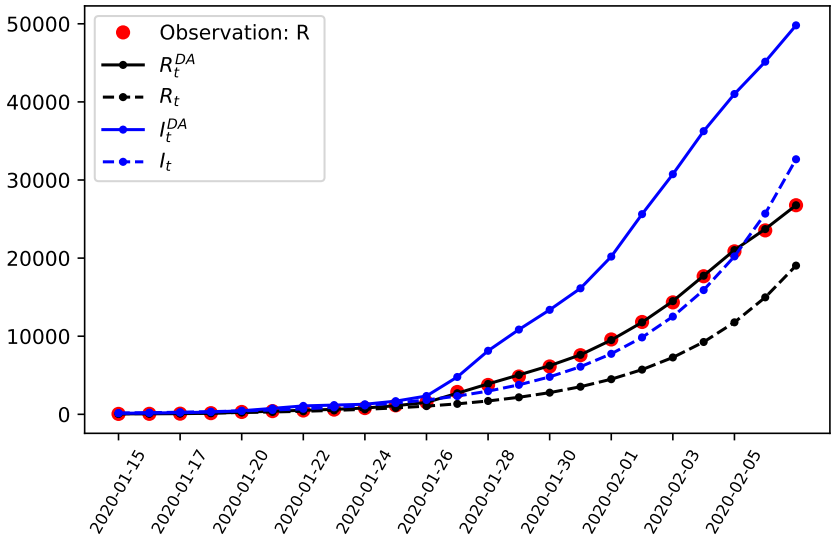

Figure 1: Comparing estimates of confirmed cases R and unobservable amount of infected people I in Wuhan, using static and updating parameters.

4 The Extended Epidemiological Assimilation Scheme

4.1 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

$$\frac{dS}{dt} = -\beta^e I \tag{9}$$

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

$$\frac{dT}{dt} = +\alpha I - \gamma T \tag{11}$$

$$\frac{dR}{dt} = \gamma T \tag{12}$$

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 β^e, α, γ time dependent, the SITR model in equations (9)-(12) can be discretised with respect to the time variable, giving the following equations:

$$S_{t+1} = S_t - \beta_t^e I_t \tag{13}$$

$$I_{t+1} = I_t + \beta_t^e I_t - \alpha_t I_t \tag{14}$$

$$T_{t+1} = T_t + \alpha_t I_t - \gamma_t T_t \tag{15}$$

$$R_{t+1} = R_t + \gamma_t T_t \tag{16}$$

which is a linearized approximation of the original SIR model with the additional compartment T. 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 $[t, t+\tau]$ and assuming to have observations of the variable T_t we denote here with T_t^{obs}, the resulting assimilation scheme is given by

$$I_t^{DA} = \underset{I}{\operatorname{argmin}} \sum_{i=t+1}^{t+\tau} ||T_t^{obs} - H(i|I, \beta_{t-1}^e)||_{\mathbf{Q}_t^{-1}} + ||I - I_t||_{\mathbf{P}_t^{-1}} \tag{17}$$

which in a first step infers the number of infected people I and where $H : I \rightarrow T$ is a linear transformation function usually called observation function [1]. To estimate the infection rate, in a second step we use minimize

$$\beta_t^e = \underset{\beta^e}{\operatorname{argmin}} = \sum_{i=t+1}^{t+\tau} ||T_t^{obs} - H(i|I, \beta^e)||_{\mathbf{Q}_t^{-1}} \tag{18}$$

which updates β conditioned on assimilated values of I. The resulting algorithm implements a 4D-VAR assimilation scheme in cost function (7), 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 **Q** and **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 **Q** and **P** before applying the assimilation scheme, which is named hybrid data assimilation.

4.2 Hybrid Data Assimilation

We estimate values for both the state and observation covariance matrices **Q** and **P** by using an ensemble approach [7]. The values for **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

$$\mathbf{x}_{(1)}^b, \mathbf{x}_{(2)}^b, .., \mathbf{x}_{(N)}^b \tag{19}$$

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

$$\mathbf{V}_{ens} = \mathbf{X}^b = \frac{1}{\sqrt{N-1}}(\mathbf{x}_{(1)}^b - \overline{\mathbf{x}}^b, \mathbf{x}_{(2)}^b - \overline{\mathbf{x}}^b, ..., \mathbf{x}_{(N)}^b - \overline{\mathbf{x}}^b) \tag{20}$$

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

$$\mathbf{Q}^* = \mathbf{X}^{b^T}\mathbf{X}^b \tag{21}$$

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 $\underline{\mathbf{x}}_{(i)}^b$ 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.

5 Results

5.1 Sensitivity Analysis

As we mentioned in Section 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. To illustrate the hyperparameter sensitivity we compare the number of estimated infected people and we apply a mean root squared forecasting error (MRSFE) metric:

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

where $\hat{y}_{t,n}$ represents the model prediction, $y^r_{t,n}$ the real observation with forecast horizons defined by $h = 1$, and $\tau_0 = 1$ the starting period of the forecast for n variables. The results are given in table 2. The model fit follows no clear pattern in the combination of covariance matrices, with the smallest forecasting error of 288.10 being a value of one for the observation noise and 100 for the model covariance matrix.

P Value	0.1	1	1	1	100	100
Q Value	0.5	100	1	10	1	10
Infections	13348	13580	13176	13320	13453	13199
MRSFE	321.90	288.10	329.48	316.59	328.64	314.11

Table 2: Sensitivity analysis for different values of observation and model error covariance matrices. The number of infected people predicted for the 12.2.2020

Fig. 2 depicts different infection curves given the results in table 2 and show that the dynamics are affected by the choice of the covariance matrices. The sensitivity of results confirm the need for a more rigorous choice of covariance estimates given the few and noisy datapoints, which we discuss next.

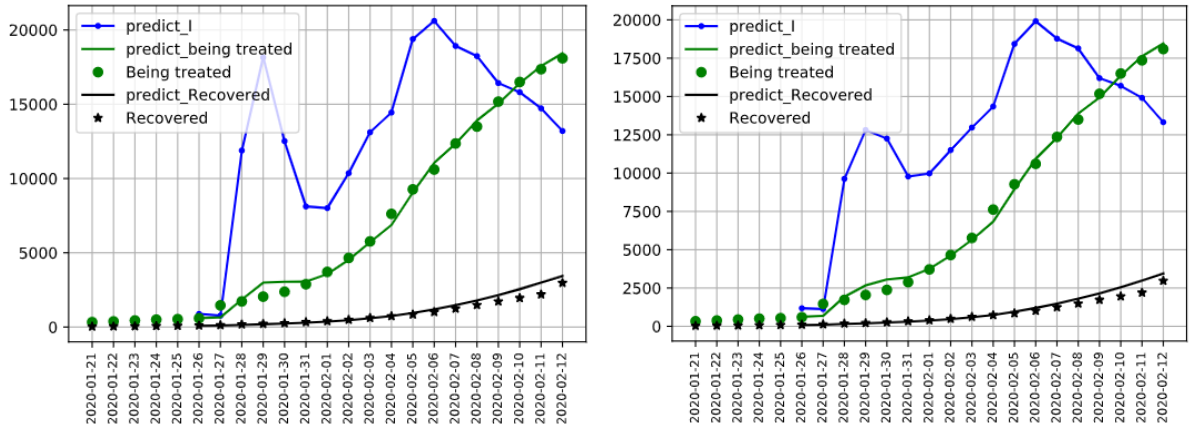

Figure 2: Estimated short run dynamics for $\mathbf{P}$ = 1 and $\mathbf{Q} = 10$ (left) and estimated short run dynamics for $\mathbf{P}$ = 100 and $\mathbf{Q} = 10$ (right)

Date	01-26	01-29	02-01	02-05	02-08	02-11
Infected Patients	1077	15587	7936	19389	18199	14899
Treated Patients	631	2778	3534	9069	13899	17634
RSFE	38.82	728.03	179.30	201.88	402.86	274.04

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.

5.2 Hybrid Assimilation Results

We present our results for the final hybrid assimilation scheme, and report the short- and long run predicted dynamics of the virus spread and show additional parameter estimates. We describe the results of the Hybrid ensemble covariance method, comparing it first with a naive unit covariance setup:

The results in the top left column of Fig. 3 show a peak of infections around early and mid February for Wuhan, with a strong dip in the January period. This is possibly due to an overweight of observation over model dynamics, showing that the initial unit covariance does not account for model and observation uncertainty. The introduction of the hybrid data assimilation method as given in the top right column of Fig. 3 makes the assimilation scheme more robust to initial values and introduces a more realistic and smooth development of the number of infected cases in Wuhan, without an unlikely strong decrease of infections end of January.

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 Wuhan. The total number of people being treated in hospitals follows with a small lag and will be reached in early march.

Comparing the left and right bottom figures of Fig. 3, the transmissibility rate β shows less variation over time, which differs from the model without ensembles where strong variation is visible. High variability implies that the transmissiblity 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 stable and unaffected by external policy factors within Wuhan city. Also 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 top left of Fig. 3 causes a large spike in forecasting errors for treated people end of January. Comparing it with a hybrid assimilation approach in table 4 reveals an overall lower number of forecasting error and better fit.

5.3 Other Chinese Cities

To compare the predicted dynamics of our model in regions hit less sever by the outbreak we extend the analysis beyond Wuhan. Following the same framework, we investigate the COVID-19 spread in Beijing. As given in Fig. 4, the infection rate β reached the high level 0.27 from Jan 26th to Feb 2nd, followed by a gradual decrease period with an

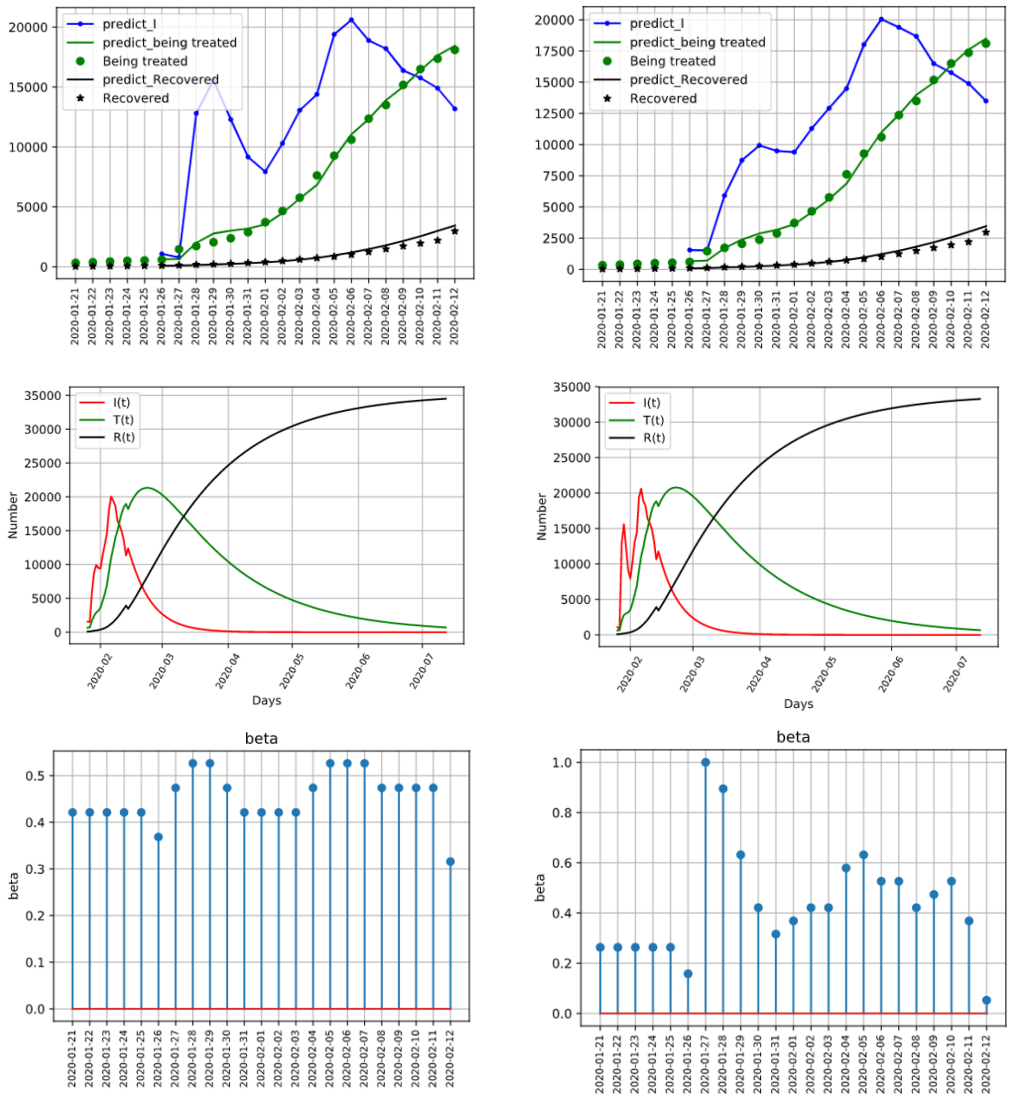

Figure 3: The left figures presents short- and longrun dynamics of the infection rates with a simple unit covariance matrix, the right figures present results using the hybrid assimilation approach

infection rate bottoming out at a value of 0.05. The predicted infected people in the community touched the peak on Feb 3rd with a daily incidence around 440. When comparing parameter estimates in Fig. 6 to the estimates of Wuhan, the β in Beijing is smaller and the incidence peak is 3 days earlier, which indicated the effects of intervention in early epidemic stages. The long range forecasts in Fig. 5 indicate a gradual decrease of the number of treated patients as well as the number of infections, resembling the dynamics for Wuhan on a lower level.
Compared to Wuhan, the peak of the infection occurs five days before Wuhan on the 2nd of February, showing that low early infection levels as well as quarantine measures introduced by the government led to a rapid decline of infection cases. To put results into perspective, the next section applies the methodology to international data, giving estimates of peaks of covid19 globally.

5.4 Analysis of International Data

To illustrate the flexibility of our approach we add a brief international comparison with additional results for multiple countries. Since country level data is more readily available than city specific data but lacks more granular data such as patients under treatment we conduct inference using the DA-SIR model. We focus solely on the number of infected

Date	01-26	01-29	02-01	02-05	02-08	02-11
Infected Patients	1552	8743	9397	18025	18685	14887
Treated Patients	650	2391	3613	8983	13982	17606
RSFE	57.25	341.86	100.66	287.33	485.25	246.29

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

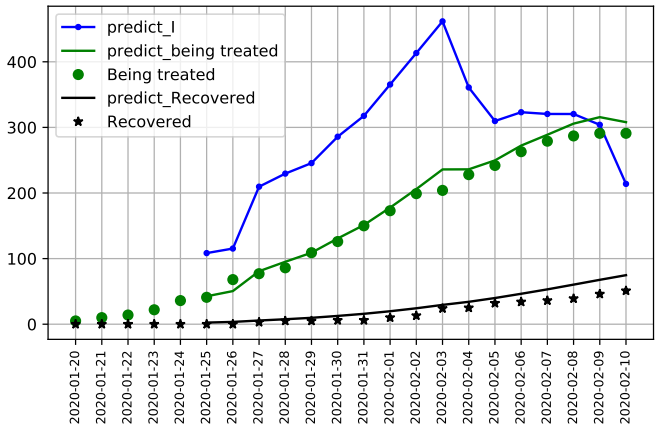

Figure 4: SITR results for Beijing, showing short run dynamics

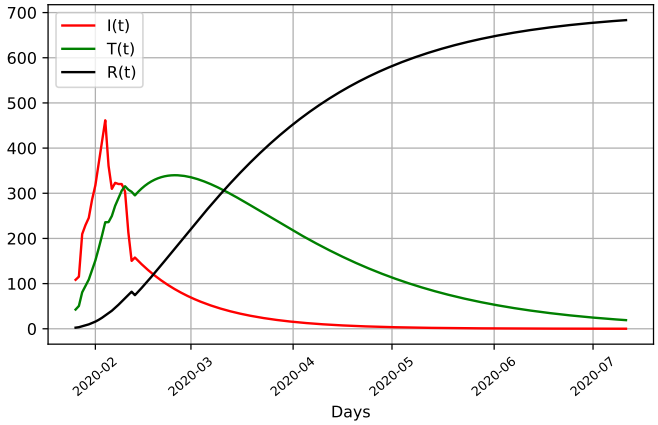

Figure 5: The long run dynamics of the epidemic in Beijing generated by the SITR model

people extrapolated from the number of confirmed cases and recoveries. The data was obtained from the John Hopkins University Coronavirus Resource Center[2]. We run test results for developed and developing countries. 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 sub-sample of the data, omitting observations after the 17/04/20 to evaluate the model out of sample.
We compare cases for the United States and Italy. Fig. 7 depicts the dynamics of the epidemic in Italy, where according to the model the peak of infections has already occured at the beginning of April, with a gradual decrease in infection numbers afterwards.

The predicted confirmed cases deviate most visibly from the observations during the peak period, since according to the SIR specifications a high absolute amount of infections affects the marginal increase in new infections. For nationwide estimates, constrained testing capabilities to confirm cases are unlikely to increase with the rate of infections, possibly

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

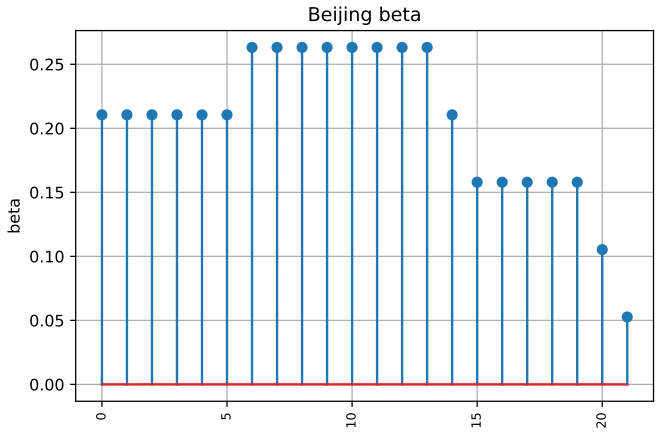

Figure 6: The beta estimates of the epidemic in Beijing, R-Ratio

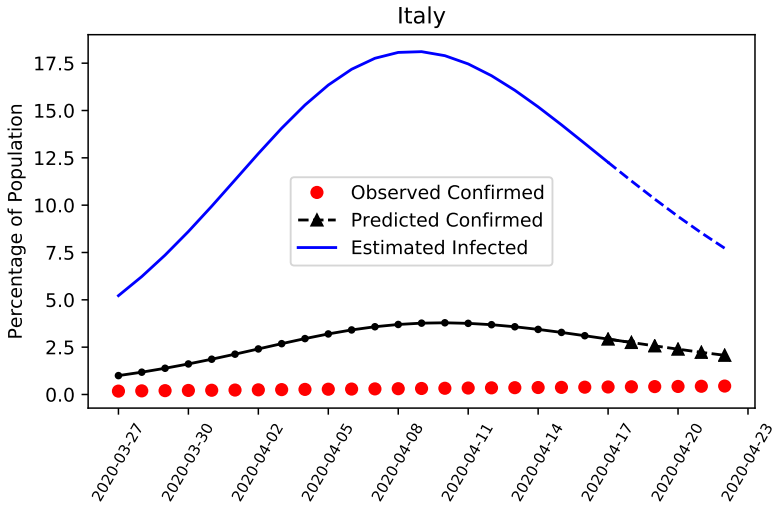

Figure 7: SIR results and forecasts for Italy, showing estimates of infections and confirmed cases.

explaining the difference in predicted and observed values. The forecasts are given in table 5. Forecasing errors on a nationwide level are larger compared to city bases estimates. In the data assimilation correction cycle forecasting errors are feed back into the model and lead to a correction of parameters which leads to model adjustments and decreases in infection estimates.According to the model estimates, the infection rate at the end of the sample is around 10 percent of the population, which is forecasted to drop to 6 percent in a 5 day ahead forecast.

Comparing results to the United States in Fig. 8, the trajectories of both countries are in different stages. The number

Date	04-18	04-19	04-20	04-21	04-22	Average
US RSFE	15.76e3	59.09e3	29.85e3	32.12e3	89.77e3	73.69e3
Italy RSFE	14.03e5	12.91e5	11.82e5	10.75e5	09.74e5	11.85e5

Table 5: Root squared forecasting errors in five day out of sample forecasting period for the United States and Italy.

of infections are increasing within sample, as is the forecasted number of infections. At the end of the sample, the infection rate is 0.6 percent of the population, which is forecasted to increase to around one percent within five days. The predicted confirmed cases are closer to observed values due to percent wise relatively low infection numbers compared to Italy, although the model predicts an increasing trend given current model dynamics. The trend is upward sloping, but the overall level as percentage of total population is lower than in Italy. The different levels of infections are likely due to different inception dates of the pandemic, as well as a high amount of uncertainty given very limited

testing capabilities in Italy and especially the United States.

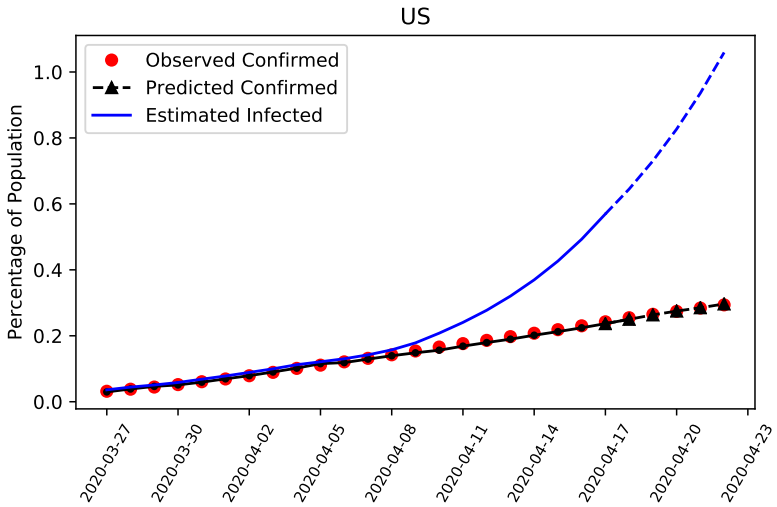

Figure 8: SIR results and forecasts for the United States, showing estimates of infections and confirmed cases.

The results indicate that the pandemic has reached a peak in Italy recently, the dynamics for the United States indicate that no plateau has been reached yet and that the number of infections is likely to increase. The results show how the assimilation framework can be extended to multiple countries and provide robust results given the large uncertainty in infection estimates.

6 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 Wuhan, China. 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 introduce 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 Wuhan the peak of infections is reached end of February, with the number of patients being treated peaking early march.
The generalisability of our model allows the model to be implemented in other cases and countries such as Italy and the United States where the model indicates further growth in the United States and a decline of total infection cases in Italy. 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 a comparative study across countries focusing on the epidemiological implications. 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 conference "Scientists against CoViD-19 and beyond"

References

[1] Mark Asch, Marc Bocquet, and Maelle Nodet. *Data assimilation: methods, algorithms, and applications*. 12 2016.

[2] Natsuko Imai, Ilaria Dorigatti, Anne Cori, Steven Riley, and Neil M Ferguson. Estimating the potential total number of novel coronavirus cases in wuhan city, china, 2020.

[3] Qun Li, Xuhua Guan, Peng Wu, Xiaoye Wang, Lei Zhou, Yeqing Tong, Ruiqi Ren, Kathy SM Leung, Eric HY Lau, Jessica Y Wong, et al. Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*, 2020.

[4] Joseph T Wu, Kathy Leung, and Gabriel M Leung. Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in wuhan, china: a modelling study. *The Lancet*, 2020.

[5] CJ Rhodes and T Déirdre Hollingsworth. Variational data assimilation with epidemic models. *Journal of theoretical biology*, 258(4):591–602, 2009.

[6] Luis MA Bettencourt, Ruy M Ribeiro, Gerardo Chowell, Timothy Lant, and Carlos Castillo-Chavez. Towards real time epidemiology: data assimilation, modeling and anomaly detection of health surveillance data streams. *NSF Workshop on Intelligence and Security Informatics*, pages 79–90, 2007.

[7] Xuguang Wang, David Parrish, Daryl Kleist, and Jeffrey Whitaker. Gsi 3dvar-based ensemble–variational hybrid data assimilation for ncep global forecast system: Single-resolution experiments. *Monthly Weather Review*, 141(11):4098–4117, 2013.

[8] Massimo Bonavita, Elias Hólm, Lars Isaksen, and Mike Fisher. The evolution of the ecmwf hybrid data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 142(694):287–303, 2016.

[9] Luis MA Bettencourt and Ruy M Ribeiro. Real time bayesian estimation of the epidemic potential of emerging infectious diseases. *PLoS One*, 3(5), 2008.

[10] Loren Cobb, Ashok Krishnamurthy, Jan Mandel, and Jonathan D Beezley. Bayesian tracking of emerging epidemics using ensemble optimal statistical interpolation. *Spatial and spatio-temporal epidemiology*, 10:39–48, 2014.

[11] Roy M Anderson. Discussion: the kermack-mckendrick epidemic threshold theorem. *Bulletin of mathematical biology*, 53(1-2):1, 1991.

[12] Salvatore Cuomo, Ardelio Galletti, Giulio Giunta, and Livia Marcellino. Numerical effects of the gaussian recursive filters in solving linear systems in the 3dvar case study. *Numerical Mathematics: Theory, Methods and Applications*, 10(3):520–540, 2017.

[13] Rossella Arcucci, Luisa D'Amore, Jenny Pistoia, Ralf Toumi, and Almerico Murli. On the variational data assimilation problem solving and sensitivity analysis. *Journal of Computational Physics*, 335:311–326, 2017.

[14] National health commission of the people's republic of china. `http://web.archive.org/web/20080207010024/`. Accessed: 2020-02-07.