



This is a repository copy of *Bayesian emulation and history matching of JUNE*.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/220115/>

Version: Published Version

Article:

Vernon, I. orcid.org/0000-0002-9161-9946, Owen, J. orcid.org/0009-0000-9618-9589, Aylett-Bullock, J. orcid.org/0000-0001-7551-3423 et al. (11 more authors) (2022) Bayesian emulation and history matching of JUNE. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 380 (2233). 20220039. ISSN 1364-503X

<https://doi.org/10.1098/rsta.2022.0039>

Reuse

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:

<https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



eprints@whiterose.ac.uk
<https://eprints.whiterose.ac.uk/>

Research



Cite this article: Vernon I *et al.* 2022 Bayesian emulation and history matching of JUNE.

Phil. Trans. R. Soc. A **380**: 20220039.

<https://doi.org/10.1098/rsta.2022.0039>

Received: 18 February 2022

Accepted: 7 June 2022

One contribution of 18 to a theme issue 'Technical challenges of modelling real-life epidemics and examples of overcoming these'.

Subject Areas:

statistics

Keywords:

disease models, Bayes linear, emulation, calibration, history matching

Author for correspondence:

I. Vernon

e-mail: i.r.vernon@durham.ac.uk

Bayesian emulation and history matching of JUNE

I. Vernon^{1,2}, J. Owen^{1,2}, J. Aylett-Bullock^{1,3},

C. Cuesta-Lazaro^{1,4}, J. Frawley^{1,5},

A. Quera-Bofarull^{1,4}, A. Sedgewick^{1,6}, D. Shi^{1,4},

H. Truong^{1,3}, M. Turner^{1,5}, J. Walker^{1,3}, T. Caulfield⁷,

K. Fong^{8,9} and F. Krauss^{1,3}

¹Institute for Data Science, ²Department of Mathematical Sciences,

³Institute for Particle Physics Phenomenology, ⁴Institute for

Computational Cosmology, ⁵Advanced Research Computing,

⁶Centre for Extragalactic Astronomy, ⁷Department of Computer Science, Durham University, Durham DH13LE, UK

⁸Department of Science, Technology, Engineering and Public Policy, University College London, London WC1E6BT, UK

⁹Department of Anaesthesia, University College London Hospital, London NW12BU, UK

IV, 0000-0002-9161-9946; JA-B, 0000-0001-7551-3423;

AQ-B, 0000-0001-5055-9863; HT, 0000-0002-0105-1278;

FK, 0000-0001-5043-3099

We analyze JUNE: a detailed model of COVID-19 transmission with high spatial and demographic resolution, developed as part of the RAMP initiative. JUNE requires substantial computational resources to evaluate, making model calibration and general uncertainty analysis extremely challenging. We describe and employ the uncertainty quantification approaches of Bayes linear emulation and history matching to mimic JUNE and to perform a global parameter search, hence identifying regions of parameter space that produce acceptable matches to observed data, and demonstrating the capability of such methods.

This article is part of the theme issue 'Technical challenges of modelling real-life epidemics and examples of overcoming these'.

© 2022 The Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License <http://creativecommons.org/licenses/by/4.0/>, which permits unrestricted use, provided the original author and source are credited.

1. Introduction: Bayesian emulation and uncertainty quantification

The COVID-19 pandemic disrupted healthcare systems and caused substantial fatalities around the globe. Various models have been developed to aid decision makers in the assessment of policy options, from simple analytic models up to complex agent-based models (ABMs). JUNE, introduced in [1], is of the latter type, and its high level of demographic and spatial resolution demands substantial computational resources to evaluate. A critical component in the uncertainty analysis, and subsequent use for decision support of a complex epidemiological model such as JUNE, is the process of model calibration: the matching of the model to observed data from the real system.

This process can be extremely challenging, and in many cases, its intractability precludes the full exploitation of sophisticated models, which may otherwise contain nuanced insights into the system of interest. The problem of calibrating a complex and computationally demanding model is not unique to epidemiology and occurs in a wide range of scientific disciplines including cosmology, climate, systems biology, geology and energy systems [2–5]. To solve this problem, an area of Bayesian statistics arose, sometimes referred to as the uncertainty analyses of complex computer models, or to use its more recent (and slightly more general name): the area of uncertainty quantification (UQ) [5–7]. UQ provides a statistical methodology combining a large number of efficient techniques with a set of overarching principles that address how to analyze complex models rigorously, transparently and robustly, for use in scientific investigations, for making real-world predictions and for subsequent decision support. A core goal of this work is to demonstrate the capability of such methods for use with complex epidemiological models. A full analysis of the behaviour of models with a large number of input parameters and possibly several outputs, and their subsequent calibration, encounters the following three major issues:

- (i) For complex models, the evaluation time of the model is often so long that an exhaustive exploration of the model's behaviour over its full input parameter space is infeasible.
- (ii) When comparing models to observed real-world data, an adequate statistical description of the link between model and reality, covering all major uncertainties and allowing for the rigorous use of an *imperfect* model, is required.
- (iii) When calibrating, the appropriate scientific goal should be to identify *all* locations in input parameter space that lead to acceptable fits between model and observed data, and not just to find the location of a single good match.

We summarize in the next section three UQ methods: (a) Bayes linear emulation, (b) linking models to reality and (c) Bayesian history matching, which address the aforementioned three problems. We then apply these UQ methods to the JUNE model in §3.

2. Bayesian emulation and history matching

(a) Bayes linear emulation

Complex models typically have runtimes that can vary from seconds to days or even weeks, greatly inhibiting full model exploration, calibration, forecasting etc. Many of the techniques in UQ therefore revolve around the construction of Bayesian *emulators*: statistical constructs that mimic the scientific model in question, providing predictions of the model outputs with associated uncertainty, at as yet unevaluated input parameter settings [8]. The emulators provide insight into the model's core structure and, unlike the models they mimic, are extremely fast to evaluate, typically being several orders of magnitude faster. Hence, they facilitate previously infeasible model exploration and global parameter searches. As an emulator makes predictions that have an associated (input dependent) uncertainty statement, they naturally fit within an overarching Bayesian uncertainty analysis, in which the impact of using an emulator instead of

the model, can be understood and quantified. Emulators can be built for deterministic models, stochastic models, multilevel models (composed of models of increasing fidelity) and networks of models, providing a flexible and powerful set of tools to deal with a large class of scientific scenarios. Here, we outline the construction of Bayes Linear emulators, a robust form of emulator, based on a partial specification, which has been successfully employed in several settings [2,5].

We represent a general scientific model as the function $f(x)$. Here, $x = (x_1, \dots, x_d)$ is a vector composed of all the input parameters. For example, x_1 may represent an infectivity parameter, x_2 a social distancing parameter, etc. $f(x) = (f_1(x), \dots, f_q(x))$ is the vector of all model outputs of interest, so, for example, $f_1(x)$ may represent the number of people hospitalized in England on a particular day, $f_2(x)$ may represent the number of deaths on that day, all as a function of the inputs x . We denote the general component of $f(x)$ as $f_i(x)$, where the index i will cycle through the full list of outputs of interest, for example, in the application to JUNE in §3, i cycles through the set $i \in \{\text{type, region, time}\}$. We anticipate that, due to limited computational resources, we will only be able to evaluate the model at a finite (and possibly small) number of input parameter locations $x^{(1)}, x^{(2)}, \dots, x^{(n)}$ giving rise to model outputs $D_i = (f_i(x^{(1)}), f_i(x^{(2)}), \dots, f_i(x^{(n)}))^T$, where $i = 1, \dots, q$, and ' T ' denotes the transpose. Therefore, at a new unevaluated input location, x , even say for a deterministic (i.e. repeatable) model, we will still be uncertain about the output value of $f(x)$, as we will be for the majority of the input space. We take a subjective Bayesian view and treat the unknown $f(x)$ as a random quantity and construct an emulator that represents our beliefs about possible reasonable forms that this function $f(x)$ could take. A popular emulator form for each output $f_i(x)$ is as follows [2]:

$$f_i(x) = \sum_j b_{ij} g_{ij}(x_{A_i}) + u_i(x_{A_i}) + w_i(x), \quad (2.1)$$

where we have selected a subset of the inputs, x , known as the active variables, x_{A_i} , that are most influential for output $f_i(x)$. The first term on the right-hand side of equation (2.1) is a regression term, where g_{ij} are appropriately selected known deterministic functions of x_{A_i} , a common choice being low-order polynomials, and b_{ij} are unknown scalar regression coefficients. The second term, $u_i(x_{A_i})$, is a weakly second-order stationary process over x_{A_i} , for which we only need to specify its second-order structure, choosing $E[u_i(x_{A_i})] = 0$ and utilizing an appropriate covariance function: a classic example suitable for smooth functions is the squared exponential:

$$\text{Cov}(u_i(x_{A_i}), u_i(x'_{A_i})) = \sigma_{u_i}^2 \exp \left\{ -\frac{\|x_{A_i} - x'_{A_i}\|^2}{\theta_i^2} \right\}, \quad (2.2)$$

where $\sigma_{u_i}^2$ and θ_i are the variance and correlation length of $u_i(x_{A_i})$, respectively, which may be specified *a priori* [2], or fitted using, e.g. MLE or MAP [4]. This simple covariance function may be enough, especially if the emulators regression term captures much of the model's behaviour; however, if not, various extensions are available, e.g. individual correlation lengths θ_i^j for each input $x_{A_i}^j$ [9]. The third term, $w_i(x)$, is a white noise process uncorrelated with b_{ij} , $u_i(x_{A_i})$, and itself such that

$$\text{Cov}(w_i(x), w_i(x')) = \begin{cases} \sigma_{w_i}^2 & \text{if } x = x' \\ 0 & \text{otherwise} \end{cases}, \quad (2.3)$$

with expectation zero, and $\text{Var}(w_i(x)) = \sigma_{w_i}^2$. $w_i(x)$ represents the effects of the remaining inactive inputs not included in the list of x_{A_i} and formally facilitates a type of dimensional reduction [2].

The emulator form, as represented by equation (2.1), has various desirable features and exploits our beliefs about the general anticipated behaviour of physically realistic models. The regression term, $\sum_j b_{ij} g_{ij}(x_{A_i})$, attempts to mimic the large-scale global behaviour of the function $f_i(x)$: often substantial in physical models. The second term, $u_i(x_{A_i})$, the weakly stationary process, mimics the local behaviour of $f_i(x)$, again exploiting concepts of smoothness of either $f_i(x)$ or attributes of $f_i(x)$ (if, say, $f_i(x)$ is stochastic). Such terms are highly versatile and can fit a large class of models; however, they require a sufficient density of runs to be suitably informed (regulated

by the correlation length parameters θ_i). In the literature, there is sometimes an over-reliance on similar Gaussian process style terms and a neglect of the regression terms, which may be unwise, as GPs of this form are typically capable of capturing the broad global behaviour, or the more complex local behaviour, but rarely both. We deliberately use the regression terms for the global structure and utilize the $u_i(x_{A_i})$ to capture the local behaviour.

We can select the list of active inputs, x_{A_i} , using various statistical techniques. For example, these could consist of classical linear model fitting criteria such as AIC or BIC, which have the benefit of speed and reasonable accuracy when applied to appropriate (nonlinear) sets of regression functions g_{ij} [2], or approaches such as automatic relevance determination [9], which can give increased accuracy provided the assumed form of the covariance function is suitable. In addition, we would also seek to incorporate expert knowledge of the model into the active input selection process, either by directly incorporating 'known' active inputs or by using a more nuanced Bayesian approach, of particular importance for expensive models. A list of p active inputs for a particular physical output, $f_i(x)$, means that we have notably reduced the input dimensionality from d to p , which can result in large efficiency gains in subsequent calculations. The small remaining effect of the inactive inputs is not ignored, but is captured by the third term $w_i(x)$ in equation (2.1), whose variance $\sigma_{w_i}^2$ represents the added uncertainty induced by the dimensional reduction.

(i) What to emulate

A major issue when emulating complex models is the choice of the set of attributes/outputs of the model to emulate. For example, often an objective function describing the mismatch between model and data has been emulated (e.g. a simple chi-squared metric, or a more complex likelihood function). However, despite being deceptively simple having just a single output, the objective function typically has a complex form, as it depends on the union of all active inputs and possesses numerous local maxima/minima [2], rendering this an inefficient strategy. Instead, we prefer to emulate the physical outputs of the model directly, as these tend to have (a) a smaller list of active inputs per output allowing a nuanced and sometimes substantial dimensional reduction tailored to each individual output, and (b) a simpler functional dependence on the input parameters that is often well represented by the regression terms in the emulator. Further choices are required when emulating stochastic models, where we can choose to emulate summaries of outputs of interest such as the mean, the variance or quantiles if required, possibly conditioning on key events such as epidemic take-off, and extend for example to covariance structures between groups of outputs if needed. In these cases, the role of $w_i(x)$ is extended to also incorporate the uncertainties induced by using estimates from finite samples [10] or to employ full variance emulation as in ref. [11].

(ii) Designing batches of model evaluations

We begin by specifying the region of input space of interest, typically a d -dimensional hypercube, and denote this $\mathcal{X}_0 \subset \mathbb{R}^d$. We then design a set of 'space filling' runs over \mathcal{X}_0 , constructed to be well spread out, without any large holes. For example, we may use a maximin Latin hypercube design, an approximately orthogonal design, also desirable for emulator construction [12,13].

(iii) Updating the emulator

We then update our prior emulator structure given by equation (2.1) with the information from the set of model runs using our favourite statistical tools. Specifically, we would prefer a fully probabilistic Bayesian approach if we required full probability distributions on all emulated outputs, $f_i(x)$ [14], and were we prepared to specify full joint probabilistic priors.

However, in most cases, our preferred choice is to use Bayes Linear methods, a more tractable version of Bayesian statistics, which requires a far simpler prior specification and analysis [15,16]. It deals purely with expectations, variances and covariances of all uncertain quantities of interest,

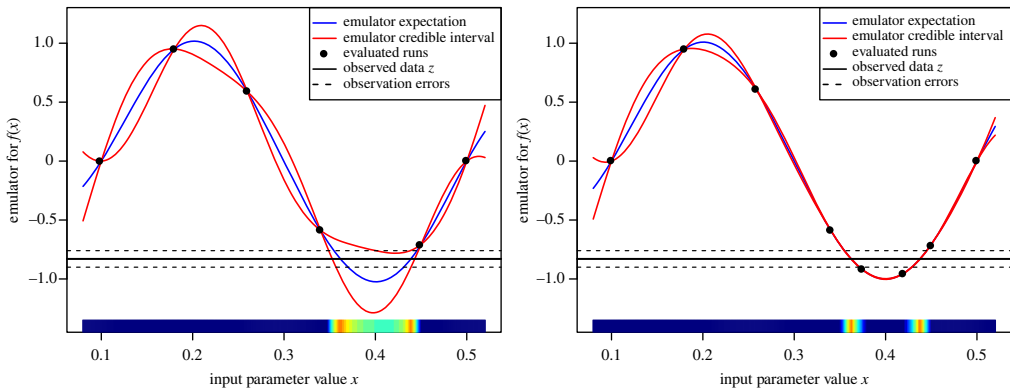


Figure 1. An emulator of a one-dimensional toy model, where $f(x) = \sin(2\pi(x - 0.1)/0.4)$, for the first wave/iteration, using just six runs (left panel), and for the second wave, using two additional runs (right panel). The emulator's expectation $E_D[f(x)]$ and credible intervals $E_D[f(x)] \pm 3\sqrt{\text{Var}_D(f_i(x))}$ are given by the blue and red lines, respectively, with the observed data z that we wish to match to as the black horizontal line (with errors). The implausibility $I(x)$ is represented by the coloured bar along the x -axis, with dark blue implying $I(x) > 3$, light blue $2.5 < I(x) < 3$ and yellow $I(x) < 1$. (Online version in colour.)

and uses the following Bayes linear update equations, derived from foundational arguments [16], to adjust our beliefs in the light of new data. When emulating the i th output $f_i(x)$ of a complex model, where we had performed an initial batch of n runs giving a vector of model output values $D_i = (f_i(x^{(1)}), f_i(x^{(2)}), \dots, f_i(x^{(n)}))^T$, we obtain the adjusted expectation, $E_{D_i}(f_i(x))$, and adjusted variance, $\text{Var}_{D_i}(f_i(x))$, for $f_i(x)$ at new input point x using:

$$E_{D_i}(f_i(x)) = E(f_i(x)) + \text{Cov}(f_i(x), D_i)\text{Var}(D_i)^{-1}(D_i - E(D_i)) \quad (2.4)$$

and

$$\text{Var}_{D_i}(f_i(x)) = \text{Var}(f_i(x)) - \text{Cov}(f_i(x), D_i)\text{Var}(D_i)^{-1}\text{Cov}(D_i, f_i(x)). \quad (2.5)$$

All quantities on the right-hand side of equations (2.4) and (2.5) can be calculated from equations (2.1) and (2.2) combined with prior specifications for $E(b_{ij})$, $\text{Var}(b_{ij})$, $\sigma_{u_i}^2$, $\sigma_{w_i}^2$ and θ_i . $E_{D_i}(f_i(x))$ and $\text{Var}_{D_i}(f_i(x))$ provide a prediction for $f_i(x)$ with associated uncertainty and are used directly in the implausibility measures used for the global parameter searches described in §2c. Note that multivariate versions of equations (2.4) and (2.5) are available. In addition, we may make certain pragmatic choices in the emulator construction process, for example, while we typically keep the regression coefficients b_{ij} uncertain, we may directly specify $\sigma_{u_i}^2$, $\sigma_{w_i}^2$ and θ_i a priori, or use suitable plugin estimates as described in ref. [2]. We can test the emulators using a series of diagnostics, for example checking their prediction accuracy over a new batch of runs [17]. An example of a one-dimensional emulator is given in figure 1, cf. ref. [8] for an introduction and refs. [2,14,18] for details. The above Bayes linear emulation framework is fully implemented in the 'hmer' R package [19].

(b) Assessing uncertainties: linking the model to the real world

Most epidemiology models are developed to help explain, understand and predict the behaviour of the corresponding real world system of interest, typically in terms of the progression through a population of an infectious disease. An essential part of determining whether such a model is adequate for this task is the comparison of the model with data formed from observations of the real system. However, this comparison involves several uncertainties that must be accounted for to provide a meaningful definition of an 'acceptable' match between a model run and the observed data. Hence, it is vital to define a clear statistical model describing the difference between the epidemiological model, $f(x)$, and the observed data denoted as the vector z . While

more complex statistical models are available [20], here, we describe a simple but powerful version that has been successfully used in a variety of scientific disciplines, for example climate, cosmology, oil reservoirs, epidemiology and systems biology [2–5,10,21].

The most familiar source of uncertainty is observational or experimental. We represent the features of interest of the real system as a vector of uncertain quantities, y , which will be measured imperfectly involving a vector of errors e , to give the vector of observations, z , as follows:

$$z = y + e. \quad (2.6)$$

We represent the errors as additive here, but could use a more complex form if necessary. Depending on the scientific context, we then make judgements about the relationship between y and e , e.g. a common specification [2] is to judge the errors e to be independent from y , with expectation, $E(e) = \mu_e$ and $\text{Var}(e) = \Sigma_e$, a $q \times q$ covariance matrix. Setting $\mu_e = 0$ corresponds to the judgement that the observations were unbiased, and setting $\Sigma_e = \text{diag}(\sigma_{e,1}^2, \dots, \sigma_{e,q}^2)$, that is a diagonal matrix, corresponds to uncorrelated observation errors, etc.

A critical feature that we must incorporate is the difference between the epidemiological model, $f(x)$, of the system and the real system, y , itself. We represent this difference between model and reality using a *structural model discrepancy* term. First, we note that even were we to evaluate the model, $f(x)$, at its best possible choice of input, x^* , the output, $f(x^*)$, would still not be in agreement with the real epidemiological system value y , due to the many simplifications and approximations inherent to the model; therefore,

$$y = f(x^*) + \epsilon, \quad (2.7)$$

where ϵ is the structural model discrepancy: a vector of uncertain quantities that directly represents the difference between the model and the real system. Note that we are still treating y , f , x^* and ϵ as vectors of random quantities. Now we have to make judgements about their relationships: a simple and popular specification [2,5] would be to judge that ϵ is independent of $f(x^*)$, x^* and e , with $E(\epsilon) = 0$. In the case of a single output, we would then specify $\text{Var}(\epsilon) = \sigma_\epsilon^2$. However, for the full case of q outputs, we may specify $\text{Var}(\epsilon) = \Sigma_\epsilon$, a $q \times q$ covariance matrix. Σ_ϵ may have intricate structure possessing non-zero covariances between components of ϵ , to capture the heavily correlated deficiencies of the model outputs. Various structures for Σ_ϵ of increasing complexity are available [2,5,14], along with methods for their specification [2,22]. Note that typically the form of Σ_ϵ is very different from Σ_e .

While the inclusion of the structural model discrepancy is unfamiliar to most modellers, it is now of standard practice in the UQ literature for analyzing complex but imperfect models [14,18,23,24]. It facilitates a richer analysis whereby we can incorporate our necessarily uncertain knowledge of the model's deficiencies to improve our modelling of reality y . Its inclusion can prevent over-fitting when calibrating and also reduces both bias and overconfidence when predicting. It is also vital when combining the results of several models.

(c) Bayesian history matching

Due to their fast evaluation speed, emulators can be used in a variety of UQ calculations that would be otherwise infeasible. One of the most important is the problem of performing global parameter searches. Here, we outline a powerful iterative emulator-based global search method known as history matching (HM), which has been successfully employed in a variety of scientific disciplines [2,3,5,10]. HM is designed to answer the questions:

- (i) Are there any input parameter settings that lead to acceptable matches between the model output and observed data?
- (ii) If so, what is the full set \mathcal{X} that contains all such input parameter settings?

Note the emphasis on finding *all* such acceptable matches: optimizing to find a single good fit is not adequate for assessing the impact of parametric uncertainty, nor for making predictions.

HM proceeds iteratively by ruling out regions of input parameter space that can be discarded from further investigation based on *implausibility measures* [5]. For an unexplored input parameter, x , we can ask how far would the emulator's expected value for the individual function output, $f_i(x)$, be from the corresponding observed value, z_i , before we would deem it highly unlikely for $f_i(x)$ to give an acceptable match were we to actually evaluate the function at x . The implausibility measure, $I_i(x)$, captures this concept, and for an individual, output is given by the distance $E_{D_i}(f_i(x)) - z_i$ between emulator expectation and observed data, standardized by all relevant uncertainties,

$$I_i^2(x) = \frac{(E_{D_i}(f_i(x)) - z_i)^2}{\text{Var}_{D_i}(f_i(x)) + \text{Var}(\epsilon_i) + \text{Var}(e_i)}. \quad (2.8)$$

Here, $\text{Var}_{D_i}(f_i(x))$ is the emulator variance, $\text{Var}(\epsilon_i)$ is the variance of the model discrepancy and $\text{Var}(e_i)$ is the variance of the observational error, a direct consequence of equations (2.6) and (2.7). See also figure 1 (the x -axis) for a depiction of $I(x)$.

A large value of $I_i(x)$ for a particular x implies that we would be unlikely to obtain an acceptable match between $f_i(x)$ and z_i were we to run the model at x . Hence, we can discard the input, x , from the parameter search if $I_i(x) > c$, for some cutoff, c , which is often chosen by appealing to Pukelsheim's 3-sigma rule [25], a very general and powerful result, which states that for *any* continuous, unimodal distribution, 95% of its probability must lie within $\pm 3\sigma$, regardless of asymmetry or skew, suggesting that a choice of $c = 3$ may be reasonable [2]. This is the simplest univariate form, but we can combine implausibility measures from several outputs using say $I_M(x) = \max_{i \in Q} I_i(x)$ for some set Q , or employ more complex multivariate forms [2].

Before performing the k th HM iteration, we define the current set of non-implausible input points as \mathcal{X}_k and the set of outputs that we considered for emulation in the previous wave as Q_{k-1} . We proceed as follows [4]:

- (1) Design and evaluate a well chosen set of runs over the current non-implausible space \mathcal{X}_k , e.g. using a maximin Latin hypercube with rejection [2]. Combine these with any non-implausible runs surviving from previous waves.
- (2) Check if there are new, informative outputs that can now be emulated accurately (that were difficult to emulate in previous waves) and add them to the previous set Q_{k-1} , to define Q_k .
- (3) Use the runs to construct new, more accurate emulators defined only over the region \mathcal{X}_k for each output in Q_k .
- (4) The implausibility measures $I_i(x)$, $i \in Q_k$, are then recalculated over \mathcal{X}_k , using the new emulators.
- (5) Cutoffs are imposed on the implausibility measures $I_i(x) < c$ and this defines a new, smaller non-implausible volume \mathcal{X}_{k+1} , which should satisfy $\mathcal{X} \subset \mathcal{X}_{k+1} \subset \mathcal{X}_k$.
- (6) Unless (a) the emulator variances for all outputs of interest are now small in comparison to the other sources of uncertainty due to the model discrepancy and observation errors, or (b) the entire input space has been deemed implausible, and return to step 1.
- (7) If 6 (a) is true, generate as large a number as possible of acceptable runs from the final non-implausible volume \mathcal{X} , sampled depending on scientific goal.

The history matching approach is powerful for several reasons: (a) while reducing the volume of the non-implausible region, we expect the function $f(x)$ to become smoother, and hence to be more accurately approximated by the regression part of the emulator, $b_{ij}g_{ij}(x_{A_i})$. (b) At each new HM iteration, we will have a higher density of points and hence the second term, $u_i(x_{A_i})$, in the emulator should be more effective, as it depends on proximity to the nearest runs. (c) In later iterations, the previously strongly dominant active inputs from early waves will have their effects curtailed, and hence, it will be easier to select additional active inputs, unnoticed before. (d) There may be several outputs that may be difficult to emulate in early iterations (perhaps because of their erratic behaviour in uninteresting parts of the input space) but simple to emulate in later waves once we have restricted the input space to a much smaller and more epidemiologically

realistic region. See ref. [4] for further discussions comparing HM with Bayesian MCMC and ABC, ref. [26] for a direct comparison with ABC and the R package ‘hmer’ [19] for full implementation of the HM algorithm. We now apply these methods to the JUNE model.

3. Application of emulation and history matching to JUNE

(a) The JUNE model

JUNE [1] is an ABM that describes the spread of an infectious disease through large synthetic populations. Originally designed to simulate the circulation of COVID-19 through the English population, JUNE has also been adapted to capture the populations of Cox’s Bazaar [27], a refugee camp in Bangladesh, and of Rhineland-Palatinate [28], one of Germany’s federal states. JUNE’s description of the epidemic spread rests on four areas:

- the construction of a realistic synthetic population that reflects, as accurately as possible, the population demographic and their geographic distribution;
- the simulation of the population sociology, i.e. how the individuals behave: how they spend their time, whom they get into contact with and in which social environment;
- the parameterization of the infection, how it is transmitted from infected to susceptible individuals and impact it has on the health of infected individuals;
- the mitigation of spread and impact of the infection through pharmaceutical and non-pharmaceutical interventions (NPIs) such as social distancing and vaccinations, respectively.

They are discussed in more detail below.

(i) Population

JUNE builds its synthetic population based on real or parameterized census data—in the case relevant for this contribution, JUNE constructs the about 55 million residents of England based on the 2011 census data accessible through the NOMIS database provided by the ONS. The data are organized hierarchically, with Output Areas (OAs) the smallest relevant unit, comprising on average about 250 residents with relatively similar socio-economic characteristics. The OAs have a specified geographic location, and their data contain information about age, sex and ethnicity of the area’s residents [29–31] and the composition of the households they live in [32], in about 20 categories.¹ JUNE uses national averages to correlate age, sex and ethnicity of individuals, which are presented as uncorrelated distributions in the data. In a similar way, information such as the national distributions of age differences of partners [33], and of parents and their children [34], are used to assign the individuals to their households.

As additional static properties of the population, JUNE assigns school-aged children to the nearest age-appropriate school; information about school locations and the age ranges for their students is taken from ref. [35]. Within the schools, the students are grouped into class units of 20–30 individuals and have teachers assigned to them. In a similar way, universities are filled with students—the young adults—and they are grouped into year groups of about 200 students.

The OAs are part of Middle Super Output Areas (MSOAs) with about 12 500 residents and 50 OAs constituting one MSOA. The census data provide information about the sectors of companies within MSOAs and about the distribution of the working population over these sectors, using the Standard Industrial Classification (SIC) scheme [36]. The parameterization of company sizes with national sector-dependent averages allows JUNE to construct an origin-destination matrix for the employees at the level of MSOAs [37]. Information concerning the commuting habits of

¹The English and Welsh census distinguishes between children, young (dependent) and adults, such as university students, independent adults and old adults, and classifies households according to the respective numbers in them.

individuals contained in the census data [38] underpins the construction of simplified virtual public transport networks within JUNE.

(ii) Interactions

Having defined the static properties of the synthetic population—where people live, work and study—their daily lives outside work and education are filled with various activities. These activities include shopping, visiting friends and relatives in their homes, frequenting pubs and restaurants, going to the gym or cinema, to name a few. In the absence of any of these leisure activities, people are supposed to stay at home. Surveys performed, e.g. by the Office for National Statistics [39], define the average proportion of time spent with various activities, in dependence on age and sex. These averages are translated into a probabilistic treatment thereby creating a highly flexible and varied daily schedule for JUNE's virtual individuals.

These schedules are supplemented with contact matrices from POLYMOD [40] and the BBC Pandemic Project [41], which indicate the average number of daily contacts—communication or physical—of individuals of age i with individuals of age j in different social settings \mathcal{L} , for example home (H), school (S) and work (W). As the contact numbers are presented as population averages, suitable for their deployment in compartment models, they need to be renormalized for the socially more granular IBMs,² resulting in the renormalized overall contact matrices $\chi_{ij}^{(\mathcal{L})}$ and the corresponding fraction of physical contacts, $\phi_{ij}^{(\mathcal{L})}$, where $\mathcal{L} \in \{S, H, W\}$. While this introduces some uncertainty into the modelling of social interactions, the interplay of the synthetic population model with the contact matrices provides a welcome closure test for the self-consistency of the overall model.

For the purpose of fitting to data and the quantification of uncertainties in the model, we assume that the construction of the synthetic population and its interactions is well understood and robustly and well parameterized as it is driven by data of relatively high quality.

(iii) Infection

The description of the infection consists of two separate parts. First, the transmission from an infected person i to a susceptible person s needs to be simulated. In JUNE, as in many other models, this is described as a probabilistic process. The infection probability for a susceptible person s with susceptibility ψ_s during a time interval from t to $t + \Delta t$, spent with a group of individuals g in social context \mathcal{L} is given as follows:

$$\begin{aligned} P_{si}(t, t + \Delta t) &= 1 - \exp \left[-\psi_s \sum_{i \in g} \int_t^{t+\Delta t} dt' \beta_{si}^{(\mathcal{L}, g)} \mathcal{I}_i(t') \right] \\ &\approx 1 - \exp \left[-\psi_s \Delta t \sum_{i \in g} \beta_{si}^{(\mathcal{L}, g)} \mathcal{I}_i(t) \right]. \end{aligned} \quad (3.1)$$

In the aforementioned equation, $\mathcal{I}_i(t)$ denotes the time-dependent infectiousness of the infected individual i in group g . In JUNE, it follows a profile given by

$$\mathcal{I}_i(\tau) = \mathcal{I}_{\text{imax}} \frac{\tau^{a-1} e^{-\tau}}{\Gamma(a)}, \quad (3.2)$$

with $\tau = t - t_0 - t_{\text{inc}}$, and t_0 is the time of infection of the individual, t_{inc} is the incubation period and Γ is the gamma function. t_{inc} is sampled from a normal distribution. The maximal or peak value of infectiousness for individual i is sampled from a log-normal distribution with median $\exp(\mu) = 1$ and shape parameter $\sigma = 0.25$, which allows for a long but small tail of highly

²As an example consider the number of contact children have with adults in schools. Clearly the number of contacts of average adults with children in schools is much less than the number of contacts adult teachers have with the children, necessitating a renormalization of the number of contacts by the proportion of teachers in the overall adult population.

infectious individuals, which can be connected to super-spreader events. The $\beta_{si}^{(\mathcal{L},g)}$ in equation (3.1) is the contact intensity between s and i ,

$$\beta_{si}^{(\mathcal{L},g)} = \beta_{\mathcal{L}} \frac{\chi_{si}^{(\mathcal{L})}}{N_g} [1 + \phi_{si}^{(\mathcal{L})}(\alpha - 1)], \quad (3.3)$$

where β are the social location-dependent baseline intensities, N_g is the number of individuals in the group setting, normalizing the contact number χ_{si} and α parameterizes the relative increase in infection probability for the proportion of physical contacts ϕ_{si} . These parameters, the social-environment dependent β and the universal α , cannot be derived from first principles and must be obtained from fits to available data; they constitute a significant portion of the parameter space in the model and, correspondingly, a significant source of uncertainty.

Once an individual is infected, it takes some time—the incubation period—before they can infect others and some additional time before the onset of symptoms. A large range of input data has been used to derive various symptom trajectories for infected individuals, which in the case of high-income western countries in the global North depends mainly on their age and sex.³ In the original formulation of the JUNE model, significant efforts have gone into the quantification of probabilities for different health outcomes in the population, with some emphasis to also capture the health impact of COVID-19 on the highly vulnerable care home residents; we refer the reader to ref. [1] for more details. Here, it should suffice to state that in JUNE asymptomatic and symptomatic trajectories with varying severity have been identified, the latter ranging from mild, flu-like symptoms over admission to regular or intensive-care wards to death in hospital or at residence. Although there are some uncertainties related to this treatment, we usually do not consider them and treat the health outcomes as fixed by data. We seed initial infections based on the number of fatalities 2–3 weeks afterwards, by using the infection-fatality rates obtained from data and encoded in JUNE. The parameter $\alpha_{\text{seedstrength}}$ is an additional factor that modifies the resulting number of initial infections.

(iv) Interventions

Since the beginning of the COVID-19 epidemic, the UK government—like many other governments around the world—has employed a wide range of mitigation strategies. At the beginning of the pandemic, these interventions were mainly non-pharmaceutical, and these NPIs ranged from relatively simple strategies at the level of individuals, such as mask wearing and other social distancing measures, to more involved and global strategies such as partial or complete lockdowns, involving school closures and the furloughing of parts of the work force. In JUNE, these measures can easily be modelled: social distancing measures and mask wearing can be described by modifying the β 's in the corresponding social settings by a factor, $M_{\mathcal{L}}$, capturing the reduced, but non-zero, transmission probability, while the closure of schools or furloughing of the work force is easily described, based on data [42–45], by keeping the impacted population at home instead of sending them to schools or work. For a more detailed description of the translation of NPIs to the JUNE simulation, we refer the reader to ref. [1].

(b) Inputs, outputs and initial emulation

Our primary goal is to test if the JUNE model can produce acceptable matches to observed data at the national and regional level, from the first wave of the COVID-19 pandemic and the subsequent summer period. We wish to identify the region of parameter space, \mathcal{X} , leading to such acceptable fits, if it exists. We identify a large set of input parameters, x , of interest to search over, primarily composed of interaction intensity parameters at the group level, seeding and quarantine compliance parameters, and social distancing parameters (see [1] for details), and

³In other population settings, for example in refugee camps, age and sex do not constitute good proxies for the overall health of an individual, and comorbidities need to be explicitly factored in, giving rise to significant additional assumptions and uncertainties.

Table 1. The input parameters explored in the global parameter search, their type and their ranges that define the search region \mathcal{X}_0 .

input parameter (x_i)	type	range
β_{pub}	location-dependent contact intensity	[0.02,0.6]
β_{grocery}	—	[0.02,0.6]
β_{cinema}	—	[0.02,0.6]
$\beta_{\text{university}}$	—	[0.02,0.6]
$\beta_{\text{city transport}}$	—	[0.08,0.77]
$\beta_{\text{inter city transport}}$	—	[0.08,1.2]
β_{hospital}	—	[0.08,1.2]
$\beta_{\text{care home}}$	—	[0.08,1.2]
β_{company}	—	[0.08,1.2]
β_{school}	—	[0.08,1.2]
$\beta_{\text{household}}$	—	[0.08,1.2]
$\beta_{\text{carevisits}}$	—	[0.1,10]
$\beta_{\text{householdvisits}}$	—	[0.1,10]
α_{physical}	physical contact factor	[1.8,3]
$\alpha_{\text{seed strength}}$	modifies initial/seeding infections	[0.1,2]
$M_{\text{quarantine household compliance}}$	quarantine compliance	[0.034,0.26]
$M_{\text{socialdistancing } \beta_{\text{factor}}}$	social distance (1 week prior to lockdown)	[0.65,0.95]
$M_{\text{sd3 random factor all}}$	enhanced social distance (full lockdown)	[0.1,0.5]
$M_{\text{sd4 random factor all}}$	social distance (post lockdown, non-leisure)	[0.25,1]
$M_{\text{sd4 random factor leisure}}$	social distance (post lockdown, leisure)	[0.25,1]

specify ranges for each, given in table 1, which define the initial search space, \mathcal{X}_0 . These ranges were chosen to be conservative, informed in part by earlier exploratory runs while also respecting the role each parameter plays in the model, including the time period over which they operate (see [1] for details). A typical full England run of JUNE would take approximately 10 hours to complete on 64 cores (Intel Xeon Skylake) and 128 GB of memory. This substantial computational expense combined with a relatively high-dimensional input parameter space makes a global parameter search extremely challenging and necessitates the use of emulation. While there are several types of data available for the early pandemic, many of these had questions regarding reliability. For example, case data were substantially affected by the limited and evolving availability of COVID-19 tests, while hospital admission data were, especially during the peak of the first wave, collected with understandably varying levels of diligence across trusts. While it would in principle be possible to incorporate such data sets using a detailed statistical model for the measurement errors, e , in equation (2.6) that incorporated under-counting, we instead focus on hospital and total death data [46–49], which although still uncertain due to the precise definition of death with COVID-19, suffers from far fewer issues.

We define the primary JUNE outputs of interest to be the hospital deaths and total deaths from 19 March to the end of August 2020, for England and its seven regions: East of England, London, Midlands, North East and Yorkshire, North West, South East and South West. We choose a subset of dates $\{t_1, t_2, \dots\}$, shown as the vertical dashed lines in figure 4, to emulate. The observed data and JUNE output is noisy, so we smooth them both using a standard kernel smoother (Gaussian kernel, bandwidth 7 days) as we wish to compare the underlying trends, and define the

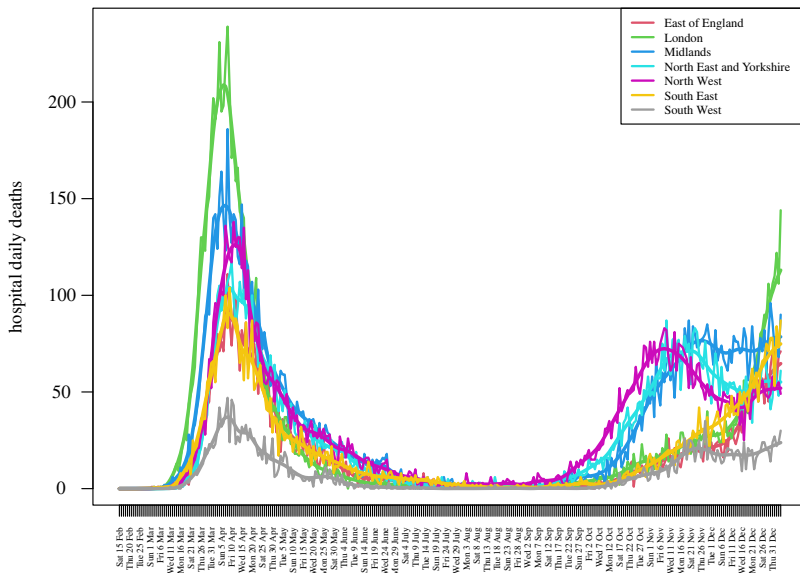


Figure 2. Daily deaths in hospital wards and ICU in 2020, by region. The smoothed version used in the HM is also shown. (Online version in colour.)

smoothed versions to be the target observed data points z_i (shown in figure 2), and the primary JUNE outputs, $f_i(x)$. Therefore, i index cycles through elements of the set $i \in \{type, region, time\}$, where $type$ labels hospital or total deaths, $region$ labels each of the seven regions of England or England itself and $time$ labels the time points $\{t_1, t_2, \dots\}$ of interest, given as the dashed lines in figure 4. We specify conservative observation error and model discrepancy variances $\sigma_{\epsilon_i}^2$ and $\sigma_{\epsilon_i}^2$ for each output as described in ref. [1], by decomposing each into multiplicative and additive components to represent possible systematic biases, in addition to a scaled \sqrt{n} component for the observation error only, to model the noisy count process.

We design and evaluate a first iteration/wave of 150 runs over the input space, \mathcal{X}_0 , using a maximin Latin hypercube design. The outputs of these runs are shown as the purple lines in figure 4. We construct emulators for each output, $f_i(x)$, as detailed in §2a (using full quadratic regression terms selected using BIC, and MAP estimates for θ_i [2]). The emulators provide insight into the behaviour of the JUNE model. For example, we can examine the coefficients b_{ij} of the linear terms $g_{ij}(x_{A_i}) = x_{A_i}^j$ for the inputs featuring in equation (2.1), to gain insight into the effect each input has on each individual output. Estimates of these are shown in figure 3 for the total deaths in England outputs, where i index therefore cycles through just the various time points: $i \in \{\text{'Total Deaths', 'England', } t'_1, t'_2, \dots\}$ with each time point, labelled on the x -axis, giving rise to a single vertical strip in the plot corresponding to a single emulator (note that a finer temporal resolution $\{t'_1, t'_2, \dots\}$ is used here for added detail, while far fewer time points are used in the HM). Conversely, j labels the active input $x_{A_i}^j$ in question as given on the y -axis. Here, red/blue represents positive/negative dependencies b_{ij} , respectively, standardized as proportions of the largest coefficient of that output. We see strong anticipated contributions from $\beta_{company}$, β_{school} and $\beta_{household}$ in the first wave of the pandemic from March to May, and more modest effects from $M_{social\ distancing\ beta\ factor}$ and $\beta_{grocery}$ throughout the summer period. The sensitivities of β_{school} and $\beta_{household}$ change to negative (blue) by May, as in many of these uncalibrated runs, herd immunity has been reached, and hence increasing β_{school} will decrease deaths (as they will be brought forward in time). Note that the parameters $\beta_{household\ visits}$ and $\beta_{care\ visits}$ are not included in figure 3 as they were only implemented prior to the second iteration of runs, but were included

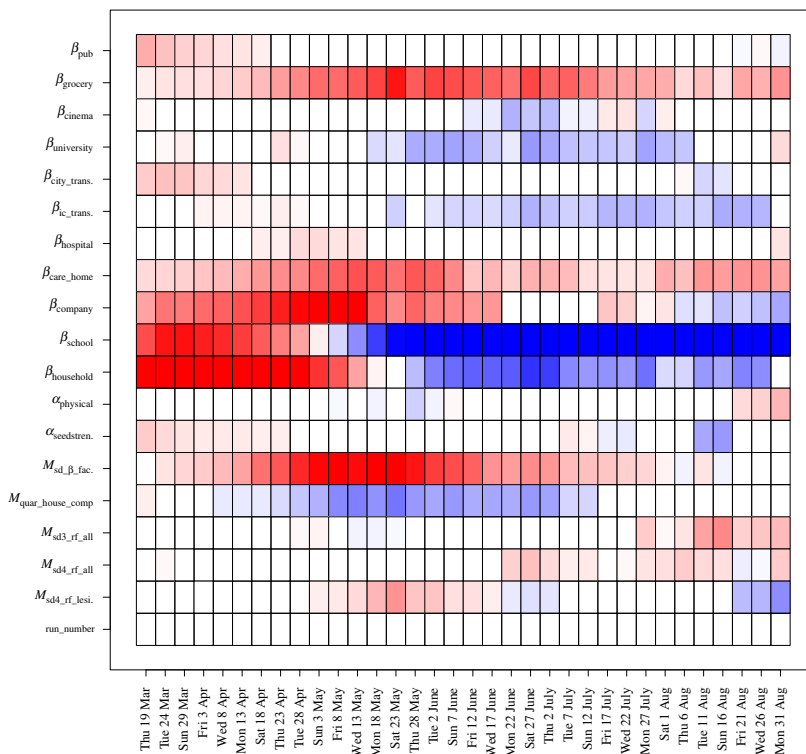


Figure 3. Estimates for the coefficients b_{ij} of the linear terms $g_{ij}(x_{A_i}) = x_{A_i}^{b_{ij}}$ that are found to feature in the emulators for total deaths in England for the first iteration/wave of runs, where i labels the time point (x -axis) and j labels the inputs (y -axis). Red/blue represents positive/negative dependencies of $f_i(x)$ on that input, respectively, standardized as proportions of the largest coefficient for that time point. A finer temporal resolution is used here for added clarity. Note that this plot shows the time-dependent sensitivity of the model to the inputs, but that the actual inputs x do not vary over time. (Online version in colour.)

in the subsequent full history match by suitable inflation of the iteration 1 emulator uncertainties [20]. More insight can be gained from full emulator sensitivity analysis [50].

(c) Iterative history matching

We now employ the history matching framework from §2c, iteratively removing parameter space based on current implausibility measures, and performing batches/iterations of further runs. Initially, in iterations 1 and 2, only the hospital and total deaths for England up to the end of May 2020 were included in the HM, to rule out the more exotic regions of parameter space, while for iterations 3–5, all the seven regions of England were also included and the time period extended to the end of August 2020. Figure 4 shows the outputs from iterations 1, 3 and 5 for total deaths in England as the purple, yellow and red lines, respectively. As the iterations proceed, the emulators become more accurate, we learn more about the global parameter space, and hence, the runs approach the observed data, yielding reasonable matches across the first COVID-19 wave. By iteration 5, the majority of emulators attained the accuracy required for the stopping criteria in the HM algorithm. The region \mathcal{X}_5 of 20-dimensional parameter space deemed non-implausible at iteration 5 is shown in figure 5 as a collection of two-dimensional *optical depth plots*, which simply show the depth in the remaining 18 dimensions of the non-implausible region (see [4]). The optical depth $\rho(x')$ is defined for each point x' in the two-dimensional space shown in each

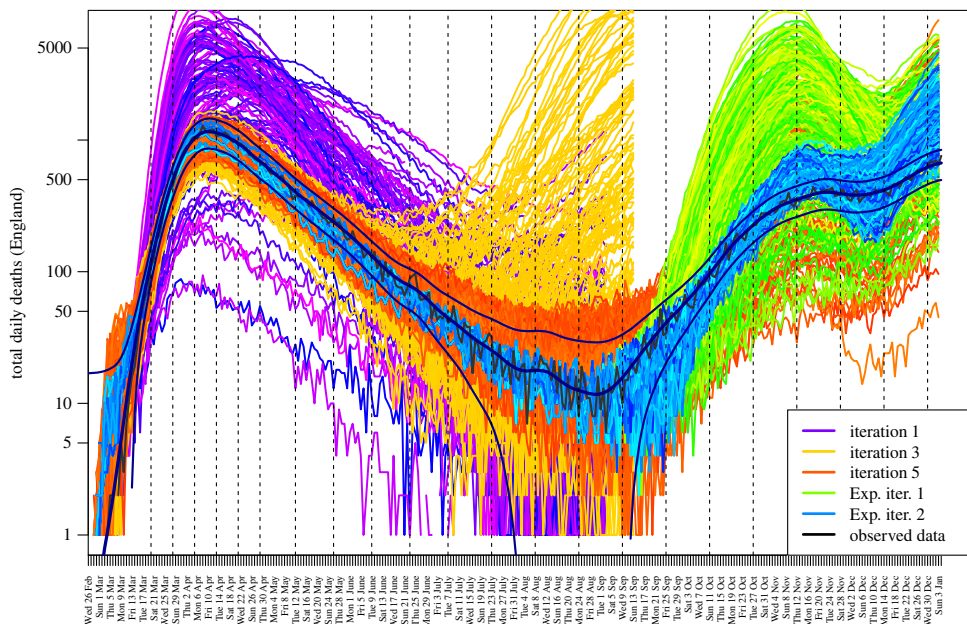


Figure 4. The JUNE output for total daily deaths in England in 2020, for several iterations of the HM process. The smoothed and noisy data, along with the combined uncertainties due to σ_ϵ and σ_ϵ , are shown in black. (Online version in colour.)

individual plot panel as follows:

$$\rho(x') = \frac{V_{18}\{x \in \mathcal{X}_5 \mid x' \text{ fixed}\}}{V_{18}\{x \in \mathcal{X}_0 \mid x' \text{ fixed}\}}, \quad (3.4)$$

where $V_{18}\{\cdot\}$ denotes the 18-dimensional volume of the remaining space. $\rho(x')$ can therefore show where large or small amounts of non-improbable points can be found, conditioned on x' , providing further insight into the structure of \mathcal{X}_5 . Figure 5 gives insights into the constraints imposed on the parameters by the death data and corresponding uncertainty specification. For example, we see that we learn a lot about certain influential parameters such as β_{school} , which are fairly well constrained, while others such as $\beta_{\text{care home}}$ can take a wider range of values. Provisional investigations suggest we can further constrain $\beta_{\text{care home}}$ by adding deaths in care home settings to the calibration outputs. We also see interesting relationships between pairs of parameters, e.g. the reciprocal relations between β_{company} vs. $\beta_{\text{household}}$, suggesting one or other can be high, but not both. We see similar relations between β_{company} vs. M_{social} distancing β factor. However, one should be aware that the actual constraints imposed are higher dimensional in nature and cannot be fully represented by such two-dimensional plots, but that they can be explored further e.g. by examining the eigenstructure of \mathcal{X}_5 , as done in ref. [51]. Note that in using HM in this way, we do not seek to probabilize the non-improbable region as in a full Bayesian calibration, but we could go on to do this (e.g. by routing the emulators through an MCMC algorithm) if desired, but the additional information gained may be in part an artefact of the particular additional distributional choices that such an analysis requires, which may impact robustness and predictive accuracy.

While performing a full global exploration of the input parameter space is of course preferable, it is sometimes useful to perform a fast ‘look-ahead’ stage to check if such an expensive model is capable of fitting the next period of observed data, or whether model improvements are required. Figure 4 also shows the results of such an exercise, where we took eight runs with acceptable matches to the death data up to the end of August, and performed small 30-point five-dimensional Latin hypercube designs for each of the eight cases up until December 2020,

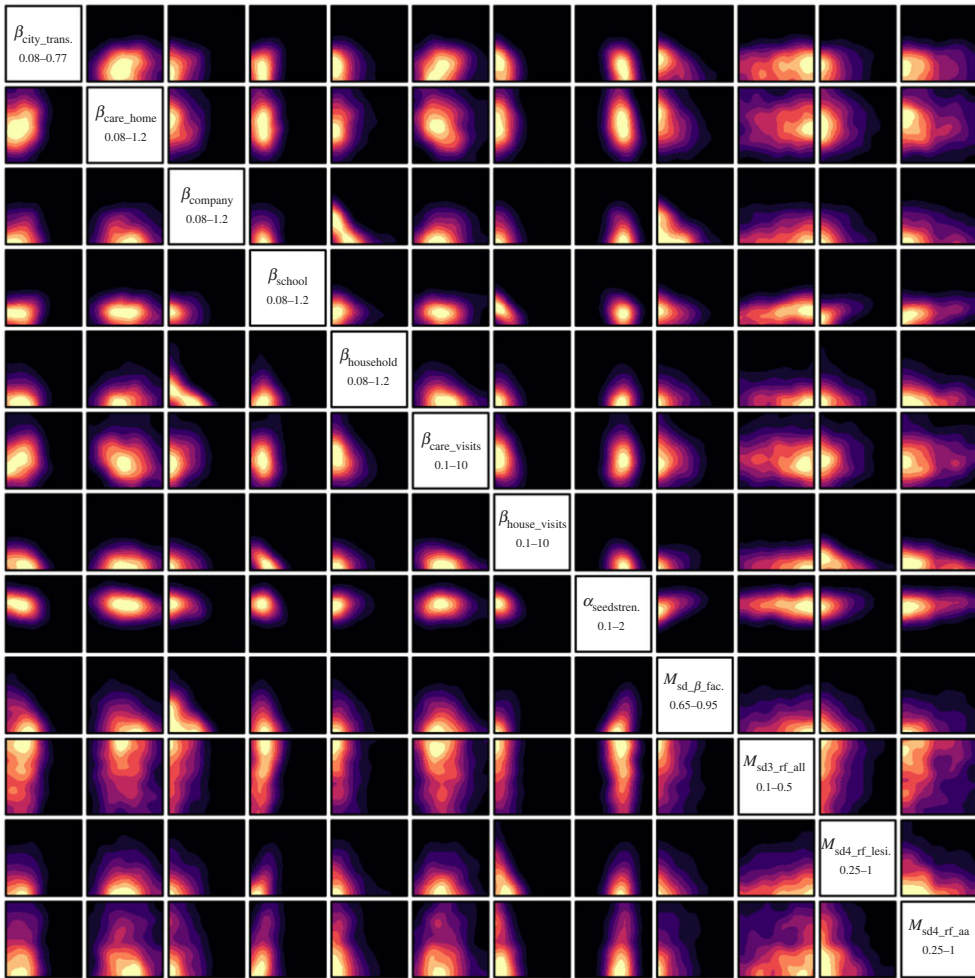


Figure 5. The optical depth $\rho(x')$ of various two-dimensional projections of the full 20-dimensional non-implausible region \mathcal{X}_5 found after the 5th iteration. The 12 most constrained inputs are shown, labelled on the diagonal (the remaining eight inputs were relatively unconstrained). The colour scales are standardized and linear in depth, with yellow showing maximum depth for that projection and purple/black showing minimum/zero depth. This region corresponds to the red runs in figure 4. (Online version in colour.)

now varying only five additional parameters relevant to the second COVID-19 wave (social distancing for schools, leisure, and non-leisure activities, November lockdown and B.1.1.7 variant infectiousness), the output of which is given by the green lines. One iteration of HM was performed to reduce the five-dimensional parameter space in each case, and a new set of runs designed, which are shown as blue lines in figure 4. We see reasonable matches to the first part of the second COVID-19 wave, with perhaps a late take-off in early September, and a partial overshoot in November to December, suggesting that JUNE may well provide acceptable matches after a full HM.

To give more detail, figure 6 shows a single unsmoothed run (red lines), from this final batch, but now for hospital deaths and total deaths for England and all seven regions, and shows the sort of quality of matches we are seeing so far. The black points give the (unsmoothed) death data and the combined uncertainties due to σ_e and σ_e shown as the blue lines. The fact that JUNE matches several regions simultaneously, at least over the first wave, without resorting to any region specific parameters, suggests that geographical variations in the relative importance

in different types of interaction drove/affected the different epidemic curves in those regions. Further, more detailed investigation to confirm this is of course required. We leave the extension to 2021 and beyond to the future work, as this requires the complex behavioural and partial restrictions (on travel, visiting relatives, etc.) imposed over the December 2020 Christmas period (the ‘cancelled Christmas’) and the January to April 2021 lockdown and subsequent staggered release to be implemented and tested, possibly requiring additional time-dependent parameters, which is the ongoing work (although we note that vaccines and multiple variants have already been implemented in JUNE). We also note that the process of using complex models combined with emulators and appropriate uncertainties to make realistic predictions over such periods is a substantive UQ topic in its own right, which deserves separate treatment [52].

(d) Discussion

Models such as JUNE, with its high level of demographic and spatial granularity, may become important tools to aid local and national decision makers. However, to fully exploit the nuances of such complex and expensive ABMs, efficient and comprehensive calibration methods are required. We demonstrated the emulation of JUNE, providing insight into the model structure, and employed HM to identify the region of parameter space yielding reasonable matches to national and regional level hospital and total death data for the first COVID-19 wave. Such techniques form an essential tool for the future use of complex epidemiological ABMs, expanding our capabilities to combine detailed models with rigorous UQ. The ability to perform global exploration of the parameter spaces of expensive models of this form and to embed this within a broader UQ framework is vital for making predictions with realistic uncertainty statements and hence vital for subsequent decision support.

4. Outlook/future directions

Our work represents an important step towards the full exploitation of highly granular and detailed ABMs in health settings and elsewhere, harnessing the full depth of their simulations in providing high-quality understanding of critical dynamics and robust quantitative projections for improved decision support.

The next steps in this project are to include further outputs of interest within the HM for JUNE, (hospitalizations, case rates, age categories, etc.) and to examine smaller geographic regions, in which the stochasticity of JUNE will become more pronounced, compared to the national/regional level where it is somewhat subdominant. This will require more sophisticated emulator strategies [11], and if we are interested in detailed spatial predictions, will require the updating of the JUNE state vector using UQ style data-augmentation techniques [53]. Beyond this, these UQ methods are currently being incorporated wherever JUNE is being employed e.g. by the UN for Cox’s Bazaar [27], a refugee camp in Bangladesh, and for Rhineland-Palatinate [28], one of Germany’s federal states.

In addition, we plan to use the model to investigate in more detail social imbalances in COVID-19 attack rates and infection-fatality ratios, which are relatively easy to trace in a model such as JUNE. Supplementing the model with the elaborate UQ techniques will allow us to identify, in more detail and with increased certainty, important correlations between socio-economic markers of the population and the infection dynamics and outcomes.

Data accessibility. A full open source code base and implementation examples are available through github: <https://github.com/IDAS-Durham/JUNE>; pypi: <https://pypi.org/project/june/>. The history matching and emulation methods are available in the ‘hmer’; R package v1.0 (available from CRAN). The version of JUNE used for this work was v1.0 [54].

Authors’ contributions. I.V.: formal analysis, funding acquisition, investigation, methodology, visualization, writing—original draft, writing—review and editing; J.O.: investigation, methodology, writing—review and editing; J.A.: conceptualization, data curation, methodology, software, writing—review and editing; C.C.: conceptualization, data curation, methodology, software, writing—review and editing; J.F.: investigation,

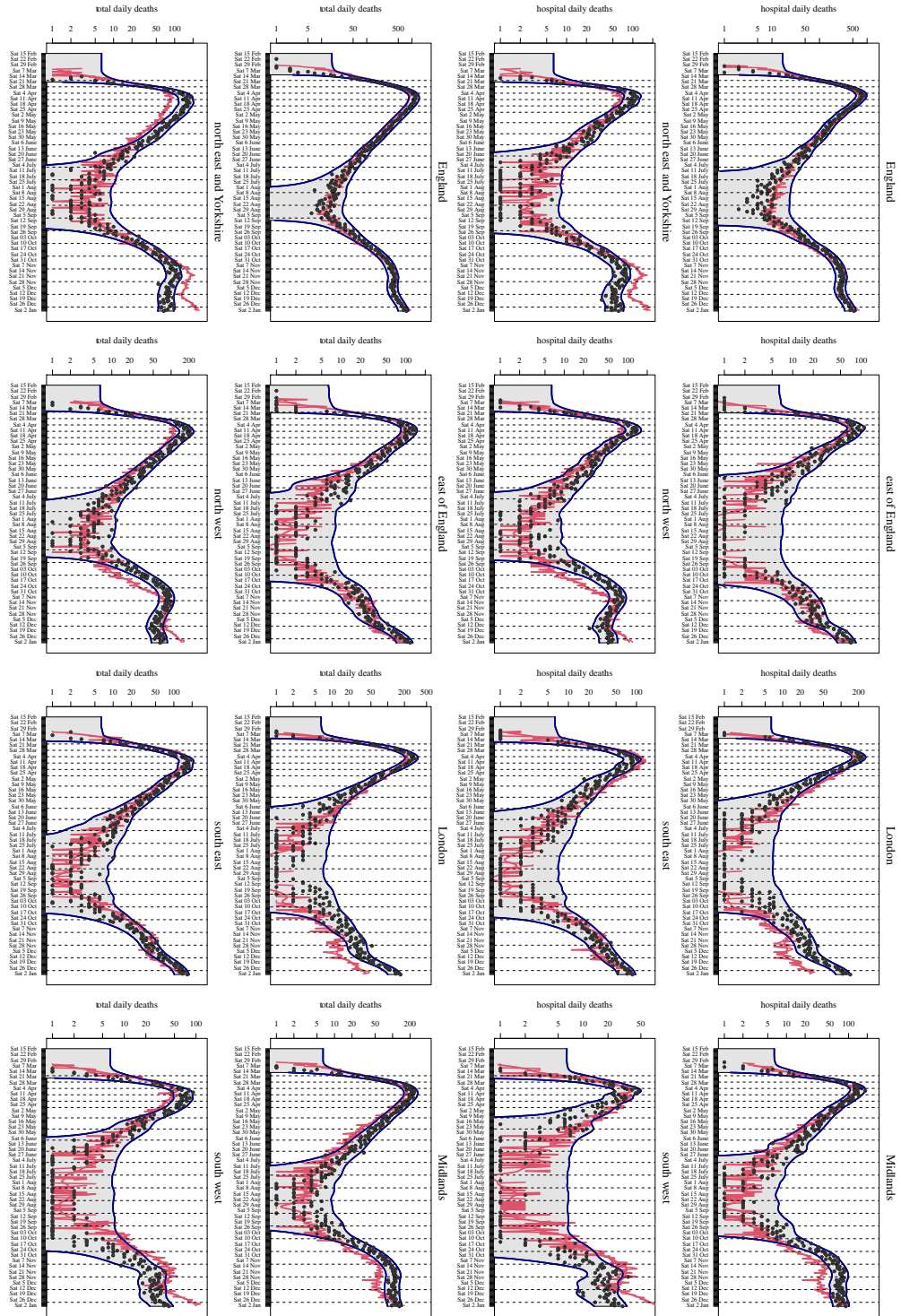


Figure 6. A single JUNE run (red lines), from the second exploratory iteration (i.e. one of the blue lines in figure 4). The panels show hospital deaths (rows 1 and 2, viewed in landscape) and total deaths (rows 3 and 4, viewed in landscape) for England and the seven regions, as given in the plot titles. The black points give the (unsmoothed) death data, and the combined uncertainties due to σ_ϵ and σ_ϵ are shown as the blue lines. (Online version in colour.)

software, writing—review and editing; A.Q.: conceptualization, data curation, investigation, methodology, software, writing—review and editing; A.S.: data curation, investigation, software, writing—review and editing; D.S.: data curation, investigation, software, writing—review and editing; H.T.: data curation, investigation, software, writing—review and editing; M.T.: data curation, investigation, software, writing—review and editing; J.W.: data curation, investigation, software, writing—review and editing; T.C.: conceptualization, data curation, investigation, resources, writing—review and editing; K.F.: conceptualization, data curation, investigation, methodology, resources, writing—review and editing; F.K.: conceptualization, data curation, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, writing—original draft, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. J.A.B., C.C.L., A.Q.B, A.S., H.T., and J.W. thank the STFC-funded Centre for Doctoral Training in Data-Intensive Science for financial support. J.F., J.O., and M.T.'s work is funded through the UKRI project 'Waves, Lock-Downs, And Vaccines - Decision Support And Model With Superb Geographical And Sociological Resolution' (EP W011956), and D.S. and J.W. are supported through an STFC Impact Acceleration Award. F.K. gratefully acknowledges funding as Royal Society Wolfson Research fellow. I.V. gratefully acknowledges Wellcome funding (218261/Z/19/Z). This work used the DiRAC@Durham facility managed by the Institute for Computational Cosmology on behalf of the STFC DiRAC HPC Facility (www.dirac.ac.uk) and additional computing resources provided by the Hartree Centre and the JASMIN facility. Durham's equipment was funded by BEIS capital funding via STFC capital grant nos. ST/K00042X/1, ST/P002293/1, ST/R002371/1 and ST/S002502/1, Durham University and STFC operations grant no. ST/R000832/1. DiRAC is part of the National e-Infrastructure.

Acknowledgements. This work was undertaken as a contribution to the Rapid Assistance in Modelling the Pandemic (RAMP) initiative, coordinated by the Royal Society. We are indebted to a number of people who shared their insights into various aspects of the project with us: we would like to thank Sinclair Sutherland for his patience and support in using the ONS database of the census data. We are grateful to Bryan Lawrence, Grenville Lister, Sadie Bartholomew and Valeriu Predoi from the National Centre of Atmospheric Science and the University of Reading for assistance in improving the computational performance of the model. The authors are grateful for inspiring collaboration with Richard Bower, Aoife Curran, Edward Elliott, Miguel Icaza-Lizaola, and Julian Williams in initial phases of the project.

References

1. Aylett-Bullock J *et al.* 2021 JUNE: open-source individual-based epidemiology simulation. *R. Soc. Open Sci.* **8**, 210506. (doi:10.1098/rsos.210506)
2. Vernon I, Goldstein M, Bower RG. 2010 Galaxy formation: a Bayesian uncertainty analysis. *Bayesian Anal.* **5**, 619–670. (doi:10.1214/10-ba524)
3. Williamson D, Goldstein M, Allison L, Blaker A, Challenor P, Jackson L, Yamazaki K. 2013 History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble. *Clim. Dyn.* **41**, 1703–1729. (doi:10.1007/s00382-013-1896-4)
4. Vernon I, Liu J, Goldstein M, Rowe J, Topping J, Lindsey K. 2018 Bayesian uncertainty analysis for complex systems biology models: emulation, global parameter searches and evaluation of gene functions. *BMC Syst. Biol.* **12**, 1–29. (doi:10.1186/s12918-017-0484-3)
5. Craig PS, Goldstein M, Seheult AH, Smith JA. 1997 Pressure matching for hydrocarbon reservoirs: a case study in the use of Bayes linear strategies for large computer experiments (with discussion). In *Case Studies in Bayesian Statistics* (eds C Gatsonis, JS Hodges, RE Kass, R McCulloch, P Rossi, ND Singpurwalla), vol. 3, pp. 36–93. New York: Springer-Verlag.
6. Santner TJ, Williams BJ, Notz WI. 2003 *The design and analysis of computer experiments*. New York: Springer-Verlag.
7. Kennedy MC, O'Hagan A. 2000 Predicting the output from a complex computer code when fast approximations are available. *Biometrika* **87**, 1–13. (doi:10.1093/biomet/87.1.1)
8. O'Hagan A. 2006 Bayesian analysis of computer code outputs: a tutorial. *Reliab. Eng. Syst. Saf.* **91**, 1290–1300. (doi:10.1016/j.res.2005.11.025)
9. Rasmussen CE, Williams CKI 2006 *Gaussian processes for machine learning*. Massachusetts Institute of Technology. MIT Press.

10. Andrianakis I, Vernon IR, McCreesh N, McKinley TJ, Oakley JE, Nsubuga RN, Goldstein M, White RG. 2015 Bayesian history matching of complex infectious disease models using emulation: a tutorial and a case study on HIV in Uganda. *PLoS Comput. Biol.* **11**, e1003968. (doi:10.1371/journal.pcbi.1003968)
11. Andrianakis I, Vernon I, McCreesh N, McKinley TJ, Oakley JE, Nsubuga RN, Goldstein M, White RG. 2017 History matching of a complex epidemiological model of human immunodeficiency virus transmission by using variance emulation. *J. R. Stat. Soc.: Ser. C (Appl. Stat.)* **66**, 717–740. (doi:10.1111/rssc.12198)
12. Sacks J, Welch WJ, Mitchell TJ, Wynn HP. 1989 Design and analysis of computer experiments. *Stat. Sci.* **4**, 409–435.
13. Currin C, Mitchell T, Morris M, Ylvisaker D. 1991 Bayesian prediction of deterministic functions with applications to the design and analysis of computer experiments. *J. Am. Stat. Assoc.* **86**, 953–963. (doi:10.1080/01621459.1991.10475138)
14. Kennedy MC, O'Hagan A. 2001 Bayesian calibration of computer models. *J. R. Stat. Soc., Ser. B* **63**, 425–464. (doi:10.1111/1467-9868.00294)
15. Goldstein M. 1999 Bayes linear analysis. In *Encyclopaedia of Statistical Sciences* (eds S Kotz *et al.*), pp. 29–34. Hoboken, NJ: Wiley.
16. Goldstein M, Wooff DA. 2007 *Bayes linear statistics: theory and methods*. Chichester: Wiley.
17. Bastos TS, O'Hagan A. 2008 Diagnostics for Gaussian process emulators. *Technometrics* **51**, 425–438. (doi:10.1198/TECH.2009.08019)
18. Vernon I, Goldstein M, Bower RG. 2014 Galaxy formation: Bayesian history matching for the observable universe. *Stat. Sci.* **29**, 81–90. (doi:10.1214/12-STS412)
19. Iskauskas A. 2022 hmer: History Matching and Emulation Package. R package version 1.0.0. See <https://CRAN.R-project.org/package=hmer>.
20. Goldstein M, Rougier JC. 2009 Reified Bayesian modelling and inference for physical systems (with discussion). *J. Stat. Plann. Inference* **139**, 1221–1239. (doi:10.1016/j.jspi.2008.07.019)
21. Jackson SE, Vernon I, Liu J, Lindsey K. 2020 Understanding hormonal crosstalk in Arabidopsis root development via emulation and history matching. *Stat. Appl. Genet. Mol. Biol.* **19**, 20180053. (doi:10.1515/sagmb-2018-0053)
22. Goldstein M, Seheult A, Vernon I. 2013 Assessing model adequacy. In *Environmental Modelling: Finding Simplicity in Complexity* (eds J Wainwright, M Mulligan), 2nd ed. Chichester, UK: John Wiley & Sons, Ltd.
23. Craig PS, Goldstein M, Rougier JC, Seheult AH. 2001 Bayesian forecasting for complex systems using computer simulators. *J. Am. Stat. Assoc.* **96**, 717–729. (doi:10.1198/016214501753168370)
24. Brynjarsdottir J, O'Hagan A. 2014 Learning about physical parameters: the importance of model discrepancy. *Inverse Prob.* **30**, 24. (doi:10.1088/0266-5611/30/11/114007)
25. Pukelsheim F. 1994 The three sigma rule. *Am. Stat.* **48**, 88–91.
26. McKinley TJ, Vernon I, Andrianakis I, McCreesh N, Oakley JE, Nsubuga RN, Goldstein M, White RG. 2018 Approximate Bayesian computation and simulation-based inference for complex stochastic epidemic models. *Stat. Sci.* **33**, 4–18. (doi:10.1214/17-STS618)
27. Aylett-Bullock J *et al.* 2021 Operational response simulation tool for epidemics within refugee and IDP settlements: a scenario-based case study of the Cox's Bazar settlement. *PLoS Comput. Biol.* **10**, e1009360. (doi:10.1371/journal.pcbi.1009360)
28. Schott M *et al.* 2021 JUNE-Germany: an agent-based epidemiology simulation including multiple virus strains, vaccinations and testing campaigns for Germany (accessed on 1 October 2021).
29. Office for National Statistics. QS103EW (Age by single year); see <https://www.nomisweb.co.uk/census/2011/qs103ew>.
30. Office for National Statistics. Sex by age; see <https://www.nomisweb.co.uk/census/2011/lc1117ew>.
31. Office for National Statistics. DC2101EW (Ethnic group by sex by age); see <https://www.nomisweb.co.uk/census/2011/dc2101ew>.
32. Office for National Statistics. LC1109EW (Household composition by age by sex). See <https://www.nomisweb.co.uk/census/2011/lc1109ew>.
33. Office for National Statistics. Marriages in England and Wales; 2017. See <https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/marriagecohabitationandcivilpartnerships/datasets/marriagesinenglandandwales2013>.

34. Office for National Statistics. Birth characteristics in England and Wales: 2017. See <https://www.ons.gov.uk/releases/birthcharacteristicsinenglandandwales2017>. 2017.
35. Education and Skills Funding Agency. UK Register of Learning Providers. See <https://www.ukrlp.co.uk/>.
36. Office for National Statistics. UK SIC 2007. <https://www.ons.gov.uk/methodology/classificationsandstandards/ukstandardindustrialclassificationofeconomicactivities/uksic2007>. 2007.
37. Office for National Statistics. WU01EW (Location of usual residence and place of work by sex). See <https://www.nomisweb.co.uk/census/2011/wu01ew>.
38. Office for National Statistics. QS701EW (Method of travel to work); see <https://www.nomisweb.co.uk/census/2011/qs701ew>.
39. Gershuny JI, Sullivan O. 2015 United Kingdom Time Use Survey, 2014-2015. UK Data Service. SN: 8128. Available from: (doi:10.5255/UKDA-SN-8128-1)
40. Mossong J *et al.* 2008 Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med.* **5**, e74. (doi:10.1371/journal.pmed.0050074)
41. Klepac P, Kissler S, Gog J. 2018 Contagion! The BBC Four Pandemic—The model behind the documentary. *Epidemics* **24**, 49–59. (doi:10.1016/j.epidem.2018.03.003)
42. UK Department for Education. Attendance in education and early years settings during the coronavirus (COVID-19) outbreak. See <https://www.gov.uk/government/collections/attendance-in-education-and-early-years-settings-during-the-coronavirus-covid-19-outbreak>. 2020.
43. Office for National Statistics. Coronavirus and key workers in the UK. See <https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/articles/coronavirusandkeyworkersintheuk/2020-05-15>. 2020.
44. Institute for Fiscal Studies. Sector shutdowns during the coronavirus crisis: which workers are most exposed?. <https://www.ifs.org.uk/publications/14791>. 2020.
45. YouGov. YouGov COVID-19 behaviour changes tracker: avoiding going to work. See <https://yougov.co.uk/topics/international/articles-reports/2020/03/17/personal-measures-taken-avoid-covid-19>. 2020.
46. Office for National Statistics. Deaths registered weekly in England and Wales, provisional. See <https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/datasets/weeklyprovisionalfiguresondeathsregisteredinenglandandwales>.
47. Office for National Statistics. Deaths involving COVID-19 in the care sector, England and Wales. See <https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/articles/deathsinvolvingcovid19inthecaresectorenglandandwales/deathsoccurringupto12june2020andregisteredupto20june2020provisional#characteristics-of-care-home-residents-who-died-from-covid-19>.
48. Scientific Advisory Group for Emergencies. Dynamic CO-CIN report to SAGE and NERVTAG - 30 June 2020; see <https://www.gov.uk/government/publications/dynamic-co-cin-report-to-sage-and-nervtag-30-june-2020>.
49. Scientific Advisory Group for Emergencies. Dynamic CO-CIN report to SAGE and NERVTAG, 13 May 2020. See <https://www.gov.uk/government/publications/dynamic-co-cin-report-to-sage-and-nervtag-13-may-2020>.
50. Oakley J, O'Hagan A. 2002 Bayesian inference for the uncertainty distribution of computer model outputs. *Biometrika* **89**, 769–784. (doi:10.1093/biomet/89.4.769)
51. Bower RG, Vernon I, Goldstein M, Benson AJ, Lacey CG, Baugh CM, Cole S, Frenk CS. 2010 The parameter space of galaxy formation. *MonNotRoyAstronSoc* **96**, 717–729.
52. Goldstein M, Rougier JC. 2006 Bayes linear calibrated prediction for complex systems. *J. Am. Stat. Assoc.* **101**, 1132–1143. (doi:10.1198/016214506000000203)
53. Sinsbeck M, Tartakovsky DM. 2015 Impact of data assimilation on cost-accuracy tradeoff in multifidelity models. *SIAM/ASA J. Uncertain. Quantification* **3**, 954–968. (doi:10.1137/141001743)
54. Quera-Bofarull A *et al.* 2021 JUNE: open-source individual-based epidemiology simulation, v1. See <https://doi.org/10.5281/zenodo.4925939>.