Annual Reviews in Control xxx (xxxx) xxx

Contents lists available at ScienceDirect

Annual Reviews in Control

journal homepage: www.elsevier.com/locate/arcontrol

Monitoring and forecasting the COVID-19 epidemic in the UK

Peter C. Young [a,*], Fengwei Chen [b]

[a] Lancaster Environment Centre and the Data Science Institute, Lancaster University, UK
[b] School of Electrical Engineering and Automation, Wuhan University, Wuhan 430072, China

ARTICLE INFO

Keywords:
Monitoring
Forecasting
Recursive estimation
Fixed interval smoothing
Hybrid Box–Jenkins model
Dynamic harmonic regression
Dynamic linear regression
State-dependent parameter estimation

ABSTRACT

This paper shows how existing methods of time series analysis and modeling can be exploited in novel ways to monitor and forecast the COVID-19 epidemic. In the past, epidemics have been monitored by various statistical and model metrics, such as evaluation of the effective reproduction number, $R(t)$. However, $R(t)$ can be difficult and time consuming to compute. This paper suggests two relatively simple data-based metrics that could be used in conjunction with $R(t)$ estimation and provide rapid indicators of how the epidemic's dynamic behavior is progressing. The new metrics are the epidemic rate of change (RC) and a related state-dependent response rate parameter (RP), recursive estimates of which are obtained from dynamic harmonic and dynamic linear regression (DHR and DLR) algorithms. Their effectiveness is illustrated by the analysis of COVID-19 data in the UK and Italy. The paper also shows how similar methodology, combined with the refined instrumental variable method for estimating hybrid Box–Jenkins models of linear dynamic systems (RIVC), can be used to relate the daily death numbers in the Italian and UK epidemics and then provide 15-day-ahead forecasts of the UK daily death numbers. The same approach can be used to model and forecast the UK epidemic based on the daily number of COVID-19 patients in UK hospitals. Finally, the paper speculates on how the state-dependent parameter (SDP) modeling procedures may provide data-based insight into a nonlinear differential equation model for epidemics such as COVID-19.

1. Introduction

The best known COVID-19 mathematical models are the nonlinear epidemiological models, an example of which is the very well known, nonlinear *Susceptible, Infectious, Recovered* (SIR) differential equation model proposed by Kermack and McKendrick (1927). Such models can be deterministic, as in the original SIR model; stochastic, exploiting numerical Bayesian methods of analysis (see e.g. Flaxman, Mishra, & Gandy et al, 2020); or, stochastic, agent-based, microsimulation models, such as Hoertel et al. (2020) and Venkatramanan et al. (2018).

Such 'mechanistic' models are quite complicated, particularly the agent-based versions, because they try to explain the behavior of the epidemic, in various degrees of detail, as it develops in a country or region. While there are papers on more conventional statistical analysis (see e.g. Murray, 2020; Sahoo & Sapra, 2020) applied directly to epidemic time series, it is difficult to find any that consider epidemic data analysis using the data-based dynamic systems analysis methods that have evolved mainly within the systems and control community.

One exception is the paper by Lega and Brown (2016), which does not directly exploit any of the dynamic systems analysis methods used in the present paper but it does use an approach to data-based ('data-driven') modeling that is *similar* to the state-dependent parameter approach to modeling nonlinear stochastic systems (Priestley, 1980; Young, 1993b, 2000; Young, McKenna, & Bruun, 2001) that is utilized in Section 4 of the present paper. Lega and Brown exploit Matlab routines for time series analysis and the results are linked with the estimation of parameters in the well known logistic equation description of epidemic behavior (see e.g. Banks, 1994).

Given this existing epidemiological context, it must be emphasized that the present paper is not concerned with such mechanistic models except, as mentioned above, in Section 4, where nonlinear dynamic models are discussed in a more speculative manner. Rather, it discusses an alternative and totally data-based approach to the evaluation, monitoring, modeling and forecasting of COVID-19 time series. The time series models are identified statistically from the COVID-19 data and, except for the fact that they are continuous-time, differential equation models, they cannot, and should not, be compared with the epidemiological models. The aim of the paper is simply to exploit freely available methods of recursive time series analysis for either estimating the changes in certain metrics that can provide continuing insight into the behavior of the epidemic; or identifying models that can yield short-term (1–4 week) adaptive forecasts of the COVID-19 death series in the UK. These methods are intended as useful additions to the existing

* Corresponding author.
E-mail addresses: p.young@lancaster.ac.uk (P.C. Young), fengwei.chen@whu.edu.cn (F. Chen).

https://doi.org/10.1016/j.arcontrol.2021.01.004
Received 29 June 2020; Received in revised form 21 January 2021; Accepted 22 January 2021
Available online 18 February 2021

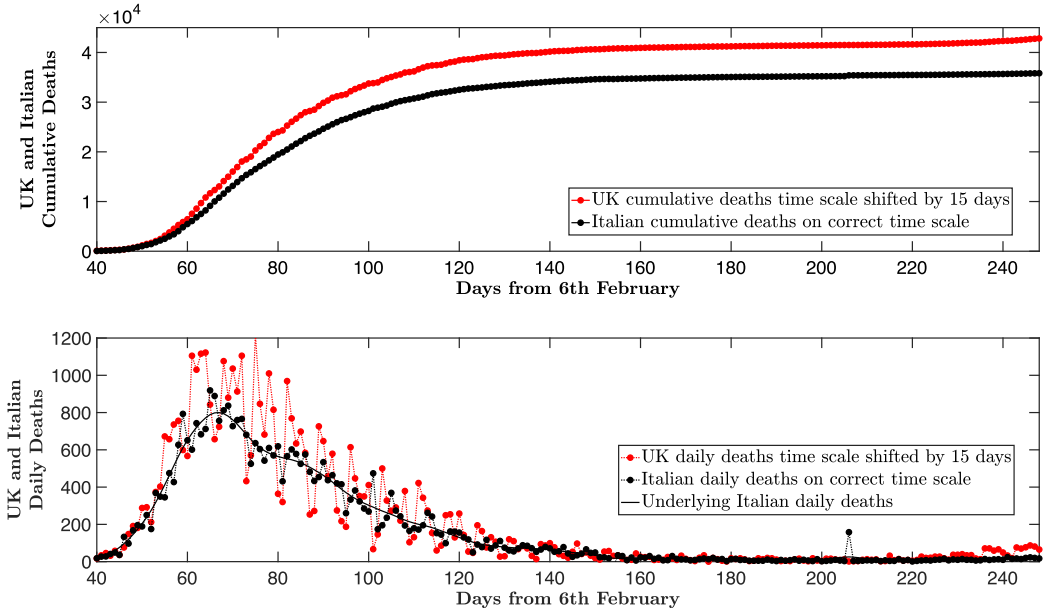

Fig. 1. Cumulative and daily UK Deaths data used in the paper. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

methods used in epidemic data analysis and, together with the results obtained from epidemiological modeling, it is felt that such information could be helpful in the monitoring, forecasting and management of an epidemic.

Section 2 of the paper considers the COVID-19 epidemic as a classical, noisy, dynamic system whose progress can be monitored by two new metrics: one based on optimally smoothed estimation of the epidemic *Rate of Change* (RC); and the other on a related state-dependent *Response rate Parameter* (RP). These metrics are evaluated and discussed in relation to the daily confirmed cases of COVID-19 and the daily deaths arising from these in the UK and Italy.[1] A typical example of these data is shown in Fig. 1, where the UK and Italian data are plotted in red and black colors, respectively. Note that the UK data, in particular, are characterized by a changing weekly cycle caused by the methods of data collection and delays that occur in this operation, particularly at weekends. In this figure, the upper panel shows the cumulative series, while the lower panel shows the daily deaths. The UK epidemic started 15 days after the Italian epidemic so the UK series has been shifted by 15 days to illustrate the similarity in the series.

Although there is a weekly cycle in the Italian series, it is most pronounced in the UK series and, clearly, it needs to be accounted for in the data analysis. Consequently, both metrics are based on optimal time-variable parameter *Dynamic Harmonic Regression* (DHR) estimation, which provides a model of the measured series $y(k)$ that includes an estimate of the changing weekly cycle. In particular, the DHR algorithm decomposes $y(k)$, at sampling instant k, where $k = 1, 2, \ldots, N$ and N is the sample size, into a number of 'unobserved components' that include this weekly cyclical behavior $W(k)$, as well as the underlying changes in the series being considered, $x(k)$, and additive noise, $\xi(k)$.

Effectively, $x(k)$ is an optimally smoothed estimate of the changes in $y(k)$ after the weekly cycle and noise have been removed. Most importantly for the purposes of the present paper, the DHR analysis also yields an optimal estimate of the continuous-time temporal rate of change $\frac{dx}{dt}(k)$, which defines the RC metric at the kth sampling instant.

The time-variable estimates of the temporal changes in the RP metric are based on the same DHR estimates but use these in association with the *Dynamic Linear Regression* (DLR) algorithm, as discussed in Section 2.2 of the paper, where Eqs. (1) and (7) show that RP is a state-dependent parameter, usually referred to in the biological literature as the 'specific growth rate'. The DHR and DLR estimation is carried out using the dhr and dlr routines in the CAPTAIN Toolbox[2] for Matlab, in conjunction with their associated dhropt and dlropt optimization routines.

Section 3 of the paper exploits the similarity of the Italian and UK epidemics and shows how similar methodology to that used in Section 2, combined with the refined instrumental variable method for estimating hybrid Box–Jenkins *Transfer Function* (TF) models of linear dynamic systems (RIVC), as implemented in the CAPTAIN rivcbjid and rivcbj routines, can be used to relate the epidemics in Italy and the UK. This TF model, which is simply an operational representation of a differential equation model between the Italian and UK time series, is then used to provide adaptive, 15-day-ahead forecasts of the UK epidemic on a 'rolling' basis, i.e. the forecast is updated continually every few days, with the model re-estimated at each update.

The linear dynamic models in Section 3 yield useful forecasts but these are based on the fortuitous similarity of the UK and Italian epidemics and the associated 15 day lead time that this provided over the whole of the first wave of the COVID-19 epidemic in the UK. As a new wave of the virus began in early October, 2020, it became quite clear that this lead time had evaporated and the approach was no longer useful. It is important to note, however, that the TF modeling approach discussed in this section can be used in any situation where there are sufficient data available for a linear TF relationship to be identified between the measured time series, such as the relationship between the daily number of COVID-19 hospital patients and daily COVID-19 deaths in the UK, which is discussed at the end of Section 3.

Section 4 outlines briefly the results of initial, on-going research that considers the data-based identification of a *State-Dependent Parameter* (SDP) nonlinear model for forecasting the UK deaths series. This SDP

[1] Most of the COVID-19 data used in this paper were downloaded from the GitHub Web Site at https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series.

[2] The CAPTAIN Toolbox is a free toolbox that can be downloaded from http://wp.lancs.ac.uk/captaintoolbox/. From hereon in this paper, the term 'CAPTAIN' will be used when making reference to the CAPTAIN Toolbox.

model is self-generating from specified initial conditions and, unlike the method used in Section 3, it does not rely on any fortuitous relationships with other series.

2. The metrics and their estimation

Most of the optimal estimation methods used in this paper utilize an approach to time series analysis that exploits optimal recursive *Fixed Interval Smoothing* (FIS) (Jazwinski, 1970; Norton, 1986; Young, 1984, 2011) and so yields an estimate of the metric that is available immediately upon receipt of new data. As such, it can provide additional, useful and powerful, indicators to enhance the information provided by $R(t)$ estimation, which can be difficult and time consuming to compute (Cori, Ferguson, Fraser, & Cauchemez, 2013).

The recursive FIS estimation is an essential element in the estimation of the time-variable parameters that characterize the DHR and DLR models (Young, Pedregal, & Tych, 1999), where the nature of the 'unobserved components' in these models are allowed to change over time to reflect changes in the time series being analyzed. In this paper these are the daily confirmed cases of the COVID-19 virus infection and the daily deaths arising from these, considered initially in terms of the number of daily deaths in the UK population as shown in Fig. 1. For comparative purposes, however, these data are standardized in relation to the population of the country being considered, so that they refer to the deaths (or confirmed) cases per million in the population.

It is widely recognized that the behavior of epidemics can be described by a nonlinear differential equation which, for a single variable, takes the general form:

$$\frac{dx(t)}{dt} = r(t)x(t) \tag{1}$$

where the response rate parameter RP, denoted by $r(t)$, changes as a function of the changes in the state of the system; i.e. it can be considered as a 'state-dependent parameter' (see also the later Section 4). The state of this system can be defined in various ways and has been the subject of much research over many years. However, this paper does not address the nature of the any such state equations until later, in Section 4, where it is considered in initial, speculative terms. Rather it considers directly the dynamic nature of the response $x(t)$ and the associated rate parameter $r(t)$, based on the daily measured response $y(k)$ where, on the kth day,

$$y(k) = x(k) + \xi(k) \tag{2}$$

Here, $\xi(k)$ represents disturbances that affect our observation of the 'underlying', but not directly observable, response $x(t)$. These are caused mainly, in the present context, by a pronounced and changing weekly cycle associated with the method of data collection.

2.1. The Rate of Change (RC) metric

Both metrics are obtained by considering the observed series $y(k)$ as the sum of unobserved components that are assembled in the following DHR model:

$$y(k) = T(k) + W(k) + e(k) \qquad e(k) = \mathcal{N}(0, \sigma^2) \tag{3}$$

In the present context, $T(k)$ is a low frequency 'trend' component; while $W(k)$ is the cyclical component that models the weekly cycle and is modeled as:

$$W(k) = \sum_{i=1}^{R_c} \left\{ \alpha(i,k)\cos\left(\omega_i k\right) + \beta(i,k)\sin\left(\omega_i k\right) \right\} \tag{4}$$

where each $\alpha(i,k)$ and $\beta(i,k)$ is a stochastic *Time-Variable Parameter* (TVP) and $\omega_i, i = 1, 2, \ldots, R_c$, are the fundamental and harmonic frequencies associated with the cyclicity in the series; here in the region of 0.286 and 0.143 cycles/day, associated with harmonic periods of 3.5 and 7 days in a weekly cycle. Finally, the irregular component $e(k)$ represents the stochastic variations in $y(k)$ that have not been explained by all the other components. Normally, if the spectral peaks have been identified well, then it will be a zero mean, normally distributed and serially uncorrelated, white noise process. Each TVP in the DHR model is assumed to be one member of the *Generalized Random Walk* (GRW) family. The model is optimized using the CAPTAIN dlropt routine, to yield the optimal hyper-parameters required to generate the estimates of the unobserved components. The GRW model and the associated hyper-parameters are discussed briefly in Appendix.

In the case of the COVID-19 data, the $T(k)$ component accounts for the underlying changes in the series, and is represented by the estimate $\hat{x}(k)$ of $x(k)$ in Eq. (2). The DHR analysis also provides a daily sampled estimate $\frac{d\hat{x}}{dt}(k)$ which is an estimate of $\frac{dx(t)}{dt}$, the rate of change of the daily deaths, $x(t)$ in Eq. (1), on the kth day. This is the RC metric and, as we shall see, it provides revealing insight into the behavior of the epidemic, with $\frac{d\hat{x}}{dt}(k)$ values greater than zero representing growth in the epidemic; values less than zero indicating its decline; and zero positive values showing where the epidemic has peaked or reached a period of no growth. Indeed, this derivative measure has 'phase advance' properties so that, as we shall see, it can be helpful in providing prior warning of the peak being achieved.

2.2. The Response rate Parameter (RP) metric

Given the DHR estimates of the underlying epidemic response $x(t)$ and its rate of change $\frac{dx(t)}{dt}$, it is possible to formulate the following model for the relationship in Eq. (1)

$$\frac{d\hat{x}}{dt}(k) = r(k)\hat{x}(k) + \eta(k) \qquad \eta(k) = \mathcal{N}(0, \sigma_\eta^2) \tag{5}$$

where $\frac{d\hat{x}}{dt}(k)$ and $\hat{x}(k)$ are the daily sampled values of the DHR estimates $\frac{d\hat{x}(t)}{dt}$ and $\hat{x}(t)$, while $\eta(k)$ is the associated measurement noise. If $\hat{x}(k)$ is noise-free (see later) and $\eta(k)$ is white noise, as shown, then Eq. (5) will be recognized as a TVP regression relationship: a special, single parameter example of the DLR model (see section 5.1 in Young, 2011):

$$y_r(k) = \sum_{i=1}^{np} \alpha_i(k) z_i(k) + e(k); \;\; e(k) = \mathcal{N}(0, \sigma_e^2) \tag{6}$$

where $y_r(k)$ is the dependent variable; the $z_i(k), i = 1, \ldots, np$ are the regression variables; and $\alpha_i(k)$ are the TVPs. In this simple example, $np = 1$, $\alpha_i(k) = r(k)$ and $y_r(k) = \frac{d\hat{x}}{dt}(k)$. As in the case of the DHR model, the TVPs $\alpha_i(k)$ can be selected from the GRW family of random walk models and optimized accordingly, in this case using the CAPTAIN dlropt routine. The resulting estimate $\hat{r}(k)$ of $r(k)$ is then obtained using the dlr routine and this provides the required temporal changes in the RP estimate. Once again, $r(k)$ has a critical value of zero marking the boundary between the growth and decline of the epidemic.

One feature of the DLR estimation in Eq. (5) is that it is using FIS estimates of the variables involved in this regression relationship, as obtained from the DHR model estimation. As a result, the explanation of the dependent variables $\frac{d\hat{x}}{dt}(k)$ and $\hat{x}(k)$ is virtually perfect and so the confidence bounds on the RP metric are small, suggesting more confidence than is justified given the noisy nature of the measured data. For this reason, the confidence bounds shown later in Fig. 3 are adjusted, using Monte Carlo analysis based on the estimated standard errors in the $\hat{x}(t)$ and $\frac{d\hat{x}(t)}{dt}$, to provide a more realistic idea of this uncertainty.

Finally, another alternative to using DLR estimation is to consider the estimate obtained directly from

$$r(k) = \frac{\frac{d\hat{x}}{dt}(k)}{\hat{x}(k)} + \zeta(k) \tag{7}$$

using logarithmic transforms to compute the ratio of variables on the right-hand side of this equation. The use of this transform requires care, however, both because of its behavior when the series is near zero and because it can result (as in the present case) in a residual series that does not conform very well to the statistical estimation requirements. For these reasons, this approach is not considered further.

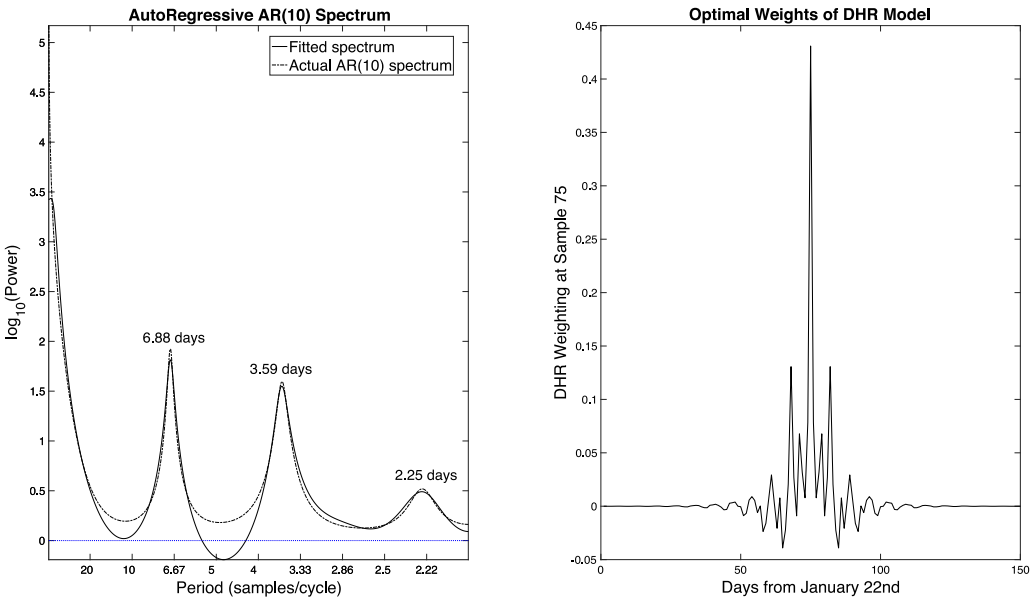

Fig. 2. Dynamic Harmonic Regression (DHR) model optimal spectrum and weights.

2.3. Analysis of the UK COVID-19 data

The DHR and DLR estimation procedures described in the previous sub-section are computationally very rapid and provide a virtually instantaneous estimate of the changing metrics every day, as soon as the latest update in the number of deaths is received. Fig. 2 presents two results that illustrate the nature of this analysis for the UK deaths series. The left panel shows how the optimization of the hyper-parameters for the DHR model produces a model pseudo-spectrum that matches the 10th order *AutoRegressive* (AR) model spectrum, as identified using the *Akaike Information Criterion* (AIC: see Akaike, 1974). Here, the spectral peaks are associated with the weekly cycle, clearly defined at 6.88 and 3.59 days, respectively. In addition, there is a smaller, short period peak at 2.25 days which helps to explain the color in the signal that is not explained by the cyclical component. The right panel shows the optimal weights generated by FIS for the DHR estimation at the 75th sample location (the response of the optimized DHR model to a unit impulse at sample 75). In effect, this FIS 'wavelet' moves through the series and extracts the cyclical component, with adjustments to its shape as either end of the data is approached (this is achieved by a two-stage recursive estimation procedure: see Young, 2011 for details).

Following DHR and DLR estimation, the estimates of the metrics RC and RP can be plotted to see how they are varying over time. An example of this is shown in Fig. 3, which provides an illustration of how the RC metric changes over the 150 day period and how these changes can be interpreted. Here, RC is plotted as a red line with its estimated uncertainty shown in light gray. Two regions are marked in light colors: a 'Danger Zone' in light red where RC $>$ 0 and the death numbers are increasing in some manner; and a 'Safe Zone' in light green, where RC $<$ 0 and the numbers are decreasing. These results are obtained from analysis of the measured UK daily deaths data, standardized to measure the deaths per million people and plotted as black dots. Such standardization is used for all of the results presented subsequently and, in this case, the UK population is rounded to 68 million. The magenta line is the full DHR estimate but its uncertainty bounds are not shown in order to simplify the plot; while the blue line with associated dark gray uncertainty bounds, is the DHR estimate of the underlying daily deaths $x(k)$, which is smooth and very effectively corrected for the cyclical factors. This same presentation of the results, with the associated line definitions and colors, is used in the subsequent Figs. 5 to 7, although the background red and green colors are omitted. Fig. 3 also shows, as a thick black line, the well known centralized 7 day moving average, which appears to be the preferred measure of the underlying behavior in epidemiology and has been shown increasingly in the material published by the UK Government.[3] It is taking the weekly cycle into account in a much cruder manner than DHR, so it is more volatile and, because of the way it is computed, it cannot be estimated at the beginning and end of the series.

The capital letters in Fig. 3 indicate the progress of the epidemic, as monitored by the RC metric: SE indicates the initial rise in RC at the start of the epidemic in the UK, at around sample 55 on March 17th 2020. PP shows where RC has detected the maximum rate of change before the epidemic starts to move towards its peak, thus predicting that the peak is being approached; LP shows the location of the peak, defined as RC crosses zero at sample 80 on April 11th. The epidemic is now declining, with RC less than zero but the metric reaches a minimum value around the April 25th and then starts to rise towards zero, showing that the decline is slowing down. And it eventual reaches zero again at sample 124 on May 25th, indicating that the epidemic numbers are about to start increasing again. Fortunately, this is predicted by RC at location LPP as a local peak, which is attained at sample 130 on June 1st; and the epidemic (if defined in terms of the daily death numbers) then declines up to sample 150 on June 20th, when the first version of this paper was finalized.

In the actual day-to-day monitoring, the results are obtained based only on the data up to each day and a movie film would be required to illustrate the full progression of the epidemic in these terms. In the absence of this, all we can do is to consider the results plotted in Fig. 4, which span each day from May 31st to June 4th and then a jump to June 8th, In this sequence of plots, we see that, on May 31st, the RC has moved very near to the zero boundary, providing a warning that something may be going to happen; which, indeed, it does on the next day when RC becomes quite strongly positive, suggesting a possible secondary epidemic. It is interesting that this conclusion coincided with the discussion in the news at the time about the increase of the $R(t)$ metric towards unity, with similar connotations, demonstrating how RC can be considered as a quick indicator of the changes in $R(t)$.

However, on the next two days, June 2nd and 3rd, a small peak forms, predicting a subsequent peak in the death numbers and a suggestion that this is a local peak, not a continuing rise in death numbers.

[3] https://coronavirus.data.gov.uk/#category=nations&map=rate&area=n92000002.

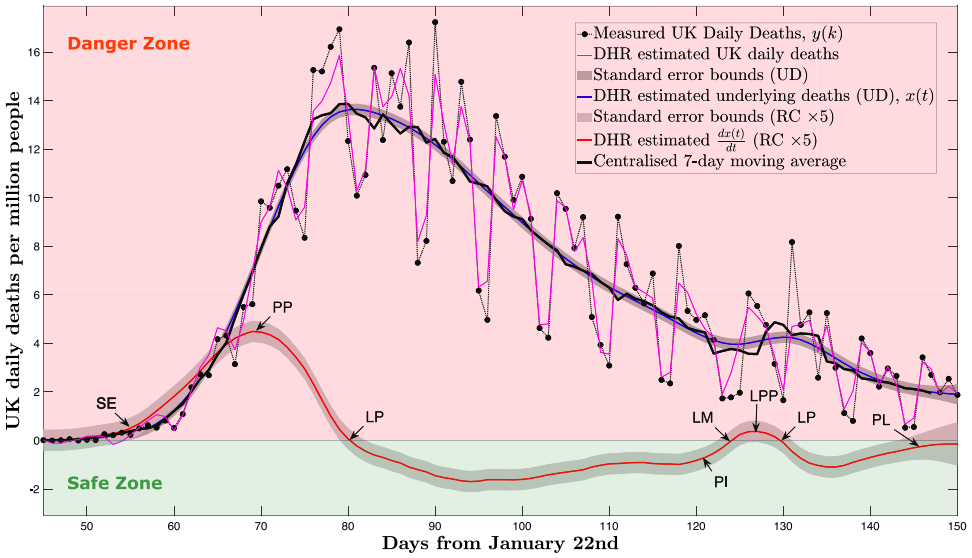

Fig. 3. Progress of the UK epidemic monitored by the RC metric. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

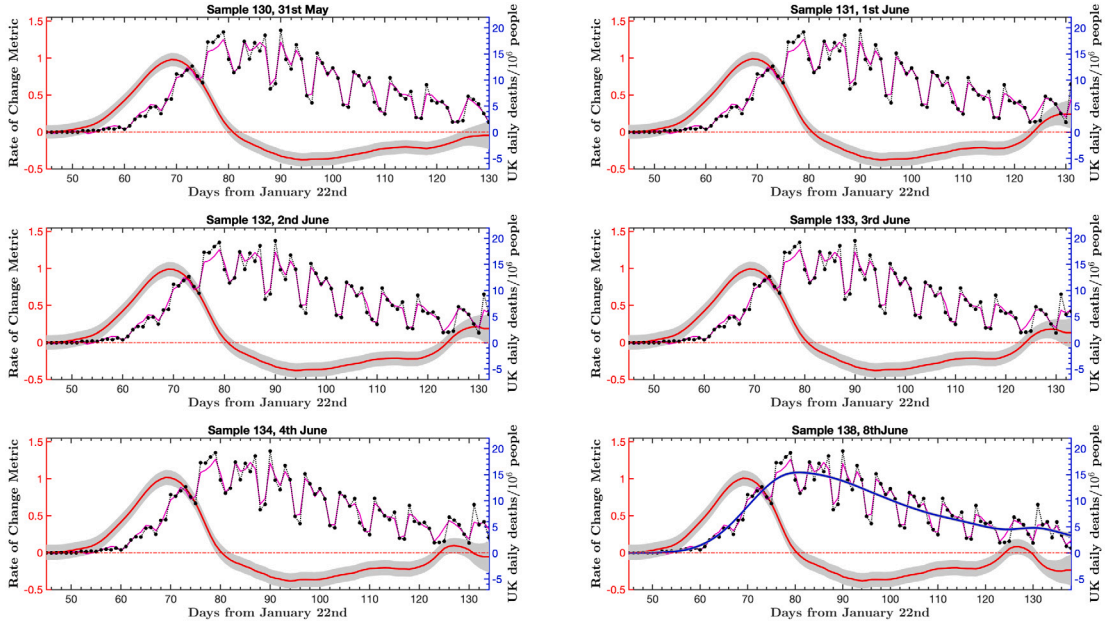

Fig. 4. Sequential progress of the RC metric applied to UK Deaths data.

The remaining two plots confirm this with the peak fully developed on June 4th, as RC goes through zero, back into the 'Safe Zone'. Finally, note that this peak is at the same location when all the data up to June 20th are analyzed (see below), showing the accuracy and stability of the monitoring as time progresses.

Fig. 5 shows the RC and RP results for the UK up to June 20th with RC, as considered above, in the right panel and the RP metric, for comparison, in the left panel. The DHR forecasts for each metric, which are plotted from a forecast origin of June 8th, are very reasonable, with the DHR model generating the weekly cycles quite well. It is clear that the RP metric provides similar results to RC, except that the estimation takes longer to settle down, with quite large standard error bounds to start off with, although these become quite narrow and are less than those for RC after this.

Fig. 6 shows what has happened to the RC metric when applied to the deaths (right panel) and confirmed cases (left panel) in the UK epidemic for the whole time up to a forecasting origin FO on September 27th. After the section of the epidemic shown in Fig. 5, the deaths and confirmed cases continued to drop and the deaths appeared to be leveling out, signaling an end to the epidemic and giving rise to a relaxation in the restrictions that had applied over the lock-down period. However, the RC metric for the confirmed cases moved into the danger zone at the location shown by the red arrow in the left-hand panel and, two days later, the confirmed cases started to rise sharply, a trend that has continued, with only small down-turns from time-to-time, until the end of the plot. On the other hand, the number of deaths continued to drop with the RC metric hovering around zero. On August 31st, however, at the location shown by the red arrow in the right-hand panel, there was an ominous, sustained rise of RC into the danger zone, as shown in the enlarged inset on the plot. This heralded a subsequent initial rise in deaths that soon produces a second wave of the epidemic (see also Section 4). This is consistent with the very large rise in confirmed cases, shown in the left-hand panel, and an associated rise in hospital admissions (not shown). In addition to the

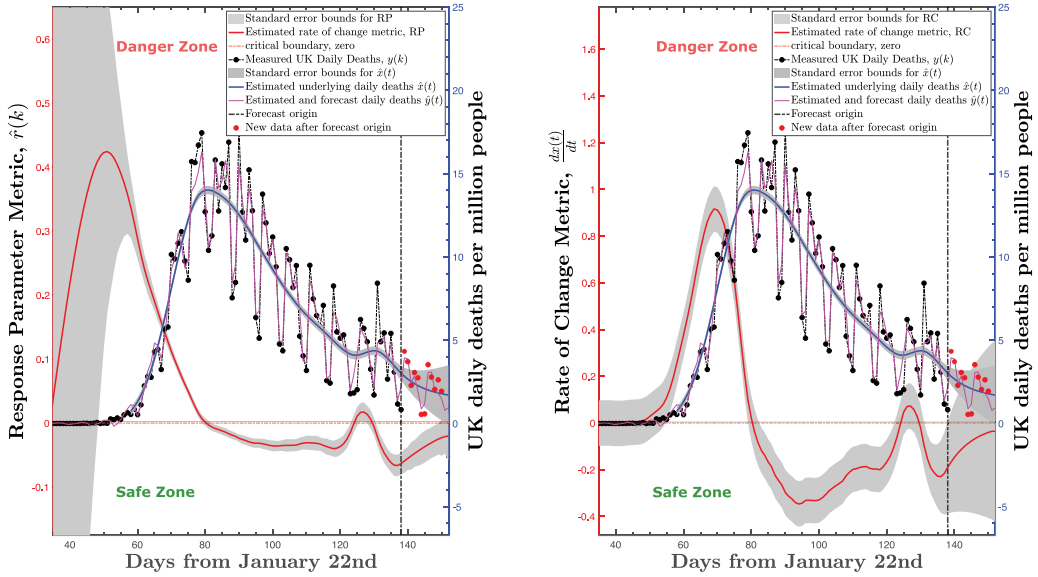

Fig. 5. Both metrics applied to UK Deaths data. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

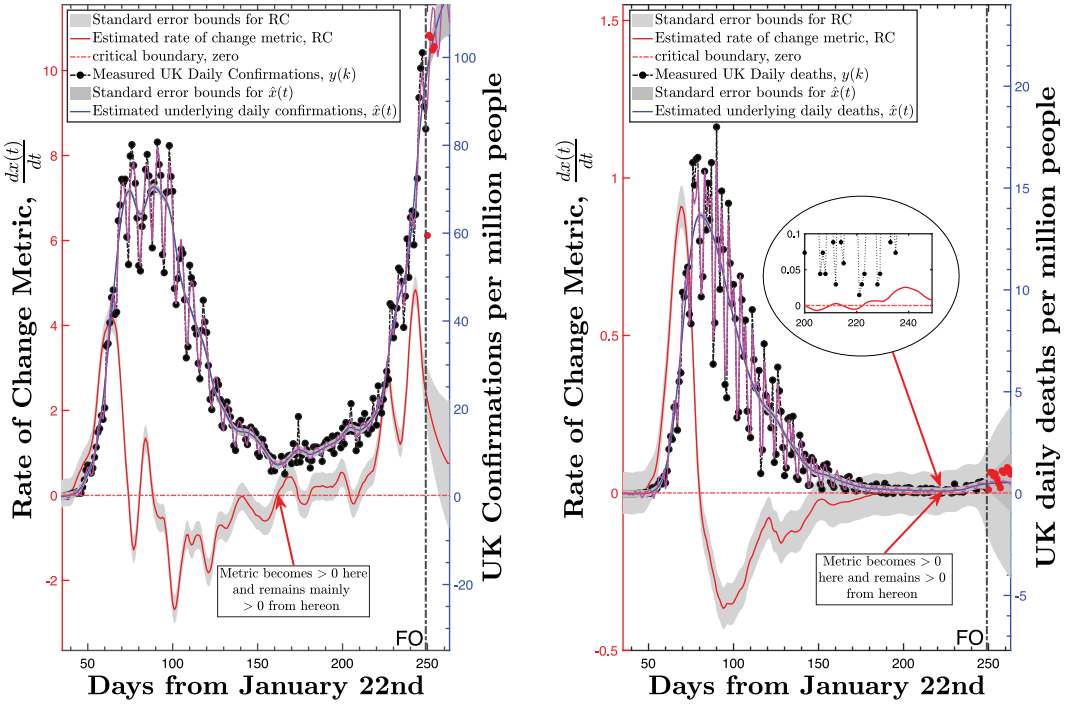

Fig. 6. Rate of Change metric applied to UK Confirmed and Deaths data. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

expected hospitalization lags in the system, the fact that it took so long for the death numbers to increase when the confirmed cases were rising so sharply has been put down to the inaccuracy of the test-and-trace system over this period and that a lot of the confirmed cases were amongst younger people who did not develop dangerous symptoms that required hospitalization.

2.4. The situation in other countries

Fig. 7 shows the results for Italy, again considering only the RC metric. Here, the population used in the standardization is 60.5 million. Comparing Figs. 5 and 7, we see that the Italian epidemic progressed in a similar manner to that in the UK, except that it preceded it by 15 days. Indeed, in the next Section 3, this similarity is exploited to produce adaptive, rolling forecasts of the progress in the UK epidemic. Some sharp upward movements in the metrics while they are in the 'Safe Zone' gave some cause for concern at the time and some consequent slowing in the decline of the epidemic. Otherwise, this decline continued steadily until the numbers of confirmations and deaths were both relatively low. As in the UK, however, these low numbers did not last and, soon after the RC metric became positive, the confirmed cases started to rise, with the metric remaining predominantly positive and the changes becoming very significant towards the end of the plot. As in the UK, however, the death numbers in the right-hand panel remain

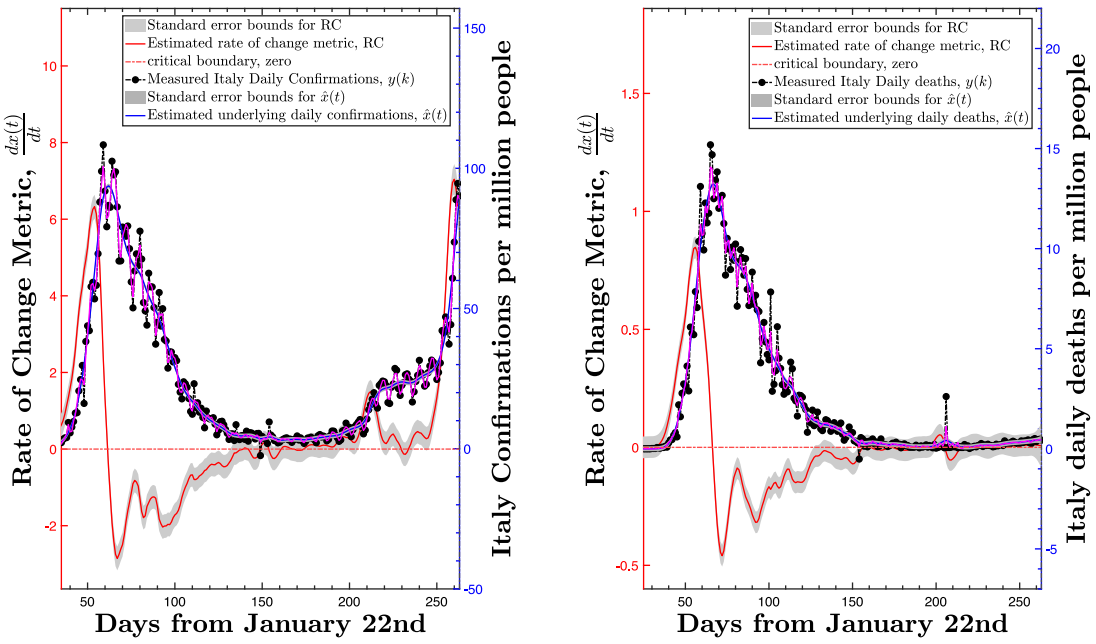

Fig. 7. Rate of Change metric applied to Italian Confirmed and Deaths data. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

steady at low numbers, with a just perceptible rise towards the end of the plot. Clearly, the close similarity between the Italian and UK epidemics in the first wave of the epidemic, with the Italian deaths leading those in the UK by 15 days, is no longer applicable and the second wave of the UK epidemic in now slightly leading that in Italy.

Similar analysis to that shown above has been applied to the data from Germany and the USA and this is available in a technical note (Young, 2020).

3. Exploiting the relationship between the Italian and UK epidemics

Fig. 1 has shown the similarity of the Italian and UK epidemic behavior in the first wave of the epidemic, but with the UK having larger numbers of confirmations and deaths, which is partly consistent with the slightly larger UK population (68 *cf* 60.5 million, a ratio of 1.124). Given this similarity, it is interesting, over this first wave, to consider relating the two epidemics in some manner. The simplest approach is to estimate a DLR model of the following form between the cumulative death numbers:

$$y_c(k) = \beta_0 + \beta_1(k) z_c(k) + e(k) \tag{8}$$

where $y_c(k)$ and $z_c(k)$ are, respectively the cumulative deaths in the UK and Italy.

When the dlr routine in CAPTAIN is applied to the cumulative data, it produces the estimates of the $\beta_1(k)$ shown in Fig. 8, with the left panel showing the estimate obtained under the assumption that the parameter is constant; while, in the right panel, the parameter is assumed to vary, with the hyper-parameter defining this variation optimized by the dlropt routine. It is clear from these plots that this parameter needs to vary: although the estimate initially converges to a value of about 1.2, it increases after this, reaching a value of 1.253 by the end of the data (because the data are shifted to accommodate the 15 day difference between the two epidemics, the plot shows the days from February 6th). This rise is due to the transient increase in the UK death numbers that occurred between March 25th and 31st, i.e. days 109 and 115 (as discussed in Section 2.3, with the associated plot in Fig. 3), which resulted in a permanent increase in the cumulative death numbers.

This DLR model can be used directly for modeling and forecasting the daily death series in the UK by exploiting the 15 day lead provided by the Italian data to achieve 15-day-ahead forecasting. However, the same objective can also be achieved using another, more interesting approach that utilizes linear dynamic model identification. This is based on the following hybrid model (see Young, 2015) for the UK daily death series $y(k)$ shown below in Eq. (9). Here, the underlying model output $x(t)$, in (9)(i), is the output of a continuous-time TF model; while the discrete-time output equation, in (9)(ii), defines $y(k)$ as the sum of $x(k)$, the sampled value of $x(t)$ on the kth day; the weekly cycle component $W(k)$, obtained from the DHR model; and the residual noise, modeled as a standard discrete-time *AutoRegressive Moving Average* (ARMA) process.

$$\begin{aligned} x(t) &= \frac{B(s)}{A(s)} u(t) && (i) \\ y(k) &= x(k) + W(k) + \frac{D(z^{-1})}{C(z^{-1})} e(k) && (ii) \\ e(k) &= \mathcal{N}(0, \sigma^2) && (iii) \end{aligned} \tag{9}$$

Here, $s = \frac{d}{dt}$ is the derivative operator; z^{-1} is the backward shift operator; $e(k)$ is zero mean white noise with variance σ^2; and $u(t)$ represents the underlying behavior of the Italian series, as identified using an associated hybrid model of the form:

$$\begin{aligned} u(t) &= \frac{B_I(s)}{A_I(s)} I(t) \\ y_I(k) &= u(k) + \xi(k) \end{aligned} \tag{10}$$

in which $y_I(k)$ are the daily samples of the Italian data; $I(t)$ is a unit impulse input signal; and $\xi(k)$ is additive noise. The full model identification procedure then consists of the following four simple steps:

1. Identify the structure of the model (10) and estimate the associated parameters.

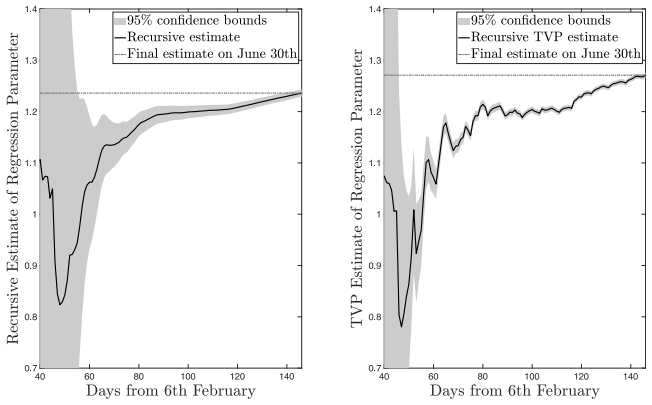

Fig. 8. Recursive estimation of the regression parameter in Eq. (8).

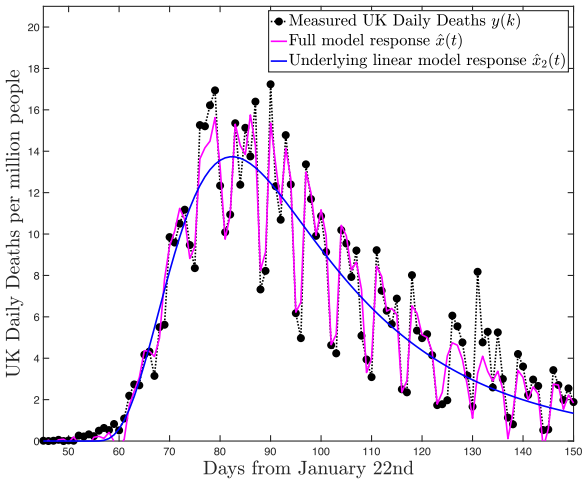

Fig. 9. Comparison of the model (12) response and the standardized COVID-19 daily death numbers in the UK. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2. Identify a continuous-time, linear TF model for the underlying behavior $x(t)$ of the UK series by employing the deterministic output $\hat{u}(t)$ of the model in step 1 as the input signal (thereby avoiding 'errors-in-variables' problems (Söderström, 2007; Young, 2011) caused by the weekly cycle and noise on the Italian series: see Remark 3.1 below). The response of this model is denoted by $\hat{x}_1(t)$.
3. Identify a DHR model for the error $\hat{W}(k) = y(k) - \hat{x}_1(k)$ associated with the model obtained in step 2. The sampled model output is then $\hat{x}(k) = \hat{x}_1(k) + \hat{W}(k)$, where $\hat{W}(k)$ is the deterministic output of the DHR model and an estimate of the weekly cycle.
4. Remove $\hat{W}(k)$ from the measured $y(k)$ to yield the 'de-cycled' series $y_{dc}(k) = y(k) - \hat{W}(k)$; and then repeat step 2 using this series in place of $y(k)$.

Remark 3.1. It may seem odd to use the unit impulse input model (10) to generate an estimate of the underlying behavior in the Italian series but this works well. This is because the Italian series, in this first wave of the epidemic, resembles an impulse response and can be considered as the response to the impulse injection of the COVID-19 virus into Italy. This approach handles the errors-in-variables problems by ensuring that there are no errors on the input $\hat{u}(t)$; errors that would arise from the weekly cycle and the measurement noise $\xi(k)$ if these were retained by direct use of the Italian series as the input to the model (9), rather than $\hat{u}(t)$. Both of these errors-in-variables effects on the input could cause problems of bias on the parameter estimates and so have a deleterious effect on the forecasting.

Remark 3.2. Step 4 is required to refine the estimate of the TF model which, with the estimated cyclical component removed, takes the form;

$$\begin{aligned} x(t) &= \frac{B(s)}{A(s)}\hat{u}(t-\tau) && (i) \\ y_{dc}(k) &= x(k) + \frac{D(z^{-1})}{C(z^{-1})}e(k);\ e(k) = \mathcal{N}(0,\sigma^2) && (ii) \end{aligned} \tag{11}$$

This is now in the standard hybrid Box–Jenkins model form (for details, see Young, 2015) and so it can be identified by RIVC estimation using the rivcbj routine in CAPTAIN.

Using the above procedure, the model (11) is identified as:

$$\begin{aligned} x(t) &= \frac{b_0}{s^2 + a_1 s + a_2}\hat{u}(t-7) && (i) \\ y_{dc}(k) &= x(k) + \frac{1}{C(z^{-1})}e(k); && (ii) \\ e(k) &= \mathcal{N}(0,\sigma^2) && (iii) \end{aligned} \tag{12}$$

where the estimated parameters in the main TF (12)(i) are as follows, with the standard errors shown in parentheses:

$$\begin{aligned} &\hat{a}_1 = 0.250(0.016);\ \hat{a}_2 = 0.0433(0.0025); \\ &\hat{b}_0 = 0.0456(0.0028);\ \hat{\sigma}^2 = 0.226; \end{aligned} \tag{13}$$

In this case, the ARMA noise model is identified as a simpler AR process, with polynomial $C(z^{-1})$ of 7th order (estimate not shown). The associated residual series $e(k)$ is uncorrelated both serially and with the input signal $\hat{u}(t)$, as required.

The DHR model of the weekly cycle, as identified using the CAPTAIN dhr routine, has harmonic components with periods of 6.95, 3.54, 2.48 and 2 days and the associated hyper-parameter vectors, as optimized by the associated dhropt routine are:

$$\begin{aligned} \mathbf{nvr} &= [0.056\ 0.010\ 0.094\ .024]^T \\ \boldsymbol{\alpha} &= [0.98\ 0.97\ 0.81\ 0.77]^T \end{aligned} \tag{14}$$

Although this identification procedure is sub-optimal in statistical terms, Fig. 9 shows that the resulting model explains the data well, except for the short period when there is a transient rise in the series, as discussed earlier in Section 2.3, that cannot be accounted for by this model. The overall coefficient of determination, based on the model response $\hat{x}(k)$ is $R_T^2 = 0.985$, i.e. 98.5% of the variance in the measured data $y(k)$ is explained by the deterministic output of the model; and the variance of the error $y(k) - \hat{x}(k)$ associated with this is 0.25. Most importantly in the present context, the model provides reasonable adaptive 15-day-ahead, rolling forecasts of the UK data, again by exploiting the 15 day lead time provided by the Italian series. This is illustrated by the three plots in Fig. 10, which show 15-day-ahead forecasts made from forecast origin locations spaced at 7 day intervals.

There are various measures of forecast accuracy used in the forecasting literature but two of these are:

1. the *Root Mean Square Error* (RMS)

$$\text{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[y(k+i) - \hat{y}(k+i)]^2} \tag{15}$$

2. the forecast *Coefficient of Determination* (CoD or R_T^2) is defined as:

$$R_T^2 = 1 - \frac{1}{f}\sum_{i=1}^{f}\frac{[y(k+i) - \hat{y}(k+i)]^2}{[y(k+i) - \bar{y}]^2} \tag{16}$$

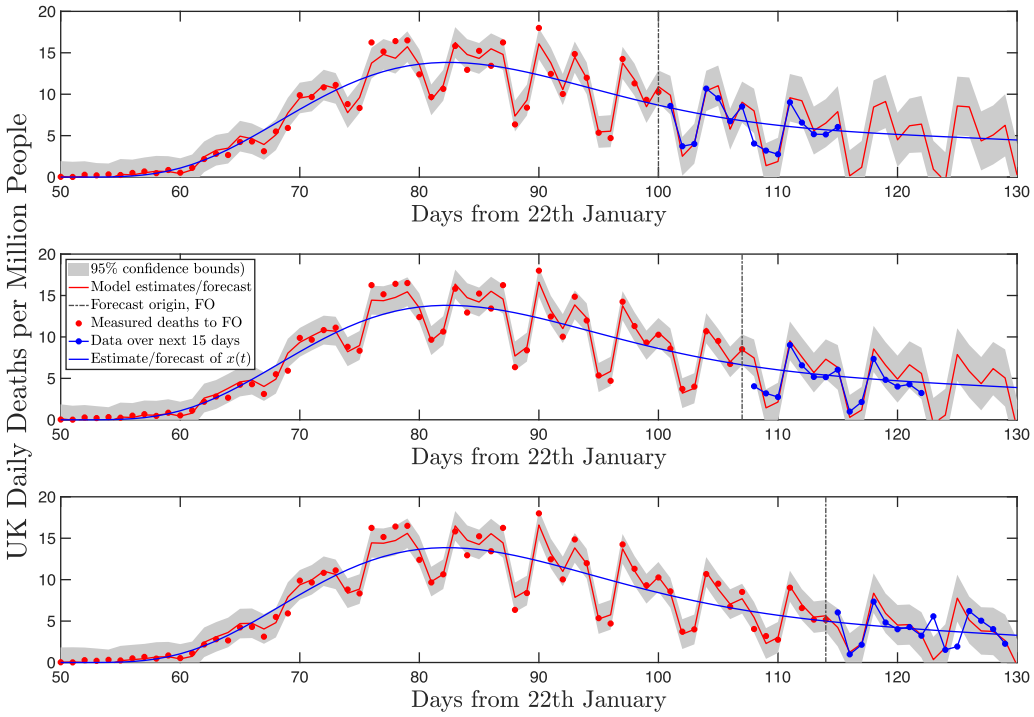

Fig. 10. Adaptive, rolling estimation and forecasting results for UK daily death numbers during the COVID-19 epidemic. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Comparison of forecast accuracy measures.

7-Day-Ahead Forecast from:	R_T^2 Model fit	R_T^2 Forecast	RMS
May 1st	0.985	0.849	0.98
May 7th	0.982	0.375	1.75
May 14th	0.986	0.800	0.92

where $\bar{y}$ denotes the mean value of $y(k)$. In 2., the second term on the right-hand side of the equation provides measure of the ratio of the forecasting error variance to the variance of the UK deaths time series. Consequently, when this is subtracted from unity, the closer R_T^2 is to unity, the better the forecast in these variance terms. This is also called the 'Skill Score' in weather forecasting (Murphy, 1995) and the *Nash–Sutcliffe model Efficiency coefficient* (NSE) in hydrological modeling and forecasting (Nash & Sutcliffe, 1970). Note that this R_T^2 measure is more often used as a measure of how the estimated model explains all the data. i.e. with $f = N$, where N is the sample size used for estimation.

A 7-day-ahead forecast is probably the most useful in practical terms because the forecast begins to deteriorate a little after this. Table 1 shows the R_T^2 and RMS measures for this forecast interval at the three locations in Fig. 10. The first column reports the R_T^2 values for the whole of the data, *including the forecast*: these all show a good explanation of the data, with values all greater than $R_T^2 = 0.98$, i.e. 98% of the daily deaths explained by the model and its forecast. The second column shows the R_T^2 values for just the 7-day-ahead forecast: here, the forecasts from May 1st and May 14th are quite good, with $R_T^2 > 0.8$. The forecast from May 7th is not so good but visually the forecast is quite acceptable. The poorer measures are largely due to the higher forecasting errors on the first and last days and: e.g. the R_T^2 and RMS measures are improved to 0.61 and 1.27, respectively, if the first day's forecast is removed. Finally the third RMS column confirms the results in the second column, with the smallest RMS errors for the May 1st and May 14th forecasts.

Given the changing nature of the weekly cycle, the forecasting performances illustrated by this table and Fig. 10 are quite reasonable and, together with the RC metric, provide a good idea of what to expect over the next two weeks of the epidemic, as required for management and decision making during an epidemic such as COVID-19. It is felt that such results could enhance the information gained from more traditional epidemic models, particularly as they account so well for the changing weekly cycle.

Finally, it is important to note that the transfer function modeling approach discussed in this section can be used in any situation where there are sufficient data available for a linear TF relationship to be identified between the measured time series. One interesting illustration of this approach is based on the estimated TF relationship between the data on the number of COVID-19 patients in UK hospitals[4] and the subsequent COVID-19 deaths. This example is discussed fully in a technical note (Young, 2020) but Fig. 11 shows an example of three-week-ahead forecasts obtained in this manner for two dates, seven days apart, with the actual UK daily death numbers after the forecast origin (FO) shown as blue dots. These forecasts exploit the identified 15 day pure delay in the TF model, with the additional days forecast by the associated DHR model of the weekly cycle. As in the case of the Italy–UK analysis, there is clear evidence of parametric change in the model, so that adaptive, rolling forecasting of this kind is necessary.

4. What is the differential equation model of the epidemic?

This is not a question that can be answered in the present paper. Research on a fully data-based answer to this question is on-going but, based on initial results, it is interesting to speculate on what kind of dynamic model might be appropriate. Before this, however, it is worth considering briefly the nature of more traditional epidemiological models that are normally identified using the traditional 'Popperian' hypothetico-deductive approach (Popper, 1959). Here the hypothetical model assumed by the modeler on the basis of epidemiological theory is related to data from epidemics in various ways. Probably the simplest,

[4] Data obtained from https://www.ecdc.europa.eu/en/publications-data?s=hospital+admissions.

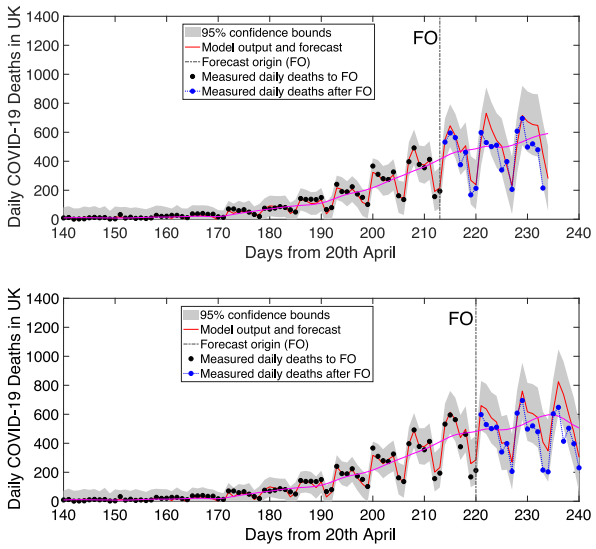

Fig. 11. Two examples of the UK COVID-19 forecasting results based on the number of COVID-19 patients in UK hospitals. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

well known model of this type is the SIR model proposed by Kermack and McKendrick (1927):

$$\begin{aligned} \frac{dS}{dt} &= -\frac{\beta I S}{N}, \\ \frac{dI}{dt} &= \frac{\beta I S}{N} - \gamma I, \\ \frac{dR}{dt} &= \gamma I, \end{aligned} \tag{17}$$

where S is the number of the susceptible people in the population, I is the number of the infected people, R is the number of the people 'removed', either by death or recovery, and N is the sum of these three.

An alternative approach to modeling is the hypothetico-inductive procedure proposed by the present first author (Young, 2013). The associated *Data-Based Mechanistic* (DBM) modeling methodology has been used successfully in the modeling of dynamic systems in a number of different scientific disciplines (see e.g.Young, 1993a, 1998, 2018). In hypothetico-inductive DBM modeling, the idea is to follow Isaac Newton's philosophy of '*hypotheses non fingo*' i.e.[5]

> *... avoid initial, possibly prejudicial assumptions about the nature of the model and to attempt to infer this from the data*

Consequently, prior hypotheses are considered, but only *after* a solely data-based model has been identified and estimated directly from the available data. And then, an existing model hypothesis is only accepted if it can be reconciled with the DBM model. A good illustration is the DBM model for an Australian blow-fly (*Lucilia Cuprina*) population (Young, 2000), where the DBM model is found to be consistent with the hypothetical model suggested previously by Gurney, Blythe, and Nisbet (1980).

Such a DBM modeling approach will not be pursued in detail here but, given the previous discussion in Section 2.2 on the state-dependent response parameter $r(t)$, it is interesting to consider the standard first step in DBM nonlinear system modeling: SDP estimation (Young, 2000, 2011), which provides initial insight into the nature of a purely data-based nonlinear differential equation model of an epidemic.

[5] see reference 1 in https://en.wikipedia.org/wiki/General_Scholium.

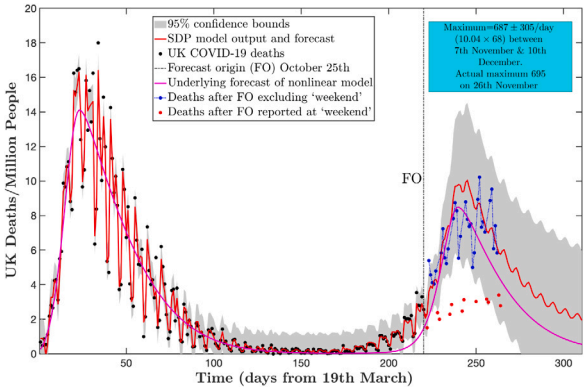

Fig. 12. SDP model response and forecast. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In the case of the UK COVID-19 epidemic, initial research has identified the following SDP differential equation:

$$\frac{d^2x(t)}{dt^2} = a_1\left\{\chi_1(t)\right\}\frac{dx(t)}{dt} + a_2\left\{\chi_2(t)\right\}x(t) \tag{18}$$

where the second derivative $\frac{d^2x(t)}{dt^2}$ is estimated from the DHR estimated $\frac{dx(t)}{dt}$, using FIS estimation based on an *Integrated Random Walk* (IRW) model (see Young, 2011, section 4.5.4; and the Appendix), as coded in the CAPTAIN routine irwsm. In Eq. (18), $a_1\left\{\chi_1(t)\right\}$ and $a_2\left\{\chi_2(t)\right\}$ are both parameters that are assumed to vary as a function of state, or other, variables, on which they may depend. These variables are specified in the vectors of variables χ_1 and χ_2. In particular, identification analysis based on the UK daily deaths series has found that both are dependent on the first derivative, i.e. the $\chi_i(t)$, $i = 1, 2$ are single variables with $\chi_1(t) = \chi_2(t) = \frac{dx(t)}{dt}$. In particular,

$$\begin{aligned} a_1 &= \alpha_{1p} \text{ and } a_2 = \alpha_{2p} + \beta_{2p}\frac{dx(t)}{dt} \text{ if } \frac{dx(t)}{dt} > 0 \\ a_1 &= \alpha_{1n} \text{ and } a_2 = \alpha_{2n} + \beta_{2n}\frac{dx(t)}{dt} \text{ if } \frac{dx(t)}{dt} < 0 \end{aligned} \tag{19}$$

where the parameters are estimated as follows:

$$\begin{aligned} &\hat{\alpha}_{1p} = 0.2548(0.010);\ \hat{\alpha}_{2p} = \hat{\beta}_{2p} = 0.0151(0.0009) \\ &\hat{\alpha}_{1n} = -0.1777(0.085);\ \alpha_{2n} = \beta_{2n} = -0.0070(0.004) \end{aligned} \tag{20}$$

and again the standard errors are shown in parentheses.

This SDP model explains the UK data well, as shown in Fig. 12, where, as in the previous Section 3, a DHR model is incorporated to handle the weekly cycle and the forecast is a combination of the SDP model and the DHR model forecasts. The figure also shows a forecast made by the model on October 25th which anticipates quite well the daily deaths for the 44 days up to December 7th, when this paper was being finalized. The plot shows the deaths after the forecasting origin (FO) as blue and red dots: the blue dots are measurements made from Tuesday to Saturday each week, while the red dots are those made on Sunday and Monday. The latter data illustrate a clear 'weekend' effect, where the significantly lower values are presumably the result of changed procedures over the weekend; while, as a result, the Tuesday measurements are always the highest in the week because those missed over the weekend are added to the total.

Note that the amplitude of the weekly cycle increases quite rapidly after the forecast origin, leading to 3 measurements outside the 95% bounds. Also, the forecast does not capture the very large movements in the measured weekly cycle because it is based only on the data up to the FO, where the cycle is significantly smaller. The model was simulated in Simulink and estimated by nonlinear least squares optimization using the lsqnonlin routine in Matlab and data up to the forecast origin FO. Note that, although the model is self-generating from specified initial

conditions, it was found necessary to stimulate the second wave of the epidemic by a very small step input of 0.0002 applied on August 24th (equivalent to only one additional death every 73 days). Both the date of application and the magnitude of the impulse were included in the optimization and August 24th happened to coincide with the first signs of problems arising from the relaxation after 'lock-down'.[6]

It must be emphasized that this model is being developed for short-term adaptive forecasting with the model parameters updated as new data are received. The longer term forecast, with its large confidence bounds, is shown in Fig. 12 simply to illustrate that the response displays epidemic-type behavior, recognizing that both the date and magnitude of this second peak may be poorly defined. Indeed, quite a number of the Monte Carlo realizations, as used to evaluate the effect of parametric uncertainty on the 95% forecasting bounds, displayed further epidemic waves after this second one, as data in early 2021 have in fact revealed.

On the other hand, the forecast in Fig. 12 is probably quite reasonable in comparison with other forecasts generated at this FO time. Moreover, its inherent ability to generate multiple epidemic waves means that the model makes reasonable sense in nonlinear dynamic terms. Also, the individual epidemic response is as expected: for the relatively short periods (24% of the time) when $\frac{dx(t)}{dt}$ is greater than zero, the model is exponentially unstable; while, when it is less than zero, the model is stable, with a response quite similar to a second order linear system (demonstrating why it was possible to identify the linear second order model in Eq. (12) of Section 3).

In accordance with the objectives of this paper, the SDP model (18) is completely data-based. However, it provides a basis for further hypothetico-inductive DBM modeling, where it is hoped that, given additional data, superior SDP models may be developed and investigated further in epidemiological terms to see how they relate to theoretical nonlinear growth and epidemic models that have been suggested previously in the epidemiological literature (e.g. Heesterbeek & Roberts, 2015 and the prior references therein; Banks, 1994).

5. Conclusions

The main objective of this paper has been to evaluate whether existing and powerful estimation and modeling procedures, developed largely within a systems, control and forecasting context, can be exploited to good advantage in monitoring and forecasting the behavior of COVID-19 epidemics, particularly the severe one that has so badly affected the inhabitants of the UK since the middle of March 2020. The results obtained in the paper suggest that the time-variable parameter regression algorithms (DHR and DLR), which rely heavily on optimal recursive fixed interval smoothing algorithms, provide additional monitoring and forecasting information that should be useful in the management of epidemics such as COVID-19. In particular, these algorithms have been used to obtain estimates of the two metrics, RC and RP, that allow for continuing monitoring of the epidemic's progress, unhindered by the weekly cycles and noise affecting the measured time series on which they are based, as well as prior indications of changes occurring in these series.

The paper has also shown how the refined instrumental variable method of continuous-time model identification (RIVC) can be used to identify and estimate hybrid transfer function models (i.e. continuous-time differential equations with discrete-time difference equations for additive noise) that describe the dynamic relationship between epidemic-related time series. It has also shown that, if these models contain suitably long pure time delays, they can be exploited successfully for forecasting purposes. Other analysis, not reported in the present paper, has shown how the recursive RIVC algorithm, as well as its counterpart for fully discrete-time models RIV, can be used in other ways that allow the parameters in the model to change: e.g. to relate the confirmed and death series, where the dynamic relationship is changing as time progresses. This is a current subject of research by the authors, who have applied such TVP estimation methods in other application areas (see e.g. Padilla, Garnier, Young, Chen, & Yuz, 2019)

The initial data-based mechanistic modeling analysis, obtained using state-dependent parameter estimation, has yielded promising results. These suggest that this approach may be able to identify a parametrically efficient nonlinear epidemiological model that can be compared with the existing nonlinear simulation models. In order to be fully successful, however, such a study would require access to detailed information on the management actions taken (or not taken) during the epidemic. For example, such information would be required in any attempt to explain the transient growth period that occurred during late March in the UK (see Section 2.3), or the second wave of the UK epidemic discussed in Section 4.

Finally, it is clear that the various statistical estimation procedures discussed in this paper are not known within the epidemiological and medical communities, who rely on traditional simulation modeling methods and statistical tools, such a regression analysis and centralized moving averaging. Hopefully, the present paper may come to their attention and encourage them try out the various algorithms that have been used in the paper, all of which are available in the CAPTAIN Toolbox for Matlab, which can be downloaded without cost from the web site noted in the paper.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We are grateful for the comments of the referees, which helped in the preparation of the final version of this paper. Fengwei Chen was supported in part by the National Natural Science Foundation of China under Grant 62073246.

Appendix

In a DHR model, each TVP is modeled by one member of the GRW family. Here, the stochastic evolution of each parameter sub-vector $\mathbf{a}_i(k)$, where,

$$\mathbf{a}_i(k) = \begin{bmatrix} \alpha_i(k) \\ \nabla\alpha_i(k) \end{bmatrix}$$

is assumed to follow the following state space model:

$$\mathbf{a}_i(k) = \mathbf{A}_i\mathbf{a}_i(k-1) + \mathbf{D}_i\boldsymbol{\eta}_i(k-1) \qquad i = 1, 2, \ldots, n, \tag{21}$$

where

$$\mathbf{A}_i = \begin{bmatrix} \alpha & \beta \\ 0 & \gamma \end{bmatrix}, \qquad \mathbf{D}_i = \begin{bmatrix} \delta & 0 \\ 0 & \varepsilon \end{bmatrix}$$

and $\boldsymbol{\eta}_i(k) = [\eta_{1i}(k)\ \eta_{2i}(k)]^T$ is a 2×1, zero mean, white noise vector that allows for stochastic variability in the parameters and is assumed to be characterized by a (normally diagonal) covariance matrix $\mathbf{Q}_{\eta_i}$.

This GRW model includes a number of special cases (see Young, 2011), but the most important of these in the present context are the IRW: $\alpha = \beta = \gamma = \varepsilon = 1$; $\delta = 0$; the scalar *Random Walk* (RW): scalar but equivalent to (21) with $\beta = \gamma = \varepsilon = 0$; $\alpha = \delta = 1$: i.e. just the first equation in (21)); and the intermediate case of *Smoothed Random Walk* (SRW): $0 < \alpha < 1$; $\beta = \gamma = \varepsilon = 1$ and $\delta = 0$. The hyper-parameters in this case are the various, normally assumed constant, coefficients in this GRW model (α, β, γ, δ, ε, as well as the elements of $\mathbf{Q}_{\eta_i}$ or, more

[6] 126 illegal gatherings broken up by police: see https://en.wikipedia.org/wikiTimeline_of_the_COVID-19_pandemic_in_England.

conveniently, the *Noise-Variance Ratio* (NVR) matrix $\mathbf{Q}_{nvr} = \mathbf{Q}_{\eta_i}/\sigma^2$, where σ^2 is the variance of the white noise in the observation equation. These are also assumed to be unknown *a priori* and need to be specified by the user or optimized in relation to the data; e.g. using the dhropt routine in CAPTAIN. There will be a set of hyper-parameters required for each parameter and these are accommodated in a vector: in relation to the RW, IRW and SRW models, these are the NVR vector of $\mathbf{Q}_{\eta_i}$ values, denoted by **nvr**; and the vector of α_i values, denoted by $\boldsymbol{\alpha}$.

The estimated frequency values are chosen by reference to the spectral properties of the time series, as quantified by the AR(n) spectrum, with the order n identified by reference to the Akaike AIC (see Fig. 2 in the paper). This DHR model can be considered as a straightforward extension of the classical, constant parameter, harmonic regression (or Fourier series) model, in which the gain and phase of the harmonic components can vary as a result of estimated temporal changes in the parameters $a(i,k)$, $b(i,k)$, $\alpha(i,k)$ and $\beta(i,k)$.

Postscript

A paper like the present one that is trying to describe an on-going epidemic like COVID-19 is always out of date because, by the time the analysis is completed and the paper is written, the epidemic continues take its course. Without knowledge of what might happen in 2021, the author decided to conclude the paper on December 7th 2020 and no further analysis and development of the SDP model in Section 4 has been carried out since then. However, after this date a new variant of the COVID-19 virus emerged in the UK and led to a considerable surge in deaths within the second wave of the epidemic that has led to death numbers in excess of those encountered in the first wave (1820 per day on January 20th; i.e. 26.8 per million of the population, so off the scale of the plot in Fig. 12). This would have to be accommodated in the SDP model, for example by the insertion of inputs such as that applied on August 24th to promote the second wave shown in Fig. 12.

References

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Banks, R. G. (1994). *Growth and diffusion phenomena*. Berlin: Springer-Verlag.

Cori, A., Ferguson, N. M., Fraser, C., & Cauchemez, S. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. *American Journal of Epidemiology*, *178*(9), 1505–1512.

Flaxman, S., Mishra, S., & Gandy et al, A. (2020). Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*, *584*, 257–261.

Gurney, W. S. C., Blythe, S. P., & Nisbet, R. M. (1980). Nicholson's blowflies revisited. *Nature*, *287*, 17–21.

Heesterbeek, J. A. P., & Roberts, M. G. (2015). How mathematical epidemiology became a field of biology: a commentary on Anderson and May (1981). *Philosophical Transactions of the Royal Society, Series B (Biological Sciences)*, *370*(1666), Article 20140307.

Hoertel, N., Blachier, M., Blanco, C., Olfson, M., Massetti, M., Rico, M. S., et al. (2020). A stochastic agent-based model of the SARS-CoV-2 epidemic in France. *Nature Medicine*, *26*(9), 1417–1421.

Jazwinski, A. H. (1970). *Stochastic processes and filtering theory*. San Diego: Academic Press.

Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London, Series A (Mathematical and Physical Sciences)*, *115*, 700–721.

Lega, J., & Brown, H. E. (2016). Data-driven outbreak forecasting with a simple nonlinear growth model. *Epidemics*, *17*, 19–26.

Murphy, A. H. (1995). The coefficients of correlation and determination as measures of performance in forecast verification. *Weather and Forecasting*, *10*, 681–688.

Murray, C. J. L. (2020). Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months. medRxiv.

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models: discussion of principles. *Journal of Hydrology*, *10*, 282–290.

Norton, J. P. (1986). *An introduction to identification*. New York: Academic Press, (reprinted by Dover Publications, Inc., 2009).

Padilla, A., Garnier, H., Young, P. C., Chen, F., & Yuz, J. I. (2019). Identification of continuous-time models with slowly time-varying parameters. *Control Engineering Practice*, *93*, Article 104165.

Popper, K. (1959). *The logic of scientific discovery*. London: Hutchinson.

Priestley, M. B. (1980). State-dependent models: A general approach to non-linear time series analysis. *Journal of Time Series Analysis*, *1*, 47–71.

Sahoo, B. K., & Sapra, B. K. (2020). A data driven epidemic model to analyse the lockdown effect and predict the course of COVID-19 progress in India. *Chaos, Solitons & Fractals*, *139*, Article 110034.

Söderström, T. (2007). Errors-in-variables methods in system identification. *Automatica*, *43*, 939–958.

Venkatramanan, S., Lewis, B., Chen, J., Higdon, D., Vullikanti, A., & Marathe, M. (2018). Using data-driven agent-based models for forecasting emerging infectious diseases. *Epidemics*, *22*, 43–49, The RAPIDD Ebola Forecasting Challenge.

Young, P. C. (1984). *Recursive estimation and time-series analysis: An introduction*. Berlin: Springer-Verlag.

Young, P. C. (1993a). Data-based mechanistic models. In P. C. Young (Ed.), *Concise encyclopedia of environmental systems* (pp. 137–142). Oxford: Pergamon Press.

Young, P. C. (1993b). Time variable and state dependent modelling of nonstationary and nonlinear time series. In T. Subba Rao (Ed.), *Developments in time series analysis* (pp. 374–413). London: Chapman and Hall.

Young, P. C. (1998). Data-based mechanistic modeling of environmental, ecological, economic and engineering systems. *Environmental Modelling & Software*, *13*, 105–122.

Young, P. C. (2000). Stochastic, dynamic modelling and signal processing: time variable and state dependent parameter estimation. In W. J. Fitzgerald, A. Walden, R. Smith, & P. C. Young (Eds.), *Nonlinear and nonstationary signal processing* (pp. 74–114). Cambridge: Cambridge University Press.

Young, P. C. (2011). *Recursive estimation and time-series analysis: An introduction for the student and practitioner*. Berlin.: Springer-Verlag.

Young, P. C. (2013). Hypothetico-inductive data-based mechanistic modeling of hydrological systems. *Water Resources Research*, *49*(2), 915–935.

Young, P. C. (2015). Refined instrumental variable estimation: Maximum likelihood optimization of a unified Box-Jenkins model. *Automatica*, *52*, 35–46.

Young, P. C. (2018). Data-based mechanistic modelling and forecasting globally averaged surface temperature. *International Journal of Forecasting*, *34*, 315–334.

Young, P. C. (2020). *Additional results for the paper Monitoring and forecasting the COVID-19 epidemic in the UK*: *Technical report*, Lancaster University Environment Centre and the Data Science Institute (Available from Author: Email p.young@lancaster.ac.uk).

Young, P. C., McKenna, P., & Bruun, J. (2001). Identification of nonlinear stochastic systems by state dependent parameter estimation. *International Journal of Control*, *74*, 1837–1857.

Young, P. C., Pedregal, D. J., & Tych, W. (1999). Dynamic harmonic regression. *Journal of Forecasting*, *18*, 369–394.