

LONDON  
SCHOOL of  
HYGIENE  
& TROPICAL  
MEDICINE



Funk, Sebastian; Camacho, Anton; Kucharski, Adam; Lowe, Rachel; Eggo, Rosalind; Edmunds, John (2017) Assessing the performance of real-time epidemic forecasts: A case study of the 2013-16 Ebola epidemic. bioRxiv. DOI: <https://doi.org/10.1101/177451> (Unpublished)

Downloaded from: <http://researchonline.lshtm.ac.uk/4650144/>

DOI: [10.1101/177451](https://doi.org/10.1101/177451)

#### Usage Guidelines

Please refer to usage guidelines at <http://researchonline.lshtm.ac.uk/policies.html> or alternatively contact [researchonline@lshtm.ac.uk](mailto:researchonline@lshtm.ac.uk).

Available under license: <http://creativecommons.org/licenses/by/2.5/>

1                                 Assessing the performance of  
2   real-time epidemic forecasts:  
3                     A case study of the 2013–16 Ebola epidemic

4         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>

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

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

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

10         <sup>4</sup> Barcelona Institute for Global Health (ISGLOBAL), Barcelona, Spain

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

12   **Abstract**

13             Real-time forecasts based on mathematical models can inform criti-  
14             cal decision-making during infectious disease outbreaks. Yet, epidemic  
15             forecasts are rarely evaluated during or after the event, and there is  
16             little guidance on what the best metrics for assessment are. Here,  
17             we propose to disentangle different components of forecasting ability  
18             by using metrics that separately assess the calibration, sharpness and  
19             unbiasedness of forecasts. We used this approach to analyse the per-  
20             formance of weekly forecasts generated in real time in Western Area,  
21             Sierra Leone, during the 2013–16 Ebola epidemic in West Africa. We  
22             found that probabilistic calibration was good at short time horizons  
23             but deteriorated for long-term forecasts. This suggests that forecasts  
24             provided usable performance only a few weeks ahead of time, reflecting  
25             the high level of uncertainty in the processes driving the trajectory of  
26             the epidemic. Comparing the semi-mechanistic model we used during  
27             the epidemic to simpler null models showed that the our model per-  
28             formed better with respect to probabilistic calibration, and that this  
29             would have been identified from the earliest stages of the outbreak.  
30             As forecasts become a routine part of the toolkit in public health,  
31             standards for evaluation of performance will be important for assess-  
32             ing quality and improving credibility of mathematical models, and for  
33             elucidating difficulties and trade-offs when aiming to make the most  
34             useful and reliable forecasts.

## 35 Introduction

36 Forecasting the future trajectory of cases during an infectious disease out-  
37 break can make an important contribution to public health and interven-  
38 tion planning. Infectious disease modellers are now routinely asked for  
39 predictions in real time during emerging outbreaks (Heesterbeek et al.,  
40 2015). Forecasting targets usually revolve around expected epidemic du-  
41 ration, size, or peak timing and incidence (Goldstein et al., 2011; Nsoesie  
42 et al., 2013; Yang et al., 2015; Dawson et al., 2015), geographical distribu-  
43 tion of risk (Lowe et al., 2014), or short-term trends in incidence (Johansson  
44 et al., 2016; Liu et al., 2015). Despite the increase in activity, however,  
45 forecasts made during an outbreak is rarely investigated during or after the  
46 event for their accuracy.

47 The growing importance of infectious disease forecasts is epitomised by  
48 the growing number of so-called forecasting challenges. In these, researchers  
49 compete in making predictions for a given disease and a given time hori-  
50 zon. Such initiatives are difficult to set up during unexpected outbreaks,  
51 and are therefore usually conducted on diseases known to occur seasonally,  
52 such as dengue (Johansson et al., 2016; National Oceanic and Atmospheric  
53 Administration, 2017; Centres for Disease Prevention and Control, 2017)  
54 and influenza (Biggerstaff et al., 2016). The *Ebola forecasting challenge* was  
55 a notable exception, triggered by the 2013–16 West African Ebola epidemic  
56 and set up in June 2015. Since the epidemic had ended in most places at  
57 that time, the challenge was based on simulated data designed to mimic the  
58 behaviour of the true epidemic instead of real outbreak data (Viboud et al.,  
59 2017).

60 Providing accurate forecasts during emerging epidemics comes with par-  
61 ticular challenges as uncertainties about the processes driving growth and  
62 decline in cases, in particular human behavioural changes and public health  
63 interventions, can preclude reliable long-term predictions (Moran et al.,  
64 2016; Funk et al., 2017b). Short-term forecasts with an horizon of a few  
65 generations of transmission (e.g., a few weeks in the case of Ebola), on the  
66 other hand, can yield important information on current and anticipated  
67 outbreak behaviour and, consequently, guide immediate decision making.

68 The most recent example of large-scale outbreak forecasting efforts was  
69 during the 2013–16 Ebola epidemic, which vastly exceeded the burden of  
70 all previous outbreaks with almost 30,000 reported cases of the disease, re-  
71 sulting in over 10,000 deaths in the three most affected countries: Guinea,  
72 Liberia and Sierra Leone. During the epidemic, several research groups pro-  
73 vided forecasts or projections at different time points, either by generating  
74 scenarios believed plausible, or by fitting models to the available time series  
75 and projecting them forward to predict the future trajectory of the out-

76 break (Fisman et al., 2014; Lewnard et al., 2014; Nishiura and Chowell,  
77 2014; Rivers et al., 2014; Towers et al., 2014; Camacho et al., 2015b; Dong  
78 et al., 2015; Drake et al., 2015; Merler et al., 2015; Siettos et al., 2015; White  
79 et al., 2015). (Chretien et al., 2015; Chowell et al., 2017). One forecast that  
80 gained attention during the epidemic was published in the summer of 2014,  
81 projecting that by early 2015 there might be 1.4 million cases (Meltzer et al.,  
82 2014). While this number was based on unmitigated growth in the absence  
83 of further intervention and proved a gross overestimate, it was later high-  
84 lighted as a “call to arms” that served to trigger the international response  
85 that helped avoid the worst-case scenario (Frieden and Damon, 2015).

86 Traditionally, epidemic forecasts are assessed using aggregate metrics  
87 such as the mean absolute error (MAE, Chowell, 2017; Pei and Shaman,  
88 2017; Viboud et al., 2017). These, however, often only assess how close the  
89 most likely or average predicted outcome is to the true outcome. The ability  
90 to correctly forecast uncertainty, and to quantify confidence in a predicted  
91 event, is not assessed by such metrics. Appropriate quantification of uncer-  
92 tainty, especially of the likelihood and magnitude of worst case scenarios,  
93 is crucial in assessing potential control measures. Methods to assess proba-  
94 bilistic forecasts are now being used in other fields, but are not commonly  
95 applied in infectious disease epidemiology (Gneiting and Katzfuss, 2014;  
96 Held et al., 2017). It is worth noting that good predictive ability need not  
97 coincide with good fit, as statistical evidence may not translate into forecast  
98 capability because of model uncertainty and noisy, incomplete data.

99 We produced weekly sub-national real-time forecasts during the Ebola  
100 epidemic, starting on 28 November 2014. These were published on a dedi-  
101 cated web site and updated every time a new set of data were available (Cen-  
102 ter for the Mathematical Modelling of Infectious Diseases, 2015). They were  
103 generated using a model that has, in variations, been used to forecast bed  
104 demand during the epidemic in Sierra Leone (Camacho et al., 2015b) and  
105 the feasibility of vaccine trials later in the epidemic (Camacho et al., 2015a;  
106 Camacho et al., 2017). During the epidemic, we provided sub-national fore-  
107 casts for three most affected countries (at the level of counties in Liberia,  
108 districts in Sierra Leone and prefectures in Guinea).

109 Here, we apply assessment metrics that elucidate different properties of  
110 forecasts, in particular their probabilistic calibration, sharpness and bias.  
111 Using these methods, we retrospectively assess the forecasts we generated  
112 for Western Area in Sierra Leone, an area that saw one of the greatest  
113 number of cases in the region and where our model informed bed capacity  
114 planning.

## 115 **Materials and Methods**

### 116 **Data sources**

117 Numbers of suspected, probable and confirmed cases at sub-national levels  
118 were initially compiled from daily *Situation Reports* (or *SitReps*) provided  
119 in PDF format by Ministries of Health of the three affected countries during  
120 the epidemic (Camacho et al., 2015b). Data were automatically extracted  
121 from tables included in the reports wherever possible and otherwise man-  
122 ually converted by hand to machine-readable format and aggregated into  
123 weeks. From 20 November 2014, the World Health Organization (WHO)  
124 provided tabulated data on the weekly number of confirmed and probable  
125 cases. These were compiled from the patient database, which was contin-  
126 uously cleaned and took into account reclassification of cases avoiding po-  
127 tential double-counting. However, the patient database was updated with  
128 substantial delay so that the number of reported cases would typically be  
129 underestimated in the weeks leading up to the date of the forecast. Because  
130 of this, we used the SitRep data for the most recent weeks until the latest  
131 week in which the WHO case counts either equalled or exceeded the SitRep  
132 counts. For all earlier times, the WHO data were used.

### 133 **Transmission model**

134 We used a semi-mechanistic stochastic model of Ebola transmission de-  
135 scribed previously (Camacho et al., 2015b; Funk et al., 2017a). Briefly,  
136 the model was based on a Susceptible-Exposed-Infectious-Recovered (SEIR)  
137 model with fixed incubation period of 9.4 days (WHO Ebola Response Team,  
138 2014), following an Erlang distribution with shape 2. The country-specific  
139 infectious period was determined by adding the average delay to hospitalisa-  
140 tion to the average time from hospitalisation to death or discharge, weighted  
141 by the case-fatality rate. Cases were assumed to be reported with a stochas-  
142 tic time-varying delay. On any given day, this was given by a gamma distri-  
143 bution with mean equal to the country-specific average delay from onset to  
144 hospitalisation and standard deviation of 0.1 day. We allowed transmission  
145 to vary over time, in order to be able to capture behavioural changes in the  
146 community, public health interventions or other factors affecting transmis-  
147 sion for which information was not available at the time. The time-varying  
148 transmission rate was modelled using a daily Gaussian random walk with  
149 fixed volatility (or standard deviation of the step size) which was estimated  
150 as part of the inference procedure (see below). To ensure the transmission  
151 rate remained positive, we log-transformed it, so that its behaviour in time

152 can be written as

$$153 \quad d \log \beta_t = \sigma dW_t \quad (1)$$

154 where  $\beta_t$  is the time-varying transmission rate,  $W_t$  is the Wiener process  
155 and  $\sigma$  the volatility of the transmission rate. In fitting the model to the  
156 time series of cases we extracted posterior predictive samples of trajectories,  
157 which we used to generate forecasts.

## 158 **Model fitting**

159 Each week, we fitted the model to the available case data leading up to  
160 the date of the forecast. Observations were assumed to follow a negative  
161 binomial distribution, approximated as a discretised normal distribution for  
162 numerical convenience. Four parameters were estimated in the process: the  
163 basic reproduction number  $R_0$  (uniform prior within (1, 5)), initial num-  
164 ber of infectious people (uniform prior within (1, 400)), overdispersion of  
165 the (negative binomial) observation process (uniform prior within (0, 0.5))  
166 and volatility of the time-varying transmission rate (uniform prior within  
167 (0, 0.5)). We confirmed from the posterior distributions of the parameters  
168 that these priors did not set any problematic bounds. Samples of the pos-  
169 terior distribution of parameters and state trajectories were extracted using  
170 particle Markov chain Monte Carlo (Andrieu et al., 2010) as implemented  
171 in the *ssm* library (Dureau et al., 2013). For each forecast, 50,000 samples  
172 were extracted and thinned to 5000.

## 173 **Predictive model variants**

174 We used the samples of the posterior distribution generated using the Monte  
175 Carlo sampler to produce a range of predictive trajectories, using the final  
176 values of estimated state trajectories as initial values for the forecasts and  
177 simulating the model forward for up to 10 weeks. While all model fits were  
178 generated using the same model described above, we tested a range of dif-  
179 ferent predictive model variants to assess the quality of ensuing predictions.  
180 We tested variants where trajectories were stochastic (with demographic  
181 stochasticity and a noisy reporting process), as well as ones where these  
182 sources of noise were removed for predictions. We further tested predictive  
183 model variants where the transmission rate continued to follow a random  
184 walk (unbounded, on a log-scale), as well as ones where the transmission rate  
185 stayed fixed during the forecasting period. Where the transmission rate re-  
186 mained fixed for prediction, we tested variants where we used the final value  
187 of the transmission rate and ones where this value would be averaged over

188 a number of weeks leading up to the final fitted point, to reduce the poten-  
189 tial influence of the last time point, where the transmission rate may not  
190 have been well identified. We tested variants where the predictive trajectory  
191 would be based on the final values and start at the last time point, and ones  
192 where they would start at the penultimate time point, which could, again,  
193 be expected to be better informed by the data. For each model and forecast  
194 horizon, we generated point-wise medians and credible intervals from the  
195 sample trajectories.

## 196 Null models

197 To assess the performance of the semi-mechanistic transmission model we  
198 compared it to simpler null models: two representing the constituent parts  
199 of the semi-mechanistic model, and a non-mechanistic time series model.  
200 As first null model, we used a *deterministic* model that only contained the  
201 mechanistic core of the semi-mechanistic model with a fixed transmission  
202 rate. As second null model, we used an *unfocused* model where the num-  
203 ber of cases itself was modelled using a stochastic volatility model (without  
204 drift), that is a daily Gaussian random walk, and forecasts generated as-  
205 suming the weekly number of new cases was not going to change. Lastly, we  
206 used a null model based on a non-mechanistic Bayesian *autoregressive* linear  
207 model. The deterministic and models were implemented in *libbi* (Murray,  
208 2015) via the *RBi* (Jacob and Funk, 2017) and *RBi.helpers* (Funk, 2016) *R*  
209 packages (R Core Team, 2017). The autoregressive model was implemented  
210 using the *bsts* package (Scott, 2017).

## 211 Metrics

212 The paradigm for assessing probabilistic forecasts is that they should max-  
213 imise the sharpness of predictive distributions subject to calibration (Gneit-  
214 ing et al., 2007). We therefore first assessed whether models were calibrated  
215 at a given forecasting horizon, before assessing their sharpness and other  
216 properties.

217 *Calibration* or reliability (Friederichs and Thorarinsdottir, 2012) of fore-  
218 casts is the ability of a model to correctly identify its own uncertainty in  
219 making predictions. In a perfectly calibrated model, the data at each time  
220 point look as if they came from the predictive probability distribution at  
221 that time. Equivalently, one can inspect the probability integral transform  
222 of the predictive distribution at time  $t$  (Dawid, 1984),

$$223 \quad u_t = F_t(x_t) \tag{2}$$

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

229 To assess calibration, we applied the Anderson-Darling test of unifor-  
230 mity to the probabilities  $u_t$ . The resulting p-value was a reflection of how  
231 compatible the forecasts were with the null hypothesis of uniformity of the  
232 PIT, or of the data coming from the predictive probability distribution. We  
233 considered a model to be calibrated if the p-value found was greater than a  
234 threshold of  $p \geq 0.1$ , possibly calibrated if  $0.01 < p < 0.1$ , and uncalibrated  
235 if  $p \leq 0.01$ .

236 *Sharpness* is the ability of the model to generate predictions within a  
237 narrow range of possible outcomes. It is a data-independent measure, that  
238 is, it is purely a feature of the forecasts themselves. To evaluate sharpness at  
239 time  $t$ , we used the median absolute deviation about the median (MADM)  
240 of  $y$

$$241 \quad S_t(F_t) = m(|y - m(y)|) \quad (3)$$

242 where  $y$  is a variable distributed according to  $F_t$ , and  $m(y)$  is the median  
243 of  $y$ . The sharpest model would focus all forecasts on one point and have  
244  $S = 0$ , whereas a completely blurred forecast would have  $S \rightarrow \infty$ . Again,  
245 we used Monte-Carlo samples  $X$  from  $F_t$  to estimate sharpness.

246 We further assessed the *bias* of forecasts to assess whether a model sys-  
247 tematically over- or underpredicted. We defined bias at time  $t$  as

$$248 \quad B_t(F_t, x_t) = 2 \left( \int_{-\infty}^{\infty} F_t(y) H(y - x_t) dy - 0.5 \right) \quad (4)$$

249 where  $H(x)$  is the Heaviside step function with the half-maximum conven-  
250 tion  $H(0) = 1/2$ . This metric is equivalent to

$$251 \quad B_t(F_t, x_t) = 2 (E_{F_t} [H(X - x_t)] - 0.5) \quad (5)$$

252 which can be estimated using a finite number of samples, such as the Monte-  
253 Carlo samples generated in our inference procedure. Here,  $x_t$  are the ob-  
254 served data points,  $E_{F_t}$  is the expectation with respect to the predictive  
255 CDF  $F_t$  and  $X$  are independent realisations of a variable with distribution  
256  $F_t$ . The most unbiased model would have exactly half of forecasts above or  
257 equal to the data at time  $t$  and  $B_t = 0$ , whereas a completely biased model  
258 would yield either all forecasts above ( $B_t = 1$ ) or below ( $B_t = -1$ ) the data.  
259 To get a single bias score  $U$ , we took the mean across forecast time

$$260 \quad B(F_t, x_t) = \frac{1}{T} \sum_t B_t(F_t, x_t), \quad (6)$$



261 where  $T$  is the number of forecasting time points.

262 Lastly, we evaluated forecasts using the *Continuous Ranked Probability*  
263 *Score* (CRPS, Hersbach, 2000). CRPS is a distance measure that measures  
264 forecasting performance at the scale of the predicted data, combining an  
265 assessment calibration and sharpness. It is a *strictly proper forecasting score*,  
266 that is one which is optimised if the predictive distribution is the same as  
267 the one generating the data, with 0 being the ideal score. CRPS reduces  
268 to the mean absolute error (MAE) if the forecast is deterministic and can  
269 therefore be seen as its probabilistic generalisation. It is defined as

$$270 \quad \text{CRPS}(F_t, x_t) = - \int_{-\infty}^{\infty} (F_t(y) - H(y - x_t))^2 dy, \quad (7)$$

271 A convenient equivalent formulation using independent samples from  $F_t$   
272 was suggested by Gneiting et al. (2007) and is given by

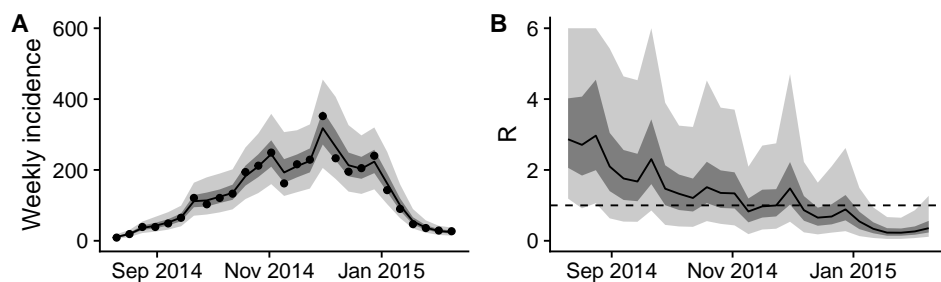
$$273 \quad \text{CRPS}(F_t, x_t) = E_{F_t} |X - x_t| - \frac{1}{2} E_{F_t} |X - X'|, \quad (8)$$

274 where  $X$  and  $X'$  are independent realisations of a random variable with  
275 CDF  $F_t$ .

## 276 Results

277 The semi-mechanistic model used to generate real-time forecasts during the  
278 epidemic was able to reproduce the trajectories up to the date of each fore-  
279 cast, following the data closely by means of the smoothly varying transmis-  
280 sion rate (Fig. 1). The overall behaviour of the reproduction number (ig-  
281 noring depletion of susceptibles which did not play a role at the population  
282 level given the relatively small proportion of the population infected) was  
283 one of a near-monotonic decline, from a median estimate of 2.9 (interquartile  
284 range (IQR) 2.2–3.8, 95% credible interval (CI) 1.1–7.8) in the first fitted  
285 week (beginning 10 August, 2014) to a median estimate of 1.3 (IQR 0.9–1.9,  
286 95% CI 0.3–3.9) in early October, 1.4 (IQR 1.0–2.0, 95% CI 0.4–4.6) in early  
287 November, 1.1 (IQR 0.7–1.4, 95% CI 0.2–3.0) in early December, 0.6 in early  
288 January (IQR 0.4–0.9, 95% CI 0.1–1.9) and 0.3 at the end of the epidemic  
289 in early February (IQR 0.2–0.5, 95% CI 0.1–1.3).

290 Forecasts from the semi-mechanistic model were calibrated for one or  
291 two weeks, but deteriorated rapidly at longer forecasting horizons (Table 1  
292 and Fig. 2). The two best calibrated models used deterministic forecasts  
293 starting at the last fitted data point. Of these two, forecasts that kept the  
294 transmission rate constant from the value at the last data point performed



**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.

295 slightly better than one that continued to change the transmission rate fol-  
296 lowing a random walk with volatility estimated from the time series. Both  
297 of the best calibrated models were calibrated for two-week ahead forecasts,  
298 and possibly calibrated for three weeks. All of the model variants were un-  
299 calibrated four weeks or more ahead, and none of the stochastic models was  
300 calibrated for any forecast horizon.

301 The best-calibrated of our semi-mechanistic forecasts was better cali-  
302 brated than any of the null models (Fig. 3A) for up to three weeks. While  
303 the autoregressive null model was calibrated for 1-week-ahead forecasts, it  
304 was not calibrated for longer forecast horizons. The unfocused null model,  
305 which assumes that the same number of cases will be reported in the weeks  
306 following the week during which the forecast was made, was only possibly  
307 calibrated for 1-week ahead and uncalibrated beyond. The deterministic  
308 null model was uncalibrated for all forecast horizons.

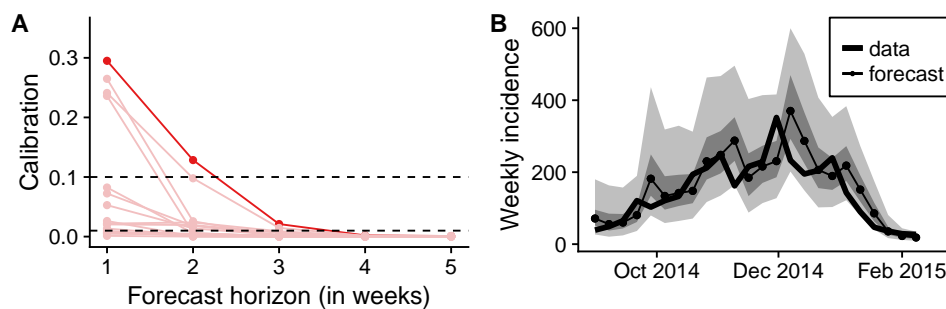
309 Our model as well as all null models except the unfocused model showed a  
310 tendency to overestimate the predicted number of cases (Fig. 3B). This bias  
311 increased with the forecast horizon. The best-calibrated semi-mechanistic  
312 model progressed from a 12% bias at 1 week ahead to 20% (2 weeks), 30% (3  
313 weeks), 40% (4 weeks) and 44% (5 weeks) overestimation. At the same  
314 time, this model showed rapidly decreasing sharpness as the forecast horizon  
315 increased (Fig. 3C). This is reflected in the mean CRPS values (Fig. 3D),  
316 which combine calibration and sharpness and reflect a probabilistic analogue

Model				Forecast horizon (weeks)			
stochasticity	start	averaged	volatility	1	2	3	4
deterministic	at last data point	no	yes	<b>0.24</b>	<b>0.1</b>	0.01	<0.01
deterministic	at last data point	no	no	<b>0.3</b>	<b>0.13</b>	0.02	<0.01
deterministic	at last data point	2 weeks	no	<b>0.26</b>	0.03	<0.01	<0.01
deterministic	at last data point	3 weeks	no	<b>0.24</b>	<0.01	<0.01	<0.01
deterministic	1 week before	no	yes	<b>0.05</b>	0.01	<0.01	<0.01
deterministic	1 week before	no	no	<b>0.07</b>	0.02	<0.01	<0.01
deterministic	1 week before	2 weeks	no	<b>0.08</b>	<0.01	<0.01	<0.01
deterministic	1 week before	3 weeks	no	0.03	<0.01	<0.01	<0.01
stochastic	at last data point	no	yes	0.02	0.02	<0.01	<0.01
stochastic	at last data point	no	no	0.02	0.02	<0.01	<0.01
stochastic	at last data point	2 weeks	no	0.01	<0.01	<0.01	<0.01
stochastic	at last data point	3 weeks	no	<0.01	<0.01	<0.01	<0.01
stochastic	1 week before	no	yes	<0.01	<0.01	<0.01	<0.01
stochastic	1 week before	no	no	<0.01	<0.01	<0.01	<0.01
stochastic	1 week before	2 weeks	no	<0.01	<0.01	<0.01	<0.01
stochastic	1 week before	3 weeks	no	<0.01	<0.01	<0.01	<0.01

**Table 1. Calibration of forecast model variants of our semi-mechanistic model.** Shown is the calibration (p-value of the Anderson-Darling test of uniformity) for deterministic and stochastic forecasts starting either at the last data point or one week before, either starting from the last value of the transmission rate or from an average over the last 2 or 3 weeks, and including volatility (in a Gaussian random walk) in the transmission rate or not, at different forecast horizons up to 4 weeks. The p-values highlighted in bold reflect predictive models we consider likely to be calibrated.

317 to the MAE. At 1-week ahead, the mean CRPS values of the autoregressive,  
318 unfocused and best semi-mechanistic forecasting models were all around 30  
319 (i.e., on average the prediction was out by approximately 30 cases). At  
320 increasing forecasting horizon, the CRPS of the semi-mechanistic model  
321 grew faster than the CRPS of the autoregressive and unfocused null models,  
322 but since these were no longer calibrated at horizons longer than one week,  
323 the semi-mechanistic model would still be preferred for forecast horizons up  
324 to three weeks.

325 We studied the calibration behaviour of the models over time, that is  
326 using the data and forecasts available up to different time points during the  
327 epidemic (Fig. 4). This shows that from very early on, not much changed  
328 in the ranking of the different semi-mechanistic model variants. Comparing  
329 the best semi-mechanistic forecasting model to the null models, again, for  
330 almost the whole duration of the epidemic the semi-mechanistic model would



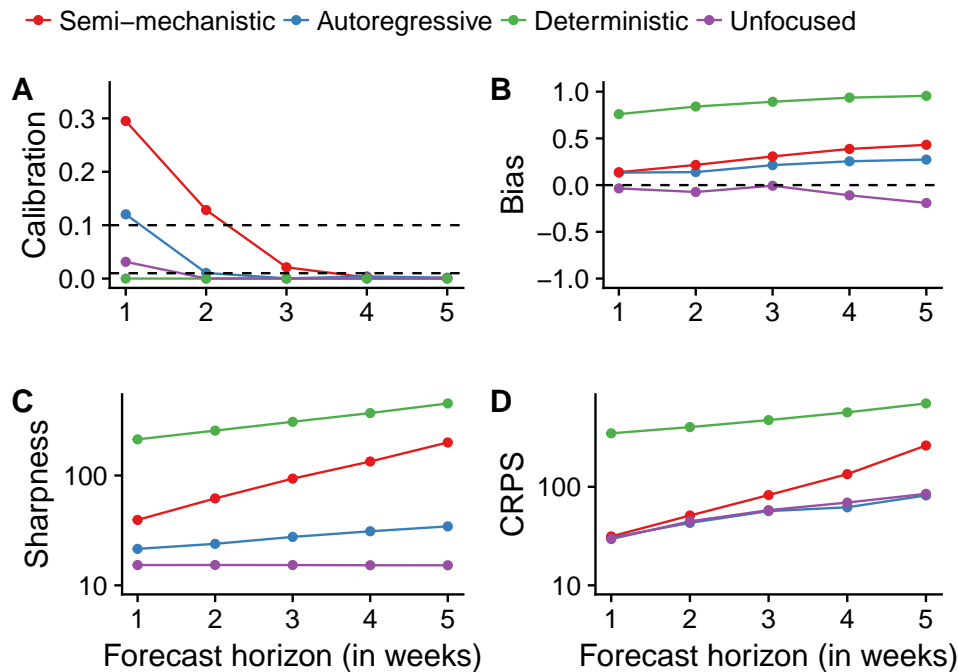
**Figure 2. Calibration of forecasts from the semi-mechanistic model.** (A) Calibration of model variants (p-value of Anderson-Darling test) as a function of the forecast horizon. Shown in dark red is the best calibrated forecasting model variant. Other model variants are shown in light red. (B) Comparison 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.

331 have been determined to be the best calibrated for forecasts 1 or 2 weeks  
332 ahead.

## 333 Discussion

334 Outbreaks of emerging infectious diseases in resource-poor settings are often  
335 characterised by limited data and a need for short-term forecasts to inform  
336 bed demands and allocation of other human and financial resources. Several  
337 groups produced and published forecasts over the course of the Ebola epi-  
338 demic, and the alleged failure of some to predict the correct number of cases  
339 by several orders of magnitude generated some controversy around the use-  
340 fulness of mathematical models (Butler, 2014; Rivers et al., 2014). To our  
341 knowledge, we were the only research team making weekly forecasts avail-  
342 able in real time, distributing them to a wide range of international public  
343 health practitioners via a dedicated email list, as well as on a publicly ac-  
344 cessible web page. Because we did not have access to data that would have  
345 allowed us to assess the importance of different transmission routes (buri-  
346 als, hospitals and the community) we relied on a relatively simple, flexible  
347 model.

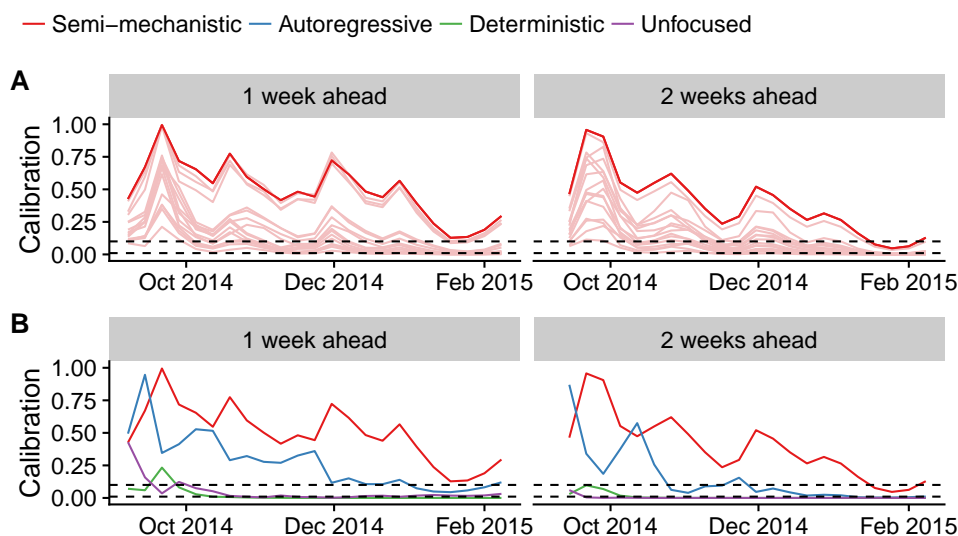
348 Applying a suite of assessment methods to our forecasting model, we  
349 found that the used semi-mechanistic model variants were probabilistically



**Figure 3. Forecasting metrics of the best semi-mechanistic model variant compared to null models.** Metrics shown are (A) calibration (p-value of Anderson-Darling test), (B) bias, (C) sharpness (MADM) and (D) CRPS, all as a function of the forecast horizon.

350 calibrated to varying degree with the best ones calibrated for up to 2-3  
 351 weeks ahead, but performance deteriorated rapidly as the forecasting horizon  
 352 increased. Since the model variants were similar enough to produce the same  
 353 mean future trajectories, differences in calibration reflected differences in the  
 354 quantification of uncertainty. The best performing forecasts were the once  
 355 generated the least variance in the trajectories, indicating that, in general,  
 356 our models overestimated the possible diversity in future trajectories. A  
 357 possible future improvement could be to post-process predictions by tuning  
 358 their variance to improve performance (Liu et al., 2015).

359 The rapid deterioration of probabilistic calibration even of our best per-  
 360 forming model variants reflects our lack of knowledge about the underlying  
 361 processes shaping the epidemic at the time, from public health interventions  
 362 by numerous national and international agencies to changes in individual and  
 363 community behaviour. During the epidemic, we only published forecasts up  
 364 to 3 weeks ahead, as longer forecasting horizons were not considered appro-  
 365 priate.



**Figure 4. Calibration over time.** Shown are 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.

366 Our forecasts suffered from bias that worsened as the forecasting horizon  
367 expanded. Generally, the forecasts tended to overestimate the number of  
368 cases to be expected in the following weeks. Log-transforming the transmis-  
369 sion rate in order to ensure positivity skewed the underlying distribution and  
370 made very high values possible. Moreover, we did not model a trend in the  
371 transmission rate, whereas in reality transmission decreased over the course  
372 of the epidemic, probably due to a combination of factors ranging from bet-  
373 ter provision of isolation beds to increasing awareness of the outbreak and  
374 subsequent behavioural changes. While our model captured changes in the  
375 transmission rate in model fits, it did not forecast any trends such as a  
376 the observed decrease over time. Capturing such trends and modelling the  
377 underlying causes would be an important future improvement of real-time  
378 infectious disease models used for forecasting.

379 There can be trade-offs between achieving good outcomes on the differ-  
380 ent forecast metrics we used, so that deciding whether the best forecast is  
381 the best calibrated, the sharpest or the least biased, or some compromise  
382 between the three, is not a straightforward task. Our assessment of fore-  
383 casts using separate metrics for probabilistic calibration, sharpness and bias  
384 highlights the underlying trade-offs. While the semi-mechanistic model we

385 used during the Ebola epidemic was better calibrated than the null mod-  
386 els, this came at the expense of a decrease in the sharpness of forecasts.  
387 Comparing the models using the CRPS alone, the best calibrated semi-  
388 mechanistic model would not necessarily have been chosen. Following the  
389 paradigm of maximising sharpness subject to calibration, we therefore rec-  
390 ommend to treat probabilistic calibration as a prerequisite to the use of  
391 forecasts, in line with what has recently been suggested for post-processing  
392 of ensemble forecasts (Wilks, 2018). Probabilistic calibration is essential for  
393 making meaningful probabilistic statements (such as the chances of seeing  
394 the number of cases exceed a set threshold in the upcoming weeks) that en-  
395 able realistic assessments of resource demand, the possible future course of  
396 the epidemic including worst-case scenarios, as well as the potential impact  
397 of public health measures.

398 Other models may have performed better than the ones presented here.  
399 The deterministic SEIR model we used as a null model performed poorly on  
400 all forecasting scores, and failed to capture the downturn of the epidemic in  
401 Western Area. On the other hand, a well-calibrated mechanistic model that  
402 accounts for all relevant dynamic factors and external influences could, in  
403 principle, have been used to predict the behaviour of the epidemic reliably  
404 and precisely. Yet, lack of detailed data on transmission routes and risk  
405 factors precluded the parameterisation of such a model and are likely to do  
406 so again in future epidemics in resource-poor settings. Future work in this  
407 area will need to determine the main sources of forecasting error, whether  
408 structural, observational or parametric, as well as strategies to reduce such  
409 errors (Pei and Shaman, 2017).

410 In practice, there might be considerations beyond performance when  
411 choosing a model for forecasting. Our model combined a mixture of a mech-  
412 anistic core (the SEIR model) with non-mechanistic variable elements. By  
413 using a flexible non-parametric form of the time-varying transmission rate,  
414 the model provided a good fit to the case series despite a high levels of uncer-  
415 tainty about the underlying process. At the same time, having a model with  
416 a mechanistic core came with the advantage of enabling the assessment of  
417 interventions just as with a traditional mechanistic model. For example, the  
418 impact of a vaccine could be modelled by moving individuals from the sus-  
419 ceptible into the recovered compartment (Camacho et al., 2015a; Camacho  
420 et al., 2017). At the same time, the model was flexible enough to visually  
421 fit a wide variety of time series, and this flexibility might mask underlying  
422 misspecifications. More generally, when choosing between forecast perfor-  
423 mance and the ability to explicitly account for the impact of interventions,  
424 a model that accounts for the latter might, in some cases, be preferable.

425 Epidemic forecasts played an important and prominent role in the re-  
426 sponse to and public awareness of the Ebola epidemic (Frieden and Damon,



427 2015). Forecasts have been used for vaccine trial planning against Zika  
428 virus (World Health Organization, 2017) and will be called upon again to  
429 inform the response to the next emerging epidemic or pandemic threat.  
430 Recent advances in computational and statistical methods now make it pos-  
431 sible to fit models in near-real time, as demonstrated by our weekly fore-  
432 casts (Center for the Mathematical Modelling of Infectious Diseases, 2015).  
433 An agreement on standards of forecasting assessment is urgently needed in  
434 infectious disease epidemiology, and retrospective or even real-time assess-  
435 ment of forecasts should become standard for epidemic forecasts to prove  
436 accuracy and improve end-user trust. The metrics we have used here or  
437 variations thereof could become measures of forecasting performance that  
438 are routinely used to evaluate and improve forecasts during epidemics. To  
439 facilitate this, outbreak data must be made available openly and rapidly.  
440 Where available, combination of multiple sources, such as epidemiological  
441 and genetic data, could increase predictive power. It is only on the basis of  
442 systematic and careful assessment of forecast performance during and after  
443 the event that predictive ability of computational models can be improved  
444 and lessons be learned to maximise their utility in future epidemics.

## 445 References

- 446 Andrieu, C., A. Doucet, and R. Holenstein (2010). “Particle Markov chain  
447 Monte Carlo methods”. *J R Stat Soc B*, 269–342.
- 448 Biggerstaff, M. et al. (2016). “Results from the centers for disease control and  
449 prevention’s predict the 2013–2014 Influenza Season Challenge”. *BMC*  
450 *Infectious Diseases* 1, 357.
- 451 Butler, D. (Nov. 2014). “Models overestimate Ebola cases.” *Nature* (7525),  
452 18. ISSN: 1476-4687.
- 453 Camacho, A. et al. (Dec. 2015a). “Estimating the probability of demonstrat-  
454 ing vaccine efficacy in the declining Ebola epidemic: a Bayesian modelling  
455 approach”. *BMJ Open* 12, e009346.
- 456 Camacho, A. et al. (2015b). “Temporal Changes in Ebola Transmission in  
457 Sierra Leone and Implications for Control Requirements: a Real-Time  
458 Modelling Study”. *PLOS Curr.: Outbreaks*.
- 459 Camacho, A. et al. (Dec. 2017). “Real-time dynamic modelling for the design  
460 of a cluster-randomized phase 3 Ebola vaccine trial in Sierra Leone”.  
461 *Vaccine*.
- 462 Center for the Mathematical Modelling of Infectious Diseases  
463 (2015). *Visualisation and projections of the Ebola outbreak*  
464 *in West Africa*. <http://ntnmcch.github.io/ebola/>. Archived at  
465 <http://www.webcitation.org/6oYEBYoeD> on Feb 24, 2017.
- 466 Centres for Disease Prevention and Control (Oct.  
467 2017). *Epidemic Prediction Initiative*. URL:



- 468 <https://predict.phiresearchlab.org/legacy/dengue/index.html>, Archived  
469 at <http://www.webcitation.org/6rsS5QDar> on 11 July, 2017.
- 470 Chowell, G. (2017). “Fitting dynamic models to epidemic outbreaks with  
471 quantified uncertainty: A primer for parameter uncertainty, identifiability,  
472 and forecasts”. *Infectious Disease Modelling* 3, 379–398.
- 473 Chowell, G. et al. (2017). “Perspectives on model forecasts of the 2014–2015  
474 Ebola epidemic in West Africa: lessons and the way forward”. *BMC Med*  
475 1, 42.
- 476 Chretien, J.-P., S. Riley, and D. B. George (Dec. 2015). “Mathematical  
477 modeling of the West Africa Ebola epidemic”. *eLife*, e09186.
- 478 Dawid, A. P. (1984). “Present Position and Potential Developments: Some  
479 Personal Views: Statistical Theory: The Prequential Approach”. *J R Stat*  
480 *Soc [Ser A]* 2, 278.
- 481 Dawson, P. M. et al. (2015). “Epidemic predictions in an imperfect world:  
482 modelling disease spread with partial data”. In: *Proc. R. Soc. B*. 1808.  
483 The Royal Society, 20150205.
- 484 Dong, F. et al. (Dec. 2015). “Evaluation of ebola spreading in west africa and  
485 decision of optimal medicine delivery strategies based on mathematical  
486 models”. *Infection, Genetics and Evolution*, 35–40.
- 487 Drake, J. M. et al. (2015). “Ebola cases and health system demand in  
488 Liberia”. *PLoS Biol* 1, e1002056.
- 489 Dureau, J., S. Ballesteros, and T. Bogich (2013). “SSM: Inference for time  
490 series analysis with State Space Models”.
- 491 Fisman, D., E. Khoo, and A. Tuite (2014). “Early epidemic dynamics of  
492 the west african 2014 ebola outbreak: estimates derived with a simple  
493 two-parameter model.” *PLOS Curr.: Outbreaks*.
- 494 Frieden, T. R. and I. K. Damon (2015). “Ebola in West Africa — CDC’s  
495 role in epidemic detection, control, and prevention”. *Emerging Infectious*  
496 *Diseases* 11, 1897.
- 497 Friederichs, P. and T. L. Thorarinsdottir (2012). “Forecast verification for  
498 extreme value distributions with an application to probabilistic peak  
499 wind prediction”. *Environmetrics* 7, 579–594.
- 500 Funk, S. (2016). *rbi.helpers: rbi helper functions*. R package version 0.2.
- 501 Funk, S. et al. (Dec. 2017a). “Real-time forecasting of infectious disease  
502 dynamics with a stochastic semi-mechanistic model”. *Epidemics*.
- 503 Funk, S. et al. (2017b). “The impact of control strategies and behavioural  
504 changes on the elimination of Ebola from Lofa County, Liberia”. *Phil*  
505 *Trans Roy Soc B* (1721), 20160302.
- 506 Gneiting, T., F. Balabdaoui, and A. E. Raftery (Apr. 2007). “Probabilistic  
507 forecasts, calibration and sharpness”. *Journal of the Royal Statistical*  
508 *Society: Series B (Statistical Methodology)* 2, 243–268.
- 509 Gneiting, T. and M. Katzfuss (Jan. 2014). “Probabilistic Forecasting”. *Annual*  
510 *Review of Statistics and Its Application* 1, 125–151.

- 511 Goldstein, E. et al. (2011). “Predicting the epidemic sizes of influenza  
512 A/H1N1, A/H3N2, and B: a statistical method”. *PLoS Med* 7, e1001051.
- 513 Heesterbeek, H. et al. (2015). “Modeling Infectious Disease Dynamics in the  
514 Complex Landscape of Global Health”. *Science* 6227, aaa4339–aaa4339.
- 515 Held, L., S. Meyer, and J. Bracher (June 2017). “Probabilistic forecasting  
516 in infectious disease epidemiology: the 13th Armitage lecture”. *Statistics  
517 in Medicine* 22, 3443–3460.
- 518 Hersbach, H. (2000). “Decomposition of the continuous ranked probability  
519 score for ensemble prediction systems”. *Weather and Forecasting* 5, 559–  
520 570.
- 521 Jacob, P. E. and S. Funk (2017). *RBi: R interface to LibBi*. R package  
522 version 0.7.0.
- 523 Johansson, M. A. et al. (2016). “Evaluating the performance of infectious  
524 disease forecasts: A comparison of climate-driven and seasonal dengue  
525 forecasts for Mexico”. *Scientific reports*.
- 526 Lewnard, J. A. et al. (Dec. 2014). “Dynamics and control of Ebola virus  
527 transmission in Montserrado, Liberia: a mathematical modelling analy-  
528 sis.” *Lancet Infect Dis* 12, 1189–1195.
- 529 Liu, F. et al. (2015). “Short-term forecasting of the prevalence of trachoma:  
530 expert opinion, statistical regression, versus transmission models”. *PLoS  
531 neglected tropical diseases* 8, e0004000.
- 532 Lowe, R. et al. (2014). “Dengue outlook for the World Cup in Brazil: an early  
533 warning model framework driven by real-time seasonal climate forecasts”.  
534 *The Lancet infectious diseases* 7, 619–626.
- 535 Meltzer, M. I. et al. (Sept. 2014). “Estimating the future number of cases  
536 in the Ebola epidemic–Liberia and Sierra Leone, 2014–2015.” *MMWR  
537 Surveill Summ*, 1–14.
- 538 Merler, S. et al. (Feb. 2015). “Spatiotemporal spread of the 2014 outbreak of  
539 Ebola virus disease in Liberia and the effectiveness of non-pharmaceutical  
540 interventions: a computational modelling analysis”. *Lancet Infect Dis* 2,  
541 204–211.
- 542 Moran, K. R. et al. (2016). “Epidemic forecasting is messier than weather  
543 forecasting: The role of human behavior and internet data streams in  
544 epidemic forecast”. *The Journal of Infectious Diseases* suppl\_4, S404–  
545 S408.
- 546 Murray, L. (2015). “Bayesian State-Space Modelling on High-Performance  
547 Hardware Using LibBi”. *Journal of Statistical Software, Articles* 10, 1–  
548 36. ISSN: 1548-7660.
- 549 National Oceanic and Atmospheric Administration (Oct. 2017). *Dengue  
550 Forecasting*. URL: <http://dengueforecasting.noaa.gov/>, Archived at  
551 <http://www.webcitation.org/6oWfUBKnC> on Feb 24, 2017.
- 552 Nishiura, H. and G. Chowell (2014). “Early transmission dynamics of Ebola  
553 virus disease (EVD), West Africa, March to August 2014”. *Euro Surveill*  
554 (36), 20894.

- 555 Nsoesie, E., M. Mararthe, and J. Brownstein (2013). “Forecasting peaks of  
556 seasonal influenza epidemics”. *PLoS currents*.
- 557 Pei, S. and J. Shaman (Oct. 2017). “Counteracting structural errors in en-  
558 semble forecast of influenza outbreaks”. *Nature Communications* 1.
- 559 R Core Team (2017). *R: A Language and Environment for Statistical Com-  
560 puting*. R Foundation for Statistical Computing. Vienna, Austria.
- 561 Rivers, C. M. et al. (2014). “Modeling the impact of interventions on an  
562 epidemic of Ebola in Sierra Leone and Liberia”. *PLOS Curr.: Outbreaks*.
- 563 Scott, S. L. (2017). *bsts: Bayesian Structural Time Series*. R package version  
564 0.7.1.
- 565 Siettos, C. et al. (2015). “Modeling the 2014 ebola virus epidemic-agent-  
566 based simulations, temporal analysis and future predictions for liberia  
567 and sierra leone”. *PLOS Curr.: Outbreaks*.
- 568 Towers, S., O. Patterson-Lomba, and C. Castillo-Chavez (2014). “Temporal  
569 variations in the effective reproduction number of the 2014 West Africa  
570 Ebola outbreak”. *PLOS Curr.: Outbreaks*.
- 571 Viboud, C. et al. (Aug. 2017). “The RAPIDD ebola forecasting challenge:  
572 Synthesis and lessons learnt”. *Epidemics*.
- 573 White, R. A. et al. (2015). “Projected treatment capacity needs in Sierra  
574 Leone”. *PLOS Curr.: Outbreaks*.
- 575 WHO Ebola Response Team (Sept. 2014). “Ebola Virus Disease in West  
576 Africa - The First 9 Months of the Epidemic and Forward Projections.”  
577 *N Engl J Med*.
- 578 Wilks, D. S. (2018). “Enforcing calibration in ensemble postprocessing”.  
579 *Quarterly Journal of the Royal Meteorological Society* 710, 76–84.
- 580 World Health Organization (2017). *Efficacy trials of ZIKV Vaccines: end-  
581 points, trial design, site selection. WHO Workshop Meeting Report*.
- 582 Yang, W. et al. (2015). “Forecasting influenza epidemics in Hong Kong”.  
583 *PLoS computational biology* 7, e1004383.

## 584 **Acknowledgements**

### 585 **Funding**

586 This work was funded by the Research for Health in Humanitarian Crises  
587 (R2HC) Programme, managed by Research for Humanitarian Assistance  
588 (Grant 13165). SF, AJK and AC were supported by fellowships from the  
589 UK Medical Research Council (SF: MR/K021680/1, AC: MR/J01432X/1,  
590 AJK: MR/K021524/1). RL was supported by a Royal Society Dorothy  
591 Hodgkins fellowship. RME was supported by the Innovative Medicines  
592 Initiative 2 (IMI2) Joint Undertaking under grant agreement EBOVAC1  
593 (Grant 115854). The IMI2 is supported by the European Union Horizon

594 2020 Research and Innovation Programme and the European Federation of  
595 Pharmaceutical Industries and Associations.

#### 596 **Author contribution**

597 SF, AC and WJE conceived the study; SF and AC analysed the data; SF  
598 wrote a first draft of the paper; AC led the generation of real-time forecasts  
599 during the Ebola epidemic; all authors contributed to the text of the final  
600 version.

#### 601 **Competing interests**

602 There are no competing interests.

#### 603 **Data and materials availability**

604 The authors declare that all data supporting the findings of this study will  
605 be available within the paper and its supplementary information files.