RESEARCH ARTICLE

A Bayesian Ensemble Approach for Epidemiological Projections

Tom Lindström[1,2,3,4]*, **Michael Tildesley**[3,5], **Colleen Webb**[2,3]

1 Department of Physics, Chemistry and Biology, Linköping University, Linköping, Sweden, **2** Department of Biology, Colorado State University, Fort Collins, Colorado, United States of America, **3** US National Institute of Health, Bethesda, Maryland, United States of America, **4** University of Exeter, Exeter, United Kingdom, **5** School of Veterinary Medicine and Science, University of Nottingham, Leicestershire, United Kingdom

* tomli@ifm.liu.se

OPEN ACCESS

Citation: Lindström T, Tildesley M, Webb C (2015) A Bayesian Ensemble Approach for Epidemiological Projections. PLoS Comput Biol 11(4): e1004187. doi:10.1371/journal.pcbi.1004187

Editor: Sergei L. Kosakovsky Pond, University of California San Diego, UNITED STATES

Received: October 1, 2014

Accepted: February 11, 2015

Published: April 30, 2015

Copyright: © 2015 Lindström et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Data available, subject to suitable confidentiality agreements, from Animal and Plant Health Agency (APHA), RADAR@apha.gsi.gov.uk

Funding: This work was supported by the RAPIDD program of the Science & Technology Directorate and the Fogarty International Center, National Institutes of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Mathematical models are powerful tools for epidemiology and can be used to compare control actions. However, different models and model parameterizations may provide different prediction of outcomes. In other fields of research, ensemble modeling has been used to combine multiple projections. We explore the possibility of applying such methods to epidemiology by adapting Bayesian techniques developed for climate forecasting. We exemplify the implementation with single model ensembles based on different parameterizations of the Warwick model run for the 2001 United Kingdom foot and mouth disease outbreak and compare the efficacy of different control actions. This allows us to investigate the effect that discrepancy among projections based on different modeling assumptions has on the ensemble prediction. A sensitivity analysis showed that the choice of prior can have a pronounced effect on the posterior estimates of quantities of interest, in particular for ensembles with large discrepancy among projections. However, by using a hierarchical extension of the method we show that prior sensitivity can be circumvented. We further extend the method to include a priori beliefs about different modeling assumptions and demonstrate that the effect of this can have different consequences depending on the discrepancy among projections. We propose that the method is a promising analytical tool for ensemble modeling of disease outbreaks.

Author Summary

Policy decisions in response to emergent disease outbreaks use simulation models to inform the efficiency of different control actions. However, different projections may be made, depending on the choice of models and parameterizations. Ensemble modeling offers the ability to combine multiple projections and has been used successfully within other fields of research. A central issue in ensemble modeling is how to weight the projections when they are combined. For this purpose, we here adapt and extend a weighting method used in climate forecasting such that it can be used for epidemiological considerations. We investigate how the method performs by applying it to ensembles of projections for the UK foot and mouth disease outbreak in UK, 2001. We conclude that the method is a promising analytical tool for ensemble modeling of disease outbreaks.

Introduction

Epidemiological forecasting is inherently challenging because the outcome often depends on largely unpredictable characteristics of hosts and pathogens as well as contact structure and pathways that mediate transmission [1]. Faced with such uncertainty, policy makers must still make decisions with high stakes, both in terms of health and economics. For instance, global annual malaria mortality was recently estimated at around 1.1 million [2] and to optimize control efforts, policy makers must make seasonal predictions about spatiotemporal patterns [3]. The prospect of an emergent pandemic influenza outbreak remains a global threat and emergency preparedness must evaluate the costs and benefits of control measures such as border control, closing of workplaces and/or schools as well as different vaccination strategies [4]. Livestock diseases are major concerns for both animal welfare and economics. As an example, the United Kingdom (UK) 2001 outbreak of foot and mouth disease (FMD) involved culling of approximately 7 million animals, either in an effort to control the disease or for welfare reasons, and the total cost has been estimated at £8 billion [5]. To minimize the size and duration of future outbreaks, various strategies for culling and vaccination must be compared [6–8]. As a tool to address these challenging tasks, mathematical models offer the possibility to explore different scenarios, thereby informing emergency preparedness and response to epidemics [1,9–12].

The predictive focus of epidemiological models can either be classified as forecasting or projecting [13]. Forecasting aims at estimating what will happen and can be used for example to predict seasonal peaks of outbreaks [3,14] or to identify geographical areas of particular concern [15]. Projecting, which is the main focus of this study, instead aims at comparing different scenarios and exploring what would happen under various assumptions of transmission, e.g. comparing the effectiveness of different control actions [7,16–19].

Whilst analytical models clearly provide important insight into observed dynamics and a theoretical understanding of epidemiology [20–22], there has been a shift in recent years towards stochastic simulation models for predictive purposes [1]. Typically, dynamic models are constructed and outbreaks are simulated repeatedly, thus generating predictive distributions of outcomes [1,17,18,23]. This variability in outcomes caused by the mere stochasticity of the transmission process includes one level of uncertainty, but still only relies on a single set of assumptions about the underlying disease transmission process. However, multiple assumptions can often be justified, leading to further uncertainty in the predictions. For instance, different models may have different projections because of different assumptions about transmission or because they incorporate different levels of detail. It may also be informative to explore different projections in terms of different parameterizations of a single model, for example corresponding to worst or best case scenarios. Faced with a set of projections, an important issue is how to combine these in a manner such that they can be used to inform policy.

The issue of multiple projections is not unique to the field of epidemiology, and various techniques of ensemble modeling have been used to merge projections based on different modeling assumptions. The key concept is that rather than relying on a single set of assumptions, a range of projections is used for predictive purposes. For instance, climate forecasting has employed ensemble techniques to account for uncertainty about initial conditions, parameter values and structure of the model design when predicting climate change [24,25]. Weather forecasting has been improved by combining the results of multiple models [26,27]. Similarly, hydrological model ensembles have been demonstrated to increase reliability of catchment forecasting [28] and have been used to assess the risk of flooding events [29]. Ensemble methods have also proven to be a powerful decision tool for medical diagnostics [30,31] and ecological considerations including management [32] and prediction of future species distribution [33].

Ensemble modeling has not yet been extensively used in epidemiology. However a few implementations exist, commonly by feeding climate or weather ensembles into disease models. Daszak et al. [34] coupled a set of climate projections to an environmental niche model of Nipah virus to predict future range distribution of the virus under climate change. Similarly, Guis et al. [35] investigated the effect of climate change on Bluetongue emergence in Europe by simulating outbreaks under different climate change scenarios. Focusing on a shorter time scale, Thomson et al. [3] used an ensemble of seasonal forecasts to predict the spatiotemporal pattern of within seasonal variation in malaria incidence. These studies all used a single disease model projection, coupled to an ensemble of climate or weather forecasts and the use of structurally different epidemiological models are to our knowledge still rare. However, Smith et al. [36] compared different malaria vaccination strategies by implementing an ensemble approach with different alterations of a base model. Also, in order to estimate global malaria mortality, Murray et al. [2] used weighted averages of different predictive models.

Given the success of ensemble methods in other fields, we expect that epidemiological implementations will increase. For that purpose however, there is a need for methods that combine multiple projections. A central issue in ensemble modeling is how to weight different projections, and we envisage four main procedures for this. Firstly, all models can be given equal weights. For instance, the IPCC 2001 report on climate change [37] used a set of climate models and gave the range of probable scenarios by averaging over different models and uncertainty by envelopes that included all scenarios. Gårdmark et al. [32] used seven ecological models for cod stock and argued that in order to prevent underestimation of uncertainty, weighted model averages are not to be used and when communicating with policy makers, it is preferable to present all included projections as well as the underlying assumptions behind them. A similar approach was also used by Smith et al. [36], who presented the prevalence of malaria under different vaccination strategies by medians of individual models and the range of the whole ensemble.

Secondly, expert opinions can be used to weight models. To our knowledge, no ensemble study has implemented weights based exclusively on expert opinion, but Bayesian model averaging can incorporate expert opinion as a subjective prior on model probabilities [38]. This approach relies on engaging stakeholders and communicating the underlying assumptions of the projections.

Thirdly, models can be weighted by agreement with other models. This approach was implemented by Räisänen and Palmer [39], who used cross-validation to weight climate models. As a more informal approach to the use of model consensus, the third IPCC report excluded two models because these predicted much higher global warming than the rest of the ensemble, thus acting as outliers [24].

Fourthly, weights can be determined by the models' ability to replicate data. If all models are fitted to the same data using likelihood based methods, weights can be given directly by Akaike or Bayesian Information Criterion (AIC or BIC) [40,41]. In the FMD context, this may be a suitable approach if all included models are data driven kernel models that estimate parameters from outbreak data, such as those proposed by Jewell et al. [42] or Tildesley et al. [43]. However, such weighting schemes would be unfeasible when including detailed simulation models that rely on a large number of parameters, that are determined by expert opinion or lab experiment, such as AusSpread [44], NAADSM [45] and InterSpread Plus [46]. We propose that the future of ensemble modeling for epidemiology will benefit from combining structurally different model types, and methods of weighting need to handle both kernel type models as well as detailed simulation models.

Thus, bias assessment is often confined to the ability of models to replicate observed summary statistics of interest, in particular when the resolution of data observation is on a courser scale than the model prediction [47]. Such methods have been developed within the field of climate forecasting. Giorgi and Mearns [48] introduced a formal framework in which model weights were assessed based on model bias compared to observed data as well as convergence, i.e. agreement with the model consensus. Tebaldi et al. [47] extended the approach to a Bayesian framework. This approach is appealing because it provides probability distributions of quantities of interest, hence uncertainty about the projected outcomes may be provided to policy makers. As such, it would be a suitable approach also for epidemiological predictions.

However, methods developed in one field might not be directly transferable to another. Tebaldi et al. [47] points out that lack of data at fine scale resolution is a limiting factor for climate forecasting. Yet, at courser resolution climate researchers have access to long time series of climate variables to assess model bias. Comparable data may be available for endemic diseases, such as malaria [36] or tuberculosis [49], or seasonally recurrent outbreaks, such as influenza [14] or measles [50]. However, for emerging diseases, long time series would rarely be available, making the lack of data an even bigger issue for epidemiology.

In this methodological paper we aim to explore the potential of using ensemble methods based on the approach presented by Tebaldi et al. [47] for epidemiological projections. The Tebaldi et al. methodology focus on ensembles where projections are made with different models and our aim is to provide a corresponding framework for disease outbreak projections. To investigate the potential of the framework for epidemiology, we here use variations of a single model as a proxy for different models, thus allowing us to investigate how the methodology performs under different levels of discrepancy among projections in the ensemble. We exemplify the implementation by using the UK 2001 FMD outbreak and projections modelled by different parameterizations of the Warwick model [7,9].

In the 2001 UK FMD outbreak, livestock on all infected premises (IPs) were culled. In addition, livestock on farms that were identified to be at high risk of infection were culled as either traditional dangerous contacts (DCs) or contiguous premises (CPs). CPs were defined as "a category of dangerous contact where animals may have been exposed to infection on neighboring infected premises" [5,8]. We start by focusing on ensemble prediction of epidemic duration under the control action employed during the 2001 outbreak compared with an alternative action that excludes CP culling. We investigate sensitivity to priors and explore a hierarchical Bayesian extension of the method to circumvent potential problems with prior sensitivity. We also explore the potential of including subjective a priori trust in the different modeling assumptions and extend the methodology further to allow incorporation of multiple epidemic quantities, here exemplified by adding number of infected and culled farms to the analysis. Through a simulation study, we finally explore the capacity and limitations of the proposed ensemble method, pinpointing some important features of ensemble modeling

Materials and Methods

We apply a terminology such that control actions refers to different strategies for disease control. In the ensemble, each action is simulated under different modeling assumptions about the underlying process, expressed as either different models or, as in the example described here, different parameterizations of the same model. We refer to the combination of control action and modeling assumption as a projection. Each projection is further simulated with several replicates, which produce different outcomes merely due to the stochasticity of simulation models. We are also interested in how discrepancy among projections influences the performance of the weighting method and refer to sets of different projections as different ensembles with small and large discrepancy. A flow chart that demonstrates the relationship between different concepts and weighting schemes are presented in Fig 1.

The Warwick model, control actions and ensembles

We focus on projections of FMD made by the Warwick model [7,9]. This model was developed in the early stages of the 2001 FMD outbreak by Keeling and coworkers to determine the potential for disease spread and the impact of intervention strategies [9]. Here, we utilized a modified version of the model used in 2001, and we briefly describe relevant aspects of the Warwick model with regard to ensemble modeling. Full details of the model can be found in [7,9]. The rate at which an infectious farm I infects a susceptible farm J is given by:

$$Rate_{IJ} = Sus_J \times Trans_I \times K(d_{IJ}) \tag{1}$$

where

$$Sus_J = ([Z_{sheep,J}]^{p_{s,S}} S_{sheep} + [Z_{cow,J}]^{p_{c,S}} S_{cow}) \tag{2}$$

is the susceptibility of farm J and

$$Trans_I = ([Z_{sheep,I}]^{p_{s,T}} T_{sheep} + [Z_{cow,I}]^{p_{c,T}} T_{cow}) \tag{3}$$

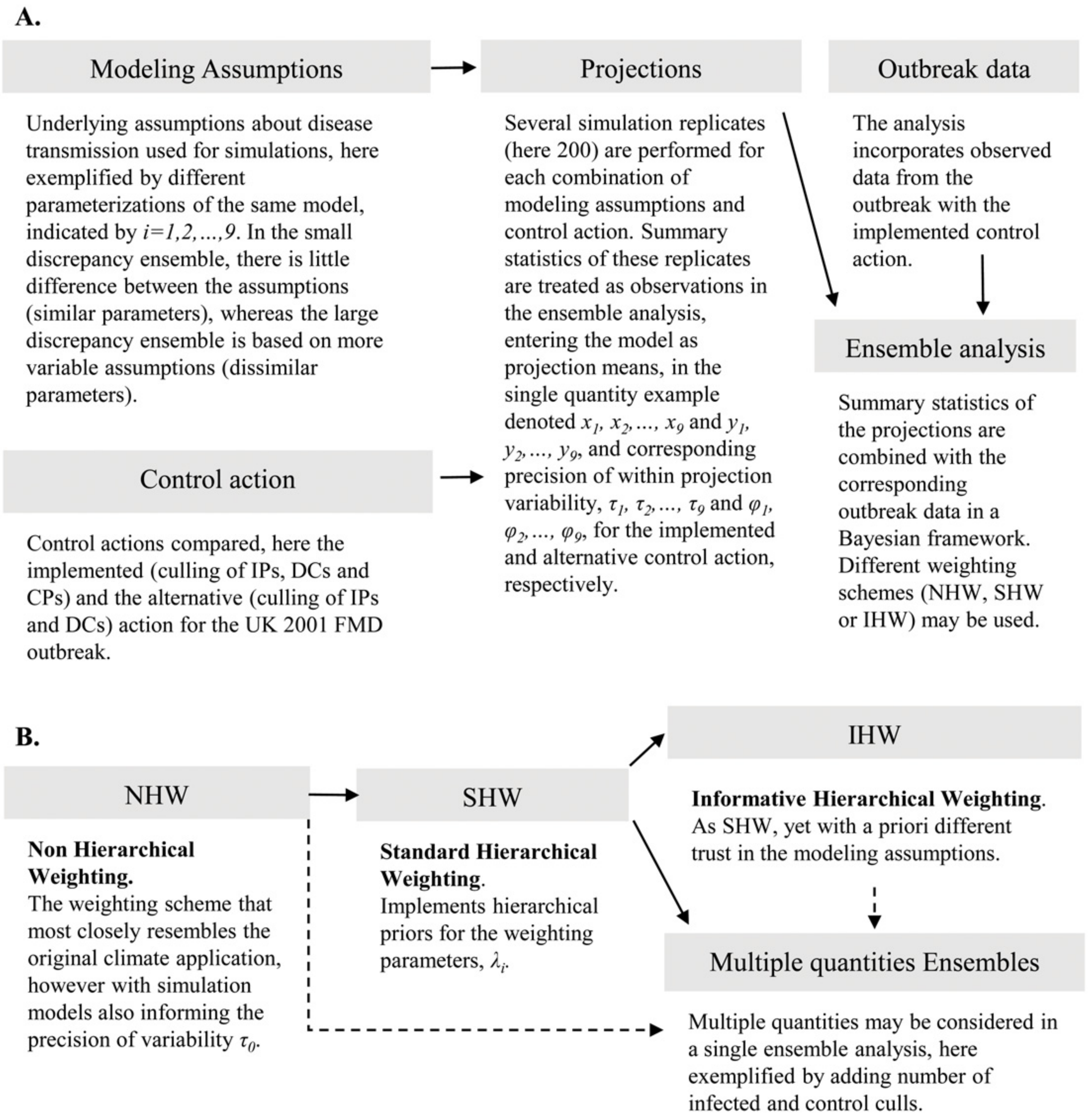
is the transmissibility of farm I and $K(d_{IJ})$ is the distance-dependent transmission kernel, estimated from contact tracing [9]. In this model $Z_{s,\ I}$ is the number of livestock species s (sheep or cow) recorded as being on farm I, S_s and T_s measure the region and species-specific susceptibility and transmissibility, d_{IJ} is the distance between farms I and J and $K(d_{IJ})$ is the distance dependent transmission kernel. The parameters, $p_{s,\ S}$, $p_{c,\ S}$, $p_{s,\ T}$ and $p_{c,\ T,}$ are power law parameters that account for a non-linear increase in susceptibility and transmissibility as animal numbers on a farm increase. Previous work has indicated that a model with power laws provides a closer fit to the 2001 data than when these powers are set to unity [43,51,52].

This version of the model has previously been parameterized to fit to the 2001 FMD outbreak [43]. Region-specific transmissibility and susceptibility parameters (and associated power laws) capture specific epidemiological characteristics and policy measures used in the main hot spots of Cumbria, Devon and the Welsh and Scottish borders. The model is therefore fitted to five regions: Cumbria, Devon, Wales, Scotland and the rest of England (excluding Cumbria and Devon). A table listing all the parameter values used in this model is given in Tildesley et al. [43].

In order to obtain multiple modeling assumptions for ensemble modeling, we specified different transmission rates, i as

$$Rate_{IJi} = Sus_J \times k_{1i} Trans_I \times K(k_{2i} d_{IJ}) \tag{4}$$

where k_{1i} and k_{2i} are constants, specific for each modeling assumption, that scale the transmissibility and the spatial kernel, respectively. $k_{1i} = k_{2i} = 1$, follow the parameterizations of Tildesley et al. [43] and by decreasing or increasing these constants, we obtain parameterizations that correspond to best or worst case expectations about the transmissibility and spatial range of transmission. We are interested in how the level of discrepancy among modeling assumptions influences the performance of the ensemble method and we therefore created two different ensembles with different k_{1i} and k_{2i}, as listed in Table 1. We refer to these as the large and small



Underlying assumptions about disease transmission used for simulations, here exemplified by different parameterizations of the same model, indicated by $i=1,2,\ldots,9$. In the small discrepancy ensemble, there is little difference between the assumptions (similar parameters), whereas the large discrepancy ensemble is based on more variable assumptions (dissimilar parameters).

Control actions compared, here the implemented (culling of IPs, DCs and CPs) and the alternative (culling of IPs and DCs) action for the UK 2001 FMD outbreak.

Several simulation replicates (here 200) are performed for each combination of modeling assumptions and control action. Summary statistics of these replicates are treated as observations in the ensemble analysis, entering the model as projection means, in the single quantity example denoted $x_1, x_2,\ldots, x_9$ and $y_1, y_2,\ldots, y_9$, and corresponding precision of within projection variability, $\tau_1, \tau_2,\ldots, \tau_9$ and $\varphi_1, \varphi_2,\ldots, \varphi_9$, for the implemented and alternative control action, respectively.

The analysis incorporates observed data from the outbreak with the implemented control action.

Summary statistics of the projections are combined with the corresponding outbreak data in a Bayesian framework. Different weighting schemes (NHW, SHW or IHW) may be used.

Non Hierarchical Weighting. The weighting scheme that most closely resembles the original climate application, however with simulation models also informing the precision of variability τ_0.

Standard Hierarchical Weighting. Implements hierarchical priors for the weighting parameters, λ_i.

Informative Hierarchical Weighting. As SHW, yet with a priori different trust in the modeling assumptions.

Multiple quantities may be considered in a single ensemble analysis, here exemplified by adding number of infected and control culls.

Fig 1. Overview of analyses and methods developments. Panel A presents a conceptual flowchart, showing that modeling assumptions are combined with control actions to simulate projections. These are then combined with observed outbreak data in the ensemble analysis. Panel B presents the methods developments in this study, indicating that we start with the Non Hierarchical Weighting (NHW) scheme, which is most similar to the original climate application. We then extend this to the Standard Hierarchical Weighting (SHW) scheme, and subsequently to Informative Hierarchical Weighting (IHW). We also extend the analysis to consider multiple epidemic quantities. The dashed lines indicate combinations for which the method and the supplied algorithm (S1 File) can be used but are not treated explicitly in the presented examples.

doi:10.1371/journal.pcbi.1004187.g001

Table 1. Scaling constants *k1*$_i$ and *k2*$_i$ used for each modeling assumption *i* in the small and large discrepancy ensemble.

Modeling assumption (*i*)	Small discrepancy		Large discrepancy	
	k_{1i}	k_{2i}	k_{1i}	k_{2i}
1	1	1	1	1
2	0.95	1	0.8	1
3	1.05	1	1.2	1
4	1	0.95	1	0.8
5	0.95	0.95	0.8	0.8
6	1.05	0.95	1.2	0.8
7	1	1.05	1	1.2
8	0.95	1.05	0.8	1.2
9	1.05	1.05	1.2	1.2

doi:10.1371/journal.pcbi.1004187.t001

discrepancy ensemble, corresponding to large and small differences, respectively, in the underlying modeling assumptions used for projections.

DCs in our model were determined based upon both prior infection by an IP and future risk of infection in the same way as in previous work [8]. CPs were defined as farms that share a common boundary and were determined on an individual basis. The model was seeded with the farms that were predicted to be infected prior to the introduction of movement restrictions on the 23rd February. For each modeling assumption *i* and control action, 200 replicates were simulated and each simulation progressed until the epidemic died out.

Adapting the Tebaldi et al. method for emerging diseases

To demonstrate concepts and explore the potential of using the Tebaldi et al. [47] approach for epidemiological considerations we initially focus on outbreak duration. This is often considered to be the most costly aspect of FMD outbreaks due to its implication for trade [53]. In section 2.7 we extend the methodology to multiple epidemic quantities. However, the outbreak duration example offers transparent transition from the original climate analysis of Tebaldi et al. [47] that considers the ensemble estimated difference between current and future mean temperatures. In order to introduce the framework to epidemiology, we consider the difference between the implemented and an alternative control action, attempting to show whether the inclusion of CP culling was an appropriate choice given the conditions at the start of the outbreak. As this is a post outbreak analysis, we know the final outbreak duration of the observed outbreak, but that is just a single realization and due to the stochastic nature of disease transmission, the exact outcome may be quite variable. We also have no observed outbreak under the alternative control action to compare with the implemented control. Under these conditions, the most appropriate quantities to compare are the mean duration of a large number of outbreaks under the two control actions, something that can only be achieved through epidemic modeling.

We are interested in comparing projections under the implemented control action to the observed data in order to estimate model weights. Using the Bayesian method of Tebaldi et al. [47], weights as well as statistics of the outbreak, like duration, are considered unknown random variables, and we denote the mean outbreak duration under the implemented and the alternative control action as μ and ν, respectively, corresponding to the mean current and future temperature, respectively, for the climate application. In order to fit with the normal assumptions of the method, we consider log-duration in the analysis.

Weights are expressed through precision, $\boldsymbol{\lambda} = \lambda_1, \lambda_2, \ldots, \lambda_n$, with λ_i denoting the precision of modeling assumption *i*. The projection specific parameter x_i indicates the mean of all replicates under the implemented control action (analog of current climate) for modeling assumption *i*. For the UK 2001 outbreak this included culling of IPs, DCs and CPs. The corresponding projection for the alternative control action (analog of future climate), that included culling of IPs and DCs is denoted y_i. The relationship between projections and ensemble parameters is expressed as

$$x_i \sim \text{Normal}(\mu, \lambda_i^{-1}) \tag{5}$$

$$y_i \sim \text{Normal}(\nu + \beta(x_i - \mu), (\theta\lambda_i)^{-1}), \tag{6}$$

with Normal(μ,λ_i^{-1}) denoting the normal distribution with mean μ and variance λ_i^{-1}. Parameter θ is included to allow for difference in overall precision of the modeling assumptions under implemented and alternative control actions. However, since projections x_i and y_i are based on simulations, it is fair to assume that modeling assumption *i* that has a high precision for the observed control action also will do well for the unobserved action. This is incorporated by the λ_i term in both Eqs 5 and 6. For the same reason, we may expect that a projection of a large x_i also corresponds to a large value for y_i and thus β is included to allow for correlation between corresponding projections for the two control actions; a projection that e.g. over-predicts duration of the outbreak for the observed control action can be expected to also over-predict the alternative control action.

The analysis of Tebaldi et al. [47] also assessed bias of projections by their ability to reproduce observed current climate by incorporating the relationship between observed current climate x_0, an unobserved true mean climate variable (μ) and the precision of natural variability τ_0 through

$$x_0 \sim \text{Normal}(\mu, \tau_0^{-1}). \tag{7}$$

In climate modeling, it is a fair assumption that τ_0 is a known, fixed parameter because it can be assessed through historical data. That would rarely be the case for the corresponding epidemic considerations, at least for emerging diseases. Using a single outbreak to evaluate bias, we clearly have no way of assessing variability in outcomes. We therefore include τ_0 as an unknown, random variable that is estimated in the analysis as described in the following section.

To aid the interpretation and transfer from the climate to the epidemiological interpretation, we have included Table 2 that lists the variables used in the analysis.

Stochasticity and variability

Our main interest in terms of outcome under the implemented control action is μ rather than x_0. However, it is clear that in addition to the mean duration of the outbreak, the uncertainty about the process also results in some variability in the outcomes that we need to consider. The stochastic simulations used to generate projections provide not only a mean simulated outbreak quantity, but also a range of outcomes that projects the variability. In the absence of repeated outbreaks to evaluate variability of outcomes, an obvious choice would be to use this information to inform the variability τ_0. Defining the variability τ_i as the precision of projections under the implemented action for modeling assumptions $i = 1,2,\ldots,n$ we include a hierarchical structure in the analysis so that for $i = 0,1,2,\ldots,n$

$$\tau_i \sim \text{Gamma}(a_\tau, b_\tau), \tag{8}$$

Table 2. Variables in the NHW analysis with the interpretation for the original climate interpretation and the epidemiological counterpart.

Variable	Climate Interpretation	Epidemiological Interpretation
x_i	Current temperature mean for modeling assumption *i*.	Mean log-duration under implemented control for modeling assumption *i*.
y_i	Future temperature mean for modeling assumption *i*.	Mean log-duration under alternative control for modeling assumption *i*.
x_0	Observed current temperature.	Observed outbreak log-duration.
τ_0	Precision of natural variability.	Precision of variability of individual outbreaks given the observed initial conditions. *
μ	Current mean temperature. *	Mean outbreak duration, implemented control. *
V	Future mean temperature. *	Mean outbreak duration, alternative control. *
λ_i	Precision of model *i* (weights). *	Precision of model *i* (weights). *
B	Correlation between current and. future projections. *	Correlation between implemented and alternative control projections. *
θ	Ratio between precision for current and future projections. *	Ratio between precision for implemented and alternative control projections. *
τ_i and φ_i	Not included.	Precision of within projection variability, calculated as the precision of individual replicates around the mean values x_i and y_i.

* Asterisk indicates parameters treated as random variables and no asterisk indicate data.

doi:10.1371/journal.pcbi.1004187.t002

where Gamma(a_τ, b_τ) indicates the gamma distribution with shape parameter a_τ and rate b_τ both of which are unknown parameters and are estimated in the analysis. Thus, as it would be cumbersome to elicit a fixed prior for τ_0 based on our prior expectations about variability, we instead assume that τ_0 comes from some, unknown distribution, and make use of τ_1, τ_2,. . ., τ_n to inform what this distribution should be.

Similarly, we need to model the variability of projections under the alternative control action, and denoting this φ_i we specify

$$\varphi_i \sim \text{Gamma}(a_\varphi, b_\varphi), i = 1, 2, \ldots, n \tag{9}$$

The parameters a_φ and b_φ are conditionally independent from all other parameters in the analysis and can be modelled separately in the Bayesian analysis. As x_i, y_i, τ_i and φ_i are calculated from a finite number of realization with each modeling assumption and control action, there is some uncertainty related to this. Tebaldi et al. [47] however points out that while it is certainly possible to construct a Bayesian model that takes this uncertainty into account, the effect is minimal if the number of replicates is large. With the $R = 200$ replicates preformed here, the uncertainty of the mean will in practice have very little effect, and we have included x_i, y_i, τ_i and φ_i as fixed observations.

Priors for a_τ and b_τ were specified as a gamma distribution with shapes $A_{a\tau}$ and $A_{b\tau}$, respectively, and rates $B_{a\tau}$ and $B_{b\tau}$, respectively. Similarly, the priors for a_φ and b_φ, were specified as a gamma distribution with shapes $A_{a\varphi}$ and $A_{b\varphi}$, respectively, and rates $B_{a\varphi}$ and $B_{b\varphi}$, respectively. We explored different parameter choices for the hyperpriors and found that the results were insensitive to the choice of prior for a wide range of values. In the analysis presented, we used $A_{a\tau} = A_{b\tau} = B_{a\tau} = B_{b\tau} = A_{a\phi}\ A_{b\phi} = B_{a\phi} = B_{b\phi} = 0.001$. This corresponds to prior distributions with a mean of one and a variance of 1000, thus allowing for a wide range of plausible values.

Bayesian model

Bayesian analysis requires the specification of prior parameters. We follow Tebaldi et al. [47] with priors specified as uniform on the real line for μ, ν, and β, and $\lambda_i \sim$Gamma(a_λ, b_λ) for i = 1,2,. . ., n and $\theta \sim$Gamma(a_θ, b_θ). We also need to specify hyperpriors for a_τ and b_τ, and we implement a_τ~Gamma($A_{a\tau}$, $B_{a\tau}$) and b_τ~Gamma($A_{b\tau}$, $B_{b\tau}$). Denoting $\boldsymbol{x} = x_1, x_2, \ldots, x_n$ and $\boldsymbol{y} = y_1, y_2, \ldots, y_n$, the full posterior distribution under these priors is given by

$$\begin{aligned} &P(\mu, v, \beta, \lambda, \theta, \tau_0 | x_0, x, y, \tau_1, \tau_2, \ldots) \\ &\propto \prod_{i=1}^{n} \left(\lambda_i^{a_\lambda - 1} e^{-b_\lambda \lambda_i} \lambda_i \theta^{1/2} \exp\left\{ -\frac{\lambda_i}{2} \left[(x_i - \mu)^2 + \theta(y_i - v - \beta(x_i - \mu))^2 \right] \right\} \right) \qquad (10) \\ &\theta^{a_\theta - 1} e^{-b_\theta \theta} \tau_0^{1/2} \exp\left\{ \frac{\tau_0}{2} (x_0 - \mu)^2 \right\} \prod_{i=0}^{n} (\tau_i^{a_\tau - 1} e^{-b_\tau \tau_i}) a_\tau^{A_{a\tau} - 1} e^{-B_{a\tau} a_\tau} b_\tau^{A_{b\tau} - 1} e^{-B_{b\tau} b_\tau} \end{aligned}$$

This posterior only differs from the one defined by Tebaldi et al. in that we include τ_0 as an unknown variable and use a hierarchical structure for its prior. Using Markov Chain Monte Carlo (MCMC) techniques as described in 2.9, we first performed the analysis with priors as specified by Tebaldi et al. [47] where applicable, i.e. $a_\lambda = b_\lambda = a_\theta = b_\theta = 0.001$, because they argue that this ensures that the prior contributes little to the posteriors.

However, we propose that this argument is not necessarily always valid. In particular λ_i could be expected to be sensitive to priors because it is essentially only fitted to two data points, x_i and y_i. Yet, based on approximations of conditional distributions, Tebaldi et al. argued that the gamma distribution with $a_\lambda = b_\lambda = 0.001$ is appropriately vague for the analysis. For transparency we here follow their approach and investigate the effect of the prior for the simplified model where $\beta = 0$. The mean of the conditional distribution of λ_i is then approximated by

$$E(\lambda_i | X_0, X, Y) \cong \frac{a_\lambda + 1}{b_\lambda + \frac{1}{2}[(x_i - \tilde{\mu})^2 + \theta(y_i - \tilde{v})^2]}, \qquad (11)$$

where $\tilde{\mu}$ is the conditional mean of the distribution of μ, given by

$$\tilde{\mu} = \left(\sum_{i=1}^{n} \lambda_i X_i + \tau_0 x_0 \right) \Big/ \left(\sum_{i=1}^{n} \lambda_i + \tau_0 \right) \qquad (12)$$

and $\tilde{v}$ the corresponding value for v, given by

$$\tilde{v} = \left(\sum_{i=1}^{n} \lambda_i y_i \right) \Big/ \left(\sum_{i=1}^{n} \lambda_i \right). \qquad (13)$$

We stress that $\Sigma\lambda_i$ need not sum to one, as might be intuitive when using weights. As given by Eqs 11 and 12, the mean of μ and v only depends on the relative values of λ_i, but the absolute values influence the width of the distribution, with the variance of the conditional distributions increasing with lower absolute values of λ_i (Table 3).

While a low value of a_λ certainly ensures little contribution to the numerator in Eq (11), it is less evident that a low value for b_λ contributes little to the denominator because if $x_i \to \tilde{\mu}$ and $y \to \tilde{v}$, the denominator actually approaches b_λ. Hence, to ensure that a low value of b_λ can be considered vague such that our posterior is informed primarily by x_0, $\boldsymbol{x}$ and $\boldsymbol{y}$, we must conclude that $|x_i - \tilde{\mu}|$ or $|y_i - \tilde{v}|$ is clearly separated from zero. However, if $\lambda_i \gg \lambda_j$ for all $i \neq j$ and $\lambda_i \gg \tau_0$, then $\tilde{\mu} \approx x_i$ and $\tilde{v} \approx y_i$ and nothing in the model prevents this relationship. In fact, if we consider the gamma prior with shape and scale parameters set to 0.001, the distribution has most of its density near zero, however with a fat tail (yet exponentially bounded) that allows for high values. In the current analysis, this corresponds to the prior belief that the majority of modeling assumptions will have very low precision while a few will have very high. Under this prior belief, it is expected that for some model i, $\lambda_i \gg \lambda_j$ for all $i \neq j$. In the instance where

Table 3. Conditional distributions used for updating via Gibbs sampling for the single quantity example.

Parameter	Conditional distribution
λ_i	$\text{Gamma}\left(a_\lambda + 1, \hat{b}_{\lambda i} + \frac{(x_i-\mu)^2 - \theta(y_i - v - \beta(x_i - \mu))^2}{2}\right)$ for $\hat{b}_{\lambda i} = b_\lambda$ for all modeling assumptions *i* in the NHW method,$\hat{b}_{\lambda i} = a_\lambda/m_\lambda$ for all modeling assumptions *i* in the SHW method and $\hat{b}_{\lambda i} = a_\lambda/\hat{m}_{\lambda i}$ in the IHW method.
μ	$\text{Normal}(\bar{\mu}, \sigma_\mu^2)$ with $\sigma_\mu^2 = (\sum\lambda_i + \theta\beta^2\sum\lambda_i + \tau_0)^{-1}\bar{\mu} = \sigma_\mu^2(\sum\lambda_i x_i - \theta\beta\sum\lambda_i(y_i - v - \beta x_i) + \tau_0 x_0)$
v	$\text{Normal}(\bar{v}, (\theta\sum\lambda_i)^{-1})$ with $\bar{v} = \frac{\sum\lambda_i(y_i - \beta(x_i-\mu))}{\sum\lambda_i}$
β	$\text{Normal}(\bar{\beta}, \sigma_\beta^2)$ with $\sigma_\beta^2 = (\theta\sum\lambda_i(x_i - \mu)^2)^{-1}\bar{\beta} = \sigma_\beta^2\theta^{-1}\sum\lambda_i(y_i - v)(x_i - \mu)$
θ	$\text{Gamma}\left(a_\theta + \frac{N}{2}, b_\theta + \frac{\sum\lambda_i(y_i - v - \beta(x_i-\mu))^2}{2}\right)$
τ_0	$\text{Gamma}\left(a_\tau + 1/2, b_\tau + \frac{(x_i-\mu)^2}{2}\right)$
φ_0	$\text{Gamma}(a_\phi, b_\phi)$

doi:10.1371/journal.pcbi.1004187.t003

instead $\tau_0 \gg \lambda_i$ for all *i*, then $\tilde{\mu} \approx x_0$ and the approximation would hold, but we cannot expect that relationship.

As we cannot a priori be sure that the choice of a_λ and b_λ does not influence our posterior as long as they are arbitrarily small, we performed a prior sensitivity analysis and re-ran the analysis with $a_\lambda = b_\lambda = 0.01$ and $a_\lambda = b_\lambda = 0.0001$. We could expect that the sensitivity to priors depends on the difference among modeling assumptions, and we investigate this by analyzing ensembles with little and large discrepancy between assumptions in the ensemble as given by Table 1.

We refer to this as the Non Hierarchical Weighting (NHW) method.

Standard Hierarchical Weighting model

If we cannot ensure that the analysis is insensitive to the choice of prior, it implies that our prior beliefs will influence how much different projections contribute to ensemble predictions with the current method. Using prior beliefs is sometimes desirable, and in section 2.6 we consider the situation where we trust some modeling assumptions more than others. However, it would rarely be the case that we would have substantial expectations that could inform the shape, a_λ, and scale, b_λ, of the prior for $\boldsymbol{\lambda}$.

A potential solution might be to extend the model to include hierarchical priors such that the prior for λ_i is estimated in the model rather than a priori fixed. We make a slight change to the parameterization of the prior distribution such that

$$\lambda_i \sim \text{Gamma}(a_\lambda, a_\lambda/m_\lambda), \tag{14}$$

i.e. specifying the distribution by its mean m_λ and shape a_λ, which are estimated in the model. In that way, we move our uncertainty up a level and express our beliefs about the distribution of m_λ and a_λ, rather than $\boldsymbol{\lambda}$. Using m_λ rather than b_λ facilitates the specification of a prior for the mean precision parameter that corresponds to the priors previously specified on individual λ_i. This parameterization further aids the use of prior beliefs about weights in section 2.6.

While we can never be strictly uninformative in Bayesian analysis, the hierarchical prior can allow for a wide range of plausible m_λ and a_λ whereas the model presented in section 2.4 requires these to be specified explicitly. This also allows for the concept of "borrowing strength" [54], such that the distribution of each λ_i can be indirectly informed by all other precisions via the hierarchical distribution. This is often beneficial in situations where individual parameters are fitted to a small amount of data [55,56], which clearly is the case for λ_i here. To extend Eq (10) to a hierarchical model, we include hyperpriors such that

$$a_\lambda \sim \text{Gamma}(A_{a\lambda}, B_{a\lambda}) \tag{15}$$

and

$$m_\lambda \sim \text{Gamma}(A_{m\lambda}, B_{m\lambda}). \tag{16}$$

We performed the corresponding sensitivity analysis for the hierarchical ensemble prediction by applying hyperpriors $A_{a\lambda} = B_{a\lambda} = A_{m\lambda} = B_{m\lambda}$ set to 0.01, 0.001 and 0.0001. We refer to this as hierarchical sensitivity set-up one. Secondly, we performed a sensitivity analysis, hierarchical sensitivity set-up two, where we fixed the shape parameters $A_{a\lambda} = A_{m\lambda} =$ 0.001 and only varied $\text{B}_{a\lambda} = \text{B}_{m\lambda}$, again set to either 0.01, 0.001 or 0.0001.

We refer to this as model as the Standard Hierarchical Weighting (SHW) method.

Informative Hierarchical Weighting model

Using expert opinions may substantially improve predictions [57], and there are several instances where incorporating prior beliefs that reflect the "trust" in different modeling assumptions could be useful. For instance, a policymaker might have more trust in one model type over another, and rather than excluding the models that are considered less reliable (i.e. giving them a priori zero weigh), it could be useful to include them, yet with less contribution to the ensemble.

In the case considered here, where modeling assumptions represent most likely, best and worst case in terms of parameterization, we might want to give the "most likely" modeling assumption higher weight. For the analysis with fixed a_λ and b_λ, described in section 2.4, we could merely elicit a different scale parameter b_λ for each λ_i, such that modeling assumptions with high trust are given a low value. However, with the shape parameter a_λ set to a low value ("vague" shape), the prior may have little effect on the posterior λ_i. Eliciting a high value of a_λ would instead result in a posterior that is merely the results of our prior beliefs and we have no foundation for which to elicit some intermediate value.

In order to combine the hierarchical approach with informative priors, we propose a modification of the analysis presented in section 2.5, where the assumption of exchangeability is relaxed in the hierarchical structure with

$$\lambda_i \sim \text{Gamma}(a_\lambda, a_\lambda / \hat{m}_{\lambda i}), \tag{17}$$

where $\hat{m}_{\lambda i} = w_i m_\lambda$ and w_i indicates the a priori trust in modeling assumption i. With $w_i = kw_j$, the prior distribution of λ_i will have a mean that is k times that of λ_j and from Eqs (12) and (13) the relationship also implies that before $\boldsymbol{\lambda}$ is estimated (i.e. involving the data x_0, $\boldsymbol{x}$ and $\boldsymbol{y}$), the outputs of modeling assumption i will contribute k times as much to μ and ν as does assumption j.

To demonstrate the effect that a priori trust in different modeling assumptions can have on the posterior estimates, we consider the case where the best, most likely and worst case scenarios for each of the two varied parameters corresponds to percentile 2.5, 50 and 97.5, respectively, of a normal distribution. Given that the density at percentiles 2.5 and 97.5 then is 0.15 of that at the mode, we specify $w_i = 0.15$ for $i = 2, 3, 4$ and 7, i.e. for modeling assumptions where one of the varied parameters follows the most likely scenario, whereas the other one is set to either worst or best case. With the same rationality, we specify $w_i = 0.021$ for $i = 5, 6, 8$ and 9, i.e. modeling assumptions where both parameters follow either best or worst case expectations.

We also investigate the case where a high weight is given to a projection x_i further away from the observed data x_0. Consistently, modeling assumptions $i = 5$ predicted the shortest duration for all actions and ensembles. We therefore also performed the analysis with $w_5 = 1$ and $w_1 = 0.021$, and all other weights are as above. This allows us to investigate the performance of the informative weighting method when an outlier is up-weighted.

We refer to this method as the Informative Hierarchical Weighting (IHW) method.

Multiple epidemic quantities

In the above examples, we focused on a single epidemic quantity, allowing for transparent transition from the original Tebaldi et al. work [47] that focused on temperature. For epidemiology, it may however be useful to consider multiple epidemic quantities. This could be done in different ways, but here we offer a straightforward multi-quantity extension of the Bayesian model for the single epidemic quantity, based on the supposition that the relative weights are equal for all quantities. As such, we implement a single weighting parameter λ, shared among all quantities. For other parameters, we use a similar notation as for the single quantity analysis, but give many of the parameters an additional index q, indicating that the parameter is quantity specific. We expand the Bayesian model by defining

$$x_{i,q} \sim \text{Normal}(\mu_q, \theta_{\mu,q}\lambda_i^{-1}), \tag{18}$$

$$y_{i,q} \sim \text{Normal}(v_q + \beta_q(x_{i,q} - \mu_q), (\theta_{v,q}\lambda_i)^{-1}), \tag{19}$$

$$x_{0,q} \sim \text{Normal}(\mu_q, \tau_{0,q}^{-1}), \tag{20}$$

where $x_{i,\ q}$ and $y_{i,\ q}$ are the mean projections of modeling assumption i for epidemic quantity q for the implemented and alternative control action, respectively, and $x_{0,\ q}$ is the corresponding observed value. As for the single epidemic quantity example, $\mu_q$and $V_q$are the expected values of quantity q, and because we cannot expect to have the same correlation between control actions for all quantities, β_q is included as unique for each q. Parameters $\theta_{\mu,\ q}$ and $\theta_{v,\ q}$ scales the precision of models between actions and quantities and the parameters of the Bayesian model are identifiable by defining $\theta_{\mu,\ 1} = 1$. Similarly, we specify quantity specific parameters

$$\begin{aligned} \tau_{i,q} &\sim \text{Gamma}(a_{\tau,q}, b_{\tau,q}), i = 0, 1, 2, \ldots, n, \\ \varphi_{i,q} &\sim \text{Gamma}(a_{\varphi,q}, b_{\varphi,q}), i = 1, 2, \ldots, n. \end{aligned} \tag{21}$$

The conditional distributions for the multi-quantity extension are provided in Table 4. We denote the total number of quantities in an analysis as Q.

Analysis of simulated outbreaks

The above examples focus on the UK 2001 FMD outbreak and show how the introduced framework can be applied to actual outbreak data. However, a limitation to this approach is that we are confined to investigating the behavior of the ensemble methodology for that particular outbreak. To further investigate the potential and limitations of the proposed methods, we also performed analysis of simulated outbreak data. With simulated data, we have "true" estimates of μ and v, and we want to explore the ability of the ensemble to predict these under two different conditions; when the true values lies within the range of **X** and **Y** predicted by the

Table 4. Conditional distributions used for updating via Gibbs sampling for *Q* epidemic quantities.

Parameter	Conditional distribution
λ_i	Gamma$\left(a_{\lambda}+Q,\hat{b}_{\lambda i}+\sum_{q=1}^{Q}\frac{\theta_{\mu,q}(x_{i,q}-\mu_q)^2-\theta_{v,q}(y_{i,q}-v_q-\beta_q(x_{i,q}-\mu_q))^2}{2}\right)$ for $\hat{b}_{\lambda i}=b_{\lambda}$ for all modeling assumptions *i* in the NHW method,$\hat{b}_{\lambda i}=a_{\lambda}/m_{\lambda}$ for all modeling assumptions *i* in the SHW method and $\hat{b}_{\lambda i}=a_{\lambda}/\hat{m}_{\lambda i}$ in the IHW method.
μ_q	Normal$(\bar{\mu},\sigma_{\mu}^2)$ with $\sigma_{\mu}^2=(\sum\theta_{\mu,q}\lambda_i+\theta_{v,q}\beta_q^2\sum\lambda_i+\tau_{0,q})^{-1}\bar{\mu}=\sigma_{\mu}^2(\sum\theta_{\mu,q}\lambda_i x_{i,q}-\theta_{v,q}\beta_q\sum\lambda_i(y_{i,q}-v_q-\beta_q x_{i,q})+\tau_{0,q}x_{0,q})$
v_q	Normal$(\bar{v},(\theta_{v,q}\sum\lambda_i)^{-1})$ with $\bar{v}=\frac{\sum\lambda_i(y_{i,q}-\beta_q(x_{i,q}-\mu_q))}{\sum\lambda_i}$
β_q	Normal$(\bar{\beta},\sigma_{\beta}^2)$ with $\sigma_{\beta}^2=(\theta_{v,q}\sum\lambda_i(x_{i,q}-\mu_q)^2)^{-1}\bar{\beta}=\sigma_{\beta}^2\theta_{v,q}^{-1}\sum\lambda_i(y_{i,q}-v_q)(x_{i,q}-\mu_q)$
$\theta_{\mu,\,q}$ for $q \neq 1$	Gamma$\left(a_{\theta}+\frac{N}{2},b_{\theta}+\frac{\sum\lambda_i(x_{i,q}-\mu_q)^2}{2}\right)$
$\theta_{v,\,q}$	Gamma$\left(a_{\theta}+\frac{N}{2},b_{\theta}+\frac{\sum\lambda_i(y_{i,q}-v_q-\beta_q(x_{i,q}-\mu_q))^2}{2}\right)$
$\tau_{0,\,q}$	Gamma$\left(a_{\tau,q}+1/2,b_{\tau,q}+\frac{(x_{i,q}-\mu_q)^2}{2}\right)$
$\varphi_{0,\,q}$	Gamma$(a_{\phi,\,q}, b_{\phi,\,q})$

doi:10.1371/journal.pcbi.1004187.t004

individual models of the ensemble and when it does not. For multi-model ensembles, this corresponds to the situation where the true behavior of the outbreak is encapsulated within the range of underlying assumptions of the individual models and when it is not.

Here we explore the outcome of these conditions by first simulating outbreaks with the parameterizations of modeling assumption 1 ($k_1 = k_2 = 1$), i.e. located in the center of both the small and large discrepancy ensemble. This simulates outbreaks where the true behavior of the outbreak is encapsulated within the range of underlying assumptions of the individual projections for both ensembles. We also simulate outbreaks with a parameterization where both k_1 and k_2 are set to 0.9. This produces outbreaks where the true behavior is outside of the assumptions of the projections for the small discrepancy ensemble, yet inside the range of the large discrepancy ensemble.

The exact behavior of the ensemble depends on the actual realization of the individual outbreak, because the observed values x_0 are different due to the stochastic disease transmission process. We therefore apply both the small and large discrepancy ensembles to ten realizations of each of the simulation parameterizations. We implement both the single and multiple epidemic quantity analysis, thus further highlighting the effect of using multiple quantities.

Computation

We use Markov Chain Monte Carlo (MCMC) techniques to obtain samples from the full posterior distribution of the proposed Bayesian models (NHW, SHW and IHW). For many parameters, the conditional distribution belongs to a standard parametric family, thus allowing for Gibbs sampling. We list these conditional distributions in Table 3 for single quantity analysis and Table 4 for multiple quantities.

We also rely on Metropolis-Hastings (M-H) updates, and with the computation used for multi-quantity analysis being a straightforward extension of that used for the single quantity, we start by describing the update scheme for the single quantity analysis. The conditional distribution of b_{τ} has a known form, $P(b_{\tau}|\ldots)=\text{Gamma}(A_{b\tau}+(N+1)a_{\tau},B_{b\tau}+\sum_{i=0}^{N}\tau_i)$, that would allow for Gibbs sampling of b_τ, whereas M-H updates need to be implemented for a_τ. We however found strong correlation between the marginal posterior estimates of a_τ and b_τ, and mixing was improved by performing joint M-H updates of these parameters by multivariate Random Walk (RW) proposals. Mixing can be further improved by updating parameters on a transform that resembles a Gaussian distribution, and we therefore performed updates on the log-transform, i.e. based on current values of a_τ and b_τ We proposed candidate parameters $[\log(a_\tau^*), \log(b_\tau^*)]$ from MVN([log(a_τ),log(b_τ)],Σ_τ). Here MVN indicates the multivariate normal distribution and Σ_τ the covariance matrix. Candidate values are accepted with the probability

$$\min\left(1, \frac{\text{Gamma}(a_\tau^*|A_{a\tau}, B_{a\tau})\text{Gamma}(b_\tau^*|A_{b\tau}, B_{b\tau}) \prod_{i=0}^{N} \text{Gamma}(\tau_i|a_\tau^*, b_\tau^*)}{\text{Gamma}(a_\tau|A_{a\tau}, B_{a\tau})\text{Gamma}(b_\tau|A_{b\tau}, B_{b\tau}) \prod_{i=0}^{N} \text{Gamma}(\tau_i|a_\tau, b_\tau)} |J_\tau|\right), \tag{22}$$

where $|J_\tau| = a_\tau^* b_\tau^* (a_\tau b_\tau)^{-1}$ indicates the Jacobian determinant of the log-transform.

Mixing can be improved if the covariance matrix Σ_τ is proportional to the covariance of the marginal posterior of [log(a_τ),log(b_τ)], here indicated as Φ. However, this is not known prior to the analysis. We therefore implement an optimized method of the Robbins-Monroe search process as presented by Garthwaite et al. [58]. This estimates the covariance during the MCMC and finds the scaling parameter ρ such that $\Sigma_\tau = \rho\Phi$ provides a chosen long term acceptance rate, here set to 0.234 based on suggestions by Roberts et al. [59]. The method has been demonstrated to be appropriate also for RW on transformed scales of the parameters [60].

The corresponding updates of a_φ and b_φ were also performed with M-H updates and we proposed candidate parameters $[\log(a_\varphi^*), \log(b_\varphi^*)]$ from MVN([log(a_ϕ),log(b_ϕ)],Σ_ϕ). and accepted them with probability

$$\min\left(1, \frac{\text{Gamma}(a_\varphi^*|A_{a\varphi}, B_{a\varphi})\text{Gamma}(b_\varphi^*|A_{b\varphi}, B_{b\varphi}) \prod_{i=1}^{N} \text{Gamma}(\varphi_i|a_\varphi^*, b_\varphi^*)}{\text{Gamma}(a_\varphi|A_{a\varphi}, B_{a\varphi})\text{Gamma}(b_\varphi|A_{b\varphi}, B_{b\varphi}) \prod_{i=1}^{N} \text{Gamma}(\varphi_i|a_\varphi, b_\varphi)} |J_\varphi|\right). \tag{23}$$

We used a similar approach for updates of a_λ and m_λ in the hierarchical methods (SHW and IHW) and proposed $[\log(a_\lambda^*), \log(m_\lambda^*)]$ from MVN([log(a_λ),log(b_λ)],Σ_λ). Candidate parameters were accepted with probability

$$\min\left(\frac{1, \text{Gamma}(a_\lambda^*|A_{a\lambda}, B_{a\lambda})\text{Gamma}(m_\lambda^*|A_{b\lambda}, B_{b\lambda}) \prod_{i=1}^{N} \text{Gamma}(\lambda_i|a_\lambda^*, \hat{b}_{\lambda i}^*)}{\text{Gamma}(a_\lambda|A_{a\lambda}, B_{a\lambda})\text{Gamma}(m_\lambda|A_{b\lambda}, B_{b\lambda}) \prod_{i=1}^{N} \text{Gamma}(\lambda_i|a_\lambda, \hat{b}_{\lambda i})|J_\lambda|}\right), \tag{24}$$

where $\hat{b}_{\lambda i} = a_\lambda / m_\lambda$ for all modeling assumptions i in the SHW method and $\hat{b}_{\lambda i} = a_\lambda / \hat{m}_{\lambda i}$ with $\hat{m}_{\lambda i} = w_i m_\lambda$ in the IHW method. As above, we used the method of Garthwaite et al. [58] to determine Σ_λ to obtain a long term acceptance rate of 0.234.

We also found strong correlation between μ and ν. In order to improve mixing, we repeated the updates of these parameters five times for each iteration of the MCMC.

The same update scheme was used for the multi-quantity consideration, yet with a separate $\Sigma_{\tau,\,q}$, $\Sigma_{\phi,\,q}$ and $\Sigma_{\lambda,\,q}$ adaptively estimated for each quantity q.

The algorithm was implemented in MATLAB (The MathWorks, Inc., Natick, Massachusetts, United States) and code is available as supplementary information (S1 File).

Results

Single quantity analysis

We start by presenting the results for the single quantity analysis, highlighting the behavior of the method for the NHW, SHW and IHW schemes. Fig 2, panels A and B show the estimates of outbreak duration for the two control actions for the large discrepancy ensembles using the NHW method and reveals rather large prior sensitivity. Note that we plot marginal posteriors of $M = e^{\mu}$ and $N = e^{\nu}$, respectively. As such, the posteriors represent the geometric mean outbreak duration. The corresponding arithmetic mean can be calculated as $e^{\mu+1/(2\tau_0)}$ and $e^{\nu+1/(2\phi_0)}$, respectively, yet we use the geometric mean as it more clearly shows the relationship with individual projections, here presented by $x_i = e^{x_i}$ and $Y_i = e^{y_i}$, respectively. For $a_\lambda = b_\lambda = 0.0001$ (solid gray lines), the distributions are multimodal with peaks at individual model predictions, whereas a more smooth shape is obtained for $a_\lambda = b_\lambda = 0.01$ (dashed black lines) and $a_\lambda = b_\lambda = 0.001$ (solid black line) yields an intermediate result.

With the SHW method, we instead obtain posteriors that are largely insensitive to the choice of hyperprior. Fig 2, panels E and F present the result of sensitivity set-up one, showing near identical posterior estimates when hyperparameters $A_{a\lambda}$, $B_{a\lambda}$, $A_{m\lambda}$ and $B_{m\lambda}$ are set to 0.01,

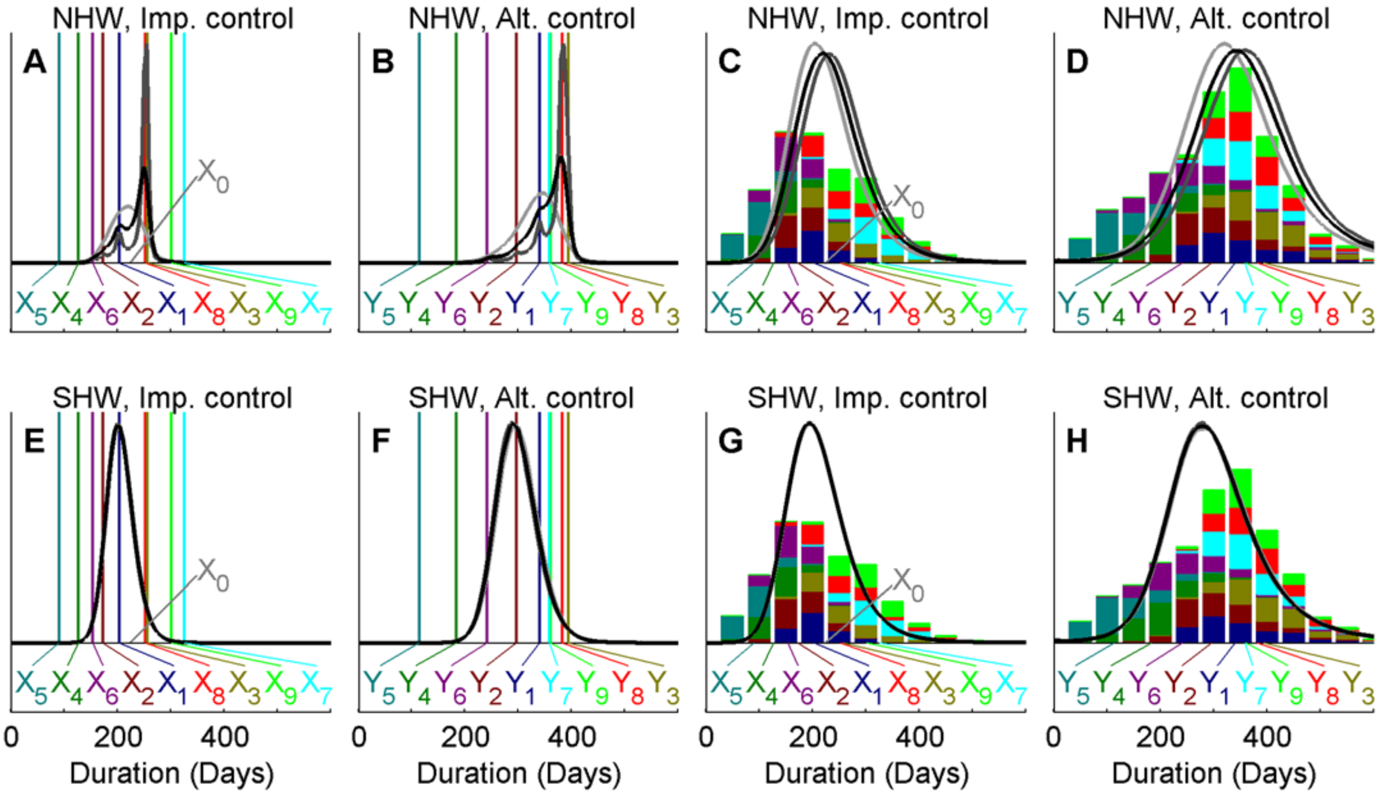


Fig 2. Ensemble prediction for the UK 2001 FMD outbreak duration. Comparing two methods of ensemble prediction of expected outbreak duration under implemented and alternative control actions (panels A, E and B, F, respectively) as well as the corresponding duration of individual outbreaks (panels C, G and D, H, respectively). The observed duration of the UK 2001 FMD outbreak is indicated by *X0* while vertical colored lines (panels A, B, E and F) and annotations *X1*, *X2*,.., *X9* and *Y1*, *Y2*,. . ., *Y9* indicate the mean outbreak duration of each projection. The total bar height (panels C, D, G, H) indicate the frequency of simulations across all projections with outbreaks ending within each 50 day interval and colors indicate the contribution of each of the projections. Results are shown for analyses of the large discrepancy ensemble using the NHW method (panels A-D) and the SHW method (panels E-H). Marginal posterior estimates correspond to different priors, with prior parameters *aλ* = *bλ* = 0.01 (light gray), *aλ* = *bλ* = 0.001 (black) and *aλ* = *bλ* = 0.0001 (dark gray) the NHW method (panels A-D) and *Aaλ* = *Abλ* = *Baλ* = *Bbλ* set to 0.01 (light gray), 0.001 (black) and 0.0001 (dark gray) for the SHW method (panels E-H). Note that posterior estimates are very similar for the latter, making the lines overlap.

doi:10.1371/journal.pcbi.1004187.g002

0.001 or 0.0001. Sensitivity set-up two produced results that were visually indistinguishable from panels E and F and are not presented. This further indicates that the hierarchical method is robust to the choice of hyperpriors.

Within epidemiology, there is clearly an interest in not just the expected outbreak duration, but also other statistics such as the probability of large outbreaks occurring. We therefore consider the posterior predictive distributions of individual outbreak durations under the two control actions. For the non-hierarchical model (Fig 2, panels C and D), there is an obvious effect of the choice of prior with higher probability of long outbreaks for lower values of a_λ and b_λ. For the hierarchical model (Fig 2, panels G and H), there is again little difference among posteriors corresponding to different priors.

When evaluating the efficiency of control actions, the difference *N-M* is of particular interest. In the example presented here, this estimates how much longer the outbreak would have been if culling of CPs had been excluded from the control. As shown in Fig 3, the estimates are again sensitive to the choice of prior with the NHW method, yet insensitive with the SHW method. The range of the posterior under the NHW method is less sensitive to the prior for the low discrepancy ensemble (panel B) than for the large discrepancy ensemble (panel A), where higher probability of less difference is estimated with $a_\lambda = b_\lambda = 0.01$ than for $a_\lambda = b_\lambda = 0.0001$. However, the multimodal behavior of the NHW method with low values of a_λ and b_λ is obtained also for the low discrepancy ensemble.

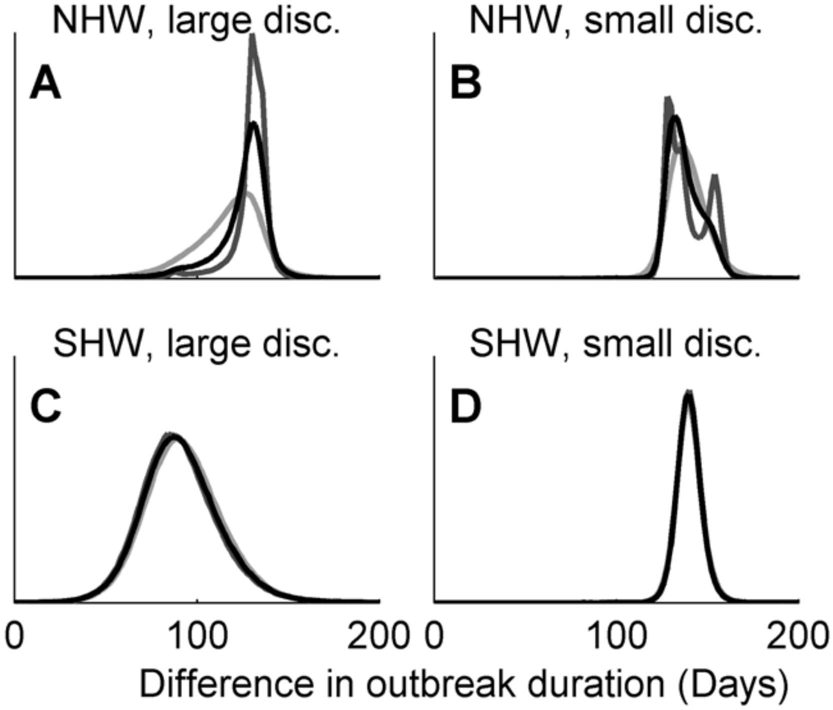


Fig 3. Ensemble predicted difference between control actions. Posterior predictive distributions of the difference in outbreak duration between implemented and alternative control actions for the 2001 UK FMD outbreak using the NHW method (panels A, B) and the SHW method (panels C, D,). Results are shown for large (panels A, C) and small (panels B, D) discrepancy among projections in the ensemble. Marginal posterior estimates correspond to different priors, with prior parameters $a\lambda = b\lambda = 0.01$ (light gray), $a\lambda = b\lambda = 0.001$ (black) and $a\lambda = b\lambda = 0.0001$ (dark gray) for the NHW method (panels A, B) and $Aa\lambda = Ab\lambda = Ba\lambda = Bb\lambda$ set to 0.01 (light gray), 0.001 (black) and 0.0001 (dark gray) for the SHW method (panels C, D). Due to high similarity of estimates, the plots are largely overlapping for the hierarchical method.

doi:10.1371/journal.pcbi.1004187.g003

Fig 4 demonstrates the effect that a priori beliefs about the weights have on the predicted outbreak duration under large and small discrepancy ensembles. When using a priori higher weights for the most likely scenarios (modeling assumption one; black dotted lines), the posterior estimates are shifted and become more centered on projections of that particular modeling assumption compared to the case where a priori weights are equal (black solid line). The outcome of up-weighting the outlier (modeling assumption five; solid gray lines) is however different between the two ensembles. For the small discrepancy ensemble (panels A and B), similar results are found as for the up-weighting of the most likely scenarios; posteriors are shifted towards the projection with a priori high weight. For the high discrepancy ensemble (panels C and D), the posterior estimates of outbreak duration instead become wider for both control actions, indicating larger uncertainty about the expected duration of outbreaks.

Fig 5 shows the marginal posterior estimates of individual weights λ_i under different discrepancy among projections and informative weighting schemes. When using a priori equal weights, there is little difference in the estimates for the small discrepancy ensemble (top left panel) whereas moderate differences are obtained for large discrepancy (bottom left panel). Note that while the error bars are overlapping, the mean estimate of the most likely scenarios

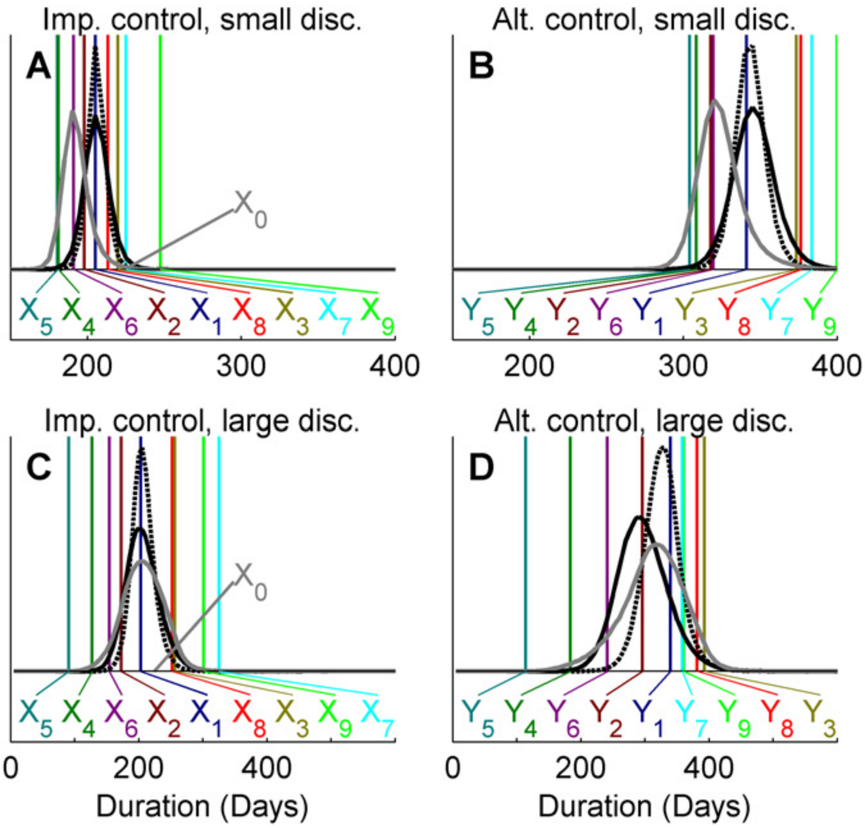


Fig 4. Ensemble prediction of expected outbreak duration with the IHW method. Panels A and C show ensemble estimates of mean duration for the implemented control action of the 2001 UK FMD outbreak under small and large discrepancy ensemble, respectively. Panels B and D shows the corresponding estimates for an alternative control action. Marginal posterior distributions indicate the expected outbreak duration estimated by the ensemble with a priori equal weights (solid black line), up-weighting of the most likely scenarios ($i = 1$; dashed black line), up-weighting of the outlier ($i = 5$; solid gray line). Observed outbreak duration is indicated by *X0* and predicted means under the implemented and alternative control action are indicated by *X1*, *X2*,. . . *X9* and *Y1*, *Y2*,. . ., *Y9*, respectively.

doi:10.1371/journal.pcbi.1004187.g004

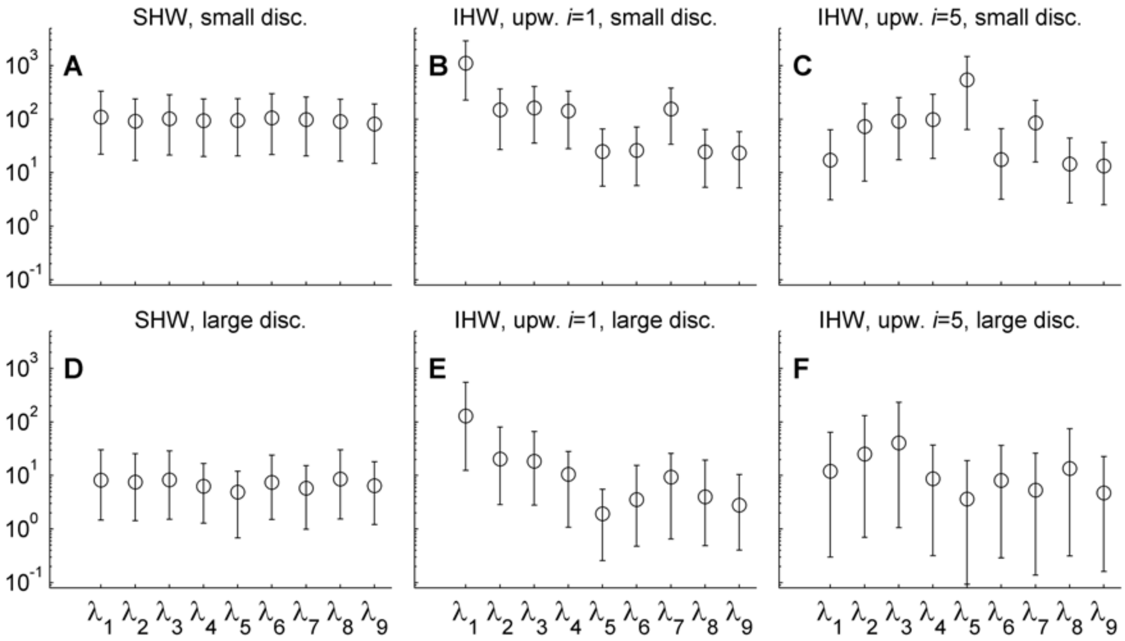


Fig 5. Marginal posterior estimates of weighting parameters, λ. Posterior means are indicated with circles and error bars indicate 95% central credibility intervals under a priori equal weights (left column panels), up-weighting of modeling assumption $i = 1$(middle column panels) and up-weighting of assumption $i = 5$ (right column panels). Results are shown for small and large discrepancy ensembles in top and bottom row panels, respectively.

doi:10.1371/journal.pcbi.1004187.g005

(modeling assumption one) is approximately 1.7 times as large as that of the outlier(modeling assumption five), meaning that the former will contribute approximately 1.7 times as much to the posterior means of μ and ν than the latter (Eqs (12) and (13)).

When giving a priori highest weight to the most likely scenario (modeling assumption one; middle column panels), the posterior estimate of λ_1 is consistently shifted upwards, meaning that the most likely scenarios (modeling assumption one) will contribute more to the posteriors of μ and ν than other projections. For the up-weighting of the outlier, projections corresponding to modeling assumption five, the same is found when there is little discrepancy among projections (lower right panel). This is however not found for the high discrepancy ensemble (top right panel), where the main effect is that compared to equal a priori weights (top left panel), the error bars are wider; this indicates larger uncertainty about weights and consequently about the contribution of individual projections to the posterior estimates of outbreak durations.

Multiple quantity analysis and simulated outbreaks

The proposed multi-quantity method can be implemented with either NHW, SHW or IHW schemes. Here we aim to illustrate the effect of using multiple quantities and focus on the SHW scheme. Fig 6 plots the marginal posterior density of mean outbreak duration under the two control actions as estimated for the multiple quantity analysis (solid) together with the corresponding estimates for the single quantity analysis (dashed). The figure illustrates how inclusion of multiple quantities in the analysis leads to tighter distributions, centered on projections for $i = 1$. The multi-quantity analysis produces a probability distribution of all considered quantities, and Fig 7 further illustrates how the marginal posterior densities are located above zero for all three considered quantities.

To illustrate the performance of the method under different conditions, we also analyzed simulated outbreaks. Fig 8 shows the posteriors of mean duration for outbreaks simulated with

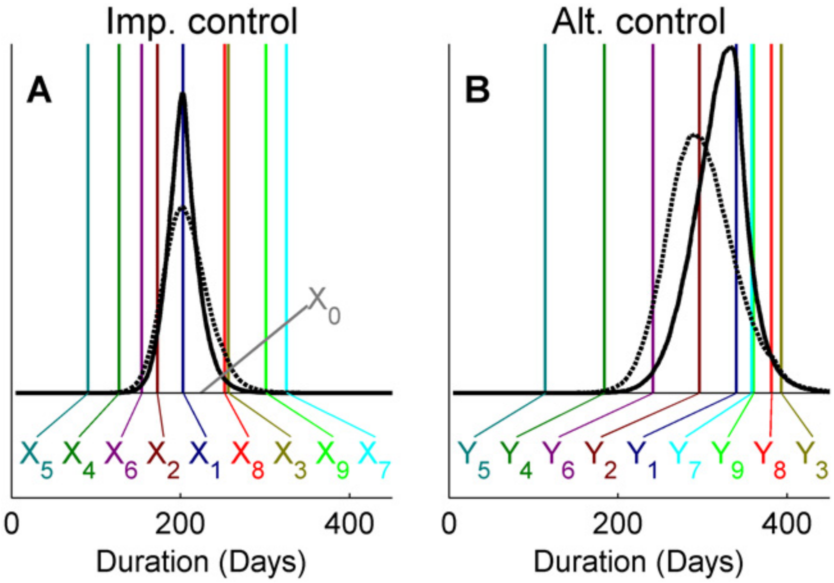


Fig 6. Ensemble prediction of outbreak duration with *Q* = 1 and *Q* = 3. Posterior estimates of mean outbreak duration for the implemented (Panel A) and alternative (Panel B) control action, with the dashed line indicating the posteriors corresponding to the single quantity analysis and the solid line indicating the corresponding posterior when number of infected and control culls are included. The figure shows the result of the large discrepancy ensemble.

doi:10.1371/journal.pcbi.1004187.g006

the $k_1 = k_2 = 1$ parameterization and applying the small (triangles) and large (circles) discrepancy ensembles, represented by the median values and error bars indicating the 95% central credibility interval. Note that individual realizations, indicated by stars, are expected to frequently be outside of the credibility envelopes. Error bars are inclusive of the true mean outbreak duration (dashed lines) for all ten analyzed realizations for both the implemented and alternative control actions. However, the credibility envelopes are tighter and medians closer to the true value for the multi-quantity analysis. This indicates that the ensemble prediction is improved by including multiple quantities.

When applying the analysis to outbreaks simulated with the $k_1 = k_2 = 0.9$ parameterization (Fig 9), the large discrepancy ensemble error bars are still consistently inclusive of the true value. As with Fig 8, credibility envelopes are tighter for the multi-quantity analysis. The error bars of the small discrepancy ensemble that all rely on simulations with parameterizations with higher k_1 and k_2 than the true value, are not inclusive of the true value, indicating that the small discrepancy ensemble fails in predicting the true values of the outbreak.

Discussion

Ensemble modeling is appealing because it offers the possibility to combine multiple projections. Within weather forecasting, the approach has given more robust predictions, and we could expect that to be the case for epidemiology as well. However, there is a need for the development of methods describing how to combine several epidemiological projections. The aim of this study was to investigate the possibility of using the Bayesian framework introduced by Tebaldi et al. [47]. We find that it is a promising approach, for primarily three reasons.

Firstly, when the methodology is implemented in a hierarchical Bayesian framework, it provides an appealing interpretation of model exchangeability. Essentially, projections and their underlying modeling assumptions are treated as random draws from a population of possible

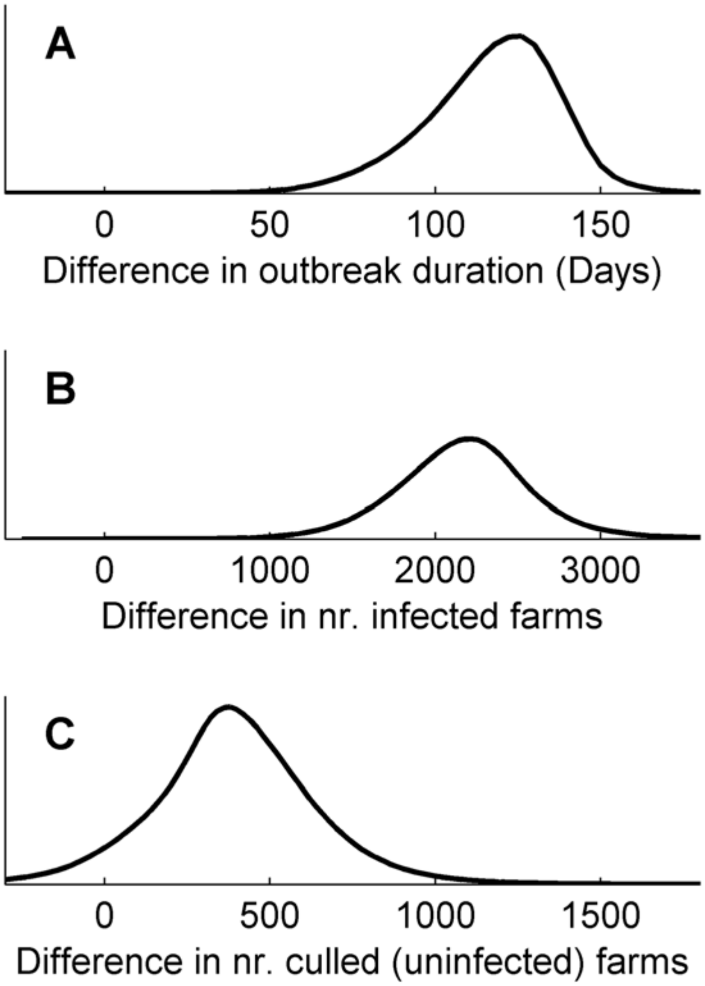


Fig 7. Posterior estimates of difference between controls. The figure shows the marginal posterior predictive estimates of difference in outbreak duration (Panel A), number of infected farms (Panel B) and number of control culls (Panel C) between the implemented and alternative control action. The figure shows the result of the large discrepancy ensemble.

doi:10.1371/journal.pcbi.1004187.g007

projections. By estimating the hierarchical parameters a_λ and m_λ jointly with individual precisions (weights) λ_i, the characteristics of this hypothetical population are estimated. Smith et al. [61] used a similar approach for climate ensembles and pointed out that this reduces the impact of which models are included or excluded in the ensemble. That is, we should expect to get similar results when using a different set of model assumptions if they are chosen independently. We stress that this interpretation is more valid for multi-model ensembles, however. Also, the term "random draws" should not be interpreted as arbitrary. Rather, the interpretation is that models should come from a population of well-informed, reasonable models. The analysis treats the outputs of the performed simulations under different assumptions as data (Eq 1, Table 1), and as such they are used to inform the quantities of interest (μ and ν). This may seem counterintuitive, yet it only serves as a formal means to combine the results of multiple projections, and by Eqs (7) and (20), these are combined with available outbreak data.

Secondly, the framework can handle several different weighting schemes simultaneously. The original methods introduced by Tebaldi et al. [47] used convergence and bias to assess weights. Here, we further extend the framework such that informative priors can be included

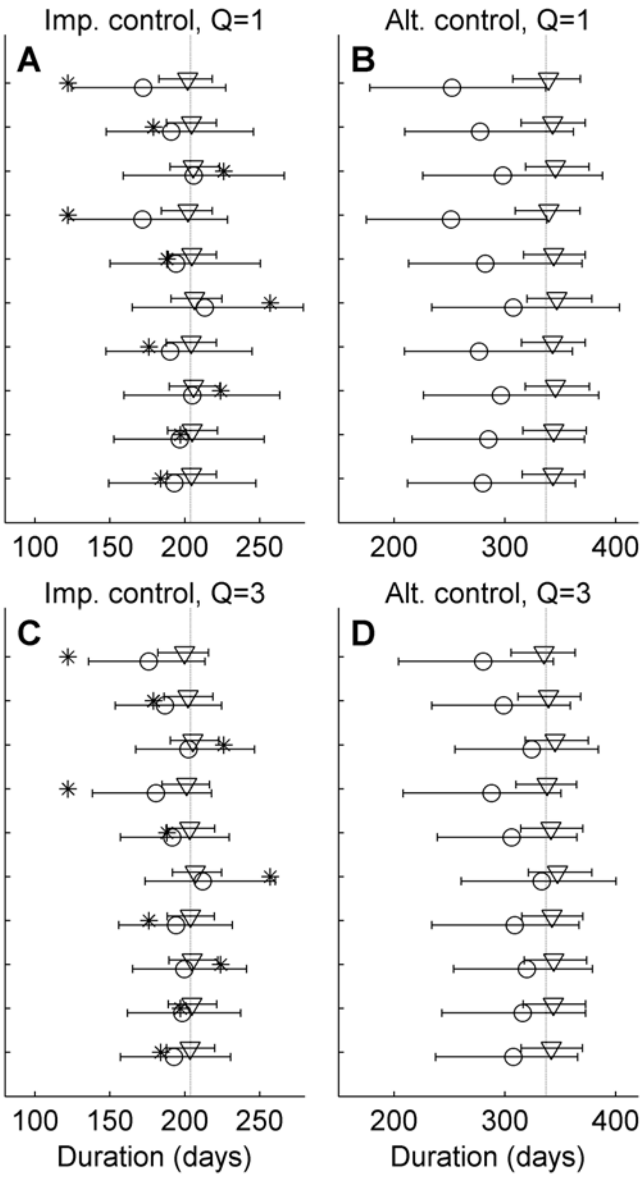


Fig 8. Analysis of synthetic data. Triangles and circles indicate the median posterior estimates of mean outbreak duration after analysis with the small and large discrepancy ensemble, respectively, under the implemented (Panels A and C) and alternative (Panels B and D) control action. Results are shown for ten realizations of synthetic data simulated with *k1* = *k2* = 1. The error bars indicate the 95% credibility interval, the dashed lines the true values and the star the individual realization (expected to frequently lie outside the predicted mean). Panels A and B show the results of single quantity analysis (only outbreak duration) and Panels C and D the corresponding results of analyses where number of infected farms and control culls are included.

doi:10.1371/journal.pcbi.1004187.g008

to inform the weights, thus relaxing the supposition that all modeling assumptions are a priori exchangeable. Epidemiological predictions suffer from lack of available data to assess model bias, and we propose that expert opinions will play a larger role than in other fields of research. With the analytical tool proposed here, a policymaker can choose to include a range of projections based on different modeling assumptions, yet give them different weights, rather than including one or a few (given a weight of one) and excluding others (given a weight of zero).

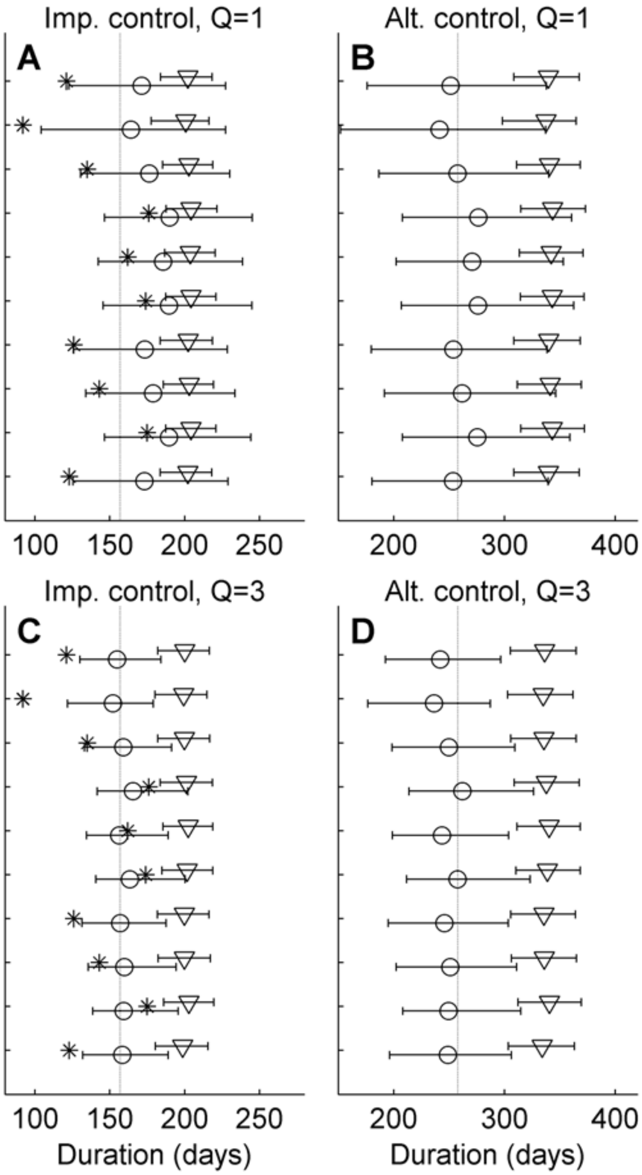


Fig 9. Analysis of synthetic data. Triangles and circles indicate the median posterior estimates of mean outbreak duration after analysis with the small and large discrepancy ensemble, respectively, under the implemented (Panels A and C) and alternative (Panels B and D) control action. Results are shown for ten realizations of synthetic data simulated with *k1* = *k2* = 0.9. The error bars indicate the 95% credibility interval, the dashed lines the true values and the star the individual realization analyzed (expected to often lie outside the predicted mean). Panels A and B show the results of single quantity analysis (only outbreak duration) and Panels C and D the corresponding results of analyses where number of infected farms and control culls are included.

doi:10.1371/journal.pcbi.1004187.g009

When using different mechanistic models, subjective trust in the different models can be incorporated by using methods of prior elicitation based on expert opinion [38]. Importantly, our methods can incorporate these subjective beliefs in the hierarchical framework, requiring only the specification of the a priori relative confidence in the underlying assumptions of the projections. Definition of an individual, fixed prior would undoubtedly be cumbersome to elicit from expert opinion; it would not be feasible to ask policymakers to define an individual gamma prior for each modeling assumption.

Here we used ensembles based on projections of the same model with different parameterizations, demonstrating the possibility to explore parameter space, yet with unequal probabilities of different parameterizations. Uncertainty about parameters will be an issue for most epidemiological models, and we propose that multi-model ensembles should incorporate projections with different models and different parameterizations. Thus, different mechanistic assumptions as well as parameter uncertainty would be incorporated in the ensemble.

Thirdly, the framework produces easily interpretable probability distributions. It is important that communication with policymakers include uncertainties about prediction rather than just the most likely outcome [16]. In the ensemble context, these uncertainties take into account different assumptions about the transmission process. Gårdmark et al. [32] suggested that uncertainty should be communicated with policymakers by presenting the full range of predicted outcomes. However, that would give equal weights to all included projections and would require that the results be communicated with a detailed description of all assumptions made, thus allowing the policymaker to decide how much to trust each modeling assumption. This would be a cumbersome task, particularly for detailed simulation models that rely on a large number of parameters. We therefore argue that it is beneficial to communicate the aggregated and weighted result as easily interpretable probability distributions. With further modifications of the methodology, we propose that the approach could also be used as a forecasting tool during an outbreak, e.g. by letting x_i and y_i denote current and future numbers of infected farms. In such a situation, there is a great need for rapid and clear communication of model results to aid policy decisions. The visual manner in which uncertainty is presented using probability distributions makes them easy to understand and communicate [62].

We here show that these distributions are sensitive to the choice of priors when using the NHW method (Fig 2, panels A-D, Fig 3, panels A, B). However, the impact of the prior is heavily reduced when using the hierarchical framework (Fig 2, panels E-H, Fig 3, panels C, D). Thus, our results demonstrate that the hierarchical approach is preferred for ensemble modeling and using the non-hierarchical approach can lead to spurious conclusions. We argue that this would also be the case for other fields, such as climate ensembles, but it is likely to be a larger concern for epidemiology where data to modify the prior are fewer. Considering Eq (11), we could ensure that b has little contribution to the denominator if $\tau_0 \gg \lambda_i$ for all i, ensuring that the prior has little contribution to the posterior. For climate considerations, we envisage that the precision of natural variability, τ_0, would be large relative to each λ_i if bias is assessed by comparing model simulations to long time series of climate data. For epidemiological considerations, this would however rarely be the case. In the proposed method, we instead inform τ_0 largely by the simulation outputs, letting the projections of the ensemble determine how variable outcomes are.

Climate modeling, from which the proposed method is adapted, is primarily concerned with differences between current and future mean climate variables [24]. Epidemiology is not only concerned with mean projections but also with other quantities such as the probability of very large or long outbreaks occurring. Fig 2, panels G and H illustrates the probability of a given epidemic duration occurring for a single outbreak under the two control actions with the preferred SHW method. Comparing the posterior predictive distribution to the density of merely lumping the results of all simulations, as illustrated by the colored bars, the posterior predictive distribution of the ensemble method has a lower probability of both very long and short outbreaks. This is because projections of such outbreaks are down-weighted when their bias is assessed in the analysis; the observed outbreak duration would be unlikely under the modeling assumptions that produce these projections. Thus, ensemble methods that give equal weights to all projections can overestimate the uncertainty about outbreaks, preventing the models from informing appropriate policy decisions.

We have further extended the methodology to allow for informative priors on the weights. Compared to climate models, epidemiology often has far less available data to assess model bias. As such, expert opinion will often play a larger role within this field. Fig 4 illustrates the behavior of the ensemble prediction under such informative priors. When up-weighting projections for $i = 1$, which is also likely under the observed outbreak duration, the posteriors are shifted towards these projections and produce tighter distributions. This is also found when up-weighting the outlier, $i = 5$, in the small discrepancy ensemble (Fig 4, panels A and B), in which no projection is particularly unlikely for the observed duration. Projection x_5 is however unlikely in the large discrepancy ensemble. As a result, the effect of up-weighting the underlying modeling assumptions of this projection primarily makes the distribution wider, resulting from a larger uncertainty about individual weights (Fig 5). This is to be interpreted such that if expert opinions a priori determine that a modeling assumption that is unlikely to predict the observed data is better than other assumptions, the conclusions should be that there is less information in the ensemble as whole. However, when expert opinions are well informed and do not contradict with observed data, they can lead to more precise predictions.

It should be stressed that discrepancy among projections in the ensembles should be viewed as relative to τ_0, the estimated variability in outbreak duration given the initial conditions. A crucial difference between the original method applied to climate change and the epidemiological consideration presented here is that τ_0 is unknown for the latter and therefore needs to be estimated. We argue that in the absence of multiple outbreaks, it is sensible to inform this by the model simulations. Stochastic simulations are often used to estimate the range of outcomes for non-ensemble projections [1,17,18,23], and we propose that when extending the use of models to the ensemble context, they can be used to estimate this feature as well. We have therefore chosen a Bayesian model structure where τ_0 is informed largely by the within projection variability, $\tau_1, \tau_2,\ldots, \tau_n$, via the Gamma($a_\tau$, b_τ) distribution in Eq (8). All projections of the implemented control action contribute equally to this distribution in the method presented here, thus we are giving equal weights to all modeling assumptions in terms of informing τ_0. Estimation of different weights in terms of informing τ_0 based on a single outbreak, analogous to the estimation of λ, would not be conceivable. However, if policymakers believe that some modeling assumptions are more reliable in terms of capturing the variability of outcomes, we envisage that the Bayesian model structure can be altered to include this. If applied to endemic disease, τ_0 could be informed similarly to the natural variability of temperature in climate application, and the algorithm we supply is set up to handle this situation. Also, data from multiple outbreaks could be used to inform τ_0 when available. Yet, data quality will rarely be comparable to climate data, which highlights one of the major challenges for epidemiological modeling.

We also provide a multi-quantity extension of the Bayesian ensemble framework. Fig 5 shows that when adding number of infected and culled farms to the analysis, the marginal posteriors of outbreak duration become narrower and centered on x_1 and y_1, i.e. the projections based on the most likely scenario. This illustrates that predictions can be improved by incorporating multiple quantities when assessing the weights.

The main scope of this study is to introduce ensemble methods to the field of epidemiology rather than to produce inference about the 2001 FMD outbreak. However, Fig 7 illustrates the types of conclusions the method can provide. The three quantities we include in the multi-quantity analysis are all of great concern to policy makers when assessing the impact of control actions. The probability distributions represent the ensemble predicted difference in the outcome of the outbreak if the control action had excluded culling of CPs. The distributions all have most of the density above zeros, indicating that excluding culling of CPs would most likely have resulted in a prolonged and larger outbreak. We should however point out that these results are based on a single model. To make more robust predictions, we propose that the same type of analysis be made with projections of different models.

We also analyzed simulated data to provide a more general depiction of the performance of the method under different conditions. Fig 8 shows the result of analysis of ten simulated outbreaks with the parameterization in the center of the ensemble, i.e. $k_1 = k_2 = 1$. As this is in agreement with both the small and large discrepancy ensemble, the true values (dashed lines) consistently lie within the 95% credibility intervals. However, when using the $k_1 = k_2 = 0.9$ parameterization, the assumptions of the model used to simulate the outbreak is only inclusive of the large discrepancy ensemble, and consequently only the large discrepancy ensemble error bars are inclusive of the true values. Noting that we primarily use the different parameterization as a proxy for different models, this simple simulation example illustrates some obvious but essential points. Ensemble modeling should not be interpreted as a remedy for models based on poor assumptions about the modeled process. It offers the ability to combine multiple assumptions, thus integrating uncertainty with regards to this in the predictions. However, if all models are based on similar but inaccurate assumptions, ensemble modeling will not improve predictions. Intentionally making models similar to each other increases this risk and should be avoided if the models are to be used for ensemble purposes.

Accepting these limitations, we argue that the ensemble approach will be beneficial to epidemiological risk assessment because rather than choosing a single model for the purpose, it offers the possibility to combine projections from models that make mechanistically different assumptions about the transmission process. Thus, uncertainty with regard to this is incorporated in the predictions, which is important as projections of different models have been reported to deviate [63–65]. The use of multi-model ensembles would rely on collaboration of modeling teams, as well as overcoming confidentiality constraints in accessing outbreak data and population demographics. The current development in FMD modeling is seeing encouraging development in that area. The Quadrilateral Epiteam [19] has compared simulation of several outbreak scenarios in a subset of the UK demographics with five different models: NAADSM [45], Netherlands CVI [66], InterSpreadPlus [46], AusSpread [44] and ExoDis [67].

This demonstrates that potential obstacles for multi-model ensembles can be overcome and we envisage that epidemiology will see a shift towards multi-model ensembles to inform policy decisions, as has been seen in climate research [24,25] and weather forecasting [26,27]. Combining the results of multiple models however requires means of weighting these. We conclude that the presented framework is a promising approach because it provides easily interpretable probability distributions of quantities of interest. It also offers an appealing interpretation of model exchangeability, while at the same time combining several different weighting schemes, including a priori beliefs when such are available.

In this study, we introduced this framework by applying it to a simple question: how would exclusion of contiguous premises culling from the control action have affected the outcome of the UK 2001 outbreak? The aim of the study has been to introduce the methodological framework to epidemiology and solve some key issues associated with this transfer, including prior sensitivity, informing weights by expert opinion, using models to inform the variability in the outcome of individual outbreaks and extension to consider multiple epidemic quantities. We have purposely chosen the simple example because it allows for a straightforward transfer from the original climate implementation, and at the same time lets us demonstrate essential concepts and the potential of the framework. Models are however used to answer a range of different questions in epidemiology, and combining multiple projections has the potential to improve the way models are used to inform policy. We argue that the framework we introduce here has great potential, and foresee that many of the questions addressed in epidemiological modeling would require further developments of the Bayesian model, structured to fit with the specific problem. To facilitate this, we have supplied the algorithm (S1 File) and hope that it will aid further development of ensemble methods for epidemiology.

Supporting Information

S1 File. MCMC code written in MatLab.
(ZIP)

Author Contributions

Conceived and designed the experiments: TL MJT CW. Performed the experiments: TL MJT. Analyzed the data: TL. Contributed reagents/materials/analysis tools: TL MJT. Wrote the paper: TL MJT CW.

References

1. Woolhouse M (2011) How to make predictions about future infectious disease risks. Philos Trans R Soc Lond B Biol Sci 366: 2045–2054. doi: 10.1098/rstb.2010.0387 PMID: 21624924
2. Murray CJL, Rosenfeld LC, Lim SS, Andrews KG, Foreman KJ, et al. (2012) Global malaria mortality between 1980 and 2010: a systematic analysis. Lancet 379: 413–431. doi: 10.1016/S0140-6736(12)60034-8 PMID: 22305225
3. Thomson MC, Doblas-Reyes FJ, Mason SJ, Hagedorn R, Connor SJ, et al. (2006) Malaria early warnings based on seasonal climate forecasts from multi-model ensembles. Nature 439: 576–579. doi: 10.1038/nature04503 PMID: 16452977
4. Ferguson NM, Cummings D a T, Fraser C, Cajka JC, Cooley PC, et al. (2006) Strategies for mitigating an influenza pandemic. Nature 442: 448–452. doi: 10.1038/nature04795 PMID: 16642006
5. Anderson I (2002) Foot and Mouth Disease 2001: Lessons to be Learned Inquiry Report. London, UK.
6. Keeling MJ, Woolhouse MEJ, May RM, Davies G, Grenfell BT (2003) Modelling vaccination strategies against foot-and-mouth disease. Nature 421: 136–142. PMID: 12508120
7. Tildesley MJ, Savill NJ, Shaw DJ, Deardon R, Brooks SP, et al. (2006) Optimal reactive vaccination strategies for a foot-and-mouth outbreak in the UK. Nature 440: 83–86. doi: 10.1038/nature04324 PMID: 16511494
8. Tildesley MJ, Bessell PR, Keeling MJ, Woolhouse MEJ (2009) The role of pre-emptive culling in the control of foot-and-mouth disease. Proc R Soc B Biol Sci 276: 3239–3248. doi: 10.1098/rspb.2009.0427 PMID: 19570791
9. Keeling MJ, Woolhouse MEJ, Shaw DJ, Matthews L, Chase-Topping M, et al. (2001) Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. Science 294: 813–817. PMID: 11679661
10. Ferguson NM, Donnelly CA, Anderson R (2001) The Foot-and-Mouth Epidemic in Great Britain: Pattern of Spread and Impact of Interventions. Science 292: 1155–1160. doi: 10.1126/science.1061020 PMID: 11303090
11. Reiner RC, Stoddard ST, Forshey BM, King A a, Ellis AM, et al. (2014) Time-varying, serotype-specific force of infection of dengue virus. Proc Natl Acad Sci U S A 111: E2694–E2702. doi: 10.1073/pnas.1314933111 PMID: 24847073
12. Shea K, Tildesley MJ, Runge MC, Fonnesbeck CJ, Ferrari MJ (2014) Adaptive management and the value of information: learning via intervention in epidemiology. PLoS Biol 12: e1001970. doi: 10.1371/journal.pbio.1001970 PMID: 25333371
13. Massad E, Burattini MN, Lopez LF, Coutinho F a B (2005) Forecasting versus projection models in epidemiology: the case of the SARS epidemics. Med Hypotheses 65: 17–22. doi: 10.1016/j.mehy.2004.09.029 PMID: 15893110
14. Shaman J, Karspeck A (2012) Forecasting seasonal outbreaks of influenza. Proc Natl Acad Sci U S A 109: 20425–20430. doi: 10.1073/pnas.1208772109 PMID: 23184969
15. Hufnagel L, Brockmann D, Geisel T (2004) Forecast and control of epidemics in a globalized world. Proc Natl Acad Sci U S A 101: 15124–15129. doi: 10.1073/pnas.0308344101 PMID: 15477600
16. Ferguson NM, Keeling MJ, Edmunds WJ, Gani R, Grenfell BT, et al. (2003) Planning for smallpox outbreaks. Nature 425: 681–685. doi: 10.1038/nature02007 PMID: 14562094

17. Chao DL, Halloran ME, Obenchain VJ, Longini IM (2010) FluTE, a publicly available stochastic influenza epidemic simulation model. PLoS Comput Biol 6: e1000656. doi: 10.1371/journal.pcbi.1000656 PMID: 20126529

18. Buhnerkempe M, Tildesley MJ, Lindström T, Grear DA, Portacci K, et al. (2014) The Impact of Movements and Animal Density on Continental Scale Cattle Disease Outbreaks in the United States. PLoS One 9: e91724. doi: 10.1371/journal.pone.0091724 PMID: 24670977

19. Roche SE, Garner MG, Sanson RL, Cook C, Birch C, et al. (2014) Evaluating vaccination strategies to control foot-and-mouth disease: a model comparison study. Epidemiol Infect: 1–20. doi: 10.1017/S0950268814001927

20. Mollison D (1977) Spatial Contact Models for Ecological and Epidemic Spread. J R Stat Soc Ser B 39: 283–326. doi: 10.2307/2985089

21. Anderson RM, May RM (1979) Population biology of infectious diseases: Part I. Nature 280: 361–367. PMID: 460412

22. May R, Anderson R (1987) Transmission dynamics of HIV infection. Nature 326: 137–142. PMID: 3821890

23. Lindström T, Stenberg Lewerin S, Wennergren U (2012) Influence on disease spread dynamics of herd characteristics in a structured livestock industry. J R Soc Interface 9: 1287–1294. doi: 10.1098/rsif.2011.0625 PMID: 22112656

24. Tebaldi C, Knutti R (2007) The use of the multi-model ensemble in probabilistic climate projections. Philos Trans A Math Phys Eng Sci 365: 2053–2075. doi: 10.1098/rsta.2007.2076 PMID: 17569654

25. IPCC (2013) Summary for Policymakers. In: Stocker TF, Qin D, Plattner G- K, Tignor M, Allen SK, et al., editors. The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

26. Palmer TN, Doblas-Reyes FJ, Hagedorn R, Alessandri a., Gualdi S, et al. (2004) Development of a European Multimodel Ensemble System for Seasonal-To-Interannual Prediction (Demeter). Bull Am Meteorol Soc 85: 853–872. doi: 10.1175/BAMS-85-6-853

27. Gneiting T, Raftery AE (2005) Weather forecasting with ensemble methods. Science 310: 248–249. PMID: 16224011

28. Velázquez J a., Anctil F, Perrin C (2010) Performance and reliability of multimodel hydrological ensemble simulations based on seventeen lumped models and a thousand catchments. Hydrol Earth Syst Sci 14: 2303–2317. doi: 10.5194/hess-14-2303-2010

29. Cloke HL, Pappenberger F (2009) Ensemble flood forecasting: A review. J Hydrol 375: 613–626. doi: 10.1016/j.jhydrol.2009.06.005

30. Mangiameli P, West D, Rampal R (2004) Model selection for medical diagnosis decision support systems. Decis Support Syst 36: 247–259. doi: 10.1016/S0167-9236(02)00143-4

31. West D, Mangiameli P, Rampal R, West V (2005) Ensemble strategies for a medical diagnostic decision support system : A breast cancer diagnosis application. Eur J Oper Res 162: 532–551. doi: 10.1016/j.ejor.2003.10.013

32. Gårdmark A, Lindegren M, Neuenfeldt S, Blenckner T, Heikinheimo O, et al. (2013) Biological ensemble modeling to evaluate potential futures of living marine resources. Ecol Appl 23: 742–754. PMID: 23865226

33. Maiorano L, Falcucci A, Zimmermann NE, Psomas A, Pottier J, et al. (2011) The future of terrestrial mammals in the Mediterranean basin under climate change. Philos Trans R Soc Lond B Biol Sci 366: 2681–2692. doi: 10.1098/rstb.2011.0121 PMID: 21844047

34. Daszak P, Zambrana-Torrelio C, Bogich TL, Fernandez M, Epstein JH, et al. (2013) Interdisciplinary approaches to understanding disease emergence: the past, present, and future drivers of Nipah virus emergence. Proc Natl Acad Sci U S A 110 Suppl: 3681–3688. doi: 10.1073/pnas.1201243109 PMID: 22936052

35. Guis H, Caminade C, Calvete C, Morse AP, Tran A, et al. (2012) Modelling the effects of past and future climate on the risk of bluetongue emergence in Europe. J R Soc Interface 9: 339–350. doi: 10.1098/rsif.2011.0255 PMID: 21697167

36. Smith T, Ross A, Maire N, Chitnis N, Studer A, et al. (2012) Ensemble Modeling of the Likely Public Health Impact of a Pre-Erythrocytic Malaria Vaccine. PLoS Med 9: e1001157. doi: 10.1371/journal.pmed.1001157 PMID: 22272189

37. IPCC (2001) Climate Change 2001: Synthesis Report. Houghton JT, Ding Y, Griggs DJ, Noguer M, van der Linden PJ, et al., editors Cambridge, UK: Cambridge University Press.

38. Ye M, Pohlmann KF, Chapman JB (2008) Expert elicitation of recharge model probabilities for the Death Valley regional flow system. J Hydrol 354: 102–115. doi: 10.1016/j.jhydrol.2008.03.001

39. Räisänen J, Palmer TN (2001) A Probability and Decision-Model Analysis of a Multimodel Ensemble of Climate Change Simulations. J Clim 14: 3212–3226. doi: 10.1175/1520-0442(2001)014<3212:APADMA>2.0.CO;2

40. Burnham KP, Anderson DR (2004) Multimodel Inference Understanding AIC and BIC in Model Selection. Sociol Methods Res 33: 261–304. doi: 10.1177/0049124104268644

41. Gibbons JM, Cox GM, Wood a T a., Craigon J, Ramsden SJ, et al. (2008) Applying Bayesian Model Averaging to mechanistic models: An example and comparison of methods. Environ Model Softw 23: 973–985. doi: 10.1016/j.envsoft.2007.11.008

42. Jewell CP, Kypraios T, Christley RM, Roberts GO (2009) A novel approach to real-time risk prediction for emerging infectious diseases: a case study in Avian Influenza H5N1. Prev Vet Med 91: 19–28. doi: 10.1016/j.prevetmed.2009.05.019 PMID: 19535161

43. Tildesley MJ, Deardon R, Savill NJ, Bessell PR, Brooks SP, et al. (2008) Accuracy of models for the 2001 foot-and-mouth epidemic. Proc R Soc B Biol Sci 275: 1459–1468. doi: 10.1098/rspb.2008.0006 PMID: 18364313

44. Garner MG, Beckett SD (2005) Modelling the spread of foot-and-mouth disease in Australia. Aust Vet J 83: 758–766. PMID: 16395942

45. Harvey N, Reeves A, Schoenbaum M a, Zagmutt-Vergara FJ, Dubé C, et al. (2007) The North American Animal Disease Spread Model: a simulation model to assist decision making in evaluating animal disease incursions. Prev Vet Med 82: 176–197. doi: 10.1016/j.prevetmed.2007.05.019 PMID: 17614148

46. Stevenson M a, Sanson RL, Stern MW, O'Leary BD, Sujau M, et al. (2013) InterSpread Plus: a spatial and stochastic simulation model of disease in animal populations. Prev Vet Med 109: 10–24. doi: 10.1016/j.prevetmed.2012.08.015 PMID: 22995473

47. Tebaldi C, Smith RL, Nychka D, Mearns LO (2005) Quantifying Uncertainty in Projections of Regional Climate Change : A Bayesian Approach to the Analysis of Multimodel Ensembles. J Clim 18: 1524–1540. doi: 10.1175/JCLI3363.1

48. Giorgi F, Mearns LO (2002) Calculation of Average, Uncertainty Range, and Reliability of Regional Climate Changes from AOGCM Simulations via the "Reliability Ensemble Averaging" (REA) Method. J Clim 15: 1141–1158. doi: 10.1175/1520-0442(2002)015!1141:COAURAO2.0.CO;2

49. Schmitt SM, O'brien DJ, Bruning-Fann CS, Fitzgerald SD (2002) Bovine tuberculosis in Michigan wildlife and livestock. Ann N Y Acad Sci 969: 262–268. PMID: 12381603

50. Grassly NC, Fraser C (2006) Seasonal infectious disease epidemiology. Proc Biol Sci 273: 2541–2550. doi: 10.1098/rspb.2006.3604 PMID: 16959647

51. Diggle PJ (2006) Spatio-temporal point processes, partial likelihood, foot and mouth disease. Stat Methods Med Res 15: 325–336. doi: 10.1191/0962280206sm454oa PMID: 16886734

52. Deardon R, Brooks SP, Grenfell BT, Keeling MJ, Tildesley MJ, et al. (2010) Inference For Individual-Level Models Of Infectious Diseases In Large Populations. Stat Sin 20: 239–261.

53. Mahul O, Durand B (2000) Simulated economic consequences of foot-and-mouth disease epidemics and their public control in France. Prev Vet Med 47: 23–38.

54. Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian Data Analysis. 2nd ed. Chapman & Hall/CRC.

55. Lindström T, Grear D a, Buhnerkempe M, Webb CT, Miller RS, et al. (2013) A bayesian approach for modeling cattle movements in the United States: scaling up a partially observed network. PLoS One 8: e53432. doi: 10.1371/journal.pone.0053432 PMID: 23308223

56. Amiel JJ, Lindström T, Shine R (2014) Egg incubation effects generate positive correlations between size, speed and learning ability in young lizards. Anim Cogn 17: 337–347. doi: 10.1007/s10071-013-0665-4 PMID: 23922118

57. Garthwaite PH, Kadane JB, O'Hagan A (2005) Statistical Methods for Eliciting Probability Distributions. J Am Stat Assoc 100: 680–701. doi: 10.1198/016214505000000105

58. Garthwaite PH, Fan Y, Sisson SA (2014) Adaptive Optimal Scaling of Metropolis-Hastings Algorithms Using the Robbins-Monro Process. Commun Stat Theory Methods In press. PMID: 25089070

59. Roberts GO, Gelman A, Gilks WR (1997) Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms. Ann Appl Probab 7: 110–120. doi: 10.2307/2245134

60. Lindström T, Brown GP, Sisson SA, Phillips BL, Shine R (2013) Rapid shifts in dispersal behavior on an expanding range edge. Proc Natl Acad Sci 110: 13452–13456. doi: 10.1073/pnas.1303157110 PMID: 23898175

61. Smith RL, Tebaldi C, Nychka D, Mearns LO (2009) Bayesian Modeling of Uncertainty in Ensembles of Climate Models. J Am Stat Assoc 104: 97–116. doi: 10.1198/jasa.2009.0007

62. Wade PR (2000) Bayesian methods in conservation biology. Conserv Biol 14: 1308–1316.

63. Dubé C, Stevenson M a, Garner MG, Sanson RL, Corso B a, et al. (2007) A comparison of predictions made by three simulation models of foot-and-mouth disease. N Z Vet J 55: 280–288. doi: 10.1080/00480169.2007.36782 PMID: 18059645

64. Gloster J, Jones A, Redington A, Burgin L, Sørensen JH, et al. (2010) Airborne spread of foot-and-mouth disease—Model intercomparison. Vet J 183: 278–286. doi: 10.1016/j.tvjl.2008.11.011 PMID: 19138867

65. Sanson RL, Harvey N, Garner MG, Stevenson MA (2011) Foot and mouth disease model verification and “relative validation.” Rev Sci Tech L`Office Int Des Epizoot 30: 527–540. PMID: 21961223

66. Backer J a, Hagenaars TJ, Nodelijk G, van Roermund HJW (2012) Vaccination against foot-and-mouth disease I: epidemiological consequences. Prev Vet Med 107: 27–40. doi: 10.1016/j.prevetmed.2012.05.012 PMID: 22749763

67. DEFRA (2005) D5100/R3. Cost Benefit Analysis of Foot and Mouth Disease Controls.