An open challenge to advance probabilistic forecasting for dengue epidemics

Michael A. Johansson[a,b,1], Karyn M. Apfeldorf[c], Scott Dobson[c], Jason Devita[c], Anna L. Buczak[d], Benjamin Baugher[d], Linda J. Moniz[d], Thomas Bagley[d], Steven M. Babin[d], Erhan Guven[d], Teresa K. Yamana[e], Jeffrey Shaman[e], Terry Moschou[f], Nick Lothian[f], Aaron Lane[f], Grant Osborne[f], Gao Jiang[g], Logan C. Brooks[h], David C. Farrow[h], Sangwon Hyun[i], Ryan J. Tibshirani[h,i], Roni Rosenfeld[h], Justin Lessler[j], Nicholas G. Reich[k], Derek A. T. Cummings[l,m], Stephen A. Lauer[k], Sean M. Moore[n,o], Hannah E. Clapham[p], Rachel Lowe[q,r], Trevor C. Bailey[s], Markel García-Díez[t], Marilia Sá Carvalho[u], Xavier Rodó[r], Tridip Sardar[v], Richard Paul[w,x], Evan L. Ray[y], Krzysztof Sakrejda[k], Alexandria C. Brown[k], Xi Meng[k], Osonde Osoba[z], Raffaele Vardavas[z], David Manheim[aa], Melinda Moore[z], Dhananjai M. Rao[bb], Travis C. Porco[cc], Sarah Ackley[cc], Fengchen Liu[cc], Lee Worden[cc], Matteo Convertino[dd], Yang Liu[ee], Abraham Reddy[ee], Eloy Ortiz[ff], Jorge Rivero[ff], Humberto Brito[ff,gg], Alicia Juarrero[ff,hh], Leah R. Johnson[ii], Robert B. Gramacy[jj], Jeremy M. Cohen[ji], Erin A. Mordecai[kk], Courtney C. Murdock[ll,mm], Jason R. Rohr[n,o], Sadie J. Ryan[m,nn,oo], Anna M. Stewart-Ibarra[pp], Daniel P. Weikel[qq], Antarpreet Jutla[rr], Rakibul Khan[rr], Marissa Poultney[rr], Rita R. Colwell[ss], Brenda Rivera-García[tt], Christopher M. Barker[uu], Jesse E. Bell[vv], Matthew Biggerstaff[ww], David Swerdlow[ww], Luis Mier-y-Teran-Romero[a,j], Brett M. Forshey[xx], Juli Trtanj[yy], Jason Asher[zz], Matt Clay[zz], Harold S. Margolis[a], Andrew M. Hebbeler[aaa,bbb], Dylan George[bbb,ccc], and Jean-Paul Chretien[bbb,ddd]

A wide range of research has promised new tools for forecasting infectious disease dynamics, but little of that research is currently being applied in practice, because tools do not address key public health needs, do not produce probabilistic forecasts, have not been evaluated on external data, or do not provide sufficient forecast skill to be useful. We developed an open collaborative forecasting challenge to assess probabilistic forecasts for seasonal epidemics of dengue, a major global public health problem. Sixteen teams used a variety of methods and data to generate forecasts for 3 epidemiological targets (peak incidence, the week of the peak, and total incidence) over 8 dengue seasons in Iquitos, Peru and San Juan, Puerto Rico. Forecast skill was highly variable across teams and targets. While numerous forecasts showed high skill for midseason situational awareness, early season skill was low, and skill was generally lowest for high incidence seasons, those for which forecasts would be most valuable. A comparison of modeling approaches revealed that average forecast skill was lower for models including biologically meaningful data and mechanisms and that both multimodel and multiteam ensemble forecasts consistently outperformed individual model forecasts. Leveraging these insights, data, and the forecasting framework will be critical to improve forecast skill and the application of forecasts in real time for epidemic preparedness and response. Moreover, key components of this project—integration with public health needs, a common forecasting framework, shared and standardized data, and open participation—can help advance infectious disease forecasting beyond dengue.

forecast | dengue | epidemic | Peru | Puerto Rico

Infectious diseases pose a continuing and dynamic threat globally. The mosquito-transmitted dengue viruses, for example, are endemic throughout the tropical regions of the world and infect millions of people each year (1). In endemic areas, dengue incidence has a clear seasonal pattern but also, exhibits strong interannual variation, with major epidemics occurring every few years (2, 3). In San Juan, Puerto Rico, hundreds of confirmed cases may be reported over an entire interepidemic season, while hundreds of cases can be reported every week during the peak of epidemics (Fig. 1). Timely and effective large-scale interventions are needed to reduce the serious impacts of dengue epidemics on health, healthcare systems, and economies (4, 5). Unfortunately, these epidemics have proven difficult to predict, hindering efforts to prevent and control their impact.

Research on the determinants of dengue epidemics has included both statistical models incorporating historical incidence and climatological determinants (6) and dynamical,

Significance

Forecasts routinely provide critical information for dangerous weather events but not yet for epidemics. Researchers develop computational models that can be used for infectious disease forecasting, but forecasts have not been broadly compared or tested. We collaboratively compared forecasts from 16 teams for 8 y of dengue epidemics in Peru and Puerto Rico. The comparison highlighted components that forecasts did well (e.g., situational awareness late in the season) and those that need more work (e.g., early season forecasts). It also identified key facets to improve forecasts, including using multiple model ensemble approaches to improve overall forecast skill. Future infectious disease forecasting work can build on these findings and this framework to improve the skill and utility of forecasts.

Fig. 1. Dengue and climate data for Iquitos, Peru and San Juan, Puerto Rico. The black and colored lines for dengue cases indicate the total and virus-specific weekly number of laboratory-confirmed cases. The yellow and red points indicate the peaks in the training and testing datasets, respectively. The climate data show the weekly rainfall (blue) and mean temperature (red) for Iquitos and San Juan, respectively, from the National Centers for Environmental Prediction Climate Forecast System Reanalysis.

mathematical models aimed at identifying both intrinsic and extrinsic drivers (7, 8). This body of research led to important insights, such as the putative influence of various climatological components (9), antibody-dependent enhancement (10, 11), serotype-specific cross-protection (12, 13), and spatial heterogeneity (14) on transmission dynamics.

Despite this substantial body of research, there are currently no operational dengue forecasts with documented prospective forecast skill, and challenges exist for both forecast development and assessment. First, the objectives of published forecasts and outcome metrics vary and are often not tied to specific public health needs. Second, there have been few accessible dengue datasets for forecasting research. Third, differences in data and metrics significantly complicate the comparison of forecasts from different research groups. Fourth, existing evaluations generally assess only point prediction accuracy, ignoring information on forecast confidence. Fifth, evaluations rarely incorporate out-of-sample testing (testing on either reserved or prospective data that were not used to develop and fit the models), the most important test for a forecasting model.

The need to systematically evaluate forecasting tools is widely recognized (15) and motivated multiple US government agencies within the Pandemic Prediction and Forecasting Science and Technology Working Group, coordinated by the White House Office of Science and Technology Policy, to launch an open forecasting challenge in 2015, the Dengue Forecasting Project. First, we worked with epidemiologists from dengue-endemic regions to identify 3 important epidemic forecasting targets: 1) the intensity of the epidemic peak (peak incidence), 2) the timing of that peak (peak week), and 3) the total number of cases expected over the duration of the season (season incidence). Reliable forecasts of these outcomes could improve the allocation of resources for primary prevention (e.g., risk communication, vector control) or secondary prevention (e.g., planning medical staffing, preparing triage units) (16). Additionally, because out-of-sample prediction is an important test of mechanistic causality, forecasts could also provide insight on key drivers of dengue epidemics and therefore, the expected impacts of interventions. Second, we identified 2 dengue-endemic cities, Iquitos, Peru (17, 18) and San

Juan, Puerto Rico (19), with serotype-specific incidence data and local climate data that could be released publicly for enough seasons (13 and 23, respectively) to allow training of models and forecasting across multiple seasons (Fig. 1). Third, we established an a priori forecasting framework, including a specific protocol for submitting and evaluating out-of-sample probabilistic forecasts made at 4-wk intervals across 4 training and 4 testing seasons for each of the 3 targets in both locations.

Results

Sixteen teams submitted binned probabilistic forecasts generated using a variety of approaches, including statistical and mechanistic models and multimodel ensembles (*SI Appendix*, Table S1). All teams used the provided dengue data, 10 (63%) used matched climate data, 2 used serotype data, and 1 used additional data on global climate (e.g., Southern Oscillation Index). Three additional models were developed for comparison: a null model (equal probability assigned to each possible outcome), a baseline statistical time series model (a seasonal autoregressive integrated moving average [SARIMA] model), and a simple ensemble (an average of the probabilities of the 16 team and baseline forecasts).

After finalizing models and submitting forecasts for 4 training seasons (2005/2006 to 2008/2009), teams received additional data and had a maximum of 2 wk to submit forecasts for the testing seasons (2009/2010 to 2012/2013). Forecasts varied widely (Fig. 2 and *SI Appendix*, Figs. S1 and S2). For example, forecasts with data up to week 12 and week 24 predicted that the peaks in the 2012/2013 season might have been among the lowest or the highest on record. Confidence also varied: some forecasts were certain of an outcome being in a particular forecast bin, while others had broad 95% prediction intervals spanning the entire range of historical values, and some assigned 0 probability to the observed outcome.

We assessed forecast skill using the logarithmic score, a proper score incorporating probabilistic accuracy and precision. High logarithmic scores indicate consistent assignment of high probability to the eventually observed outcome. Forecast skill increased as seasons progressed for most models (Fig. 3). Some submitted forecasts outperformed both the null and baseline models for early time points, with numerous models showing increased skill around the time of the observed peak (median peak weeks: 22.5 for San Juan and 28 for Iquitos). The peak incidence target for Iquitos in 2011/2012 was not scored, as no distinct peak was identifiable. Forecast calibration (e.g., assigning 70% probability to events that occurred 70% of the time) varied across teams (*SI Appendix*, Fig. S3) and was strongly associated with forecast skill (*SI Appendix*, Fig. S4).

The highest skill early season forecasts (weeks 0 to 24) for each target–location pair were submitted by Team N (University of California, San Francisco, peak week, Iquitos), Team E (VectorBiTE, peak week, San Juan) (20), Team B (Breaking Bad Bone Fever, peak incidence and total incidence, Iquitos) (21), Team G (Areté, peak incidence, San Juan), and Team J (Delphi, total incidence, San Juan) (Figs. 3 and 4 and *SI Appendix*, Table S1). Many teams outscored both the null model and for each target except peak week, the baseline model. The ensemble forecast outperformed most individual models and was the only forecast to outperform the null model for every target. Training season forecasts showed similar patterns of low early season skill and overconfidence by some models, and numerous models outperformed the baseline and null models (*SI Appendix*, Fig. S5).The top teams differed for all targets except peak week in Iquitos (Fig. 4 and *SI Appendix*, Table S2), but the ensemble forecasts outperformed the majority of individual forecasts and the null forecast for all targets for all 8 seasons.

To assess extrinsic factors that may impact forecast skill, we fitted a series of regression models to target-, location-, and season-specific variables (*SI Appendix*). Scores were higher for forecasts made later in the season (0.043 per week, 95% confidence interval
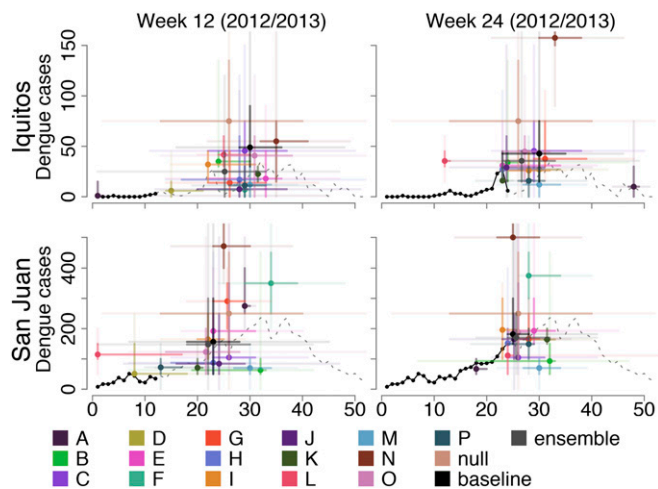
Fig. 2. Weeks 12 and 24 forecasts for the 2012/2013 dengue season in Iquitos and San Juan. The solid black lines indicate the most recent data that were available to teams to inform these forecasts, and the dashed lines indicate the data that became available later in the season. The colored points represent point estimates for each team, while the bars represent 50 and 95% prediction intervals (dark and light, respectively). Forecasts for additional time points and seasons as well as for seasonal incidence are shown in *SI Appendix*, Figs. S1 and S2, respectively.

[95% CI]: 0.039 to 0.046), seasons with lower peak incidence (0.43 per location-specific SD, 95% CI: 0.37 to 0.49), seasons with earlier peaks (0.048 per week prior to long-term location-specific mean, 95% CI: 0.040 to 0.057), San Juan (0.65, 95% CI: 0.54 to 0.76), and targets with fewer bins (peak and seasonal incidence, 0.0257 per bin, 95% CI: 0.0221 to 0.0293) (*SI Appendix*, Table S3).

Comparing high-level forecasting approaches across all targets and all 8 seasons while controlling for the differences by forecast week, season characteristics, location, and the numbers of bins (described above), we found that logarithmic scores were higher for teams using ensemble approaches (mean difference: 1.02, 95% CI: 0.91 to 1.13) (*SI Appendix*, Table S3). Forecasts from models incorporating mechanistic approaches (e.g., compartmental models or ensemble models with at least 1 mechanistic submodel) had lower logarithmic scores (−0.65, 95% CI: −0.80 to −0.49) than purely statistical approaches. Additionally, models using climate data had lower logarithmic scores (−0.14, 95% CI: −0.19 to −0.09). Relatedly, we found that forecasts using ensemble approaches tended to be better calibrated (−0.0010, 95% CI: −0.0034 to 0.0007) and that those using mechanistic approaches or climate data were less so (*SI Appendix*, Table S4). We did not compare models using serotype data or incorporating vector population dynamics, as only 2 models included serotype data (using them in different ways), and all but 1 mechanistic model included modeled vector populations (actual vector data were not available).

Discussion

Research aimed at forecasting epidemics and their impact offers tantalizing opportunities to prevent or control infectious diseases. Although many epidemic forecasting tools promise high accuracy, they have largely been fit to specific, nonpublic datasets and assessed only on historical data rather than future, unobserved outcomes. Here, we executed a multimodel assessment of out-of-sample probabilistic forecasts for key seasonal characteristics of dengue epidemics. Comparing these forecasts provides insight on current capabilities to forecast dengue, our understanding of the drivers of dengue epidemics, challenges to forecast skill, and avenues for improvement.

Good forecasts should identify possible outcomes relevant to decision makers and reliably assign probabilities to those outcomes (22). Proper scores (23, 24) of probabilistic forecasts, such as the logarithmic score used here, have distinct advantages over more common point prediction error metrics. Error only measures 1 dimension of forecast skill, the distance between the estimated and observed outcomes, and does not consider confidence, an essential characteristic for stochastic outcomes. Logarithmic scores for the submitted forecasts revealed low early season forecast skill, with many forecasts performing worse than a null forecast that assigned an equal probability to each possible outcome. Even in endemic areas with strong seasonal transmission patterns, epidemics are difficult to predict at time horizons of several months or more.

Nonetheless, several teams consistently outperformed the null model for each target–location pair, indicating that, even in early weeks, models provided some reliable information about what was likely to happen. Not surprisingly, forecasts improved substantially as seasons progressed and data accumulated. As more data are reported, the likely outcomes are reduced, and forecasting is easier (e.g., if 1,100 cases have been reported by week 40, it is impossible that the season total will be less than 1,000 and extremely unlikely that it would exceed 10,000). Despite this, some models had decreased or steady late season skill, possibly indicating that they did not fully account for data updates. Week-to-week incidence varies substantially, making peaks hard to identify in real time, and therefore, models with high midseason to late season skill may be very useful for situational awareness.

Overall scores varied by target, location, and season. Differences in target-specific scores were not associated with target-specific entropy, implying that target-specific differences were more likely due to study design than intrinsic differences in predictability. Specifically, the peak week target had more bins (52 vs. 11), and therefore, probabilities were distributed across more bins, leading to lower probabilities for the outcomes and lower scores. Higher scores for San Juan compared with Iquitos may reflect differences in dynamics, the availability of more historical data, or the location-specific bin selection. This difference was not related
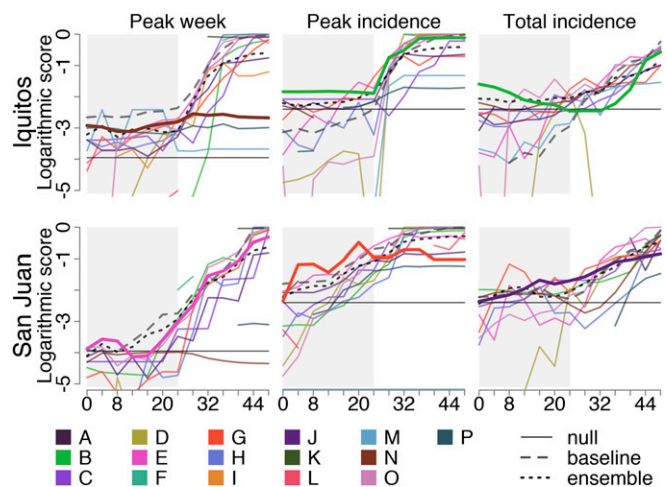
Fig. 3. Forecast skill by team, forecast week, and target in the testing seasons (2009/2010 to 2012/2013). Solid colored lines represent the scores of individual teams averaged across all testing seasons for the respective forecast week, target, and location. For each target, the top forecast for the first 24 wk (shaded) is indicated in bold (highest average early season score). The solid black lines indicate the null model (equal probability assigned to all possible outcomes), the dashed gray lines indicate the baseline model, and the dotted black lines indicate the ensemble model. Forecasts with logarithmic scores of less than −5 are not shown. Breaks in lines indicate a score of negative infinity in at least 1 of the testing seasons.
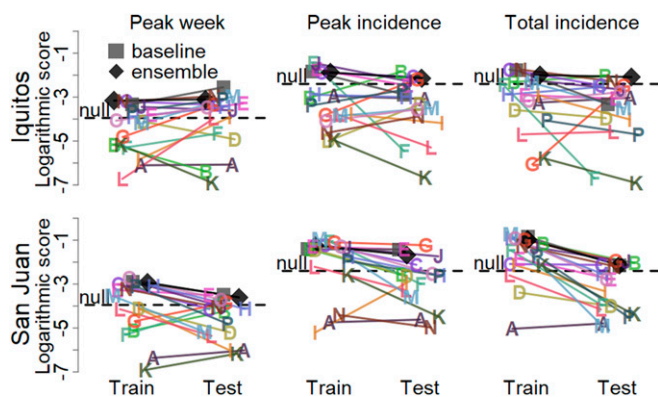
Fig. 4. Overall forecast scores for weeks 0 to 24 in the training (2005/2006 to 2008/2009) and testing (2009/2010 to 2012/2013) seasons. Each point is the average target- and location-specific log score for a model in the training (left side; light shading) and testing (right side; dark shading) seasons. The horizontal dispersion within training and testing scores is random to improve visualization. The null forecast for each target is represented by a horizontal line. Numerous forecasts assigned 0 probability to at least 1 observed outcome. Those individual forecast probabilities were changed to 0.001 before calculating the logarithmic scores.

to location-specific variability, as target-specific entropy was similar or higher for San Juan (peak week: 1.28 for Iquitos and 2.08 for San Juan; peak incidence: 1.75 and 1.73, respectively; and season incidence: 1.28 and 1.39, respectively). However, the long-term dynamics in the 2 locations were markedly different, with more recent introduction and serotype replacement in Iquitos vs. decades of hyperendemic transmission in San Juan. The effect of these differences and simply the availability of more historical data for San Juan are not distinguishable in this study. Finally, we found that forecast skill was lower for seasons with later and higher peaks. The association with later peaks may indicate a particular challenge of late seasons or a more general association with atypical peak timing rather than a late season per se, or it may simply reflect a higher proportion of forecasts being made before the peak, when there is more uncertainty. Influenza forecasts also tend to perform worse in late seasons (25). The association of low forecast skill with high incidence is also a key challenge; seasonal cycling is generally predictable, but high-incidence epidemics, the biggest challenge for public health, are the hardest to predict.

A wide variety of modeling approaches was used, including different criteria for data selection (e.g., climate data, lags), model frameworks (e.g., mechanistic, statistical), parameter assignment methods (e.g., fitting, specifying), and forecast generation procedures (e.g., model selection, combination). Because there are so many potential options for these components, the 17 models that we evaluated (teams and baseline) only represent a small subspace of all possible models. We, therefore, restricted our analysis to 3 high-level characteristics represented by multiple forecasts (climate data, a mechanistic model, or an ensemble approach), recognizing that even these findings may not be generalizable. Suitable climatic conditions are biologically necessary for dengue virus transmission, yet models including climate data had less skill than models that did not. One challenge is that climate forecasts may be more useful than historical data for dengue forecasts, but climate forecasts have their own uncertainty (26). Moreover, it is possible that better climate forecasts may not improve dengue forecasts. For example, climate may determine dengue seasonality, but models characterizing seasonality using historical dengue data alone (e.g., the baseline SARIMA model) may be able to provide equivalent information about expected future incidence (6). Incorporating additional data also increases model complexity in the form of variability in those data, parameters, and structural assumptions.

Including estimated parameters or model structures that better match historical data or biological relationships may come at the expense of lower out-of-sample forecast skill.

Our finding that statistical models generally outperformed mechanistic models is another indicator of the potential downside of overly complex forecasting models. Statistical models may have performed slightly better because robust uncertainty estimates are easier to generate with standard statistical packages compared with tailored mechanistic models. For example, the relatively simple baseline SARIMA models (4 parameters for Iquitos, 5 for San Juan) were developed with a standard statistical package and generally performed well compared with more complex models, including having the best overall calibration and the highest skill forecasts for peak week. Although simple models have also performed well in other forecasting challenges (27, 28), mechanistic models should not be dismissed. Mechanistic models allow for the incorporation of biological interactions (e.g., serotype interaction, spatial heterogeneity) and are essential for estimating the impacts of potential interventions (29). Statistical models can be used to guide development of better mechanistic models, capturing key components of good forecasts, such as seasonality, short-term autocorrelation, and accurate characterization of uncertainty. Moreover, hybrid approaches such as ensemble models, including statistical and mechanistic submodels, may be able to leverage advantages of both approaches.

Ensemble approaches were used by almost half the teams (7 of 16) (20, 21, 30) and on average, had better calibration and higher forecast skill than forecasts generated from single models. Moreover, a simple ensemble of all of the forecasts was among the highest scoring forecasts for every target and time point and was the only forecast to outperform the null forecast for all targets. Despite being a simple average of many forecasts, most of which performed substantially worse on their own, the ensemble balanced uncertainty across competing models with different assumptions and parameters, improving calibration by hedging bets when submodels disagreed and consolidating them when there was agreement. This cross-model modulation of uncertainty leads to higher skill forecasts as seen here and in other challenges (25, 28) and highlights a key advantage of multimodel and multiteam forecasting: a suite of models is likely to outperform any single approach (31). It also points to an important future research area: optimization of ensembles with fitted and dynamic weights.

While these insights can drive future research, there were also key limitations. For example, 2 potentially important dengue drivers were not assessed: vector populations and dengue virus serotypes. Vector data were simply not available on a spatiotemporal scale commensurate with the dengue data used here. Because numerous studies have shown that the interactions between dengue virus serotypes and human immunity may be a critical driver of long-term dengue dynamics (32), we provided datasets, including serotype data. However, only 2 teams chose to use them: one as an indicator of recent introduction of a serotype and the other in a complex 4-strain mechanistic compartmental model. The importance of serotype data for forecasting remains an open and important question, particularly for long-term dengue-endemic areas, such as Southeast Asia, where these effects may be strongest. Datasets with such extensive historical data are rare but offer an opportunity to identify key epidemic drivers that could inform current and future surveillance strategies in areas with less comprehensive historical data. Additionally, the comparison of approaches was only among the limited set used by the teams, not a comprehensive library of approaches. Different data and models have the potential to improve forecasts, but additional evidence is needed to understand which data and relationships are most important for dengue forecasting. Those determinations will also be key to future surveillance strategies, identifying the most important data to capture.

The challenge structure also had some limitations. Forecasts were evaluated on probabilities that were binned according to prespecified bins. Because targets are on different scales, it is not clear how to objectively define these bins to enable between-target comparison. It is also unclear how closely the bins should be tied to very specific decision-making needs, such as identifying an "outbreak," a concept with a wide variety of definitions that are intrinsically dependent both on surveillance and a threshold selection algorithm (33). Binned forecasts enable more comprehensive comparison of forecasts without selection of a specific threshold and allow scaling to higher levels, such as the binary probability of incidence exceeding a particular threshold. The datasets differed in both amount of data (13 seasons for Iquitos, 23 for San Juan) and characteristics of local dengue (serotype replacement in Iquitos, hyperendemicity in San Juan). Yet, those only represent 2 locations of the many where dengue is endemic. More datasets will be needed to determine the generalizability of forecasting tools, but few datasets with this level of detail exist. To evaluate forecasts over multiple seasons, the project was designed to use retrospective data and therefore, was not truly prospective. To facilitate forecasting at 13 time points per season, some future data were shared. To assure appropriate use, all teams agreed to forecast using data exclusively from weeks prior to and including the forecast week, and testing data were only available for 2 wk and only after selection of a final model and submission of training forecasts. Another challenge posed by these retrospective datasets is that they do not represent real-time reporting with its intrinsic reporting delays, another key forecasting challenge. Short-term forecasts for seasonal influenza show promise at helping bridge this gap (25), but comparable data were not available for this challenge, and the problem is far from solved. The datasets also do not represent all infections or even all cases, as we focused on laboratory-confirmed cases. Some cases do not seek care, do not have access to care, or are misdiagnosed. This may impact forecast model inputs and outputs, as both the underlying transmission dynamics and the case burden are imperfectly captured by data on confirmed cases.

Nonetheless, this project highlights important lessons for the larger panorama of challenges to advance the research and application of epidemic forecasting for public health. First, to make forecasts relevant to decision making in outbreak responses, targets should be clearly and quantitatively defined, and they should directly address specific public health needs. To integrate forecasts into decision making, it will be vital to refine the way that forecasts are communicated and maximize their operational relevance. Second, more participation leads to more information gain both for improved forecast skill via ensembles and also, for characterizing the strengths and weakness of different modeling approaches (25, 28, 34). Opening new data, facilitating access, and presenting engaging problems can drive participation and enable this type of research. Third, forecast skill should be openly evaluated on out-of-sample data with prespecified metrics that consider uncertainty. Self-evaluation of point predictions on data that are not openly accessible does little to characterize the utility of a forecasting tool. Good forecasts should be able to 1) differentiate between possible out-of-sample outcomes and 2) accurately express confidence in those predictions. Together, these components can be the building blocks for future forecasting systems, such as those that have transformed weather and storm forecasting (35).

Dengue remains a major public health challenge, and decades of dengue research have led to little progress in prospective prediction of dengue epidemics. Here, we identified key challenges and established a framework with datasets to help advance this research specifically toward targets that would benefit public health and forecasting science. Next generation models by the participating teams and others should adopt the testing–training framework, data, and metrics to assess forecast performance using the scores of the forecasts published here as benchmarks to measure advances in forecasting skill. At the same time, it may be important to refine targets and identify new targets to maximize public health utility. Additional datasets to retrospectively and prospectively develop and validate forecasts will be critical for demonstrating forecast skill and reliability across multiple seasons (and multiple locations for broader implementation). The recent epidemics of chikungunya and Zika viruses have further complicated clinical and laboratory-based surveillance for dengue and created a more complex immunological landscape for flaviviruses, changes that create new challenges for interpreting surveillance data and forecasting. There is also a need for improved surveillance data systems to ensure that data are machine readable and available in real time to support truly prospective, real-time forecasts. Lastly, better forecasts will drive interventions, increasing the importance of better mechanistic models that can both forecast and estimate the impact of interventions. These are formidable challenges, but through probabilistic forecasting projects, such as the one reported here, the community can move this research forward, translating the research into public health tools that can transform the way that we prepare for and respond to epidemics.

Materials and Methods

Data. Weekly laboratory-confirmed and serotype-specific dengue surveillance data were provided for 2 endemic locations: Iquitos, Peru (17, 36) and San Juan, Puerto Rico (19). Data were time referenced starting with 1 January, and data from 31 December (30 December for leap years) were removed to ensure 52 wk/y. The week with the lowest average incidence over the training period was then selected as the end week for the transmission season (week 26 in Iquitos and week 17 in San Juan) such that each dengue season began on the following week. All data were final, reflecting all cases with onset in each week regardless of reporting delays that affected the availability of data in real time. The data were divided into training data (Iquitos: 2000/2001 to 2008/2009, San Juan: 1990/1991 to 2008/2009) and testing data (2009/2010 to 2012/2013 for both locations). Climate and environmental data were provided for both locations (*SI Appendix*). Complete datasets are available at https://predict.cdc.gov and ref. 37. Participants were permitted to use other data (e.g., social media or demographic data) but not data on dengue in the study locations or nearby locations unless those data were made available to all participants.

Forecast Targets. For each season and location, the following targets were forecasted: *Peak week*, the week with the highest incidence of dengue (or undefined if more than 1 wk had the highest number of cases); *Peak incidence*, the number of dengue cases reported in the peak week; and *Total incidence*, the total number of confirmed dengue cases reported over the season.

Each forecast included a point estimate and a binned probability distribution. For peak week, each bin represented a single week (i.e., 1, 2, . . ., 52). For peak and total incidence, 11 bins were chosen empirically by setting an upper bound ~50% higher than the maxima observed in the training data. The maximum observed peak incidence in Iquitos was 116 cases, and we used bins of width 15 cases to cover up to 149 with 10 bins plus a final bin for 150 or more cases. For San Juan, with a maximum of 461 cases, we used bins of width 50 and 500 or more as the final bin. For total incidence, the maxima observed were 715 and 6,690 cases for Iquitos and San Juan, respectively. Bin widths were selected at 100 and 1,000 cases, respectively, with the last bin for >1,000 or >10,000 cases. Probabilities between 0 and 1 were assigned to each bin, summing to 1.0 for each specific forecast (e.g., the week 4 forecast for peak week in San Juan 2005/2006).

Forecasting. The forecasting project started on 5 June 2015, with public announcement of the challenge and online publication of the training datasets and forecast templates. Forecasting occurred in 2 stages. First, to participate, each team was required to submit a model description and a set of formatted forecasts for all 3 targets at both locations for the last 4 seasons of the training dataset (2005/2006 to 2008/2009) at 13 time points per season (weeks 0, 4, 8, . . ., 48) by email by 12 August 2015. Each team explicitly stated that these were out-of-sample forecasts using only the data from prior time points in all datasets used. The training forecasts and model descriptions were evaluated for adherence to the guidelines. Teams meeting those guidelines received the testing data on 19 August and had 2 wk to generate

and submit forecasts from the same model for the 4 testing seasons (2009/2010 to 2012/2013; deadline: 2 September 2015). The only incentives for participation were the provision of data, the opportunity to compare prospective forecasts, and the opportunity to participate in the development of this manuscript. Details are available at https://dengueforecasting.noaa.gov and https://predict.cdc.gov and in ref. 37.

We analyzed 3 additional models for comparison: a null model, a baseline model, and an ensemble model. The null model assigned equal probabilities to all bins (e.g., 1 of 52 for each possible peak week). The baseline models were SARIMA models, capturing seasonal trends and short-term autocorrelation [SARIMA(1, 0, 0)(4, 1, 0)$_{12}$ for San Juan and SARIMA(1, 0, 0)(3, 1, 0)$_{12}$ for Iquitos] (6). Finally, the ensemble model was created by averaging the probability bins from all team forecasts and the baseline forecast.

Evaluation. All forecasts were evaluated using the logarithmic score, a proper scoring rule based on probability densities (24, 38). The logarithmic score is the average logarithm of the probability assigned to the observed outcome bin (described above), p_i, over n predictions: $S_n = \frac{1}{n}\sum_{i=1}^{n}\log(p_i)$. We used Bayesian generalized linear models to identify season and model characteristics potentially related to forecast skill (*SI Appendix*). All analyses were performed in R (https://www.r-project.org/).

aDivision of Vector-Borne Diseases, Centers for Disease Control and Prevention, San Juan 00920, Puerto Rico; bDepartment of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA 02115; cData Analytics, Areté Associates, Northridge, CA 91324; dSystems Integration Branch, Johns Hopkins University Applied Physics Laboratory, Laurel, MD 20723; eDepartment of Environmental Health Sciences, Mailman School of Public Health, Columbia University, New York, NY 10032; fData to Decisions Cooperative Research Center, Kent Town, SA 5067, Australia; gHeinz College Information System Management, Carnegie Mellon University, Adelaide, SA 5000, Australia; hSchool of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213; iDepartment of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213; jDepartment of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205; kDepartment of Biostatistics and Epidemiology, School of Public Health and Health Sciences, University of Massachusetts, Amherst, MA 01003; lDepartment of Biology, University of Florida, Gainesville, FL 32611; mEmerging Pathogens Institute, University of Florida, Gainesville, FL 32611; nDepartment of Biological Sciences, University of Notre Dame, Notre Dame, IN 46556; oEck Institute for Global Health, University of Notre Dame, Notre Dame, IN 46556; pHospital for Tropical Diseases, Oxford University Clinical Research Unit, Ho Chi Minh City, Vietnam; qDepartment of Infectious Disease Epidemiology, London School of Hygiene & Tropical Medicine, London WC1E 7HT, United Kingdom; rClimate and Health Program, Barcelona Institute for Global Health, 08003 Barcelona, Spain; sCollege of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter EX4 4QF, United Kingdom; tPredictia Intelligent Data Solutions, 39005 Santander, Spain; uScientific Computation Program, Oswaldo Cruz Foundation, Rio de Janeiro 21040-900, Brazil; vDepartment of Mathematical Biology, Indian Statistical Institute, Kolkata, India 700108; wPasteur Kyoto International Joint Research Unit for Integrative Vaccinomics, 606-8501 Kyoto, Japan; xDepartment of Global Health, Centre National de la Recherche Scientifique, 75016 Paris, France; yDepartment of Mathematics and Statistics, Mount Holyoke College, South Hadley, MA 01075; zRAND Corporation, Santa Monica, CA 90401; aaOpen Philanthropy, San Francisco, CA 94105; bbDepartment of Computer Science and Software Engineering, Miami University, Oxford, OH 45056; ccF. I. Proctor Foundation for Research in Ophthalmology, University of California, San Francisco, CA 94122; ddInformation Science and Technology, Hokkaido University, Sapporo 060-0808, Japan; eeDivision of Environmental Health Sciences, School of Public Health, University of Minnesota, Twin Cities, MN 55455; ffVectorAnalytica, Washington, DC 20007; ggDepartment of Aeronautical Engineering, Universidade de Sao Paolo, Sao Paolo 13566-590, Brazil; hhDepartment of Philosophy, University of Miami, Coral Gables, FL 33146; iiDepartment of Statistics, Virginia Tech, Blacksburg, VA 24060; jjIntegrative Biology, University of South Florida, Tampa, FL 33620; kkDepartment of Biology, Stanford University, Stanford, CA 94305; llInfectious Diseases, College of Veterinary Medicine, University of Georgia, Athens, GA 30602; mmOdum School of Ecology, University of Georgia, Athens, GA 30602; nnDepartment of Geography, University of Florida, Gainesville, FL 32608; ooSchool of Life Sciences, University of KwaZulu, Natal 3629, South Africa; ppDepartment of Medicine, State University of New York Upstate Medical University, Syracuse, NY 13421; qqDepartment of Biostatistics, University of Michigan, Ann Arbor, MI 48109; rrDepartment of Civil and Environmental Engineering, West Virginia University, Morgantown, WV 26505; ssDepartment of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD 20742; ttPuerto Rico Department of Health, San Juan 00927, Puerto Rico; uuDepartment of Pathology, Microbiology, and Immunology, School of Veterinary Medicine, University of California, Davis, CA 95616; vvDepartment of Environmental, Agricultural, and Occupational Health, College of Public Health, University of Nebraska Medical Center, Omaha, NE 68198; wwInfluenza Division, Centers for Disease Control and Prevention, Atlanta, GA 30329; xxArmed Forces Health Surveillance Branch, Department of Defense, Silver Spring, MD 20904; yyClimate Program Office, National Oceanic and Atmospheric Administration, Silver Spring, MD 20910; zzLeidos supporting the Biomedical Advanced Research and Development Authority, Department of Health and Human Services, Washington, DC 20201; aaaBureau of Oceans, International Environmental and Scientific Affairs, US Department of State, Washington, DC 20520; bbbOffice of Science and Technology Policy, The White House, Washington, DC 20502; cccBNext, In-Q-Tel, Arlington, VA 22201; and dddNational Center for Medical Intelligence, Defense Intelligence Agency, Fort Detric, MD 21702

1. S. Bhatt et al., The global distribution and burden of dengue. *Nature* **496**, 504–507 (2013).
2. M. A. Johansson, D. A. Cummings, G. E. Glass, Multiyear climate variability and dengue–El Niño southern oscillation, weather, and dengue incidence in Puerto Rico, Mexico, and Thailand: A longitudinal data analysis. *PLoS Med.* **6**, e1000168 (2009).
3. W. G. van Panhuis et al., Region-wide synchrony and traveling waves of dengue across eight countries in Southeast Asia. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 13069–13074 (2015).
4. D. Constenla, C. Garcia, N. Lefcourt, Assessing the economics of dengue: Results from a systematic review of the literature and expert survey. *Pharmacoeconomics* **33**, 1107–1135 (2015).
5. D. S. Shepard, E. A. Undurraga, Y. A. Halasa, J. D. Stanaway, The global economic burden of dengue: A systematic analysis. *Lancet Infect. Dis.* **16**, 935–941 (2016).
6. M. A. Johansson, N. G. Reich, A. Hota, J. S. Brownstein, M. Santillana, Evaluating the performance of infectious disease forecasts: A comparison of climate-driven and seasonal dengue forecasts for Mexico. *Sci. Rep.* **6**, 33707 (2016).
7. M. A. Johansson, J. Hombach, D. A. Cummings, Models of the impact of dengue vaccines: A review of current research and potential approaches. *Vaccine* **29**, 5860–5868 (2011).
8. R. C. Reiner Jr et al., A systematic review of mathematical models of mosquito-borne pathogen transmission: 1970-2010. *J. R. Soc. Interface* **10**, 20120921 (2013).
9. R. Lowe et al., Evaluating probabilistic dengue risk forecasts from a prototype early warning system for Brazil. *eLife* **5**, e11285 (2016).
10. N. Ferguson, R. Anderson, S. Gupta, The effect of antibody-dependent enhancement on the transmission dynamics and persistence of multiple-strain pathogens. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 790–794 (1999).
11. D. A. Cummings, I. B. Schwartz, L. Billings, L. B. Shaw, D. S. Burke, Dynamic effects of antibody-dependent enhancement on the fitness of viruses. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 15259–15264 (2005).
12. H. J. Wearing, P. Rohani, Ecological and immunological determinants of dengue epidemics. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 11802–11807 (2006).
13. B. Adams et al., Cross-protective immunity can account for the alternating epidemic pattern of dengue virus serotypes circulating in Bangkok. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 14234–14239 (2006).
14. J. Lourenço, M. Recker, Natural, persistent oscillations in a spatial multi-strain disease system with application to dengue. *PLoS Comput. Biol.* **9**, e1003308 (2013).
15. Pandemic Prediction and Forecasting Science and Technology Working Group, Towards epidemic prediction: Federal efforts and opportunities in outbreak modeling. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/NSTC/towards_epidemic_prediction-federal_efforts_and_opportunities.pdf. Accessed 22 October 2019.
16. World Health Organization and Special Programme for Research and Training in Tropical Diseases, *Dengue: Guideline for Diagnosis, Treatment, Prevention and Control* (World Health Organization, 2009).
17. B. M. Forshey et al.; NMRCD Febrile Surveillance Working Group, Arboviral etiologies of acute febrile illnesses in Western South America, 2000-2007. *PLoS Negl. Trop. Dis.* **4**, e787 (2010).

18. S. T. Stoddard *et al.*, Long-term and seasonal dynamics of dengue in Iquitos, Peru. *PLoS Negl. Trop. Dis.* **8**, e3003 (2014).
19. T. M. Sharp *et al.*, Virus-specific differences in rates of disease during the 2010 Dengue epidemic in Puerto Rico. *PLoS Negl. Trop. Dis.* **7**, e2159 (2013).
20. L. R. Johnson *et al.*, Phenomenological forecasting of disease incidence using heteroskedastic Gaussian processes: A dengue case study. *Ann. Appl. Stat.* **12**, 27–66 (2018).
21. A. L. Buczak *et al.*, Ensemble method for dengue prediction. *PLoS One* **13**, e0189988 (2018).
22. L. Held, S. Meyer, J. Bracher, Probabilistic forecasting in infectious disease epidemiology: The 13th Armitage lecture. *Stat. Med.* **36**, 3443–3460 (2017).
23. G. W. Brier, Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **78**, 1–3 (1950).
24. T. Gneiting, A. E. Raftery, Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **102**, 359–378 (2007).
25. M. Biggerstaff *et al.*, Results from the second year of a collaborative effort to forecast influenza seasons in the United States. *Epidemics* **24**, 26–33 (2018).
26. R. Lowe *et al.*, Climate services for health: Predicting the evolution of the 2016 dengue season in Machala, Ecuador. *Lancet Planet. Health* **1**, e142–e151 (2017).
27. J. Lega, H. E. Brown, Data-driven outbreak forecasting with a simple nonlinear growth model. *Epidemics* **17**, 19–26 (2016).
28. C. Viboud *et al.*; RAPIDD Ebola Forecasting Challenge group, The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt. *Epidemics* **22**, 13–21 (2018).
29. D. Manheim, M. Chamberlin, O. A. Osoba, R. Vardavas, M. J. S. M. Moore, *Improving Decision Support for Infectious Disease Prevention and Control* (Rand Corporation, 2016).
30. T. K. Yamana, S. Kandula, J. Shaman, Superensemble forecasts of dengue outbreaks. *J. R. Soc. Interface* **13**, 20160410 (2016).
31. R. Polikar, Ensemble based systems in decision making. *IEEE Circuits Syst. Mag.* **6**, 21–45 (2006).
32. N. G. Reich *et al.*, Interactions between serotypes of dengue highlight epidemiological impact of cross-immunity. *J. R. Soc. Interface* **10**, 20130414 (2013).
33. O. J. Brady, D. L. Smith, T. W. Scott, S. I. J. E. Hay, Dengue disease outbreak definitions are implicitly variable. *Epidemics* **11**, 92–102 (2015).
34. P. E. Tetlock, B. A. Mellers, N. Rohrbaugh, E. Chen, Forecasting tournaments: Tools for increasing transparency and improving the quality of debate. *Curr. Dir. Psychol. Sci.* **23**, 290–295 (2014).
35. D. R. Wernly, L. W. Uccellini, "Storm forecasting for emergency response" in *Storms*, R. A. Pielke, R. A. Pielke, Eds. (Routledge, London, UK, 2000), pp. 70–97.
36. R. C. Reiner Jr *et al.*, Time-varying, serotype-specific force of infection of dengue virus. *Proc. Natl. Acad. Sci. U.S.A.* **111**, E2694–E2702 (2014).
37. M. A. Johansson, Dengue forecasting project data repository. Github. https://github.com/cdcepi/dengue-forecasting-project-2015. Deposited 25 October 2019.
38. R. Rosenfeld, J. Grefenstette, D. Burke, A proposal for standardized evaluation of epidemiological models. https://delphi.midas.cs.cmu.edu/files/StandardizedEvaluation_Revised_12-11-09.pdf. Accessed 20 October 2019.