Trade-offs between individual and ensemble forecasts of an emerging infectious disease

Rachel J. Oidtman[1,2,3,*], Elisa Omodei[2], Moritz U. G. Kraemer[4,5,6], Carlos A. Castañeda-Orjuela[7], Erica Cruz-Rivera[7], Sandra Misnaza-Castrillón[7], Myriam Patricia Cifuentes[8], Luz Emilse Rincon[8], Viviana Cañon[9], Pedro de Alarcon[10], Guido España[1], John H. Huber[1], Sarah C. Hill[4,11], Christopher M. Barker[12], Michael A. Johansson[13], Carrie A. Manore[14], Robert C. Reiner, Jr.[15], Isabel Rodriguez-Barraquer[16], Amir S. Siraj[1], Enrique Frias-Martinez[17], Manuel García-Herranz[2,*], and T. Alex Perkins[1,*]

[1] *Department of Biological Sciences and Eck Institute for Global Health, University of Notre Dame, Notre Dame, IN, USA*
[2] *UNICEF, New York, NY, USA*
[3] *Department of Ecology and Evolution, University of Chicago, Chicago, IL, USA*
[4] *Department of Zoology, University of Oxford, Oxford, UK*
[5] *Boston Children's Hospital, Boston, MA, USA*
[6] *Harvard Medical School, Boston, MA, USA*
[7] *Instituto Nacional de Salud, Bogotá, Colombia*
[8] *Ministerio de Salud y Protección Social, Bogotá, Colombia*
[9] *UNICEF, Bogotá, Colombia*
[10] *LUCA Telefonica Data Unit, Madrid, Spain*
[11] *Department of Pathobiology and Population Sciences, The Royal Veterinary College, London, UK*
[12] *Department of Pathology, Microbiology, and Immunology, School of Veterinary Medicince, University of California, Davis, CA, USA*
[13] *Division of Vector-Borne Diseases, Centers for Disease Control and Prevention, San Juan, Puerto Rico*
[14] *Information Systems and Modeling (A-1), Los Alamos National Laboratory, Los Alamos, NM, USA*
[15] *Institute for Health Metrics and Evaluation, University of Washington, Seattle, WA, USA*
[16] *Department of Medicine, University of California, San Francisco, CA, USA*
[17] *Telefonica Research, Madrid, Spain*
[*] *Corresponding authors: rjoidtman@gmail.com, mgarciaherranz@unicef.org, taperkins@nd.edu*

June 30, 2021

Abstract

When new pathogens emerge, numerous questions arise about their future spread, some of which can be addressed with probabilistic forecasts. The many uncertainties about the epidemiology of emerging pathogens can make

it difficult to choose among model structures and assumptions, however. To assess the potential for uncertainties about emerging pathogens to affect forecasts of their spread, we evaluated the performance of a suite of 16 forecasting models in the context of the 2015-2016 Zika epidemic in Colombia. Each model featured a different combination of assumptions about the role of human mobility in driving transmission, spatiotemporal variation in transmission potential, and the number of times the virus was introduced. All models used the same core transmission model and the same iterative data assimilation algorithm to generate forecasts. By assessing forecast performance through time using logarithmic scoring with ensemble weighting, we found that which model assumptions had the most ensemble weight changed through time. In particular, spatially coupled models had higher ensemble weights in the early and late phases of the epidemic, whereas non-spatial models had higher ensemble weights at the peak of the epidemic. We compared forecast performance of the equally weighted ensemble model to each individual model and identified a trade-off whereby certain individual models outperformed the ensemble model early in the epidemic but the ensemble model outperformed all individual models on average. On balance, our results suggest that suites of models that span uncertainty across alternative assumptions are necessary to obtain robust forecasts in the context of emerging infectious diseases.

1 Introduction

Pathogen emergence, or the phenomenon of novel or established pathogens invading a new host population, has been occurring more frequently in recent decades [1]. In the last 40 years, more than 150 pathogens of humans have been identified as emerging or re-emerging [2, 3]. In these situations, host populations are largely susceptible, which can result in dynamics ranging from self-limiting outbreaks, as with Lassa virus [4], to sustained pandemics, as with HIV [5], depending on the pathogen's traits and the context in which it emerges. When emergence does occur, mathematical models can be helpful for anticipating the future course of the pathogen's spread [6, 7, 8].

A necessary part of using models to forecast emerging pathogens is making decisions about how to handle the many uncertainties associated with these unfamiliar microbes [8]. Given the biological and ecological diversity of emerging pathogens, there is often considerable uncertainty about various aspects of their natural histories, such as their potential for superspreading [9], the role of human mobility in their spatial spread [10, 11], drivers of spatiotemporal variation in their transmission [6, 12], and even their modes of transmission [13]. In the case of MERS-CoV, for example, it took years to determine that the primary transmission route was spillover from camels rather than sustained human-to-human transmission [14]. A lack of definitive understanding about such basic aspects of natural history represents a major challenge for forecasting emerging pathogens.

Inevitably, different forecasters make diverse choices about how to address unknown aspects of an emerging pathogen's natural history, as they do for numerous model features. This diversity of approaches has itself been viewed as part of the solution to the problem of model uncertainty, based on the idea that the biases of different models might counteract one another to produce a reliable forecast when viewed from the perspective of an ensemble of models [15]. This idea has support in multi-model efforts to forecast seasonal transmission of endemic pathogens, such as influenza and dengue viruses [16, 17, 18, 19, 20], with ensemble forecasts routinely outperforming individual models. These successes with endemic pathogens have motivated multi-model approaches in response to several emerging pathogens, including forecasting challenges for chikungunya [21] and Ebola [22], vaccine trial site selection for Zika [23], and a multi-model decision-making framework for COVID-19 [15, 24].

Although there has been increased attention to multi-model forecasting of emerging pathogens in the last few years, these initiatives have involved significant effort to coordinate forecasts among multiple modeling groups [25, 26]. Coordination across multiple groups has clear potential to add value beyond what any single modeling group can offer alone. At the same time, using multiple models to hedge against uncertainties about a pathogen's natural history could potentially improve forecasts from a single modeling group, too [16, 18]. This could, in turn, improve ensemble forecasts based on contributions from multiple modeling groups. An ensemble-based approach by one modeling group that contributes to forecasts of seasonal influenza in the United States demonstrates the success that a single modeling group can achieve with an ensemble-based approach [27], and that such an ensemble can contribute value to an ensemble of forecasts from multiple modeling groups [18]. Similar approaches have not been widely adopted for forecasting emerging pathogens by a single modeling group (although see [28]), despite the heightened uncertainty inherent to emerging pathogens.

Here, we evaluate the potential for an ensemble of models that span uncertainties in pathogen natural history, but share a common core structure, to accurately forecast the dynamics of an emerging pathogen. We do so in the context of the 2015-2016 Zika epidemic in Colombia, which was well-characterized epidemiologically (Fig. 1) [29] and involved potentially consequential uncertainties about: i) the role of human mobility in facilitating spread across the country [30], ii) the relationship between environmental conditions and transmission of this mosquito-borne virus [6, 12], and iii) the number of times the virus was introduced into the country [31]. In this retrospective analysis, we used data assimilation to update 16 distinct models throughout the epidemic period and assessed forecast performance of all models relative to an equally weighted ensemble model. This allowed us to quantify the contribution of variants of each of the three aforementioned uncertainties to model performance during different phases of the epidemic. In doing so, we sought to not only assess the performance of the ensemble model relative to individual models, but also to learn about features of individual models that may be associated with improved forecast accuracy over the course of an epidemic.

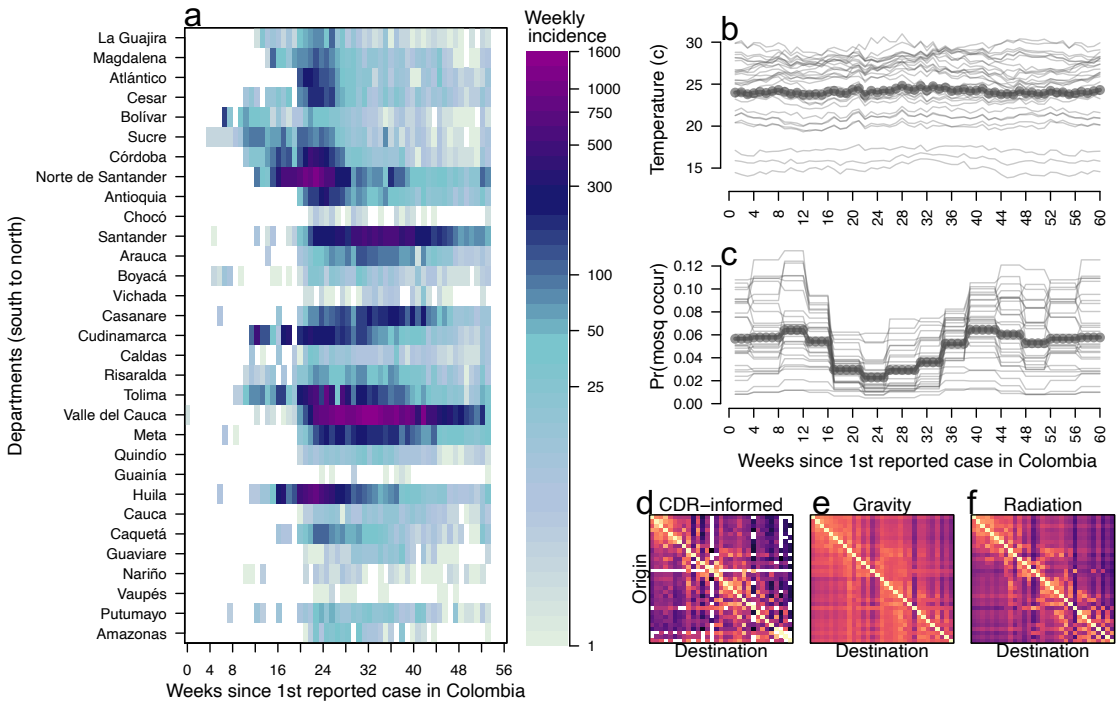

Figure 1: **Temporal and spatial variation of Zika incidence, temperature, and mosquito occurrence probability in Colombia.** *a.* Weekly Zika incidence from August 9, 2015 to October 1, 2016 with all 31 mainland departments approximately ordered from south to north. *b.* Points indicate average temperature data and lines indicate temperature by department. *c.* Points indicate average mosquito occurrence probability and lines indicate mosquito occurrence probability by department. *d-f.* Mobility matrices under three different assumptions of mobility, with departments ordered south to north on y-axis and north south on x-axis. Tan indicates high rates of mobility, dark purple indicates low rates of mobility, white indicates no movements.

2 Results

2.1 General forecast performance

Before any data assimilation had occurred, our 16 models (See Table 1) initially forecasted very low incidence across most departments over the 60-week period of our analysis (Figs. 2 top row, S12). Even so, short-term forecasts over a four-week horizon were consistent with the still-low observed incidence at that time (Figs. S3 purple, S18). By the time twelve weeks of data had been assimilated into the models, forecasts over the 60-week period of our analysis were considerably higher than the initial forecasts and better aligned with the observed trajectory of the epidemic (Figs. 2 second row, S13). Over those first twelve weeks, model parameters changed modestly (Fig. S6) and correlations among parameters began to emerge (Figs. S7, S8, S9, S10). We observed a more substantial change in the proportion of individual stochastic realizations (where the n^{th} stochastic realization is the n^{th} "particle" generated from some set of parameters $\vec{\theta}_{t,n}$ at time t) resulting in an epidemic, with those particles resulting in no epidemic being filtered out almost entirely by week 12 (Fig. S1). Because each particle retained its stochastic realization of past incidence across successive data assimilation periods, stochastic realizations of past incidence were inherited by particles much like parameter values. By week 24, many of the models correctly recognized that they were at or near the epidemic's peak and forecasted a downward trajectory for the remainder of the 60-week period of our analysis (Figs. 2 third row, S27). The particle filtering algorithm replaced nearly half of the original particles by that point (Fig. S2), with the new particles consisting of stochastic realizations of past incidence selected through data assimilation and updated every four weeks with forward simulations based on either original or new parameter combinations. As the end of the 60-week period of our analysis was approached, parameter correlations continued to strengthen (Figs. S7, S8, S9, S10), our estimate of the reporting probability increased (Fig. S6), and only around 20% of the original particles remained (Fig. S1).

Table 1: Different model assumptions regarding the role of human mobility in facilitating pathogen spread across the country, the relationship between environmental conditions and transmission of ZIKV, and the number of times the virus was introduced into Colombia. The suite of 16 models reflected factorial combinations of these three assumptions.

Human mobility	Transmission potential	Number of ZIKV introductions
CDR-informed	Fixed R [6]	One
Gravity model	Dynamic R [12]	Two
Radiation model		
No human mobility		

2.2 Model-specific forecast performance

To quantify the forecast performance of individual models over time, we used logarithmic scoring (hereafter, log scoring) to compare forecasts of cumulative incidence four weeks into the future to observed values at departmental and national levels. We assessed log scores once the first case was reported nationally for spatially coupled models (i.e., models with explicit human mobility), and once the first case was

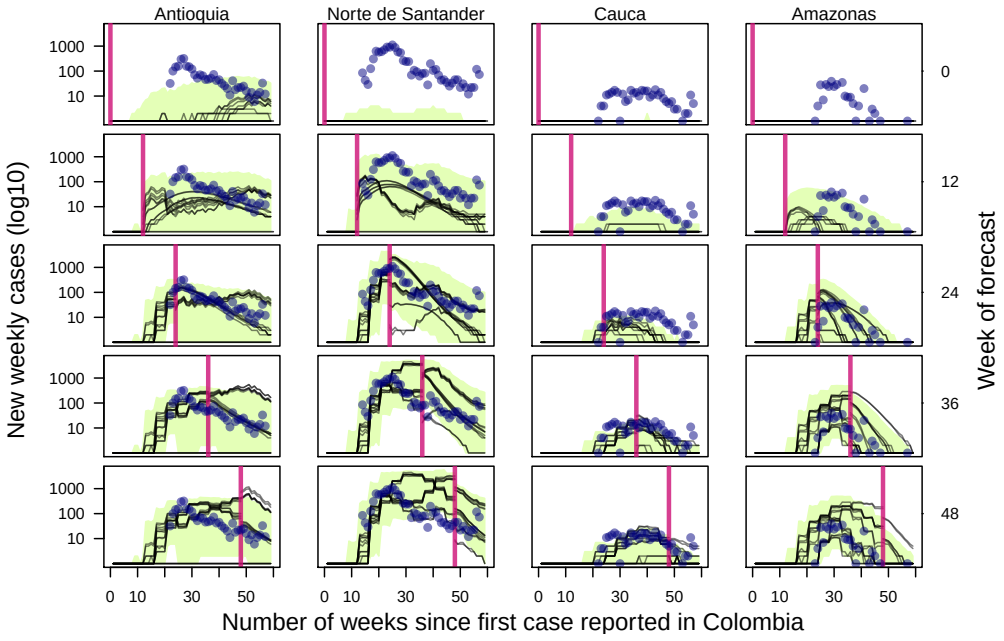

Figure 2: **Observed incidence (navy points) with the median forecast for 16 models (black lines) with the equally weighted ensemble model (green band) for Antioquia, Norte de Santander, Cauca, and Amazonas at five points throughout the epidemic.** Plotted departments reflect differences in population, epidemic size, and geographic regions of Colombia and are represented by each column. The vertical pink line indicates the point at which the forecast was made (also labeled on the right axis), with data to the left of the line assimilated into the model fit. Forecasts to the right of the vertical line change as more data is assimilated into the model, while model fits to the left of the vertical line do not change. The green band reflects the 50% credible interval of the equally weighted ensemble model.

reported in each department for non-spatial models (i.e., models with no explicit human mobility). Log scores were generally high for spatially coupled models early in the epidemic, given that observed cases and forecasts were both low at that time (Fig. S18, a-c). By week 12, as cases were reported in more departments, the accuracy of forecasts from non-spatial models improved (Fig. S18 d onward). Forecast performance around the peak of the epidemic differed considerably across models and departments, with forecasts from non-spatial models being somewhat lower than observed incidence and forecasts from spatially coupled models being somewhat higher (Fig. S14, Fig. S18 f-j). Around the peak of the epidemic, forecasts from spatially coupled models generally had higher log scores in departments with lower incidence (e.g., Nariño). Later in the epidemic (weeks 40-56), some models continued to forecast higher incidence than observed in some departments, despite having passed the peak incidence of reported cases (Fig. S16). In particular, models that used the dynamic instead of the static formulation of the reproduction number (i.e., the temporal relationship between R and environmental drivers is dynamic instead of static) were more susceptible to this behavior (note lower log scores in "Rt" versus "R" models in Fig. S18 k-o), given that their forecasts were sensitive to seasonal changes in temperature and mosquito occurrence.

Next, we used these log scores in an expectation-maximization (EM) optimization algorithm [32] to identify an optimal weighting of retrospective model-specific forecasts into an ensemble forecast (Fig. S25-S29) in each forecasting period (Fig. S17). To learn how model assumptions affected the inclusion of different models into the optimally weighted ensemble for each forecasting period, we summed and then normalized models' ensemble weights across each class of assumption (Fig. 3). Over the course of the epidemic, changes in weighting for the assumptions about human mobility and spatiotemporal variation in transmission, but not about the number of virus introductions into the country, closely followed patterns in the trajectory of the national epidemic. Spatially coupled models had most or all of the weight in the early and late stages of the epidemic, while non-spatial models had most of the weight around the peak of the epidemic (Fig. 3 b). Although the non-spatial models somewhat under-predicted incidence in the middle stages of the epidemic, this was often to a lesser extent than the spatially coupled models' over-predictions of incidence (Fig. S3). As a result, the EM algorithm achieved a balance between the over- and under-predictions of these different models.

The maximum ensemble weight in any forecasting period was 0.802, held by one model with a static R, two ZIKV introductions into the country, and CDR-informed human mobility 12 weeks after the first reported Zika case (Fig. S17). Combined, the two models with static R and CDR-informed human mobility data had the most instances of a non-zero ensemble weight (Fig. S17), occurring in 13 of 15 assimilation periods, with an average weight of 0.18. Around the peak of the epidemic, non-spatial models had the highest ensemble weight, reflecting the accuracy of short-term forecasts in some departments (e.g., Magdalena and Vaupés) and their overall accuracy in nationally-aggregated forecasts (Fig. S11). Near the end of the epidemic, the ensemble weight for models with a static R (Fig. 3 c) increased as their forecasts more closely matched the downturn of incidence later in the epidemic relative to models with dynamic R (Fig. S20). This was likely the result of mosquito occurrence probability and temperature becoming more favorable for transmission in many departments later in the epidemic (Fig. S21-S22), causing the dynamic R models to forecast a late resurgence in Zika incidence.

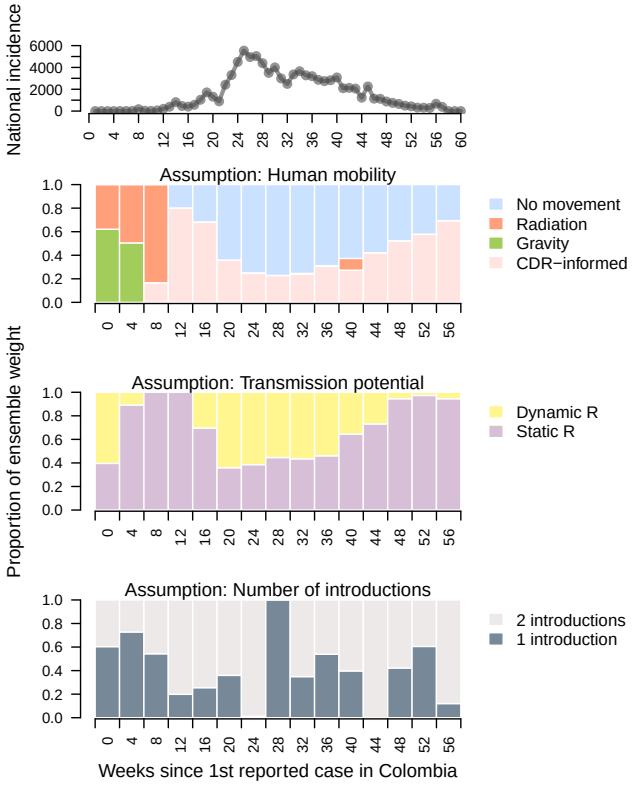

Figure 3: **Ensemble weight partitioned across assumptions about the role of human mobility in driving transmission, drivers of spatiotemporal variation in R, and the number of ZIKV introductions.** a: Weekly Zika incidence aggregated to the national scale. b-d: Proportion of ensemble weight across assumption type colored by explicit assumption.

2.3 Target-oriented forecast performance

Short-term changes in incidence are an important target of infectious disease forecasting, but there are other targets of potentially greater significance to public health decision making. To explore these, we evaluated the ability of the 16 models—and an evenly weighted ensemble—to forecast three targets at the department level: peak incidence, week of peak incidence, and onset week, which we defined as the week by which ten cases were first reported. We evaluated models based on log scores of these targets. Summing log scores across departments to allow for comparisons across different forecasting periods (Fig. 4), we found that, on average, the ensemble model outperformed every individual model for all three forecasting targets (indicated by the ensemble model's location on the y-axis). Early in the epidemic, spatially coupled models with a static R performed only slightly better (up to 1%) than the equally weighted ensemble (Fig. 4). For the remainder of the epidemic, the equally weighted ensemble model outperformed all individual models (Fig. 4). Such small changes in forecast performance when averaging over space shows that differences in forecast performance across space dominate relative to those across time.

By summing log scores across forecasting periods to allow for comparisons across departments (Fig. 5), we found that some individual models outperformed the ensemble model in forecasting the peak incidence and the week of peak incidence. In departments on the Caribbean Coast that experienced intermediate epidemic sizes (e.g., Antioquia, Sucre, Atlántico), spatially coupled models with a static R outperformed the ensemble model at forecasting the peak week by about 10% (Fig. 5 A). At those same locations, the equally weighted ensemble performed better than or similar to those same models at forecasting peak incidence and onset week (Fig. 5 b-c). Over forecasting periods and departments, the non-spatial models consistently had lower average forecast scores than the spatially coupled models (indicated by their location on the y-axis in Figs. 4-5). This trend appeared because initial forecasts from non-spatial models were not updated until the first case appeared in each department, while initial forecasts from spatially coupled models were updated when the first case appeared in the country.

3 Discussion

We assessed the potential for a suite of individual models that span a range of uncertainties, and ensembles of these models, to accurately forecast the dynamics of an emerging pathogen. Results from the general forecast performance analysis demonstrated that once we began assimilating data into models, forecasts rapidly became more accurate. Models were initialized with a wide range of parameter values [33], with many initial parameter combinations producing unrealistic forecast trajectories, but after only four assimilation periods (12 weeks), nearly 100% of those parameters that produced zero infections were dropped. Similar to other retrospective forecast analyses [16, 34], as more data were assimilated into the models over time, the model fits and forecasts generally became more closely aligned with temporal trends in the data. This was because the particle filter allowed model parameters to continually adapt to noisy data [35]. There were still some exceptions where the particle filter could not fully compensate for shortcomings of the transmission model, such as the drastic underestimates of incidence in departments with sub-optimal conditions for transmission (e.g., static R model in Risaralda in Fig. S20). At the same time,

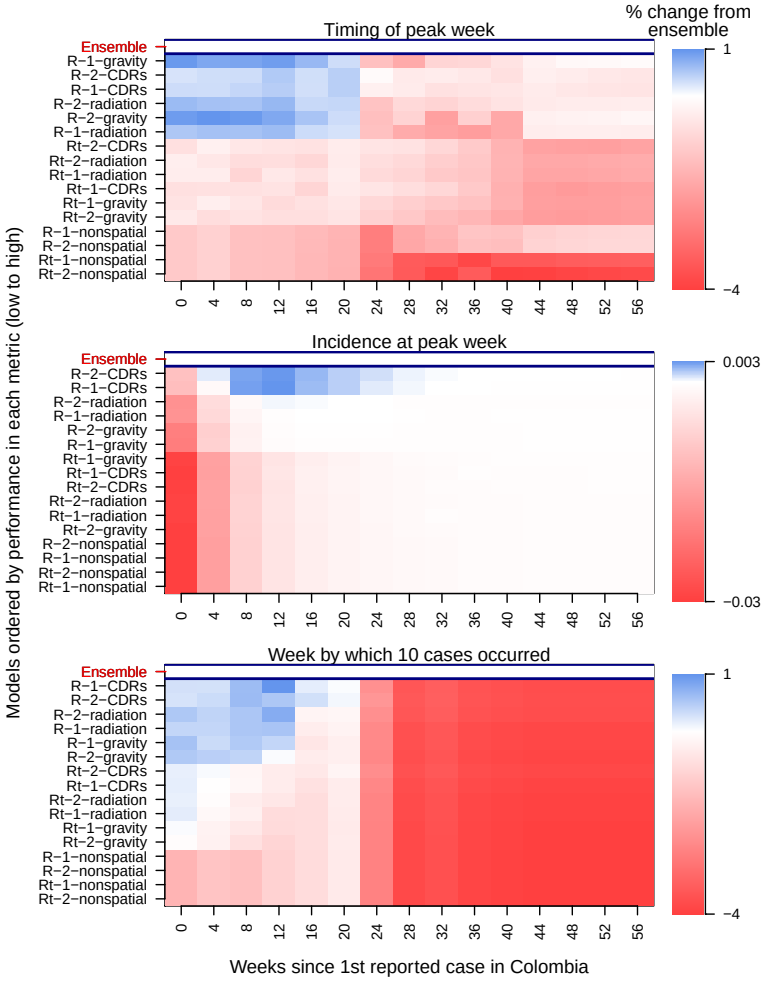

Figure 4: **Model-specific forecast scores relative to equally weighted ensemble model for each assimilation period and forecasting target.** *a.* Timing of peak week (within two weeks). *b.* Incidence at peak week. *c.* Onset week. Forecast scores are averaged over department. Models are ordered on the y-axis by average forecast score for each forecasting target. Model names on the y-axis are abbreviated such that "R" or "Rt" indicates assumption about spatiotemporal variation, "1" or "2" indicates number of introduction events, and "CDRs", "gravity", "radiation" or "nonspatial" indicates the human mobility assumption. In the heat plot, blue indicates individual model performed better than the ensemble model in a given department, red indicates individual model performed worse than ensemble model, and white indicates individual model performed roughly the same as the ensemble model.

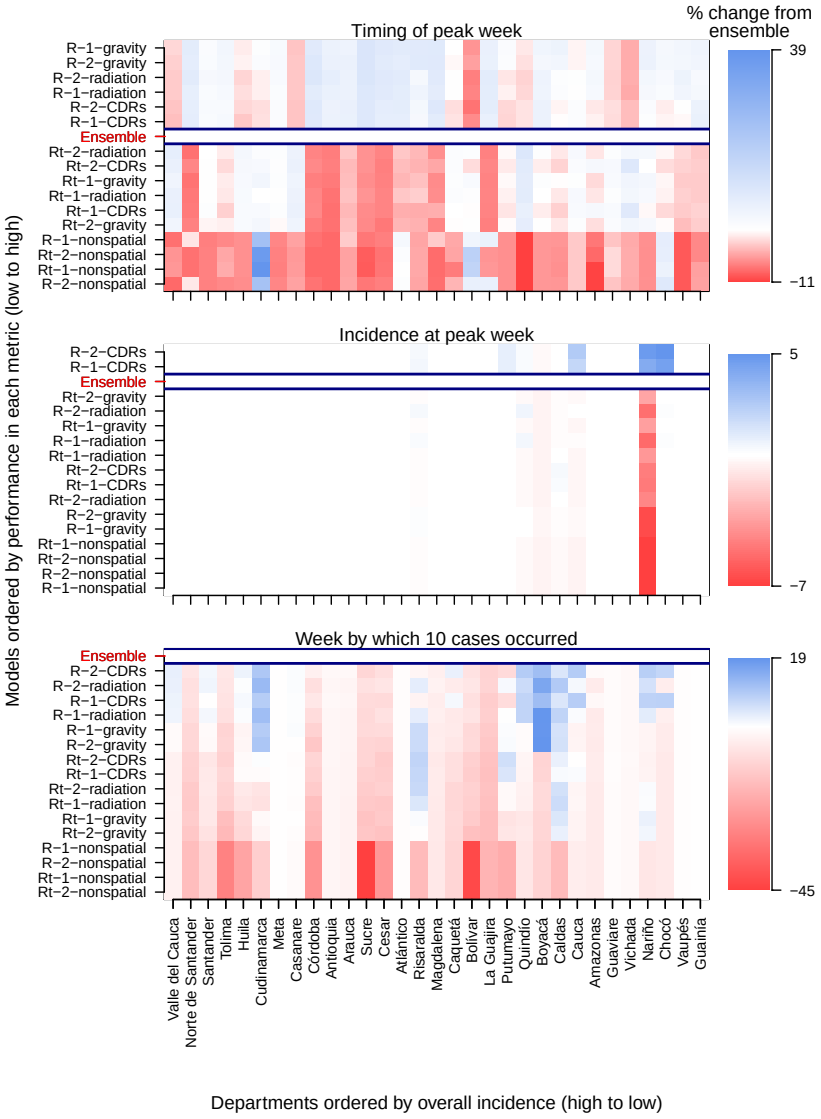

Figure 5: **Model-specific forecast scores relative to equally weighted ensemble model for each department and forecasting target.** *a.* Timing of peak week (within two weeks). *b.* Incidence at peak week. *c.* Onset week, or the week by which ten cumulative cases occurred. Forecast scores are averaged over department. Models are ordered on the y-axis by average forecast score for each forecasting target, with model names abbreviated in the same manner as Fig. 4. Departments are ordered on the x-axis from high to low for overall incidence. In the heat plot, blue indicates individual model performed better than the ensemble model in a given department, red indicates individual model performed worse than ensemble model, and white indicates individual model performed roughly the same as the ensemble model.

the broader suite of models buffered against shortcomings of any single transmission model.

In the model-specific forecast performance analysis, we identified clear temporal trends related to when models with a static R versus a dynamic R should be included in an optimally weighted ensemble. In contrast, there were no clear temporal trends in weighting regarding the assumption about the number of times the virus was introduced into the country, potentially reflecting that, even with multiple introductions, most transmission may have been linked to a single introduction [31]. Models with a dynamic R had higher weights in the ensemble at the peak of the epidemic, while models with a static R had higher weights at the beginning and end of the epidemic. This was likely due to temporal shifts in temperature and mosquito occurrence probabilities dominating forecasts of transmission potential for the models with a dynamic R. For example, in the latter parts of the epidemic when reported cases were declining, mosquito conditions and temperature became more suitable for transmission in many departments. This caused models with a dynamic R to forecast a resurgence in ZIKV transmission in those departments, while models with a static R forecasted a downturn in incidence that was more similar to the observed dynamics. This finding that susceptible depletion may have been more influential than temporal variation in environmental conditions for the Zika epidemic is consistent with recent findings for SARS-CoV-2 [36].

Through the model-specific forecast performance analysis, we also found that spatially coupled models had higher ensemble weights in the early and late stages of the epidemic, while non-spatial models had higher weights around the peak of the epidemic. The importance of including spatially coupled models in the optimally weighted ensemble early in the epidemic supports the general notion that human mobility may be particularly predictive of pathogen spread early in an epidemic [7, 30, 37, 38]. In part, temporal shifts in weighting around the peak of the epidemic were due to more accurate nationally-aggregated forecasts from the non-spatial models. This result was consistent with a previous modeling analysis of the invasion of chikungunya virus in Colombia, which showed that models fitted independently to sub-national time series recreated national-level patterns well when aggregated [39]. A shift in ensemble weights toward non-spatial models around the peak of the epidemic was also due to less accurate department-level forecasts from the spatially coupled models. At that point in the epidemic, prevalence was at its highest, which means that we would expect local epidemics to be more endogenously driven and less sensitive to pathogen introductions across departments.

In the target-oriented forecast performance analysis, we found that the equally weighted ensemble generally outperformed individual models, with a few key exceptions. In the months leading up to the peak of the epidemic, spatially coupled models with a static R had slightly, but consistently, higher forecast scores with respect to peak week and onset week. Like the model-specific analysis results, this result illustrates the importance of human mobility in facilitating the spread of an emerging pathogen across a landscape [30]. Individual models outperforming the equally weighted ensemble model in the early phase of the epidemic is not wholly surprising given that non-spatial models were represented equally in that ensemble throughout the epidemic. Non-spatial models may be realistic when locations have self-sustaining epidemics, but they are not appropriate for capturing early-phase growth and its dependence on importations [40]. Another instance when individual models had higher forecast scores than the equally weighted ensemble was with respect to peak week for spatially coupled models with a static R in departments

along the Caribbean Coast. Compared to dynamic R models, the static R models more accurately forecasted peak week in these departments (e.g., Magdalena, Cesar, Sucre), as they did not forecast a late-stage resurgence in transmission. The equal weighting of the dynamic R models in the ensemble therefore led to overall lower peak week forecast scores for the ensemble relative to static R models. Still, our results indicating that an equally weighted ensemble mostly outperformed individual models adds to the growing literature highlighting the importance of ensemble methods in epidemiological forecasting [16, 17, 27, 41, 42].

We considered both equally and optimally weighted ensembles and found that the equally weighted ensemble had a lower root mean square error than the optimally weighted ensemble (RMSE=0.640 and 0.705, respectively)—therefore providing slightly more accurate forecasts of the observed data (Fig. S23). With the optimally weighted ensemble, which we updated at each data assimilation period, we found that model weights changed over the course of the epidemic Fig. S17). Although this is intuitive given the changing nature of an emerging epidemic through time [8], it may be problematic in practice. It is almost as if the ensemble weights require their own forecast. On the one hand, promising new advances in ensemble modeling [27, 41]—such as adaptive stacking for seasonal influenza forecasting [43]—are being used to address this issue of identifying optimal, adaptive weights without training to historical data. On the other hand, in an emerging pathogen context, establishing optimal model weights by way of model fitting and forecast generation is often reliant on available incidence data (rather than historical data) that is highly variable [44], given the delayed nature of data reporting [45]. In this context, our results demonstrate that it is preferable to use an equally weighted ensemble to buffer against uncertainty in optimal ensemble weights. As is also being demonstrated in forecasts of COVID-19, equally weighted ensembles can provide accurate forecasts [26, 46, 44] and may be a better reflection of the considerable structural uncertainty inherent to models of emerging pathogen transmission [24].

A few limitations of our study should be noted. First, while an equally weighted ensemble approach allowed us to consider contributions of several alternative model assumptions, there was high uncertainty associated with these forecasts (sometimes spanning orders of magnitude, see Fig. S24). Potential end-users of these types of forecasts could consider high levels of uncertainty to be problematic for decision-making [47], though if the uncertainty does not affect the choice of a control measure, then the uncertainty may not be as relevant [48]. In the future, ensemble approaches aimed at increasing precision and reducing uncertainty [49, 27] could be used in conjunction with equally weighted ensembles. Second, we considered alternative models across only three assumptions. With ZIKV transmission, there are additional structural uncertainties that could be considered, such as the role of sexual transmission [50]. In real-time applications of our or other Zika forecasting models, it could be worthwhile to explore these types of ZIKV-specific structural uncertainties. Relatedly, the static and dynamic R had minor differences in their formulations, such that the static R also included a socioeconomic index. In future work, it could be interesting to explore if the inclusion of this time-independent variable affected the dynamic R. Third, in this analysis we did not explicitly consider delays in reporting that likely would have occurred had these forecasts been generated in real time [51]. In that context, temporally aggregating data to a wider interval (e.g., at 2-week intervals rather than 1-week intervals) could potentially help mitigate the effects of reporting delays to some extent. Fourth, we assumed that the reporting probability was constant through time. Although this is a standard assumption [52] given the

lack of data to inform a time-varying relationship for this mechanistic element [53], it would be interesting to include and test a reporting dynamics model (e.g., the reporting probability scales with incidence [54]) as an additional component included in our ensemble framework. Fifth, we conducted this analysis at the departmental level instead of this municipality level, which could obfuscate meaningful differences across regions of a single department [29]. In future work, it would be useful to test and assess our forecasting algorithm and outputs at different spatial scales [39].

As the world is reminded of on a daily basis with COVID-19, pathogen emergence is an ongoing phenomenon that will continue to pose threats in the future [55]. A better understanding of an emerging pathogen's natural history could help to reduce pathogen-specific structural uncertainties, but these insights may not always occur in time to inform model development for real-time forecasting [8]. Our results highlight important trade-offs between individual and ensemble models in this context. Specifically, we demonstrated that an equally weighted ensemble forecast was almost always more accurate than individual models. Instances in which individual models were better than the ensemble, or greatly improved the ensemble, also provided insight. For example, incorporating human mobility into models improved forecasts in the early and late phases of an epidemic, which underscores the importance of making aggregated mobility data available early in an epidemic [56]. The range of outcomes resulting from alternative modeling assumptions in model-specific forecasts demonstrates why it will continue to be important to address structural uncertainties in forecasting models in the future.

4 Materials and methods

4.1 Data

We used passive mandatory surveillance data for reported cases of Zika, from the National Surveillance System (Sivigila) at the first administrative level (31 mainland departments) in Colombia. To span the beginning, peak, and tail of the epidemic in Colombia, we focused on the 60-week period between August 9, 2015 and October 1, 2016. We used the version of these data collated by Siraj et al. [29], as well as modeled values of weekly average temperature and estimates of department-level population from that data set. For some models, we worked with monthly estimates of mosquito occurrence probability (i.e., dynamic R models) obtained from Bogoch et al. [57], and for others we worked with time-averaged estimates (i.e., static R models) from Kraemer et al. [58].

For models that relied on cell phone data to describe human mobility, we used anonymized and aggregated call detail records (CDRs). Every time a user receives or makes a call, a CDR including the time, date, ID, and the tower (BTS) providing the service is generated. The positions of the BTSs are georeferenced and so the aggregated mobility between towers can be tracked in time. We used this information to derive daily mobility matrices at the municipality level in Colombia from February 2015 to August 2015. Mobility matrices captured the number of individuals that moved in each given day from one municipality to another (i.e., that appeared in BTSs of different municipalities). The change for each day was captured by comparing the last known municipality to the current one. No individual information or records were available.

As these data did not align with the time frame of the epidemic, and to calculate a mobility matrix at a department level, we computed a representative mobility

matrix by summing all available CDRs within the municipalities of each department and normalizing them to sum to one relative to the sum of CDRs originating from that department. In five departments (Amazonas, Cudinamarca, Guainía, Vaupés, Vichada), the proportion of CDRs linking callers within the same department was below 60%. Given that this implied an unrealistically low proportion of time spent within an individual's department of residence, we interpreted those values as idiosyncrasies of the data and not representative of human mobility [59]. Thus, for those five departments, we replaced the proportion of within-department CDRs with the mean proportion of within-department CDRs from all other departments. We then re-normalized the number of CDRs originating from each department in our mobility matrix to sum to one.

4.2 Summary of models

To produce weekly forecasts of ZIKV transmission across Colombia, we sought to use a computationally efficient model with the flexibility to include relevant epidemiological and ecological mechanisms. We used a previously described semi-mechanistic, discrete-time, and stochastic model [60] that had been previously adapted and used to model mosquito-borne pathogen transmission [61, 62]. Using this model, we were able to account for the extended generation interval of ZIKV using overlapping pathogen generations across up to five weeks of the generation interval distribution of ZIKV [62]. Furthermore, we could specify this model to be either spatially connected or non-spatial—a key assumption that we considered in our analysis.

We considered a suite of 16 models that spanned all combinations of four assumptions about human mobility across Colombia's 31 mainland departments, two assumptions about the relationship between environmental conditions and the reproduction number (R), and two assumptions about how many times Zika virus was introduced to Colombia (Table 1). Twelve of 16 models allowed for spatial connectivity across departments [60], while four models were non-spatial. There were up to two steps in the transmission process: transmission across departments (for spatially connected models) and local transmission within departments.

Across departments, we simulated the movement of individuals using a spatial connectivity matrix ($\mathbf{H}$), the d^{th} column of which corresponds to the proportion of time spent by residents of department d in all departments $\vec{d}$. Using this matrix, we redistributed infections in department d in week t ($I_{d,t}$) across $\vec{d}$ as a multinomial random variable,

$$I'_{\vec{d},t} \sim \text{multinomial}(I_{d,t}, \mathbf{H}_{\vec{d},d}), \tag{1}$$

where the first and second arguments represent the number of trials and the probabilities of the outcomes, respectively. By taking this Lagrangian approach to modeling human mobility, transmission across departments can occur either by infected visitors transmitting to local susceptibles or susceptible visitors becoming infected by local infecteds. The relative occurrence of these events depends on the prevalence of infection, susceptibility, local transmission potential, and mobility patterns of a given pair of departments.

Within each department, we defined a variable representing the effective number of infections that could have generated new infections in week t ($I''_{d,t}$) as

$$I''_{d,t} = \sum_{j=1}^{5} \omega_j^{GI} I'_{d,t-j}, \tag{2}$$

where ω_j^{GI} is the probability that the generation interval is j weeks [63]. The relationship between $I''_{d,t}$ and the expected number of new local infections in week $t+1$ ($I_{d,t+1}$) follows

$$I_{d,t+1} = \beta_{d,t} \frac{I''_{d,t}}{N_d} S_{d,t}, \tag{3}$$

where $\beta_{d,t}$ is the transmission coefficient, N_d is the total population, and $S_{d,t}$ is the total susceptible population prior to local transmission in week t. We accounted for the role of stochasticity in transmission by using the stochastic analogue of Eqn. 3, such that

$$I_{d,t+1} \sim \text{negative binomial}\left(\beta_{d,t} \frac{I''_{d,t}}{N_d} S_{d,t}, I''_{d,t}\right) \tag{4}$$

where the first and second arguments are the mean and dispersion parameters, respectively [60].

To allow for comparison of the model's simulations of infections ($I_{d,t}$) with empirical data on reported cases ($y_{d,t}$), we applied a reporting probability (ρ) to simulated infections to obtain simulated cases ($C_{d,t}$), such that $C_{d,t} \sim \text{binomial}(I_{d,t}, \rho)$. Using this, we defined the contribution to the overall log-likelihood of the model and its parameters from a given department d and week t as

$$\boldsymbol{\ell}_{d,t}(\vec{\theta}_t) = \ln\left(\text{negative binomial}(y_{d,t}+1 \,|\, \phi, C_{d,t}+1)\right), \tag{5}$$

where ϕ is a dispersion parameter that we estimated to account for variability in case reporting beyond that captured by ρ. Shifting $y_{d,t}$ and $C_{d,t}$ by one in eq. (5) was intended to safeguard against $\boldsymbol{\ell}_{d,t}$ being undefined in situations where $C_{d,t} = 0$.

4.2.1 Assumptions about human mobility

We allowed for spatial coupling across departments in 12 of 16 models. In these models, we informed $\mathbf{H}$ in three alternative ways: i) with mobility data extracted from mobile phone CDRs, ii) with a gravity model, or iii) with a radiation model (Fig. 1d-f). For the gravity model, we used parameters previously fitted to CDRs from Spain and validated in West Africa [11]. For the radiation model, we calculated human mobility fluxes according to the standard formulation of this model [64], which depends only on the geographic distribution of population. In four of 16 models, we assumed that departments were spatially uncoupled (Table 1), such that each department was modeled individually with its own set of parameters. In those models, each department's epidemic was seeded independently with its own set of imported infections. Further details about the spatially uncoupled models can be found in the Supplemental Text.

4.2.2 Assumptions about environmental drivers of transmission

We parameterized the transmission coefficient, $\beta_{d,t}$, based on a description of the reproduction number, $R_{d,t}$, appropriate to environmental drivers for department d and time t. We considered two alternative formulations of $R_{d,t}$ that were informed by data that were available prior to the first reported case of Zika in Colombia. Specifically, both of these alternative formulations used different outputs from previous modeling efforts [6, 12] and because of this they contain slightly different components. Both formulations were defined such that

$$\beta_{d,t} = kR_{d,t} \tag{6}$$

where k is a scalar that we estimated over the course of the epidemic to account for the unknown magnitude of ZIKV transmission in Colombia. In addition to the summary below, further details about these formulations of $R_{d,t}$ are provided in the Supplementary Methods.

The formulation of $\beta_{d,t}$ that we refer to as "dynamic" is defined at each time t in response to average temperature at that time ($T_{d,t}$) and mosquito occurrence probability at that time ($OP_{d,t}$). This relationship can be expressed generically as

$$\beta_{d,t} = k\tilde{R}_{d,t}(T_{d,t}, OP_{d,t}|c, \psi, \alpha, v), \tag{7}$$

where c, ψ, α, and v are parameters governing the relationship among $T_{d,t}$, $OP_{d,t}$, and $\tilde{R}_{d,t}$. We informed the component of $\tilde{R}_{d,t}$ related to mosquito density with monthly estimates of $OP_{d,t}$, which derive from geostatistical modeling of *Aedes aegypti* occurrence records globally [57]. Other components of $\tilde{R}_{d,t}$, which include several temperature-dependent transmission parameters, were informed by laboratory estimates [12]. Given that this formulation of $\tilde{R}_{d,t}$ was not validated against field data prior to the Zika epidemic in Colombia, we estimated values of c, ψ, α, and v over the course of the epidemic.

The formulation of $\beta_{d,t}$ that we refer to as "static" is defined as a time-averaged value that is constant across all times t. Temporal variation in $T_{d,t}$ is still accounted for, but its time-varying effect on $R_{d,t}$ is averaged out over all times $\vec{t}$ to result in a temporally constant R_d. Mosquito occurrence probability is also incorporated through a temporally constant value (OP_d) [58]. The relationship among these variables can be expressed generically as

$$\beta_{d,t} = k\bar{R}_d(T_{d,\vec{t}}, OP_d, PPP_d), \tag{8}$$

where PPP_d is purchasing power parity in department d (a feature not included in the dynamic model) [65]. This input is an economic index that was intended to serve as a proxy for spatial variation in conditions that could affect exposure to mosquito biting, such as housing quality or air conditioning use [6]. Given that this formulation of $\bar{R}_d$ was informed by data from outbreaks of Zika and chikungunya prior to the Zika epidemic in Colombia, we did not estimate its underlying parameters over the course of the epidemic in Colombia.

4.2.3 Assumptions about introduction events

Although many ZIKV infections were likely imported into Colombia throughout the epidemic, we assumed that ZIKV introductions into either one or two departments drove the establishment of ZIKV in Colombia [31]. Under the two different scenarios, there was either one introduction event into one department or there were two independent introduction events into two randomly drawn departments. For each parameter set, the initial number of imported infections was seeded randomly between one and five in a single week, the timing of which was estimated as a parameter. Following the initial introduction(s), we assumed that ZIKV transmission was driven by a combination of movement of infected people among departments and local transmission within departments, as specified by each model.

4.3 Data assimilation and forecasting

For each particle, we produced a single forecast to "initialize" the model prior to the first reported case in Colombia. Beginning with the time of the first reported

case in Colombia, we then assimilated new data, updated parameter estimates, and generated forecasts every four weeks, consistent with the four-week frequency used by Johansson et al. in an evaluation of dengue forecasts [16]. We specified 20,000 initial parameter sets ($\vec{\theta}_{1,n}$), indexed by n, by drawing independent samples from prior distributions of each parameter [66] (see Supplemental Methods). Each parameter set was used to generate a corresponding particle: a stochastic realization of the state variables ($I_{d,1,n}$ and $C_{d,1,n}$). At each assimilation period, we normalized log-likelihoods summed across departments over the preceding four weeks to generate particle weights,

$$\omega(t,n) = \frac{\sum_d \sum_{\tau=t-3}^{t} \boldsymbol{\ell}_{d,\tau}(\vec{\theta}_{t,n})}{\sum_n \sum_d \sum_{\tau=t-3}^{t} \boldsymbol{\ell}_{d,\tau}(\vec{\theta}_{t,n})}. \quad (9)$$

Proportional to these particle weights ($\omega(t,n)$), we sampled 18,000 sets of corresponding parameters ($\vec{\boldsymbol{\theta}}_t^{\text{resampled}}$) and state variables ($\{\boldsymbol{I}_{d,t}^{\text{resampled}}, \boldsymbol{C}_{d,t}^{\text{resampled}}\}$) from time t with replacement and used them at the next data assimilation step four weeks later, where boldface indicates a set of parameters or state variables, respectively, over all n. In doing so, information including the initial prior assumptions ($\vec{\theta}_{1,n}$) and the likelihoods at each four-week period was assimilated into the model sequentially over time. Given that particle filtering algorithms are susceptible to particle drift—or the convergence of particles onto very few states through iterative re-sampling [33]—we also generated 2,000 new parameter sets at each data assimilation step. To do so, we drew random samples of model parameters from a multivariate normal distribution with parameter means and covariances fitted to the resampled 18,000 parameter sets ($\vec{\boldsymbol{\theta}}_t^{\text{resampled}}$). Whereas the 18,000 resampled parameter sets already included simulated values of state variables $I_{d,t,n}$ and $C_{d,t,n}$ through time t, the 2,000 new parameter sets did not and so we informed initial conditions of $\boldsymbol{I}_{d,t}^{\text{new}}$ with draws from $\boldsymbol{I}_{d,t:t}^{\text{resampled}}$ for those parameter sets at the time they were created. Together, the 18,000 resampled parameter sets ($\vec{\boldsymbol{\theta}}_t^{\text{resampled}}$) and the 2,000 new parameter sets ($\vec{\boldsymbol{\theta}}_t^{\text{new}}$) constituted the set of parameter sets used as input for the next data assimilation step ($\vec{\boldsymbol{\theta}}_{t+4} = \{\vec{\boldsymbol{\theta}}_t^{\text{resampled}}, \vec{\boldsymbol{\theta}}_t^{\text{new}}\}$). We also used this new set of parameters as the basis for forecasts made at time t, which simply involved simulating forward a single realization of the model for each parameter set.

4.4 Evaluating forecast performance

At each of the 15 time points at which we performed data assimilation through the 60-week forecasting period, we created an ensemble forecast that evenly weighted contributions from each of the 16 models [46]. To populate this ensemble, we specified 20,000 total samples, with 1,250 samples from each model. We assessed the model-specific performance of individual and ensemble forecasts using log scores, which are forecast scoring rules that assess both the precision and accuracy of forecasts [67]. For a specific forecasting target, z, and model, m, the log score is defined as $\log f_m(z^*|\mathbf{x})$, where $f_m(z|\mathbf{x})$ is the predicted density conditioned on the data, $\mathbf{x}$, and z^* is the empirical value of the target Z [16].

We computed log scores for departmental and national incidence over each four-week assimilation period. Following [17], we used an expectation-maximization algorithm to generate ensemble weights for each model in each assimilation period. For each model, we computed 32 log scores (i.e., one for each department and one

nationally). To compute the ensemble weight for a given model feature, such as the static R assumption, we summed the weights of all models with that feature.

We assessed target-oriented forecast performance using log scores for three forecasting targets: timing of peak week (within two weeks), incidence at peak week, and onset week, which we defined as the week by which ten cumulative cases occurred. These choices were motivated by forecasting assessments for influenza and dengue [16, 17, 18, 68] and deemed applicable to public health objectives for forecasting an emerging pathogen such as Zika. For peak week and onset week, we used modified log scores [18], such that predictions within two weeks of the correct week were considered accurate. We evaluated a total of 7,680 log scores, reflecting three targets for each of 16 models in each of 31 departments plus at the national level and at each of 15 time points at which data assimilation occurred.

As log scores only yield a relative measure of model performance, we used forecasting scores [18] as a way to retrospectively compare forecast performance for different forecasting targets. Whereas log scores are preferable for comparing performance across models on the same data, forecasting scores are preferable for comparing forecast performance across data composed of different locations and times. A forecasting score is defined simply as the exponential of the average log score, where the latter reflects an average over one or more indices, such as models, time points, targets, or locations.

Data availability

The mobile phone data set used in this study is proprietary and subject to strict privacy regulations. The access to this data set was granted after reaching a non-disclosure agreement with the proprietor, who anonymized and aggregated the original data before giving access to the authors. The data could be available on request after negotiation of a non-disclosure agreement. The contact person is Enrique Frías-Martínez (enrique.friasmartinez@telefonica.com). Epidemiological, meteorological, and demographic data are available from Siraj et al. [29] and additionally available on https://github.com/roidtman/eid_ensemble_forecasting.

Author contributions

RJO, EO, MUGK, CMB, MAJ, CAM, RCR, IR-B, MG-H, and TAP conceptualized the study; RJO, EO, MUGK, CAC-O, EC-R, AM-C, PC, LER, VC, PA, GE, JHH, SCH, ASS, EF-M, and MG-H provided and / or processed data; RJO, EO, MUGK, CA-O, EC-R, SM-C, PC, LER, VC, PA, EF-M, MG-H, and TAP participated in biweekly meetings to provide feedback on research; RJO, EO, MUGK, MG-H, and TAP developed the model and wrote the first draft of the manuscript; RJO, EO, MUGK, JHH, and SCH analyzed data; EO, MG-H, and TAP supervised the research; all authors reviewed the manuscript.

Acknowledgements

The authors would like to thank Clara Palau Montava for help with managing the early stages of this project. The authors would additionally like to thank the UNICEF-Colombia Representative, Aida Oliver Arostegui, INS Director, Martha

Lucia Ospina Martinez, and the past and present Ministers of the Ministry of Health, Juan Pablo Uribe Restrepo and Fernado Ruiz Gomez.

RJO acknowledges support from an Eck Institute for Global Health Fellowship, GLOBES grant, Arthur J. Schmitt Fellowship, and the UNICEF Office of Innovation. MUGK is supported by The Branco Weiss Fellowship - Society in Science, administered by the ETH Zurich and acknowledges funding from the Oxford Martin School and the European Union Horizon 2020 project MOOD (#874850). SCH is supported by the Wellcome Trust (220414/Z/20/Z). This research was funded in whole, or in part, by the Wellcome Trust [Grant number 220414/Z/20/Z]. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission

References

[1] Kate E. Jones et al. "Global trends in emerging infectious diseases". In: *Nature* 451.7181 (Feb. 2008), pp. 990–993. ISSN: 1476-4687. DOI: 10.1038/nature06536. URL: https://doi.org/10.1038/nature06536.

[2] Katherine F. Smith et al. "Global rise in human infectious disease outbreaks". In: *Journal of The Royal Society Interface* 11.101 (2014), p. 20140950. DOI: 10.1098/rsif.2014.0950. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2014.0950.

[3] Juliet Bedford et al. "A new twenty-first century science for effective epidemic response". In: *Nature* 575.7781 (Nov. 2019), pp. 130–136. DOI: 10.1038/s41586-019-1717-y. URL: https://doi.org/10.1038/s41586-019-1717-y.

[4] Giovanni Lo Iacono et al. "Using modelling to disentangle the relative contributions of zoonotic and anthroponotic transmission: the case of Lassa fever". In: *PLOS Neglected Tropical Diseases* 9.1 (Jan. 2015), pp. 1–13. DOI: 10.1371/journal.pntd.0003398. URL: https://doi.org/10.1371/journal.pntd.0003398.

[5] Thomas C Quinn. "Global burden of the HIV pandemic". In: *The Lancet* 348.9020 (1996), pp. 99–106. ISSN: 0140-6736. DOI: https://doi.org/10.1016/S0140-6736(96)01029-X. URL: http://www.sciencedirect.com/science/article/pii/S014067369601029X.

[6] T. Alex Perkins et al. "Model-based projections of Zika virus infections in childbearing women in the Americas". In: *Nature Microbiology* 1.9 (2016), p. 16126. DOI: 10.1038/nmicrobiol.2016.126. URL: https://doi.org/10.1038/nmicrobiol.2016.126.

[7] Moritz U. G. Kraemer et al. "Spread of yellow fever virus outbreak in Angola and the Democratic Republic of the Congo 2015-2016: a modelling study". In: *The Lancet Infectious Diseases* 17.3 (Mar. 2017), pp. 330–338. DOI: 10.1016/S1473-3099(16)30513-8. URL: https://doi.org/10.1016/S1473-3099(16)30513-8.

[8] C. Jessica E. Metcalf and Justin Lessler. "Opportunities and challenges in modeling emerging infectious diseases". In: *Science* 357.6347 (2017), pp. 149–152. DOI: 10.1126/science.aam8335. URL: https://science.sciencemag.org/content/357/6347/149.

[9] J. O. Lloyd-Smith et al. "Superspreading and the effect of individual variation on disease emergence". In: *Nature* 438.7066 (Nov. 2005), pp. 355–359. ISSN: 1476-4687. DOI: 10.1038/nature04153. URL: https://doi.org/10.1038/nature04153.

[10] Amy Wesolowski et al. "Impact of human mobility on the emergence of dengue epidemics in Pakistan". In: *Proceedings of the National Academy of Sciences* 112.38 (2015), pp. 11887–11892. DOI: 10.1073/pnas.1504964112. URL: https://www.pnas.org/content/112/38/11887.

[11] M. U. G. Kraemer et al. "Utilizing general human movement models to predict the spread of emerging infectious diseases in resource poor settings". In: *Scientific Reports* 9.1 (2019), p. 5151. ISSN: 2045-2322. DOI: 10.1038/s41598-019-41192-3.

[12] Erin A. Mordecai et al. "Detecting the impact of temperature on transmission of Zika, dengue, and chikungunya using mechanistic models". In: *PLOS Neglected Tropical Diseases* 11.4 (Apr. 2017), pp. 1–18. DOI: 10.1371/journal.pntd.0005568. URL: https://doi.org/10.1371/journal.pntd.0005568.

[13] Birgit Nikolay et al. "Transmission of Nipah Virus — 14 Years of Investigations in Bangladesh". In: *New England Journal of Medicine* 380.19 (2019). PMID: 31067370, pp. 1804–1814. DOI: 10.1056/NEJMoa1805376. URL: https://doi.org/10.1056/NEJMoa1805376.

[14] Gytis Dudas et al. "MERS-CoV spillover at the camel-human interface". In: *eLife* 7 (Jan. 2018), e31257. ISSN: 2050-084X. DOI: 10.7554/eLife.31257. URL: https://doi.org/10.7554/eLife.31257.

[15] Katriona Shea et al. "Harnessing multiple models for outbreak management". In: *Science* 368.6491 (2020), pp. 577–579. DOI: 10.1126/science.abb9934. URL: https://science.sciencemag.org/content/368/6491/577.

[16] Michael A. Johansson et al. "An open challenge to advance probabilistic forecasting for dengue epidemics". In: *Proceedings of the National Academy of Sciences* 116.48 (2019), pp. 24268–24274. DOI: 10.1073/pnas.1909865116. URL: https://www.pnas.org/content/116/48/24268.

[17] Nicholas G. Reich et al. "Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the U.S." In: *PLOS Computational Biology* 15.11 (Nov. 2019), pp. 1–19. DOI: 10.1371/journal.pcbi.1007486. URL: https://doi.org/10.1371/journal.pcbi.1007486.

[18] Nicholas G. Reich et al. "A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States". In: *Proceedings of the National Academy of Sciences* 116.8 (2019), pp. 3146–3154. DOI: 10.1073/pnas.1812594116. URL: https://www.pnas.org/content/116/8/3146.

[19] Craig J. McGowan et al. "Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016". In: *Scientific Reports* 9.1 (Jan. 2019), p. 683. DOI: 10.1038/s41598-018-36361-9. URL: https://doi.org/10.1038/s41598-018-36361-9.

[20] Leah R. Johnson et al. "Phenomenological forecasting of disease incidence using heteroskedastic Gaussian processes: A dengue case study". In: *The Annals of Applied Statistics* 12.1 (2018), pp. 27–66. DOI: 10.1214/17-AOAS1090. URL: https://doi.org/10.1214/17-AOAS1090.

[21] Sara Y. Del Valle et al. "Summary results of the 2014-2015 DARPA Chikungunya challenge". In: *BMC Infectious Diseases* 18.1 (May 2018), p. 245. DOI: 10.1186/s12879-018-3124-7. URL: https://doi.org/10.1186/s12879-018-3124-7.

[22] Cécile Viboud et al. "The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt". In: *Epidemics* 22 (2018). The RAPIDD Ebola Forecasting Challenge, pp. 13–21. DOI: https://doi.org/10.1016/j.epidem.2017.08.002. URL: http://www.sciencedirect.com/science/article/pii/S1755436517301275.

[23] ZIKAVAT Collaboration et al. "Preliminary results of models to predict areas in the Americas with increased likelihood of Zika virus transmission in 2017". In: *bioRxiv* (2017). DOI: 10.1101/187591. URL: https://www.biorxiv.org/content/early/2017/09/29/187591.

[24] Katriona Shea et al. "COVID-19 reopening strategies at the county level in the face of uncertainty: Multiple Models for Outbreak Decision Support". In: *medRxiv* (2020). DOI: 10.1101/2020.11.03.20225409. URL: https://www.medrxiv.org/content/early/2020/11/05/2020.11.03.20225409.

[25] Dylan B. George et al. "Technology to advance infectious disease forecasting for outbreak management". In: *Nature Communications* 10.1 (Sept. 2019), p. 3932. ISSN: 2041-1723. DOI: 10.1038/s41467-019-11901-7. URL: https://doi.org/10.1038/s41467-019-11901-7.

[26] Evan L Ray et al. "Ensemble Forecasts of Coronavirus Disease 2019 (COVID-19) in the U.S." In: *medRxiv* (2020). DOI: 10.1101/2020.08.19.20177493. URL: https://www.medrxiv.org/content/early/2020/08/22/2020.08.19.20177493.

[27] Logan C. Brooks et al. "Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions". In: *PLOS Computational Biology* 14.6 (June 2018), pp. 1–29. DOI: 10.1371/journal.pcbi.1006134. URL: https://doi.org/10.1371/journal.pcbi.1006134.

[28] G. Chowell et al. "Real-time forecasting of epidemic trajectories using computational dynamic ensembles". In: *Epidemics* 30 (2020), p. 100379. ISSN: 1755-4365. DOI: https://doi.org/10.1016/j.epidem.2019.100379. URL: http://www.sciencedirect.com/science/article/pii/S1755436519301112.

[29] Amir S. Siraj et al. "Spatiotemporal incidence of Zika and associated environmental drivers for the 2015-2016 epidemic in Colombia". In: *Scientific Data* 5.1 (2018), p. 180073. DOI: 10.1038/sdata.2018.73. URL: https://doi.org/10.1038/sdata.2018.73.

[30] Moritz U. G. Kraemer et al. "The effect of human mobility and control measures on the COVID-19 epidemic in China". In: *Science* 368.6490 (2020), pp. 493–497. DOI: 10.1126/science.abb4218. URL: https://science.sciencemag.org/content/368/6490/493.

[31] Allison Black et al. "Genomic epidemiology supports multiple introductions and cryptic transmission of Zika virus in Colombia". In: *BMC Infectious Diseases* 19.1 (2019), p. 963. DOI: 10.1186/s12879-019-4566-2. URL: https://doi.org/10.1186/s12879-019-4566-2.

[32] Roni Rosenfeld. *The EM Algorithm*. 1997. URL: http://www.cs.cmu.edu/afs/cs.cmu.edu/academic/class/11761-s97/WWW/tex/EM.ps.

[33] Michael C. Dietze. *Ecological Forecasting*. Princeton University Press, 2017. ISBN: 9780691160573. URL: http://www.jstor.org/stable/j.ctvc7796h.

[34] Nicholas B. DeFelice et al. "Ensemble forecast of human West Nile virus cases and mosquito infection rates". In: *Nature Communications* 8.1 (Feb. 2017), p. 14592. ISSN: 2041-1723. DOI: 10.1038/ncomms14592. URL: https://doi.org/10.1038/ncomms14592.

[35] Wan Yang, Alicia Karspeck, and Jeffrey Shaman. "Comparison of Filtering Methods for the Modeling and Retrospective Forecasting of Influenza Epidemics". In: *PLOS Computational Biology* 10.4 (Apr. 2014), pp. 1–15. DOI: 10.1371/journal.pcbi.1003583. URL: https://doi.org/10.1371/journal.pcbi.1003583.

[36] Rachel E. Baker et al. "Susceptible supply limits the role of climate in the early SARS-CoV-2 pandemic". In: *Science* 369.6501 (2020), pp. 315–319. DOI: 10.1126/science.abc2535. URL: https://science.sciencemag.org/content/369/6501/315.

[37] S. Cauchemez et al. "Local and regional spread of chikungunya fever in the Americas". In: *Euro surveillance* 19.28 (2014), pp. 20854–20854. ISSN: 1560-7917. DOI: 10.2807/1560-7917.es2014.19.28.20854. URL: https://pubmed.ncbi.nlm.nih.gov/25060573.

[38] Michael A. Johansson et al. "Nowcasting the Spread of Chikungunya Virus in the Americas". In: *PLOS ONE* 9.8 (Aug. 2014), pp. 1–8. DOI: 10.1371/journal.pone.0104915. URL: https://doi.org/10.1371/journal.pone.0104915.

[39] Sean M. Moore et al. "Local and regional dynamics of chikungunya virus transmission in Colombia: the role of mismatched spatial heterogeneity". In: *BMC Medicine* 16.1 (Aug. 2018), p. 152. ISSN: 1741-7015. DOI: 10.1186/s12916-018-1127-2. URL: https://doi.org/10.1186/s12916-018-1127-2.

[40] Shengjie Lai et al. "Seasonal and interannual risks of dengue introduction from South-East Asia into China, 2005-2015". In: *PLOS Neglected Tropical Diseases* 12.11 (Nov. 2018), pp. 1–16. DOI: 10.1371/journal.pntd.0006743. URL: https://doi.org/10.1371/journal.pntd.0006743.

[41] Tom Lindström, Michael Tildesley, and Colleen Webb. "A Bayesian ensemble approach for epidemiological projections". In: *PLOS Computational Biology* 11.4 (Apr. 2015), pp. 1–30. DOI: 10.1371/journal.pcbi.1004187. URL: https://doi.org/10.1371/journal.pcbi.1004187.

[42] Teresa K. Yamana, Sasikiran Kandula, and Jeffrey Shaman. "Superensemble forecasts of dengue outbreaks". In: *Journal of The Royal Society Interface* 13.123 (2016), p. 20160410. DOI: 10.1098/rsif.2016.0410. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2016.0410.

[43] Thomas McAndrew and Nicholas G. Reich. *Adaptively stacking ensembles for influenza forecasting with incomplete data*. 2020. arXiv: 1908.01675 [stat.AP].

[44] Evan L. Ray et al. "Challenges in training ensembles to forecast COVID-19 cases and deaths in the United States". In: *International Institute of Forecasters* (2021). URL: https://forecasters.org/blog/2021/04/09/challenges-in-training-ensembles-to-forecast-covid-19-cases-and-deaths-in-the-united-states/.

[45] T. Alex Perkins et al. "Estimating unobserved SARS-CoV-2 infections in the United States". In: *Proceedings of the National Academy of Sciences* 117.36 (2020), pp. 22597–22602. ISSN: 0027-8424. DOI: 10.1073/pnas.2005476117. eprint: https://www.pnas.org/content/117/36/22597.full.pdf.

[46] Thomas McAndrew et al. "Aggregating predictions from experts: A review of statistical methods, experiments, and applications". In: *WIREs Computational Statistics* (), e1514. DOI: 10.1002/wics.1514. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.1514.

[47] Korryn Bodner, Marie-Josée Fortin, and Péter K. Molnár. "Making predictive modelling ART: accurate, reliable, and transparent". In: *Ecosphere* 11.6 (2020), e03160. DOI: 10.1002/ecs2.3160. URL: https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1002/ecs2.3160.

[48] Shou-Li Li et al. "Essential information: Uncertainty and optimal control of Ebola outbreaks". In: *Proceedings of the National Academy of Sciences* 114.22 (2017), pp. 5659–5664. ISSN: 0027-8424. DOI: 10.1073/pnas.1617482114. eprint: https://www.pnas.org/content/114/22/5659.full.pdf. URL: https://www.pnas.org/content/114/22/5659.

[49] Sen Pei and Jeffrey Shaman. "Counteracting structural errors in ensemble forecast of influenza outbreaks". In: *Nature Communications* 8.1 (Oct. 2017), p. 925. ISSN: 2041-1723. DOI: 10.1038/s41467-017-01033-1. URL: https://doi.org/10.1038/s41467-017-01033-1.

[50] Antoine Allard et al. "The risk of sustained sexual transmission of Zika is underestimated". In: *PLOS Pathogens* 13.9 (Sept. 2017), pp. 1–12. DOI: 10.1371/journal.ppat.1006633. URL: https://doi.org/10.1371/journal.ppat.1006633.

[51] Axel Bonačić Marinović et al. "Quantifying Reporting Timeliness to Improve Outbreak Control". In: *Emerging Infectious Disease journal* 21.2 (2015), p. 209. DOI: 10.3201/eid2102.130504. URL: https://doi.org/10.3201/eid2102.130504.

[52] Matt J. Keeling and Pejman Rohani. *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press, 2008. ISBN: 9780691116174. URL: http://www.jstor.org/stable/j.ctvcm4gk0.

[53] LT Figueiredo, SM Cavalcante, and MC Simões. "Dengue serologic survey of schoolchildren in Rio de Janeiro, Brazil, in 1986 and 1987". In: *Bull Pan Am Health Organ.* 24.2 (1990), pp. 217–225. URL: https://iris.paho.org/bitstream/handle/10665.2/27164/ev24n2p217.pdf?sequence=1.

[54] Jue Tao Lim et al. "Time varying methods to infer extremes in dengue transmission dynamics". In: *PLOS Computational Biology* 16.10 (Oct. 2020), pp. 1–19. DOI: 10.1371/journal.pcbi.1008279. URL: https://doi.org/10.1371/journal.pcbi.1008279.

[55] Honglei Sun et al. "Prevalent Eurasian avian-like H1N1 swine influenza virus with 2009 pandemic viral genes facilitating human infection". In: *Proceedings of the National Academy of Sciences* (2020). ISSN: 0027-8424. DOI: 10.1073/pnas.1921186117. eprint: https://www.pnas.org/content/early/2020/06/23/1921186117.full.pdf. URL: https://www.pnas.org/content/early/2020/06/23/1921186117.

[56] Caroline O. Buckee et al. "Aggregated mobility data could help fight COVID-19". In: *Science* 368.6487 (2020), pp. 145–146. DOI: 10.1126/science.abb8021. URL: https://science.sciencemag.org/content/368/6487/145.2.

[57] Isaac I. Bogoch et al. "Potential for Zika virus introduction and transmission in resource-limited countries in Africa and the Asia-Pacific region: a modelling study". eng. In: *The Lancet. Infectious diseases* 16.11 (Nov. 2016), pp. 1237–1245. DOI: 10.1016/S1473-3099(16)30270-5. URL: https://doi.org/10.1016/S1473-3099(16)30270-5.

[58] Moritz UG Kraemer et al. "The global distribution of the arbovirus vectors *Aedes aegypti* and *Ae. albopictus*". In: *eLife* 4 (June 2015), e08347. ISSN: 2050-084X. DOI: 10.7554/eLife.08347. URL: https://doi.org/10.7554/eLife.08347.

[59] Amy Wesolowski et al. "Heterogeneous mobile phone ownership and usage patterns in Kenya". In: *PLOS ONE* 7.4 (Apr. 2012), pp. 1–6. DOI: 10.1371/journal.pone.0035319. URL: https://doi.org/10.1371/journal.pone.0035319.

[60] Yingcun Xia, Ottar N. Bjørnstad, and Bryan T. Grenfell. "Measles metapopulation dynamics: A gravity model for epidemiological coupling and dynamics". English (US). In: *American Naturalist* 164.2 (Aug. 2004), pp. 267–281. ISSN: 0003-0147. DOI: 10.1086/422341.

[61] Rachel J. Oidtman et al. "Inter-annual variation in seasonal dengue epidemics driven by multiple interacting factors in Guangzhou, China". In: *Nature Communications* 10.1 (Mar. 2019). DOI: 10.1038/s41467-019-09035-x. URL: https://doi.org/10.1038/s41467-019-09035-x.

[62] T.A. Perkins et al. "Estimating drivers of autochthonous transmission of chikungunya virus in its invasion of the Americas". In: *PLOS Currents Outbreaks* (2017). DOI: 10.1371/currents.outbreaks.a4c7b6ac10e0420b1788c9767946d1fc.

[63] Amir S. Siraj et al. "Temperature modulates dengue virus epidemic growth rates through its effects on reproduction numbers and generation intervals". In: *PLOS Neglected Tropical Diseases* 11.7 (July 2017), pp. 1–19. DOI: 10.1371/journal.pntd.0005797. URL: https://doi.org/10.1371/journal.pntd.0005797.

[64] Filippo Simini et al. "A universal model for mobility and migration patterns". In: *Nature* 484.7392 (2012), pp. 96–100. ISSN: 1476-4687. DOI: 10.1038/nature10856. URL: https://doi.org/10.1038/nature10856.

[65] William D. Nordhaus. "Geography and macroeconomics: New data and new findings". In: *Proceedings of the National Academy of Sciences* 103.10 (2006), pp. 3510–3517. ISSN: 0027-8424. DOI: 10.1073/pnas.0509842103. eprint: https://www.pnas.org/content/103/10/3510.full.pdf. URL: https://www.pnas.org/content/103/10/3510.

[66] M. Sanjeev Arulampalam et al. "A tutorial on particle filters for online nonlinear/non-gaussian Bayesian tracking". In: *IEEE Transactions on Signal Processing* 50.4 (Feb. 2002), pp. 174–188.

[67] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian Raftery. "Probabilistic forecasts, calibration and sharpness". In: *Journal of the Royal Statistical Society: Series B* 69 (2007), pp. 243–268. URL: https://doi.org/10.1371/journal.pcbi.1004187.

[68] Sen Pei et al. "Forecasting the spatial transmission of influenza in the United States". In: *Proceedings of the National Academy of Sciences* 115.11 (2018), pp. 2752–2757. DOI: 10.1073/pnas.1708856115. URL: https://www.pnas.org/content/115/11/2752.

Supplemental methods

Priors on parameters common to all models

In each model that we considered, we iteratively estimated the reporting probability (ρ), dispersion parameter of the negative binomial distribution (ϕ), R multiplier (k), and the timing of the first infection (ι). When possible, we leveraged previous estimates of parameters to inform prior distributions for model parameters. In some instances, we used dengue-specific parameter estimates as priors for Zika-specific parameters [1].

For the reporting probability (ρ), we assumed a mean of 0.2 and a variance of 0.05. Although this mean and variance are not directly informed by an empirical study of Zika reporting, they are in line with what we would expect for dengue [2, 3] and Zika [4]. We assumed that ρ was a beta random variable and, using the method of moments, we specified a prior distribution for ρ such that

$$\rho \sim \text{beta}(0.44, 1.76) \tag{10}$$

which resulted in mean and variance consistent with our prior assumptions.

The dispersion parameter of the negative binomial distribution accounts for variability in case reporting beyond that captured by ρ. Lower values of the dispersion parameter indicate overdispersion, such that variability in cases cannot be explained by a single rate of case incidence, as would be generated by a Poisson distribution with rate $\rho I_{d,t}$. Given the likelihood of variation in the reporting probability over the course of the epidemic [5] and across departments [6], we specified a uniform prior for ϕ,

$$\phi \sim \text{uniform}(0, 1), \tag{11}$$

which resulted in a level of overdispersion in reporting equal to at least a geometric distribution ($\phi = 1$) but potentially greater ($\phi < 1$).

To relate the transmission coefficient ($\beta_{d,t}$) to environmentally driven descriptions of the reproduction number ($R_{d,t}$), we used a multiplier (k) that applied to both the static and dynamic models of R. We specified a gamma prior distribution,

$$k \sim \text{gamma}(2.25, 0.75), \tag{12}$$

the parameters of which were chosen by moment matching to result in a mean of three and a variance of two.

To seed the Zika epidemic in Colombia, we assumed that undetected transmission could have been occurring at any time throughout the first 34 weeks of 2015. For reference, the first case was not reported until August 9, 2015 (week 35). Thus, we specified a uniform prior for the date of the initial introduction (ι_1) between weeks 1 and 34 of 2015. We assumed that the location of the first introduction (l_1) could have been any of the departments in Colombia with equal prior probability.

Assumptions about human mobility

Spatially coupled models

For the spatially coupled models, we assumed that transmission was coupled across departments by human mobility. In these models, we informed the spatial connectivity matrix in three ways: i) aggregated mobility patterns extracted from mobile phone call detail records (CDRs), ii) a gravity model, or iii) a radiation model. In

the gravity and radiation models, $T_{i,j}$ is defined as the total number of trips from department i to department j. This takes the form $T_{i,j} = c\frac{N_i^\alpha N_j^\beta}{d_{i,j}^\gamma}$ in the gravity model and $T_{i,j} = T_i \frac{N_i N_j}{(N_i+s_{i,j})(N_i+N_j+s_{i,j})}$ in the radiation model, where N_i and N_j are population sizes at the origin i and destination j, $d_{i,j}^\gamma$ is the distance between i and j, $s_{i,j}$ is the total population within radius $d_{i,j}$ from i, and T_i is the total number of individuals who make a trip. The parameters c, α, β, and γ were fitted to CDRs from Spain and validated in West Africa [7]. All three mobility models were row-normalized to correspond to the proportion of time spent by residents of department i in department j.

Spatially uncoupled models

For the spatially uncoupled models, we assumed that each department's epidemic occurred independently of all other departments. We used the same prior distributions as described above for ρ, ϕ, k for each department. Under this assumption, we did not include a parameter for the location of the initial introduction into Colombia, as we instead were concerned with the initial introduction into that department. It was still necessary to specify the timing of that introduction and, for models that considered it, the timing of a second introduction. Following the rationale that undetected transmission could have occurred for up to 34 weeks prior to the first reported case in a given department, we assumed an even prior for the timing of the introduction(s) into a given department.

Assumptions about environmental drivers of transmission

Dynamic model

The environmentally driven component of $\beta_{d,t}$ for the dynamic model, $\tilde{R}_{d,t}(T_{d,t}, OP_{d,t}|c, \psi, \alpha, v)$, was defined as the product of two functions: one that depends on $T_{d,t}$ and one that depends on $OP_{d,t}$.

The function of $OP_{d,t}$ was defined as $c(-\log(1 - OP_{d,t}))$, which converts occurrence probability into an expectation of mosquito abundance [8]. The constant c scales this expectation to account for uncertainty in its magnitude, which we estimated and assumed to have a gamma-distributed prior with a shape parameter of 16 and a rate parameter of 0.07. This choice of prior resulted in average $\tilde{R}_{d,t}$ being equal to one when evaluated at the mean values of the prior for c.

The function of $T_{d,t}$ was based on a temperature-dependent description of the basic reproduction number by Mordecai et al. [9]. Specifically, we used the version of that model based on Briere functions for parameters that were not otherwise accounted for in estimates of mosquito occurrence probability [10]. Those parameters include the mosquito biting rate (a), mosquito-to-human transmission probability (b), human-to-mosquito transmission probability (c), average adult mosquito lifespan (lf), and parasite development rate (PDR). Those functions combine to form the temperature-dependent component of $\tilde{R}_{d,t}$,

$$a(T_{d,t})b(T_{d,t})c(T_{d,t})e^{-1/(lf(T_{d,t})PDR(T_{d,t}))}PDR(T_{d,t}). \quad (13)$$

We did not include other parameters from Mordecai et al. [9] related to mosquitoes, such as immature development rate and egg-to-adult survival rate, as the effects of those parameters were accounted for in estimates of $OP_{d,t}$ from Kraemer et al. [10].

To reduce the number of parameters that we needed to estimate, we worked with a simplified description of the temperature curves produced by eq. (13) (Fig. S5). To do so, we took random draws from the posterior distribution of parameters from Mordecai et al. [9], computed functions of temperature according to eq. (13), and fitted parameters of a skew normal distribution to those curves by least squares. The skew normal distribution has three parameters—location (ψ), scale (α), shape (υ)—that together control the mean, variance, and skew of the distribution, which is sufficient to approximate the uncertainty in posterior predictions of the temperature curves described by eq. (13). We then took the mean and covariance across those estimates of ψ, α, and υ to describe their variation with a multivariate normal distribution, which was the prior distribution we used for those parameters at the first step of our particle filter.

Static model

The environmentally driven component of β_d for the dynamic model, $\bar{R}_d(T_{d,\vec{t}}, OP_d, PPP_d)$, was defined as the product of three functions: one that depends on $T_{d,\vec{t}}$, one that depends on OP_d, and one that depends on PPP_d. We used values of $\bar{R}_d$ at the 5 km x 5 km grid cell level as calculated by Perkins et al. [8]. To aggregate them to the department level, we took a population-weighted mean of $\bar{R}_d$ across 5 km x 5 km grid cells within each department. Although we made no other modifications to the calculation of $\bar{R}_d$, we summarize the methodology used by Perkins et al. [8] in the interest of comparability with the dynamic model.

The function of OP_d was defined as $-\log(1 - OP_d))$, which was the same procedure used to convert occurrence probability into an expectation of mosquito abundance as used in the dynamic model. For the static model, however, we followed Perkins et al. [8] and relied on a description of occurrence probability that was not defined on a time-varying basis [10].

Rather than estimate a scaling parameter like c in the dynamic model, we relied on a scaling parameter defined as a function of PPP_d by Perkins et al. [8]. This function took the form of a monotonically decreasing, cubic spline function estimated with a shape-constrained additive model. The data that informed this estimate of both the relationship with PPP_d and the magnitude of $\bar{R}_d$ were outbreak sizes of 12 chikungunya outbreaks and one Zika outbreak [8] that occurred prior to the ZIKV invasion of the Americas.

The function of $T_{d,\vec{t}}$ was based on a temperature-dependent description of the basic reproduction number that includes temperature-dependent descriptions of mosquito mortality (μ) [11] and the extrinsic incubation period (n) [12]. Those functions contribute to an expression for monthly contributions to $R_{d,t}$,

$$\frac{bca^2 e^{-\mu(T_{d,t})n(T_{d,t})}}{\mu(T_{d,t})r}, \tag{14}$$

along with constants that represent the mosquito-to-human transmission probability (b), human-to-mosquito transmission probability (c), mosquito biting rate (a), and rate of recovery from human infection (r). For each department d, we took the average of the six largest values of eq. (14) as the contribution of $T_{d,\vec{t}}$ to $\bar{R}_d$.

Assumptions about introduction events

One-introduction models

Each of the one-introduction models assumed infections were seeded at one point in time (ι_1) in one location (l_1), as specified in the general model parameters section of the Supplementary Methods.

Two-introduction models

Each of the two-introduction models made similar assumptions about ι_1 and l_1 for the first introduction. For these models, we additionally specified the timing (ι_2) and location (l_2) of the second introduction events, for which we assumed even priors. In reality, there were likely more than only one or two introductions throughout the epidemic, but genomic epidemiology suggests the majority of local transmission resulted from only one or two introductions [13].

Supplemental tables

Symbol	Definition	Class
ρ	Reporting probability	Estimated parameter
ϕ	Overdispersion in reporting	Estimated parameter
k	$R_{d,t}$ multiplier	Estimated parameter
ι_1	Timing of first infection	Estimated parameter
ι_2	Timing of second infection	Estimated parameter
l_1	Location of first infection	Estimated parameter
l_2	Location of second infection	Estimated parameter
c	Dynamic R scalar	Estimated parameter
ψ	Location for skew normal	Estimated parameter
α	Scale for skew normal	Estimated parameter
υ	Shape for skew normal	Estimated parameter
θ	Parameter set	Set of estimated parameters
$I_{d,t}$	Simulated infections	State variable
$I'_{d,t}$	Redistributed infections	State variable
$I''_{d,t}$	Effective number of infections	State variable
$C_{d,t}$	Simulated cases	State variable
$y_{d,t}$	Observed data	Data
d	Department index	Data
t	Time; $t = 0, ..., T$	Data
n	Particle index; $n = 1, ..., N$	Data
η	Timing of first reported case	Data

Table S1: Mathematical symbols for parameters, state variables, and data with their respective meanings.

Supplemental figures

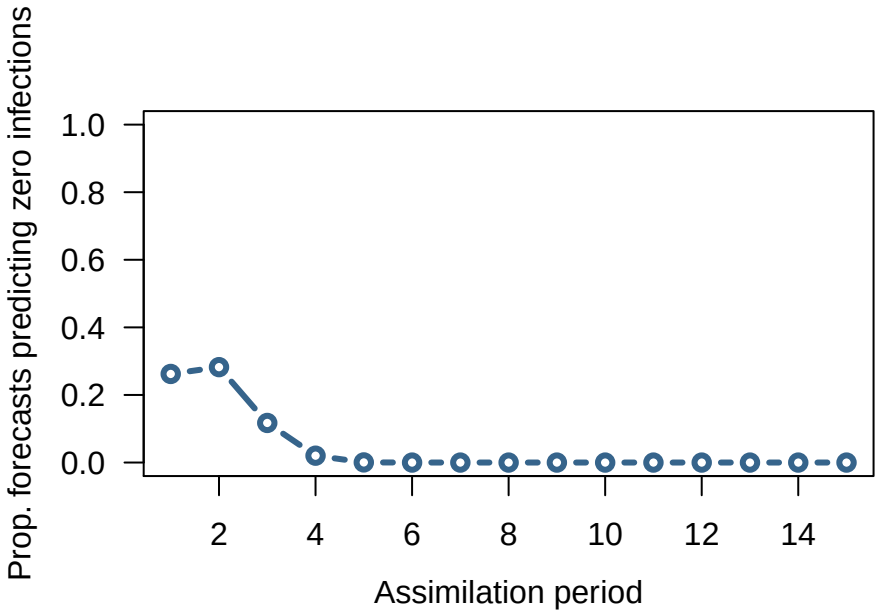

Figure S1: **Proportion of forecast trajectories predicting zero infections.** Forecast trajectories are specific to each particle and this example is from the model using CDR-derived mobility data, static R, and one importation event, corresponding to Fig. S2.

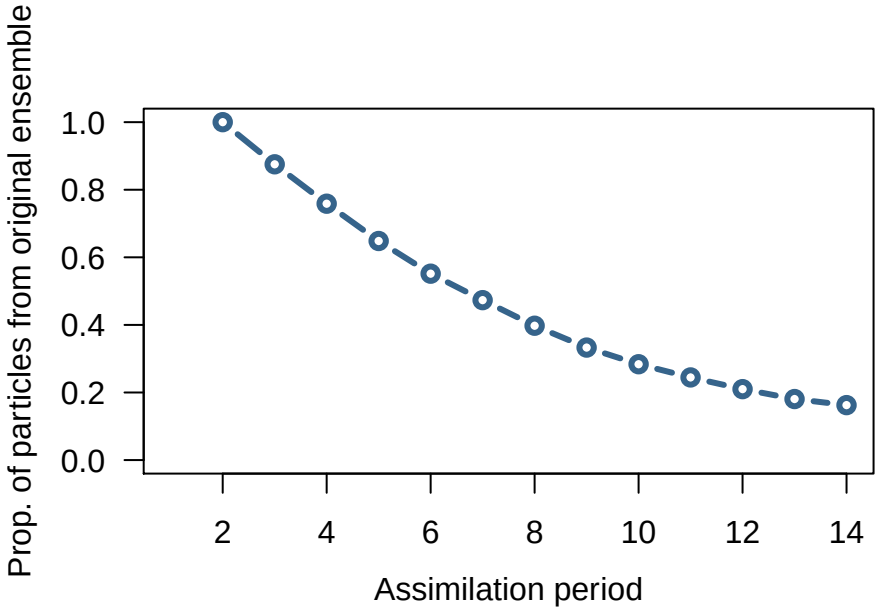

Figure S2: **Proportion of particles that remain from the original particle ensemble present in the retained ensemble at each assimilation period.** This example is from the model using CDR-derived mobility data, static R, and one importation event, corresponding to Fig. S1.

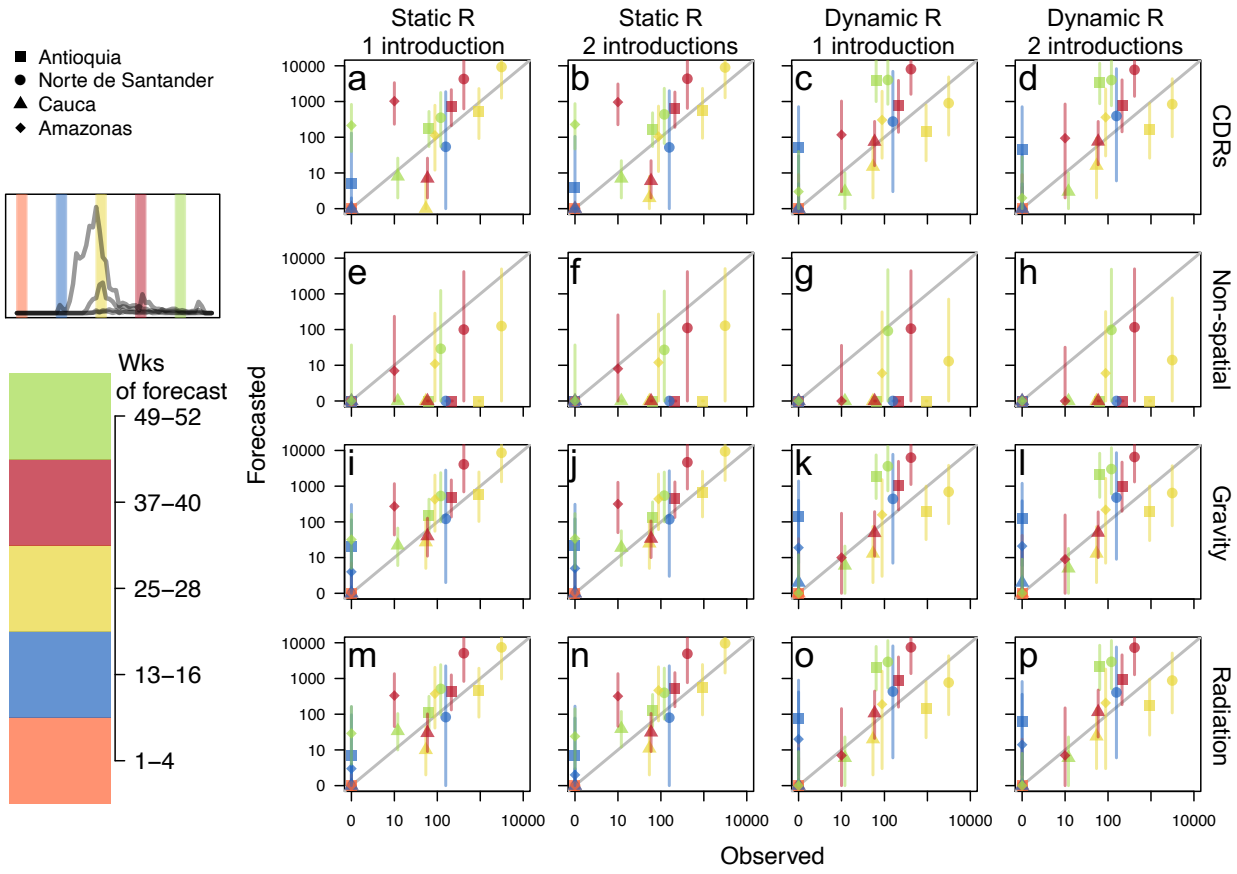

Figure S3: **Cumulative observed versus forecasted incidence at 4-week ahead intervals for each individual model for Antioquia, Norte de Santander, Cauca, and Amazonas at five points throughout the epidemic.** *a-p.* Each plot represents a different model, with model features labels on rows and columns. Plotted departments reflect differences in population, epidemic size, and geographic regions of Colombia and are denoted by point type. Point shape denotes department. Point color indicates time at which forecasts were generated and is visually denoted in inset plot and color bar. 1-, 2-, 3-, and 4-week ahead forecasts and observed incidence were aggregated for ease of comparison. Points are the median values and lines are the 50% credible interval. 1:1 line is in grey.

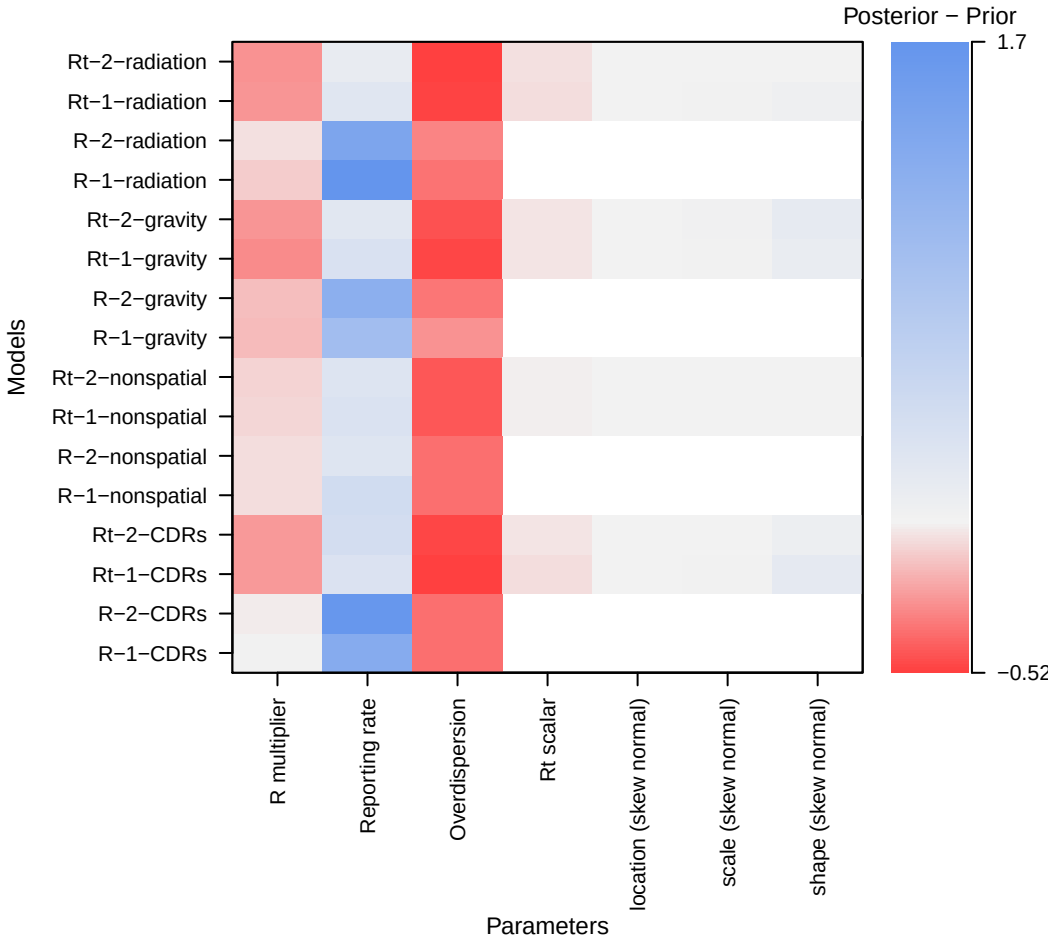

Figure S4: **Relative difference between the prior estimates of the parameters and the posterior estimates at the final time point in the epidemic.** Parameters include the R multiplier (k), reporting rate (ρ), overdispersion parameter (ϕ), R_t scalar (c), and the location (ψ), shape (α), and scale (ν) parameters for the skew normal distribution. Blue indicates posterior estimates were higher than prior estimates, red indicates posterior estimates were lower than prior estimates, and grey indicates no difference. Areas in white indicate the corresponding model does not use that parameter. To calculate the relative difference, we subtracted the prior estimates of parameters from the posterior estimates of parameters and divided the difference by the prior to standardize over different parameter magnitudes. For comparison purposes, we left out the initial timing and initial location of ZIKV introduction parameters.

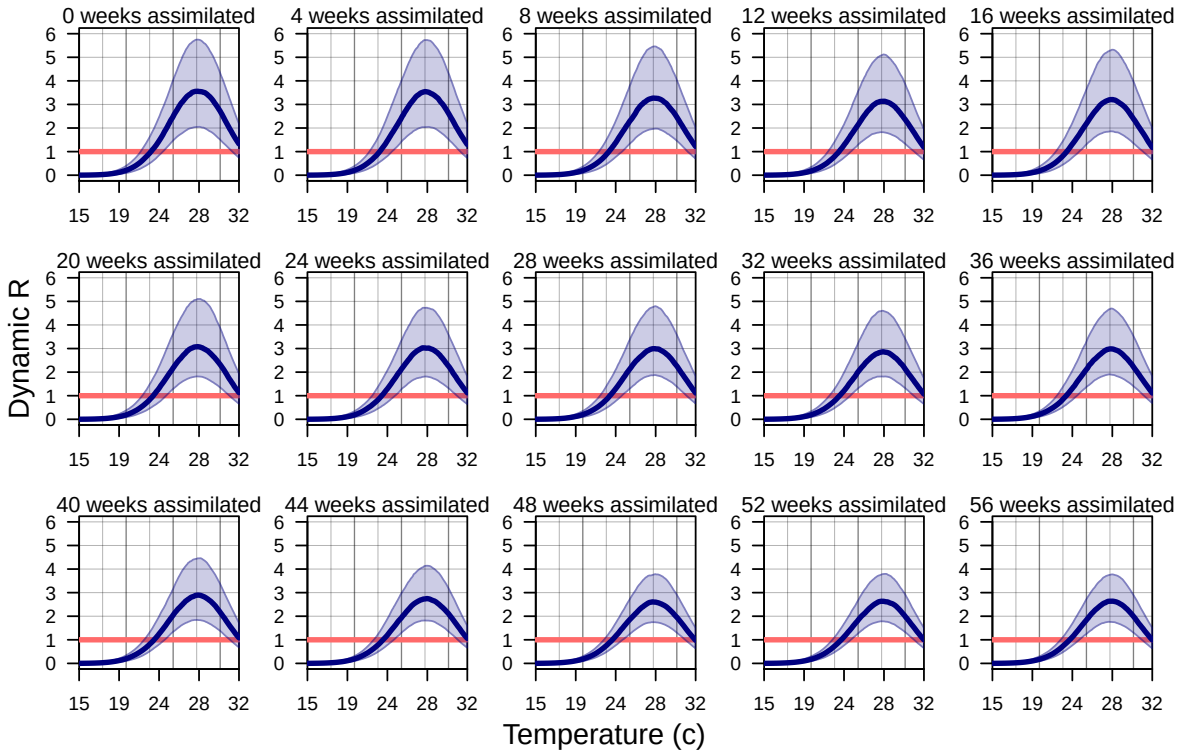

Figure S5: **Estimates of R at each assimilation week across temperatures for average mosquito occurrence probability in Colombia.** Blue bands indicate 95% credible interval, with think navy line indicating the median estimate of R. Horizontal red line indicates $R = 1$.

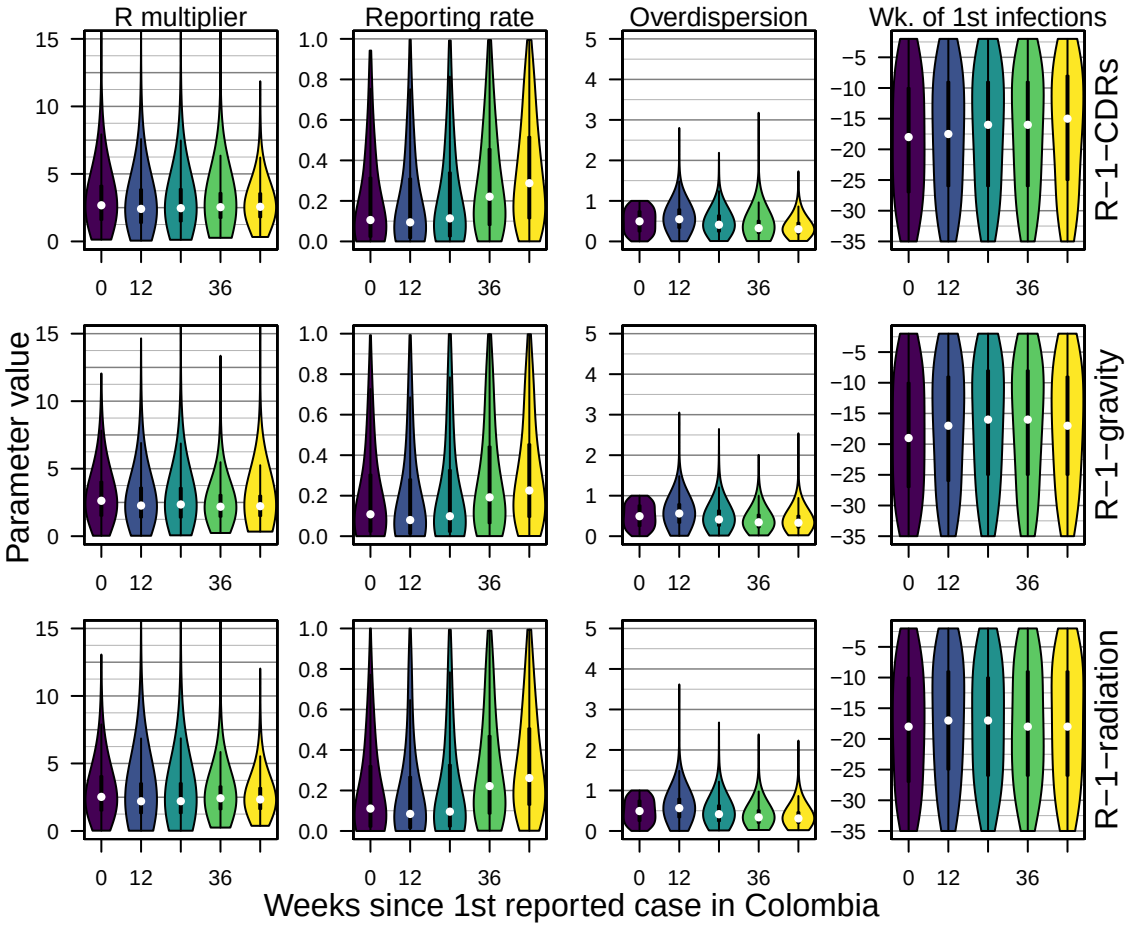

Figure S6: **Posterior distributions of model parameters for the spatially coupled models with a static R at five different points in time.** The R multiplier (k), reporting rate (ρ), the overdispersion parameter (ϕ), and the timing of the first importations (ι_1) were model parameters represented in each model, although we are showing the posterior distribution only for a subset of the models. Violin plots are colored by time at which the forecasts were made, and correspond to time points in Fig. S3 and Fig. S23, S10

.

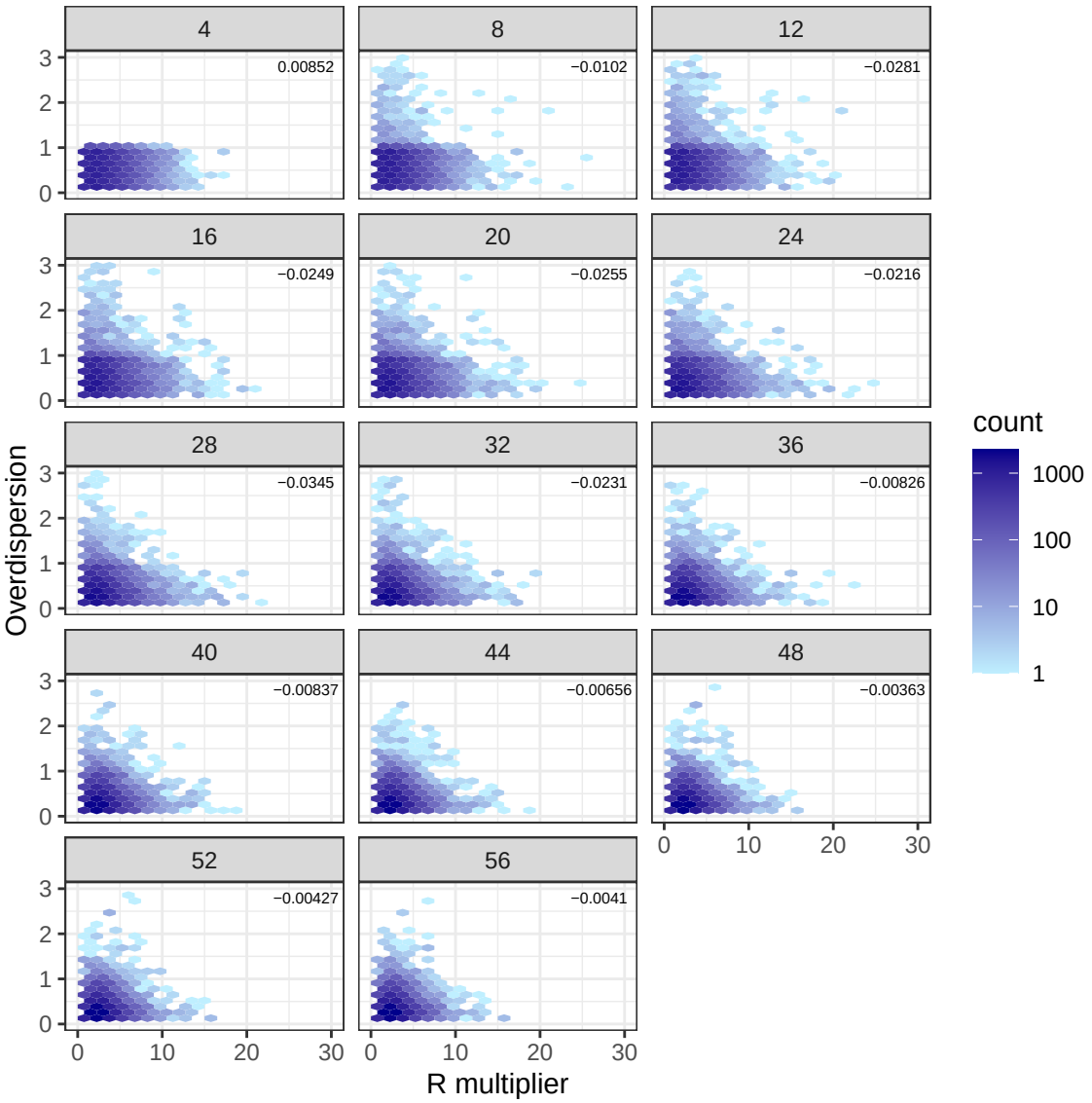

Figure S7: R_d **multiplier versus the overdispersion parameter in the negative binomial distribution.** The overdispersion parameter in the negative binomial distribution represents the variability in the reporting probability. Pearson's correlation coefficient for the two parameters within a parameter set is listed in the top right corner. This example is from the model using CDR-derived mobility data, static R, and one importation event, corresponding to Fig. S1.

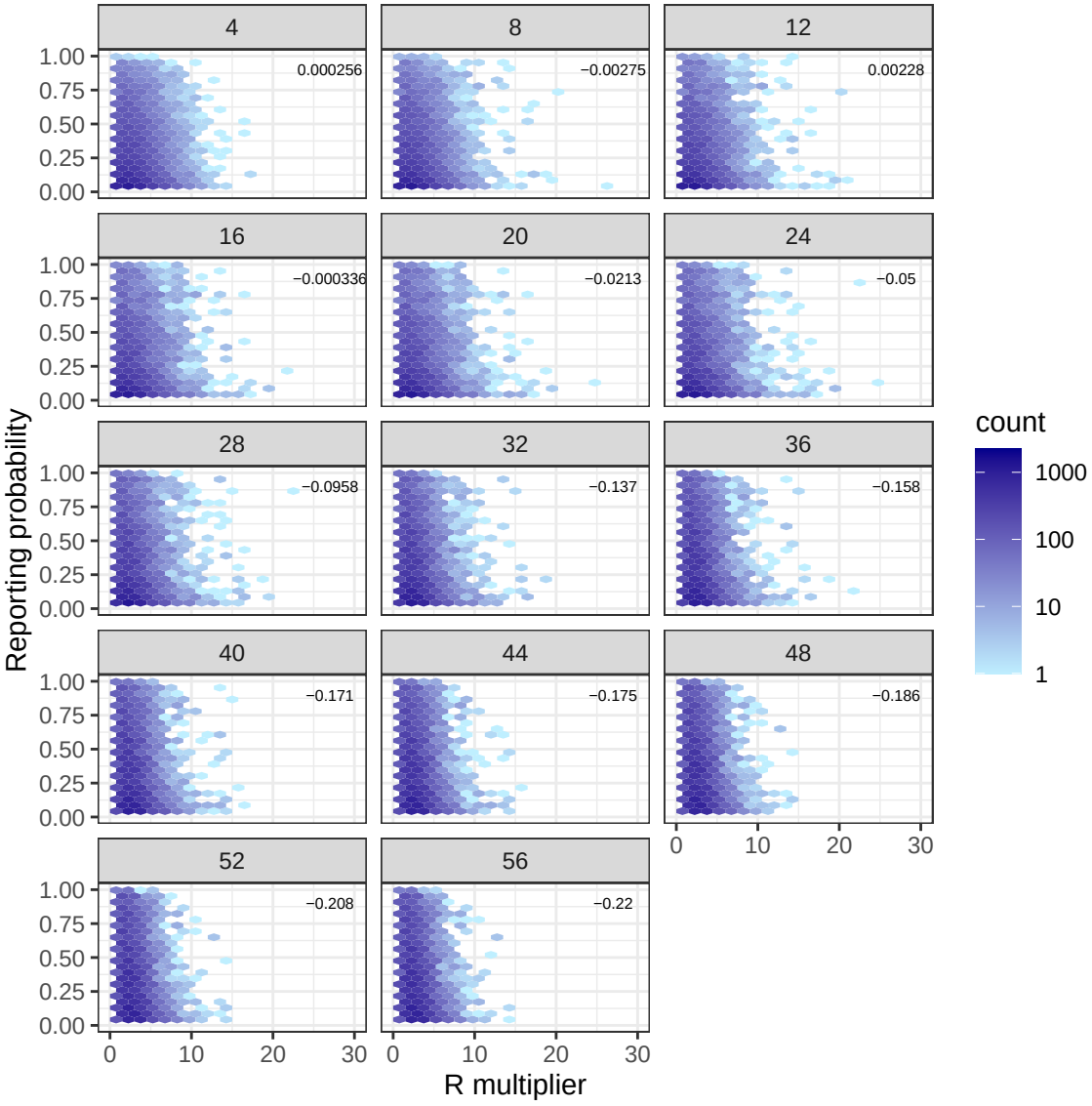

Figure S8: R_d **multiplier versus the reporting probability.** Pearson's correlation coefficient for the two parameters within a parameter set is listed in the top right corner. This example is from the model using CDR-derived mobility data, static R, and one importation event, corresponding to Fig. S1.

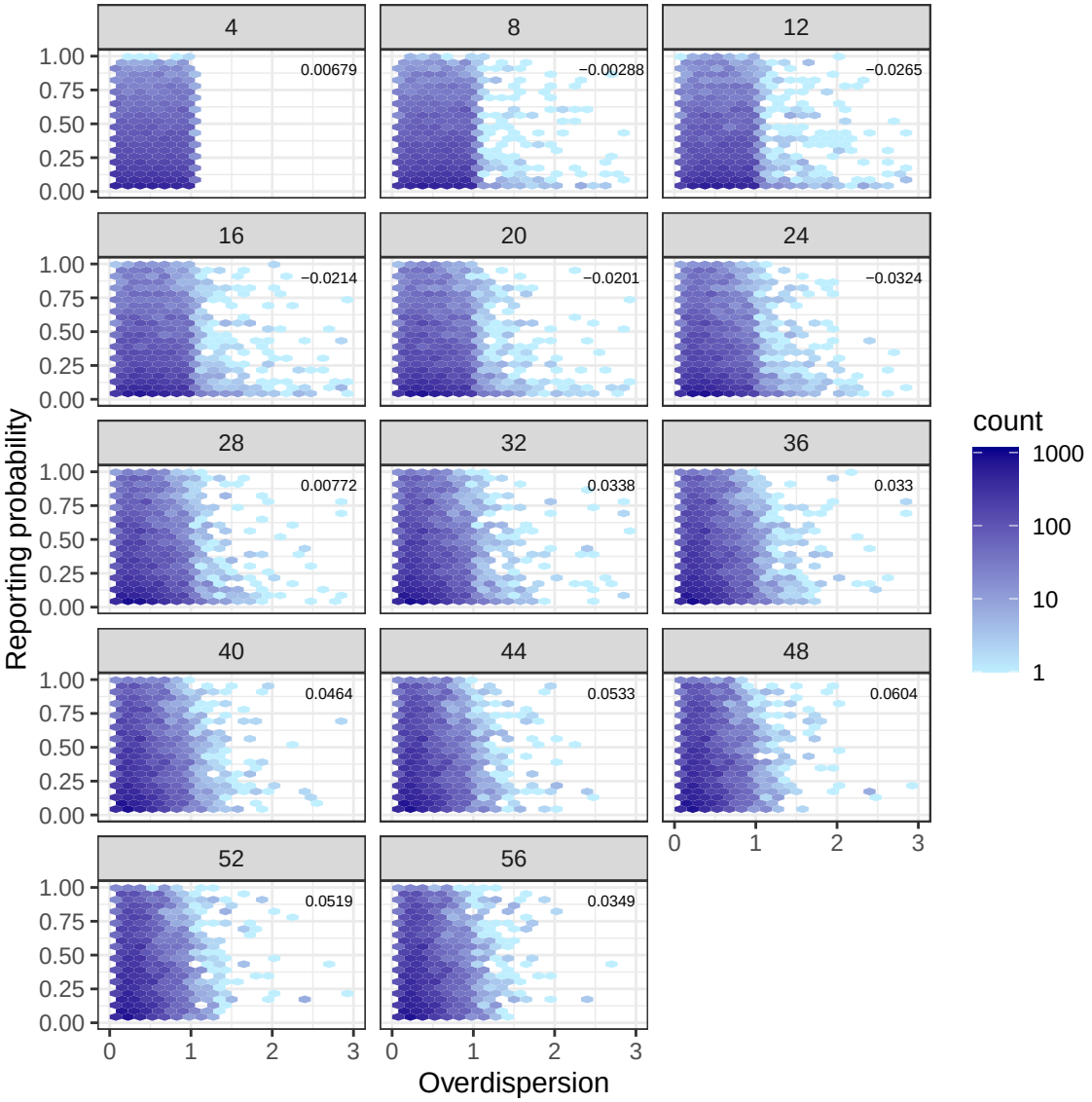

Figure S9: **The negative binomial overdispersion parameter versus the reporting probability for one example model at each assimilation period.** The overdispersion parameter in the negative binomial distribution represents the variability in the reporting probability. Pearson's correlation coefficient for the two parameters within a parameter set is listed in the top right corner. This example is from the model using CDR-derived mobility data, static R, and one importation event, corresponding to Fig. S1.

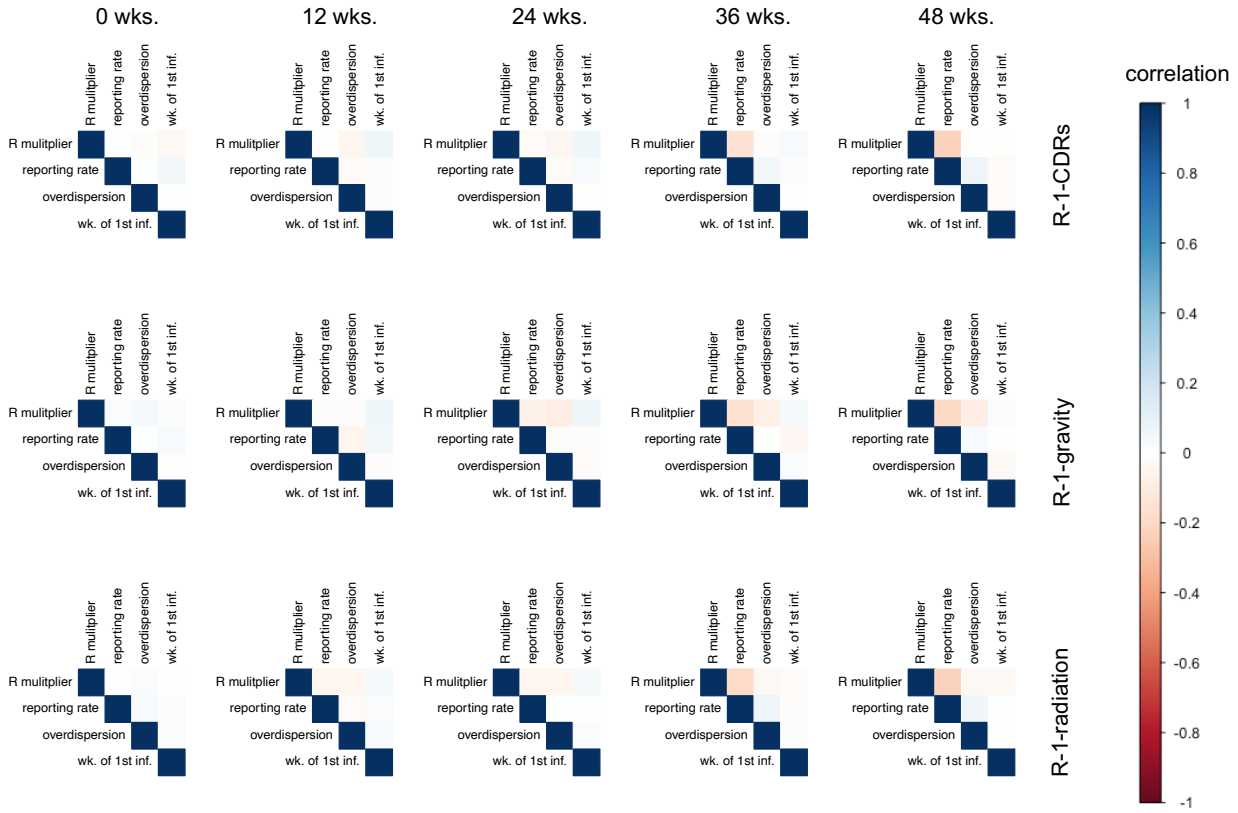

Figure S10: **Correlation of model parameters across within a parameter set through time for the spatially coupled models with a static R at five different points in time.** Parameters include, the R multiplier (k), reporting rate (ρ), the overdispersion parameter (ϕ), and the timing of the first importations (ι_1) as they were represented in each model, although we are showing the correlations only for a subset of the models. Time points shown here correspond to those time points in Fig. S3 and Fig. S23, S6.

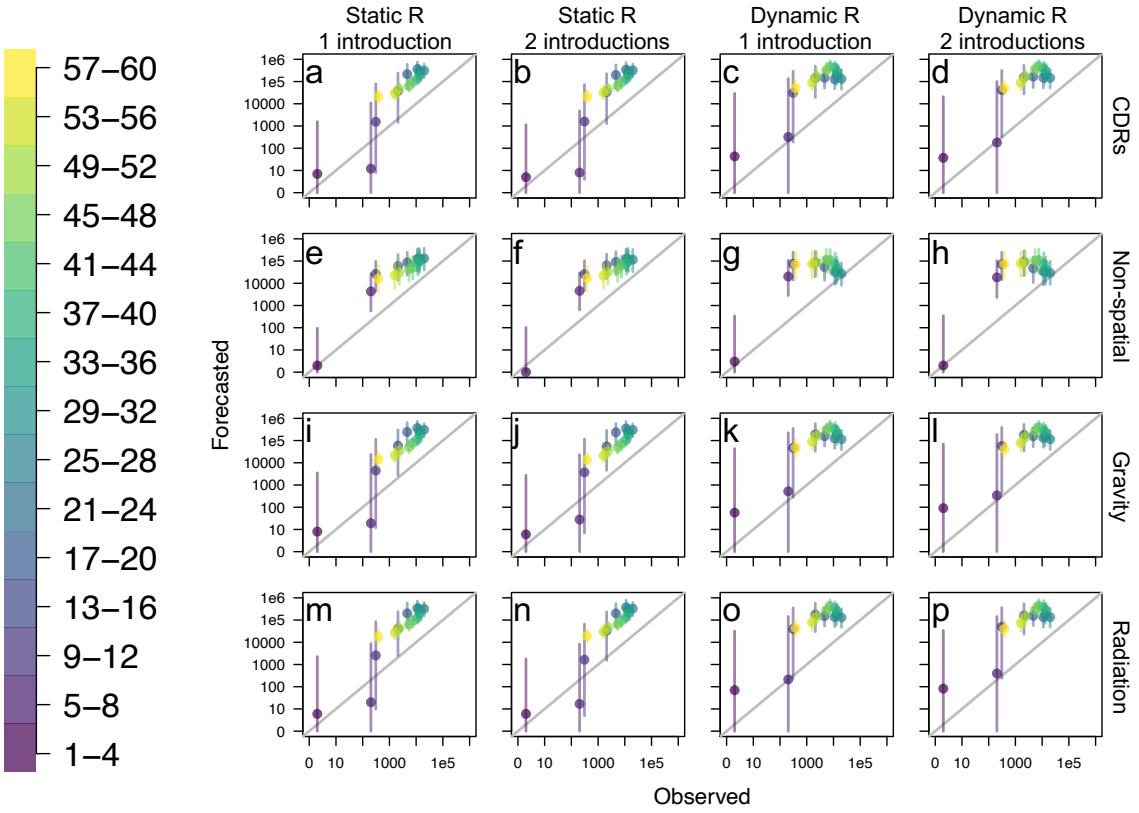

Figure S11: **Cumulative observed versus forecasted incidence at 1-, 2-, 3-, 4- week ahead intervals for each individual model aggregated nationally at each point throughout the epidemic.** *a-p.* Each plot represents a different model, with model feature labels on rows and columns. Point color indicates time at which forecasts were generated. Points are the median values and lines are the 50% credible interval. 1:1 line is in grey.

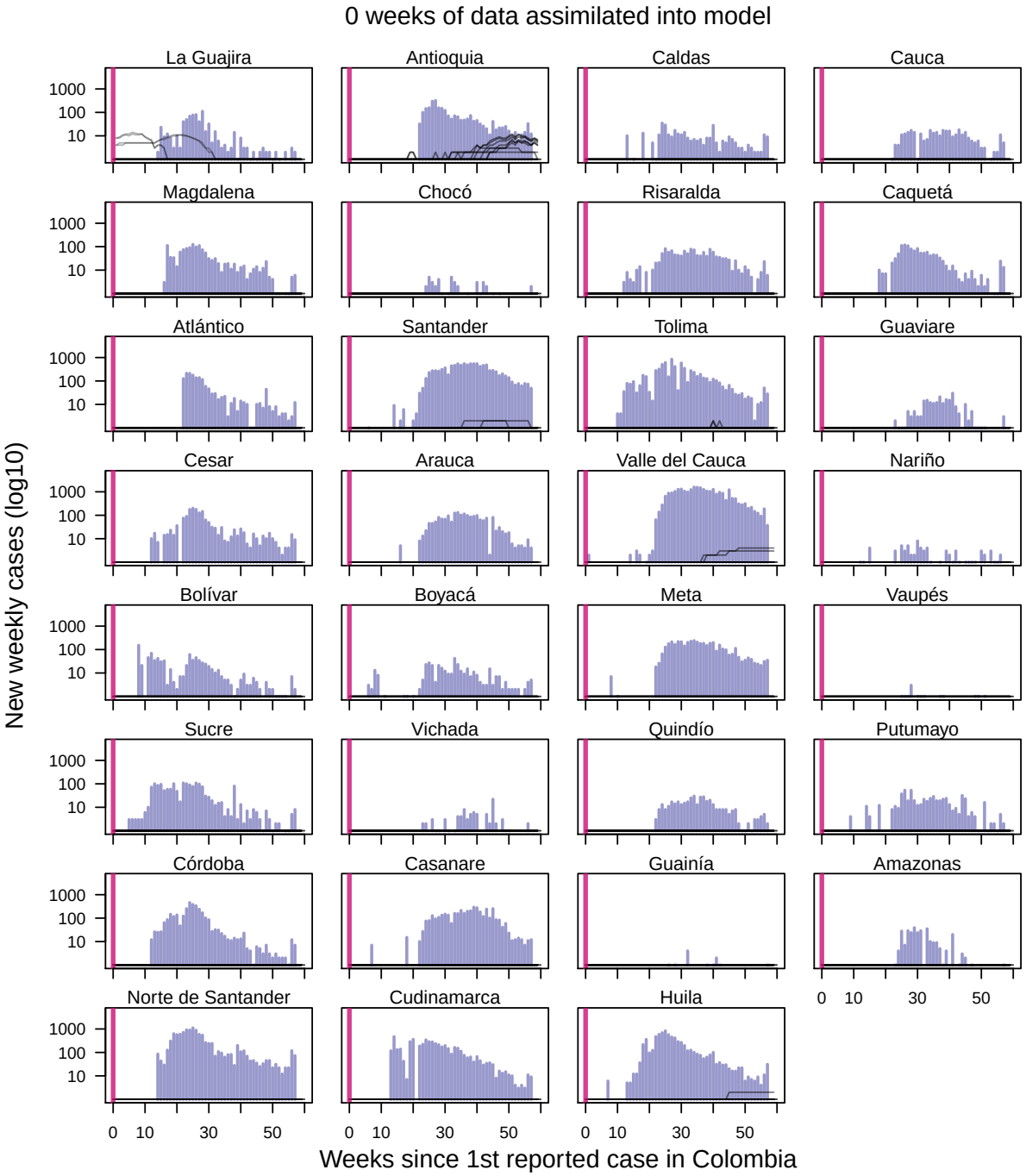

Figure S12: **Observed incidence with initial median forecast for each of the 16 models with no weeks of data yet assimilated forecasting models.** Navy bars indicate Zika incidence data at weekly time interval [14]. Vertical line indicates the point at which the forecast was made, with data to the left of the line assimilated into the model. Forecasts to the right of the vertical line change as more data is assimilated into the model, while model fits to the left of the vertical line do not change. With the forecasts generally being at zero cases in the majority of the departments, we see that models were unlikely to forecast an epidemic to occur when no data was yet to be assimilated.

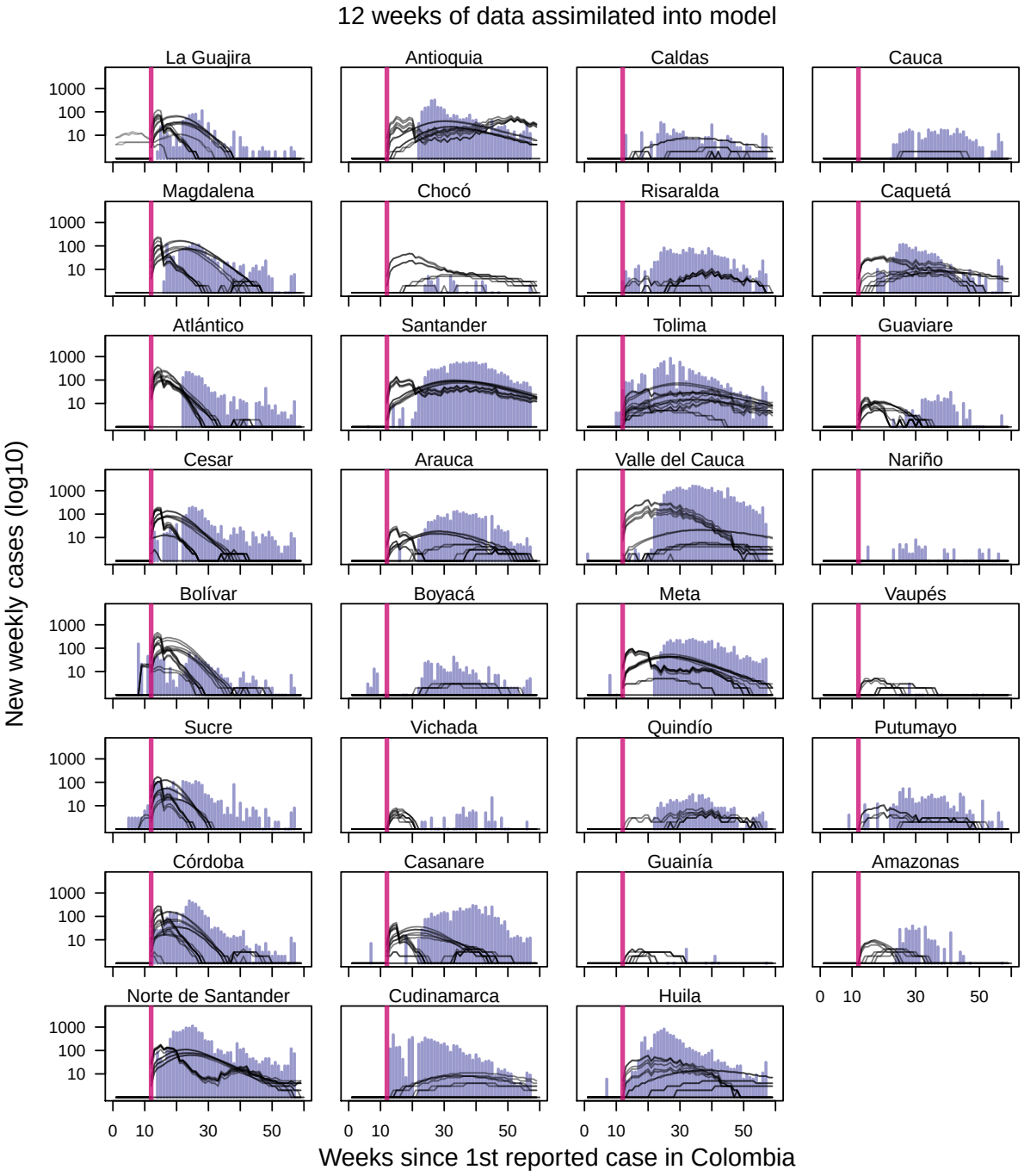

Figure S13: **Observed incidence with median forecast for each of the 16 models with 12 weeks of data assimilated into forecasting models.** Navy bars indicate Zika incidence data at weekly time interval [14]. Vertical line indicates the point at which the forecast was made, with data to the left of the line assimilated into the model. Forecasts to the right of the vertical line change as more data is assimilated into the model, while model fits to the left of the vertical line do not change.

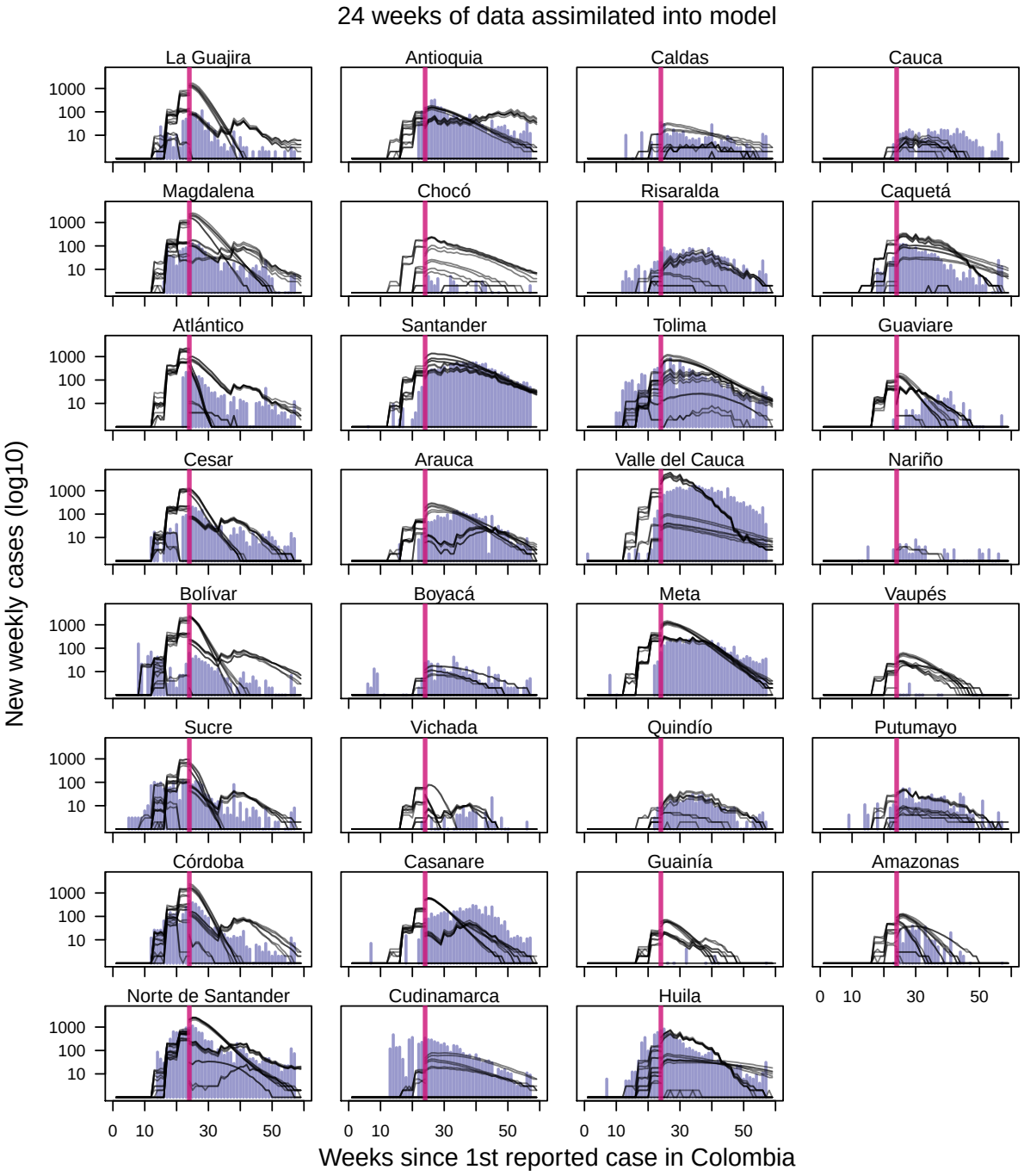

Figure S14: **Observed incidence with median forecast for each of the 16 models with 24 weeks of data assimilated into forecasting models.** Navy bars indicate Zika incidence data at weekly time interval [14]. Vertical line indicates the point at which the forecast was made, with data to the left of the line assimilated into the model. Forecasts to the right of the vertical line change as more data is assimilated into the model, while model fits to the left of the vertical line do not change.

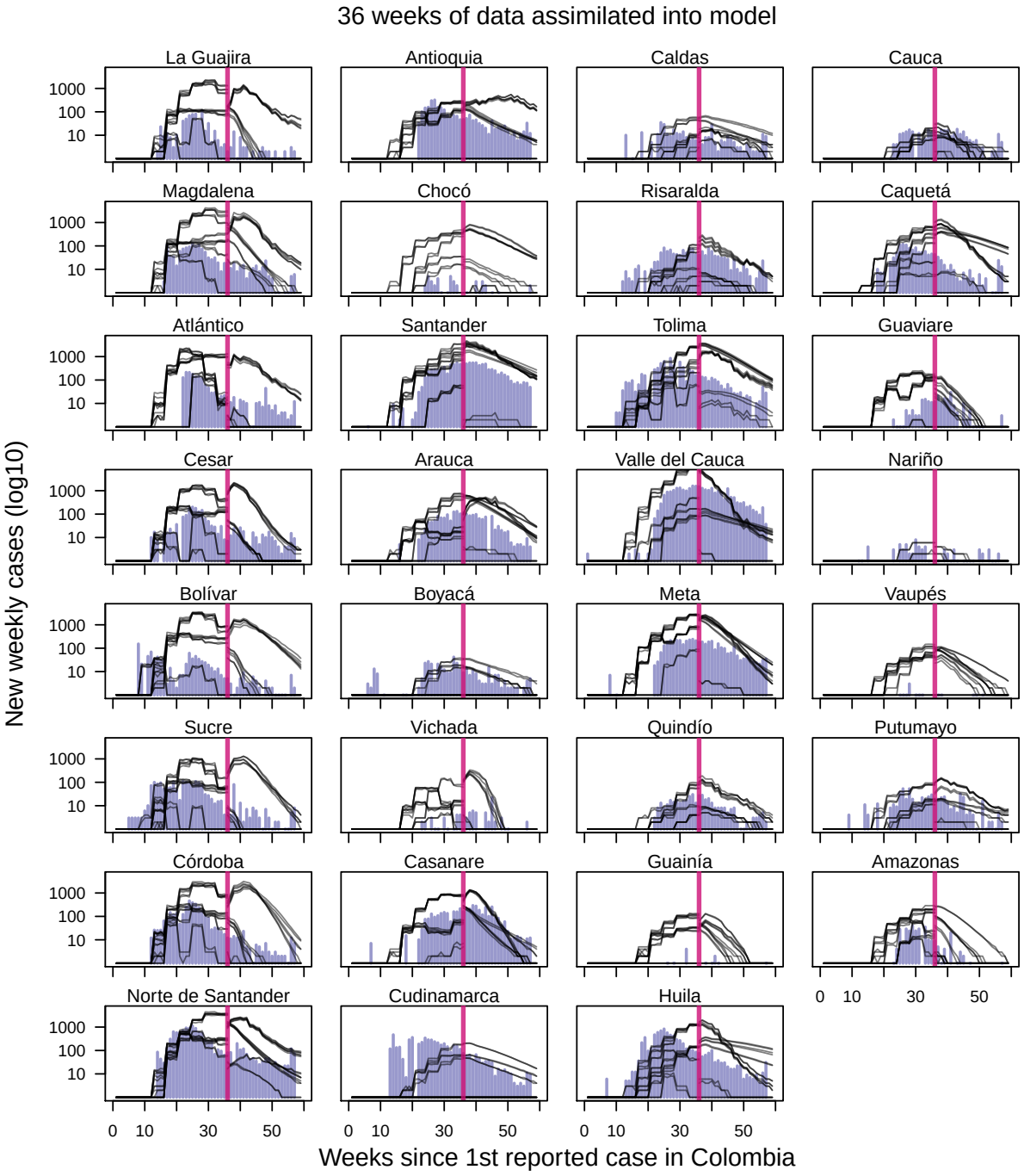

Figure S15: **Observed incidence with median forecast for each of the 16 models with 36 weeks of data assimilated into forecasting models.** Navy bars indicate Zika incidence data at weekly time interval [14]. Vertical line indicates the point at which the forecast was made, with data to the left of the line assimilated into the model. Forecasts to the right of the vertical line change as more data is assimilated into the model, while model fits to the left of the vertical line do not change.

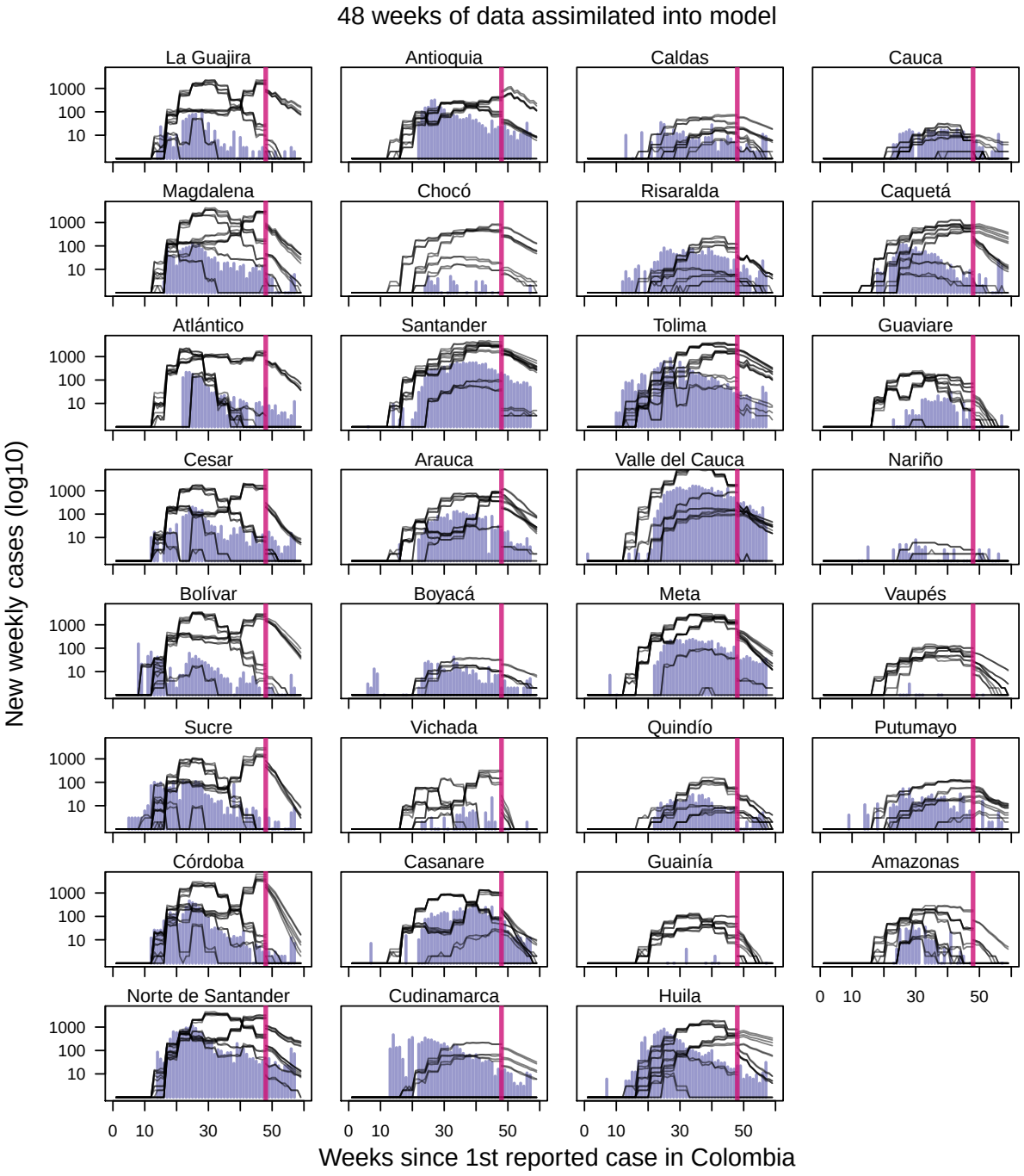

Figure S16: **Observed incidence with median forecast for each of the 16 models with 48 weeks of data assimilated into forecasting models.** Navy bars indicate Zika incidence data at weekly time interval [14]. Vertical line indicates the point at which the forecast was made, with data to the left of the line assimilated into the model. Forecasts to the right of the vertical line change as more data is assimilated into the model, while model fits to the left of the vertical line do not change.

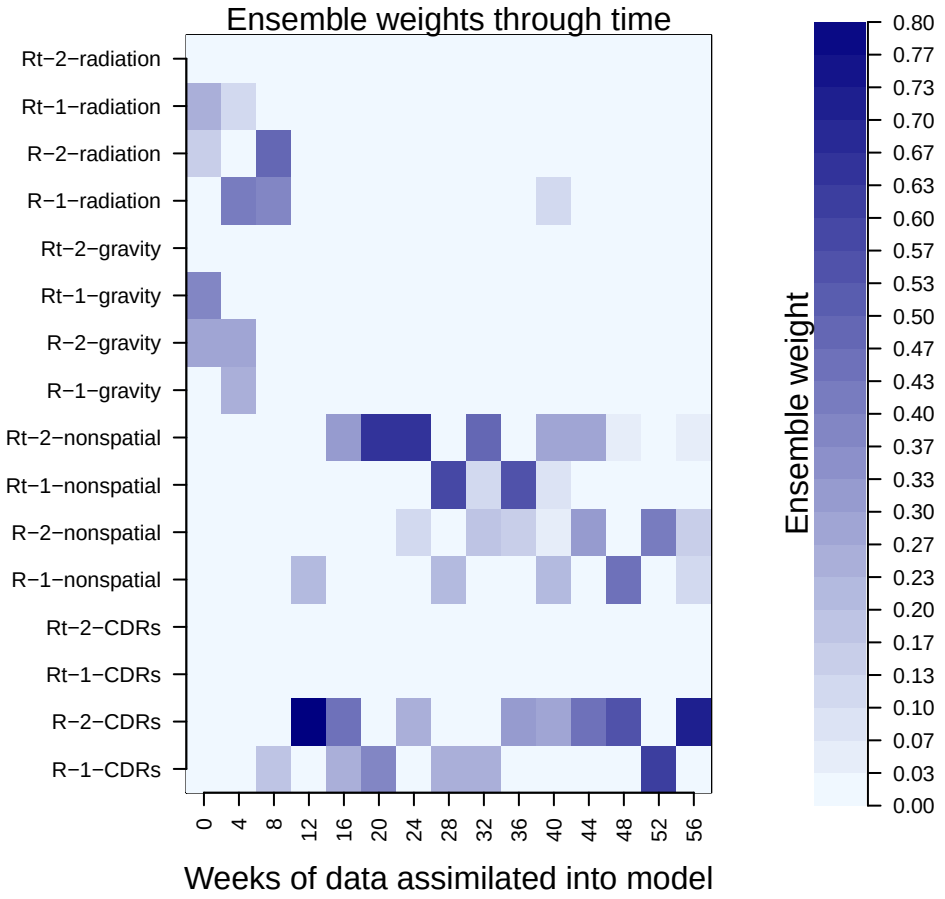

Figure S17: **Ensemble weights of each model at each assimilation period.** Ensemble weights were calculated using the expectation-maximization algorithm on short-term forecast performances, where forecast performance was assessed against 4-wk ahead incidence in a given department and nationally.

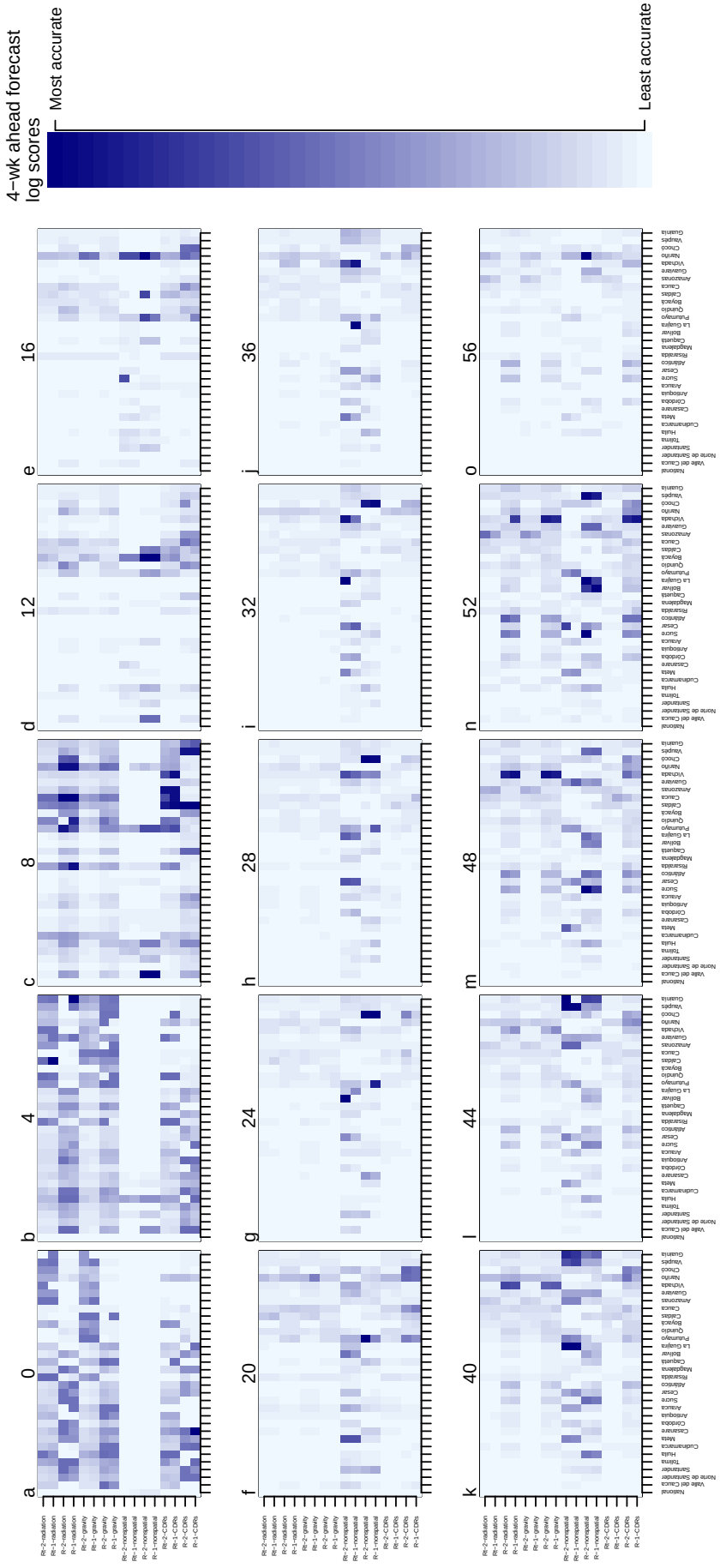

Figure S18: **Log scores of each model and department at each assimilation period.** Log scores were assessed using cumulative 4-wk ahead forecasts for each department and aggregated nationally. These log scores were then used in the expectation-maximization algorithm.

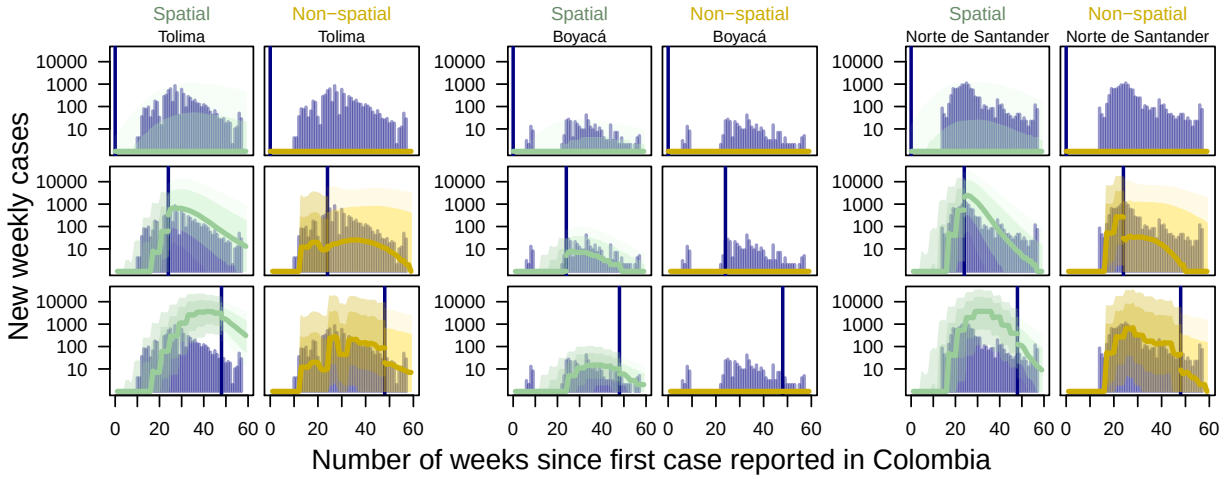

Figure S19: **Forecasts for a spatially coupled (green) and uncoupled (yellow) models with one introduction and static R for three departments.** Navy bars denote incidence data. Large bands denote the 75% CrI, darker band denotes the 50% CrI, and thick line denotes the median forecast. Each row is from the same time point. Time points were chosen to be equally spaced out through the epidemic, with the first set of forecasts from the week of the first case, the second set of forecasts generated at 24 weeks after the first case was reported in Colombia, and the third set of forecasts generated at 48 weeks after the first case was reported in Colombia.

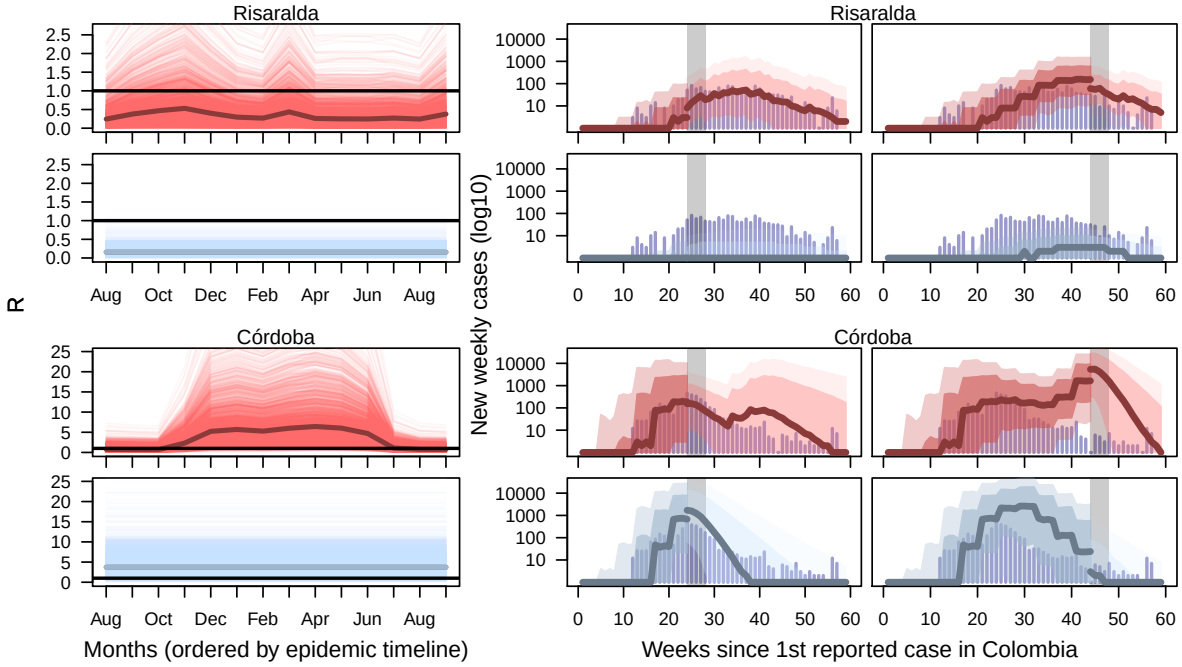

Figure S20: **Fitted R and forecasts in two departments for models with two ZIKV introductions, cell phone mobility informed human movement, and a dynamic or static R.** Models with dynamic R are depicted in red and models with a static R are depicted in blue. In plots of R, each line is a draw from the posterior, with a bold median line; horizontal black line depicts $R = 1$. The first set of forecasts in the middle column are from the peak week in both Risaralda and Córdoba, 24-28 weeks after the first case was reported in Colombia, and the second set are from 44-48 weeks after the first case was reported in Colombia. Vertical grey bars depict the forecasts and data considered when assessing short-term forecast performance.

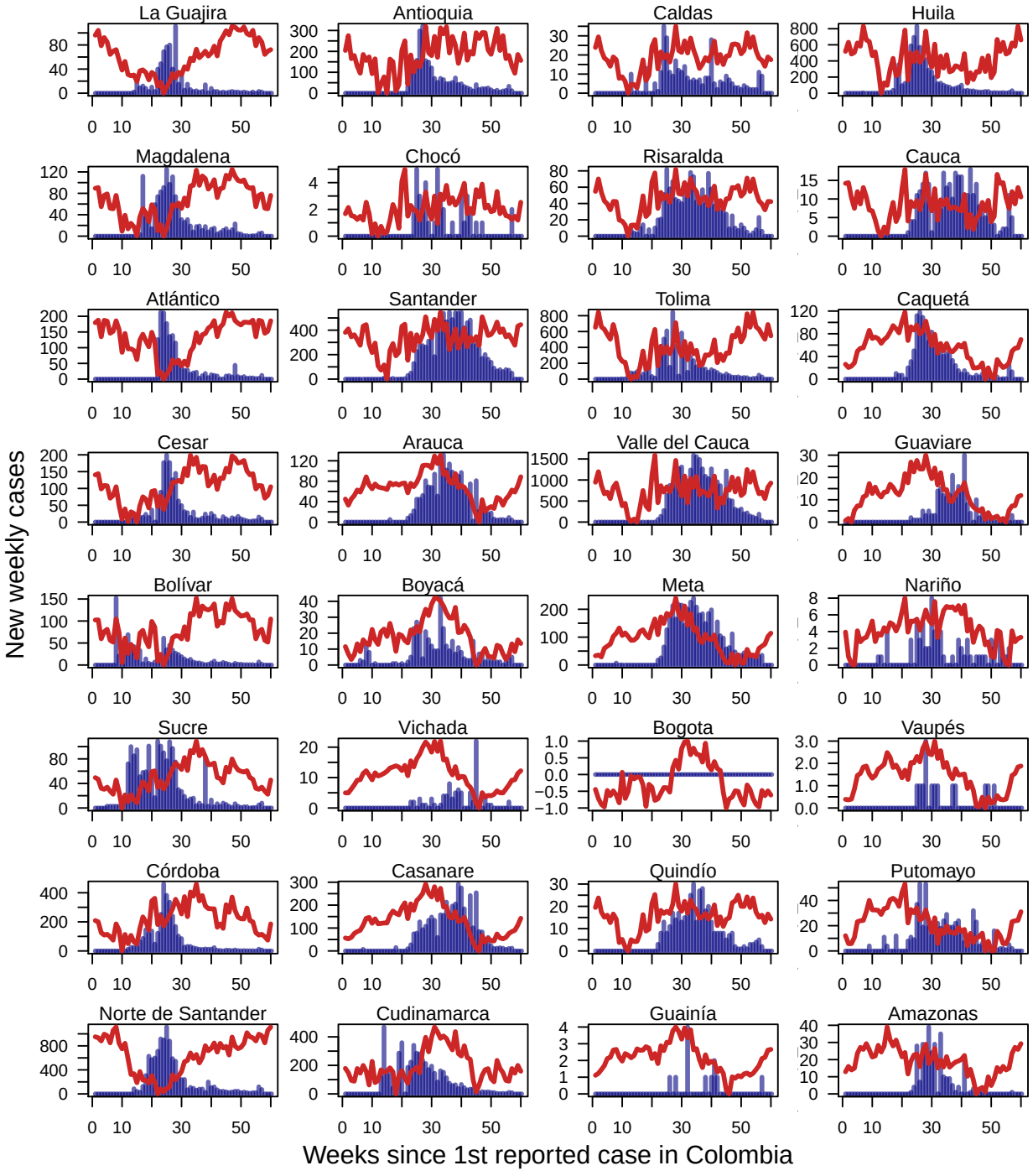

Figure S21: **Incidence by department with temperature trends.** Blue bars denote weekly Zika incidence and red line denotes temperature trends.

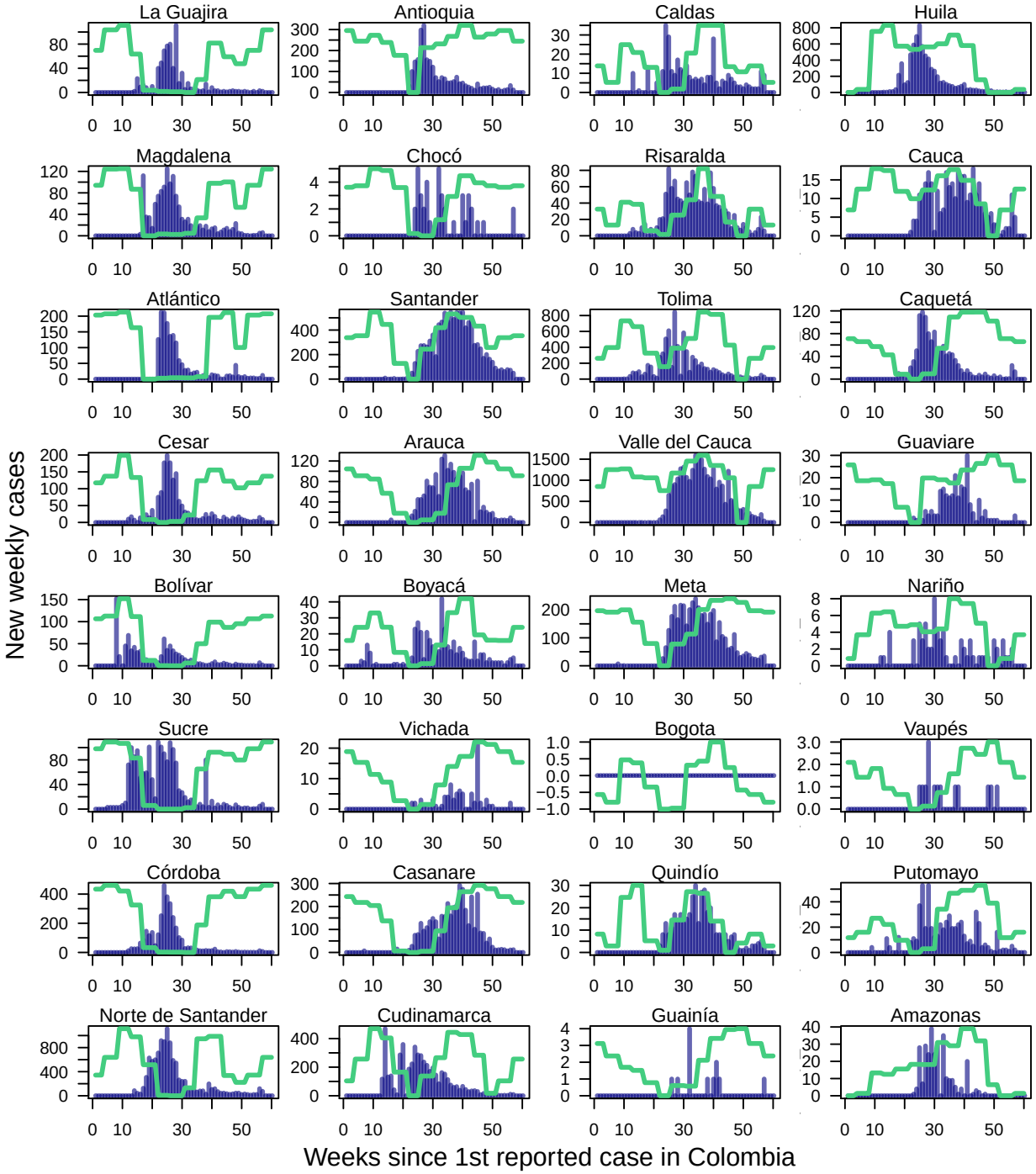

Figure S22: **Incidence by department with mosquito occurrence probability trends.** Blue bars denote weekly Zika incidence and green line denotes mosquito occurrence trends.

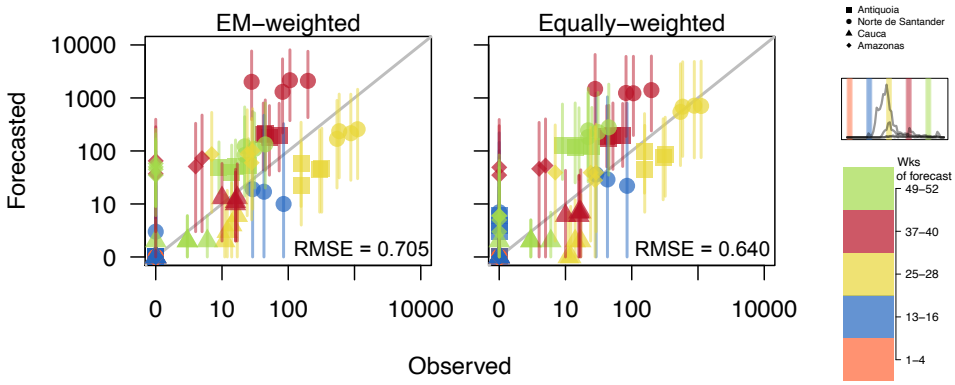

Figure S23: **Observed versus forecasted incidence at 1-, 2-, 3-, and 4-week ahead intervals for EM-weighted and equally weighted ensemble models for Antioquia, Norte de Santander, Cauca, and Amazonas.** Plotted departments reflect differences in population, epidemic size, and geographic regions of Colombia and are denoted by point type. Point shape denotes department. Point color indicates time at which the forecast was made (visually denoted in inset plot and color bar). Point is the median value and lines are the 50% credible interval. 1:1 line is in grey. The root mean square errors for the EM-weighted and equally weighted forecasts shown here are 0.705 and 0.640, respectively.

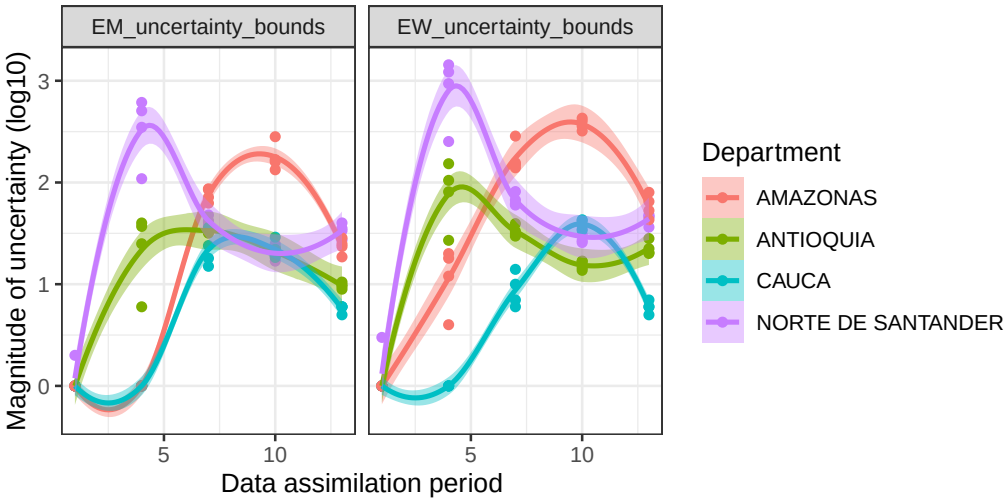

Figure S24: **Magnitude of the 50% uncertainty bounds (as shown in Fig. S23) for 1-, 2-, 3-, and 4- week ahead forecasts in five different data assimilation periods for the EM-weighted and equally weighted ensemble models for Amazonas, Antioquia, Cauca, and Norte de Santander.** The four points per data assimilation period represent the 1-, 2-, 3-, and 4- week ahead forecasts for each of the four departments denoted by color. Smoothed loess lines are shown to demonstrate how the magnitude of uncertainty changes through time for each department.

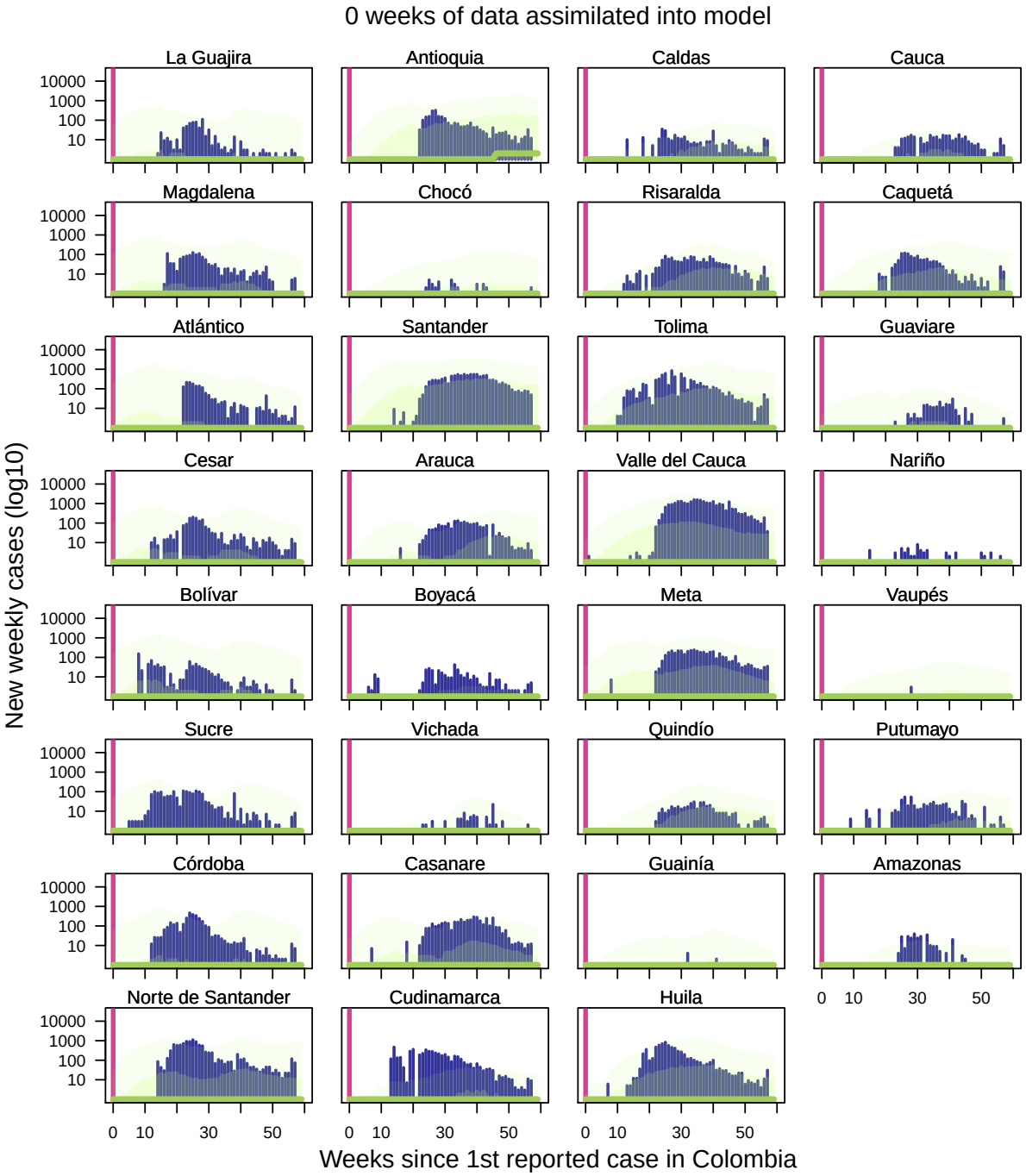

Figure S25: **Observed incidence with initial expectation maximization algorithm-weighted ensemble forecast with no weeks of data yet assimilated into forecasting models.** Navy bars indicate Zika incidence data at weekly time interval [14]. Light green band denotes 75% credible interval, darker green band denotes 50% credible interval, and the dark green line denotes median ensemble forecast. Vertical line indicates the point at which the forecast was made, with data to the left of the line assimilated into the model. Forecasts to the right of the vertical line change as more data is assimilated into the model, while model fits to the left of the vertical line do not change.

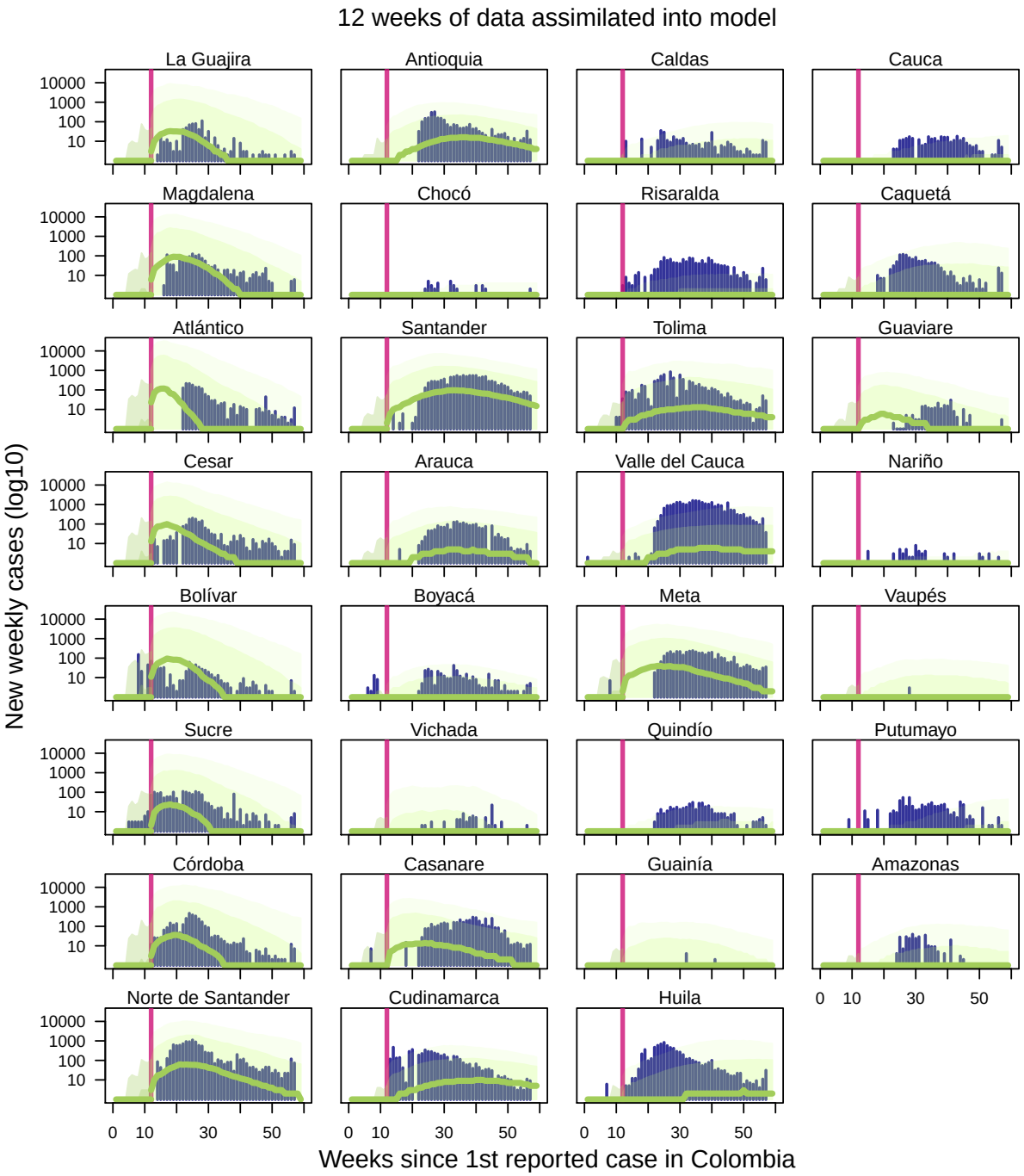

Figure S26: **Observed incidence with expectation maximization algorithm-weighted ensemble forecast with 12 weeks of data assimilated into forecasting models.** Navy bars indicate Zika incidence data at weekly time interval [14]. Light green band denotes 75% credible interval, darker green band denotes 50% credible interval, and the dark green line denotes median ensemble forecast. Vertical line indicates the point at which the forecast was made, with data to the left of the line assimilated into the model. Forecasts to the right of the vertical line change as more data is assimilated into the model, while model fits to the left of the vertical line do not change.

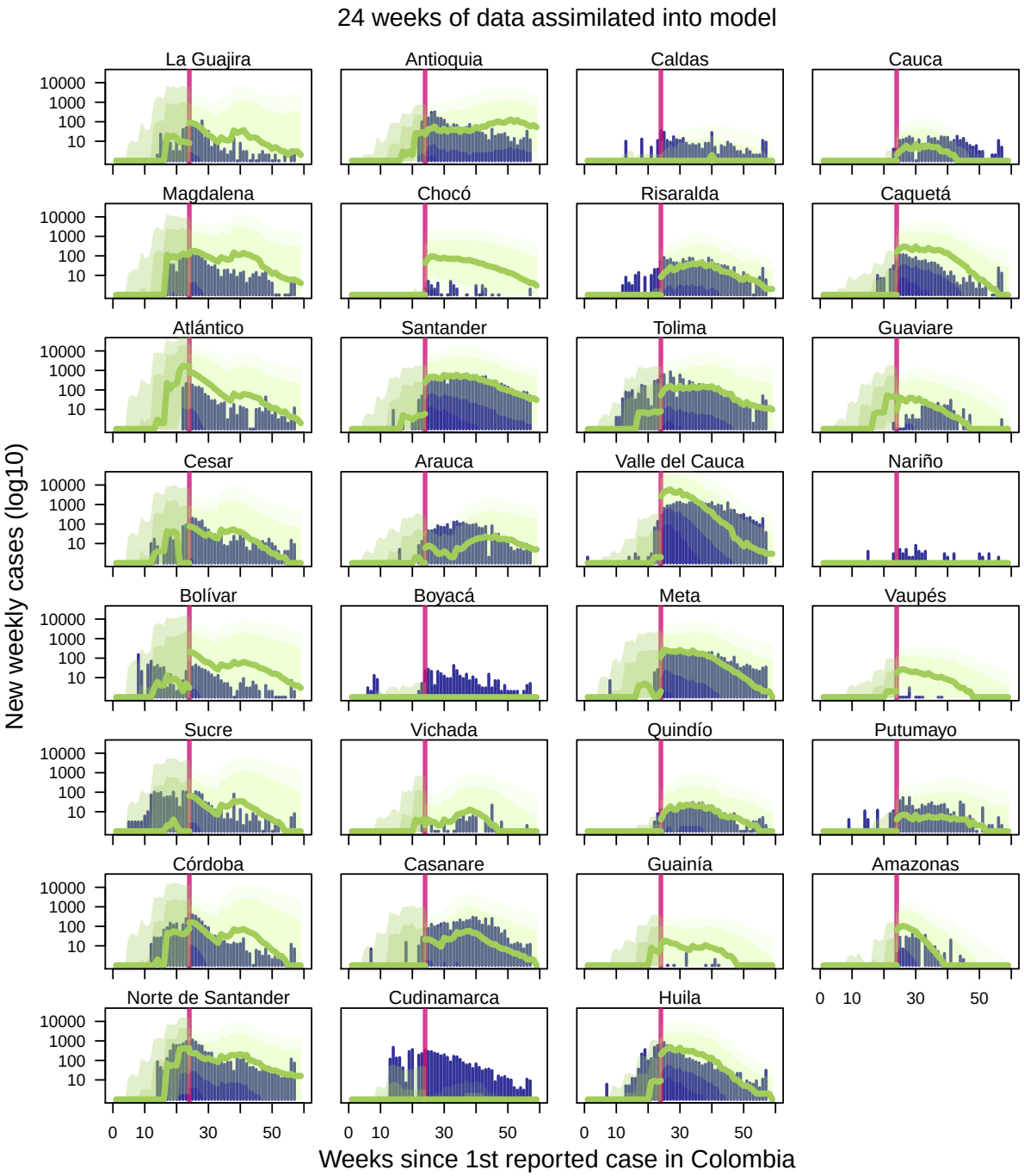

Figure S27: **Observed incidence with expectation maximization algorithm-weighted ensemble forecast with 24 weeks of data assimilated into forecasting models.** Navy bars indicate Zika incidence data at weekly time interval [14]. Light green band denotes 75% credible interval, darker green band denotes 50% credible interval, and the dark green line denotes median ensemble forecast. Vertical line indicates the point at which the forecast was made, with data to the left of the line assimilated into the model. Forecasts to the right of the vertical line change as more data is assimilated into the model, while model fits to the left of the vertical line do not change.

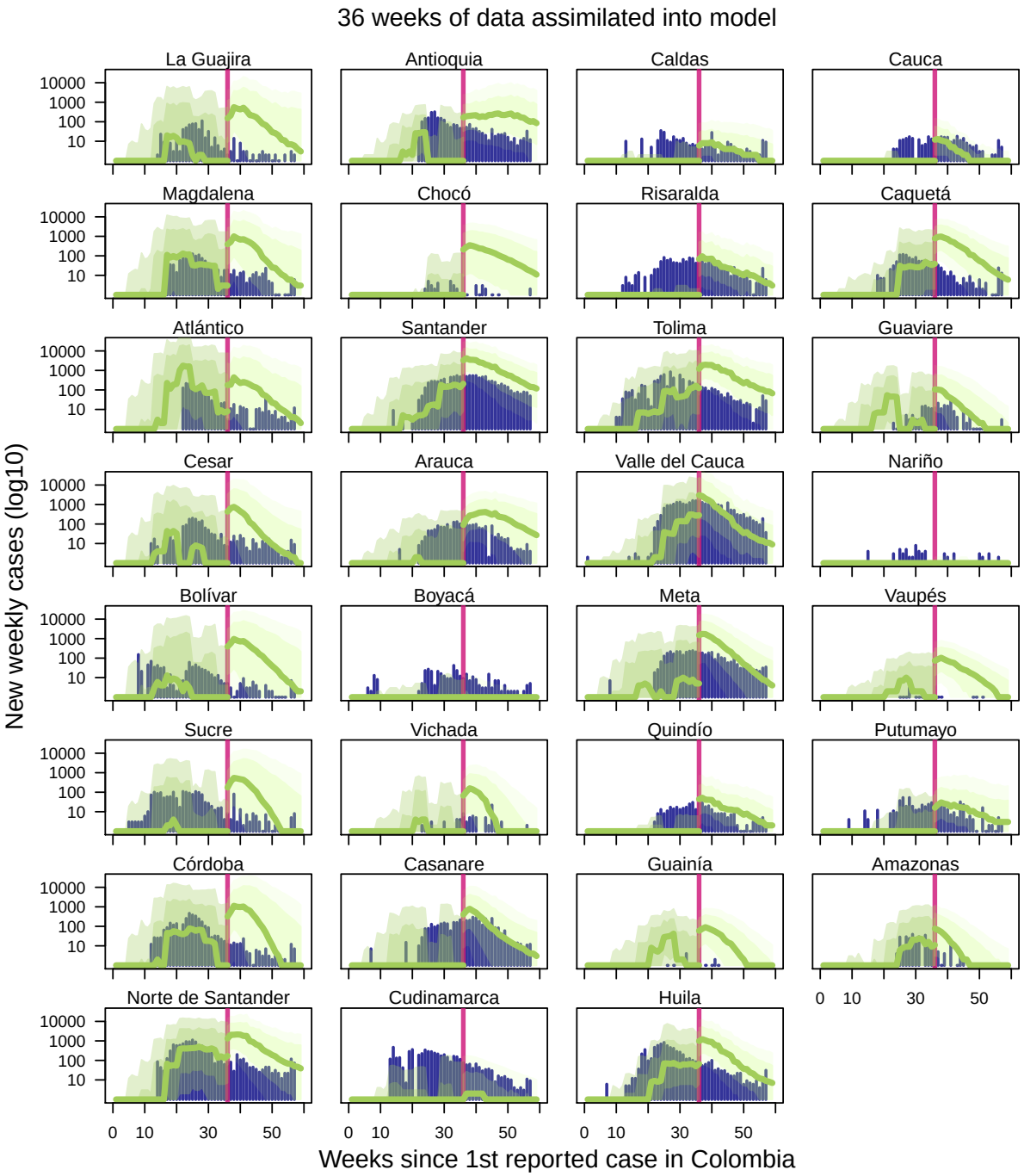

Figure S28: **Observed incidence with expectation maximization algorithm-weighted ensemble forecast with 36 weeks of data assimilated into forecasting models.** Navy bars indicate Zika incidence data at weekly time interval [14]. Light green band denotes 75% credible interval, darker green band denotes 50% credible interval, and the dark green line denotes median ensemble forecast. Vertical line indicates the point at which the forecast was made, with data to the left of the line assimilated into the model. Forecasts to the right of the vertical line change as more data is assimilated into the model, while model fits to the left of the vertical line do not change.

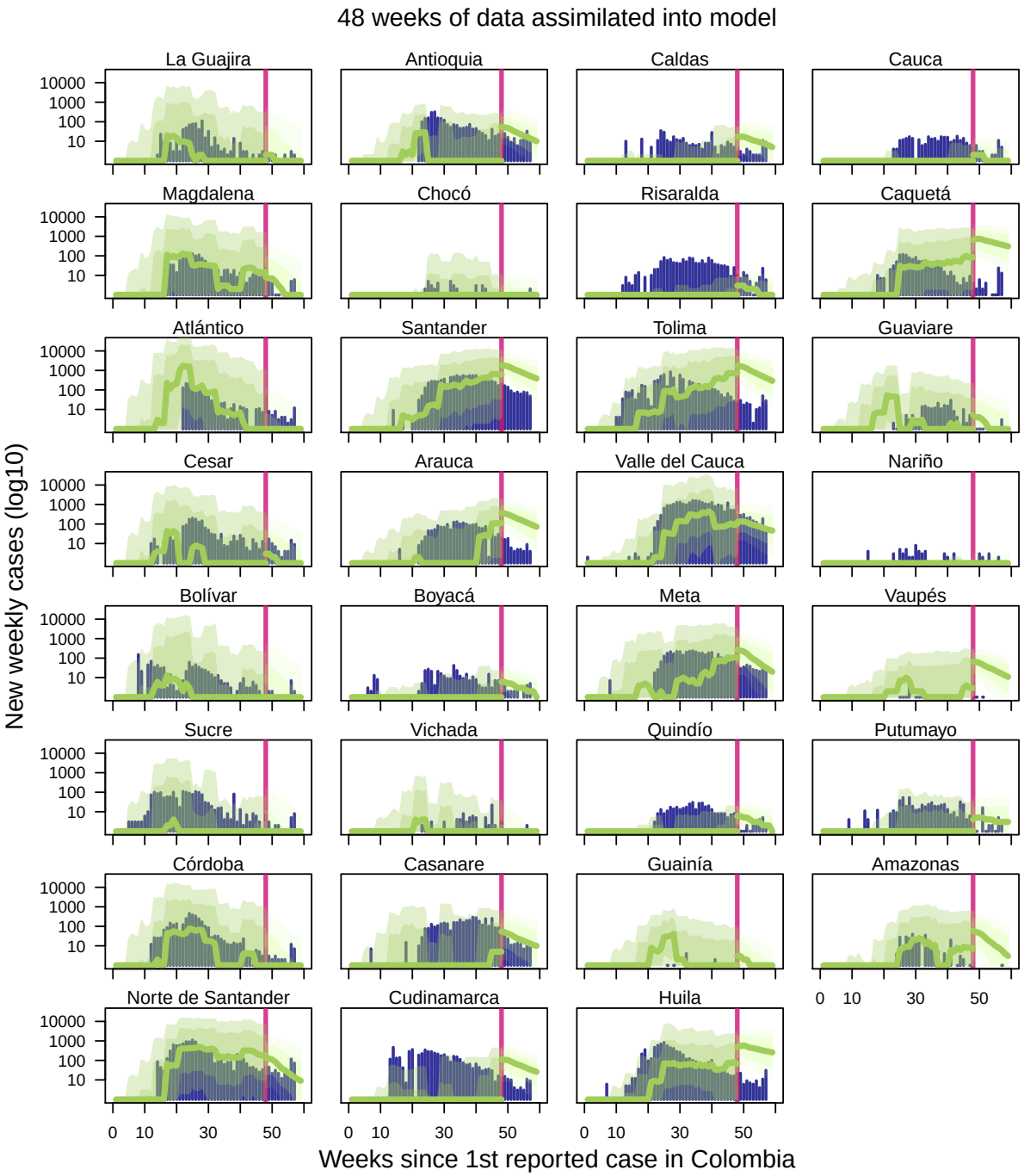

Figure S29: **Observed incidence with expectation maximization algorithm-weighted ensemble forecast with 48 weeks of data assimilated into forecasting models.** Navy bars indicate Zika incidence data at weekly time interval [14]. Light green band denotes 75% credible interval, darker green band denotes 50% credible interval, and the dark green line denotes median ensemble forecast. Vertical line indicates the point at which the forecast was made, with data to the left of the line assimilated into the model. Forecasts to the right of the vertical line change as more data is assimilated into the model, while model fits to the left of the vertical line do not change.

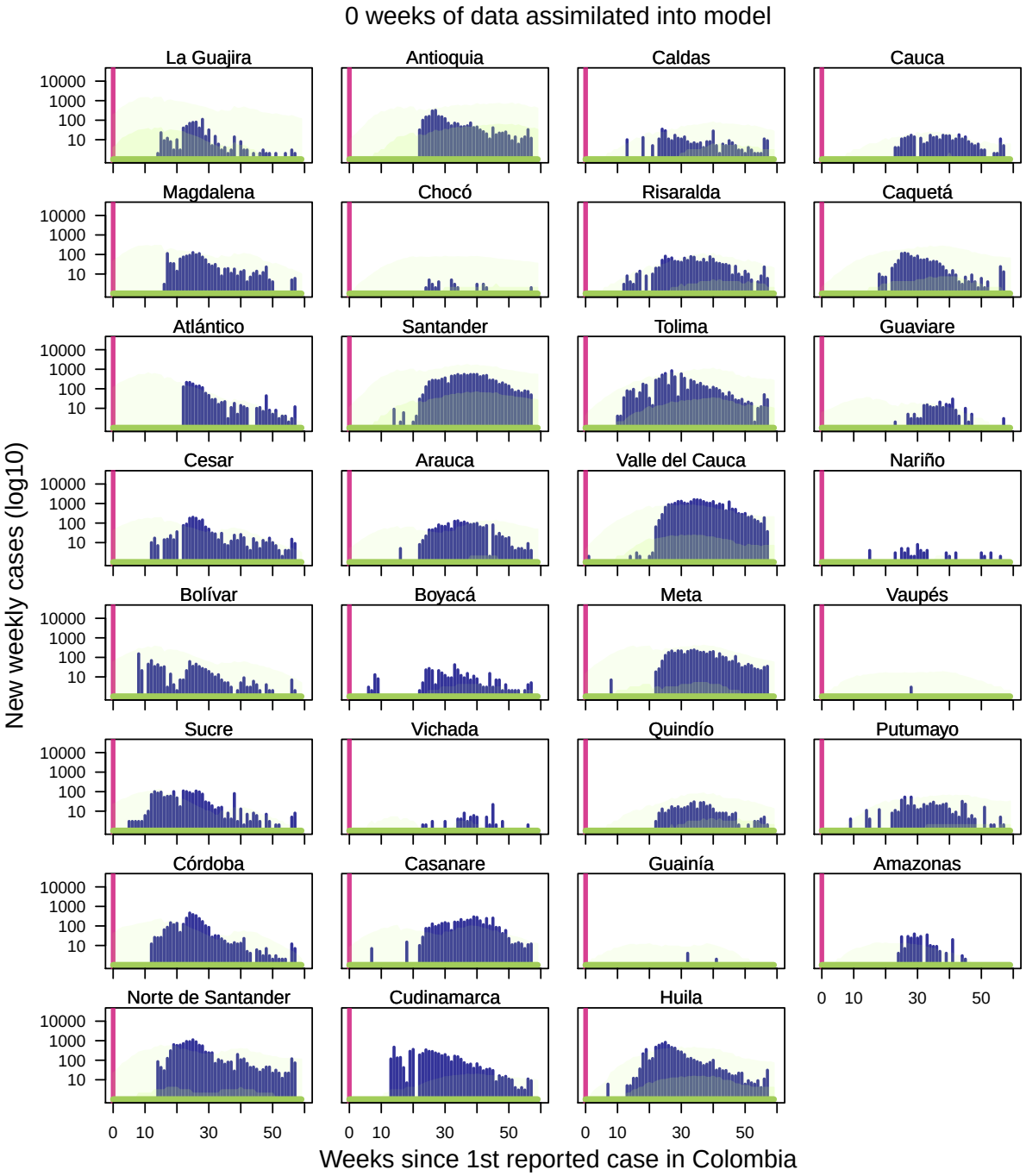

Figure S30: **Observed incidence with initial equally weighted ensemble forecast with no weeks of data yet assimilated into forecasting models.** Navy bars indicate Zika incidence data at weekly time interval [14]. Light green band denotes 75% credible interval, darker green band denotes 50% credible interval, and the dark green line denotes median ensemble forecast. Vertical line indicates the point at which the forecast was made, with data to the left of the line assimilated into the model. Forecasts to the right of the vertical line change as more data is assimilated into the model, while model fits to the left of the vertical line do not change.

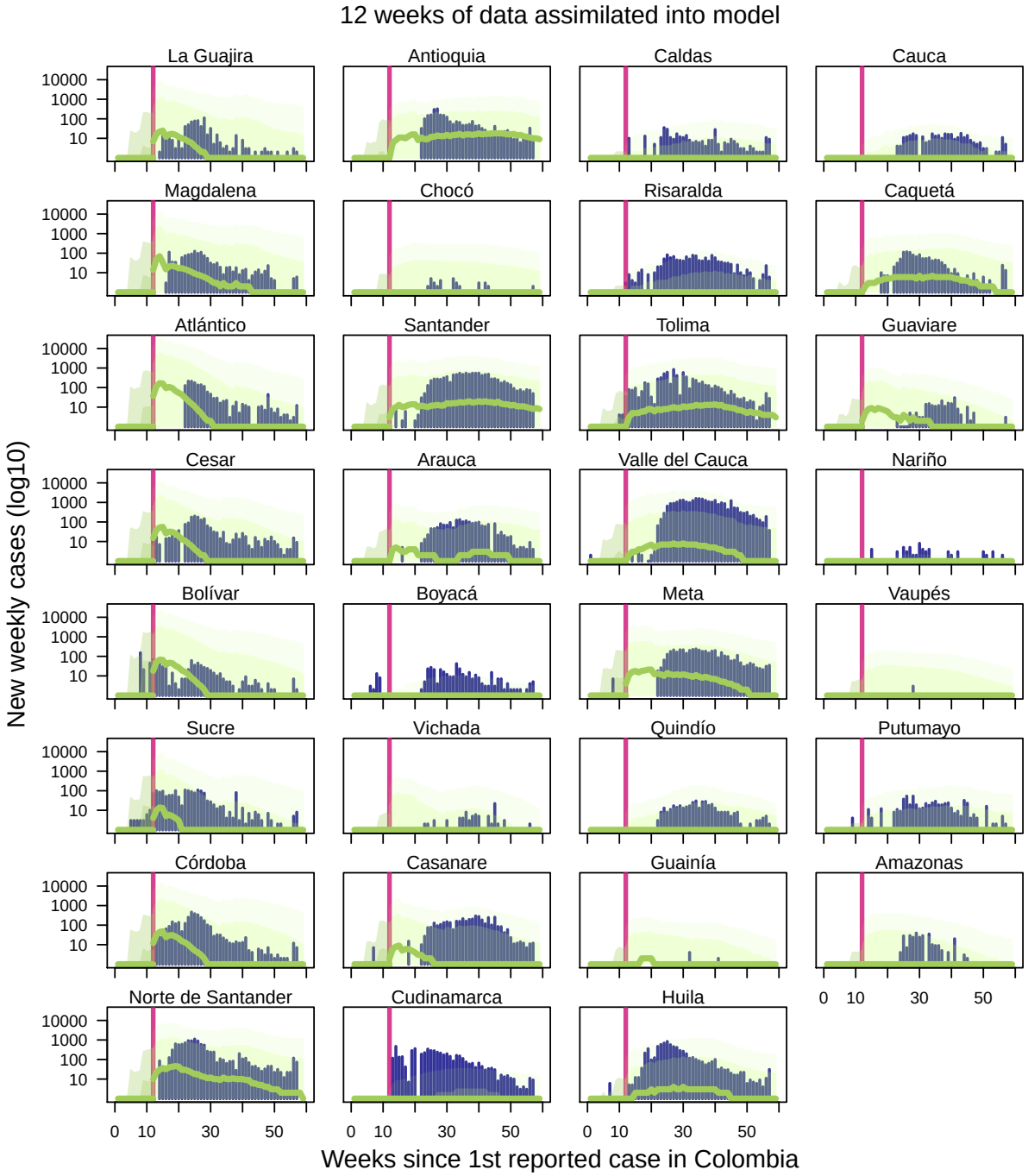

Figure S31: **Observed incidence with equally weighted ensemble forecast with 12 weeks of data assimilated into forecasting models.** Navy bars indicate Zika incidence data at weekly time interval [14]. Light green band denotes 75% credible interval, darker green band denotes 50% credible interval, and the dark green line denotes median ensemble forecast. Vertical line indicates the point at which the forecast was made, with data to the left of the line assimilated into the model. Forecasts to the right of the vertical line change as more data is assimilated into the model, while model fits to the left of the vertical line do not change.

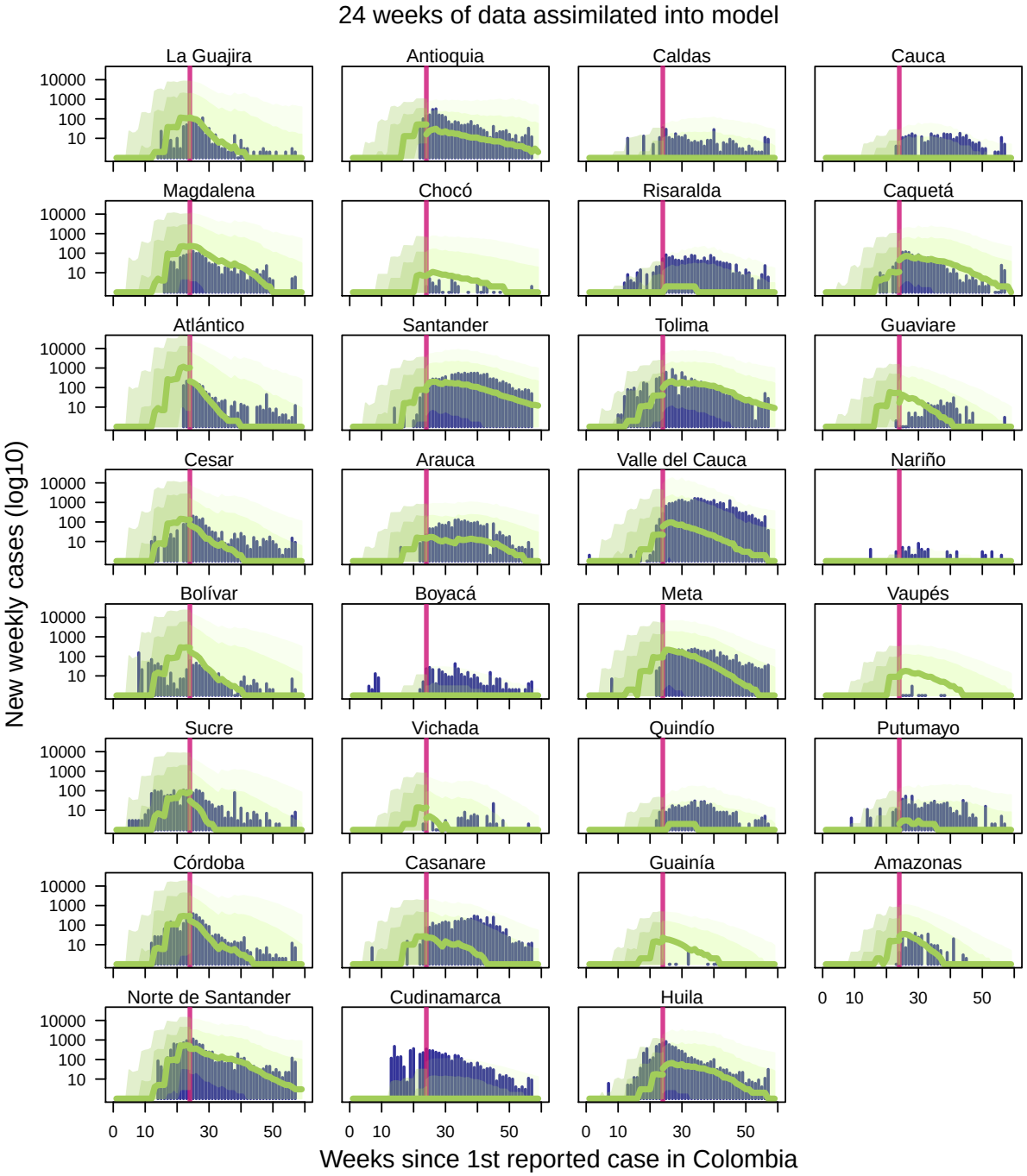

Figure S32: **Observed incidence with equally weighted ensemble forecast with 24 weeks of data assimilated into forecasting models.** Navy bars indicate Zika incidence data at weekly time interval [14]. Light green band denotes 75% credible interval, darker green band denotes 50% credible interval, and the dark green line denotes median ensemble forecast. Vertical line indicates the point at which the forecast was made, with data to the left of the line assimilated into the model. Forecasts to the right of the vertical line change as more data is assimilated into the model, while model fits to the left of the vertical line do not change.

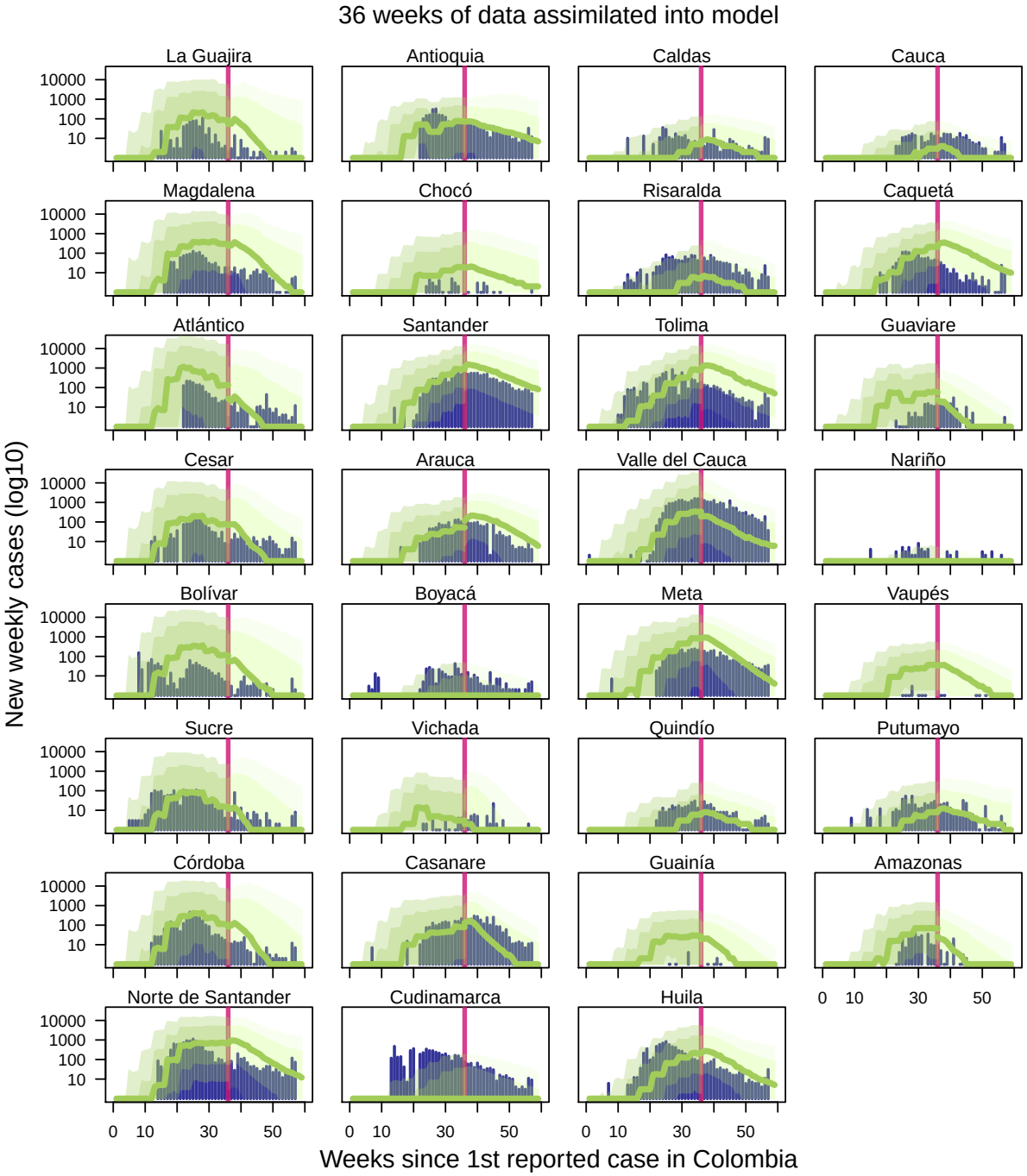

Figure S33: **Observed incidence with equally weighted ensemble forecast with 36 weeks of data assimilated into forecasting models.** Navy bars indicate Zika incidence data at weekly time interval [14]. Light green band denotes 75% credible interval, darker green band denotes 50% credible interval, and the dark green line denotes median ensemble forecast. Vertical line indicates the point at which the forecast was made, with data to the left of the line assimilated into the model. Forecasts to the right of the vertical line change as more data is assimilated into the model, while model fits to the left of the vertical line do not change.

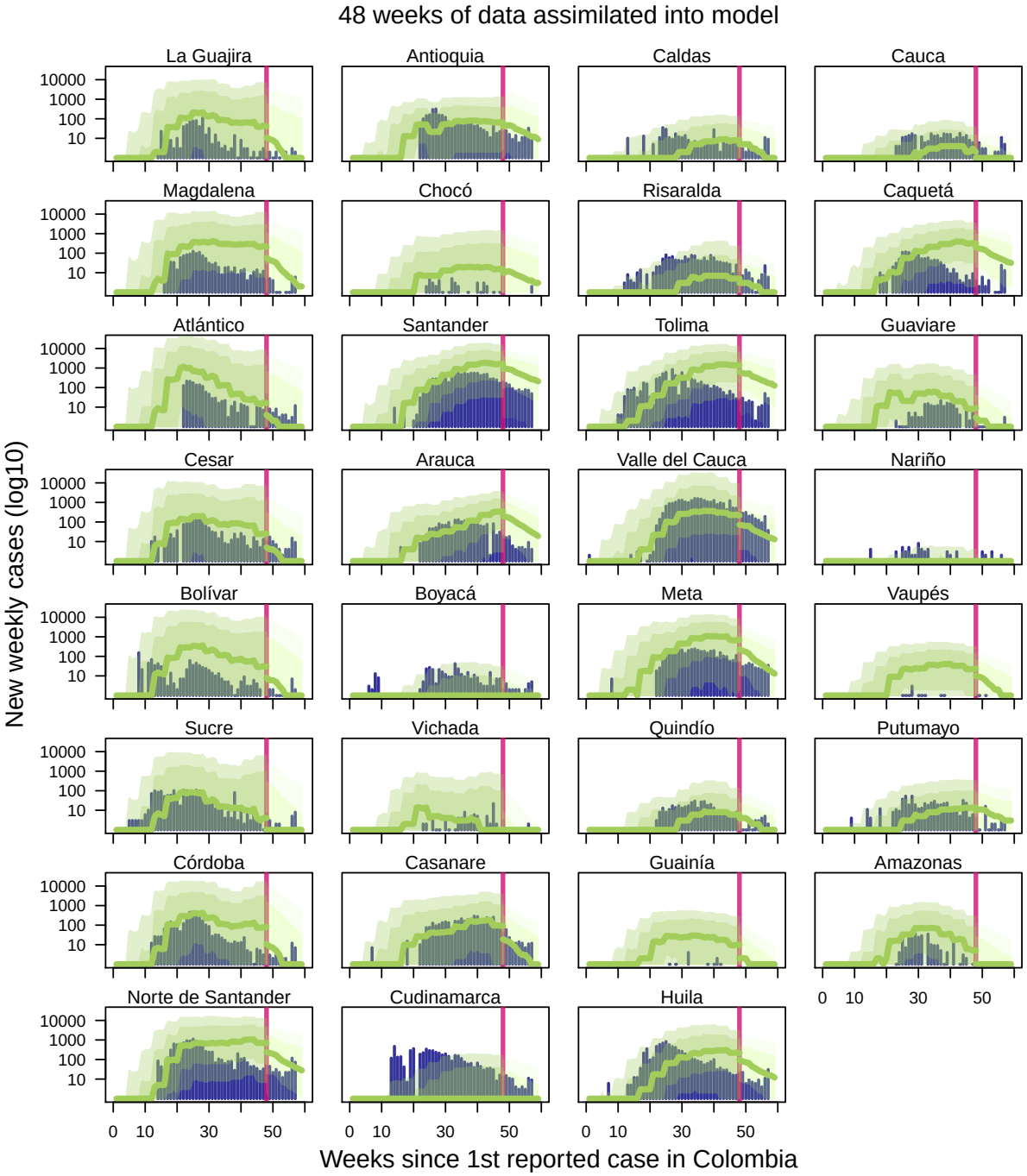

Figure S34: **Observed incidence with equally weighted ensemble forecast with 48 weeks of data assimilated into forecasting models.** Navy bars indicate Zika incidence data at weekly time interval [14]. Light green band denotes 75% credible interval, darker green band denotes 50% credible interval, and the dark green line denotes median ensemble forecast. Vertical line indicates the point at which the forecast was made, with data to the left of the line assimilated into the model. Forecasts to the right of the vertical line change as more data is assimilated into the model, while model fits to the left of the vertical line do not change.

References

[1] Sebastian Funk et al. "Comparative Analysis of Dengue and Zika Outbreaks Reveals Differences by Setting and Virus". In: *PLOS Neglected Tropical Diseases* 10.12 (Dec. 2016), pp. 1–16. DOI: 10.1371/journal.pntd.0005173. URL: https://doi.org/10.1371/journal.pntd.0005173.

[2] Samir Bhatt et al. "The global distribution and burden of dengue". In: *Nature* 496.7446 (Apr. 2013), pp. 504–507. ISSN: 1476-4687. DOI: 10.1038/nature12060. URL: https://doi.org/10.1038/nature12060.

[3] Monaise M.O. Silva et al. "Accuracy of dengue reporting by national surveillance system, Brazil". In: *Emerging Infectious Diseases* 22.2 (Feb. 2016). DOI: https://dx.doi.org/10.3201/eid2202.150495. URL: https://wwwnc.cdc.gov/eid/article/22/2/15-0495_article.

[4] Deborah P. Shutt et al. "Estimating the reproductive number, total outbreak size, and reporting rates for Zika epidemics in South and Central America". In: *Epidemics* 21 (Dec. 2017), pp. 63–79. ISSN: 1755-4365. URL: http://www.sciencedirect.com/science/article/pii/S1755436517300257.

[5] Rachel J Oidtman, Guido España, and T Alex Perkins. "Co-circulation and misdiagnosis led to underestimation of the 2015-2017 Zika epidemic in the Americas". In: *medRxiv* (2020). DOI: https://doi.org/10.1101/19010256. URL: https://www.medrxiv.org/content/10.1101/19010256v1.

[6] Sean M. Moore et al. "Leveraging multiple data types to estimate the size of the Zika epidemic in the Americas". In: *PLOS Neglected Tropical Diseases* 14.9 (Sept. 2020), pp. 1–25. DOI: 10.1371/journal.pntd.0008640. URL: https://doi.org/10.1371/journal.pntd.0008640.

[7] M. U. G. Kraemer et al. "Utilizing general human movement models to predict the spread of emerging infectious diseases in resource poor settings". In: *Scientific Reports* 9.1 (2019), p. 5151. ISSN: 2045-2322. DOI: 10.1038/s41598-019-41192-3.

[8] T. Alex Perkins et al. "Model-based projections of Zika virus infections in childbearing women in the Americas". In: *Nature Microbiology* 1.9 (2016), p. 16126. DOI: 10.1038/nmicrobiol.2016.126. URL: https://doi.org/10.1038/nmicrobiol.2016.126.

[9] Erin A. Mordecai et al. "Detecting the impact of temperature on transmission of Zika, dengue, and chikungunya using mechanistic models". In: *PLOS Neglected Tropical Diseases* 11.4 (Apr. 2017), pp. 1–18. DOI: 10.1371/journal.pntd.0005568. URL: https://doi.org/10.1371/journal.pntd.0005568.

[10] Moritz UG Kraemer et al. "The global distribution of the arbovirus vectors *Aedes aegypti* and *Ae. albopictus*". In: *eLife* 4 (June 2015), e08347. ISSN: 2050-084X. DOI: 10.7554/eLife.08347. URL: https://doi.org/10.7554/eLife.08347.

[11] Oliver J. Brady et al. "Global temperature constraints on Aedes aegypti and Ae. albopictus persistence and competence for dengue virus transmission". In: *Parasites & Vectors* 7.1 (July 2014), p. 338. ISSN: 1756-3305. DOI: 10.1186/1756-3305-7-338. URL: https://doi.org/10.1186/1756-3305-7-338.

[12] Miranda Chan and Michael A. Johansson. "The Incubation Periods of Dengue Viruses". In: *PLOS ONE* 7.11 (Nov. 2012), pp. 1–7. DOI: 10.1371/journal.pone.0050972. URL: https://doi.org/10.1371/journal.pone.0050972.

[13] Allison Black et al. "Genomic epidemiology supports multiple introductions and cryptic transmission of Zika virus in Colombia". In: *BMC Infectious Diseases* 19.1 (2019), p. 963. DOI: 10.1186/s12879-019-4566-2. URL: https://doi.org/10.1186/s12879-019-4566-2.

[14] Amir S. Siraj et al. "Spatiotemporal incidence of Zika and associated environmental drivers for the 2015-2016 epidemic in Colombia". In: *Scientific Data* 5.1 (2018), p. 180073. DOI: 10.1038/sdata.2018.73. URL: https://doi.org/10.1038/sdata.2018.73.