Assessing the Uncertainty of Epidemiological Forecasts with Normalised Estimation Error Squared

Robert E. Moore, Conor Rosato and Simon Maskell

Department of Electrical Engineering and Electronics, University of Liverpool, Brownlow Hill, Liverpool, L69 3GJ, UK

November 2021

Abstract

Estimates from infectious disease models have constituted a significant part of the scientific evidence used to inform the response to the COVID-19 pandemic in the UK. These estimates can vary strikingly in their precision, with some being over-confident and others over-cautious. The uncertainty in epidemiological forecasts should be commensurate with the errors in their predictions. We propose Normalised Estimation Error Squared (NEES) as a metric for assessing the consistency between forecasts and future observations. We introduce a novel infectious disease model for COVID-19 and use it to demonstrate the usefulness of NEES for diagnosing over-confident and over-cautious predictions resulting from different values of a regularization parameter.

1 Introduction

Several epidemiological modelling groups use statistical models of infectious disease to generate forecasts that contribute to a body of scientific evidence that informs the response to the COVID-19 pandemic in the UK. The models developed by the University of Cambridge MRC Biostatistics Unit and Public Health England (PHE) [1], the University of Warwick [2], and the London School of Hygiene & Tropical Medicine (LSHTM) [3] provide three notable examples of statistical models used to produce such estimates.

It does not appear to be standard practice to assess the uncertainty of epidemiological forecasts by scrutinising them to ensure that the uncertainty is commensurate with the disparity between predictions and future observations. Maishman et al. [4] provide a set of anonymised estimates for the effective reproduction number $R_t$ that highlights the striking differences in estimate uncertainty that different epidemiological models can produce.

We make two contributions to the collective effort of modelling COVID-19 in this paper. First, we introduce a statistical model for multisource COVID-19 surveillance data. This model is novel as a result of its use of symptom report data from the NHS 111 telephone service, its compartments for terminally ill individuals, and its implementation, which uses a bespoke numerical integrator to solve the system of ODEs for the transmission model. Second, we propose the Normalised Estimation Error Squared (NEES), which is popular in the tracking and data fusion communities, as a metric for diagnosing inconsistencies between epidemiological forecasts of observable quantities and their future values.

We begin by describing the novel statistical model in Section 2 before defining the NEES metric in Section 3. In Section 4, we make our proposal to use NEES more compelling by demonstrating its use in diagnosing issues with forecasts generated by the statistical model. Finally, in Section 5, we bring the paper to an end by presenting our conclusions.
2 Statistical Model

The model we describe in this section builds on a previous version whose implementation we contributed to the CoDatMo (Covid Data Models) organisation on GitHub [5]. The purpose of CoDatMo is to provide a collection of COVID-19 models, all written in the statistical modelling language Stan [6]. In addition to models originally implemented in Stan, CoDatMo provides Stan instantiations of COVID-19 models, initially written in other programming languages. By hosting a set of well-documented models, all implemented in a common language, CoDatMo hopes to improve understanding of the set of existing models and make it easier for newcomers and established researchers within the domain of epidemiology to make extensions and potential improvements.

We note that CoDatMo has already had some success against its objectives, with a group, primarily based at Universidade Nove de Julho in Brazil, creating a related but simpler model [7]. We also note that the UK Health Security Agency uses a slightly more sophisticated version of the model presented in this paper to generate weekly estimates of the effective reproduction number and growth rate for regions of the UK [8].

The statistical model consists of two parts: a transmission model that captures a simplified mechanism for the spread of coronavirus through the population; and an observation model that encapsulates the assumptions about the connection between the states of the transmission model and the observed surveillance data used to calibrate the model.

2.1 Transmission Model

The simple SIR compartmental model developed by Kermack and McKendrick [9] provides the theoretical basis for the transmission model. As with the SIR model, we assume a single geographical region with a large population of identical individuals who come into contact with one another uniformly at random but do not come into contact with individuals from other areas. In contrast to some other models, for example, those developed by Birrel et al. [1] and Keeling et al. [2], we treat the population as identical in terms of age and sex and only discriminate between individuals based on their disease states. We also assume that the population is closed, meaning that no births or deaths occur and no migration in or out of the population occurs.

The SIR model divides the population into three disease states: individuals who are (S) susceptible to infection; individuals who have been infected with the disease and are (I) infectious to other people; and individuals who have (R) “recovered” either by recuperating from the disease or dying. We augment these compartments in the transmission model by adding disease states for individuals who have been (E) exposed to the virus but are not yet infectious, individuals who are (T) terminally ill as a result of infection, and individuals who have (D) died of the disease. The exposed and dead compartments are standard extensions to the original SIR model. In contrast, we believe the terminally ill compartments to be a novel aspect of the model. In addition to adding these compartments, we also redefine the (R) recovered population to include the living only.

Inspired by the work of the University of Cambridge MRC Biostatistics Unit and Public Health England (PHE) [1], we partition each of the intermediate disease states (E, I, T) into two sub-states. Partitioning the sub-states in this way makes the model more realistic by implicitly constraining the times spent in each of these disease states to have Erlang rather than Exponential distributions.

We assume that there is at least one individual in each susceptible, exposed, and infectious compartment on the 17th of February 2020, which is the beginning of time in the model. At this stage, we assume that no population members have recovered, become terminally ill, or died. Two parameters, $\alpha_1$ and $\alpha_2$, determine the allocation of the rest of the population to the first five compartments of the transmission model. $\alpha_1$ is the proportion of the remaining population initially in the susceptible compartment, and $\alpha_2$ is the proportion of the infected population that is not yet infectious at time zero. For simplicity, we divide the exposed and infectious populations equally between the respective sub-states. We can express the exact relationship between the two parameters, $\alpha_1$ and $\alpha_2$, and the initial state of the transmission model as a set of equations:
\[ S(0) = (N - 5) \alpha_1 + 1, \]
\[ E_1(0) = \frac{1}{2} (N - 5) (1 - \alpha_1) \alpha_2 + 1, \]
\[ E_2(0) = \frac{1}{2} (N - 5) (1 - \alpha_1) \alpha_2 + 1, \]
\[ I_1(0) = \frac{1}{2} (N - 5) (1 - \alpha_1) (1 - \alpha_2) + 1, \]
\[ I_2(0) = \frac{1}{2} (N - 5) (1 - \alpha_1) (1 - \alpha_2) + 1, \]
\[ R(0) = 0, \]
\[ T_1(0) = 0, \]
\[ T_2(0) = 0, \]
\[ D(0) = 0. \]

Figure 1 provides a graphical illustration of the transmission model that captures the assumptions relating to the flow of individuals between disease states.

We assume that the population randomly mixes as time elapses, with infectious and susceptible individuals coming into contact with one another, potentially transmitting the virus. Susceptible people who have become exposed through these contacts are not initially infectious. The virus replicates in their bodies for a time, known as the latent period, before they become infectious and have the potential to transmit the virus onto members of the remaining susceptible population. We assume that after being infectious for some time, individuals either recover and become indefinitely immune to reinfection or become terminally ill for a while before, unfortunately, dying.

The number of individuals in each disease state varies with time according to a system of ordinary differential equations:
\[
\frac{dS(t)}{dt} = -\beta(t) \frac{I_1(t) + I_2(t)}{N} S(t),
\]
\[
\frac{dE_1(t)}{dt} = \beta(t) \frac{I_1(t) + I_2(t)}{N} S(t) - \frac{2}{d_L} E_1(t),
\]
\[
\frac{dE_2(t)}{dt} = \frac{2}{d_L} \left[ E_1(t) - E_2(t) \right],
\]
\[
\frac{dI_1(t)}{dt} = \frac{2}{d_I} E_2(t) - \frac{2}{d_I} I_1(t),
\]
\[
\frac{dI_2(t)}{dt} = \frac{2}{d_I} \left[ I_1(t) - I_2(t) \right],
\]
\[
\frac{dR(t)}{dt} = \frac{2}{d_I} I_2(t) \left[ 1 - \omega \right],
\]
\[
\frac{dT_1(t)}{dt} = \frac{2}{d_T} I_2(t) \omega - \frac{2}{d_T} T_1(t),
\]
\[
\frac{dT_2(t)}{dt} = \frac{2}{d_T} \left[ T_1(t) - T_2(t) \right],
\]
\[
\frac{dD(t)}{dt} = \frac{2}{d_T} T_2(t),
\]

where

- \( S(t) \) is the number of susceptible individuals who have not yet been infected and are at risk of infection,
- \( E_1(t) + E_2(t) \) is the number of exposed individuals who have been infected but are not yet infectious,
- \( I_1(t) + I_2(t) \) is the number of infectious individuals,
- \( R(t) \) is the number of recovered individuals,
- \( T_1(t) + T_2(t) \) is the number of terminally ill individuals who have been infected and will die,
- \( D(t) \) is the number of dead individuals,
- \( N = S(t) + E_1(t) + E_2(t) + I_1(t) + I_2(t) + R(t) + T_1(t) + T_2(t) + D(t) \) is the constant total number of individuals in the population,
- \( d_L \) is the mean time between infection and onset of infectiousness,
- \( d_I \) is the mean time for which individuals are infectious,
- \( d_T \) is the mean time for which individuals are terminally ill,
- \( \omega \), the infection fatality ratio (IFR), is the proportion of infected individuals who will die,
- \( \beta(t) \) is the mean rate of contacts between individuals per unit time that are sufficient to lead to transmission if one of the individuals is infectious and the other is susceptible. \( \beta(t) \) is a continuous piecewise linear function of time:

\[
\beta(t) = \sum_{j=1}^{J} \beta_j(t) \chi_{[t_{j-1}, t_j)}(t),
\]

where the mean rate of effective contacts during the \( j \)th time interval, \( \beta_j(t) \), is given by
\[
\beta_j(t) = \frac{\beta_{j+1} - \beta_j}{t - t_{j-1}} (t - t_{j-1}) + \beta_j,
\]
and
\[
\chi_{[t_{i-1}, t_i)}(t) = \begin{cases} 
1 & \text{if } t \in [t_{i-1}, t_i), \\
0 & \text{if } t \notin [t_{i-1}, t_i). 
\end{cases}
\]

The effective contact rate parameters \(\beta_1, \beta_2, \ldots, \beta_{J+1}\) in Equation 20 are the values \(\beta(t)\) takes on a set of predefined dates \(t_0, t_1, \ldots, t_J\). The first date is the 17th of February 2020. The second date is the 24th of March 2020, the first day after the Prime Minister announced the first national lockdown. The third date is seven days after the second. The fourth date is seven days after the third and so on.

2.2 Observation Model

Epidemiological modelling groups use many types of surveillance data to calibrate statistical models of infectious diseases. An observation model captures the assumptions about the relationship between the states of the transmission model and the surveillance data that we use for calibration.

We have designed the observation model to be extensible. Here, the model only has death data and symptom report data components to keep things as simple as possible. Nonetheless, the observation model can be extended to assimilate additional types of surveillance data, such as case data or hospital admissions data, by appending extra components that mirror the structure for ingesting symptom report data.

2.2.1 Death Data

On their official website for coronavirus data [10], the UK government publishes a daily time series of the number of deaths of individuals whose death certificate mentioned COVID-19 as one of the causes. We assume that the number of deaths on day \(t\), according to this definition, \(d_{\text{obs}}(t)\), has a negative binomial distribution parameterised by a mean \(d(t)\) and parameter \(\phi_{\text{deaths}}\) which affects overdispersion:

\[
d_{\text{obs}}(t) \sim \text{NegativeBinomial}(d(t), \phi_{\text{deaths}}),
\]

where we use the alternative parameterisation of the negative binomial distribution as defined by the Stan Development Team [11]:

\[
\text{NegativeBinomial}(\mu, \phi) = \left( \frac{n + \phi - 1}{n} \right) \left( \frac{\mu}{\mu + \phi} \right)^n \left( \frac{\phi}{\mu + \phi} \right)^\phi.
\]

In Equation 22, \(d(t)\) is the difference between the population of the D state of the transmission model between days \(t - 1\) and \(t\): \(d(t) = D(t) - D(t - 1)\).

2.2.2 Symptom Report Data

Every weekday up to the previous calendar day, NHS Digital publishes a daily time series of the number of assessments completed through the NHS 111 telephone service where callers reported potential coronavirus (COVID-19) symptoms [12]. Leclerc et al. found a strong correlation between the volume of these symptom reports and the number of COVID-19 deaths reported 16 days later [13].

NHS Digital provides the symptom report data with two important caveats, which we keep in mind while constructing the model. First, each time series element is the number of callers reporting potential COVID-19 symptoms, not testing positive for the virus. Second, there are no restrictions on the number of assessments that members of the public can complete. Consequently,
an individual can increment the count on a single day more than once by reporting potential COVID-19 symptoms through multiple assessments.

We use $c_{\text{obs}}(t)$ to denote the number of assessment calls to NHS 111 on day $t$ where callers reported potential COVID-19 symptoms. As with the number of deaths on day $t$, we assume that $c_{\text{obs}}(t)$ has a negative binomial distribution

$$c_{\text{obs}}(t) \sim \text{NegativeBinomial}(c(t), \phi_{\text{calls}}), \quad (24)$$

with parameters $c(t)$ and $\phi_{\text{calls}}$, where the negative binomial distribution is defined by the alternative parameterisation given in Equation (23). For the symptom reports, however, the mean parameter $c(t)$ has a more complicated relationship with the transmission model than being the difference between consecutive values of the deceased population.

There is an unknown latency between individuals being infected with the virus and calling NHS 111 to report their symptoms. We embody this unknown latency by making the mean number of symptom reports on day $t$, $c(t)$, depend on a weighted sum of current and lagged values of the number of new daily infections at time $t$,

$$\lambda(t) = -\beta \frac{I_1 + I_2}{N} S, \quad (25)$$

Even after sufficient time has passed since an individual was exposed to the virus to call NHS 111 and report potential coronavirus symptoms, there is no guarantee that they will do so. On aggregate, we assume that the ratio of symptom reports to potential symptom reporters, $\rho_{\text{calls}}(t)$, will vary with time due to two key phenomena. First, along with attitudes to the pandemic, the propensity of individuals to report their symptoms via the NHS 111 service changes with time. Second, the classification of reported symptoms indicative of potential coronavirus has changed at two points during the pandemic. Changes made to assessment on the 18th of May 2020 reduced the number of symptom reports, and changes made on the 26th of June 2020 increased the number of symptom reports. Full details of these changes are available on the NHS Digital website [12].

Combining our assumptions leads to the expression for the mean number of symptom reports

$$c(t) = \rho_{\text{calls}}(t) \sum_{l=0}^{L} w_l \lambda(t - l), \quad (26)$$

where $w_l$ is the lag weight placed on the number of new daily infections $l$ days before day $t$, and $L$ is the maximum lag, which we choose to be 13 days. These lag weights form a unit 14-simplex $w = (w_0, w_1, ..., w_L)$.

We select a discontinuous piecewise constant functional form for the ratio of symptom reports to potential symptom reporters

$$\rho_{\text{calls}}(t) = \sum_{k=1}^{K} \rho_{\text{calls,k}}(t) \chi_{[t_{k-1}, t_k)}(t), \quad (27)$$

where $\rho_{\text{calls,k}}(t)$ is the ratio of symptom reports to potential symptom reporters during the $k$th time interval, and the indicator function for the $k$th time interval, $\chi_{[t_{k-1}, t_k)}(t)$, is defined in Equation (21).

The first interval $[t_0, t_1)$ is from the 18th of March 2020 to the 21st of April 2020. The following interval $[t_1, t_2)$ runs for four weeks from the 21st of April 2020 and so on.

3 Normalised Estimation Error Squared

The tracking and data fusion community has studied performance measures for evaluating estimation algorithms, such as Li and Zhao’s absolute [14] and relative [15] error measures. Blasch, Rice, and Yang [16] and Chen et al. [17] describe a now-popular relative error measure called the
Normalised Estimation Error Squared (NEES), which researchers have used extensively as an easily understood approach to arguing the merits of, for example, different extensions of the Kalman filter to specific non-linear settings where the Extended Kalman Filter is routinely over-confident \[18\]. The NEES is

\[
\text{NEES} = \sum_t \left( \bar{X}_t - x_t^\dagger \right)^T \Sigma_t^{-1} \left( \bar{X}_t - x_t^\dagger \right),
\]

where \( \bar{X}_t \), \( x_t^\dagger \), and \( \Sigma_t \) are the estimate mean, the true value of the estimand \( x_t \), and the covariance matrix of the estimate at time \( t \). We note that although many researchers use NEES to refer to the metric in Equation 28 which is an average over multiple points in time, some researchers use it to refer to the summand. For example, Bar-Shalom et al. uses NEES in this way to refer to the summand of Equation 28 for a distinct point in time \[19\].

In general, \( \bar{X}_t \), \( x_t^\dagger \), and \( x_t \) are \( d \)-dimensional vectors and \( \Sigma_t \) is a \( d \times d \)-matrix. Nevertheless, in this paper, we use the NEES to assess forecasts for the daily number of deaths caused by COVID-19, which is a scalar quantity. It follows that we can simplify each term in Equation 28 to be the ratio of the squared estimation error, \( (\bar{X}_t - x_t^\dagger)^2 \), to the estimate variance, \( \sigma^2_{X_t} \). This simplification leads us to the specific form of the NEES,

\[
\text{NEES} = \frac{1}{T+7} \sum_{t=T+1}^{T+7} \frac{(\bar{X}_t - x_t^\dagger)^2}{\sigma^2_{X_t}},
\]

that we use in this paper to evaluate 7-day forecasts from day \( T+1 \) to day \( T+7 \). We compute the estimate variance,

\[
\sigma^2_{X_t} = \frac{1}{N} \sum_{i=1}^{N} (X_t^i - \bar{X}_t)^2,
\]

and the estimate mean,

\[
\bar{X}_t = \frac{1}{N} \sum_{i=1}^{N} X_t^i,
\]

in Equation 30 with the random sample \( X_1^1, X_1^2, ..., X_N^N \).

Taking a step back, we can see that each term in Equation 28 quantifies the level of agreement between the squared error of an estimate and its variance. On average, we hope the terms will be approximately one, with their numerators and denominators being similar. However, there are instances in which estimators are over-cautious and over-confident. Over-cautious estimators generate estimates with variances that, on average, dominate their squared errors, resulting in NEES values less than 1. Vice versa, over-confident estimators, which are arguably more damaging in terms of their impact on decision making, generate estimates with squared errors that, on average, dominate their variances, resulting in NEES values greater than 1.

4 Computational Experiments

4.1 Setup

We implement the statistical model described in Section 2 in Stan \[10\]. Stan is a probabilistic programming language that allows users to articulate statistical models and calibrate them with data using a Markov chain Monte Carlo (MCMC) method called the No-U-Turn Sampler (NUTS), proposed by Hoffman and Gelman \[20\]. Stan also provides diagnostic information, such as warnings about divergent transitions, to help users check it has sampled the posterior faithfully.
In addition to encoding the statistical model, the Stan implementation includes prior distributions for the model parameters. Table 1 gives details of the prior distributions we have chosen to use in the computational experiments, along with the reasoning behind their selection.

The software implementation has two idiosyncrasies worthy of discussion. First, the current implementation does not use any of the integrators provided by Stan to solve the system of ordinary differential equations (ODEs) in Equation 10. Instead, a bespoke implementation of the explicit trapezoidal method [21] solves the system of ODEs for the transmission model. Anecdotally, the trapezoidal integrator significantly reduces runtime while producing acceptable numerical errors. Second, the current implementation only uses Stan’s default initialisation strategy for $1/\phi_{\text{deaths}}$, $1/\phi_{\text{calls}}$, $\rho_{\text{calls},1}$, ..., $\rho_{\text{calls},K}$, and $\text{w}$ by drawing values uniformly between -2 and 2 on the unconstrained parameter space. Rather than doing this for $\alpha_1$, $\alpha_2$, $\beta_1$, ..., $\beta_J$, $d_L$, $d_I$, $d_T$, and $\omega$, the implementation draws uniformly from custom intervals to prevent initialisation failures caused by unrealistic parameter values.

We calibrate the model with data for seven NHS regions: the East of England, London, the Midlands, the North East and Yorkshire, the North West, the South East, and the South West. For each region, we use NUTS to draw six independent Markov chains for each of eight smoothness hyperparameter, $\sigma_\beta$, values. Each chain draws 512 samples and discards the first 256 drawn during warmup.

The data we use for calibration comprises death data, as described in Section 2.2.1, that starts on the 24th of March 2020 and finishes on 30th of December 2020, and symptom report data, as described in Section 2.2.2, that begins on 18th of March 2020 and ends on the 6th of January 2021. We chose to use this period because it includes the first two coronavirus waves in the UK.

Each calibration job generates a posterior predictive distribution for the daily number of deaths. The last seven elements of these distributions are forecasts for the 31st of December to the 6th of January 2021, for which we calculate NEES and root-mean-square error (RMSE) values.

We ran the computational experiments for this paper on two dedicated nodes of the University of Liverpool’s primary high-performance computing (HPC) resource, the Barkla cluster. Each of Barkla’s nodes has Intel(R) Xeon(R) Gold 6138 CPU @ 2.00GHz processors, a total of 40 cores and 384 GB of memory. We were unable to exploit the full potential of the hardware available because out-of-the-box Stan can only use one CPU per calibration job, which takes on average 2 hours and 20 minutes to complete.

We have made the code and data for the computational experiments publically available on GitHub.

### 4.2 Results

The NEES and RMSE values from the computational experiments are presented in the top and bottom of Table 2, respectively. The columns contain results for different values of the smoothness hyperparameter $\sigma_\beta$ defined in Table 1. Smaller values of $\sigma_\beta$ make the random-walk prior on the effective contact rate $\beta(t)$ tighter, causing it to vary more slowly and, if low enough, to underfit the data. Conversely, larger values of the smoothness hyperparameter loosen the random-walk prior and allow overfitting of the data if $\sigma_\beta$ is high enough.

Generally, the NEES values in Table 2 become closer to one as $\sigma_\beta$ increases. For the lowest values of the smoothness hyperparameter, the NEES values are either significantly below or above one, indicating large discrepancies between the forecasts and future observations. The trend is clearer for the RMSE, with the values decreasing quickly as $\sigma_\beta$ increases.

We have highlighted two NEES values for each NHS region in Table 2. Based on the definition of NEES in Section 3, we expect the values highlighted in blue to be associated with forecasts that are consistent with the future observations that they predict. However, we have different expectations for the values highlighted in orange and red, which should correspond to over-cautious and over-confident forecasts, respectively. Figure 2 displays plots of the forecasts associated with

[1]https://github.com/codatmo/UniversityOfLiverpool_PaperSubmission
Table 1: Prior distributions for the parameters of the statistical model with the rationale for their selection.

| Parameter(s) | Prior Distribution | Comment |
|--------------|--------------------|---------|
| $\alpha_1$   | Beta (5.0, 0.5)    | This reflects our belief that most of the population is initially susceptible. |
| $\alpha_2$   | Beta (1.1, 1.1)    | This reflects our uninformed beliefs about the initial division of the infected population into those who are infectious and those who are not. |
| $\beta_1$    | Normal (0.0, 0.5)  | This is a generic, weakly informative prior inspired by the work of Gelman [22]. |
| $\beta_2, ..., \beta_J$ | Normal ($\beta_{i-1}, \sigma_\beta$) | This is random-walk prior with smoothness hyperparameter $\sigma_\beta$ that enforces correlation between contiguous $\beta_i$. |
| $d_L$        | Normal(4.0, 3.0)   | This is based on an estimate of the incubation period provided by Pellis et al. [23]. |
| $d_I$        | Normal(5.0, 4.0)   | This is based on an estimate of the delay from onset of symptoms to hospitalisation provided by Pellis et al. [23]. |
| $d_T$        | Normal(13.0, 4.0)  | This is based on an estimate of the delay from hospitalisation to death provided by Linton et al. [24]. |
| $\omega$     | Beta (3.8, 571.6)  | This is based on an estimate provided by Verity et al. [25]. |
| $\varphi_{\text{deaths}}$ | Exponential (5.0) | This is a containment prior for the overdispersion parameter, which is discussed by Simpson [26]. |
| $\varphi_{\text{calls}}$ | Exponential (5.0) | This is a containment prior for the overdispersion parameter, which is discussed by Simpson [26]. |
| $\rho_{\text{calls,1} - \rho_{\text{calls,}K}}$ | HalfNormal (0.0, 0.5) | This is a generic, weakly informative prior inspired by the work of Gelman [22]. |
| $w$          | Dirichlet (0.1)    | This is a distribution chosen to encourage the unit 14-simplex of lag weights to gravitate towards its corners. |
Table 2: Top: NEES values for different combinations of NHS region and $\sigma_\beta$ value. The orange, red and blue cells highlight selected NEES values that are significantly below, significantly above, and approximately equal to one. Bottom: RMSE values for different combinations of NHS region and $\sigma_\beta$ value.

| NHS Region              | $\sigma_\beta$ | 0.0005 | 0.001  | 0.0025 | 0.005  | 0.01   | 0.025  | 0.05   |
|-------------------------|----------------|--------|--------|--------|--------|--------|--------|--------|
| East of England         | NEES           | 5.670  | 2.707  | 2.370  | 1.758  | 0.986  | 0.665  | 0.562  |
| London                  |                | 5.843  | 6.379  | 5.418  | 3.140  | 1.731  | 0.650  | 0.334  |
| Midlands                |                | 0.122  | 0.185  | 0.292  | 1.039  | 1.023  | 0.512  | 0.435  |
| North East and Yorkshire|                | 0.067  | 0.222  | 0.181  | 0.544  | 0.675  | 0.659  | 0.715  |
| North West              |                | 0.309  | 0.131  | 0.594  | 1.981  | 1.371  | 1.063  | 0.934  |
| South East              |                | 3.081  | 0.871  | 0.690  | 0.659  | 0.405  | 0.433  | 0.409  |
| South West              |                | 1.654  | 0.088  | 0.179  | 0.820  | 1.639  | 1.849  | 1.561  |

| NHS Region              | RMSE           | 79.091 | 42.212 | 35.380 | 29.295 | 20.753 | 16.458 | 14.637 |
|-------------------------|----------------|--------|--------|--------|--------|--------|--------|--------|
| East of England         |                | 108.138| 69.868 | 61.482 | 47.094 | 31.116 | 17.190 | 12.837 |
| London                  |                | 37.212 | 31.866 | 17.460 | 22.984 | 19.405 | 13.972 | 13.300 |
| Midlands                |                | 21.301 | 31.451 | 11.452 | 11.899 | 11.774 | 10.742 | 10.653 |
| North East and Yorkshire|                | 42.787 | 17.964 | 17.177 | 21.620 | 15.457 | 12.201 | 11.528 |
| North West              |                | 94.362 | 33.681 | 27.819 | 23.301 | 16.650 | 15.918 | 15.254 |
| South West              |                | 24.899 | 5.763  | 6.090  | 9.694  | 11.358 | 11.264 | 10.451 |

The plots behind the values highlighted in orange and red are in the left column, and the plots behind the values highlighted in blue are in the right column.

We can see that the plots in the left column of Figure 2 associated with NEES values significantly greater than one correspond to genuinely over-confident forecasts. The NEES and RMSE values for the East of England, London and the South East behave similarly and correctly identify the significant bias and over-confidence in the forecasts. However, the two metrics behave differently in the over-confidence forecasts for the North West and South West. The NEES values are high and quantify the over-confidence apparent where RMSE alone would simply identify that these are relatively low-bias forecasts.

We can also see that the plots in the left column of Figure 2 behind NEES values much less than one correspond to genuinely over-cautious forecasts. These forecasts for the Midlands and the North East and Yorkshire underscore NEES’s ability to distinguish between over-cautious and over-confident forecasts, which the RMSE alone cannot quantify.

As anticipated, the plots in the right column of Figure 2 behind NEES values closer to one correspond to forecasts consistent with the observations they predict. The results are less variable than those in the left column, and, in the main, neither metric identifies any issues.

While there will be cases where the RMSE alone is able to identify deficiencies of a model, these results show the potential for NEES to provide an additional useful and simple to calculate diagnostic.

5 Conclusion

We have shown that NEES is a useful metric for assessing epidemiological forecasts. The comparison with the RMSE presented in Section 4 highlights how the NEES can complement the widespread use of RMSE. We, therefore, advocate both the NEES and RMSE as simple metrics for evaluating forecasts to establish if they are over-cautious, over-confident, or capture the uncertainty associated with future observations.

One of the significant limitations of NEES is that we can only use it to assess forecasts of
Figure 2: Plots of the forecasts behind the NEES values highlighted in Table 2. The red line shows the forecast mean, and the orange ribbon covers values one standard deviation from the mean. The black and green points are the true number of deaths before and after the forecasting date, which is shown as a vertical dashed blue line.
observable variables. Epidemiological modellers cannot apply it to important latent quantities, such as the effective reproduction number $R_t$ and the growth rate $r$, which they often forecast. Accordingly, the epidemiological community needs a method for assessing forecasts of quantities for which the truth is unknown. Simulation-Based Calibration (SBC) \cite{27} is a candidate method for this task, worth investigating further.

There are three worthwhile directions in which we can extend the statistical model presented in Section 2. The first direction for extending the model involves adding additional components to the observation model described in Section 2.2 to allow calibration with a greater quantity and diversity of surveillance data. The second fruitful direction for extending the model involves modifying it to accommodate surveillance data from other countries. The final direction for extending the model entails making the disease-specific, geography-independent parameters $d_L$, $d_I$, $d_T$, and $\omega$ global to facilitate information sharing between regions, as is done by Birrell et al. \cite{1}.

**Data Accessibility**

The code and data for the computational experiments are publically available on GitHub (https://github.com/codatmo/UniversityOfLiverpool_PaperSubmission).

**Authors’ Contributions**

REM designed and implemented the statistical model, participated in the analyses and drafted the manuscript; CR carried out the computational experiments, participated in the analyses and helped draft the manuscript; SM conceived and supervised the work, participated in the analyses and critically revised the manuscript. All authors gave final approval for publication and agree to be held accountable for the work performed therein.

**Competing Interests**

We declare that we have no competing interests.

**Funding**

This work was supported by an ICASE Research Studentship jointly funded by EPSRC and AWE [EP/R512011/1]; a Research Studentship jointly funded by EPSRC and the ESRC Centre for Doctoral Training on Quantification and Management of Risk and Uncertainty in Complex Systems Environments [EP/L015927/1]; and EPSRC through the Big Hypotheses grant [EP/R018537/1].

**Acknowledgements**

We thank Public Health England (PHE), the Joint Biosecurity Centre (JBC), and the UK Health Security Agency (UKHSA) for their support. We thank Breck Baldwin and Jose Storopoli for their help in advancing CoDatMo. We also thank Veronica Bowman, Alexander Phillips and John Harris for their suggestions and Matthew Carter for configuring the HPC environment.

**References**

\cite{1} Birrell P, Blake J, van Leeuwen E, Gent N, De Angelis D. 2021 Real-time nowcasting and forecasting of COVID-19 dynamics in England: the first wave. Phil. Trans. R. Soc. B 376: 20200279. https://doi.org/10.1098/rstb.2020.0279.
[2] Keeling MJ, Dyson L, Guyver-Fletcher G, Holmes A, Semple MG, ISARIC4C Investigators, Tildesley M, Hill EM. 2021 Fitting to the UK COVID-19 outbreak, short-term forecasts and estimating the reproductive number. medRxiv, 2020.08.04.20163782.

[3] Abbott S, Hellewell J, Thompson RN et al. Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts [version 2; peer review: 1 approved with reservations]. Wellcome Open Res 2020, 5:112 (https://doi.org/10.12688/wellcomeopenres.16006.2)

[4] Maishman T, Schaap S, Silk DS, Nevitt SJ, Woods DC, Bowman VE. 2021 Statistical methods used to combine the effective reproduction number, $R(t)$, and other related measures of COVID-19 in the UK. arXiv preprint, [arXiv:2103.01742]. URL: https://arxiv.org/abs/2103.01742. Preprint.

[5] CoDatMo. 2021 Welcome to the CoDatMo site. See https://codatmo.github.io (accessed 9 September 2021).

[6] Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P, Riddell A. 2017 Stan: A Probabilistic Programming Language. Journal of Statistical Software, 76(1), 1 - 32. http://dx.doi.org/10.18637/jss.v076.i01

[7] Storopoli J, Santos ALMF, Pellini ACG, Baldwin B. 2021 Simulation-Driven COVID-19 Epidemiological Modeling with Social Media. arXiv, 2106.11686.

[8] UK Government. 2021 Reproduction number (R) and growth rate: methodology. See https://www.gov.uk/government/publications/reproduction-number-r-and-growth-rate-methodology/reproduction-number-r-and-growth-rate-methodology (accessed 21 September 2021).

[9] Kermack WO and McKendrick AG. 1927 A contribution to the mathematical theory of epidemics. Phil. Trans. R. Soc. A. https://doi.org/10.1098/rspa.1927.0118.

[10] UK Government. 2021 Coronavirus (COVID-19) in the UK. See https://coronavirus.data.gov.uk (accessed 13 September 2021).

[11] Stan Development Team. 2021 Negative binomial distribution (alternative parameterization). See https://mc-stan.org/docs/2.x/functions-reference/nbalt.html (accessed 13 September 2021).

[12] NHS Digital. 2021 Potential Coronavirus (COVID-19) symptoms reported through NHS Pathways and 111 online. See https://digital.nhs.uk/data-and-information/publications/statistical/ni-potential-covid-19-symptoms-reported-through-nhs-pathways-and-111-online/latest (accessed 14 September 2021).

[13] Leclerc QJ, Nightingale ES, Abbott S, CMMID COVID-19 Working Group, Jombart T. Analysis of temporal trends in potential COVID-19 cases reported through NHS Pathways England. Sci Rep 11, 7106 (2021). https://doi.org/10.1038/s41598-021-86266-3.

[14] Li XR and Zhao Z. 2001 Measures of performance for evaluation of estimators and filters. Proc. SPIE 4473, Signal and Data Processing of Small Targets 2001. https://doi.org/10.1117/12.492751.

[15] Li XR and Zhao Z. 2005 Relative error measures for evaluation of estimation algorithms. 7th International Conference on Information Fusion, 2005, pp. 8 pp.-. doi: 10.1109/ICIF.2005.1591857.

[16] Blasch EP, Rice A, Yang C. 2006 Nonlinear tracking evaluation using absolute and relative metrics. Proc. SPIE 6236, Signal and Data Processing of Small Targets 2006, 62360L. https://doi.org/10.1117/12.666463.
[17] Chen Z, Heckman C, Julier S, Ahmed N. 2018 Weak in the NEES?: Auto-Tuning Kalman Filters with Bayesian Optimization. 21st International Conference on Information Fusion (FUSION), 2018, pp. 1072-1079. doi: 10.23919/ICIF.2018.8454982.

[18] Longbin M, Xiaoquan S, Yiyu Z, Kang SZ and Bar-Shalom Y, "Unbiased converted measurements for tracking," in IEEE Transactions on Aerospace and Electronic Systems, vol. 34, no. 3, pp. 1023-1027, July 1998, doi: 10.1109/7.705921.

[19] Bar-Shalom Y, Li XR, Kirubarajan T. 2004 Estimation with applications to tracking and navigation: theory algorithms and software. John Wiley & Sons.

[20] Hoffman MD, Gelman A. 2014 The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. J. Mach. Learn. Res. 15: 1593–1623. https://dl.acm.org/doi/10.5555/2627435.2638586.

[21] Ascher UM and Petzold LR. 1998 Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations. Philadelphia: Society for Industrial and Applied Mathematics, ISBN 978-0-89871-412-8.

[22] Gelman A. 2020 Prior Choice Recommendations. See https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations (accessed 16 September 2021).

[23] Pellis L et al. 2021 Challenges in control of COVID-19: short doubling time and long delay to effect of interventions. Phil. Trans. R. Soc. B 376: 20200264. https://doi.org/10.1098/rstb.2020.0264.

[24] Linton NM, Kobayashi T, Yang Y, Hayashi K, Akhmetzhanov AR, Jung S-m, Yuan B, Kinoshita R, Nishiura H. 2020 Incubation Period and Other Epidemiological Characteristics of 2019 Novel Coronavirus Infections with Right Truncation: A Statistical Analysis of Publicly Available Case Data. Journal of Clinical Medicine. 2020; 9(2):538. https://doi.org/10.3390/jcm9020538.

[25] Verity R, Okell LC, Dorigatti I, Winskill P, Whittaker C, Imai N, Cuomo-Dannenburg G, Thompson H, Walker PGT, Fu H, Dighe A, Griffin JT, Baguelin M, Bhatia S, Boonyasiri A, Cori A, Cucunubá Z, FitzJohn R, Gaythorpe K, Green W, Hamlet A, Hinsley W, Laydon D, Nedjati-Gilani G, Riley S, van Elsland S, Volz E, Wang H, Wang Y, Xi X, Donnelly CA, Ghani AC, Ferguson NM. 2020 Estimates of the severity of coronavirus disease 2019: a model-based analysis. The Lancet Infectious Diseases, Volume 20, Issue 6, 2020, Pages 669-677, ISSN 1473-3099. https://doi.org/10.1016/S1473-3099(20)30243-7.

[26] Simpson D. 2018 Justify my love. See https://statmodeling.stat.columbia.edu/2018/04/03/justify-my-love/ (accessed 16 September 2021).

[27] Talts S, Betancourt M, Simpson D, Vehtari A, Gelman A. 2020 Validating Bayesian Inference Algorithms with Simulation-Based Calibration. arXiv preprint, arXiv:1804.06788 URL: https://arxiv.org/abs/1804.06788. Preprint.