

1       **EFFICIENT HISTORY MATCHING OF A HIGH DIMENSIONAL**  
2       **INDIVIDUAL BASED HIV TRANSMISSION MODEL\***

3       IOANNIS ANDRIANAKIS<sup>†</sup>, NICKY MCCREESH<sup>†</sup>, IAN VERNON<sup>‡</sup>, TREVELYAN J.  
4       MCKINLEY<sup>§</sup>, JEREMY E. OAKLEY<sup>¶</sup>, REBECCA N. NSUBUGA<sup>||</sup>, MICHAEL GOLDSTEIN<sup>‡</sup>  
5       , AND RICHARD G. WHITE<sup>†</sup>

6       **Abstract.** History matching is a model (pre-)calibration method that has been applied to  
7       computer models from a wide range of scientific disciplines. In this work we apply history matching  
8       to an individual based epidemiological model of HIV that has 96 input and 50 output parameters,  
9       a model of much larger scale than others that have been calibrated before, using this or similar  
10      methods. Apart from demonstrating that history matching can analyse models of this complexity, a  
11      central contribution of this work is that the history match is carried out using linear regression, an  
12      elementary and easier to implement statistical tool compared to the Gaussian process based emulators  
13      that have previously been used. Furthermore, we address a practical difficulty of history matching,  
14      namely, the sampling of tiny, non-implausible spaces, by introducing a sampling algorithm adjusted  
15      to the specific needs of this method. The effectiveness and simplicity of the history matching method  
16      presented here shows that it is a useful tool for the calibration of computationally expensive, high  
17      dimensional individual based models.

18      **Key words.** Emulation, Calibration, Gaussian processes, Linear regression

19      **AMS subject classifications.** 62-07, 62P10, 62J05

20      **1. Introduction.** Approximately 1.5 million people died from AIDS-related ill-  
21      nesses in 2013, with sub-Saharan Africa accounting for 74% of deaths [24]. In the  
22      same year, 2.1 million people were newly infected with HIV. Although both HIV inci-  
23      dence and mortality have fallen in recent years, more intensive treatment and control  
24      strategies are needed to accelerate the decline. Antiretroviral therapy (ART) is known  
25      to suppress the virus and stop the progression of the disease, and it can also prevent  
26      onward transmission. This therapy is available in various sub-Saharan countries and  
27      is typically administered when the CD4 count of a patient falls below a threshold.  
28      There is however an ongoing discussion about removing this threshold, and the effect  
29      such a policy would have on the general population.

30      Modelling offers one way of studying the effect of different interventions. An indi-  
31      vidual based model (simulator) has been developed at the London School of Hygiene  
32      and Tropical Medicine which can simulate HIV transmission and the effects of ART,  
33      and predict the effect of different interventions over a horizon of 10-15 years. The  
34      simulator has a number of input parameters, the values of which are uncertain, and  
35      this uncertainty should be included in any predictions we wish to make. The availabil-  
36      ity of historical data (observations) allows us to learn about the values of the input

---

\*Submitted to the editors 07/09/2016.

**Funding:** This work was funded by the UK Medical Research Council (MRC) and the UK Department for International Development (DFID) under the MRC/DFID Concordat agreement that is also part of the EDCTP2 programme supported by the European Union (MR/J005088/1). RGW is additionally funded by the Bill and Melinda Gates Foundation (TB Modelling and Analysis Consortium: OPP1084276) and UNITAID (4214-LSHTM-Sept15; PO #8477-0-600).

<sup>†</sup>Dept. of Infectious Disease Epidemiology, London School of Hygiene & Tropical Medicine, London, UK. ([andrianakis@yahoo.com](mailto:andrianakis@yahoo.com))

<sup>‡</sup>Dept. of Mathematical Sciences, Durham University, Durham, UK.

<sup>§</sup>College of Engineering, Mathematics and Physical Sciences, University of Exeter, UK.

<sup>¶</sup>School of Mathematics and Statistics, University of Sheffield, Sheffield, UK.

<sup>||</sup>Medical Research Council/Uganda Virus Research Institute, Uganda Research Unit on AIDS, Entebbe, Uganda.

37 parameters, by *calibrating* the simulator to the observations. By calibrating we mean  
 38 finding a subset of input parameter values for which the simulator’s outputs closely  
 39 match historical observations on demography, HIV prevalence, mortality etc.

40 Calibrating such a simulator is challenging mainly due to the large number of  
 41 input (96) and output (50) parameters. Having to simultaneously match a large  
 42 number of outputs means that there are many constraints that need to be satisfied,  
 43 and this can result in a very small region of the input space where the simulator  
 44 matches the observations. In high dimensional input spaces, the search for a small  
 45 part that will generate output matches can require a prohibitively large number of  
 46 simulator runs. A further complication in our case is that the simulator is stochastic,  
 47 therefore repeated evaluations are required for the same input values to estimate the  
 48 mean values of the outputs.

49 A large number of methodologies for calibrating simulators are available, which  
 50 vary from simple least squares techniques, MCMC based methodologies [7, 19], parti-  
 51 cle filters [3], and Approximate Bayesian Computation (ABC) [23, 17] amongst others.  
 52 For various reasons these methodologies are extremely difficult to apply in our case.  
 53 Data augmentation approaches would require reconstruction of the likelihood function  
 54 and integration over a very large hidden state space, whilst simulation-based methods  
 55 would require a very large number of simulator runs. The latter have only really been  
 56 applied to relatively small scale simulators (in terms of the number of inputs and  
 57 outputs; usually around 5-10 in each case).

58 The problem of calibrating a simulator could also be thought of as an optimisation  
 59 problem: finding simulator inputs to minimise the difference between the simulator  
 60 outputs and some observed target values. The optimisation of an expensive simulator  
 61 with a univariate output is considered in [10], and an extension for the multivariate  
 62 case is given in [12]. Again, we think these approaches would be difficult to apply in  
 63 our case, given the high dimensional input and output, and it is not obvious how one  
 64 would account for simulator input uncertainty, if the aim was simply to find a ‘best’  
 65 value.

66 History matching [6] is a (pre-)calibration method that has been applied with  
 67 success to slow simulators with typically larger numbers of inputs/outputs than the  
 68 simulators used in the methods mentioned above. History matching tries to identify  
 69 those parts of the simulator’s input space where, if evaluated, the simulator is likely  
 70 to match the observations. This goal is achieved via identifying regions of the input  
 71 space where a match is unlikely to be found (these regions are known as *implausible*)  
 72 and discarding them in iterations known as *waves*. History matching can deal with  
 73 simulators that are slow to evaluate by employing statistical models of the simulator  
 74 (known as *emulators*), whose key characteristic is the trivial evaluation time.

75 History matching was first applied in the field of oil simulator modelling [6],  
 76 and has since found applications in areas as diverse as galaxy formation [26, 28],  
 77 environmental models [9], systems biology [25, 29], ocean modelling [32] and epidemi-  
 78 ology [2]. The dimensionality (i.e. number of inputs (P) and outputs (R)) of the  
 79 simulators analysed in those works was considerably smaller. For example, [26, 28]  
 80 ( $P = 17, R = 11$ ), [9] ( $P = 17, R = 13$ ), [25] ( $P = 8, R = 15$ ), [32] ( $P = 26, R = 4$ ), [2]  
 81 ( $P = 22, R = 18$ ). An exception is [6] who study an oil reservoir model with  $P = 40$   
 82 and  $R = 77$ . However, [6] use the technique of active inputs to perform a substantial  
 83 dimensional reduction of the simulator, showing that it is accurately described by a  
 84 series of outputs possessing only 3 input dimensions each. However, many high di-  
 85 mensional simulators, such as the HIV model we are concerned with here, possess a far  
 86 more intricate input-output structure, for which such a large dimensional reduction

87 is not possible.

88 In this work we apply history matching to a stochastic agent based simulator with  
89 more input and output parameters than any other simulator that has been calibrated  
90 before with this or with other methods that we know of. A key contribution of  
91 this work is that we calibrate the simulator using elementary statistical tools and  
92 methodologies, that should be known to all statisticians and most modellers with  
93 basic statistical training.

94 The emulators used in history matching are typically built using Gaussian pro-  
95 cesses (GPs) [2]. In our experience, we have often found this to be an obstacle in  
96 applying the method, as not everyone is familiar with this elegant but non-trivial  
97 statistical model. The emulators we use in this work are based on linear regression  
98 models, that are much easier to code, fit, and interpret. History matching typically  
99 requires repeated fitting of emulators, either multivariate or a large number of uni-  
100 variate ones at each wave. Because its emphasis is on excluding the implausible input  
101 space iteratively, it does not depend on the availability of very precise emulators to  
102 achieve this; the same result can sometimes be achieved with less precise emulators  
103 and a few additional iterations. Therefore, linear regression models are likely to be  
104 sufficient, especially in the first waves of a history match. While we do not argue  
105 against the use of more complex statistical models for building emulators, we demon-  
106 strate that even a simple and well known tool, such as linear regression, can be used  
107 to make considerable progress in calibrating complex and computationally demanding  
108 simulators. Furthermore, the history matching framework facilitates the use of vari-  
109 ous statistical models at different stages of the process. Therefore, linear regression  
110 can be used in the initial waves and more complex regression tools, such as GPs, can  
111 be introduced later on, if some outputs are hard to emulate, or greater accuracy is  
112 required.

113 Another contribution of this work is the proposal of an algorithm that can uni-  
114 formly sample the non-implausible space. After several waves of history matching,  
115 the remaining non-implausible space can be a tiny proportion of the original (i.e. at  
116 wave 0). In general, there is no analytical description of this space and the only way  
117 to describe it is via sampling. However, this can be challenging, since this space has  
118 as many dimensions as the simulator’s inputs, has an unknown, (perhaps multimodal)  
119 shape and can be several orders of magnitude smaller than the original input space.  
120 The algorithm we propose is based on the slice sampler, is straightforward to imple-  
121 ment, and takes advantage of the specific needs of history matching to increase its  
122 efficiency. Another sampling algorithm that addresses the same problem has been pro-  
123 posed in [33], although that algorithm is intricate and significantly more challenging  
124 to implement.

125 History matching can give valuable insight into the simulator’s structure and  
126 the structure of the constraints imposed by the data. For example, by studying the  
127 correlation patterns of the fitted inputs and outputs one can understand or verify  
128 how the simulator’s internal processes interact, information that could be useful in  
129 developing the simulator further or in deriving model discrepancy terms. In the  
130 case of the simulator studied here, we also learn about epidemiological processes,  
131 the interaction between epidemiologically meaningful parameters and their plausible  
132 values, which can then be compared to those found in the literature. Finally, history  
133 matching can provide large numbers of calibrated input values, which can be used to  
134 run the simulator into the future, allowing predictions to incorporate the uncertainty  
135 about the simulator’s input values. In our case, calibrated input values are fed into  
136 several other research projects, one of which is [14], a complex decision analysis on

137 predicting the costs and effects of different ways of scaling-up access to HIV treatment  
138 in Uganda.

139 This paper is structured as follows: Section 2 describes the individual based simu-  
140 lator. Section 3 gives an overview of history matching and details the methodological  
141 additions of this paper: the use of linear regression models as emulators and the  
142 sampling algorithm. Section 4 shows the fit of the outputs to the observations, the  
143 reduction of the non-implausible space, and presents key conclusions on the simula-  
144 tor’s behaviour that were drawn from history matching. It also discusses the benefits  
145 of linear regression based emulators and the effectiveness of the proposed sampling  
146 scheme. Section 5 concludes this work.

147 **2. Simulator and problem description.** The simulator we developed was an  
148 individual-based model, written in NetLogo [31]. It simulates births, deaths, and  
149 population growth between 1950-2030, the formation and dissolution of sexual part-  
150 nerships, HIV transmission, disease progression and mortality, the HIV/ART care  
151 pathway, the effects of ART on HIV mortality and transmission, and the development  
152 and transmission of drug resistance. Key pathways into and through care are explic-  
153 itly simulated, to allow the effects of different ways of scaling up ART coverage to be  
154 accurately captured. In the simulator, people can be tested for HIV through routine,  
155 intervention, or antenatal HIV testing programs, or after experiencing HIV-related  
156 morbidity. Upon testing positive, a proportion of people are successfully linked to  
157 care. Once linked to care, a proportion of people who are eligible start ART, and  
158 the remainder receive pre-ART care. People can move from pre-ART care to ART  
159 once they are identified as eligible following a CD4 test, or due to severe morbidity.  
160 People can drop out of care at any stage. People who drop out of pre-ART care can  
161 re-enter through the same pathways used by people to enter initially. People who  
162 drop out of ART, restart ART at a rate determined by input parameters. Changes in  
163 ART eligibility criteria over time in Uganda are also simulated. In total, 50 simulator  
164 outputs and acceptable ranges were selected to enable the model, once calibrated, to  
165 accurately reflect key features of HIV epidemiology and care in Uganda. These consist  
166 of:

- 167 • Four demographic outputs, which captured key aspects of non-HIV mortality  
168 and population growth in Uganda.
- 169 • Nine sexual behaviour outputs, which captured patterns of sexual behaviour  
170 in the country.
- 171 • Five HIV prevalence outputs, to ensure that the model reflects trends in male  
172 and female HIV prevalence in Uganda over time.
- 173 • The median survival with HIV before the introduction of ART.
- 174 • Eight HIV testing outputs, reflecting trends in rates of HIV testing in HIV  
175 positive and negative men and women over time.
- 176 • Four pre-ART care coverage and twelve ART coverage outputs, to capture  
177 the scale-up of HIV care in Uganda in 2003-2014.
- 178 • Five ART retention outputs, to capture ART drop out and restart rates.
- 179 • Three second line ART outputs, to capture the proportion of people on second  
180 line ART.

181 The simulator was designed and parameterised to represent the population of  
182 Uganda as a whole. To improve simulator run times, however, only 1/2000th of the  
183 population of Uganda was simulated in each model run. As a result, we were only  
184 interested in modelling the mean output of the simulator, while the variance was  
185 deemed to have no real-world meaning in this case.

186 Once calibrated, the model can be used to simulate different ART scale-up strate-  
 187 gies, and to estimate their effects on HIV incidence, mortality, morbidity, and drug  
 188 resistance, and on ART programme and other healthcare costs. A full list of the  
 189 simulator’s inputs and outputs is given in the supplementary material. A detailed  
 190 description of the simulator can be found in [14].

### 191 3. Methods.

192 **3.1. History matching.** History matching assumes the existence of a physical  
 193 process  $y$  which is observed through observations  $z$  and a computer model (simulator)  
 194 that models  $y$ . The goal of history matching is to identify the regions of input space  
 195 corresponding to acceptable matches, and this is performed by ruling out the *implau-*  
 196 *sible* regions iteratively in waves. A brief description of history matching is given in  
 197 the following and a more detailed exposition can be found in [26, 2].

198 **3.1.1. Linking the emulator to observations.** Let

$$199 (1) \quad z = y + \phi,$$

200 where  $z$  is an observation of the physical process  $y$  which is done with some measure-  
 201 ment error  $\phi$ . Also let

$$202 (2) \quad y = g(\mathbf{x}^*) + \delta,$$

203 where  $g(\mathbf{x}^*)$  is the simulator’s output when this is evaluated at input  $\mathbf{x}^*$ . The term  
 204  $\delta$  is the model error term, which represents the discrepancy between the simulator’s  
 205 output when this is evaluated at its ‘best’ input  $\mathbf{x}^*$  and the physical process  $y$ . This  
 206 discrepancy often arises because either some parts of the physical process are not  
 207 completely understood and is not possible to include them in the simulator or they  
 208 have been deliberately left out e.g. for mathematical or computational tractability.  
 209 For more on model discrepancy the reader may consult [11, 8, 5].

210 Evaluating  $g(\mathbf{x})$  can be time consuming and exploration of the simulator’s input  
 211 space can require a very large number of evaluations. For this reason, a surrogate  
 212 statistical model (*emulator*) is built for the simulator, which a) can predict  $g(\mathbf{x})$  for  
 213 any  $\mathbf{x}$  of interest very quickly and b) can quantify the uncertainty of its predictions.  
 214 We will return to emulation in section 3.2, but for the moment let us just say that  
 215 the emulator’s predictions are linked to  $g(\mathbf{x})$  via

$$216 (3) \quad g(\mathbf{x}) = E^*[g(\mathbf{x})] + \zeta(\mathbf{x}),$$

217 where  $E^*[g(\mathbf{x})]$  is the emulator’s prediction for  $g(\mathbf{x})$  and  $\zeta(\mathbf{x})$  is the estimation error,  
 218 whose statistical characteristics can vary with  $\mathbf{x}$ .

219 Combining equations 1, 2 and 3 we can write

$$220 (4) \quad z = E^*[g(\mathbf{x}^*)] + \zeta(\mathbf{x}^*) + \delta + \phi.$$

221 The above equation refers to a single output simulator and accordingly to a scalar  
 222 observation. In case the simulator has  $R$  outputs and there are  $R$  observations  $z$   
 223 available, there will be  $R$  instances of equation 4, where each quantity will be indexed  
 224 by the output index  $r$ .

225 **3.1.2. The implausibility measure.** History matching works by rejecting the  
 226 input space which is found to be implausible. This characterisation is done using the  
 227 *implausibility measure* which is based on equation 4. For the  $r$ -th output we can write

228 the variance of the error terms in equation 4 as  $V_{o,r} = \text{Var}[\phi_r]$ ,  $V_{m,r} = \text{Var}[\delta_r]$  and  
 229  $V_{c,r} = \text{Var}[\zeta_r(\mathbf{x})]$ . We can then formulate a natural metric for the distance between  
 230 the observation  $z_r$  and the emulator's prediction at  $\mathbf{x}$  as

$$231 \quad (5) \quad I_r(\mathbf{x}) = \frac{|z_r - \mathbb{E}^*[g_r(\mathbf{x})]|}{(V_{o,r} + V_{m,r} + V_{c,r})^{1/2}}.$$

232 This is the basic form of the implausibility measure for one output, which is essentially  
 233 the distance between  $z_r$  and  $\mathbb{E}^*[g_r(\mathbf{x})]$ , standardised by all the uncertainties that might  
 234 be present in the system: the uncertainty due to observation error  $V_{o,r}$ , model error  
 235  $V_{m,r}$  and the code uncertainty  $V_{c,r}$ , which arises because we cannot evaluate the  
 236 simulator (code) for every  $\mathbf{x}$  and we substitute it with the emulator.

237 Simple distributional assumptions on the form of the various error terms, namely  
 238 a zero mean and a unimodal distribution, allow us to use the powerful and underused  
 239 Pukelsheim's rule [21] to derive cut-off values for the implausibility. That is, to come  
 240 up with thresholds such that if  $I_r(\mathbf{x})$  exceeds them we can be fairly confident that the  
 241 simulator's output  $g(\mathbf{x})$  will not be close to the observations  $z$  for this particular value  
 242 of  $\mathbf{x}$ . Pukelsheim's 3 sigma rule states that any unimodal continuous distribution  
 243 contains 95% of its probability mass within 3 standard deviations from its mean,  
 244 regardless of its skewness or higher moments. Therefore, for  $I_r(\mathbf{x}) > 3$  it will be  
 245 highly unlikely that the simulator's  $r$ th output will match the respective observation  
 246 for that particular  $\mathbf{x}$ .

247 A simple extension of the single output implausibility (Eq. 5) to multiple outputs  
 248 can be found by maximising across all outputs, i.e.

$$249 \quad I(\mathbf{x}) = \max_r I_r(\mathbf{x}).$$

250 The implausibility measure has several other extensions, some of which can incor-  
 251 porate correlation structures between outputs. For more information, the interested  
 252 reader is referred to the detailed discussion in [26].

253 **3.1.3. Procedure.** History matching iteratively discards parts of the input space  
 254 which are calculated as implausible and therefore highly unlikely to contain matches  
 255 between the simulator's outputs and the observations. In wave  $\eta$ , the search for ac-  
 256 ceptable matches is limited to the previous wave's non-implausible space ( $\mathcal{X}_{\eta-1}$ ) and  
 257 as a result the non-implausible space shrinks with each iteration (i.e.  $\mathcal{X}_\eta \subset \mathcal{X}_{\eta-1}$ ).  
 258 An outline of the procedure is given in the following:

- 259 1. Define the initial  $P$ -dimensional non-implausible space  $\mathcal{X}_{\eta=0}$ .
- 260 2. Select  $N$  training points from the current non-implausible space  $\mathcal{X}_\eta$ , using a  
 261 space filling design or some other method that aims to cover  $\mathcal{X}_\eta$ .
- 262 3. Evaluate the simulator at each of the  $N$  points. If the model is stochastic, run  
 263 the simulator  $K$  times at each of the  $N$  points. Denoting by  $\hat{g}(\mathbf{x})$  the averaged  
 264 simulator output evaluated at  $\mathbf{x}$ , form the training data  $D = \{\mathbf{x}_n, \hat{g}_n(\mathbf{x})\}_{n=1}^N$ .
- 265 4. Build and validate an emulator for as many of the simulator's  $R$  outputs as  
 266 is possible. Denote the set of emulated outputs as  $R_{\eta+1}$ . The emulators of  
 267 wave  $\eta + 1$  are defined only over  $\mathcal{X}_\eta$ , and should be more accurate than the  
 268 emulators of the previous wave, as  $\mathcal{X}_\eta$  is smaller than  $\mathcal{X}_{\eta-1}$ .
- 269 5. Evaluate the implausibility measure  $I(\mathbf{x})$  over all  $r \in R_{\eta+1}$  for a large number  
 270 of  $\mathbf{x} \in \mathcal{X}_\eta$  such that  $\mathcal{X}_\eta$  is represented with sufficient accuracy.  $\mathcal{X}_{\eta+1}$  is defined  
 271 as the set of  $\mathbf{x} \in \mathcal{X}_\eta$  for which  $I(\mathbf{x})$  is less than a chosen threshold.  $\mathcal{X}_{\eta+1}$   
 272 should be smaller than  $\mathcal{X}_\eta$ .

- 273 6. Unless one of the following conditions is true, increase wave counter  $\eta$  by 1  
 274 and repeat steps 2-5.
- 275 (a) The emulator's uncertainty  $V_c$  is smaller than the other uncertainties  
 276 (e.g.  $V_o$  or  $V_m$ ), therefore more waves would most likely lead to little  
 277 further reduction of  $\mathcal{X}_\eta$ .
  - 278 (b) All  $\mathcal{X}_\eta$  is implausible (i.e. all  $\mathcal{X}_{\eta+1}$  is empty).
  - 279 (c) A sufficient number of points  $\mathbf{x}$  that match the observation data have  
 280 been collected for the purposes of subsequent analyses.

281 Some comments on the above procedure: in step 2, a reasonable method for  
 282 choosing the  $N$  points at which the simulator is to be evaluated is a uniform design  
 283 with some space filling properties. Were we to know the particular type of regression  
 284 that we would fit, perhaps alternative designs could be more appropriate. But in  
 285 the absence of such information, a uniform, space filling design is a good all purpose  
 286 choice that is informative about the whole space. A standard choice for creating  
 287 such a design is via a maximin Latin hyper-cube [16] and can be used when possible.  
 288 This design however, is challenging to create in high dimensional spaces of arbitrary  
 289 shapes, as  $\mathcal{X}_\eta$  is likely to be. A simple but effective alternative is the following: start  
 290 with a large number of points distributed uniformly in  $\mathcal{X}_\eta$ , e.g. as provided by the  
 291 slice sampler of section 3.3. Choose the first point at random and choose the second  
 292 as the one that is the furthest apart from the first, in the sense of the Euclidean  
 293 distance. For each of the remaining points, calculate the distance to the closest of  
 294 the two first points (minimum distance) and choose as third the one with the largest  
 295 minimum distance to the first two (i.e. maximum minimum distance - maximin).  
 296 Similarly, choose as fourth the point with a maximin distance to the first 3 and so  
 297 on until  $N$  points are collected. This is a simple procedure that returns points that  
 298 are sufficiently well-spread and cover the entire input space, assuming that enough  
 299 samples from  $\mathcal{X}_\eta$  are available.

300 Condition (6a) implies that decreases in the  $V_c$  term (code uncertainty), which  
 301 should come with additional waves and improved emulators, are unlikely to contribute  
 302 in shrinking  $\mathcal{X}_\eta$  further, as the denominator of the implausibility  $I_r(\mathbf{x})$ , will be domi-  
 303 nated by  $V_o$  and  $V_m$ . If condition (6a) occurs, most of the simulator runs should fall  
 304 within the observations and history matching can be stopped. At this point, sampling  
 305 the non-implausible space  $\mathcal{X}_\eta$  should provide as many input parameters that match  
 306 the observations as required by the application.

307 Condition (6b) is an indication that the simulator cannot match the observations,  
 308 unless the errors  $V_o$ ,  $V_m$  are revised and perhaps increased. Flagging a simulator's  
 309 possible inability to match a particular calibration data set is a strong point of history  
 310 matching, which is contrasted to more traditional Bayesian calibration approaches  
 311 that will return an input parameter posterior distribution regardless of the quality of  
 312 the match.

313 Regarding step 3, outputs that are hard to emulate in the initial waves, perhaps  
 314 due to the large variation of the inputs, can become easier to handle in later waves,  
 315 when the inputs are confined in more interesting input space parts, where the simu-  
 316 lator's response can be smoother. Additionally, inputs that have a strong effect on  
 317 outputs in the initial waves, can become less important in later waves when their range  
 318 has been reduced, and other, previously unnoticed inputs, can start having a greater  
 319 impact on the simulator's behaviour, allowing more detailed emulator construction.

320 Finally, the non-implausible space can be reduced by several orders of magnitude  
 321 at each iteration, as will be demonstrated in the results section. As a result, even  
 322 a very dense initial design can end up having all its points outside the region of

323 interest within a very small number of waves. Therefore, the strategy of evaluating  
 324 the simulator at each wave for a relatively smaller number of times, allows us to focus  
 325 the computational effort in input space areas that are more likely to produce a match  
 326 to the data.

327 **3.1.4. Convergence.** History matching continues until one of the three condi-  
 328 tions mentioned in the above procedure are satisfied. That is, the history matching  
 329 waves proceed until the emulator uncertainty is smaller than the other uncertainties  
 330 in the implausibility measure, or until all  $\mathcal{X}_\eta$  has been characterised as implausible,  
 331 or until enough matches to the observation data have been found for the purposes of  
 332 the application.

333 A natural question that can arise at this point is whether history matching will  
 334 successfully identify all input space regions that match the observation data and  
 335 whether some areas might be missed. Technically, it is possible to miss some areas of  
 336 the input space that result in a match, i.e. incorrectly identify them as implausible, if  
 337 the emulators do not represent accurately the uncertainty about the simulator’s be-  
 338 haviour. Additionally, the implausibility can be seen as a statistical test that predicts  
 339 whether a particular input  $\mathbf{x}$  will match the outputs to the calibration data. Even  
 340 conservative (i.e. large) cut-off values imply that there is a small but nevertheless  
 341 non-zero probability that a good input is left out.

342 We can guard against the first condition by ensuring that the emulators are val-  
 343 idated. That is, that the simulator’s outputs fall within the uncertainty intervals  
 344 provided by the emulator. The second condition can be guarded against by choosing  
 345 suitably high implausibility cut-offs, especially in the initial waves, where the uncer-  
 346 tainty about the simulator’s behaviour is large. Finally, the smooth behaviour that  
 347 a biological system simulator should possess offers additional confidence that input  
 348 regions of interest have not been left out.

349 **3.2. Linear regression emulators.** As mentioned in the introduction, history  
 350 matching typically relies on emulators to ease the computational burden of having  
 351 to evaluate a potentially slow simulator a large number of times. Gaussian processes  
 352 (GPs) have been used extensively to build emulators, as they are a flexible statistical  
 353 model with extensive presence in recent literature. In this work however, we are  
 354 using linear regression as the model of choice for building emulators. Even though  
 355 linear regression models tend to be less flexible compared to GPs, they do offer some  
 356 advantages for history matching. First, they are generally easier to fit than GPs,  
 357 which is of assistance in the presence of a large number of simulator outputs, each of  
 358 which requires its own emulator. Second, in complex high dimensional simulators it is  
 359 common that each output is not influenced by every input of the simulator, but there  
 360 tends to be a subset of inputs that affects more the behaviour of a particular output.  
 361 These inputs are generally referred to as *active inputs*. Active inputs are not always  
 362 known *a priori*, and different sets of inputs might affect the same output at different  
 363 waves, as the input space shrinks due to history matching. Linear regression models  
 364 offer a simple and established way of choosing the active inputs for each output at  
 365 each wave. Although similar results could be achieved using GPs and Automatic  
 366 Relevance Determination (ARD) [34], doing so with linear regression models can be  
 367 more straightforward. Finally, linear models are considerably easier to implement.

368 The fundamental equation for linear regression is

369 (6) 
$$g(\mathbf{x}) = \sum_{i=1}^q h_i(\mathbf{x})\beta_i + \epsilon,$$

370 where  $h_i(\mathbf{x})$  are functions of the inputs  $\mathbf{x}$ ,  $\beta_i$  are their respective coefficients, and  $\epsilon$  is  
 371 residual, uncorrelated noise. The functions  $h_i(\cdot)$  can take any form (linear, quadratic,  
 372 interaction term between components of  $\mathbf{x}$ , or other non-linear transformation). The  
 373 term ‘linear’ in the description ‘linear regression model’ therefore refers to the lin-  
 374 ear relationship between the arbitrary functions  $h_i(\mathbf{x})$  and the coefficients  $\beta_i$ . De-  
 375 termining the exact form of the  $h_i(\mathbf{x})$  functions is essentially fitting the (statisti-  
 376 cal) model and the strategy we follow here is presented in section 3.2.1. Denoting  
 377  $h(\mathbf{x}) = [h_1(\mathbf{x}), \dots, h_q(\mathbf{x})]$  and  $\beta = [\beta_1, \dots, \beta_q]^T$ , where  $[\cdot]^T$  is vector transpose, we  
 378 can write equation 6 as

$$379 \quad (7) \quad g(\mathbf{x}) = h(\mathbf{x})\beta + \epsilon.$$

380 At each wave the simulator is evaluated  $K$  times at  $N$  points, thus producing the  
 381 training data  $D = \{\mathbf{x}_n, \hat{g}(\mathbf{x}_n)\}_{n=1}^N$ , which we also denote for brevity as  $D = (X, Y)$ .  
 382 If  $H$  is an  $N \times q$  matrix whose rows are the vectors  $h(\mathbf{x}_1), \dots, h(\mathbf{x}_N)$ , the maximum  
 383 likelihood estimate (m.l.e.) of  $\beta$  is given by the well known equation

$$384 \quad \hat{\beta} = (H^T H)^{-1} H^T Y.$$

385 Similarly, the model’s prediction at an untested  $\mathbf{x}$  is simply given by

$$386 \quad E^*[\mathbf{x}] = h(\mathbf{x})\hat{\beta}.$$

387 Finally, the m.l.e. of the uncertainty about a prediction at a new input  $\mathbf{x}$  is given by

$$388 \quad \hat{\sigma}^2 = (Y^T Y - Y^T H (H^T H)^{-1} H^T Y) / N.$$

389 The  $\hat{\sigma}^2$  term represents the code uncertainty of section 3.1.1. As a more subtle point  
 390 here we could mention that  $\hat{\sigma}^2$  also includes also the uncertainty that arises from  
 391 estimating  $g(\mathbf{x})$  from the averages  $\hat{g}(\mathbf{x})$ , which in a GP emulator would have been  
 392 modelled by the nugget term [1].

393 **3.2.1. Fitting strategy.** The main tool we use in determining the exact func-  
 394 tionals for the  $h(\mathbf{x})$  terms is based on the Bayesian Information Criterion (BIC) [22],  
 395 which in this case is given by

$$396 \quad BIC = N[\ln(2\pi\hat{\sigma}^2) + 1] + (q + 1) \ln(N).$$

397 According to this, when presented with two alternative sets of functions  $\{h_i(\mathbf{x})\}_{i=1}^{q_1}$   
 398 and  $\{h'_i(\mathbf{x})\}_{i=1}^{q_2}$  the one that results in a lower score for the BIC is to be preferred.  
 399 Using this as our main fitting tool we develop the following strategy:

400 *Zero order:* Always include a constant term to account for the overall mean of  
 401 the data. The *current regression matrix* is set to  $h(\mathbf{x}) = 1$ .

402 *First order:* Form the regression matrix  $h'(\mathbf{x}) = [h(\mathbf{x}), x_p]$  for each of the  $P$   
 403 inputs in the model and compare the resulting BIC to that of the model that only  
 404 includes  $h(\mathbf{x})$ . The input  $p$  that results in the largest drop in the BIC is added in the  
 405 emulator’s *active input* set and the matrix  $[h(\mathbf{x}), x_p]$  becomes the current regression  
 406 matrix i.e.  $h(\mathbf{x}) \leftarrow [h(\mathbf{x}), x_p]$ . The procedure is repeated for the rest of the inputs,  
 407 and stops when no additional input decreases the BIC further. The inputs included  
 408 in the emulator in this way form its *active input* set, which is often much smaller than  
 409 the full input set of the simulator.

410 *Higher orders:* Here by higher order terms we mean functions  $h_i(\mathbf{x})$  that include  
 411 either powers of  $x_p$  greater than 1 or products of  $x_p$  and their higher order terms. We  
 412 refer to a term as having  $n$ -th order when the sum of the powers of the terms involved  
 413 equals  $n$ . A second order term for example, can contain interactions  $x_i x_j$  or squares  
 414 such as  $x_i^2$ . We follow two strategies for the higher order terms.

415 *Exhaustive:* If the number of all possible combinations of  $n$ -order terms is not  
 416 prohibitively large, these terms are included one by one in the current regression  
 417 matrix and the one that results in the biggest drop in the BIC is added. The procedure  
 418 is followed for all the remaining terms.

419 *Incremental:* Suppose that we have a number of  $n - 1$  order terms and we are  
 420 trying to investigate whether any  $n$  order terms would help improve the model. We  
 421 create an  $n$ -th order term by multiplying one  $n - 1$  order term with a first order  
 422 term that is already included in the model (i.e. active input). We test if adding  
 423 this new term in the model reduces the BIC and include it in the current regression  
 424 matrix if it does or discard it if it does not. We repeat the same procedure for all the  
 425 remaining combinations of  $n - 1$  and first order terms. This procedure allows finding  
 426 high order terms which can improve the model without having to test all possible  
 427 high order combinations of powers of  $\mathbf{x}$ , which very soon become so numerous that  
 428 their evaluation is prohibitive and can lead to serious overfitting concerns.

429 *Pruning:* Before looking into increasing the order of the terms that are included  
 430 in the model, we check whether some currently included terms could be removed.  
 431 We do this by removing one term at a time from the current model. If this removal  
 432 decreases the BIC, the respective term is removed from the current regression matrix.  
 433 This reduction can help build a more parsimonious model.

434 *Validation:* The testing and addition of a large number of polynomial terms, can  
 435 lead to a model that is overfitted. We guard against this with a leave one out vali-  
 436 dation. Once the final form of the regression matrix has been decided, one training  
 437 data point is left out and the regression coefficients are calculated using the remain-  
 438 ing ones. The prediction errors are then calculated and divided by  $\hat{\sigma}^2$ . The result  
 439 is compared to a normal distribution. If the errors do not deviate significantly from  
 440 a normal distribution the emulator is declared valid. If not, we remove the highest  
 441 order terms that are currently included in the model and repeat this until the emu-  
 442 lator validates. If the emulator cannot validate despite this model simplification, this  
 443 particular output is not considered in the current wave. This is an important strength  
 444 of history matching, as outputs that are more difficult to emulate can be left until  
 445 later waves, when they may be much easier to deal with.

446 **3.3. Sampling algorithm.** After a number of history matching waves, the non-  
 447 implausible space is typically a very tiny portion of the initial non-implausible space  
 448  $\mathcal{X}_0$  and it can be several orders of magnitude smaller than the *minimum enclosing*  
 449 *hyperrectangle*. We define the latter as the smallest hyperrectangle that encloses all  
 450 known non-implausible samples. The minimum enclosing hyperrectangle coincides  
 451 with  $\mathcal{X}_0$  at wave 0 and cannot increase from one wave to the next. The shape of  
 452 the enclosed non-implausible space is generally unknown and can only be described  
 453 by the collection of  $\mathbf{x}$ 's which satisfy the implausibility conditions for *all* emulators  
 454 across *all* waves. For simplicity we define an indicator function  $\mathcal{I}(\mathbf{x})$ , which takes the  
 455 value 1 if  $\mathbf{x}$  is non-implausible for all emulators across all waves and 0 otherwise.

456 For history matching to advance, it is necessary to have a large enough number  
 457 of samples for which  $\mathcal{I}(\mathbf{x}) = 1$ , such that they cover as much of the non-implausible  
 458 region as possible. Furthermore, we would like these samples to be uniformly dis-

459 tributed over the non-implausible region, as we have no reason to favour some of  
 460 its parts over others and for design reasons mentioned in the comments of section  
 461 3.1.3. In the following we describe a method for drawing uniform samples from such  
 462 spaces in an efficient manner, and is based on an adaptation of the slice sampler to  
 463 the specific requirements of history matching. We stress that the exact shape of the  
 464 non-implausible space cannot be described analytically and can only be known via  
 465 sampling.

466 The one dimensional slice sampler [13] works as follows: suppose we have a sample  
 467  $x_i$  from a distribution  $p(x)$  and we want to draw another sample from the same  
 468 distribution.  $p(x_i)$  is evaluated and a sample  $u$  is uniformly drawn from the interval  
 469  $[0, p(x_i)]$ . Left and right sampling limits  $x_l$  and  $x_r$  are proposed and are incrementally  
 470 expanded until  $p(x_l) < u$  and  $p(x_r) < u$  are satisfied. This is the ‘stepping out’ part  
 471 of the algorithm. In the ‘shrinking’ part of the algorithm, a sample  $x_{i+1}$  is uniformly  
 472 drawn between  $x_l$  and  $x_r$ . If  $p(x_{i+1}) > u$  then  $x_{i+1}$  is accepted as the next sample. If  
 473 not,  $x_l$  is set to  $x_{i+1}$  if  $x_{i+1} < x_i$  or  $x_r = x_{i+1}$  if  $x_{i+1} > x_i$  and the process is repeated  
 474 until  $p(x_{i+1}) > u$ .

475 History matching permits two simplifications to the algorithm sketched above.  
 476 The region of interest for each sample is known and is defined by the limits of the min-  
 477 imum enclosing hyperrectangle. This makes the ‘stepping out’ part of the algorithm  
 478 redundant, as we can always set  $x_l$  and  $x_r$  for each dimension (i.e. simulator input)  
 479 to those limits. The second simplification comes from the fact that we want to give  
 480 uniform weights to any point  $\{\mathbf{x} : \mathcal{I}(\mathbf{x}) = 1\}$ . We can therefore set  $p(x) = \text{const}$ . This  
 481 drops the need for calculating the uniform number  $u$  and the condition  $p(x_{i+1}) > u$   
 482 becomes simply  $\mathcal{I}(\mathbf{x}) = 1$ . That is, if the proposed sample is non-implausible it is  
 483 accepted and it is rejected otherwise.

484 The case described above refers to one input. Higher dimensions can be accom-  
 485 modated by updating each dimension sequentially. The new sample  $\mathbf{x}_{i+1}$  is accepted  
 486 when all dimensions have been updated. The algorithm is outlined in the following  
 487 **s0** Assume the existence of one  $\mathbf{x}$  such that  $\mathcal{I}(\mathbf{x}) = 1$  and a minimum enclosing  
 488 hyperrectangle with upper and lower limits for each dimension  $p$  denoted as  
 489  $x_{p,\max}$  and  $x_{p,\min}$  respectively.  $x_p$  is the  $p$ -th element of  $\mathbf{x}$ .

490 **s1** let  $\mathbf{x}' = \mathbf{x}$   
 491 **s2** for  $p = 1 : P$   
 492 **s3**     set  $x_l = x_{p,\min}$ ,  $x_r = x_{p,\max}$   
 493 **s4**     do  
 494 **s5**         set  $x'_p \sim \text{Unif}[x_l, x_r]$   
 495 **s6**         if  $\mathcal{I}(\mathbf{x}') = 0$   
 496 **s7**             if  $x'_p < x_p$ ,  $x_l = x'_p$  else  $x_r = x'_p$   
 497 **s8**         while  $\mathcal{I}(\mathbf{x}') = 0$   
 498 **s9**  $\mathbf{x}'$  is the new non-implausible sample. Store, set  $\mathbf{x} = \mathbf{x}'$  and go to **s2** for drawing  
 499 another sample.

500 Evaluation of the membership function  $\mathcal{I}(\mathbf{x}')$  typically requires calculating the  
 501 implausibility  $I(\mathbf{x}')$  using all the emulators of the current and all previous waves:  
 502 sample  $\mathbf{x}'$  is non-implausible at wave  $\eta$  if it is non-implausible for that and all the  
 503 waves that precede it. Although it may seem paradoxical that an emulator declaring  $\mathbf{x}$   
 504 as non-implausible can be ‘over-ruled’ by another one that says it is not, two examples  
 505 can demonstrate that this can actually happen: a wave 1 emulator will typically have  
 506 larger code uncertainty compared to a wave  $\eta$  emulator as the first is trained over  
 507 the entire input space  $\mathcal{X}_0$  and the latter over the smaller  $\mathcal{X}_{\eta-1}$ . As a result, a point  
 508 within  $\mathcal{X}_{\eta-1}$  is more likely to be rejected by the wave  $\eta$  emulator which is more certain

509 about its predictions. At the same time, a point that is outside  $\mathcal{X}_{\eta-1}$  was rejected by  
 510 definition by the wave 1 or a subsequent wave emulator (it would otherwise belong in  
 511  $\mathcal{X}_{\eta-1}$ ). This point might still be considered as non-implausible by a wave  $\eta$  emulator,  
 512 as this emulator was trained only over the  $\mathcal{X}_{\eta-1}$  region and its estimates outside this  
 513 are either very uncertain or not reliable. Therefore, determining whether  $\mathbf{x}'$  is non-  
 514 implausible normally requires the evaluation of all the emulators that have been built  
 515 so far.

516 In our case, this represents a worst case scenario, and fewer evaluations are re-  
 517 quired in practice. In the previous section, we mentioned that an emulator does not  
 518 include all inputs, but only those considered active for that particular output at that  
 519 particular wave. When the slice sampler proposes a move across the  $p$ -th dimension,  
 520 evaluation of  $\mathcal{I}(\mathbf{x}')$  requires invoking only the emulators that include  $p$  in their active  
 521 input list. The remaining emulators need not be used as their results remain un-  
 522 changed. Therefore, large computational savings can be achieved if, for every input,  
 523 a list of all emulators that include input  $p$  in their active input list is made, and only  
 524 these are evaluated when the slice sampler proposes a move across the  $p$ -th dimension.

525 Additional savings in computation time can arise if the emulators are ranked  
 526 according to the space they reject. That is, given a set of non-implausible samples  
 527 from the previous wave, we can rank all the emulators of the current wave according  
 528 to the proportion of samples they reject from higher to lower. When the membership  
 529 function  $\mathcal{I}(\mathbf{x})$  needs to be evaluated, the emulator with the highest rejection rate  
 530 is invoked first. This can lead to substantially fewer emulator evaluations, as the  
 531 ones invoked first have a greater probability of rejecting a sample and once a sample  
 532 is deemed implausible by any single emulator, there is no need to evaluate it any  
 533 further. Finally, invoking the emulators of different waves in reverse order (i.e. last  
 534 wave first), can also help speed up the evaluation of  $\mathcal{I}(\mathbf{x})$ , as the later wave emulators  
 535 should be more precise over the current non-implausible region.

536 The proposed method with the computational shortcuts described above is quite  
 537 efficient, requires no tuning (e.g. in contrast to the proposal kernel of the Metropo-  
 538 lis Hastings) and can successfully sample very small spaces. It is nevertheless, an  
 539 MCMC method, and it can still be affected by poor mixing, especially when inputs  
 540 are highly correlated. The quality of the mixing therefore needs to be evaluated after  
 541 the sampling is completed and the chains should be thinned if the mixing is found to  
 542 be poor.

543 Additionally, the slice sampler can capture some disconnected regions, but not  
 544 all. As the inputs are updated on a one-by-one basis, the disconnected regions would  
 545 need to have overlapping projections in all but one input dimensions for a jump to  
 546 be possible. Starting the algorithm from a large number of non-implausible samples  
 547 reduces the probability that a disconnected region is not sampled.

548 In a typical application of history matching, we have a few thousand non-implausible  
 549 samples at each wave. In this case, a large number of parallel chains can be run, each  
 550 one initialised from a non-implausible sample. The availability of a multi-core ma-  
 551 chine or a multi-node high performance computing cluster can significantly speed up  
 552 the sampling process. The easy parallelisation of this method increases further its  
 553 efficiency and improves the handling of input space features such as disconnected  
 554 regions.

555 **4. Results.** The epidemiological simulator we study has 96 inputs and 50 out-  
 556 puts. A full list of the inputs, along with the initial non-implausible ranges is given  
 557 in the supplementary material, as well as a full list of the simulator outputs with the

558 observation data and their ranges. The range for the latter is assumed to represent  
 559  $\pm 2(V_m + V_o)^{1/2}$ , that is, 4 standard deviations calculated from the sum of the obser-  
 560 vation and the model error. 3000 training points were chosen for the first four waves  
 561 to help explore the simulator’s behaviour, but these were reduced to 1000 from wave 5  
 562 onwards, as the resulting emulators were still successful in rejecting input space. The  
 563 simulator was run for  $K = 30$  times at each design point to allow the estimation of its  
 564 mean output value. The goal of history matching was to find input parameter values  
 565 that would lead to mean outputs (as opposed to individual runs) that fell within the  
 566 observation ranges. A total of 13 waves were carried out.

567 **4.1. Output matching.** Figure 1 shows 10 simulator outputs which come from  
 568 two different time series: the proportion of HIV positive people on ART and the pro-  
 569 portion of people starting ART with low CD4 counts ( $< 250$  cells/ $\mu$ l). The observation  
 570 ranges are shown using black bars and the simulator’s output at four different waves  
 571 (1, 4, 8 and 14) are shown using darkening shades of blue. The figure demonstrates  
 572 that as the input space shrinks, the simulator’s output converges to the observations.

573 Figure 2 shows histograms of simulator outputs for 5 different output pairs across  
 574 waves 1, 8 and 14. The observation ranges are shown with black rectangles. The  
 575 figure shows again the convergence of the simulator outputs towards the observations  
 576 as waves progress. It is interesting to note that in the first wave (1st column of  
 577 Figure 2), outputs 17, 18, 45 and 51 are completely off target. In the final wave, the  
 578 majority of the simulator’s outputs are within the targets, with output 26 only being  
 579 slightly off. Incidentally, this was the output with the poorest matching among all  
 580 50. Histograms of all simulator outputs at wave 14 with their observation ranges are  
 581 given in the supplementary material.

582 Apart from convergence to the observations, Figure 2 also reveals correlations  
 583 between the outputs, which are interesting especially in wave 14 (rightmost column).  
 584 The top right panel shows that there is a strong positive correlation ( $r=0.84$ ) between  
 585 the proportion of HIV negative women and the proportion of HIV positive women who  
 586 have ever been tested in 2011. This reflects two factors. The first is the way that HIV  
 587 testing rates were controlled in the simulator. One input parameter controlled the  
 588 absolute rate of testing in HIV negative women. Another controlled the rate of testing  
 589 in HIV positive women, *relative* to the rate in HIV negative women. This introduced  
 590 a correlation between the two outputs. The second factor is that HIV negative women  
 591 could become HIV positive through transmission of the virus. Women who were tested  
 592 for HIV when they were still uninfected would remain ‘ever tested’ after becoming  
 593 HIV positive. This introduced a further correlation between the two outputs.

594 In Uganda, the numbers of women starting ART each year are higher than the  
 595 numbers of men starting. This is due both to the higher prevalence of HIV in women,  
 596 and due to the fact that they are more likely to be diagnosed (e.g. through an-  
 597 tenatal HIV testing). From 2013, a change in policy meant that all HIV positive  
 598 pregnant women became eligible for treatment, regardless of how far their disease  
 599 had progressed. This increased the numbers of women starting ART. We therefore  
 600 calibrate the simulator to the proportion of people newly starting ART in 2010 who  
 601 were women (output 25) and the increase in the proportion of new starts who were  
 602 women between 2010 and 2014 (output 26). There was a clear negative correlation  
 603 between the two outputs as shown in Figure 2, ( $r = 0.55$ ). This is because if both  
 604 outputs had high values, a very high proportion of people starting ART in 2014 would  
 605 be women. To achieve this, very few men would be able to start ART in 2013, and the  
 606 large increase in ART coverage between 2011 and 2013 could not be achieved. The

607 proportion of people newly starting ART in 2010 who were women (output 25) was  
 608 also weakly positively correlated with the HIV prevalence in women in 2011 (output  
 609 18) (see Figure 2,  $r=0.21$ ), and weakly negatively correlated with the HIV prevalence  
 610 in men in 2011 (output 17, see Figure 2,  $r=-0.21$ ). This reflects the fact that, all else  
 611 being equal, the proportion of people starting ART who are women will be higher  
 612 when HIV prevalence in women is high relative to the prevalence in men.

613 Finally, there is a positive correlation between ART coverage in 2013 (output 31)  
 614 and the proportion of people who remain on ART 12 months after first starting it  
 615 (output 51) (see Figure 2,  $r=0.22$ ). This is because ART coverage will fall as people  
 616 stop taking it. The correlation is only weak however, as the number of people newly  
 617 starting ART has a larger effect on overall coverage than the rate at which people  
 618 drop out. It can therefore be seen that the results of history matching can give useful  
 619 insight into the model’s structure.

620 Figures 1, 2 and the histograms on the online material offer evidence that the  
 621 final simulator runs match the observations. We can quantify this evidence using the  
 622 *simulator run implausibility* [26, 2], which quantifies how close an *actual* simulator  
 623 run is to the observations. For the  $r$ -th output we define this measure as

$$624 \quad (8) \quad I_{\mathcal{R},r}(\mathbf{x}) = \frac{|z_r - \hat{g}_r(\mathbf{x})|}{(V_{o,r} + V_{m,r} + \hat{s}^2(\mathbf{x})/K)^{1/2}},$$

625 where,  $\hat{g}(\mathbf{x})$  is an estimate of the simulator’s mean and  $\hat{s}^2(\mathbf{x})$  an estimate of its variance  
 626 evaluated at  $\mathbf{x}$  and calculated using actual simulator runs. The rest of the terms were  
 627 defined in section 3.1. Equation 8 is similar to the implausibility measure defined in  
 628 section 3.1.2. The difference is that this term does not involve any emulators and is  
 629 a metric that quantifies how close the mean of the  $r$ th simulator’s output is to the  
 630 observations for a particular  $\mathbf{x}$ . Also, equation 8 is not part of the history matching  
 631 algorithm, but it is just a convenient way of evaluating the closeness of the simulator’s  
 632 outputs to the observation data.

633 The simulator was evaluated 30 times at 22000 different non-implausible inputs at  
 634 wave 14. The measure in equation 8 was evaluated for each of the 50 simulator outputs  
 635 and each of the 22000 runs. Figure 3 shows the percentage of those runs that had  
 636  $I_{\mathcal{R},r}(\mathbf{x}) < 2$ . This can be interpreted as runs that would fall within the observation  
 637 ranges roughly 95% of the time if the distribution of the individual simulator runs  
 638 (repetitions) for a fixed  $\mathbf{x}$  followed a normal distribution. The results show that for  
 639 half of the outputs, more than 95% of the 22000 runs had  $I_{\mathcal{R},r}(\mathbf{x}) < 2$ . This percentage  
 640 was higher than 80% for 44 out of the 50 outputs. For 6 outputs the scores were as  
 641 follows: 14: 49%, 15: 43%, 16: 43%, 17: 54%, 18: 59% and 26: 69%. This means  
 642 that although history matching indicated that all these outputs would fall within or  
 643 just outside the observation ranges, this was true only between 40-60% of the time for  
 644 five outputs and around 70% for the sixth. These outputs were hard to emulate using  
 645 linear models, in the sense that the prediction uncertainty  $V_c$  would not drop beyond  
 646 a certain magnitude, which was comparable to that of the other two error terms  $V_m$   
 647 and  $V_o$ . The difficulty in emulating these particular outputs is not surprising, as the  
 648 majority of these outputs represented male and female HIV prevalences at different  
 649 time points, which are outputs that depend on a large number of inputs and their  
 650 interactions. Output 26 was highly stochastic (i.e. the samples of individual simulator  
 651 runs had a large variance), which implies that more simulator evaluations per design  
 652 point would be needed to increase the accuracy in the estimation of the means, and  
 653 the subsequent emulation.

654 Not being able to emulate an output is not fatal for history matching. It simply  
 655 means that fewer simulator runs will be close to the observations than the emulators  
 656 predicted. If the emulators have been set up and validated correctly, an inaccurate  
 657 emulator should still not miss the good runs if its uncertainty properly covers the  
 658 training and validation data. In other words, an emulator that is very uncertain will  
 659 be unable to rule out regions of input space that actually contain bad matches, but  
 660 should not incorrectly rule out regions containing matches that are acceptable.

661 **4.2. Input space shrinkage.** In this section we examine the shrinking of the in-  
 662 put space during the course of waves and present the main epidemiological conclusions  
 663 that were extracted. The minimum enclosing hyperrectangle at wave 13 was  $10^{-33}$   
 664 times smaller than the initial non implausible space  $\mathcal{X}_0$ . A very small number, which  
 665 however arises from the large number of inputs and the multiple constraints imposed  
 666 by the simulator’s outputs. Even within this hyperrectangle however, a tiny propor-  
 667 tion of points was non-implausible. The calculated volume of the non-implausible  
 668 space was  $\approx 10^{-45}$  times smaller than  $\mathcal{X}_0$ . That is only 1 point in  $10^{12}$  (one in a tril-  
 669 lion) is non-implausible if selected at random between the narrowest limits suggested  
 670 by the last wave’s non-implausible samples. The ratio of the volumes of the final non-  
 671 implausible space and  $\mathcal{X}_0$  are shown in Table 1. Note that the non-implausible region  
 672 at wave 13 is substantially smaller than those found in previous history matching  
 673 applications in the literature.

674 The range of 5 out of 96 inputs was reduced to less than 1% of the original,  
 675 while for around 25 it decreased to less than 50%. For approximately 50 inputs, the  
 676 range remained similar to the original. These inputs either do not substantially af-  
 677 fect the history matched outputs of the simulator, or there are combinations of these  
 678 inputs with others that are implausible but cannot be visualised in the 1 dimensional  
 679 projection of the non-implausible space that these histograms represent. The supple-  
 680 mentary material includes histograms of the non-implausible samples at wave 13 for  
 681 all 96 inputs, which demonstrate the overall input space reduction.

682 We now focus our attention on a small set of inputs, track their shrinkage through  
 683 the waves and draw some conclusions based on their correlation patterns. The lower  
 684 triangle of the lattice in Figure 4 shows scatter plots of non-implausible samples for  
 685 pairs of inputs across 4 waves. The light blue colour is the initial non-implausible  
 686 region and waves 1,4,8 and 13 are shown in darkening shades of blue. The baseline  
 687 transmission [55] range is reduced to a third as early as wave 2 and by wave 5 it is  
 688 down to 10% of the original range. Similar conclusions can be drawn about the other  
 689 inputs. The upper triangle of the lattice shows 2 dimensional histograms of wave 13  
 690 non-implausible points as a function of pairs of inputs. The colour scale represents  
 691 the  $\log_{10}$  probability of finding a non-implausible sample if we fix the values of the  
 692 inputs that lie across the axes to a particular value, and choose the rest of the inputs  
 693 randomly. For example, the red region in the panel that corresponds to inputs 75  
 694 and 55 means that fixing these inputs to those values and varying the others freely,  
 695 gives us a  $10^{-43}$  probability of finding a non-implausible sample. The grey area of  
 696 these plots indicates that for those values of the respective inputs, no samples that  
 697 match the simulator’s output to the observations were found. All the axes in Figure  
 698 4, correspond to the initial range of the inputs, as is shown in the supplementary  
 699 material.

700 Figure 5 is a zoomed-in version of some of the histograms shown in Figure 4 to  
 701 allow for a more detailed analysis of the correlation patterns. The left panel of Figure  
 702 5 shows a histogram of non-implausible samples for the baseline HIV transmission

703 probability (Input 55) and the rate at which men in one of the two sexual behaviour  
 704 risk groups form new partnerships during a particular time period (Input 75). The  
 705 overall range of input 55 is constrained to between 0.0005 and 0.0027, despite the  
 706 broad initial plausible range of 0-1. This final range is consistent with empirical data  
 707 from Uganda, which estimated the per-sex-act transmission probability to be 0.0011  
 708 (95% CI 0.00080-0.0015) [30]. The plausible range for the contact rate (input 75) in  
 709 the final wave was constrained to be between 0 and 0.3, a large reduction from its  
 710 original plausible range of 0-1. There is a clear negative correlation between the two  
 711 inputs ( $r = -0.36$ ), demonstrating that fits are unlikely to be found when the rate  
 712 at which new partnerships form (and therefore the amount of sex occurring in the  
 713 model) and the per-sex-act transmission probability are both high or both low.

714 The middle panel of Figure 5 shows the final wave distribution of the baseline  
 715 transmission probability (Input 55) against the relative increase in transmission prob-  
 716 ability for people with low CD4 counts (advanced infection) (Input 58). Unlike the  
 717 baseline transmission probability, the overall range of the latter input parameter did  
 718 not change during calibration, indicating that model fits can be found throughout the  
 719 initial plausible range. The figure shows that there is a negative correlation between  
 720 the two input parameters. This occurred as, all else being equal, increasing the value  
 721 of one parameter and simultaneously decreasing the value of the other will result in  
 722 similar overall levels of HIV transmission in the model.

723 Finally, the right panel of the same figure shows the final wave distribution of the  
 724 proportion of (low risk) men who were able to be in more than one partnership at  
 725 the same time (Input 70), against the associated concurrency input parameter (Input  
 726 66). The purpose of the concurrency parameter was to influence how likely it was that  
 727 these men who *could* form additional partnerships *would actually* do so. The graph  
 728 shows that model fits were unlikely to be found when both the proportion of men  
 729 who could form additional partnerships was low, and when it was not very likely that  
 730 men who could form additional partnerships would do so. This is because the model  
 731 was calibrated to sexual behaviour data from Uganda that indicates that around 9%  
 732 of men aged 15-49 have more than 1 ongoing partnership at any point in time [15].

733 **4.3. The case for linear models.** Gaussian processes have been extensively  
 734 used for building computer model emulators in the context of history matching and  
 735 beyond. Gaussian processes are very flexible statistical models, but at the same  
 736 time more complex and less universal and understood than the ubiquitous linear  
 737 regression, that we employ in this work. We make here the case that linear models  
 738 can be useful in history matching and they can go a long way into calibrating high  
 739 dimensional simulators. Their simplicity and widespread usage can also have some  
 740 advantages over GPs. Moreover there is no binary decision that has to be made (i.e.  
 741 use linear regression or GPs) as both statistical models can be used in the same history  
 742 match. For example, linear regression can be used at the initial waves, if it is found  
 743 to efficiently reject large portions of the input space, and GP based emulators can be  
 744 introduced at later stages if linear regression fails to provide emulators of sufficient  
 745 accuracy to reduce space further.

746 As an example, we show results from two emulators built for output 15 at wave  
 747 7 using 1000 training points. The first is a linear regression emulator containing 33  
 748 terms up to third order. The second is a GP based emulator, with a 3rd order poly-  
 749 nomial mean function and the Matérn correlation function. The GP correlation function  
 750 parameters (correlation lengths) were estimated from the data by maximising their  
 751 likelihood. The GP's mean function parameters (regression coefficients) were inte-

752 grated out. For more details on this type of GP emulators, the reader can consult  
 753 [2]. Estimating the correlation lengths on the GP emulator took a little more than an  
 754 hour while building the linear regression model required a few seconds. The compu-  
 755 tational load in the estimation of the correlation lengths was due to the optimisation  
 756 algorithm that looked for a mode in a 96-dimensional likelihood function (i.e. one  
 757 dimension per simulator input). Moreover, it is possible that the optimisation al-  
 758 gorithm found a suboptimal mode as the likelihood is almost certainly multimodal.  
 759 Finding a good mode among several would increase the computational load as a more  
 760 detailed exploration of the likelihood surface would be required.

761 Both the linear regression and the GP based emulators were then used to predict  
 762 the simulator’s output for the wave 8 runs. Histograms of the standardised errors  
 763 (i.e. the difference of the simulator’s output with the emulator’s prediction, divided  
 764 by the emulator’s standard deviation for the prediction, [4]) are given in Figure 6.  
 765 The standardised errors take values mostly in the region  $[-2, 2]$ , an indication that  
 766 both emulators are valid. The GP based emulator however, resulted in larger code un-  
 767 certainty (the  $V_{c,r}$  term in Equation 5) when evaluated at the wave 8 non-implausible  
 768 samples. As a result, the calculated implausibility was smaller and it rejected 7% of  
 769 the wave 8 non-implausible samples compared to 15% for the linear regression based  
 770 emulator. We should also note here that calculating the implausibility for  $\approx 20000$   
 771 wave 8 non-implausible samples took a few milliseconds for the linear regression em-  
 772 ulator and around 15 seconds for the GP based emulator. This is because the GP  
 773 based emulator needs to create a  $N \times Np$  correlation matrix, where  $N = 1000$  are the  
 774 training and  $Np = 20000$  were the testing points.

775 As a second example, we used GPs to model the residual between the linear  
 776 model’s predictions and the simulator’s outputs in wave 13. That is, equation 6 is  
 777 now changed to

$$778 \quad g(\mathbf{x}) = \sum_{i=1}^q h_i(\mathbf{x})\beta_i + \eta(\mathbf{x})$$

779 where  $\eta(\mathbf{x})$  is a zero mean Gaussian process, instead of the uncorrelated noise error  
 780 term  $\epsilon$  of equation 6. The rationale here is that some correlation must still exist in  
 781 the linear regression model’s residual  $\epsilon$  and capturing this with a GP should reduce  
 782 the overall uncertainty. Indeed, modelling the residual with a GP resulted in a 22%  
 783 further shrinkage of the non-implausible space compared to using the linear regression  
 784 models alone. This example shows that the two models can be combined within history  
 785 matching to increase the rejection rate at the expense of the additional computational  
 786 cost of training a GP.

787 Gaussian processes are clearly more flexible models and will outperform linear  
 788 regression in low dimensional regression problems. In high dimensions however, it  
 789 is very difficult to have a sufficient number of training points such that the GP can  
 790 accurately describe the simulator’s response surface. As a result, the performance  
 791 gap between the GPs and the less flexible linear regression becomes smaller. This  
 792 theoretical argument can support to some extent the usefulness of linear regression  
 793 based emulators in history matching of large models. Furthermore, from our experi-  
 794 ence, a major stumbling block in the adoption of history matching by practitioners  
 795 has been the requirement to understand and implement a GP-based regression model.  
 796 Demonstrating that history matching can be carried out using a much simpler and  
 797 better understood model such as linear regression can increase its adoption as a useful  
 798 tool to analyse and calibrate complex models.

799 **4.4. The sampling algorithm.** In this section we evaluate the performance  
800 of the sampling algorithm and compare it with a simple Metropolis-Hastings (MH)  
801 sampling scheme. The MH algorithm uses a transition kernel that is a zero-mean  
802 multivariate normal with a covariance matrix estimated from a 1000 non-implausible  
803 samples of the current wave, scaled to result in an acceptance rate of approximately  
804 25%. The target distribution was uniform defined over all non-implausible space. Per  
805 sample, the slice sampler requires roughly  $P = 96$  times more emulator evaluations  
806 because the inputs are updated sequentially. To allow for a fair comparison the  
807 Markov chains in the MH algorithm are run longer, such that the emulator evaluations  
808 between the two algorithms are roughly equal. The results are evaluated using the  
809 effective sample size (ESS) for each chain and input, which was calculated with the  
810 `effectiveSize` function from the R package CODA [20]. This function provides  
811 an estimate of the number of samples that can be considered uncorrelated from a  
812 Markov chain. Both algorithms were compared at waves 4, 7, 11 and 13 using 1000  
813 non-implausible samples as starting points, i.e. 1000 chains were run for each case.  
814 The ESS scores were averaged across the 1000 chains for each input.

815 Figure 7 shows the averaged ESS for the MH and the slice sampling algorithms  
816 for waves 4, 7, 11 and 13. The effective sample sizes are sorted in increasing order to  
817 facilitate the comparison. The figure demonstrates that, in general, the slice sampler  
818 results in samples that are less correlated for the same amount of computational effort,  
819 often by a very large margin. The only exceptions are inputs 55, 58 and 1 in wave  
820 13 and input 55 in wave 11. The lowest ESS score in wave 13 for the slice sampler  
821 was 5 and for the Metropolis-Hastings was 15. For wave 11, the worst components in  
822 both cases had an ESS of around 10. The low ESS scores of the slice sampler were  
823 most likely due to the fact that these particular inputs were very correlated and the  
824 correlation information, which was provided to the MH algorithm, was not available  
825 to the slice sampler. Providing the slice sampler with this information or using an  
826 extension such as [18], could help improve mixing. For the inputs that were affected  
827 the most, we have tried to mitigate the low ESS scores by drawing large numbers of  
828 samples and visually verifying that the sampled inputs spanned the entire range of  
829 non-implausible samples. Overall, the proposed sampling algorithm gave reasonably  
830 good results with the additional benefit of requiring no tuning or manual intervention,  
831 while being trivial to implement once the implausibility function is coded up.

832 **5. Conclusion.** History matching is a (pre-)calibration method capable of find-  
833 ing parts of a simulator’s input space that are likely to match the observations. We  
834 have applied this method to a simulator that is larger than any other that history  
835 matching has been applied to before. This scaling up was facilitated by the use of  
836 linear regression models as emulators and a sampling algorithm that was capable of  
837 sampling high dimensional and very small non-implausible spaces.

838 The calibrated simulator was an HIV stochastic individual based model with 96  
839 inputs and 50 outputs. The simulator’s input space was reduced by a factor of  $10^{-45}$   
840 after 13 waves of history matching. In the final wave, the majority of outputs had a  
841 more than a 90% chance of falling within the observation ranges when the simulator  
842 was run at inputs suggested by history matching. Despite the high success for each  
843 individual output, getting a simulator run that would match all the observations was  
844 relatively rare (around 5 in a 1000). However, considering the size of the problem and  
845 the number of outputs, this was still an acceptance proportion that was considered  
846 useful for the epidemiologists using the simulator. The simulator could be evaluated  
847 approximately 20000 times per day on a high performance computing cluster, and

848 despite the low overall acceptance rate it was possible to obtain a few hundred runs  
849 that simultaneously matched all the observations in reasonable time.

850 These runs are fed into a number of other research projects, such as [14], that make  
851 predictions about the trajectory of HIV in the next 10-15 years, taking into account the  
852 uncertainty that is introduced by the simulator’s unknown input parameter values.  
853 In the past, when making predictions using simulators of this scale, a single input  
854 parameter set that would fit the historical data was used, typically found by hand  
855 using prior knowledge from the model developer. This approach did not explicitly  
856 acknowledge the fact that the simulator’s input parameters are uncertain quantities -  
857 an uncertainty that was left out of any predictions. The methodology presented here  
858 comes to address this point. History matching provided us with a few hundred input  
859 parameter values, from different parts of the input space, that matched the calibration  
860 data, which were then used to run the simulator up to 2030 under 27 different ART  
861 scale-up interventions. It therefore offered a method of quantifying the effect of the  
862 uncertainty about the input parameter values, on the predicted outcome of the ART  
863 interventions.

864 Apart from the large numbers of input values that fit the observations, history  
865 matching also provided insights into the simulator’s structure. The active input se-  
866 lection methodology revealed the inputs that influenced an output the most, and  
867 reductions in the non-implausible space showed which inputs were affected by the  
868 constraints imposed by the observations. Both of these features are very useful in  
869 analysing simulators of this scale. The correlation patterns that emerged between in-  
870 puts and between outputs highlighted the existence of various structures and processes  
871 in the simulator. This information can be used to understand the internal workings of  
872 a simulator, or indeed, verify that everything works as intended, knowledge that could  
873 lead to the discovery of simulator coding errors, suggest ways in which the simulator  
874 can be improved, or even help derive appropriate model discrepancy terms in case the  
875 simulator is not capable of matching the observations.

876 Methodologically, a key feature of this work was the use of linear regression models  
877 for building emulators, instead of the GPs that were typically being used in previous  
878 history matching applications. Even though linear regression models are less flexible  
879 than GPs, they are generally easier to fit, interpret and implement. Despite their  
880 simplicity, they did cover a lot of ground towards calibrating a very complex simulator.  
881 This was also facilitated by the history matching philosophy, which does not require  
882 an emulator to describe the simulator everywhere in great precision: as long as the  
883 simulator runs fall within the uncertainty bounds of the emulator (i.e. the code  
884 uncertainty  $V_c$  is correctly specified), history matching can proceed. See [27] for  
885 further discussions on the topic.

886 Using a more advanced statistical model for building an emulator can generally  
887 reduce the code uncertainty. It is possible however, especially in the first waves of  
888 a history match, that the simulator’s outputs  $g(\mathbf{x})$  are so far from the observations  
889  $z$ , such that a moderate reduction in the code uncertainty  $V_c$  (Equation 5) will not  
890 have an appreciable effect in reducing the input space further. In later waves, when  
891  $E^*[g(\mathbf{x})]$  and  $z$  converge, the effort of building a more sophisticated emulator with  
892 smaller  $V_c$  could pay dividends. We tried this at the last wave of our history match  
893 and indeed the GP based emulator resulted in a further shrinkage of the input space.  
894 Hence, we do not try to argue against the use of GPs in building emulators for history  
895 matching, but note that linear regression models offer an alternative that is faster and  
896 more straightforward to implement.

897 The availability of a large number of non-implausible samples is critical in the

898 application of history matching. Sampling the non-implausible space can be challeng-  
 899 ing as this is high dimensional and can be quite small. A simple MCMC algorithm  
 900 that tackles this problem was proposed in this work, that was simple to implement,  
 901 requiring virtually no tuning and was successful in drawing uniform samples from  
 902 very small non-implausible spaces. The correlation between some inputs meant that  
 903 the mixing was slightly poor for a small number of inputs, something that could be  
 904 addressed using block updating.

905 In conclusion, the effectiveness and simplicity of the history matching method  
 906 presented here shows that it is a useful tool for the calibration of computationally  
 907 expensive, high dimensional individual based models.

908

## REFERENCES

- 909 [1] I. ANDRIANAKIS AND P. CHALLENGOR, *The effect of the nugget on Gaussian process emulators*  
 910 *of computer models*, Computational Statistics & Data Analysis, 56 (2012), pp. 4215–4228.
- 911 [2] I. ANDRIANAKIS, I. VERNON, N. MCCREESH, T. J. MCKINLEY, J. E. OAKLEY, R. NSUBUGA,  
 912 M. GOLDSTEIN, AND R. G. WHITE, *Bayesian history matching and calibration of complex*  
 913 *infectious disease models using emulation: a tutorial and a case study on HIV in Uganda*,  
 914 PLoS Computational Biology, 11 (2015), pp. 1–18.
- 915 [3] C. ANDRIEU, A. DOUCET, AND R. HOLENSTEIN, *Particle Markov chain Monte Carlo methods*,  
 916 Journal of the Royal Statistical Society, Series B (Methodological), 72 (2010), pp. 269–342.
- 917 [4] L. S. BASTOS AND A. O’HAGAN, *Diagnostics for Gaussian process emulators*, Technometrics,  
 918 51 (2009), pp. 425–438.
- 919 [5] J. BRYNJARSDOTTIR AND A. O’HAGAN, *Learning about physical parameters: The importance*  
 920 *of model discrepancy.*, tech. report, <http://www.tonyogahan.co.uk/academic/pub.html>, 10  
 921 2010.
- 922 [6] P. S. CRAIG, M. GOLDSTEIN, A. H. SEHEULT, AND J. A. SMITH, *Pressure matching for hydro-*  
 923 *carbon reservoirs: a case study in the use of Bayes linear strategies for large computer*  
 924 *experiments*, (with discussion) in Case Studies in Bayesian Statistics, eds. C. Gastonis et  
 925 al. Springer-Verlag, III (1997), pp. 37–93.
- 926 [7] G. J. GIBSON AND E. RENSHAW, *Estimating parameters in stochastic compartmental models using*  
 927 *Markov chain methods*, IMA Journal of Mathematics Applied in Medicine and Biology,  
 928 15 (1998), pp. 19–40.
- 929 [8] M. GOLDSTEIN AND J. ROUGIER, *Reified Bayesian modelling and inference for physical systems*,  
 930 Journal of Statistical Planning and Inference, 139 (2009), pp. 1221–1239.
- 931 [9] M. GOLDSTEIN, A. SEHEULT, AND I. VERNON, *Assessing model adequacy*, in Environmental  
 932 Modelling: Finding Simplicity in Complexity, Second Edition, J. Wainwright and M. Mul-  
 933 ligan, eds., Wiley-Blackwell, John Wiley & Sons, Ltd, Chichester, UK, 2013.
- 934 [10] D. R. JONES, M. SCHONLAU, AND W. J. WELCH, *Efficient global optimization of expensive*  
 935 *black-box functions*, Journal of Global Optimization, 13 (1998), pp. 455–492.
- 936 [11] M. C. KENNEDY AND A. O’HAGAN, *Bayesian calibration of computer models*, Journal of the  
 937 Royal Statistical Society. Series B, 63 (2001), pp. 425–464.
- 938 [12] J. KNOWLES, *A hybrid algorithm with on-line landscape approximation for expensive multiob-*  
 939 *jective optimization problems*, IEEE Transactions on Evolutionary Computation, 10 (2005),  
 940 pp. 50–66.
- 941 [13] D. J. C. MACKAY, *Information Theory, Inference, and Learning Algorithms*, Cam-  
 942 bridge University Press, 2003, <http://www.cambridge.org/0521642981>. Available from  
 943 <http://www.inference.phy.cam.ac.uk/mackay/itila/>.
- 944 [14] N. MCCREESH, I. ANDRIANAKIS, , R. N. NSUBUGA, M. STRONG, I. VERNON, T. MCKINLEY,  
 945 J. E. OAKLEY, M. GOLDSTEIN, R. HAYES, AND R. G. WHITE, *Universal test, treat, and*  
 946 *keep: improving ART retention is key in cost-effective HIV control in Uganda*, PLOS  
 947 Medicine, Submitted for publication, (July 2016).
- 948 [15] N. MCCREESH, K. O’BRIEN, R. N. NSUBUGA, L. A. SHAFER, R. BAKKER, J. SEELEY, R. J.  
 949 HAYES, AND R. G. WHITE, *Exploring the potential impact of a reduction in partnership*  
 950 *concurrency on HIV incidence in rural Uganda: a modeling study*, Sexually Transmitted  
 951 diseases, 39 (2012), pp. 407–413.
- 952 [16] M. D. MCKAY, R. J. BECKMAN, AND W. J. CONOVER, *A comparison of three methods for*  
 953 *selecting values of input variables in the analysis of output from a computer code*, Tech-  
 954 nometrics, 21 (1979), pp. 239–245.

- 955 [17] T. J. MCKINLEY, A. R. COOK, AND R. DEARDON, *Inference in epidemic models without like-*  
956 *lihoods*, The International Journal of Biostatistics, 5 (2009).
- 957 [18] I. MURRAY, R. P. ADAMS, AND D. J. MACKAY, *Elliptical slice sampling*, JMLR: W&CP, 9  
958 (2010), pp. 541–548.
- 959 [19] P. D. O’NEILL AND G. O. ROBERTS, *Bayesian inference for partially observed stochastic epi-*  
960 *demics*, Journal of the Royal Statistical Society. Series A (General), 162 (1999), pp. 121–  
961 129.
- 962 [20] M. PLUMMER, N. BEST, K. COWLES, AND K. VINES, *Coda: Convergence diagnosis and out-*  
963 *put analysis for MCMC*, R News, 6 (2006), pp. 7–11, [http://CRAN.R-project.org/doc/](http://CRAN.R-project.org/doc/Rnews/)  
964 [Rnews/](http://CRAN.R-project.org/doc/Rnews/).
- 965 [21] F. PUKELSHEIM, *The three sigma rule*, The American Statistician, 48 (1994), pp. 88–91.
- 966 [22] G. SCHWARZ, *Estimating the dimension of a model*, The annals of statistics, 6 (1978), pp. 461–  
967 464.
- 968 [23] T. TONI, D. WELCH, N. STRELKOWA, A. IPSEN, AND M. P. H. STRUMPF, *Approximate Bayesian*  
969 *computation scheme for parameter inference and model selection in dynamical systems*,  
970 Journal of the Royal Society Interface, 6 (2009), pp. 187–202.
- 971 [24] J. UNITED NATIONS PROGRAMME ON HIV/AIDS, *The gap report*, Geneva, UNAIDS, (2014).
- 972 [25] I. VERNON AND M. GOLDSTEIN, *A Bayes linear approach to systems biology*, tech. report,  
973 MUCM Technical Report, 10 2010.
- 974 [26] I. VERNON, M. GOLDSTEIN, AND R. G. BOWER, *Galaxy formation: a Bayesian uncertainty*  
975 *analysis*, Bayesian Analysis, 5 (2010), pp. 619–670.
- 976 [27] I. VERNON, M. GOLDSTEIN, AND R. G. BOWER, *Rejoinder for galaxy formation: a Bayesian*  
977 *uncertainty analysis*, Bayesian analysis, 5 (2010), pp. 697–708.
- 978 [28] I. VERNON, M. GOLDSTEIN, AND R. G. BOWER, *Galaxy formation: Bayesian history matching*  
979 *for the observable universe*, Statistical science, 29 (2014), pp. 81–90.
- 980 [29] I. VERNON, J. LIU, M. GOLDSTEIN, J. ROWE, J. TOPPING, AND K. LINDSEY, *Bayesian uncer-*  
981 *tainty analysis for complex systems biology models: emulation, global parameter searches*  
982 *and evaluation of gene functions.*, BMC Systems Biology in submission, (2016).
- 983 [30] M. WAWER ET AL., *Rates of HIV-1 transmission per coital act, by stage of HIV-1 infection, in*  
984 *Rakai, Uganda*, J. Infectious Diseases, 191(9) (2005), pp. 1403–9.
- 985 [31] U. WILENSKY, *Netlogo*, Center for connected learning and computer-based modelling, North-  
986 western university, Evanston, IL, (2009).
- 987 [32] D. WILLIAMSON, M. GOLDSTEIN, L. ALLISON, A. BLAKER, P. CHALLENGER, L. JACKSON, AND  
988 K. YAMAZAKI, *History matching for exploring and reducing climate model parameter space*  
989 *using observations and a large perturbed physics ensemble*, Climate Dynamics, 41 (2013),  
990 pp. 1703–1729.
- 991 [33] D. WILLIAMSON AND I. VERNON, *Efficient uniform designs for multi-wave computer experi-*  
992 *ments*. in revision, arXiv:1309.3520, 2013.
- 993 [34] D. P. WIPF AND S. S. NAGARAJAN, *A new view of automatic relevance determination*, in  
994 Advances in Neural Information Processing Systems 20, J. C. Platt, D. Koller, Y. Singer,  
995 and S. T. Roweis, eds., Curran Associates, Inc., 2008, pp. 1625–1632.

Wave 1	$1.8 \cdot 10^{-05}$	Wave 8	$5.2 \cdot 10^{-30}$
Wave 2	$2.6 \cdot 10^{-08}$	Wave 9	$4.5 \cdot 10^{-33}$
Wave 3	$1.6 \cdot 10^{-09}$	Wave 10	$1.2 \cdot 10^{-35}$
Wave 4	$1.7 \cdot 10^{-10}$	Wave 11	$2.9 \cdot 10^{-37}$
Wave 5	$5.2 \cdot 10^{-14}$	Wave 12	$2.9 \cdot 10^{-41}$
Wave 6	$7.7 \cdot 10^{-20}$	Wave 13	$2.4 \cdot 10^{-45}$
Wave 7	$1.1 \cdot 10^{-24}$		

TABLE 1

Ratio of the non-implausible space volume at each wave compared to the initial non-implausible space  $\mathcal{X}_0$ . This table also expresses the probability of finding a non-implausible sample at wave  $n$  if we randomly draw samples from  $\mathcal{X}_0$ .

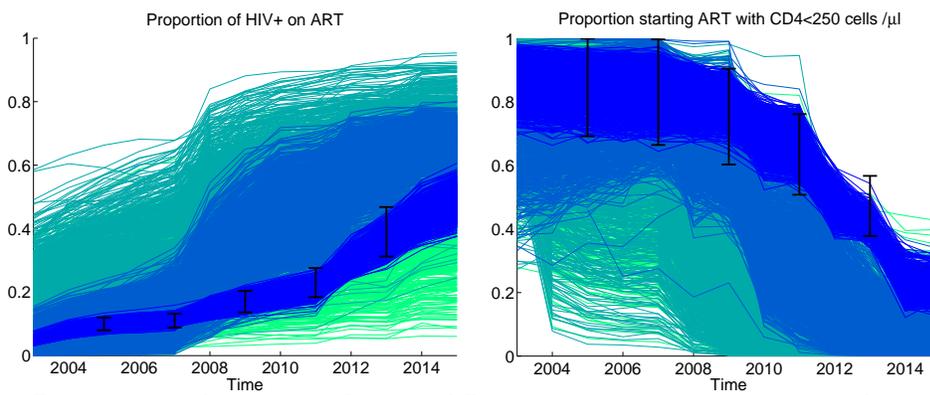


FIG. 1. 10 simulator outputs from two different time series at waves 1, 4, 8 and 14. The observations ranges for the 10 outputs are shown using the black bars. The simulator's output at the 4 different waves is shown using darkening shades of blue.

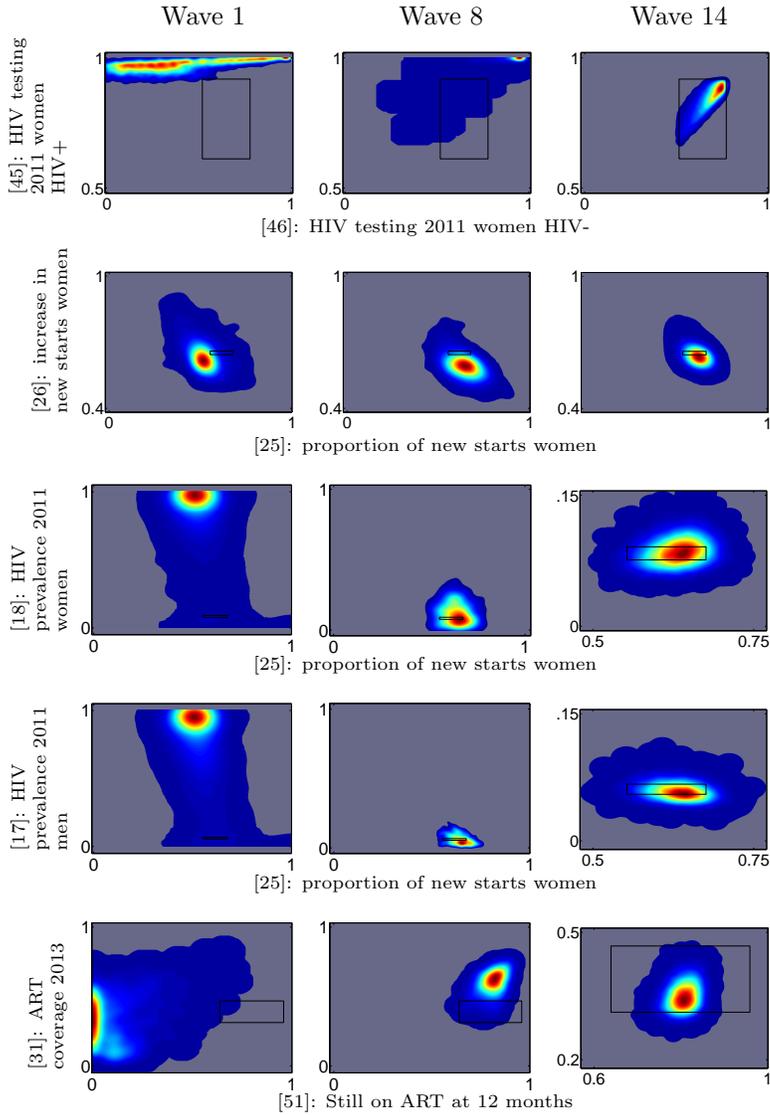


FIG. 2. Histograms of simulator outputs for 5 different output pairs across waves 1, 8 and 14. The calibration targets are shown with black rectangles. Note the different scale in some panels of the rightmost column.

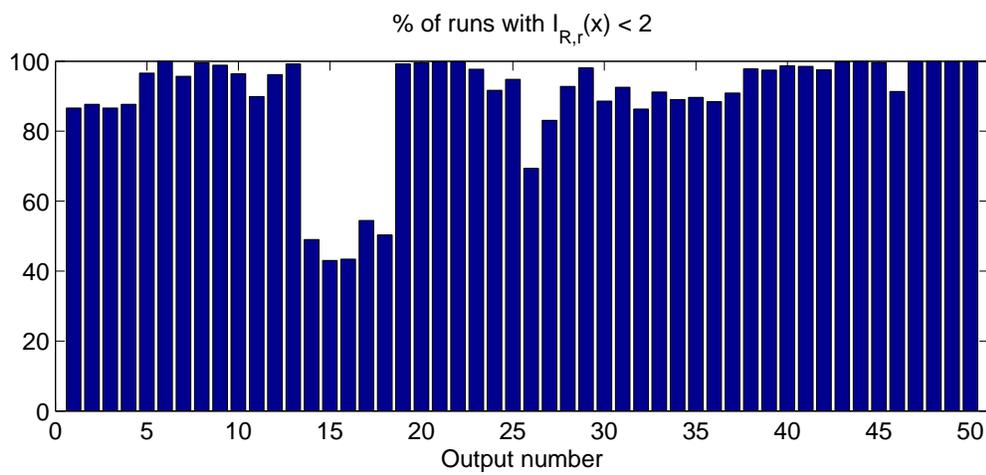


FIG. 3. Percentage of 20000 wave 14 simulator runs with a simulator run implausibility that is less than 2, which can roughly be interpreted as the simulator's output estimated mean falling within the observation interval.

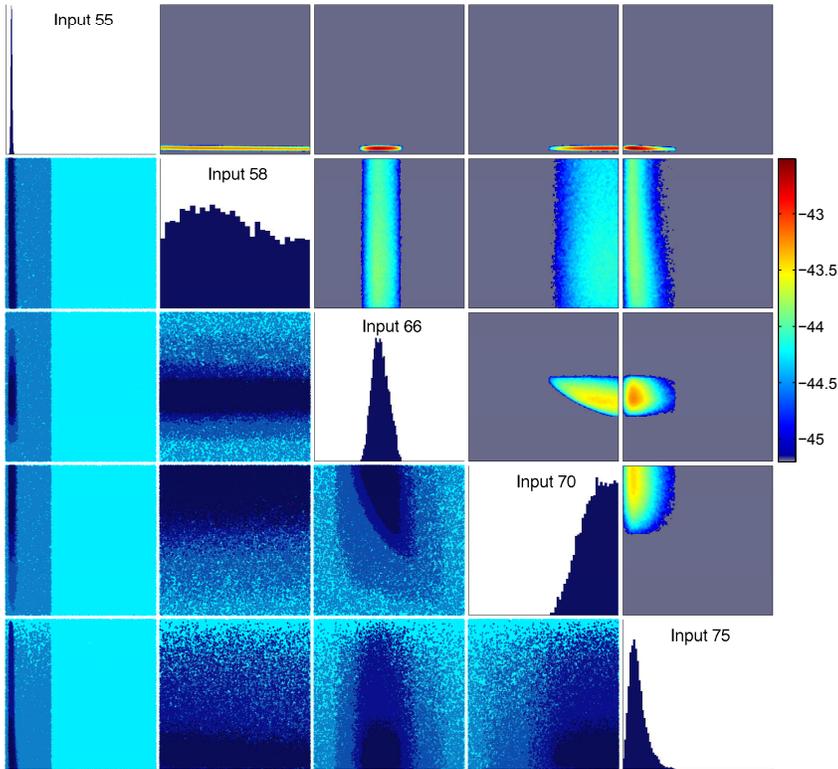


FIG. 4. Summary of the input space shrinking across waves: The lower triangle of the above lattice shows pair plots of non-implausible samples for 5 different inputs at waves 1, 4, 8 and 13, in darkening shades of blue. The upper triangle shows an estimate of the  $\log_{10}$  probability of finding a non-implausible sample after fixing the respective input pairs to a particular value. The gray area indicates that it is virtually impossible to obtain a match for these values of the input pairs. The diagonal shows 1-D histograms of the wave 13 non-implausible samples for the respective inputs. All axes range between the initial minimum and maximum value of each input.

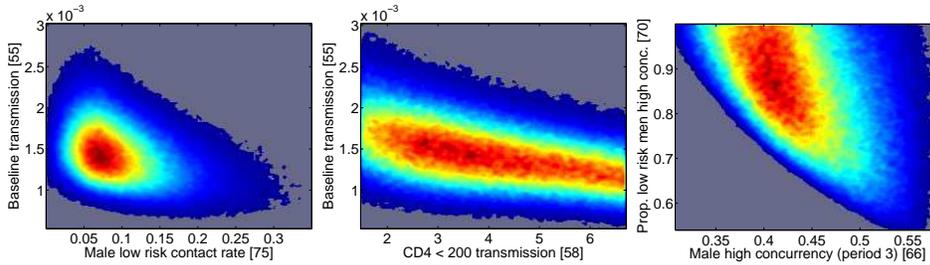


FIG. 5. Histograms of non-implausible samples at wave 13 showing correlation patterns between inputs. See text for an analysis.

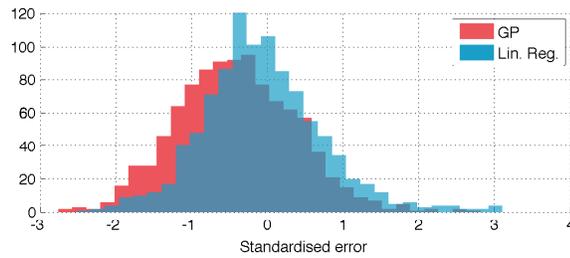


FIG. 6. *Standardised errors for the GP and linear regression based emulators for predictions of the wave 8 simulator runs. Most errors lie within the  $[-2, 2]$  interval. The GP emulator's errors have a slight negative bias, and the linear regression emulator errors are slightly skewed towards positive values.*

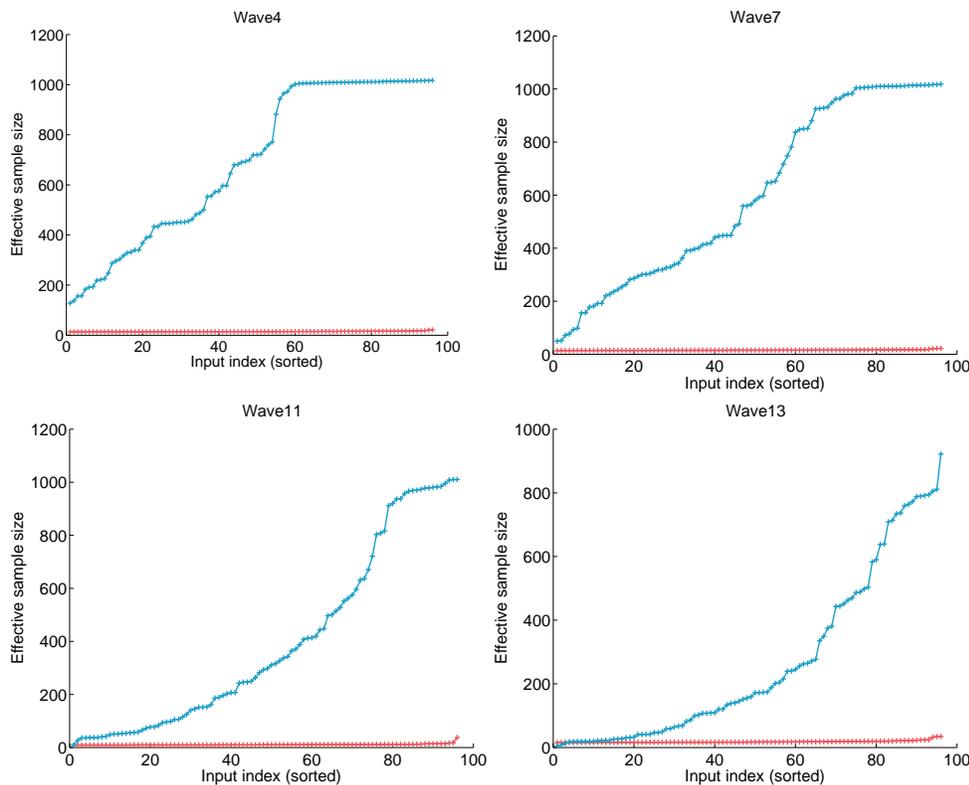


FIG. 7. *Averaged effective sample sizes for the slice sampler (blue) and the Metropolis-Hastings (red) algorithms at 4 different waves. The effective sample sizes were averaged across 1000 different chains. The slice sampler chains contained 1000 samples each and the Metropolis Hastings contained the number of samples required to match the computational effort of the slice sampler in terms of emulator evaluations. Each point in the 4 panels above corresponds to one of the 96 inputs, with their indices sorted to facilitate comparison. The slice sampler resulted in chains with less correlation, as indicated by the higher effective sample size, while in some cases the chains were nearly uncorrelated (effective sample size of  $\sim 1000$  in a 1000 sample chain). In the case of highly correlated inputs in later waves, the performance of the two algorithms was similar, although the Metropolis-Hastings algorithm was aware of the correlation structure but the slice sampler was not.*