

1                                    Assessing the performance of  
2                                    real-time epidemic forecasts:  
3    A case study of Ebola in the Western Area Region  
4                                    of Sierra Leone, 2014–15

5       Sebastian Funk<sup>1,2,\*</sup>, Anton Camacho<sup>1,2,3</sup>, Adam J. Kucharski<sup>1,2</sup>,  
      Rachel Lowe<sup>1,2,4</sup>, Rosalind M. Eggo<sup>1,2</sup>, W. John Edmunds<sup>1,2</sup>

6    <sup>1</sup> Center for the Mathematical Modelling of Infectious Diseases, London School of  
7                                    Hygiene & Tropical Medicine, London, United Kingdom

8    <sup>2</sup> Infectious Disease Epidemiology, London School of Hygiene & Tropical  
9                                    Medicine, London, United Kingdom

10                                    <sup>3</sup> Epicentre, Paris, France

11    <sup>4</sup> Barcelona Institute for Global Health, ISGLOBAL, Barcelona, Spain

12                                    \* Corresponding author. Email: [sebastian.funk@lshtm.ac.uk](mailto:sebastian.funk@lshtm.ac.uk)

13                                    **Abstract**

14                                    Real-time forecasts based on mathematical models can inform criti-  
15                                    cal decision-making during infectious disease outbreaks. Yet, epidemic  
16                                    forecasts are rarely evaluated during or after the event, and there is  
17                                    little guidance on the best metrics for assessment. Here, we propose an  
18                                    evaluation approach that disentangles different components of forecast-  
19                                    ing ability using metrics that separately assess the calibration, sharp-  
20                                    ness and unbiasedness of forecasts. This makes it possible to assess not  
21                                    just how close a forecast was to reality but also how well uncertainty

22 has been quantified. We used this approach to analyse the perfor-  
23 mance of weekly forecasts we generated in real time in Western Area,  
24 Sierra Leone, during the 2013–16 Ebola epidemic in West Africa. We  
25 investigated a range of forecast model variants based on the model fits  
26 generated at the time with a semi-mechanistic model, and found that  
27 good probabilistic calibration was achievable at short time horizons  
28 of one or two weeks ahead but models were increasingly inaccurate  
29 at longer forecasting horizons. This suggests that forecasts may have  
30 been of good enough quality to inform decision making requiring pre-  
31 dictions a few weeks ahead of time but not longer, reflecting the high  
32 level of uncertainty in the processes driving the trajectory of the epi-  
33 demic. Comparing forecasts based on the semi-mechanistic model to  
34 simpler null models showed that the best semi-mechanistic model vari-  
35 ant performed better than the null models with respect to probabilistic  
36 calibration, and that this would have been identified from the earliest  
37 stages of the outbreak. As forecasts become a routine part of the  
38 toolkit in public health, standards for evaluation of performance will  
39 be important for assessing quality and improving credibility of math-  
40 ematical models, and for elucidating difficulties and trade-offs when  
41 aiming to make the most useful and reliable forecasts.

## 42 **Introduction**

43 Forecasting the future trajectory of cases during an infectious disease out-  
44 break can make an important contribution to public health and intervention  
45 planning. Infectious disease modellers are now routinely asked for predic-  
46 tions in real time during emerging outbreaks (Heesterbeek et al., 2015).  
47 Forecasting targets can revolve around expected epidemic duration, size, or  
48 peak timing and incidence (Goldstein et al., 2011; Nsoesie et al., 2013; Yang

49 et al., 2015; Dawson et al., 2015), geographical distribution of risk (Lowe  
50 et al., 2014), or short-term trends in incidence (Johansson et al., 2016; Liu  
51 et al., 2015). However, forecasts made during an outbreak are rarely in-  
52 vestigated during or after the event for their accuracy, and only recently  
53 have forecasters begun to make results, code, models and data available for  
54 retrospective analysis.

55 The growing importance of infectious disease forecasts is epitomised by  
56 the growing number of so-called forecasting challenges. In these, researchers  
57 compete in making predictions for a given disease and a given time hori-  
58 zon. Such initiatives are difficult to set up during unexpected outbreaks,  
59 and are therefore usually conducted on diseases known to occur seasonally,  
60 such as dengue (Johansson et al., 2016; National Oceanic and Atmospheric  
61 Administration, 2017; Centres for Disease Control and Prevention, 2017)  
62 and influenza (Biggerstaff et al., 2016). The *Ebola Forecasting Challenge*  
63 was a notable exception, triggered by the 2013–16 West African Ebola epi-  
64 demic and set up in June 2015. Since the epidemic had ended in most  
65 places at that time, the challenge was based on simulated data designed  
66 to mimic the behaviour of the true epidemic instead of real outbreak data.  
67 The main lessons learned were that 1) ensemble estimates outperformed all  
68 individual models, 2) more accurate data improved the accuracy of forecasts  
69 and 3) considering contextual information such as individual-level data and  
70 situation reports improved predictions (Viboud et al., 2017).

71 In theory, infectious disease dynamics should be predictable within the  
72 timescale of a single outbreak (Scarpino and Petri, 2017). In practice, how-  
73 ever, providing accurate forecasts during emerging epidemics comes with  
74 particular challenges such as data quality issues and limited knowledge about

75 the processes driving growth and decline in cases. In particular, uncertainty  
76 about human behavioural changes and public health interventions can pre-  
77 clude reliable long-term predictions (Moran et al., 2016; Funk et al., 2017b).  
78 Yet, short-term forecasts with an horizon of a few generations of transmis-  
79 sion (e.g., a few weeks in the case of Ebola), can yield important information  
80 on current and anticipated outbreak behaviour and, consequently, guide im-  
81 mediate decision making.

82 The most recent example of large-scale outbreak forecasting efforts was  
83 during the 2013–16 Ebola epidemic, which vastly exceeded the burden of  
84 all previous outbreaks with almost 30,000 reported cases of the disease, re-  
85 sulting in over 10,000 deaths in the three most affected countries: Guinea,  
86 Liberia and Sierra Leone. During the epidemic, several research groups pro-  
87 vided forecasts or projections at different time points, either by generating  
88 scenarios believed plausible, or by fitting models to the available time series  
89 and projecting them forward to predict the future trajectory of the out-  
90 break (Fisman et al., 2014; Lewnard et al., 2014; Nishiura and Chowell,  
91 2014; Rivers et al., 2014; Towers et al., 2014; Camacho et al., 2015b; Dong  
92 et al., 2015; Drake et al., 2015; Merler et al., 2015; Siettos et al., 2015;  
93 White et al., 2015). One forecast that gained attention during the epidemic  
94 was published in the summer of 2014, projecting that by early 2015 there  
95 might be 1.4 million cases (Meltzer et al., 2014). This number was based  
96 on unmitigated growth in the absence of further intervention and proved  
97 a gross overestimate, yet it was later highlighted as a “call to arms” that  
98 served to trigger the international response that helped avoid the worst-case  
99 scenario (Frieden and Damon, 2015). While that was a particularly dras-  
100 tic prediction, most forecasts made during the epidemic were later found  
101 to have overestimated the expected number of cases, which provided a case

102 for models that can generate sub-exponential growth trajectories (Chretien  
103 et al., 2015; Chowell et al., 2017).

104 Traditionally, epidemic forecasts are assessed using aggregate metrics  
105 such as the mean absolute error (MAE, Chowell, 2017; Pei and Shaman,  
106 2017; Viboud et al., 2017). This, however, only assesses how close the most  
107 likely or average predicted outcome is to the true outcome. The ability  
108 to correctly forecast uncertainty, and to quantify confidence in a predicted  
109 event, is not assessed by such metrics. Appropriate quantification of uncer-  
110 tainty, especially of the likelihood and magnitude of worst case scenarios,  
111 is crucial in assessing potential control measures. Methods to assess proba-  
112 bilistic forecasts are now being used in other fields, but are not commonly  
113 applied in infectious disease epidemiology (Gneiting and Katzfuss, 2014;  
114 Held et al., 2017).

115 We produced weekly sub-national real-time forecasts during the Ebola  
116 epidemic, starting on 28 November 2014. Plots of the forecasts were pub-  
117 lished on a dedicated web site and updated every time a new set of data  
118 were available (Center for the Mathematical Modelling of Infectious Dis-  
119 eases, 2015). They were generated using a model that has, in variations,  
120 been used to forecast bed demand during the epidemic in Sierra Leone (Ca-  
121 macho et al., 2015b) and the feasibility of vaccine trials later in the epi-  
122 demic (Camacho et al., 2015a; Camacho et al., 2017). During the epidemic,  
123 we provided sub-national forecasts for the three most affected countries (at  
124 the level of counties in Liberia, districts in Sierra Leone and prefectures in  
125 Guinea).

126 Here, we apply assessment metrics that elucidate different properties of  
127 forecasts, in particular their probabilistic calibration, sharpness and bias.

128 Using these methods, we retrospectively assess the forecasts we generated  
129 for Western Area in Sierra Leone, an area that saw one of the greatest  
130 number of cases in the region and where our model informed bed capacity  
131 planning.

## 132 **Materials and Methods**

### 133 **Data sources**

134 Numbers of suspected, probable and confirmed Ebola cases at sub-national  
135 levels were initially compiled from daily *Situation Reports* (or *SitReps*) pro-  
136 vided in PDF format by Ministries of Health of the three affected countries  
137 during the epidemic (Camacho et al., 2015b). Data were automatically  
138 extracted from tables included in the reports wherever possible and other-  
139 wise manually converted by hand to machine-readable format and aggre-  
140 gated into weeks. From 20 November 2014, the World Health Organization  
141 (WHO) provided tabulated data on the weekly number of confirmed and  
142 probable cases. These were compiled from the patient database, which was  
143 continuously cleaned and took into account reclassification of cases avoiding  
144 potential double-counting. However, the patient database was updated with  
145 substantial delay so that the number of reported cases would typically be  
146 underestimated in the weeks leading up to the date of the forecast. Because  
147 of this, we used the SitRep data for the most recent weeks until the latest  
148 week in which the WHO case counts either equalled or exceeded the SitRep  
149 counts. For all earlier times, the WHO data were used.

## 150 **Transmission model**

151 We used a semi-mechanistic stochastic model of Ebola transmission de-  
152 scribed previously (Camacho et al., 2015b; Funk et al., 2017a). Briefly,  
153 the model was based on a Susceptible-Exposed-Infectious-Recovered (SEIR)  
154 model with fixed incubation period of 9.4 days (WHO Ebola Response Team,  
155 2014), following an Erlang distribution with shape 2. The country-specific  
156 infectious period was determined by adding the average delay to hospitalisa-  
157 tion to the average time from hospitalisation to death or discharge, weighted  
158 by the case-fatality rate. Cases were assumed to be reported with a stochas-  
159 tic time-varying delay. On any given day, this was given by a gamma distri-  
160 bution with mean equal to the country-specific average delay from onset to  
161 hospitalisation and standard deviation of 0.1 day. We allowed transmission  
162 to vary over time, to capture behavioural changes in the community, public  
163 health interventions or other factors affecting transmission for which infor-  
164 mation was not available at the time. The time-varying transmission rate  
165 was modelled using a daily Gaussian random walk with fixed volatility (or  
166 standard deviation of the step size) which was estimated as part of the in-  
167 ference procedure (see below). We log-transformed the transmission rate to  
168 ensure it remained positive. The behaviour in time can be written as

$$169 \quad d \log \beta_t = \sigma dW_t \quad (1)$$

170 where  $\beta_t$  is the time-varying transmission rate,  $W_t$  is the Wiener process and  
171  $\sigma$  the volatility of the transmission rate. The basic reproduction number  $R_{0,t}$   
172 at any time was obtained by multiplying  $\beta_t$  with the infectious period. In  
173 fitting the model to the time series of cases we extracted posterior predictive  
174 samples of trajectories, which we used to generate forecasts.

## 175 **Model fitting**

176 Each week, we fitted the model to the available case data leading up to  
177 the date of the forecast. Observations were assumed to follow a negative  
178 binomial distribution. Since the *ssm* software used to fit the model only  
179 implemented a discretised normal observation model, we used a normal ap-  
180 proximation of the negative binomial for observations, potentially introduc-  
181 ing a bias at small counts. Four parameters were estimated in the process:  
182 the initial basic reproduction number  $R_0$  (uniform prior within  $(1, 5)$ ), initial  
183 number of infectious people (uniform prior within  $(1, 400)$ ), overdispersion of  
184 the (negative binomial) observation process (uniform prior within  $(0, 0.5)$ )  
185 and volatility of the time-varying transmission rate (uniform prior within  
186  $(0, 0.5)$ ). We confirmed from the posterior distributions of the parameters  
187 that these priors did not set any problematic bounds. Samples of the pos-  
188 terior distribution of parameters and state trajectories were extracted using  
189 particle Markov chain Monte Carlo (Andrieu et al., 2010) as implemented  
190 in the *ssm* library (Dureau et al., 2013). For each forecast, 50,000 samples  
191 were extracted and thinned to 5000.

## 192 **Predictive model variants**

193 We used the samples of the posterior distribution generated using the Monte  
194 Carlo sampler to produce predictive trajectories, using the final values of es-  
195 timated state trajectories as initial values for the forecasts and simulating  
196 the model forward for up to 10 weeks. While all model fits were generated  
197 using the same model described above, we tested a range of different predic-  
198 tive model variants to assess the quality of ensuing predictions. We tested  
199 variants where trajectories were stochastic (with demographic stochasticity



200 and a noisy reporting process), as well as ones where these sources of noise  
201 were removed for predictions. We further tested predictive model variants  
202 where the transmission rate continued to follow a random walk (unbounded,  
203 on a log-scale), as well as ones where the transmission rate stayed fixed dur-  
204 ing the forecasting period. Where the transmission rate remained fixed for  
205 prediction, we tested variants where we used the final value of the trans-  
206 mission rate and ones where this value would be averaged over a number  
207 of weeks leading up to the final fitted point, to reduce the potential influ-  
208 ence of the last time point, where the transmission rate may not have been  
209 well identified. We tested variants where the predictive trajectory would be  
210 based on the final values and start at the last time point, and ones where  
211 they would start at the penultimate time point, which could, again, be ex-  
212 pected to be better informed by the data. For each model and forecast  
213 horizon, we generated point-wise medians and credible intervals from the  
214 sample trajectories.

## 215 **Null models**

216 To assess the performance of the semi-mechanistic transmission model we  
217 compared it to three simpler null models: two representing the constituent  
218 parts of the semi-mechanistic model, and a non-mechanistic time series  
219 model. For the first null model, we used a *deterministic* model that only con-  
220 tained the mechanistic core of the semi-mechanistic model, that is a deter-  
221 ministic SEIR model with fixed transmission rate and parameters otherwise

222 the same as in the model described elsewhere (Camacho et al., 2015b):

$$223 \quad \frac{dS}{dt} = -\frac{R_0}{\Delta} \frac{I_c + I_h}{N} S \quad (2)$$

$$224 \quad \frac{dE_1}{dt} = -\frac{R_0}{\Delta} \frac{I_c + I_h}{N} S - 2\nu E_1 \quad (3)$$

$$225 \quad \frac{dE_2}{dt} = 2\nu E_1 - 2\nu E_2 \quad (4)$$

$$226 \quad \frac{dI_c}{dt} = 2\nu E_2 - \tau I_c \quad (5)$$

$$227 \quad \frac{dI_h}{dt} = \tau I_c - \gamma I_h \quad (6)$$

$$228 \quad \frac{dR}{dt} = \gamma I_h \quad (7)$$

$$229 \quad \frac{dA}{dt} = \tau I_c \quad (8)$$

$$230 \quad Y_t \sim \text{NB}(A_t - A_{t-1}, \phi) \quad (9)$$

231

232 where  $Y_t$  are observations at times  $t$ ,  $S$  is the number susceptible,  $E$  the  
233 number incubating (split into two compartments for Erlang-distributed per-  
234 manence times with shape 2),  $I_c$  is the number infectious and not yet no-  
235 tified,  $I_h$  is the number infectious and notified,  $R$  is the number recovered  
236 or dead,  $A$  is an accumulator for incidence,  $R_0$  is the basic reproduction  
237 number,  $\Delta = 1/\tau + 1/\nu$  is the mean time from onset to outcome,  $1/\nu$  is the  
238 mean incubation period,  $1/\tau + 1/\gamma$  is the mean duration of infectiousness,  
239  $1/\tau$  is the mean time from onset to hospitalisation  $1/\gamma$  the mean duration  
240 from notification to outcome and  $\text{NB}(\mu, \phi)$  is a negative binomial distribu-  
241 tion with mean  $\mu$  and overdispersion  $\phi$ . All these parameters were taken  
242 from individual patient observations (WHO Ebola Response Team, 2014)  
243 except the overdispersion in reporting  $\phi$ , and the basic reproduction num-  
244 ber  $R_0$ , which were inferred using Markov-chain Monte Carlo with the same  
245 priors as in the semi-mechanistic model.

246 For the second null model, we used an *unfocused* model where the weekly

247 incidence  $Z$  itself was modelled using a stochastic volatility model (without  
248 drift), that is a daily Gaussian random walk, and forecasts generated as-  
249 suming the weekly number of new cases was not going to change:

$$250 \quad d \log Z = \sigma dW \quad (10)$$

$$251 \quad Y_t \sim \text{NB}(Z_t, \phi) \quad (11)$$

252

253 where  $Y$  are observations,  $\sigma$  is the intensity of the random walk and  $\phi$   
254 the overdispersion of reporting (both estimated using Markov-chain Monte  
255 Carlo) and  $dW$  is the Wiener process.

256 Lastly, we used a null model based on a non-mechanistic Bayesian au-  
257 toregressive AR(1) time series model:

$$258 \quad \alpha_{t+1} \sim \mathcal{N}(\phi \alpha_t, \sigma_\alpha) \quad (12)$$

$$259 \quad Y_t^* \sim \mathcal{N}(\alpha_t, \sigma_{Y^*}) \quad (13)$$

$$260 \quad Y_t = \max(0, [Y_t^*]) \quad (14)$$

261

262 where  $\phi$ ,  $\sigma_\alpha$  and  $\sigma_{Y^*}$  were estimated using Markov-chain Monte Carlo, and  
263 [...] indicates rounding to the nearest integer. An alternative model with  
264 Poisson distributed observations was discarded as it yielded poorer predic-  
265 tive performance.

266 The deterministic and unfocused models were implemented in *libbi* (Mur-  
267 ray, 2015) via the *RBi* (Jacob and Funk, 2017) and *RBi.helpers* (Funk, 2016)  
268 *R* packages (R Core Team, 2018). The Bayesian autoregressive time series  
269 model was implemented using the *bsts* package (Scott, 2017).

## 270 Metrics

271 The paradigm for assessing probabilistic forecasts is that they should max-  
272 imise the sharpness of predictive distributions subject to calibration (Gneit-  
273 ing et al., 2007). We therefore first assessed model calibration at a given  
274 forecasting horizon, before assessing their sharpness and other properties.

275 *Calibration* or reliability (Friederichs and Thorarinsdottir, 2012) of fore-  
276 casts is the ability of a model to correctly identify its own uncertainty in  
277 making predictions. In a model with perfect calibration, the observed data  
278 at each time point look as if they came from the predictive probability dis-  
279 tribution at that time. Equivalently, one can inspect the probability integral  
280 transform of the predictive distribution at time  $t$  (Dawid, 1984),

$$281 \quad u_t = F_t(x_t) \quad (15)$$

282 where  $x_t$  is the observed data point at time  $t \in t_1, \dots, t_n$ ,  $n$  being the number  
283 of forecasts, and  $F_t$  is the (continuous) predictive cumulative probability  
284 distribution at time  $t$ . If the true probability distribution of outcomes at  
285 time  $t$  is  $G_t$  then the forecasts  $F_t$  are said to be *ideal* if  $F_t = G_t$  at all times  
286  $t$ . In that case, the probabilities  $u_t$  are distributed uniformly.

287 In the case of discrete outcomes such as the incidence counts that were  
288 forecast here, the PIT is no longer uniform even when forecasts are ideal.  
289 In that case a randomised PIT can be used instead:

$$290 \quad u_t = P_t(k_t) + v(P_t(k_t) - P_t(k_t - 1)) \quad (16)$$

291 where  $k_t$  is the observed count,  $P_t(x)$  is the predictive cumulative probability  
292 of observing incidence  $k$  at time  $t$ ,  $P_t(-1) = 0$  by definition and  $v$  is standard  
293 uniform and independent of  $k$ . If  $P_t$  is the true cumulative probability  
294 distribution, then  $u_t$  is standard uniform (Czado et al., 2009). To assess  
295 calibration, we therefore applied the Anderson-Darling test of uniformity to  
296 the probabilities  $u_t$ . The resulting p-value was a reflection of how compatible  
297 the forecasts were with the null hypothesis of uniformity of the PIT, or of  
298 the data coming from the predictive probability distribution. We considered  
299 that there was no evidence to suggest a forecasting model was miscalibrated  
300 if the p-value found was greater than a threshold of  $p \geq 0.1$ , some evidence  
301 that it was miscalibrated if  $0.01 < p < 0.1$ , and good evidence that it  
302 was miscalibrated if  $p \leq 0.01$ . In this context it should be noted, though,  
303 that uniformity of the (randomised) PIT is a necessary but not sufficient  
304 condition of calibration (Gneiting et al., 2007). The p-values calculated  
305 here merely quantify our ability to reject a hypothesis of good calibration,  
306 but cannot guarantee that a forecast is calibrated. Because of this, other  
307 indicators of forecast quality must be considered when choosing a model for  
308 forecasts.

309 All of the following metrics are evaluated at every single data point. In  
310 order to compare the forecast quality of models, they are averaged across  
311 the data set.

312 *Sharpness* is the ability of the model to generate predictions within a  
313 narrow range of possible outcomes. It is a data-independent measure, that  
314 is, it is purely a feature of the forecasts themselves. To evaluate sharpness  
315 at time  $t$ , we used the normalised median absolute deviation about the

316 median (MADN) of  $y$

$$317 \quad S_t(P_t) = \frac{1}{0.675} \text{median}(|y - \text{median}(y)|) \quad (17)$$

318 where  $y$  is a variable with CDF  $P_t$ , and division by 0.675 ensures that if  
319 the predictive distribution is normal this yields a value equivalent to the  
320 standard deviation. The MAD (i.e., the MADN without the normalising  
321 factor) is related to the interquartile range (and in the limit of infinite sample  
322 size takes twice its value), a common measure of sharpness (Gneiting and  
323 Katzfuss, 2014), but is more robust to outliers (Maronna et al., 2018). The  
324 sharpest model would focus all forecasts on one point and have  $S = 0$ ,  
325 whereas a completely blurred forecast would have  $S \rightarrow \infty$ . Again, we used  
326 Monte-Carlo samples from  $P_t$  to estimate sharpness.

327 We further assessed the *bias* of forecasts to test whether a model system-  
328 atically over- or underpredicted. We defined the forecast bias at time  $t$  as  
329

$$330 \quad B_t(P_t, x_t) = 1 - (P_t(x_t) + P_t(x_t - 1)) \quad (18)$$

331 The most unbiased model would have exactly half of predictive probabil-  
332 ity mass not concentrated on the data itself below the data at time  $t$  and  
333  $B_t = 0$ , whereas a completely biased model would yield either all predictive  
334 probability mass above ( $B_t = 1$ ) or below ( $B_t = -1$ ) the data.

335 We further evaluated forecasts using two *strictly proper scoring rules*,  
336 that is scores which are minimised if the predictive distribution is the same  
337 as the one generating the data. These scores combine the assessment of  
338 calibration and sharpness for comparison of overall forecasting skill. The

339 *Ranked Probability Score* (RPS, Epstein, 1969; Murphy, 1969) for count  
340 data is defined as (Czado et al., 2009)

$$341 \quad \text{RPS}(P_t, x_t) = \sum_{k=0}^{\infty} (P_t(k) - \mathbb{1}(k \geq x_t))^2. \quad (19)$$

342 It reduces to the mean absolute error (MAE) if the forecast is deterministic  
343 and can therefore be seen as its probabilistic generalisation for discrete fore-  
344 casts. A convenient equivalent formulation for predictions generated from  
345 Monte-Carlo samples is (Gneiting et al., 2007; Czado et al., 2009)

$$346 \quad \text{RPS}(P_t, x_t) = \mathbb{E}_{P_t} |X - x_t| - \frac{1}{2} \mathbb{E}_{P_t} |X - X'|, \quad (20)$$

347 where  $X$  and  $X'$  are independent realisations of a random variable with  
348 cumulative distribution  $P_t$ .

349 The *Dawid-Sebastiani score* (DSS) only considers the first two moments  
350 of the predictive distribution and is defined as (Czado et al., 2009)

$$351 \quad \text{DSS}(P_t, x_t) = \left( \frac{x_t - \mu_{P_t}}{\sigma_{P_t}} \right)^2 + 2 \log \sigma_{P_t} \quad (21)$$

352 where  $\mu_{P_t}$  and  $\sigma_{P_t}$  are the mean and standard deviation of the predic-  
353 tive probability distribution, respectively, estimated here using Monte-Carlo  
354 samples.

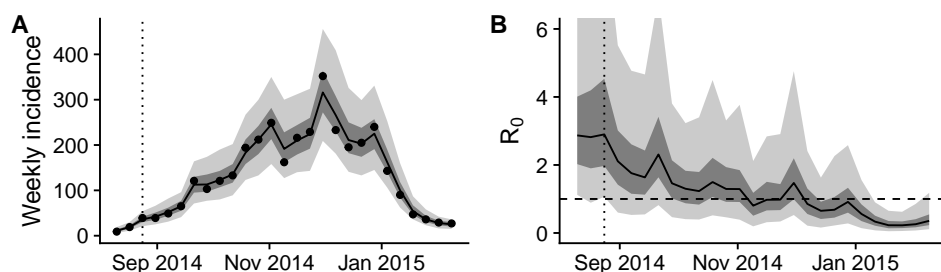
355 For comparison, we also evaluated forecasts using the *absolute error* (AE)  
356 of the median forecast, that is

$$357 \quad \text{AE}(P_t, x_t) = |\text{median}_{P_t}(X) - x_t| \quad (22)$$

358 where  $X$  is a random variable with cumulative distribution  $P_t$ .

## 359 Results

360 The semi-mechanistic model used to generate real-time forecasts during the  
361 epidemic was able to reproduce the trajectories up to the date of each fore-  
362 cast, following the data closely by means of the smoothly varying transmis-  
363 sion rate (Fig. 1). The overall behaviour of the reproduction number (ig-  
364 noring depletion of susceptibles which did not play a role at the population  
365 level given the relatively small proportion of the population infected) was  
366 one of a near-monotonic decline, from a median estimate of 2.9 (interquar-  
367 tile range (IQR) 2.1–4, 90% credible interval (CI) 1.2–6.9) in the first fitted  
368 week (beginning 10 August, 2014) to a median estimate of 1.3 (IQR 0.9–1.9,  
369 90% CI 0.4–3.7) in early November, 0.9 (IQR 0.6–1.3, 90% CI 0.2–2.2) in  
370 early December, 0.6 in early January (IQR 0.3–0.8, 90% CI 0.1–1.5) and 0.3  
371 at the end of the epidemic in early February (IQR 0.2–0.4, 90% CI 0.1–0.9).



**Figure 1. Final fit of the semi-mechanistic model to the Ebola outbreak in Western Area, Sierra Leone.** (A) Final fit of the reported weekly incidence (black line and grey shading) to the data (black dots). (B) Corresponding dynamics of the reproduction number (ignoring depletion of susceptibles). Point-wise median state estimates are indicated by a solid line, interquartile ranges by dark shading, and 90% intervals by light shading. The threshold reproduction number ( $R_0 = 1$ ), determining whether case numbers are expected to increase or decrease, is indicated by a dashed line. In both plots, a dotted vertical line indicates the date of the first forecast assessed in this manuscript (24 August 2014).



372 The epidemic lasted for a total of 27 weeks, with forecasts generated  
373 starting from week 3. For  $m$ -week ahead forecasts this yielded a sample size  
374 of  $25 - m$  forecasts to assess calibration. Calibration of forecasts from the  
375 semi-mechanistic model were good for a maximum of one or two weeks, but  
376 deteriorated rapidly at longer forecasting horizons (Fig. 2). The two semi-  
377 mechanistic forecast model variants with best calibration performance used  
378 deterministic dynamics starting at the last fitted data point (Table 1). Of  
379 these two, the forecast model that kept the transmission rate constant from  
380 the value at the last data point performed slightly better across forecast  
381 horizons than one that continued to change the transmission rate following  
382 a random walk with volatility estimated from the time series. There was  
383 no evidence of miscalibration in both of the models with best calibration  
384 performance for two-week ahead forecasts, but increasing evidence of mis-  
385 calibration for forecast horizons of three weeks or more. Calibration of all  
386 model variants was poor four weeks or more ahead, and all the stochastic  
387 model variants were miscalibrated for any forecast horizon, including the one  
388 we used to publish forecasts during the Ebola epidemic (stochastic, starting  
389 at the last data point, no averaging of the transmission rate, no projected  
390 volatility).

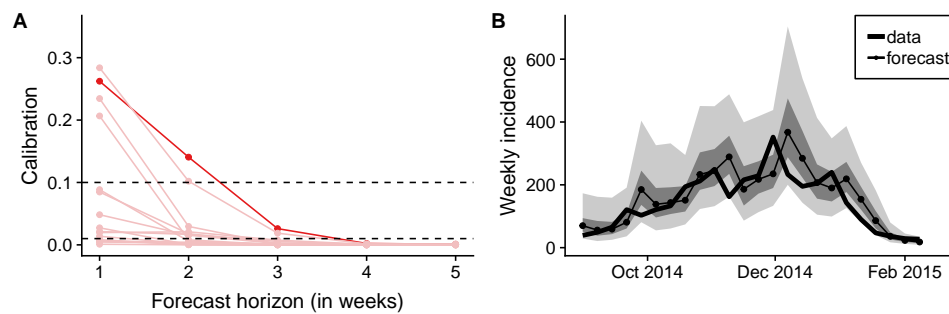
391 The calibration of the best semi-mechanistic forecast model variant (de-  
392 terministic dynamics, transmission rate fixed and starting at the last data  
393 point) was better than for any of the null models (Fig. 3A and Table 2)  
394 for up to three weeks. While there was no evidence for miscalibration of  
395 the autoregressive null model for 1-week-ahead forecasts, there was good  
396 evidence of miscalibration for longer forecast horizons. There was some ev-  
397 idence of miscalibration of the unfocused null model, which assumes that  
398 the same number of cases will be reported in the weeks following the week

stochasticity	Predictive model variant			Forecast horizon (weeks)			
	start	transmission	averaged	1	2	3	4
deterministic	at last data point	varying	no	<b>0.28</b>	<b>0.1</b>	0.02	<0.01
deterministic	at last data point	fixed	no	<b>0.26</b>	<b>0.14</b>	0.03	<0.01
deterministic	at last data point	fixed	2 weeks	<b>0.24</b>	0.03	<0.01	<0.01
deterministic	at last data point	fixed	3 weeks	<b>0.21</b>	<0.01	<0.01	<0.01
deterministic	1 week before	varying	no	0.05	0.02	<0.01	<0.01
deterministic	1 week before	fixed	no	0.09	0.02	<0.01	<0.01
deterministic	1 week before	fixed	2 weeks	0.09	<0.01	<0.01	<0.01
deterministic	1 week before	fixed	3 weeks	0.03	<0.01	<0.01	<0.01
stochastic	at last data point	varying	no	0.02	0.02	<0.01	<0.01
stochastic	at last data point	fixed	no	0.02	0.02	<0.01	<0.01
stochastic	at last data point	fixed	2 weeks	0.01	<0.01	<0.01	<0.01
stochastic	at last data point	fixed	3 weeks	<0.01	<0.01	<0.01	<0.01
stochastic	1 week before	varying	no	<0.01	<0.01	<0.01	<0.01
stochastic	1 week before	fixed	no	<0.01	<0.01	<0.01	<0.01
stochastic	1 week before	fixed	2 weeks	<0.01	<0.01	<0.01	<0.01
stochastic	1 week before	fixed	3 weeks	<0.01	<0.01	<0.01	<0.01

**Table 1. Calibration of forecast model variants of the semi-mechanistic model.** Calibration (p-value of the Anderson-Darling test of uniformity) of deterministic and stochastic forecasts starting either at the last data point or one week before, with varying (according to a Gaussian random walk) or fixed transmission rate either starting from the last value of the transmission rate or from an average over the last 2 or 3 weeks, at different forecast horizons up to 4 weeks. The p-values highlighted in bold reflect predictive models with no evidence of miscalibration. The second row corresponds to the highlighted model in Fig. 2A.

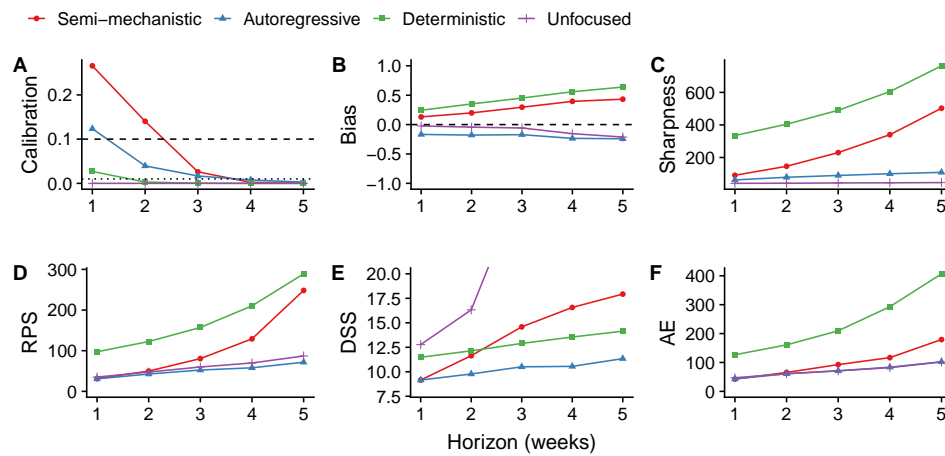
399 during which the forecast was made, for 1 week ahead and good evidence  
400 of miscalibration beyond. Calibration of the deterministic null model was  
401 poor for all forecast horizons.

402 The semi-mechanistic and deterministic models showed a tendency to  
403 overestimate the predicted number of cases, while the autoregressive and  
404 null models tended to underestimate (Fig. 3B and and Table 2). This bias  
405 increased with longer forecast horizons in all cases. The semi-mechanistic



**Figure 2. Calibration of forecasts from the semi-mechanistic model.** (A) Calibration of model variants (p-value of Anderson-Darling test) as a function of the forecast horizon. Shown in dark red is the best calibrated forecasting model variant (corresponding to the second row of Table 1). Other model variants are shown in light red. (B) Comparison of one-week forecasts of reported weekly incidence generated using the best semi-mechanistic model variant to the subsequently released data. The data are shown as a thick line, and forecasts as dots connected by a thin line. Dark shades of grey indicate the point-wise interquartile range, and lighter shades of grey the point-wise 90% credible interval.

406 model with best calibration progressed from a 12% bias at 1 week ahead to  
407 20% (2 weeks), 30% (3 weeks), 40% (4 weeks) and 44% (5 weeks) overesti-  
408 mation. At the same time, this model showed rapidly decreasing sharpness  
409 as the forecast horizon increased (Fig. 3C and and Table 2). This is re-  
410 flected in the proper scoring rules that combine calibration and sharpness,  
411 with smaller values indicating better forecasts (Fig. 3D-E and and Table 2).  
412 At 1-week ahead, the mean RPS values of the autoregressive, unfocused and  
413 best semi-mechanistic forecasting models were all around 30. At increas-  
414 ing forecasting horizon, the RPS of the semi-mechanistic model grew faster  
415 than the RPS of the autoregressive and unfocused null models. The DSS  
416 of the semi-mechanistic model, on the other hand, was very similar to the  
417 one of the autoregressive and better than that of the other null models at  
418 a forecast horizon of 1 week, with the autoregressive again performing best  
419 at increasing forecast horizons.



**Figure 3. Forecasting metrics and scores of the best semi-mechanistic model variant compared to null models.** Metrics shown are (A) calibration (p-value of Anderson-Darling test, greater values indicating better calibration, dashed lines at 0.1 and 0.01), (B) bias (less bias if closer to 0), (C) sharpness (MAD, sharper models having values closer to 0), (D) RPS (better values closer to 0), (E) DSS (better values closer to 0) and (F) AE (better values closer to 0), all as a function of the forecast horizon.

420 Focusing purely on the median forecast (and thus ignoring both cali-  
 421 bration and sharpness), the absolute error (AE, Fig.3F and Table 2) was  
 422 lowest (42) for the best semi-mechanistic model variant at 1-week ahead  
 423 forecasts, although similar to the autoregressive and unfocused null models.  
 424 With increasing forecasting horizon, the absolute error increased at a faster  
 425 rate than for the autoregressive and unfocused null models.

426 We lastly studied the calibration behaviour of the models over time;  
 427 that is, using the data and forecasts available up to different time points  
 428 during the epidemic (Fig. 4). This shows that from very early on, not much  
 429 changed in the ranking of the different semi-mechanistic model variants.  
 430 Comparing the best semi-mechanistic forecasting model to the null models,  
 431 again, for almost the whole duration of the epidemic calibration of the semi-

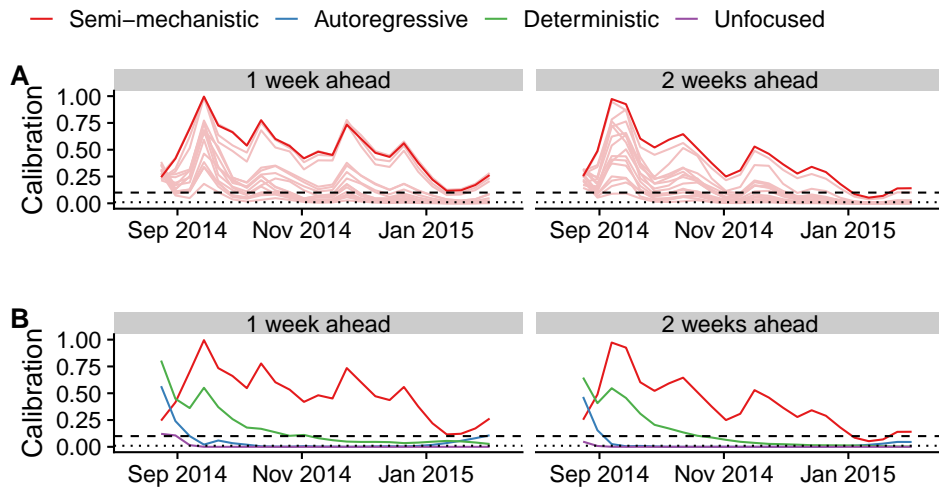
Model	Calibration	Sharpness	Bias	RPS	DSS	AE
<b>1 week ahead</b>						
Semi-mechanistic	<b>0.26</b>	91	0.13	31	9.2	42
Autoregressive	<b>0.1</b>	61	-0.17	31	9.1	43
Deterministic	0.03	340	0.24	97	11	130
Unfocused	<0.01	41	-0.024	35	13	47
<b>2 weeks ahead</b>						
Semi-mechanistic	<b>0.14</b>	150	0.2	50	12	65
Autoregressive	0.03	77	-0.18	43	9.9	60
Deterministic	<0.01	400	0.35	120	12	160
Unfocused	<0.01	42	-0.044	48	16	61
<b>3 weeks ahead</b>						
Semi-mechanistic	0.03	230	0.3	81	15	93
Autoregressive	0.02	90	-0.17	53	11	73
Deterministic	<0.01	490	0.45	160	13	210
Unfocused	<0.01	44	-0.058	60	29	71

**Table 2. Forecasting metrics and scores of the best semi-mechanistic model variant compared to null models.** The values shown are the same scores as in Fig. 3, for forecasting horizons up to three weeks. The p-values for calibration highlighted in bold reflect predictive models with no evidence of miscalibration.

432 mechanistic model was best for forecasts 1 or 2 weeks ahead.

## 433 Discussion

434 Probabilistic forecasts aim to quantify the inherent uncertainty in predicting  
435 the future. In the context of infectious disease outbreaks, they allow the  
436 forecaster to go beyond merely providing the most likely future scenario  
437 and quantify how likely that scenario is to occur compared to other possible  
438 scenarios. While correctly quantifying uncertainty in predicted trajectories  
439 has not commonly been the focus in infectious disease forecasting, it can  
440 have enormous practical implications for public health planning. Especially  
441 during acute outbreaks, decisions are often made based on so-called “worst-



**Figure 4. Calibration over time.** Calibration scores of the forecast up to the time point shown on the x-axis. (A) Semi-mechanistic model variants, with the best model highlighted in dark red and other model variants are shown in light red. (B) Best semi-mechanistic model and null models. In both cases, 1-week (left) and 2-week (right) calibration (p-value of Anderson-Darling test) are shown.

442 case scenarios” and their likelihood of occurring. The ability to adequately  
443 assess the magnitude as well as the probability of such scenarios requires  
444 accuracy at the tails of the predictive distribution, in other words good  
445 calibration of the forecasts.

446 Probabilistic forecasts need to be assessed using metrics that go beyond  
447 the simple difference between the central forecast and what really happened.  
448 Applying a suite of assessment methods to the forecasts we produced for  
449 Western Area, Sierra Leone, we found that probabilistic calibration of semi-  
450 mechanistic model variants varied, with the best ones showing good calibra-  
451 tion for up to 2-3 weeks ahead, but performance deteriorated rapidly as the  
452 forecasting horizon increased. This reflects our lack of knowledge about the  
453 underlying processes shaping the epidemic at the time, from public health

454 interventions by numerous national and international agencies to changes in  
455 individual and community behaviour. During the epidemic, we only pub-  
456 lished forecasts up to 3 weeks ahead, as longer forecasting horizons were not  
457 considered appropriate.

458 Our forecasts suffered from bias that worsened as the forecasting hori-  
459 zon expanded. Generally, the forecasts tended to overestimate the number  
460 of cases to be expected in the following weeks, as did most other forecasts  
461 generated during the outbreak (Chretien et al., 2015). This is in line with  
462 previous findings where our model was applied to predict simulated data of  
463 a hypothetical Ebola outbreak (Funk et al., 2017a). Log-transforming the  
464 transmission rate in order to ensure positivity skewed the underlying dis-  
465 tribution and made very high values possible. Moreover, we did not model  
466 a trend in the transmission rate, whereas in reality transmission decreased  
467 over the course of the epidemic, probably due to a combination of factors  
468 ranging from better provision of isolation beds to increasing awareness of  
469 the outbreak and subsequent behavioural changes. While our model cap-  
470 tured changes in the transmission rate in model fits, it did not forecast any  
471 trends such as the observed decrease over time. Capturing such trends and  
472 modelling the underlying causes would be an important future improvement  
473 of real-time infectious disease models used for forecasting.

474 There are trade-offs between achieving good outcomes for the different  
475 forecast metrics we used. Deciding whether the best forecast is the best cal-  
476 ibrated, the sharpest or the least biased, or some compromise between the  
477 three, is not a straightforward task. Our assessment of forecasts using sep-  
478 arate metrics for probabilistic calibration, sharpness and bias highlights the  
479 underlying trade-offs. While the best calibrated semi-mechanistic model

480 variant showed better calibration performance than the null models, this  
481 came at the expense of a decrease in the sharpness of forecasts. Compar-  
482 ing the models using the RPS alone, the semi-mechanistic model of best  
483 calibration performance would not necessarily have been chosen. Following  
484 the paradigm of maximising sharpness subject to calibration, we therefore  
485 recommend to treat probabilistic calibration as a prerequisite to the use of  
486 forecasts, in line with what has recently been suggested for post-processing  
487 of forecasts (Wilks, 2018). Probabilistic calibration is essential for mak-  
488 ing meaningful probabilistic statements (such as the chances of seeing the  
489 number of cases exceed a set threshold in the upcoming weeks) that enable  
490 realistic assessments of resource demand, the possible future course of the  
491 epidemic including worst-case scenarios, as well as the potential impact of  
492 public health measures. Once calibration is ensured, other criteria such as  
493 the RPS or DSS can be used to select the best model for forecasts, or to  
494 generate weights for ensemble forecasts combining several models. Such en-  
495 semble forecasts have become a standard in weather forecasting (Gneiting  
496 and Raftery, 2005) and have more recently shown promise for infectious  
497 disease forecasts (Yamana et al., 2016; Yamana et al., 2017; Viboud et al.,  
498 2017).

499 Other models may have performed better than the ones presented here.  
500 Because we did not have access to data that would have allowed us to assess  
501 the importance of different transmission routes (burials, hospitals and the  
502 community) we relied on a relatively simple, flexible model. The determinis-  
503 tic SEIR model we used as a null model performed poorly on all forecasting  
504 scores, and failed to capture the downturn of the epidemic in Western Area.  
505 On the other hand, a well-calibrated mechanistic model that accounts for  
506 all relevant dynamic factors and external influences could, in principle, have



507 been used to predict the behaviour of the epidemic reliably and precisely.  
508 Yet, lack of detailed data on transmission routes and risk factors precluded  
509 the parameterisation of such a model and are likely to do so again in future  
510 epidemics in resource-poor settings. Future work in this area will need to  
511 determine the main sources of forecasting error, whether structural, obser-  
512 vational or parametric, as well as strategies to reduce such errors (Pei and  
513 Shaman, 2017).

514 In practice, there might be considerations beyond performance when  
515 choosing a model for forecasting. Our model combined a mechanistic  
516 core (the SEIR model) with non-mechanistic variable elements. By using  
517 a flexible non-parametric form of the time-varying transmission rate, the  
518 model provided a good fit to the case series despite a high levels of uncer-  
519 tainty about the underlying process. Having a model with a mechanistic  
520 core came with the advantage of enabling the assessment of interventions  
521 just as with a traditional mechanistic model. For example, the impact of a  
522 vaccine could be modelled by moving individuals from the susceptible into  
523 the recovered compartment (Camacho et al., 2015a; Camacho et al., 2017).  
524 At the same time, the model was flexible enough to visually fit a wide variety  
525 of time series, and this flexibility might mask underlying misspecifications.  
526 More generally, when choosing between forecast performance and the ability  
527 to explicitly account for the impact of interventions, a model that accounts  
528 for the latter might, in some cases, be preferable. Where possible, the guid-  
529 ing principle in assessing real-time models and predictions for public health  
530 should be the quality of the recommended decisions based on the model  
531 results (Probert et al., 2018).

532 Epidemic forecasts played a prominent role in the response to and public

533 awareness of the Ebola epidemic (Frieden and Damon, 2015). Forecasts have  
534 been used for vaccine trial planning against Zika virus (World Health Orga-  
535 nization, 2017) and will be called upon again to inform the response to the  
536 next emerging epidemic or pandemic threat. Recent advances in computa-  
537 tional and statistical methods now make it possible to fit models in near-real  
538 time, as demonstrated by our weekly forecasts (Center for the Mathematical  
539 Modelling of Infectious Diseases, 2015). Such repeated forecasts are a pre-  
540 requisite for the use of metrics that assess not only how close the predictions  
541 were to reality, but also how well uncertainty in the predictions has been  
542 quantified. An agreement on standards of forecast assessment is urgently  
543 needed in infectious disease epidemiology, and retrospective or even real-  
544 time assessment of forecasts should become standard for epidemic forecasts  
545 to prove accuracy and improve end-user trust. The metrics we have used  
546 here or variations thereof could become measures of forecasting performance  
547 that are routinely used to evaluate and improve forecasts during epidemics.

548 For forecast assessment to happen in practice, evaluation strategies must  
549 be planned before the forecasts are generated. In order for such evaluation  
550 to be performed retrospectively, all forecasts as well as the data, code and  
551 models they were based on should be made public at the time, or at least  
552 preserved and decisions recorded for later analysis. We published weekly up-  
553 dated aggregate graphs and numbers during the Ebola epidemic, yet for full  
554 transparency it would have been preferable to allow individuals to download  
555 raw forecasts for further analysis.

556 If forecasts are not only produced but also evaluated in real time, this  
557 can give valuable insights into strengths, limitations, and reasonable time  
558 horizons. In our case, by tracking the performance of our forecasts, we

559 would have noticed the poor calibration of the model variant chosen for  
560 the forecasts presented to the public, and instead selected better calibrated  
561 variants. At the same time, we did not store the predictive distribution  
562 samples for any area apart from Western Area in order to better use available  
563 storage space, and because we did not deem such storage valuable at the  
564 time. This has precluded a broader investigation of the performance of our  
565 forecasts.

566 At the same time, research into modelling and forecasting methodology  
567 and predictive performance at times during which there is no public health  
568 emergency should be part of pandemic preparedness activities. To facilitate  
569 this, outbreak data must be made available openly and rapidly. Where avail-  
570 able, combination of multiple sources, such as epidemiological and genetic  
571 data, could increase predictive power. It is only on the basis of systematic  
572 and careful assessment of forecast performance during and after the event  
573 that predictive ability of computational models can be improved and lessons  
574 be learned to maximise their utility in future epidemics.

## 575 **References**

- 576 Andrieu, C., A. Doucet, and R. Holenstein (2010). “Particle Markov chain  
577 Monte Carlo methods”. *J R Stat Soc B*, 269–342.
- 578 Biggerstaff, M. et al. (2016). “Results from the centers for disease control and  
579 prevention’s predict the 2013–2014 Influenza Season Challenge”. *BMC  
580 Infectious Diseases* 1, 357.
- 581 Camacho, A. et al. (Dec. 2015a). “Estimating the probability of demonstrat-  
582 ing vaccine efficacy in the declining Ebola epidemic: a Bayesian modelling  
583 approach”. *BMJ Open* 12, e009346.

- 584 Camacho, A. et al. (2015b). “Temporal Changes in Ebola Transmission in  
585 Sierra Leone and Implications for Control Requirements: a Real-Time  
586 Modelling Study”. *PLOS Curr.: Outbreaks*.
- 587 Camacho, A. et al. (Dec. 2017). “Real-time dynamic modelling for the design  
588 of a cluster-randomized phase 3 Ebola vaccine trial in Sierra Leone”.  
589 *Vaccine*.
- 590 Center for the Mathematical Modelling of Infectious Diseases  
591 (2015). *Visualisation and projections of the Ebola outbreak*  
592 *in West Africa*. <http://ntncmch.github.io/ebola/>. Archived at  
593 <http://www.webcitation.org/6oYEBYoeD> on Feb 24, 2017.
- 594 Centres for Disease Control and Prevention (Oct.  
595 2017). *Epidemic Prediction Initiative*. URL:  
596 <https://predict.phiresearchlab.org/legacy/dengue/index.html>, Archived  
597 at <http://www.webcitation.org/6rsS5QDar> on 11 July, 2017.
- 598 Chowell, G. (2017). “Fitting dynamic models to epidemic outbreaks with  
599 quantified uncertainty: A primer for parameter uncertainty, identifiabil-  
600 ity, and forecasts”. *Infectious Disease Modelling* 3, 379–398.
- 601 Chowell, G. et al. (2017). “Perspectives on model forecasts of the 2014–2015  
602 Ebola epidemic in West Africa: lessons and the way forward”. *BMC Med*  
603 1, 42.
- 604 Chretien, J.-P., S. Riley, and D. B. George (Dec. 2015). “Mathematical  
605 modeling of the West Africa Ebola epidemic”. *eLife*, e09186.
- 606 Czado, C., T. Gneiting, and L. Held (2009). “Predictive model assessment  
607 for count data”. *Biometrics* 4, 1254–1261.
- 608 Dawid, A. P. (1984). “Present Position and Potential Developments: Some  
609 Personal Views: Statistical Theory: The Prequential Approach”. *J R Stat*  
610 *Soc [Ser A]* 2, 278.

- 611 Dawson, P. M. et al. (2015). “Epidemic predictions in an imperfect world:  
612 modelling disease spread with partial data”. In: *Proc. R. Soc. B*. 1808.  
613 The Royal Society, 20150205.
- 614 Dong, F. et al. (Dec. 2015). “Evaluation of ebola spreading in west africa and  
615 decision of optimal medicine delivery strategies based on mathematical  
616 models”. *Infection, Genetics and Evolution*, 35–40.
- 617 Drake, J. M. et al. (2015). “Ebola cases and health system demand in  
618 Liberia”. *PLoS Biol* 1, e1002056.
- 619 Dureau, J., S. Ballesteros, and T. Bogich (2013). “SSM: Inference for time  
620 series analysis with State Space Models”.
- 621 Epstein, E. S. (1969). “A scoring system for probability forecasts of ranked  
622 categories”. *Journal of Applied Meteorology* 6, 985–987.
- 623 Fisman, D., E. Khoo, and A. Tuite (2014). “Early epidemic dynamics of  
624 the west african 2014 ebola outbreak: estimates derived with a simple  
625 two-parameter model.” *PLOS Curr.: Outbreaks*.
- 626 Frieden, T. R. and I. K. Damon (2015). “Ebola in West Africa — CDC’s  
627 role in epidemic detection, control, and prevention”. *Emerging Infectious  
628 Diseases* 11, 1897.
- 629 Friederichs, P. and T. L. Thorarinsdottir (2012). “Forecast verification for  
630 extreme value distributions with an application to probabilistic peak  
631 wind prediction”. *Environmetrics* 7, 579–594.
- 632 Funk, S. (2016). *rbi.helpers: rbi helper functions*. R package version 0.2.
- 633 Funk, S. et al. (Dec. 2017a). “Real-time forecasting of infectious disease  
634 dynamics with a stochastic semi-mechanistic model”. *Epidemics*.
- 635 Funk, S. et al. (2017b). “The impact of control strategies and behavioural  
636 changes on the elimination of Ebola from Lofa County, Liberia”. *Phil  
637 Trans Roy Soc B* (1721), 20160302.

- 638 Gneiting, T., F. Balabdaoui, and A. E. Raftery (Apr. 2007). “Probabilistic  
639 forecasts, calibration and sharpness”. *Journal of the Royal Statistical  
640 Society: Series B (Statistical Methodology)* 2, 243–268.
- 641 Gneiting, T. and M. Katzfuss (Jan. 2014). “Probabilistic Forecasting”. *An-  
642 nual Review of Statistics and Its Application* 1, 125–151.
- 643 Gneiting, T. and A. E. Raftery (2005). “Weather forecasting with ensemble  
644 methods”. *Science* 5746, 248–249.
- 645 Goldstein, E. et al. (2011). “Predicting the epidemic sizes of influenza  
646 A/H1N1, A/H3N2, and B: a statistical method”. *PLoS Med* 7, e1001051.
- 647 Heesterbeek, H. et al. (2015). “Modeling Infectious Disease Dynamics in the  
648 Complex Landscape of Global Health”. *Science* 6227, aaa4339–aaa4339.
- 649 Held, L., S. Meyer, and J. Bracher (June 2017). “Probabilistic forecasting  
650 in infectious disease epidemiology: the 13th Armitage lecture”. *Statistics  
651 in Medicine* 22, 3443–3460.
- 652 Jacob, P. E. and S. Funk (2017). *RBi: R interface to LibBi*. R package  
653 version 0.7.0.
- 654 Johansson, M. A. et al. (2016). “Evaluating the performance of infectious  
655 disease forecasts: A comparison of climate-driven and seasonal dengue  
656 forecasts for Mexico”. *Scientific reports*.
- 657 Lewnard, J. A. et al. (Dec. 2014). “Dynamics and control of Ebola virus  
658 transmission in Montserrado, Liberia: a mathematical modelling analy-  
659 sis.” *Lancet Infect Dis* 12, 1189–1195.
- 660 Liu, F. et al. (2015). “Short-term forecasting of the prevalence of trachoma:  
661 expert opinion, statistical regression, versus transmission models”. *PLoS  
662 neglected tropical diseases* 8, e0004000.

- 663 Lowe, R. et al. (2014). “Dengue outlook for the World Cup in Brazil: an early  
664 warning model framework driven by real-time seasonal climate forecasts”.  
665 *The Lancet infectious diseases* 7, 619–626.
- 666 Maronna, R. et al. (2018). *Robust Statistics: Theory and Methods (with R)*.  
667 Wiley. ISBN: 9781119214687.
- 668 Meltzer, M. I. et al. (Sept. 2014). “Estimating the future number of cases  
669 in the Ebola epidemic—Liberia and Sierra Leone, 2014–2015.” *MMWR*  
670 *Surveill Summ*, 1–14.
- 671 Merler, S. et al. (Feb. 2015). “Spatiotemporal spread of the 2014 outbreak of  
672 Ebola virus disease in Liberia and the effectiveness of non-pharmaceutical  
673 interventions: a computational modelling analysis”. *Lancet Infect Dis* 2,  
674 204–211.
- 675 Moran, K. R. et al. (2016). “Epidemic forecasting is messier than weather  
676 forecasting: The role of human behavior and internet data streams in  
677 epidemic forecast”. *The Journal of Infectious Diseases* suppl\_4, S404–  
678 S408.
- 679 Murphy, A. H. (1969). “On the “ranked probability score””. *Journal of Ap-*  
680 *plied Meteorology* 6, 988–989.
- 681 Murray, L. (2015). “Bayesian State-Space Modelling on High-Performance  
682 Hardware Using LibBi”. *Journal of Statistical Software, Articles* 10, 1–  
683 36. ISSN: 1548-7660.
- 684 National Oceanic and Atmospheric Administration (Oct. 2017). *Dengue*  
685 *Forecasting*. URL: <http://dengueforecasting.noaa.gov/>, Archived at  
686 <http://www.webcitation.org/6oWfUBKnC> on Feb 24, 2017.
- 687 Nishiura, H. and G. Chowell (2014). “Early transmission dynamics of Ebola  
688 virus disease (EVD), West Africa, March to August 2014”. *Euro Surveill*  
689 (36), 20894.

- 690 Nsoesie, E., M. Mararthe, and J. Brownstein (2013). “Forecasting peaks of  
691 seasonal influenza epidemics”. *PLoS currents*.
- 692 Pei, S. and J. Shaman (Oct. 2017). “Counteracting structural errors in en-  
693 semble forecast of influenza outbreaks”. *Nature Communications* 1.
- 694 Probert, W. J. M. et al. (July 2018). “Real-time decision-making during  
695 emergency disease outbreaks”. *PLoS Computational Biology* 7. Ed. by  
696 K. Koelle, e1006202.
- 697 R Core Team (2018). *R: A Language and Environment for Statistical Com-  
698 puting*. R Foundation for Statistical Computing. Vienna, Austria.
- 699 Rivers, C. M. et al. (2014). “Modeling the impact of interventions on an  
700 epidemic of Ebola in Sierra Leone and Liberia”. *PLOS Curr.: Outbreaks*.
- 701 Scarpino, S. V. and G. Petri (Mar. 21, 2017). “On the predictability of  
702 infectious disease outbreaks”.
- 703 Scott, S. L. (2017). *bsts: Bayesian Structural Time Series*. R package version  
704 0.7.1.
- 705 Siettos, C. et al. (2015). “Modeling the 2014 ebola virus epidemic—agent-  
706 based simulations, temporal analysis and future predictions for liberia  
707 and sierra leone”. *PLOS Curr.: Outbreaks*.
- 708 Towers, S., O. Patterson-Lomba, and C. Castillo-Chavez (2014). “Temporal  
709 variations in the effective reproduction number of the 2014 West Africa  
710 Ebola outbreak”. *PLOS Curr.: Outbreaks*.
- 711 Viboud, C. et al. (Aug. 2017). “The RAPIDD ebola forecasting challenge:  
712 Synthesis and lessons learnt”. *Epidemics*.
- 713 White, R. A. et al. (2015). “Projected treatment capacity needs in Sierra  
714 Leone”. *PLOS Curr.: Outbreaks*.



- 715 WHO Ebola Response Team (Sept. 2014). “Ebola Virus Disease in West  
716 Africa - The First 9 Months of the Epidemic and Forward Projections.”  
717 *N Engl J Med*.
- 718 Wilks, D. S. (2018). “Enforcing calibration in ensemble postprocessing”.  
719 *Quarterly Journal of the Royal Meteorological Society* 710, 76–84.
- 720 World Health Organization (2017). *Efficacy trials of ZIKV Vaccines: end-  
721 points, trial design, site selection. WHO Workshop Meeting Report*.
- 722 Yamana, T. K., S. Kandula, and J. Shaman (Oct. 2016). “Superensemble  
723 forecasts of dengue outbreaks”. *Journal of The Royal Society Interface*  
724 123, 20160410.
- 725 — (Nov. 2017). “Individual versus superensemble forecasts of seasonal in-  
726 fluenza outbreaks in the United States”. *PLOS Computational Biology*  
727 11. Ed. by J. Lessler, e1005801.
- 728 Yang, W. et al. (2015). “Forecasting influenza epidemics in Hong Kong”.  
729 *PLoS computational biology* 7, e1004383.