Superensemble forecasts of dengue outbreaks

Teresa K. Yamana, Sasikiran Kandula and Jeffrey Shaman
Journal of the Royal Society Interface

Supplementary Material

Supplemental Methods

Historical likelihood forecasts

Peak timing

The initial historical likelihood forecast for peak week m is constructed by fitting a normal distribution to historically observed peak weeks and is denoted as $p_{init}(m)$. As the season progresses, the probability assigned to weeks that are observed not to be the peak of the outbreak are set to zero and that residual probability is distributed among the remaining candidate weeks. Specifically, given a set of $1{:}t$ observations for an outbreak in progress reported at week t, the maximum weekly number of cases observed thus far is x' and is observed at week m'. In the event that there is a tie for peak week, with exactly x' cases reported on multiple weeks, then m' is set to the first of these weeks.

Using our initial distribution of peak timing, we compute the probability that the peak has already occurred, $p_{init}(m \leq t)$. We then update the probability distribution as follows:

$$p(\mathrm{m}|\mathrm{m}') = \begin{cases} p_{init}(m \leq t), m = m' \\ 0, \;\; m \leq t \text{ and } m \neq m' \\ \quad p_{init}(m), m > t \end{cases}$$

In other words, the probability that the peak week as thus far observed is in fact the overall peak is $p_{init}(m \leq t)$. The probabilities for the peak occurring in weeks t+1 through 52 remain unchanged.

The historical likelihood forecast for peak timing is the expected value of p(m|m').

Peak incidence

We describe the initial probability distribution function of peak incidence, $p_{init}(x)$, as a gamma distribution fit to historically observed peak incidence. As cases are observed, we adjust the PDF in response to eliminated values and the likelihood that the peak has already passed. The PDF of peak incidence conditioned on the peak having passed is:

$$p(x|\, x', m \leq t) = \begin{cases} 1, \;\; x = x' \\ 0, \;\; x \neq x' \end{cases}$$

where t is the week the forecast is being made, x is peak incidence, x' is peak incidence observed through week t, and m is the peak week.

The PDF of peak incidence conditioned on the peak being in the future is the initial gamma distribution with values of x that have already been exceeded given zero probability, and the remaining values scaled such that the sum of the conditional PDF is equal to 1:

$$p(x|\, x', m > t) = \begin{cases} 0\,, & x \leq x' \\ \dfrac{p_{init}(x)}{\sum_{x=x'+1}^{\infty} p_{init}(x)}, & x > x' \end{cases}$$

The conditional probability distributions are combined as follows to produce an updated PDF of peak incidence:

$$p(x|x') = p(x|x', m \leq t) * p_{init}(m \leq t) + p(x|x', m > t) * [1 - p_{init}(m \leq t)]$$

The forecast for peak incidence is the expected value of x computed from this updated probability distribution function.

Total incidence

The historical likelihood forecasts for total incidence are obtained by adding cases observed to date to the expected future cases based on observations from years 1 through N-1. For every week t, we calculate future cases, $s_k(t)$, which is the total number of cases observed between weeks t+1 and the end of the season, for year k.

$$s_k(t) = \sum_{i=t+1}^{52} f_k(i)$$

where f_k is observed weekly dengue incidence for year k.

We then obtain a probability distribution for expected future cases by fitting a gamma distribution to $s_k(t)$. We add the total cases observed in the current season thus far (weeks 1 through t) to obtain a PDF of total incidence. The expected value of this distribution is the historical likelihood forecast of total cases.

Supplemental results

Forecast mean absolute error grouped by week relative to actual peak

In evaluating the performance of each forecast, it can be useful to compare results grouped by lead time with respect to the observed outbreak peak, defined as the number of weeks between when the forecast is made and the true peak (see Fig. S4).

Peak timing

F1 had larger errors than the F2 and F3 forecasts of peak timing made during the 8 weeks leading up to the observed peak. However, this forecasting system was faster than the other two in detecting the true peak, and the mean absolute error (MAE) dropped to zero three weeks after the true peak. F3 was the least accurate forecast after the outbreak peak. Before the outbreak peak, the superensemble forecasts generally had MAE similar to or less than that of the individual forecasts. After the outbreak peak, SE(F1,F2) had MAE roughly equal to F2, and SE(F1,F2,F3) had greater MAE than both F1 and F2.

Peak incidence

F2 had the lowest MAE for predictions of peak incidence made four or more weeks before the observed peak. F1 did poorly early in the season, but had the smallest MAE when the outbreak peak was less than four weeks in the future. F3 forecasts were much less accurate than the other two systems for forecasts made after the outbreak peak. As with forecasts of peak timing, superensemble forecasts of peak incidence had similar or smaller MAE than the best individual forecast before the outbreak peak. SE(F1,F2,F3) had higher MAE than any other forecast except F3.

Total incidence

The relative performance of the three individual forecasts fluctuated in the four months preceding the outbreak peak. After the peak, F1 had the smallest MAE. Superensemble forecasts were generally comparable to the best forecast at any given time before the outbreak peak, and better than F2 and F3 two or more weeks after the peak.

Forecast mean absolute error grouped by F2 predicted peak

We can also group MAE by lead time relative to the *predicted* outbreak peak. While the predicted peak is not necessarily the same as the observed peak, it is a useful measure as this information is available when a real-time forecast is produced, while lead time relative to the actual outbreak peak would not yet be known. Since F2 produced the most reliable forecasts of peak timing, we show MAE grouped by lead time with respect to the F2 forecast peak in Fig. S5. This grouping showed similar patterns as were observed in Fig. S4.

Sensitivity Analysis

We tested the sensitivity of our results to model assumptions and initial conditions. We varied each of the fixed parameters in the F1 forecast, as well as the length of the training window and the distribution of weekly dengue incidence in the F2 forecast. We repeated the entire analysis described in the main text a total of 4374 times, once for each combination of parameters. The varied conditions are as follows:

F1 options:

a) Percentage of weekly infections that are observed: 10%, 20%, 30%
b) Observational error variance: 40, 65, 100 cases^2
c) Initial proportion of susceptible people drawn from uniform distribution ranging from 0 to 0.4, 0 to 0.6, or 0 to 0.8.
d) Initial number of infected persons drawn from exponential distribution with mean 20, 40 or 100 people.
e) Initial values of R0 drawn from the following distributions: U[0.5,3.5],U[1,4] and U[1.5,4.5]

F2 options:

a) Training window of 4, 8 or 12 weeks
b) Distribution of weekly dengue incidence: normal or gamma

We did not alter the probability distribution functions used for the forecast F3, as this method is based on a simple resampling of historical outcomes. The distributions described in the paper are the ones that best fit the historical data.

For each combination of parameters and initial conditions, we computed the mean absolute error of each forecast for the three target metrics (peak timing, peak incidence and total incidence) (Figure S3). We then ranked the forecasts (both individual and superensemble) generated within each set of parameters from lowest (rank=1) to highest (rank=7) MAE (Table S2). For the 4374 sets of forecasts, the relative performance of each forecast method was consistent with our original results. Overall, superensemble forecasts provided a clear and consistent advantage over their contributing individual forecasts.

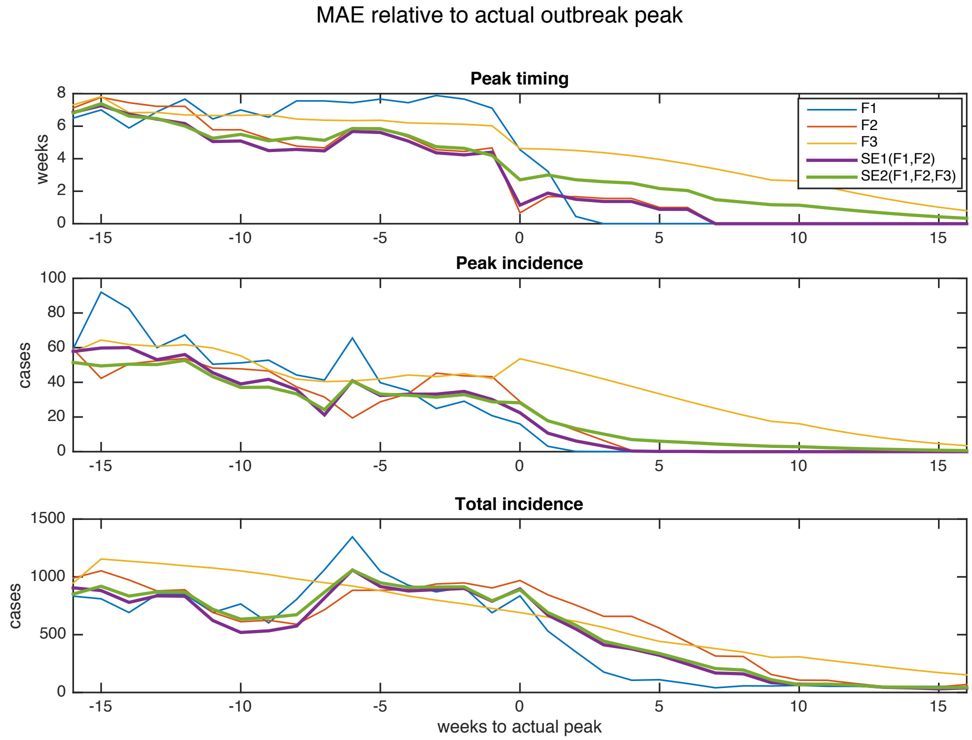

Figure S1. Mean absolute error of forecasts relative to observed outbreak peak. Negative values on the x-axis indicate forecasts made prior to the observed peak, and positive values are forecasts made after the peak has passed. Forecasts made more than 16 weeks before or after the outbreak peak are omitted.

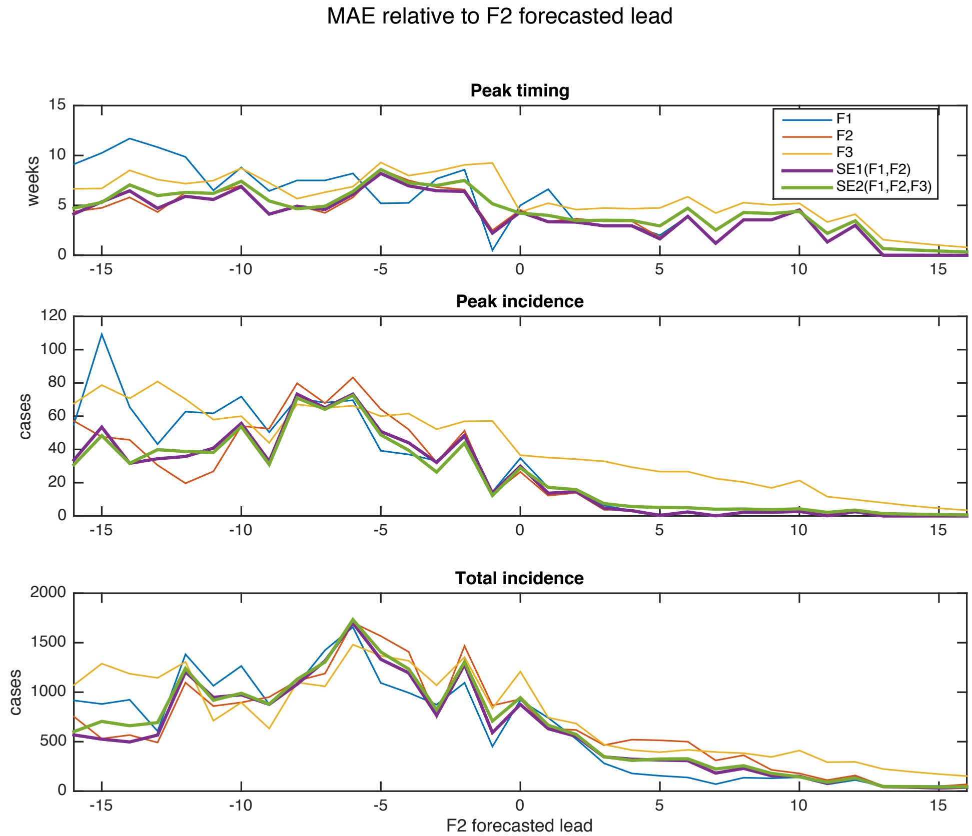

Figure S2. Mean absolute error of forecasts relative to the peak predicted by F2 forecast. Negative values on the x-axis indicate forecasts made prior to the predicted peak, and positive values are forecasts made after the predicted peak has passed. Forecasts made more than 16 weeks before or after the predicted peak are omitted.

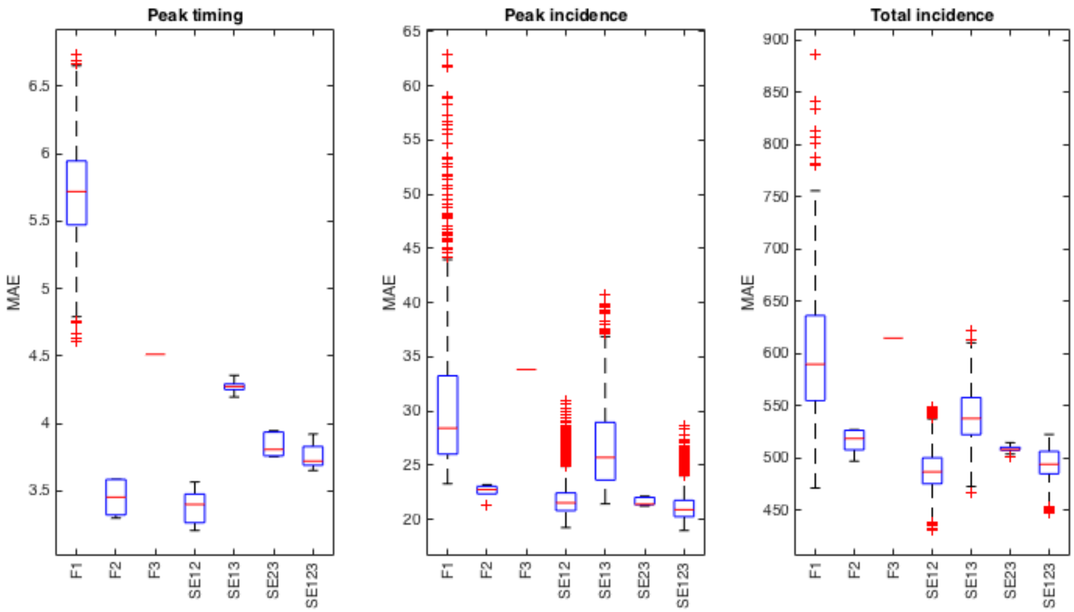

Figure S3. Results of sensitivity analysis. Box plots indicate forecast mean absolute error over the 9-year testing period under different combinations of parameters and initial conditions.

Table S1. P-values obtained using paired t-test statistic testing for differences in mean absolute error between forecasts. Results that are significant at the 90% confidence level are shown in bold.

<u>P-values t-test peak timing</u>

	F1	F2	F3	SE(F1,F2)	SE(F1,F3)	SE(F2,F3)	SE(F1,F2,F3)
F1	-	**0.000**	0.196	**0.000**	**0.017**	**0.000**	**0.000**
F2	**0.000**	-	**0.000**	**0.023**	**0.000**	**0.000**	**0.003**
F3	0.196	**0.000**	-	**0.000**	**0.000**	**0.000**	**0.000**
SE(F1,F2)	**0.000**	**0.023**	**0.000**	-	**0.000**	**0.000**	**0.000**
SE(F1,F3)	**0.017**	**0.000**	**0.000**	**0.000**	-	**0.000**	**0.000**
SE(F2,F3)	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	-	**0.004**
SE(F1,F2,F3)	**0.000**	**0.003**	**0.000**	**0.000**	**0.000**	**0.004**	-

<u>P-values t-test peak incidence</u>

	F1	F2	F3	SE(F1,F2)	SE(F1,F3)	SE(F2,F3)	SE(F1,F2,F3)
F1	-	0.192	**0.000**	**0.000**	**0.005**	**0.021**	**0.000**
F2	0.192	-	**0.000**	**0.047**	0.860	**0.032**	**0.008**
F3	**0.000**	**0.000**	-	**0.000**	**0.000**	**0.000**	**0.000**
SE(F1,F2)	**0.000**	**0.047**	**0.000**	-	**0.031**	0.602	**0.052**
SE(F1,F3)	**0.005**	0.860	**0.000**	**0.031**	-	0.157	**0.001**
SE(F2,F3)	**0.021**	**0.032**	**0.000**	0.602	0.157	-	**0.059**
SE(F1,F2,F3)	**0.000**	**0.008**	**0.000**	**0.052**	**0.001**	**0.059**	-

<u>P-values paired t-test total incidence</u>

	F1	F2	F3	SE(F1,F2)	SE(F1,F3)	SE(F2,F3)	SE(F1,F2,F3)
F1	-	0.934	**0.002**	**0.036**	0.366	0.625	**0.097**
F2	0.934	-	**0.016**	**0.001**	0.643	0.255	**0.037**
F3	**0.002**	**0.016**	-	**0.000**	**0.000**	**0.000**	**0.000**
SE(F1,F2)	**0.036**	**0.001**	**0.000**	-	0.143	**0.019**	0.110
SE(F1,F3)	0.366	0.643	**0.000**	0.143	-	0.925	0.222
SE(F2,F3)	0.625	0.255	**0.000**	**0.019**	0.925	-	**0.068**
SE(F1,F2,F3)	**0.097**	**0.037**	**0.000**	0.110	0.222	**0.068**	-

Table S2. Mean rank of forecast MAE over the 4374 combinations of model parameters and initial conditions tested during sensitivity analysis. Rank corresponds to lowest (1) to highest (7) MAE. For comparison, we show the ranking of original forecasts as described in the main text of the paper.

	Peak timing		Peak incidence		Total incidence	
	Mean rank	Original rank	Mean rank	Original rank	Mean rank	Original rank
Individual forecasts						
F1	7.0	7	6.2	6	6.3	5
F2	2.0	2	3.6	4 (tie)	3.8	6
F3	6.0	6	6.7	7	6.6	7
Superensemble forecasts						
SE12	1.0	1	2.7	2 (tie)	1.5	1
SE13	5.0	5	5.0	4 (tie)	4.7	4
SE23	4.0	4	2.4	2 (tie)	3.2	3
SE123	3.0	3	1.5	1	2.0	2

Table S3. Comparison of MAE between weighted average and simple average forecasts. P-values indicate paired t-test statistic of difference in MAE between the two forecasts; results that are significant at the 90% level are shown in bold.

	Peak timing		Peak incidence		Total incidence	
	MAE	p-value	MAE	p-value	MAE	p-value
SE(F1,F2)	3.3	**0.098**	21	0.533	473	0.675
Avg(F1,F2)	3.5		21		472	
SE(F1,F3)	4.3	**0.008**	23	**0.006**	507	0.116
Avg(F1,F3)	4.0		25		513	
SE(F2,F3)	3.8	**0.000**	21	**0.017**	505	0.482
Avg(F2,F3)	3.9		23		510	
SE(F1,F2,F3)	3.7	**0.009**	20	**0.067**	486	0.811
Avg(F1,F2,F3)	3.5		21		488	