A Spatio-temporal Approach to Short-Term Prediction of Visceral Leishmaniasis Diagnoses in India

E. S. Nightingale*[1], L. A. C. Chapman[1], S. Srikantiah[2], S. Subramanian[3], P. Jambulingam[3], J. Bracher[4], M. M. Cameron[5], G. F. Medley[1]

1 Centre for Mathematical Modelling of Infectious Disease and Department of Global Health and Development, London School of Hygiene and Tropical Medicine, London, UK

2 CARE India, Patna, Bihar, India

3 Vector Control Research Centre, Puducherry, Chennai, India

4 Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Zurich, Switzerland

5 Department of Disease Control, London School of Hygiene and Tropical Medicine, London, UK

*Corresponding author Email: Emily.Nightingale@lshtm.ac.uk

Abstract

Background

The elimination programme for visceral leishmaniasis (VL) in India has seen great progress, with total cases decreasing by over 80% since 2010 and many blocks now reporting zero cases from year to year. Prompt diagnosis and treatment is critical to continue progress and avoid epidemics in the increasingly susceptible population. Short-term forecasts could be used to highlight anomalies in incidence and support health service logistics. The model which best fits the data is not necessarily most useful for prediction, yet little empirical work has been done to investigate the balance between fit and predictive performance.

Methodology/Principal Findings

We developed statistical models of monthly VL case counts at block level. By evaluating a set of randomly-generated models, we found that fit and one-month-ahead prediction were strongly correlated and that rolling updates to model parameters as data accrued were not crucial for accurate prediction. The final model incorporated auto-regression over four months, spatial correlation between neighbouring blocks, and seasonality. Ninety-four percent of 10-90% prediction intervals from this model captured the observed count during a 24-month test period. Comparison of one-, three- and four-month-ahead predictions from the final model fit demonstrated that a longer time horizon yielded only a small sacrifice in predictive power for the vast majority of blocks.

Conclusions/Significance

The model developed is informed by routinely-collected surveillance data as it accumulates, and predictions are sufficiently accurate and precise to be useful. Such forecasts could, for example, be used to guide stock requirements for rapid diagnostic tests and drugs. More comprehensive data on factors thought to influence geographic variation in VL burden could be incorporated, and might better explain the heterogeneity between blocks and improve uniformity of predictive performance. Integration of the approach in the management of the VL programme would be an important step to ensuring continued successful control.

Author summary

This paper demonstrates a statistical modelling approach for forecasting of monthly visceral leishmaniasis (VL) incidence at block level in India, which could be used to tailor control efforts according to local estimates and monitor deviations from the currently decreasing trend. By fitting a variety of models to four years of historical data and assessing predictions within a further 24-month test period, we found that the model which best fit the observed data also showed the best predictive performance, and predictive accuracy was maintained when making rolling predictions

up to four months ahead of the observed data. Since there is a two-month delay between reporting and processing of the data, predictive power more than three months ahead of current data is crucial to make forecasts which can feasibly be acted upon. Some heterogeneity remains in predictive power across the study region which could potentially be improved using unit-specific data on factors believed to be associated with reported VL incidence (e.g. age distribution, socio-economic status and climate).

1 Introduction

1.1 Visceral leishmaniasis in India

The short-term forecasting of diseases targeted for elimination can be a important management tool. Visceral leishmaniasis (VL) is the acute disease caused by *Leishmania donovani*, which is transmitted through infected female *Phlebotomus argentipes* sandflies. In India, the burden of disease is largely contained within the four northeastern states of Bihar, Jharkhand, Uttar Pradesh and West Bengal, with the rural state of Bihar most broadly affected [1–3].

Incidence of VL in India has decreased substantially since the initiation of the regional Kala-Azar Elimination Programme (KEP), which aims to tackle the disease across the Indian subcontinent through enhanced case detection and treatment and reduction of vector density [4]. As a result, reported cases have fallen from 29,000 in 2010 to less than 5,000 in 2018 [3, 4]. The overall target of the programme is to reduce incidence to less than 1 case/10,000 people/year within each "block". Blocks are administrative sub-divisions of a district with population sizes varying from twenty thousand to several million, depending on geographic area and the proportion of urban and rural habitation. As a consequence, the target equates to an absolute total of between three and two hundred cases per year. To support the elimination effort, data are reported to a central repository (Kala-Azar Management Information System, KA-MIS) to construct line lists including the date and location of every diagnosed case.

Despite the overall decrease in incidence, there is considerable heterogeneity

between blocks (Fig. 1). In some blocks cases are now few and far between, while others remain substantially affected from year to year. The combination of the decrease and the heterogeneity raises the need for a more targeted approach; the finite resources available must be distributed efficiently to continue progress. Additionally, history has shown that VL has the potential to develop into large epidemics [5–7] and hence it is important that localised pockets of incidence are not overlooked. Intervention when incidence is low is required to prevent the trajectory from turning upwards again, as cycles of VL incidence appear to occur with a frequency of 10-20 years [8].

Fig 1. Estimated incidence per 10,000 population per block in 2018, for Bihar and the four endemic districts of Jharkhand (Dumka, Godda, Sahibganj and Pakur). Incidence is estimated according to reported cases in KA-MIS with diagnosis date in between 01/01/2018 and 31/12/2018 and block populations projected from the 2011 census according to decadal, block-level growth rates [9]. Black lines indicate block boundaries. The affected blocks of Jharkhand on average have much higher incidence than Bihar and can be seen in the bottom right of the map. Blocks marked grey had no reported cases during the study period.

The primary aim of this paper is to ascertain the potential utility of predictions based solely on routinely-collected surveillance data, within a ready-made, rapid and relatively easy-to-use framework. Such predictions could serve two purposes; firstly to support logistics, for example in setting minimum stock levels of rapid diagnostic tests and drugs, and secondly to provide an early warning if the number of cases starts to resurge. For this modelling framework to be useful to the elimination programme, it is essential that its predictions are sufficiently accurate. Hence we make predictive accuracy of the forecasting approach the focus of the model selection.

1.2 Forecasting and spatio-temporal analysis

There have been many attempts at forecasting the various forms of leishmaniasis across the three affected continents. Lewnard et al. (2014) [10] employ a seasonal ARIMA model to predict cutaneous leishmaniasis in Brazil, incorporating meterological data and evaluating one, two and three month ahead forecasts. More recently, Li et al. used an extended ARIMA model to predict incidence in Kashgar prefecture, China [11]. However, neither of these attempts to capture spatial variation. Epidemiological data, in particular regarding infectious disease, are often *both*

temporally and spatially correlated. That is to say, as well as incidence at one point in time being related to incidence in the past, incidence in one area is also related to incidence in nearby areas. Mapping reported VL incidence in India at the block level demonstrates the presence of spatial correlation (Fig. 1), with concentrated regions of high incidence appearing in East Bihar and Jharkhand. This could be due to similar geographic and demographic characteristics of neighbouring blocks, or the spread of infection by regular population movement. The latter can induce a spatio-temporal pattern in which pockets of high incidence appear to "step" between neighbouring blocks. The seasonal cycle of incidence and overall decreasing trend (Fig. 2) are clearly evident in aggregated case counts.

Fig 2. Total monthly reported cases across the study region. The annual cycle (peaking between January and April) and overall decreasing trend are clear at this aggregate level.

Several statistical approaches have been developed to model count data in space and time. These methods have been largely developed and used for understanding the drivers of patterns, often incorporating additional covariate information describing climate, geography or demography [12, 13] Dewan et al. [14] employ scanning techniques for a regional analysis solely of case data, but do not utilise the approach for prediction. Paixão-Seva et al. (2017) [15] simultaneously model the infected human, vector and dog populations in relation to landscape, climatic and economic factors, and in particular use proximity to a highway and gas pipeline as indicators of human movement. Where aetiology is not the focus, analyses often incorporate GPS locations of cases to identify hotspots and predict disease spread at a local village or household level [16], or across health facilities [17] .

In the case of VL on the Indian subcontinent, environmental data are difficult to obtain in real-time at a sufficient spatial and temporal scale for forecasting purposes, and GPS data have not been routinely or uniformly collected across the affected region. As such, statistical approaches to spatio-temporal analysis have been broadly limited to specific study regions within which additional data were collected [18]. Predictions on a regional level have so far been the remit of transmission dynamic modelling [19]. We aim to make use of the reliable and near-complete date and area data within the KA-MIS system, for the whole state of Bihar and the affected region

of Jharkhand, to understand how well future cases could be predicted solely from the surveillance data of previous cases. As far as we are aware, no previous attempt has been made to forecast VL at this spatial scale and with this level of coverage for the Indian endemic region.

Often the model which best fits observed data is selected for forecasting, yet goodness of fit does not guarantee predictive power. We therefore also investigate the relationship between the fit and predictive power.

1.3 Model Framework

A natural modelling approach is to consider the cases in each month in each block as a function of cases in the previous month and in neighbouring blocks. A model framework developed in [20, 21] has been applied previously for modelling cutaneous leishmaniasis in Afghanistan [22]. This framework decomposes the distribution of counts at each point in space and time into three components (auto-regressive, neighbourhood and endemic):

- **Auto-regressive (AR)** *The contribution of previous incidence in the same block to current incidence. A choice must be made about time period of previous incidence considered (i.e. the number of months).*
- **Neighbourhood (NE)** *The contribution of previous incidence in surrounding blocks to current incidence. A choice must be made about both the time period and spatial extent considered (i.e. neighbours, neighbours of neighbours etc.), with indirect neighbours assigned decaying weights, for example, according to a power law.*
- **Endemic (END)** *A function describing the intrinsic incidence related to block factors (such as geography or demography) or seasonality.*

The sum of these components forms the mean structure for a negative binomial distribution used to model the count in each block and month. The epidemic component consists of both auto-regression and spatial/spatio-temporal regression. The maximum distance in space or time at which we assume one block-and-month count affects another is referred to as the maximum spatial or temporal *lag*. The

endemic component attempts to explain any remaining variation, potentially due to overall temporal trends, population size and other unit-specific factors.

In addition to the genuine epidemiology of VL, there is an intermediary process of detection and reporting which contributes to the distribution of case counts. A new case in a previously unaffected area triggers active case detection (ACD) which continues for twelve months, therefore contributing to the pattern of temporal correlation. In other words, one case is likely to be promptly followed by more cases - not only because of transmission but also as a result of increased, localised detection effort. We therefore explored a flexible, distributed lag structure [23] which extends the range of spatio-temporal interaction by allowing incidence over multiple previous months to contribute to both the auto-regressive and spatial elements. The selection of an optimal lag length has been investigated for distributed lag models in one dimension (i.e. time alone) [24], but the impact of introducing a spatial component has not been thoroughly discussed. A strong interdependence between the autoregressive and neighbourhood components is introduced by simultaneously incorporating past information from the same block and the neighbourhood of that block in a distributed lag model; each block affects subsequent incidence in its neighbours, which in turn affects subsequent incidence in the original block. We apply a semi-systematic approach which attempts to optimise the temporal and spatial lags simultaneously such that one does not mask the effect of the other.

1.4 Evaluation of forecasts

The three components described in the previous section have arbitrary complexity and lead to a large number of candidate models. A key issue is therefore to identify the best-fitting model, or a set of well-fitting models, and to assess to which degree good in-sample (or retrodiction) performance translates to out-of-sample forecasting performance. In-sample performance is widely assessed via the Akaike information criterion (AIC). The AIC balances the model fit and complexity, and has been recommended for model selection for prediction purposes [25]. To assess performance of probabilistic forecasts it is standard to use proper scoring rules [21, 26–29], which offer more detailed scrutiny of the prediction than measures of absolute or squared

error (as used, for example, in [30]) by taking into account the whole predicted distribution. In fact, the ranked probability score (RPS) can be considered a generalisation of absolute error, to which it reduces if the forecast distribution consists of a single point. Proper scoring rules measure simultaneously the calibration and sharpness of forecast distributions; they capture the model's ability to predict both accurately and precisely but also to identify its own uncertainty in that prediction [28]. With a well-calibrated model the observed values should appear as having come from the predicted distribution at that point, and we want as precise or sharp a predicted distribution as possible while maintaining that calibration. In contrast, the mean absolute error for example only evaluates how well the central tendency of predictions aligns with the observations. We utilise the ranked probability score (RPS) [26] averaged over all predicted time points (502 blocks * 24 months, so 12048 test predictions), which for a predictive distribution P and an observation x is defined as

$$\overline{\mathrm{RPS}}(P, x) = \sum_{k=0}^{\infty} [F_P(k) - \mathbb{1}(x \leq k)]^2 \tag{1}$$

Here, F_P is the cumulative distribution function of P and 1 is the indicator function. The RPS thus compares the cumulative distribution function of P to that of an "ideal" forecast with all probability mass assigned to the observed outcome x. We use this score rather than the logarithmic score as it is considered more robust [31], and we wish to assign some credit to forecasts near the observed value. The score is negatively oriented, meaning that smaller values are better.

Calibration can in addition be assessed using probability integral transform (PIT) histograms. The PIT histogram shows the empirical distribution of $F_{P;i(x_i)}$ for a set of independent forecasts $i = 1, ..., I$. We here use an adapted version for count data suggested by Czado et al [26]. If the forecasts are calibrated, the histogram should be approximately uniform. U and inverse U-shaped PIT histograms indicate that the forecasts imply too little or too much variability, respectively.

A closely-related summary measure which is easy to communicate are empirical coverage probabilities [31]. We will provide coverage probabilities of central 50% and 80% prediction intervals (reaching from the 25% to 75% and 10% to the 90% quantiles of the predictive distribution, respectively). For a calibrated forecast, the empirical

coverage probabilities should be close to the nominal levels. However, in the context of sparse, low counts the discreteness of the data often prevents achieving exactly the nominal coverage level. Prediction intervals can then either be slightly conservative (too high coverage), which is usually preferred in practice, or slightly liberal.

Our hypothesis is that models constructed with the *surveillance* framework to accommodate spatio-temporal correlation in disease incidence can provide significantly more accurate (in terms of sharpness and calibration) predictions than a purely parameter-driven (i.e. independent of history and spatial context) model with overall mean and linear time trend. Initially, we examine and discuss the relationship between model complexity, its ability to describe past data (i.e. its fit) and its ability to predict the next month. We then apply this understanding to select an optimal model for prediction with a semi-systematic approach, before comparing its predictive ability for different time horizons.

2 Materials and methods

2.1 Data

Access to the KA-MIS database of VL cases was provided by the National Vector Borne Disease Control Programme (NVBDCP) and facilitated by CARE India. Individual case records were downloaded for Bihar and Jharkhand, restricted to diagnosis date between 01/01/2013 and 31/12/2018 and then aggregated by block and diagnosis month. This gave reported case counts for 441 blocks. The KA-MIS data were merged with data from the 2011 census [9] (compiled by CARE India) for the two states to produce the final data set, including endemic blocks which had no reported cases during the study period and hence did not appear in KA-MIS. Because we incorporate spatial correlation into the model, it is necessary to not have "holes" of missing data in the map. For individual blocks within the assumed "endemic" region without any reported cases in certain months, case counts were assumed to be "true zeros" since detection efforts should be consistent with the affected neighbouring blocks. The time series for these blocks were imputed with zeros and therefore contributed to the fit of the model. Four entire districts of Bihar, at the edge of the

"endemic" region, (Gaya, Jamui, Kaimur and Rohtas) had no reported cases during the period, and were excluded from the analysis.

The final analysis data set included 502 blocks across 38 districts of Bihar and Jharkhand over 72 months.

2.2 Model Structure

Due to considerable temporal variation in incidence within blocks, as a result of detection effort and cases arising in "clumps", the block-level monthly case counts are widely dispersed. A negative binomial distribution was therefore used to model the block-level case counts throughout.

All models fitted conform to the same negative binomial structure for case counts Y_{it} given previous incidence:

$$Y_{it} \mid \text{past} \sim \text{NegBin}(\mu_{it}, \psi_i) \tag{2}$$

$$\mu_{it} = \underbrace{\lambda_t \sum_{q=1}^{Q} u_q Y_{i,t-q}}_{\text{AR}} + \underbrace{\phi_t \sum_{j \neq i} \sum_{q=1}^{Q} w_{ij} u_q Y_{j,t-q}}_{\text{NE}} + \underbrace{\nu_t e_{it}}_{\text{END}}. \tag{3}$$

where Y_{it} denotes the reported case count in block i in month t with population e_{it}, neighbourhood weights w_{ij} for neighbours j of block i, and overdispersion parameter $\psi_i > 0$ such that $\text{Var}(Y_{it}) = \mu_{it}(1 + \psi_i \mu_{it})$. Normalised weights u_q for distributed lags $q = 1, ..., Q$ are defined according to a scalar parameter p which is estimated from the data.

$$u_q^0 = p(1-p)^{q-1}, \quad u_q = \frac{u_q^0}{\sum_{q=1}^{Q} u_q^0} \tag{4}$$

The log-transformed parameter of each model component is then defined by a linear regression on any relevant covariates, $\mathbf{X}_{it}$; in this case we consider time with sine and cosine terms to replicate seasonal waves.

$$\log(\lambda_t) = \boldsymbol{\beta}^{\lambda} \mathbf{X}_{it}^{\lambda}, \tag{5}$$

$$\log(\phi_t) = \boldsymbol{\beta}^{\phi} \mathbf{X}_{it}^{\phi}, \tag{6}$$

$$\log(\nu_t) = \boldsymbol{\beta}^{\nu} \mathbf{X}_{it}^{\nu}, \tag{7}$$

where $\boldsymbol{\beta}$ are the regression coefficients.

All models were fit using the R package *surveillance* [32] and its extension *hhh4addon* [33] in R version 3.5.1 (2018-07-02) [34].

Investigating fit and prediction

Thirty random models were drawn from the set of possible formulations (where all three of the endemic-epidemic components are included in some form) and compared on the metrics of interest. This informed the subsequent selection process for the final prediction model.

Code used to produce the results in this paper is available from `https://github.com/esnightingale/VL_prediction_paper`, along with a simulated version of the dataset from the final selected model.

2.3 Model selection

During the selection process, all models were fit to the subset of months 5 to 48 in order to make comparisons between temporal lags up to four months. The remaining 24 months were then predicted sequentially in a "one-step-ahead" (OSA) approach to assess predictive power (as was applied in [10]), either with rolling updates to the fit (incorporating each month's data into parameter estimates to predict the next) or without (using only the training set of data for all predictions) [22, 26]. The average RPS of these predictions served as the primary criteria for model selection, comparing by permutation test between models of increasing complexity with a significance cut-off at 0.001. At the same time, average RPS was compared to AIC from the model's training period fit to assess the relationship between fit to the "observed" data and future prediction.

The following elements were considered for inclusion in the model:

- Log of population density as a covariate in the endemic component, in place of population fraction offset.
- Seasonal variation and linear trend within the coefficients of all three

components, serving to vary the relative strength of each component over time.

- Distributed temporal lags up to 4 months, with decaying weights according to a geometric distribution.

- Spatial lags up to maximum of 7th order neighbours, with weights decaying according to a power law ($w_{ij} = o_{ij}^{-d}$, where o_{ij} is the neighbourhood order of blocks i and j, and the decay exponent d is to be estimated).

- Intercept of log population density in the neighbourhood component (*Gravity Law*), to reflect that blocks of high population density may be more strongly influenced by their neighbours due to migration.

- District and state-specific dispersion, allowing the variation in incidence to differ between spatial units.

It was not feasible to allow a block-specific dispersion parameter since many blocks had too few cases to obtain stable estimates.

Finer details of the model selection process are included in Appendix A.

2.3.1 Empirical Coverage Probabilities

As an alternative measure of prediction utility, we calculated the empirical coverage of prediction intervals produced by each model, with respect to the observed counts. This describes the proportion of points in the test period for which the observed count fell within the middle 50% or 80% of the predicted distribution. For an ideal forecast the empirical coverage will match the nominal level. An empirical coverage probability cannot be considered "strictly proper" [21, 26, 31], as the RPS score is, and hence does not favour sharpness in addition to calibration. However, a high coverage quantile interval may provide useful lower and upper bounds for expected incidence. For more detail see Appendix A.

2.3.2 Longer Prediction Horizons

For the final model, further predictions were calculated based on a rolling window of three and four months. As with the rolling OSA approach, the model was initially fit to the training set (months $1, \ldots, t$) and this fit used to predict month $t + 3$. The

model was then updated with the data from $t+1$ in order to predict $t+4$, and so on, in a similar fashion to Lewnard et al. [10]. The RPS of one, three and four month ahead predictions were compared to assess the loss in accuracy with a longer time horizon.

3 Results

Preliminary analyses of dispersion and exploration of temporal lags are described in Appendix B.

3.1 Random model assessment

According to the thirty random models drawn, fit and prediction were found to be strongly correlated (Fig. 3A). Predictions were calculated based on either a rolling fit (incorporating each month's data into parameter estimates to predict the next month) or fixed fit (using parameters fit to the training set only for all predictions). The scores for both prediction approaches were very similar for most models, suggesting that the processes defined in these models are consistent over time and hence the quality of prediction does not depend on regular model updates (Fig. 3B). This is noteworthy since in practice it may not be possible to update the fits on such a regular basis. Selecting the model based on RPS of predictions from a fixed model fit would best reflect the constraints of reality and be the more conservative approach.

Fig 3. Comparison of predictive performance and model fit, and predictive performance for training period fit and rolling fit updates, for models with randomly selected components. (A) AIC versus RPS for 30 randomly selected models. AIC is calculated from the fit to the training period only (months 13 to 48) and RPS from one-step-ahead predictions (months 49 to 72) based on the same fit. According to this random sample, fit and prediction are strongly correlated; the model which fits best to the observed data produces the best one-step-ahead predictions. (B) RPS of predictions based on the fixed training set fit versus rolling fit updates. Predictive power is very similar between the two prediction approaches.

3.2 Model selection

As was found with the random model set, the final selected model which demonstrated the highest predictive power as measured by RPS also achieved the closest fit to existing data. Initially, no more than two distributed AR lags could be added to the model without yielding evidence of miscalibration in the predictions. However, once

the neighbourhood component was added in the third stage of selection, increasing the AR lags to four months significantly improved both AIC and RPS with no evidence of miscalibration. At this point the endemic linear trend lost significance and therefore was removed in subsequent models. The AIC, RPS and empirical coverage probabilities for all models considered in the selection process are shown in Fig. 4. Fit and prediction metrics for all models are given in Table S1 and PIT histograms for the models selected at each stage are compared in Fig. S3.

Fig 4. Measures of fit and predictive power throughout the model selection process. Figures illustrate the models tested in chronological order from left to right, with each stage indicated by a different colour. Models were selected at each stage based on the biggest reduction in RPS, subject to calibration; these are identified by hollow points, and the final selected model by a star. For the two variants on the coverage probability, average quantile interval width (representing uncertainty in the predicted case count) is shown on the right axis and by the grey dashed line. Interval width is determined by the count at the upper quantile minus the count at the lower, hence an interval width of two covers three possible count values (e.g. 2, 3, 4).

We found that as RPS and AIC were improved, the empirical coverage probabilities of prediction intervals were increased far beyond their nominal level. With the final model (Model no. 42), only 5.4% (652/12048) of observations fell outside the 10-90% interval, with an average interval width of just three possible case counts. This predicted distribution is much more conservative in its coverage than a simple linear trend model (coverage 10-90% = 0.905) but attains substantially better fit and RPS, suggesting that more of the improvement comes in the form of calibration. The conservative 90% predicted quantile provides a reliable upper limit for the next month's incidence, to which a management plan could be defined accordingly. The 25-75% prediction interval was found to be of limited use since, with very low counts across the majority of the region, this interval often consists of only a single value. The median would be a more interpretable value to report.

3.3 Final model

The final model consists of a negative binomial distribution with a single dispersion parameter and the following mean structure:

$$\mu_{it} = \lambda_{it} \sum_{q=1}^{4} u_q Y_{i,t-q} + \phi_{it} \sum_{j \neq i} \sum_{q=1}^{4} w_{ij} u_q Y_{j,t-q} + e_{it} \nu_{it} \tag{8}$$

$$\log(\nu_{it}) = \alpha^{\nu} \tag{9}$$

$$\log(\lambda_{it}) = \alpha^{\lambda} + \gamma_1^{\lambda} \sin\left(\frac{2\pi}{12}t\right) + \delta_1^{\lambda} \cos\left(\frac{2\pi}{12}t\right) \tag{10}$$

$$\log(\phi_{it}) = \alpha^{\phi} + \gamma_1^{\phi} \sin\left(\frac{2\pi}{12}t\right) + \delta_1^{\phi} \cos\left(\frac{2\pi}{12}t\right) \tag{11}$$

The model fit is dominated by auto-regression; the majority of information with which to predict the current month comes from incidence in the previous four months, with seasonally-varying strength. Since the contribution of each component is modelled on a log scale these parameters have a multiplicative effect, hence the range of the seasonal AR component (approx. [0.6, 0.8]; see supplementary Fig. S4) indicates that each month's count is expected to be a certain fraction of the weighted average of the counts over the last four months. This occurs over all blocks and therefore amounts to an overall decreasing trend. After accounting for auto-regression, it was found that the neighbourhood effect did not extend beyond directly bordering blocks with respect to prediction. Seasonality within this component also serves to vary the magnitude of the effect throughout the year.

The contribution of an endemic trend was found to be negligible, reflecting the lack of homogeneity across blocks, and was therefore not included; the reduction in total incidence comes entirely from each block's autoregressive pattern. Block-specific covariate data (e.g. relating to socio-economic or geographic features of the area) would contribute to this component and potentially reveal associations which are consistent across blocks. Random intercepts were tested in the endemic component to capture unexplained block variation, yet did not improve predictive power in a basic model and caused convergence issues in more complex, distributed-lag models.

The relative contributions of the three model components are illustrated for the four blocks with highest average monthly incidence (Gopikandar, Kathikund, Boarijor and Sundarpahari) in Fig. 5.

3.3.1 Predictive performance

The final model achieved an overall $\overline{\text{RPS}}$ for one-step-ahead prediction of 0.420, 36% lower than the null (non-spatial and non-autoregressive) model and 8% lower than the

Fig 5. Model fit for the four blocks with highest average monthly incidence (Gopikandar, Kathikund, Boarijor, and Sundarpahari, all in Jharkhand). The observed case counts are indicated by black points and the coloured regions illustrate the relative contribution of the different model components. The contribution of the endemic component is negligible therefore barely visible. The fitted value from the model falls at the upper edge of the coloured region.

best non-spatial model, with individual block-wise averages ranging from 4.3×10^{-5} to 3.47. This equates to a mean absolute error of 0.58, a 30% reduction from the null model. That the RPS is lower than the MAE implies the probabilistic forecast is preferable to a simple point forecast.

Model selection was performed based on the model's *mean* RPS across all blocks and the whole test period but beneath this overall score is a broader distribution of scores for each block-month prediction, influenced by peaks, troughs and otherwise unusual incidence patterns. The histogram in Fig. 6 illustrates the distribution over blocks, demonstrating that the final model is able to predict accurately and precisely across the majority of the region, yet there is a small subset of blocks with more widely varying RPS. It should be noted that the overall performance of the model is strongly influenced by blocks with almost no incidence as these yield the very lowest scores. Similarly, there is some correlation between the blocks for which the model performs least well, and the blocks which have historically demonstrated the highest average incidence since higher counts are harder to predict than zeros or single cases. The blocks with the highest RPS also tend to exhibit sporadic patterns or have experienced sudden, sharp changes in incidence (potentially outbreaks) within the test period, which cannot be reproduced by a model primarily informed by an average of past incidence. Examples of these patterns are illustrated in Fig. S5.

Fig 6. Distribution of time-averaged ranked probability scores across all 502 blocks. Low values reflect accurate and precise prediction. The majority of blocks fall below 1 with a subset for which predictive power varies widely.

Pakur, Maheshpur, Boarijor and Sundarpahari in Jharkhand ($\overline{\text{RPS}} = 3.47$, 2.70, 2.58 and 2.58, resp.) experienced substantial jumps in incidence between May and July 2017, constituting differences of up to 27 cases from one month to the next. Paroo ($\overline{\text{RPS}} = 3.07$) showed a particularly erratic pattern of cases within the test period, with spikes of 21 and 19 cases separated by a few months of $\sim$5 cases and a

subsequent fall to just one case by December 2018. Incidence in Garkha has also been inconsistent and appeared to have been on the rise in recent years, until a similar fall at the end of 2018. It should be noted that additional case detection efforts in Jharkhand at the start of 2017 will likely have contributed substantially to the observed spikes at this time.

3.3.2 Three- and four-month-ahead prediction

For the final model, further predictions were calculated based on rolling windows of three and four months. Fig. 7 illustrates that the longer time window did not result in a substantial loss in predictive power, with block-wise RPS very similar for the majority of blocks. When compared over the same predicted months, the differences in $\overline{\text{RPS}}$ between one-month-ahead prediction and three-/four-month-ahead were found to be small but statistically significant (-0.024 and -0.028, resp.; $p < 0.0001$ for both). In terms of the empirical coverage, 85.4% of test period observations were captured in the middle 50% of the predicted distribution based on a three month window, and 85.7% with a four month window.

Fig 7. Time-averaged (over months 52-72 for comparability) RPS for three- (A) and four-month-ahead (B) predictions versus one-month-ahead. Scores are closely matched for the majority of blocks (where $\overline{\text{RPS}} < 1.5$) but the differences increase for blocks which are harder to predict.

Figs 8 and 9 illustrate the coverage of 45-55%, 25-75% and 10-90% prediction intervals for the block with the highest $\overline{\text{RPS}}$ of 3.47 (Pakur, Jharkhand) and a block with $\overline{\text{RPS}}$ of 1 (Bhagwanpur, Bihar). For Pakur, RPS is strongly influenced by the model's inability to match the spike in 2017, yet the incidence in surrounding months is well represented.

Fig 8. One-, three- and four-step-ahead predictions (solid white line) with 10-90%, 25-75% and 45-55% quantile intervals, for Pakur block in Jharkhand ($\overline{\text{RPS}} = 3.47$ for one-step-ahead over months 49-72). Observations which fall outside the outer prediction interval are indicated by a cross.

Fig 9. Corresponding predictions for Bhagwanpur block in Bihar ($\overline{\text{RPS}} = 1.00$).

4 Discussion

We have presented the evaluation of a predictive model of VL in Bihar and four endemic districts in Jharkhand, demonstrating a substantial (36% lower RPS) benefit from incorporating spatial and historical case information when compared to a non-spatial, linear trend model. We have empirically investigated the performance of different models on prediction performance rather than model fit and produced a statistical model that is capable of accurate forecasting. To the best of our knowledge, this is the first time the spatio-temporal correlation of incidence at block level across all the endemic districts of Bihar and Jharkhand has been quantified. Methods such as these can be an important tool for management of endemic diseases.

Given the lack of an effective vaccine and evidence that indoor residual spraying of insecticide fails to significantly reduce sandfly densities and VL incidence in sprayed villages [35, 36], rapid diagnosis and treatment is currently our best method of control. With a block-level estimate of the likely number of cases to arise over the next few months, local management teams could take steps to ensure they are prepared. For example, the 90% quantile of the predicted distribution could be used to inform block-specific minimum stock levels for drugs.

In practice, the prediction interval is constrained by the efficiency of the reporting process; the time taken to process diagnosis reports and input the information into the database sets a minimum horizon at which predictions would be genuinely prospective and therefore of practical use. In this paper we have assumed a delay of two months until a month's data can be considered complete, which would necessitate making predictions at least three months ahead of that point. However, conservative predictions based on preliminary month totals would still likely be of use to the programme.

We have demonstrated here that rolling three-month-ahead predictions are a reasonable approximation to one-month-ahead, but confidence is sacrificed for a minority of blocks as the time horizon is increased. There is a need for discussion with local disease management teams to determine the optimal balance between practicality and uncertainty. Moreover, the way in which we quantify the accuracy and utility of predictions would benefit from some public health insight; it is highly

likely that over- and under-estimation would need to be weighted differently, which may alter which model is deemed preferable. Ideally, the model structure would have been optimised according to predictive power on this slightly longer time horizon, but this is not a trivial task and was deemed beyond the scope of this paper.

There are also potential issues with movement of VL cases across international borders; in particular, the international boundary with Nepal cuts through a VL endemic area, artificially removing some aspects of spatial correlation. Ideally, we would take a regional perspective and also include areas in neighbouring states that have some more sporadic VL reporting.

It could be argued that the block-level is too coarse a spatial scale for modelling the spread of an infectious disease. Outbreaks of VL occur on a smaller spatial and temporal scale than has been applied here, therefore cannot be anticipated by this model. The transmission dynamic models which are usually employed for this type of problem can be defined on a village, household or even individual level [37], yet this more detailed picture demands many more assumptions which are difficult to justify in this context. The sparseness of cases at this point in the elimination process also means that aggregation at a finer temporal scale might lead to issues with parameter estimation. The block is the unit at which control efforts are co-ordinated, disease burden is monitored, and control targets are set, therefore predictions at this level could prove to be a worthwhile compromise while more realistic transmission models are out of reach. With more detailed location data, the spread of disease can be modelled as a point process at the village or household level, potentially giving insight into the size and movement of disease clusters or "hot-spots" over time. This technique has previously been applied to the case of VL [38] and may be possible to extend to a larger study region in the near future, following a recent effort to collect GPS co-ordinates of affected villages across Bihar.

In this case the best-fitting model was found to be the best-predicting model. The similarity of prediction and fitting results perhaps reflects the continuity of the processes creating the data. However, consideration of predictive power across the whole range of possible values was key to determining an optimal temporal lag length for short-term prediction. Fit and overall predictive power favoured a high number of lags in order to best capture the spatio-temporal correlation between neighbouring

block counts, which appears to contribute to prediction of sudden changes in incidence. However, auto-regression is the dominant model component and appears to be captured by lags up to four months. It would be preferable to specify a different lag length for the auto-regressive and spatial components but this is not currently implemented in the *surveillance* framework. By inspection of PIT histograms, we were able to select the lag length which balanced overall predictive power with capacity to predict at the upper end of the range.

The model selection approach taken in this analysis is semi-systematic; it was not feasible to assess every possible combination of model components. Therefore we aimed to home in on a suitable model by adding components which gave the biggest improvement in predictive performance out of a range of likely options. It was found that once the major components were included in some form, further adjustment largely had the effect of redistributing the variation attributed to each component and did not substantially alter fit or prediction. There is only so much information within the time series of cases to feed the model, so predictive power quickly reaches a limit.

The analysis presented here aims to demonstrate the best that can be done with the minimal information routinely collected by the current programme, but there is evidence that this model still cannot fully account for the heterogeneity in incidence across the region. The lack of geographic and/or demographic covariates beyond population size means that the endemic component in this model is negligible; almost all our information comes from the spatio-temporal correlations, underlining the need for up-to-date data in order to make accurate predictions. Associations between VL incidence and, for example, age and socio-economic quintiles have been demonstrated [18, 39], which may give rise to varied endemic patterns at the block level. This unknown variation could in theory be quantified by random effects within this model framework, but convergence issues (likely due to the large number of zero-counts) made this infeasible in practice.

There is clearly a limitation of fitting such a model over a large number of highly heterogeneous units with minimal unit-specific information. Model selection was performed based on an average score over all blocks and time points for which predictions were made; a model is therefore chosen which predicts well overall, but in doing so sacrifices predictive power for a minority of blocks which do not follow the

general trend. Zero counts dominate over all time and space, and the variance of the negative binomial distribution with a universal dispersion parameter is still too restrictive to account for blocks with the highest counts. It is in these areas where additional information on potential predictors of incidence could prove most valuable.

The variation in case counts may be better explained by a zero-inflated process, and the extent of zero-inflation will likely become more prominent as elimination is approached. Bayesian hierarchical models can be used to distinguish sources of variation at different levels and have the benefit of accommodating any informal or incomplete understanding of the transmission process within prior distributions for model parameters. These models have until recently been commonly implemented using Markov Chain Monte Carlo (MCMC) [40], which is computationally intensive for data rich in both space and time. They are however becoming increasingly accessible as a tool for inference and prediction, thanks to user-friendly wrappers which take advantage of fast computation using Integrated Nested Laplace Approximations (INLA) [41]. We hope to explore this approach in future work.

Conclusion

We have demonstrated a framework for forecasting VL incidence at subdistrict level in India which achieves good predictive performance based on the available routinely collected surveillance data. This framework could be used to make short-term forecasts to provide an early indication of where case numbers are higher (or lower) than expected and to support the logistics of the elimination programme.

Acknowledgments

The authors gratefully acknowledge the support and permission of Dr Dhingra, Director of the National Vector-Borne Disease Control Programme (NVBDCP), Government of India, to use the KAMIS data.

We thank Joy Bindroo and the staff of CARE India, Patna, for their help with data access and queries. We are also very grateful to Tim Pollington of the University of Warwick for helpful input on the *surveillance* package and model selection.

Data Availability

The data from the Kala-Azar Management Information System (KA-MIS) underlying the results in this manuscript cannot be shared publicly because of patient confidentiality and privacy concerns. KA-MIS data are property of the National Vector-Borne Disease Control Programme (NVBDCP, Govt of India), and are managed by CARE India. The data are available from NVBDCP (Dr Neerah Dhingra - dhingradr@hotmail.com) for researchers who meet the criteria for access to confidential data. A simulated version of the dataset used in this manuscript is available at `https://github.com/esnightingale/VL_prediction_paper`

Funding Statement

This study was supported by the Bill and Melinda Gates Foundation (https://www.gatesfoundation.org/) through the SPEAK India consortium [OPP1183986] (ESN, LACC, SS, SS, PJ, MMC, GFM). The views, opinions, assumptions or any other information set out in this article are solely those of the authors and should not be attributed to the funders or any person connected with the funders.

Ethical Approval

Ethical clearance was granted by the Observational/Interventions Research Ethics Committee at LSHTM (ref: 14674), subject to local approval. Local approval to use this data was granted by Dr Neeraj Dhingra, director of the National Vector Borne Disease Control Programme (GoI). Individual consent was not required as all data were analysed anonymously.

Competing Interests

The authors have declared that no competing interests exist.

References

1. Alvar J, Vélez ID, Bern C, Herrero M, Desjeux P, Cano J, et al. Leishmaniasis Worldwide and Global Estimates of Its Incidence. PLoS ONE. 2012;7(5):e35671. doi:10.1371/journal.pone.0035671.

2. Ready P. Epidemiology of visceral leishmaniasis. Clinical Epidemiology. 2014;6(1):147–154. doi:10.2147/CLEP.S44267.

3. Singh NS, Singh DP. A Review on Major Risk Factors and Current Status of Visceral Leishmaniasis in North India. American Journal of Entomology. 2019;3(1):6–14. doi:10.11648/j.aje.20190301.12.

4. National Vector Borne Disease Control Programme. Kala-Azar Situation in India; 2018. Available from: `https://www.nvbdcp.gov.in/index4.php?lang=1{&}level=0{&}linkid=467{&}lid=3750`.

5. Dye C, Wolpert DM. Earthquakes, influenza and cycles of Indian kala-azar. Transactions of the Royal Society of Tropical Medicine and Hygiene. 1988;82:843–850. doi:10.1016/0035-9203(88)90013-2.

6. Bora D. Epidemiology of visceral leishmaniasis in India. The National Medical Journal of India. 1999;12(2):62–68.

7. Courtenay O, Peters NC, Rogers ME, Bern C. Combining epidemiology with basic biology of sand flies, parasites, and hosts to inform leishmaniasis transmission dynamics and control. PLoS Pathogens. 2017;13(10):e1006571.

8. Rijal S, Sundar S, Mondal D, Das P, Alvar J, Boelaert M. Eliminating visceral leishmaniasis in South Asia: the road ahead. BMJ (Clinical research ed). 2019;364:k5224. doi:10.1136/bmj.k5224.

9. Ministry of Home Affairs, Government of India. C.D. Block Wise Primary Census Abstract Data; 2011. Available from: http://censusindia.gov.in/pca/cdb_pca_census/cd_block.html.

10. Lewnard JA, Jirmanus L, Júnior NN, Machado PR, Glesby MJ, Ko AI, et al. Forecasting Temporal Dynamics of Cutaneous Leishmaniasis in Northeast Brazil. PLoS Neglected Tropical Diseases. 2014;8(10):e3283. doi:10.1371/journal.pntd.0003283.

11. Li HL, Zheng RJ, Zheng Q, Jiang W, Zhang XL, Wang WM, et al. Predicting the number of visceral leishmaniasis cases in Kashgar, Xinjiang, China using the ARIMA-EGARCH model. Asian Pacific Journal of Tropical Medicine. 2020;13(2):81–90. doi:10.4103/1995-7645.275416.

12. Lowe R, Bailey TC, Stephenson DB, Graham RJ, Coelho CAS, Sá Carvalho M, et al. Spatio-temporal modelling of climate-sensitive disease risk: Towards an early warning system for dengue in Brazil. Computers & Geosciences. 2011;37(3):371–381. doi:10.1016/J.CAGEO.2010.01.008.

13. Amro A. Epidemiology and spatiotemporal analysis of visceral leishmaniasis in Palestine from 1990 to 2017. International Journal of Infectious Diseases. 2020;90:206–212. doi:10.1016/j.ijid.2019.10.044.

14. Dewan A, Abdullah AYM, Shogib MRI, Karim R, Rahman MM. Exploring spatial and temporal patterns of visceral leishmaniasis in endemic areas of Bangladesh. Tropical Medicine and Health. 2017;45(1):29. doi:10.1186/s41182-017-0069-2.

15. Sevá AdP, Mao L, Galvis-Ovallos F, Tucker Lima JM, Valle D. Risk analysis and prediction of visceral leishmaniasis dispersion in São Paulo State, Brazil. PLoS Neglected Tropical Diseases. 2017;11(2):e0005353. doi:10.1371/journal.pntd.0005353.

16. Bhunia GS, Kesari S, Chatterjee N, Kumar V, Das P. Spatial and temporal variation and hotspot detection of kala-azar disease in Vaishali district (Bihar), India. BMC Infectious Diseases. 2013;13(1):64. doi:10.1186/1471-2334-13-64.

17. Godana AA, Mwalili SM, Orwa GO. Dynamic spatiotemporal modeling of the infected rate of visceral leishmaniasis in human in an endemic area of Amhara regional state, Ethiopia. PLoS ONE. 2019;14(3):e0212934. doi:10.1371/journal.pone.0212934.

18. Bulstra CA, Le Rutte EA, Malaviya P, Hasker EC, Coffeng LE, Picado A, et al. Visceral leishmaniasis: spatiotemporal heterogeneity and drivers underlying the hotspots in Muzaffarpur, Bihar, India. PLoS Neglected Tropical Diseases. 2018;12(12):e0006888.

19. Le Rutte EA, Chapman LAC, Coffeng LE, Jervis S, Hasker EC, Dwivedi S, et al. Elimination of visceral leishmaniasis in the Indian subcontinent: a comparison of predictions from three transmission models. Epidemics. 2017;18:67–80. doi:10.1016/j.epidem.2017.01.002.

20. Meyer S, Held L, Höhle M. Spatio-Temporal Analysis of Epidemic Phenomena Using the R Package surveillance. Journal of Statistical Software. 2017;77(11). doi:10.18637/jss.v077.i11.

21. Held L, Meyer S, Bracher J. Probabilistic forecasting in infectious disease epidemiology : the 13th Armitage lecture. Statistics in Medicine. 2017;36(22):3443–3460. doi:10.1002/sim.7363.

22. Adegboye OA, Adegboye M. Spatially correlated time series and ecological niche analysis of cutaneous leishmaniasis in Afghanistan. International Journal of Environmental Research and Public Health. 2017;14(3):309. doi:10.3390/ijerph14030309.

23. Bracher J, Held L. Endemic-epidemic models with discrete-time serial interval distributions for infectious disease prediction. arXiv e-prints. 2019; p. arXiv:1901.03090.

24. Furlan CPR, Diniz CAR, Franco M. Estimation of lag length in distributed lag models: A comparative study. Advanced Applied Statistics. 2010;17(2):127–142.

25. Hyndman RJ, Athanasopoulos G. Forecasting: principles and practice. 2nd ed. OTexts: Melbourne, Australia; 2018. Available from: `OTexts.com/fpp2`.

26. Czado C, Gneiting T, Held L. Predictive model assessment for count data. Biometrics. 2009;doi:10.1111/j.1541-0420.2009.01191.x.

27. Gneiting T, Katzfuss M. Probabilistic Forecasting. Annual Review of Statistics and Its Application. 2014;1(1):125–151. doi:10.1146/annurev-statistics-062713-085831.

28. Funk S, Camacho A, Kucharski AJ, Lowe R, Eggo RM, Edmunds WJ. Assessing the performance of real-time epidemic forecasts: A case study of ebola in the Western area region of sierra leone, 2014-15. PLoS Computational Biology. 2019;15(2):e1006785. doi:10.1371/journal.pcbi.1006785.

29. Lu J, Meyer S. Forecasting Flu Activity in the United States: Benchmarking an Endemic-Epidemic Beta Model. International Journal of Environmental Research and Public Health. 2020;17(4):1381. doi:10.3390/ijerph17041381.

30. Chaves LF, Pascual M. Comparing models for early warning systems of neglected tropical diseases. PLoS Neglected Tropical Diseases. 2007;1(1):e33. doi:10.1371/journal.pntd.0000033.

31. Gneiting T, Balabdaoui F, Raftery AE. Probabilistic forecasts, calibration and sharpness. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2007;69(2):243–268. doi:10.1111/j.1467-9868.2007.00587.x.

32. Höhle M, Meyer S, Paul M. Surveillance: Temporal and Spatio-Temporal Modeling and Monitoring of Epidemic Phenomena; 2016. Available from: `https://cran.r-project.org/package=surveillance`.

33. Bracher J. hhh4addon: Extensions to endemic-epidemic timeseries modeling from package surveillance; 2018. Available from: `https://github.com/jbracher/hhh4addon`.

34. R Core Team. R: A Language and Environment for Statistical Computing; 2018. Available from: `https://www.r-project.org/`.

35. Poche DM, Garlapati RB, Mukherjee S, Torres-Poche Z, Hasker E, Rahman T, et al. Bionomics of Phlebotomus argentipes in villages in Bihar, India with

insights into efficacy of IRS-based control measures. PLoS Neglected Tropical Diseases. 2018;12(1):1–20.

36. Picado A, Dash AP, Bhattacharya S, Boelaert M. Vector control interventions for Visceral Leishmaniasis elimination initiative in South Asia, 2005-2010. Indian Journal of Medical Research. 2012;136(1):22–31.

37. Chapman LAC, Jewell CP, Spencer SEF, Pellis L, Datta S, Chowdhury R, et al. The role of case proximity in transmission of visceral leishmaniasis in a highly endemic village in Bangladesh. PLOS Neglected Tropical Diseases. 2018;12(10). doi:10.1371/journal.pntd.0006453.

38. Mandal R, Kesari S, Kumar V, Das P. Trends in spatio-temporal dynamics of visceral leishmaniasis cases in a highly-endemic focus of Bihar, India: an investigation based on GIS tools. Parasites & vectors. 2018;11(1):220. doi:10.1186/s13071-018-2707-x.

39. Chapman LAC, Morgan ALK, Adams ER, Bern C, Medley GF, Hollingsworth TD. Age trends in asymptomatic and symptomatic Leishmania donovani infection in the Indian subcontinent: A review and analysis of data from diagnostic and epidemiological studies. PLOS Neglected Tropical Diseases. 2018;doi:10.1371/journal.pntd.0006803.

40. Jewell CP, Kypraios T, Neal P, Roberts GO. Bayesian analysis for emerging infectious diseases. Bayesian Analysis. 2009;4(3):465–496. doi:10.1214/09-BA417.

41. Blangiardo M, Cameletti M, Baio G, Rue H. Spatial and spatio-temporal models with R-INLA; 2013. Available from: `https://www.sciencedirect.com/science/article/pii/S1877584512000846`.

Supporting information

Appendix A

Model Selection

Starting with a basic, endemic-only model (including a population offset and linear trend in time), potential extensions of the three core components were added in turn and measures of fit and predictive power were calculated. The addition which yielded the best improvement in the RPS of OSA predictions, subject to calibration (p not less than 0.1 for test of calibration based on RPS), was selected and then all remaining options tested again. This process was repeated until no further extension of the model made a significant ($p < 0.001$) improvement to predictive power (as determined by a permutation test on the RPS). This stringent criterium was employed in order to prioritise simplicity over complexity. If at any point an individual model parameter lost significance, the element associated with this parameter was removed in subsequent models.

Empirical Coverage Probabilities

Again using a one-step-ahead approach, the 25th and 75th quantiles of the predicted distribution were calculated and a score of 0 or 1 assigned if the observed value fell inside or outside this quantile range respectively. This binary score was assigned for each block and each month in the test set, such that we could subsequently calculate a proportion of prediction intervals which did not capture the true count. Thus, the overall score, C, is given by

$$C = \frac{1}{n_i n_t} \sum_{i,t} \mathbb{1}[y_{it} \leq q_{i,t,0.25} | y_{it} \geq q_{i,t,0.75}] \tag{12}$$

where y_{it} is the observed count for block i at month t, n_i and n_t the total number of blocks and months respectively, and $q_{i,t,p}$ the p^{th} quantile of the predicted distribution. We also investigated such a score using 10th and 90th quantiles, to ascertain whether these could be used as approximate lower and upper bounds for case counts.

Appendix B

Preliminary analyses

Dispersion

District-specific dispersion parameters were investigated, but ultimately not considered a viable option to be included in the model. Four districts in particular (Aurangabad, Banka, Jehanabad and Nawada) demonstrate extended periods of zero incidence with occasional sporadic cases or large spikes, which lead to very large dispersion estimates for these districts and therefore unrealistically high predictions. See Fig. S1 for an illustration of these patterns. Due to the neighbourhood effect, these high predictions in turn influence the predictions of any bordering blocks. Changes in detection effort could go some way to explaining these unusual patterns, however it is also likely that such patterns will become more common as elimination is approached. This suggests that an alternative modelling strategy will become necessary as cases become more sparse in space and time.

Distributed temporal lags

By sequentially adding further distributed lags to the best-fitting single-lagged model, neither a clear minimum nor an "elbow" in RPS was attained up to twelve months. The weights assigned to each lag did not show a rapid "drop-off" as a result of a high estimated decay parameter, and months substantially far back in time were still assigned non-negligible weight. PIT histograms of predictions from these lagged models are included in 4. We found that adding higher orders of distributed lags consistently improved both predictive power and fit. This appears to contradict analysis of individual block time series which suggested significant auto-correlation no more than four months back in time. In the current form of "hhh4addon", it is not possible to specify a different temporal lag length within the AR and NE components (for example, to incorporate neighbouring incidence from further back in time than within-block). Therefore, the contribution of distributed lags to both components had to be considered and a balance had to be drawn. Comparing the PIT histograms of solely auto-regressive models, the very highest counts are vastly underestimated for all lag lengths. Since the highest values in each block often reflect sudden jumps they

cannot be captured by auto-regression; more information - potentially from the surrounding area - is required to anticipate them. Models with no auto-regression but which incorporate neighbouring incidence are better able to reach the highest counts but in doing so over-estimate the moderate-to-high range. It was concluded that beyond four months of lags the improvement in prediction was small enough to discount, and much longer lags were difficult to justify epidemiologically. Therefore only four months of lags were considered for the final model.

Supplementary Figures

Fig. S1 Districts with unusual incidence patterns resulting in inflated dispersion estimates.

Fig. S2 Probability integral transform (PIT) histograms for models with increasing orders of geometric lags from 1 to 12 months (left to right, top to bottom) in the autoregressive component. The final model selection process considered up to four lags.

Table S1 Fit and prediction metrics for selected model at each stage. The reported AIC is for the fit to training data only, and RPS is of predictions made without updating this fit (i.e. fixed instead of rolling). C2575 and C1090 refer to the coverage of 50% and 80% quantile intervals, respectively, alongside the average interval width in cases. Model no. 42 is the final model.

Stage	Model No.	END	AR	NE	Dispersion	No. parameters	AIC	RPS	Calibration (p-value)	C1090	Avg. width
0	1	offset + 1 + t			1	3	65412	0.657	<0.0001	0.095	2.243
1	2	offset + 1 + t + seas(∼1, S=1)			1	5	65227	0.654	<0.0001	0.090	2.330
1	3	offset + 1			1	2	65811	0.698	<0.0001	0.044	4.158
1	4	offset + 1 + t + logpopdens			1	4	65708	0.662	<0.0001	0.094	2.180
1	5	offset + 1 + t	AR(1)		1	4	57100	0.495	0.109	0.060	2.386
1	**6**	**offset + 1 + t**	**AR(1) + seas(∼1, S=1)**		**1**	**6**	**57058**	**0.493**	**0.115**	**0.058**	**2.388**
1	7	offset + 1 + t	AR(1) + seas(∼1 + t, S=1)		1	7	57031	0.496	<0.0001	0.064	2.171
1	8	offset + 1 + t		NE(2)	1	5	56755	0.516	0.003	0.056	2.304
1	9	offset + 1 + t		NE(2) + logpopdens	1	5	56763	0.516	0.002	0.056	2.313
1	10	offset + 1 + t		NE(2) + seas(∼1, S = 1)	1	7	56685	0.515	0.001	0.054	2.308
1	11	offset + 1 + t		NE(2) + seas(∼1 + t, S = 1)	1	8	56680	0.516	0.203	0.057	2.201
1	12	offset + 1 + t			State	4	65310	0.659	<0.0001	0.098	2.145
2	13	offset + 1 + seas(∼1, S=1)	AR(1) + seas(∼1, S=1)		1	7	57024	0.502	<0.0001	0.048	2.627
2	14	offset + 1	AR(1) + seas(∼1, S=1)		1	5	57101	0.502	<0.0001	0.049	2.612
2	15	offset + 1 + t + logpopdens	AR(1) + seas(∼1, S=1)		1	6	57128	0.499	<0.0001	0.055	2.496
2	16	offset + 1 + t	AR(1) + seas(∼1 + t, S=1)		1	7	57031	0.496	<0.0001	0.064	2.171
2	17	offset + 1 + t	AR(1) + seas(∼1 + t, S=2)		1	9	56996	0.496	<0.0001	0.064	2.176
2	18	offset + 1 + t	AR(1) + seas(∼1, S=1)	NE(2)	1	8	53362	0.458	0.210	0.055	2.105
2	19	offset + 1 + t	AR(1) + seas(∼1, S=1)	NE(2) + seas(∼1, S = 1)	1	10	53300	0.457	0.294	0.053	2.101
2	20	offset + 1 + t	AR(1) + seas(∼1, S=1)	NE(2) + seas(∼1 + t, S = 1)	1	11	53301	0.458	0.125	0.053	2.122
2	21	offset + 1 + t	AR(1) + seas(∼1, S=1)	NE(2) + logpopdens	1	8	53398	0.458	0.144	0.054	2.111
2	22	offset + 1 + t	AR(1) + seas(∼1, S=1)		State	7	57059	0.493	0.123	0.058	2.389
2	**23**	**offset + 1 + t**	**AR(2) + seas(∼1, S=1)**		**1**	**6**	**53833**	**0.455**	**0.189**	**0.053**	**2.230**
2	24	offset + 1 + t	AR(3) + seas(∼1, S=1)		1	6	52279	0.439	0.005	0.061	2.017
2	25	offset + 1 + t	AR(4) + seas(∼1, S=1)		1	6	51342	0.428	<0.0001	0.064	1.877
3	26	offset + 1 + seas(∼1, S=1)	AR(2) + seas(∼1, S=1)		1	7	53806	0.457	<0.0001	0.043	2.395
3	27	offset + 1	AR(2) + seas(∼1, S=1)		1	5	53844	0.458	<0.0001	0.047	2.340
3	28	offset + 1 + t + logpopdens	AR(2) + seas(∼1, S=1)		1	6	53835	0.456	<0.0001	0.042	2.404
3	29	offset + 1 + t	AR(2) + seas(∼1 + t, S=1)		1	7	53815	0.455	0.002	0.056	2.087
3	30	offset + 1 + t	AR(2) + seas(∼1 + t, S=2)		1	9	53692	0.455	0.001	0.057	2.079

Stage	Model No.	END	AR	NE	Dispersion	No. parameters	AIC	RPS	Calibration (p-value)	C1090	Avg. width
3	31	offset + 1 + t	AR(3) + seas(∼1, S=1)		1	6	52279	0.439	0.005	0.061	2.017
3	32	offset + 1 + t	AR(2) + seas(∼1, S=1)	NE(1)	1	7	51749	0.437	0.181	0.054	1.974
3	**33**	**offset + 1 + t**	**AR(2) + seas(∼1, S=1)**	**NE(1) + seas(∼1, S = 1)**	**1**	**9**	**51675**	**0.437**	**0.122**	**0.055**	**1.966**
3	34	offset + 1 + t	AR(2) + seas(∼1, S=1)	NE(3) + seas(∼1 + t, S = 1)	1	11	51543	0.437	0.656	0.050	2.029
3	35	offset + 1 + t	AR(2) + seas(∼1, S=1)		State	7	53831	0.455	0.192	0.053	2.230
4	36	offset + 1 + seas(∼1, S=1)	AR(2) + seas(∼1, S=1)	NE(1) + seas(∼1, S = 1)	1	10	51701	0.437	0.085	0.056	1.961
4	37	offset + 1	AR(2) + seas(∼1, S=1)	NE(1) + seas(∼1, S = 1)	1	8	51673	0.437	0.194	0.055	1.969
4	38	offset + 1 + t + logpopdens	AR(2) + seas(∼1, S=1)	NE(1) + seas(∼1, S = 1)	1	9	51691	0.437	0.153	0.056	1.962
4	39	offset + 1 + t	AR(2) + t	NE(1) + seas(∼1, S = 1)	1	8	51670	0.439	0.001	0.059	1.865
4	40	offset + 1 + t	AR(2) + seas(∼1, S=2)	NE(1) + seas(∼1, S = 1)	1	11	51545	0.437	0.115	0.055	1.973
4	41	offset + 1 + t	AR(2) + seas(∼1 + t, S=2)	NE(1) + seas(∼1 + t, S = 1)	1	15	51446	0.441	0.563	0.054	1.959
4	**42**	**offset + 1**	**AR(4) + seas(∼1, S=1)**	**NE(1) + seas(∼1, S = 1)**	**1**	**8**	**50323**	**0.420**	**0.346**	**0.054**	**1.872**
4	43	offset + 1 + t	AR(2) + seas(∼1, S=1)	NE(1) + t	1	8	51749	0.437	0.545	0.053	2.003
4	44	offset + 1 + t	AR(2) + seas(∼1, S=1)	NE(1) + seas(∼logpopdens, S = 1)	1	9	51780	0.438	0.202	0.056	1.975
4	45	offset + 1 + t	AR(2) + seas(∼1, S=1)	NE(3)	1	8	51642	0.437	0.333	0.053	1.972
4	46	offset + 1 + t	AR(2) + seas(∼1, S=1)	NE(1) + seas(∼1, S = 1)	State	10	51676	0.437	0.118	0.055	1.964
4	47	offset + 1 + t	AR(2) + seas(∼1, S=1)	NE(1) + seas(∼logpopdens + t, S = 1)	State	11	51782	0.438	0.314	0.055	1.988
5	48	offset + 1 + seas(∼1, S=1)	AR(4) + seas(∼1, S=1)	NE(1) + seas(∼1, S = 1)	1	10	50342	0.420	0.297	0.055	1.867
5	49	offset + 1	AR(4) + seas(∼1 + t, S=1)	NE(1) + seas(∼1 + t, S = 1)	1	10	50296	0.424	0.614	0.052	1.864
5	50	offset + 1 + logpopdens	AR(4) + seas(∼1, S=1)	NE(1) + seas(∼1, S = 1)	1	9	50332	0.420	0.439	0.054	1.870
5	51	offset + 1	AR(4) + t	NE(1) + seas(∼1, S = 1)	1	7	50336	0.424	0.000	0.060	1.763
5	**52**	**offset + 1**	**AR(4) + seas(∼1, S=2)**	**NE(1) + seas(∼1)**	**1**	**10**	**50164**	**0.419**	**0.194**	**0.055**	**1.868**
5	53	offset + 1	AR(4) + seas(∼1 + t, S=1)	NE(1) + seas(∼1 + t, S = 2)	1	14	50097	0.423	0.782	0.052	1.851
5	54	offset + 1	AR(4) + seas(∼1, S=1)	NE(1) + seas(∼1 + t, S = 1)	1	9	50324	0.420	0.620	0.052	1.904
5	55	offset + 1	AR(4) + seas(∼1, S=1)	NE(1) + seas(∼logpopdens, S = 1)	1	8	50401	0.421	0.425	0.055	1.873
5	56	offset + 1	AR(4) + seas(∼1, S=1)	NE(1)	1	6	50416	0.420	0.251	0.054	1.877
5	57	offset + 1	AR(4) + seas(∼1, S=1)	NE(1) + seas(∼1, S = 1)	State	9	50325	0.420	0.342	0.054	1.873
5	58	offset + 1	AR(4) + seas(∼1, S=1)	NE(1) + seas(∼logpopdens + t, S = 1)	State	10	50405	0.421	0.537	0.055	1.876

PIT histograms for the selected model at each stage. Model 42 is the final model. Model 52 offered minor improvement in RPS with additional complexity.

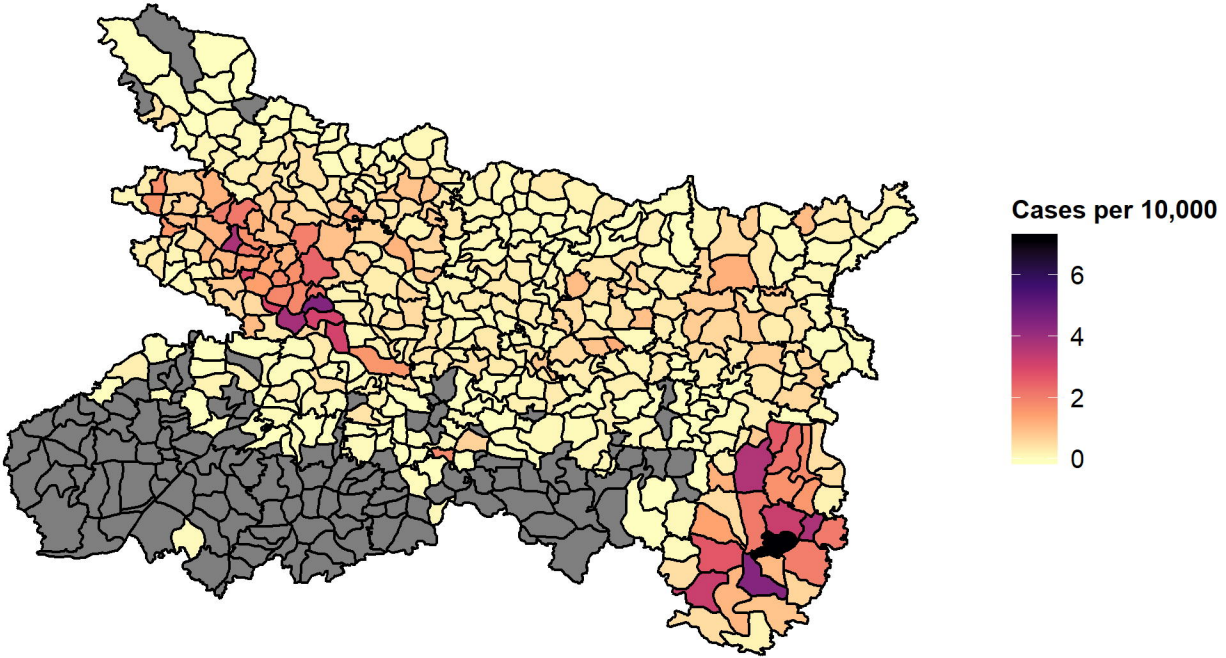
Fig. S3

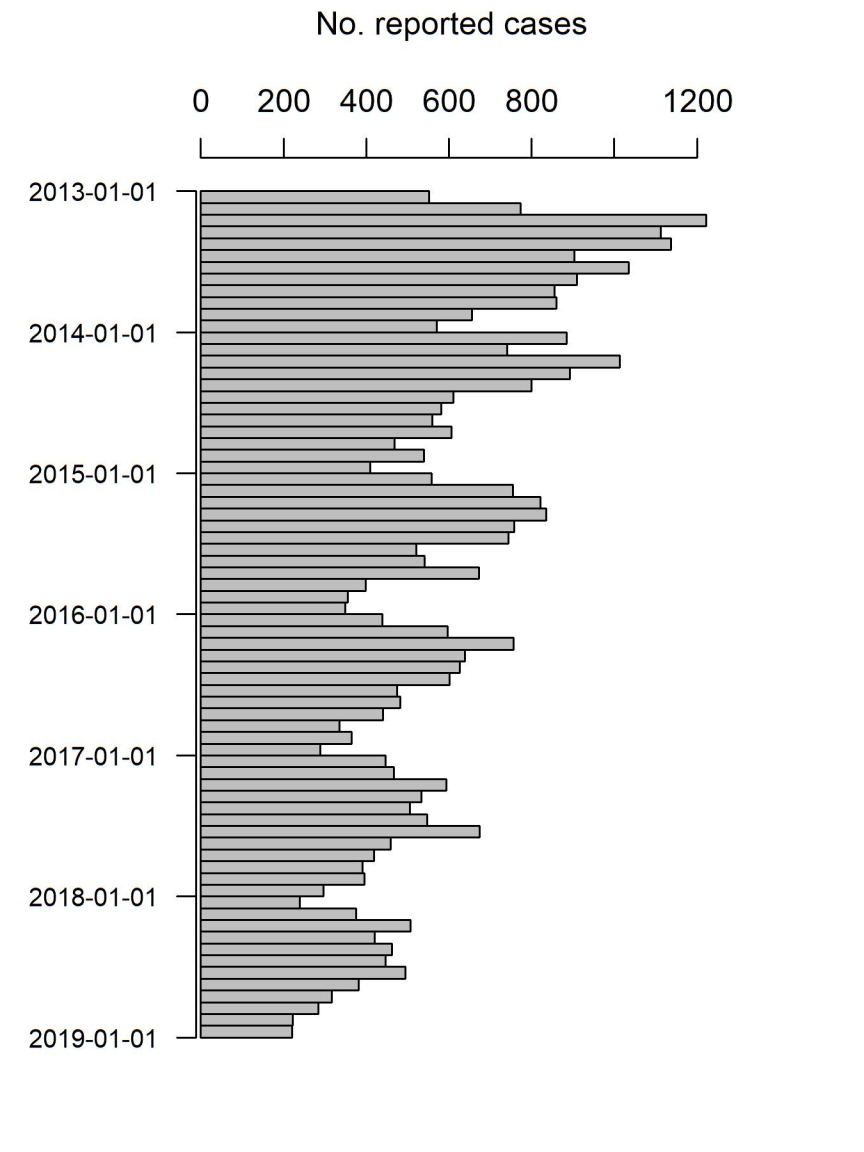
Fitted seasonal waves in the auto-regressive (AR) and neighbourhood (NE) model components. Both reflect the first-quarter peak in reported cases but the magnitude of the waves differs, with the contribution of the AR component varying more than that of the NE.

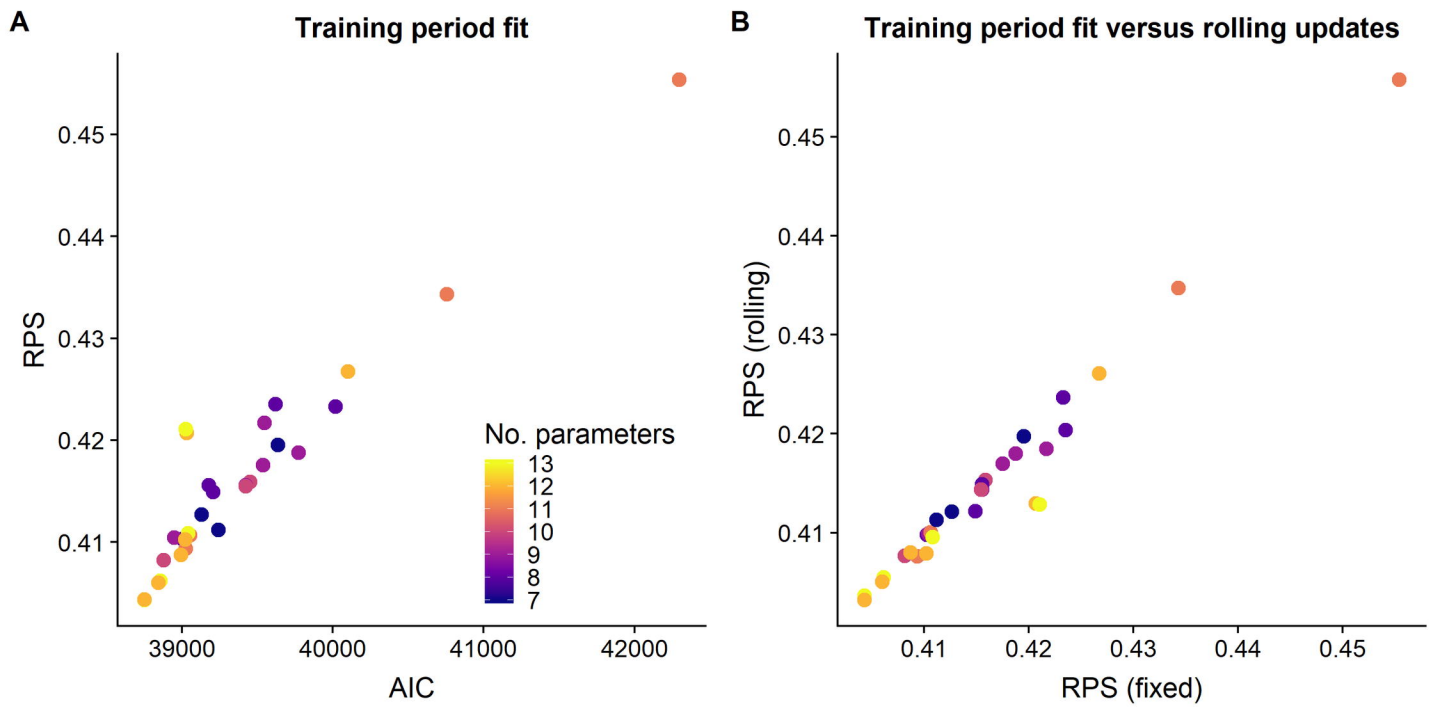
Fig. S4

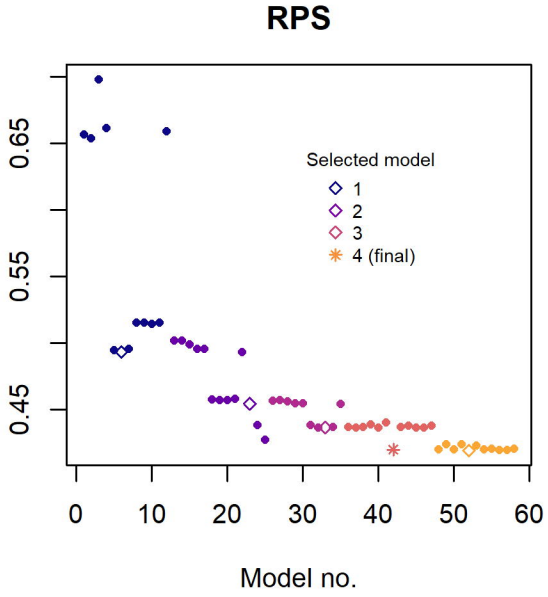
Fig. S5 Blocks with average RPS greater than 2.5 over the test period

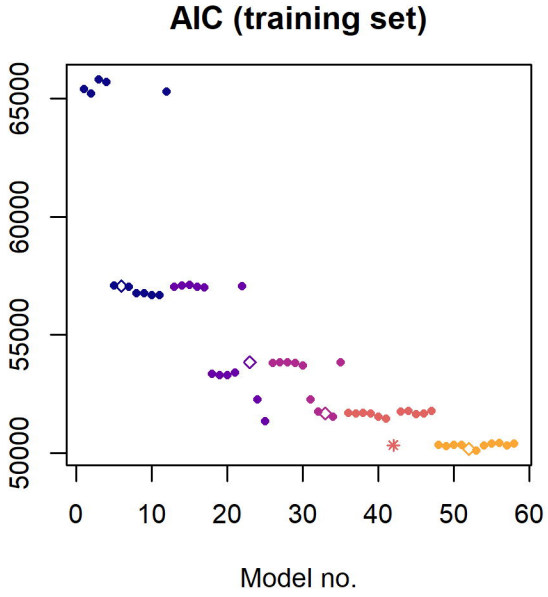
(Jan 2017 - Dec 2018)

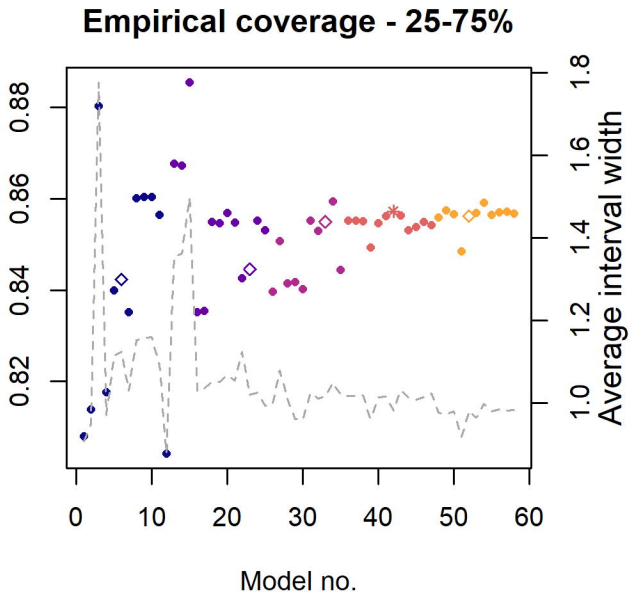

Cases per 10,000
6
4
2
0

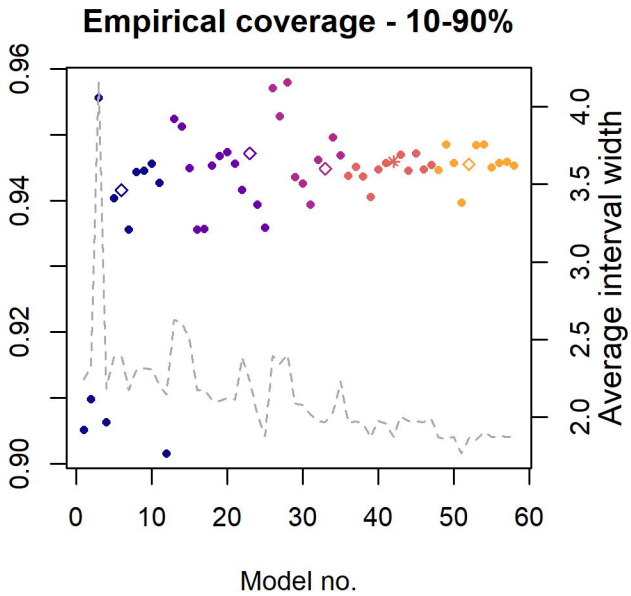

No. reported cases
0
200
400
600
800
1200
2013-01-01
2014-01-01
2015-01-01
2016-01-01
2017-01-01
2018-01-01
2019-01-01

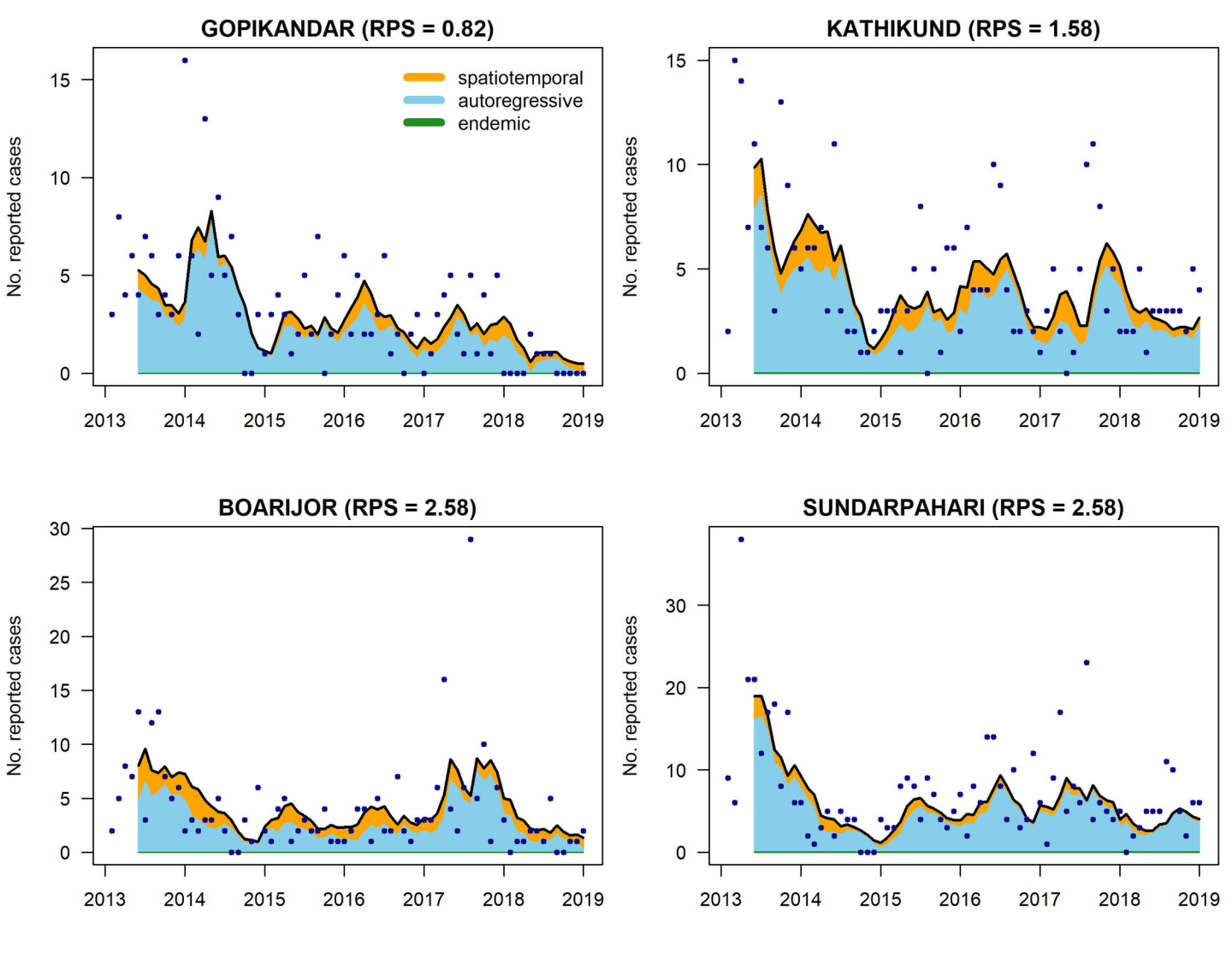

A
Training period fit
RPS
AIC
0.45
0.44
0.43
0.42
0.41
39000
40000
41000
42000
No. parameters
13
12
11
10
9
8
7
B
Training period fit versus rolling updates
RPS (rolling)
RPS (fixed)
0.45
0.44
0.43
0.42
0.41
0.41
0.42
0.43
0.44
0.45

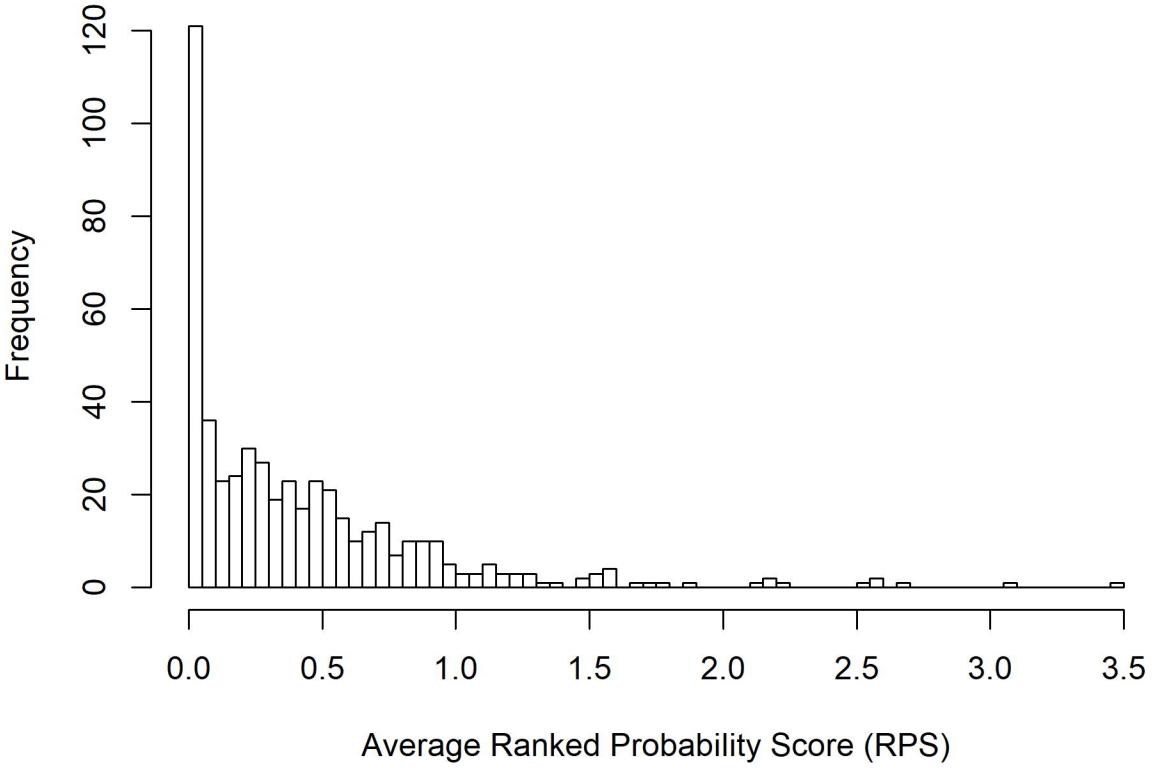

RPS
Selected model
1
2
3
4 (final)
0.45
0.55
0.65
0
10
20
30
40
50
60
Model no.

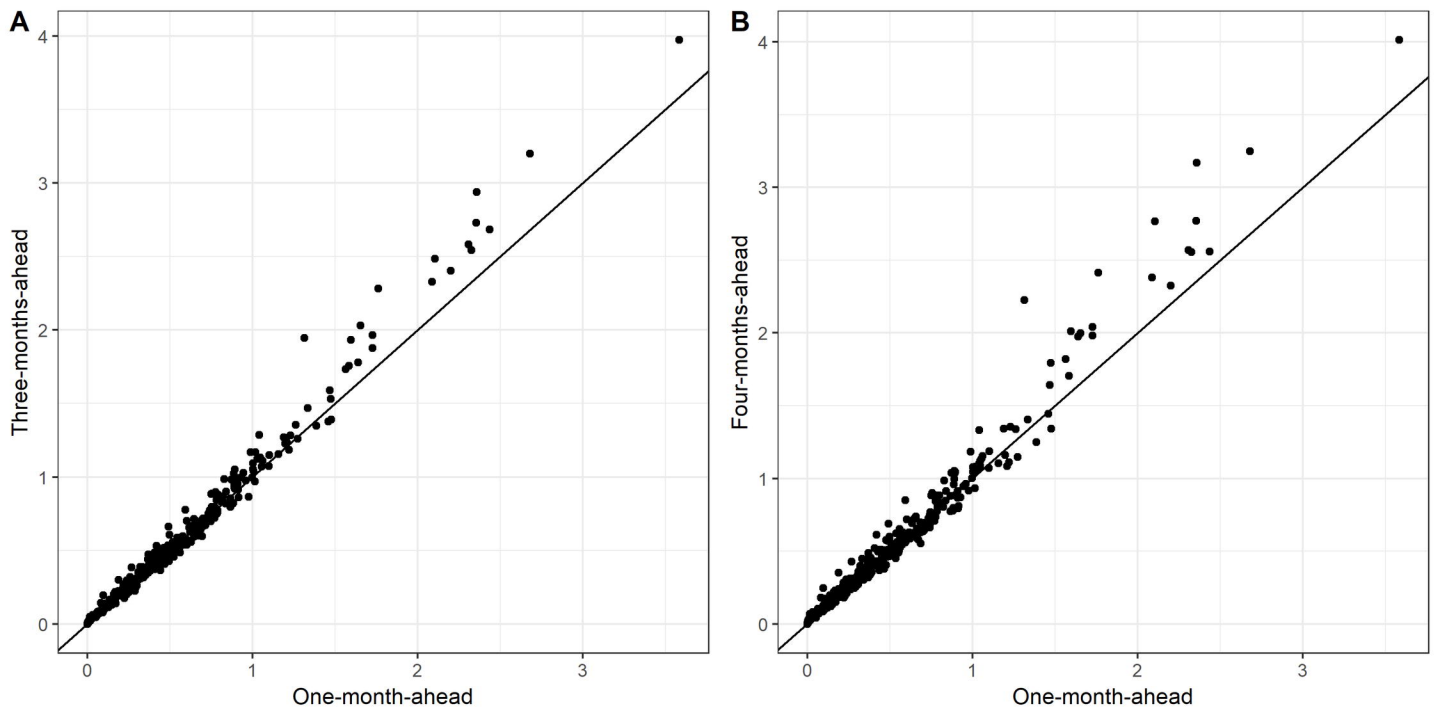

AIC (training set)
50000
55000
60000
65000
0
10
20
30
40
50
60
Model no.

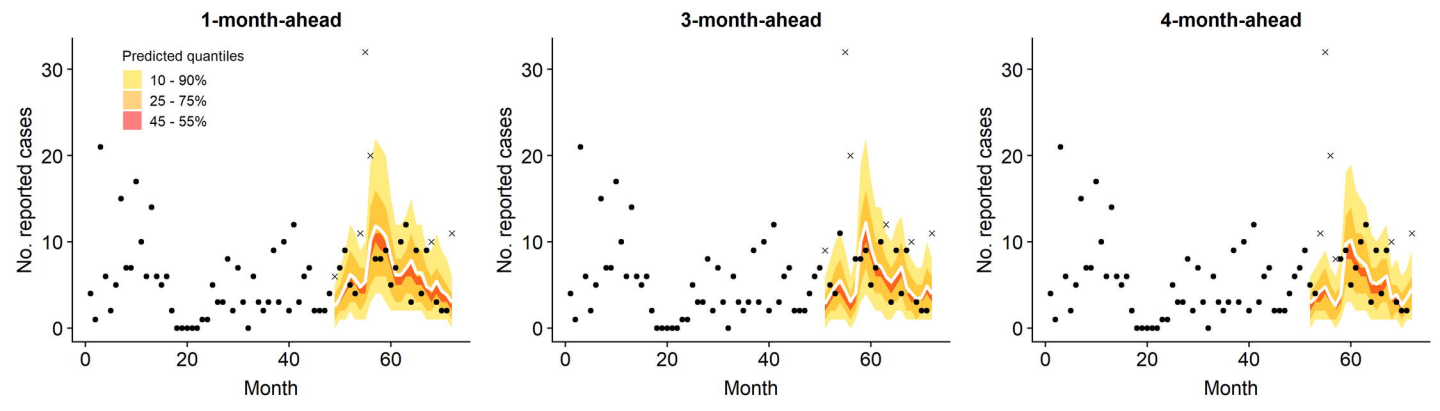

Empirical coverage - 25-75%
0.82
0.84
0.86
0.88
1.0
1.2
1.4
1.6
1.8
Average interval width
0
10
20
30
40
50
60
Model no.

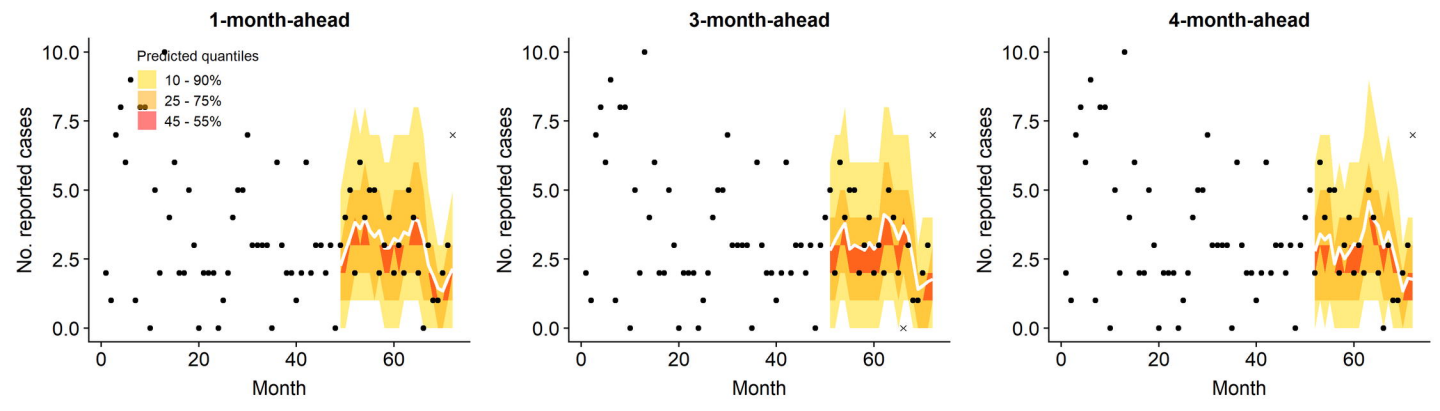

Empirical coverage - 10-90%
0.90
0.92
0.94
0.96
2.0
2.5
3.0
3.5
4.0
Average interval width
0
10
20
30
40
50
60
Model no.

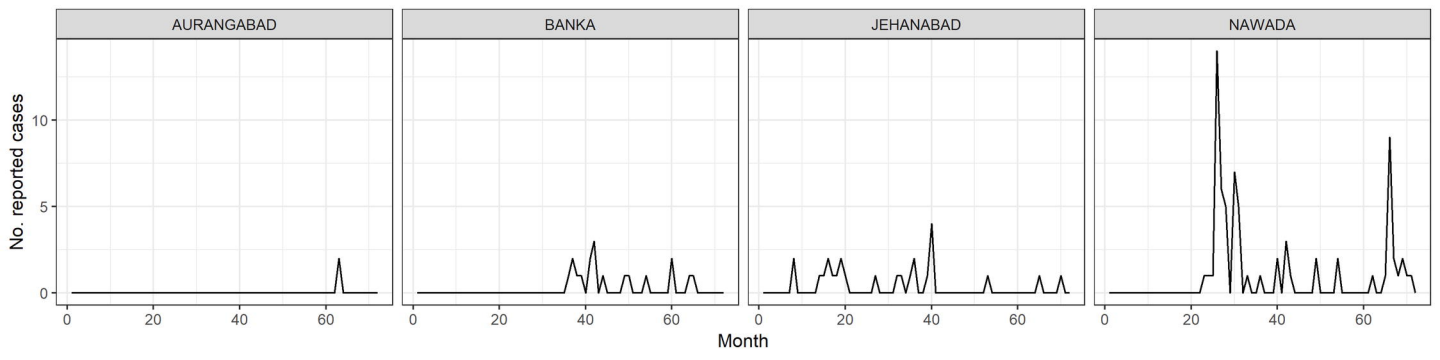

GOPIKANDAR (RPS = 0.82)
KATHIKUND (RPS = 1.58)
BOARIJOR (RPS = 2.58)
SUNDARPAHARI (RPS = 2.58)
spatiotemporal
autoregressive
endemic
No. reported cases
0
5
10
15
20
25
30
2013
2014
2015
2016
2017
2018
2019

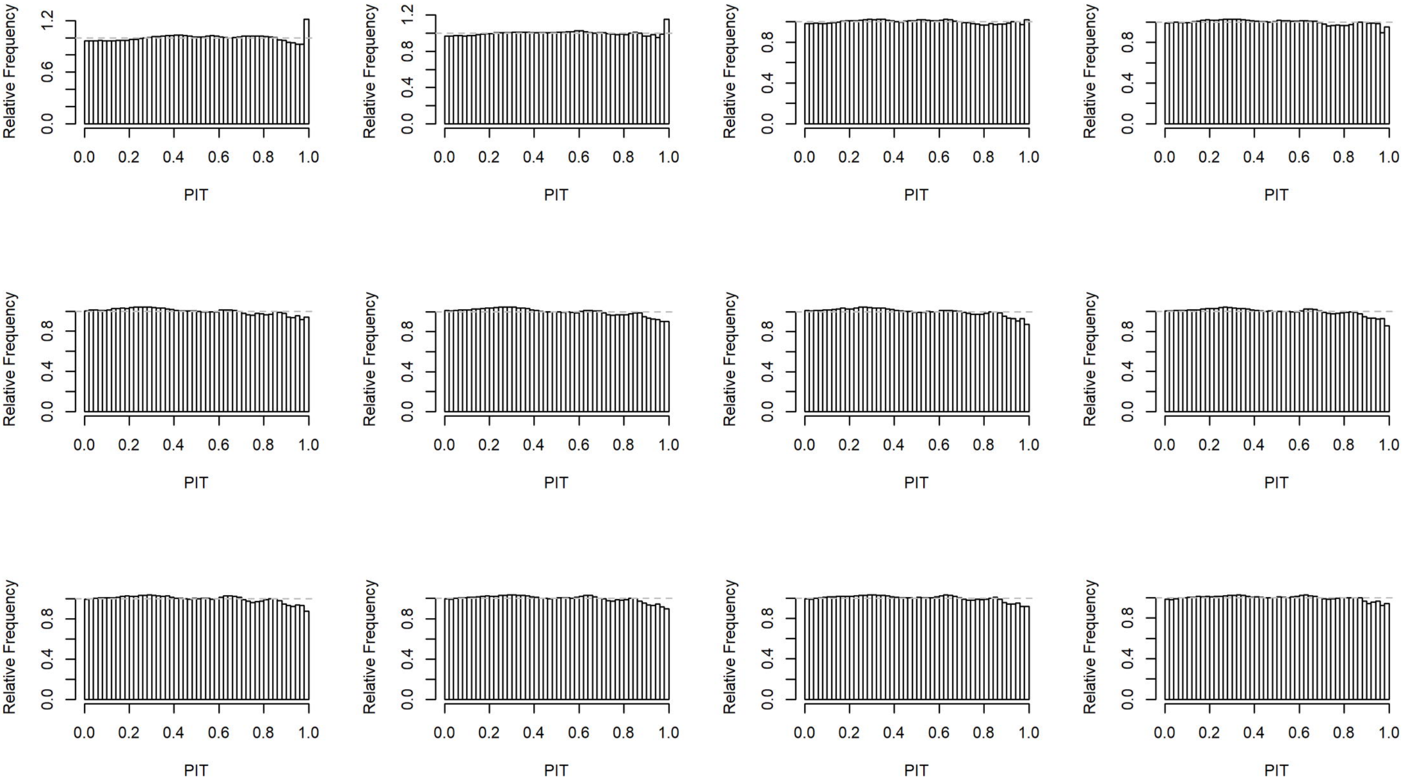

Frequency
0
20
40
60
80
100
120
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Average Ranked Probability Score (RPS)

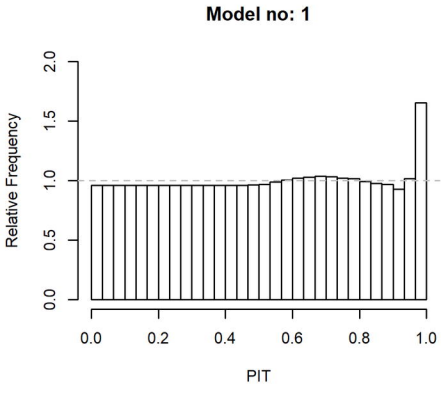

A
Three-months-ahead
One-month-ahead
0
1
2
3
4
B
Four-months-ahead
One-month-ahead
0
1
2
3
4

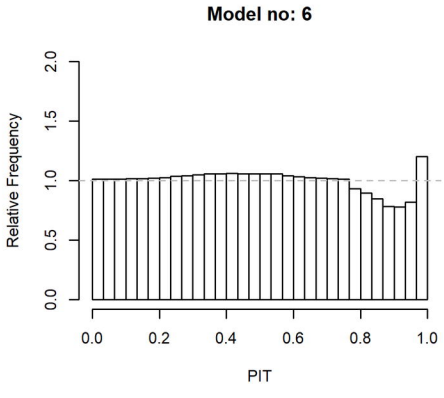

1-month-ahead
3-month-ahead
4-month-ahead
Predicted quantiles
10 - 90%
25 - 75%
45 - 55%
No. reported cases
Month
0
10
20
30
0
20
40
60

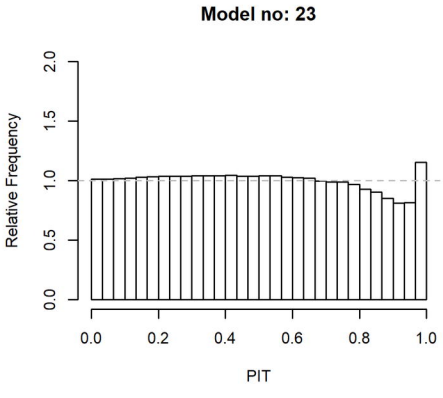

1-month-ahead
3-month-ahead
4-month-ahead
Predicted quantiles
10 - 90%
25 - 75%
45 - 55%
No. reported cases
Month
0.0
2.5
5.0
7.5
10.0
0
20
40
60

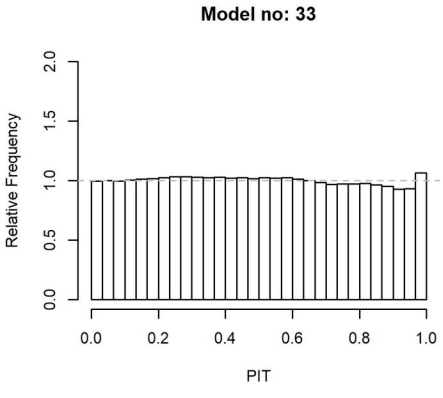

AURANGABAD
BANKA
JEHANABAD
NAWADA
No. reported cases
0
5
10
0
20
40
60
Month

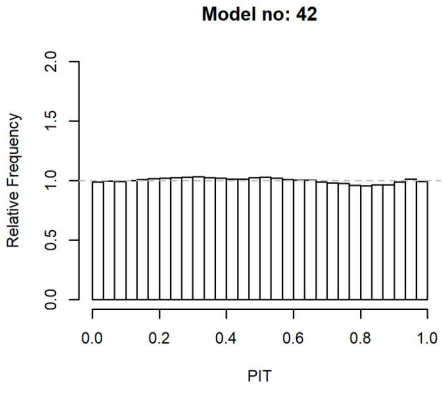

Relative Frequency
PIT
0.0
0.2
0.4
0.6
0.8
1.0
1.2

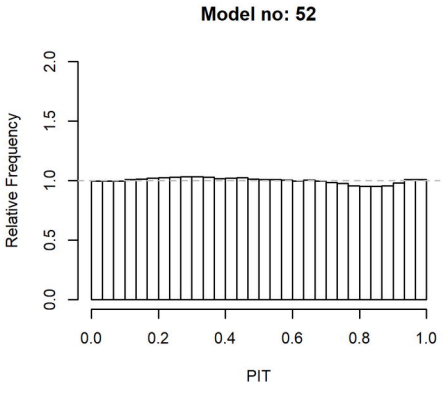

Model no: 1
Relative Frequency
0.0
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
1.0
PIT

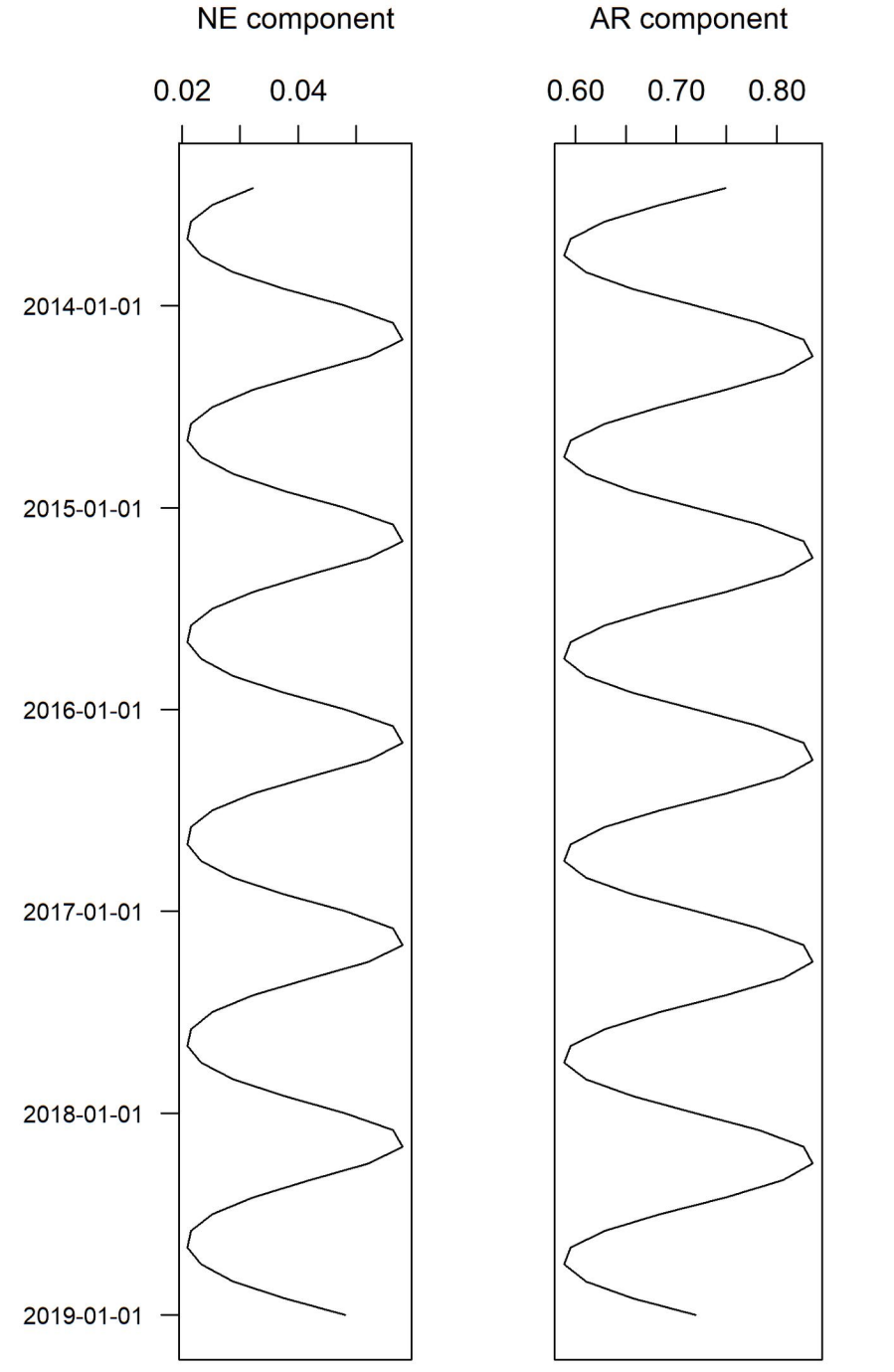

Model no: 6
Relative Frequency
0.0
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
1.0
PIT

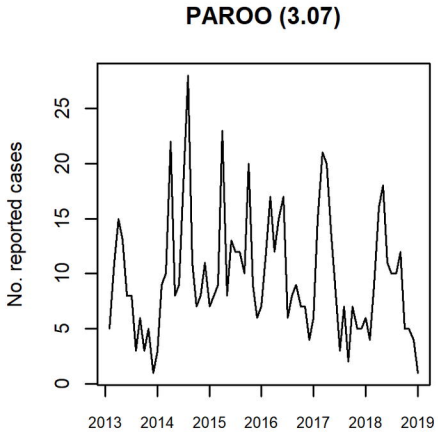

Model no: 23
Relative Frequency
0.0
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
1.0
PIT

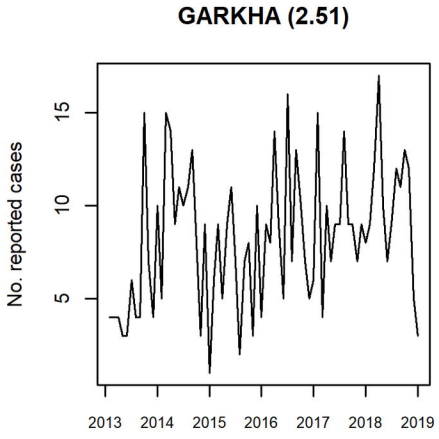

Model no: 33
Relative Frequency
0.0
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
1.0
PIT

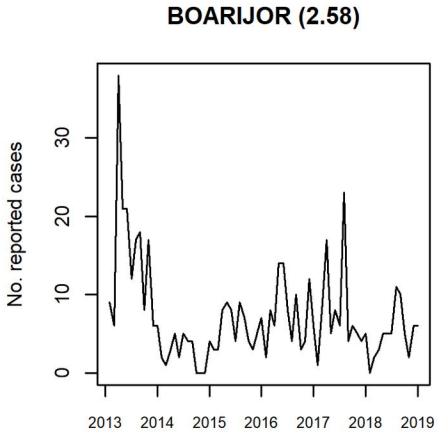

Model no: 42
Relative Frequency
0.0
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
1.0
PIT

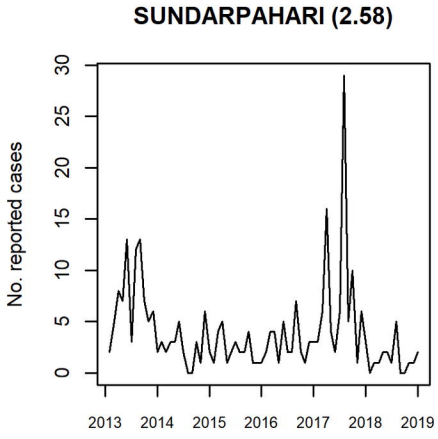

Model no: 52
Relative Frequency
0.0
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
1.0
PIT

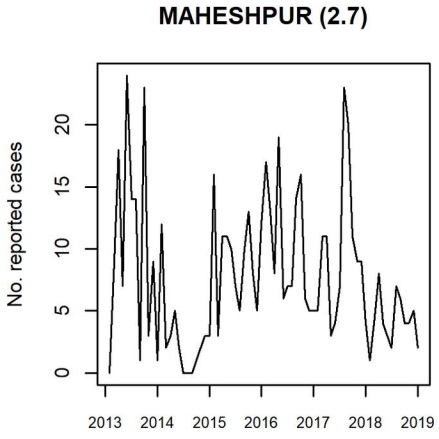

NE component
AR component
0.02
0.04
0.60
0.70
0.80
2014-01-01
2015-01-01
2016-01-01
2017-01-01
2018-01-01
2019-01-01

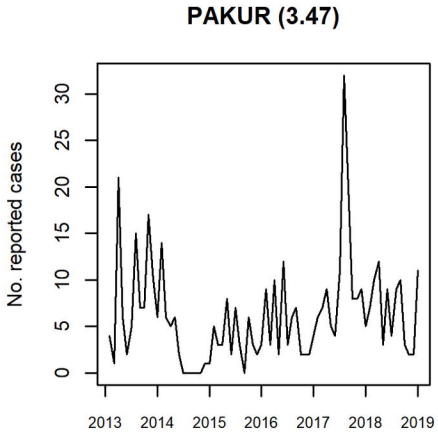

PAROO (3.07)
No. reported cases
0
5
10
15
20
25
2013
2014
2015
2016
2017
2018
2019


GARKHA (2.51)
No. reported cases
5
10
15
2013
2014
2015
2016
2017
2018
2019


BOARIJOR (2.58)
No. reported cases
0
10
20
30
2013
2014
2015
2016
2017
2018
2019


SUNDARPAHARI (2.58)
No. reported cases
0
5
10
15
20
25
30
2013
2014
2015
2016
2017
2018
2019


MAHESHPUR (2.7)
No. reported cases
0
5
10
15
20
2013
2014
2015
2016
2017
2018
2019


PAKUR (3.47)
No. reported cases
0
5
10
15
20
25
30
2013
2014
2015
2016
2017
2018
2019