Fitting mechanistic epidemic models to data: a comparison of simple Markov chain Monte Carlo approaches

Michael Li, Jonathan Dushoff, Ben Bolker

June 14, 2017

Abstract

Background

Simple mechanistic epidemic models are widely used for forecasting and parameter estimation of infectious diseases based on noisy case reporting data. Despite the widespread application of models to emerging infectious diseases, we know little about the comparative performance of standard computational-statistical frameworks in these contexts. Here we build a simple stochastic, discrete-time, discrete-state epidemic model with both process and observation error and use it to characterize the effectiveness of different flavours of Bayesian Markov chain Monte Carlo (MCMC) techniques. We use fits to simulated data, where parameters (and future behaviour) are known to explore the limitations of different platforms and quantify parameter estimation accuracy, forecasting accuracy, and computational efficiency across combinations of modeling decisions (e.g. discrete vs. continuous latent states, levels of stochasticity) and computational platforms (JAGS, NIMBLE, Stan).

Results

Models incorporating at least one source of population-level variation (i.e., dispersion in either the transmission process or the observation process) provide reasonably good forecasts and parameter estimates, while models that incorporate only individual-level variation can lead to inaccurate (or overconfident) results. Models using continuous approximations to the transmission process showed improved computational efficiency without loss of accuracy.

Conclusion

Simple models of disease transmission and observation can be fitted reliably to simple simulations, as long as population-level variation is taken into account. Continuous approximations can improve computational efficiency using more advanced MCMC techniques.

Keywords: MCMC HMC TSIR Dispersion Moment-matching

1 Introduction

Simple homogeneous population models have been widely used to study emerging infectious disease outbreaks. Although such models can provide important insights — including estimated epidemic sizes and predicted effects of intervention strategies, as well as short-term forecasts — they neglect important spatial, individual-level and other heterogeneities. Decades of work have created frameworks that enable researchers to construct models that capture many of these more realistic aspects of infectious disease epidemics. But many challenges remain. In particular, estimating parameters (and associated uncertainties) is always challenging, especially for models incorporating multiple forms of heterogeneity, and especially during the early stages of an epidemic when data are limited. Using complex models that are insufficiently supported by data can lead to imprecise and unstable parameter estimates (Ludwig and Walters, 1985) — in such cases, researchers often revert to simpler models for practical purposes.

In the past few decades, researchers have begun to adopt Bayesian approaches to disease modeling problems. Bayesian Markov Chain Monte Carlo (MCMC) is a powerful, widely used sampling-based estimation approach. Despite the widespread use of MCMC in epidemic modeling (Morton and Finkenstädt, 2005; O'Neill, 2002), however, there have been relatively few systematic studies of the comparative performance of statistical frameworks for disease modeling O'Neill et al. (2000).

In this paper, we apply relatively simple MCMC approaches to data from simulated epidemics that incorporate stochasticity in both transmission and observation, as well as variable generation-interval distributions (not assumed to be known when fitting). We compare model approaches of varying complexity, including an estimation model that matches the simulation model. For each model we quantify parameter estimation accuracy and forecasting accuracy; this sheds light on which phenomena are most important to include in models to be used for estimation and forecasting.

We also compare three different MCMC platforms: JAGS (Plummer et al., 2003), NIMBLE (de Valpine et al., 2016) and Stan (Carpenter et al., 2016). In principle, for any given model, any valid method of MCMC sampling should eventually converge on the same (correct) posterior distribution. However, even with the relatively simple models considered here, a theoretically valid software package can experience problems in practice: we wanted to investigate this phenomenon. Furthermore, even when different platforms converge to essentially the same result, they may show large differences in computational efficiency: we therefore also quantify efficiency for the models we study.

2 Methods

We generated test data using a simple framework that combines a *transmission process* based on a simple discrete-time model with an *observation process* to account for incomplete reporting. Both processes are assumed to be stochastic. We then fit the observed cases from these simulations using Bayesian methods that model the underlying true number of infections as a latent (i.e., unobserved) variable. Our Bayesian fitting models explore an approach that matches the assumptions of the simulation model, as well as various simplifications: in particular, we explore simpler methods of accounting for variation in both the transmission process and the observation process, and the use of continuous rather than discrete latent variables. For simplicity, we have here assumed that data are reported on the same discrete time scale on which the disease process is simulated (but not that the reported period is the same as the generation time of the disease; see below). This assumption requires that the generation time be at least as the reporting period. It would be relatively straightforward to relax this assumption, for example by assuming that the epidemic dynamics occur on a finer time scale than the reporting interval, or by simulating in continuous time but fitting with a discrete-time model; we do not explore these questions here.

2.1 Simulation Model

The transmission process of our dual-process framework is based on the Reed-Frost chain binomial model, which can also be described as a discrete-time, stochastic compartmental SIR model (Ludwig, 1973). To account for the possibility that some fraction of the population may be beyond the scope of the epidemic — geographically or socially isolated, genetically resistant, vaccinated or immune due to previous exposure — we assume that only a proportion P_{eff} of the total census population is actually susceptible to infection. We further assume that, in every time step, only a proportion (randomly chosen with mean P_{rep}) of new infections are actually observed. We model both transmission and observation using a beta-binomial (rather than binomial) distribution to account for additional sources of variation (i.e., overdispersion) in both processes. The equations are:

$$
\begin{aligned}
N_{\text{eff}} &= P_{\text{eff}} N && (1)\\
S_1 &= N_{\text{eff}} - I_1 && (2)\\
\Phi_t &= \sum_{i=1}^{\ell} k(i) I_{t-\ell+i} && (3)\\
I_{t+1} &\sim \text{BetaBin}(1 - e^{-\Phi_t}, S_t, \delta_P) && (4)\\
S_{t+1} &= S_t - I_{t+1} && (5)\\
\text{Obs}_t &\sim \text{BetaBin}(P_{\text{rep}}, I_t, \delta_{\text{obs}}). && (6)
\end{aligned}
$$

where Φ_t is the force of infection at time t; N_{eff} is the effective population size; and ℓ is the number of lags.

The most common parameterization of the beta-binomial comprises three parameters: the binomial size parameter N plus two additional shape parameters (α and β) that describe the Beta distribution of the per-trial probability. Uses of the beta-binomial in statistical modeling instead typically transform the shape parameters into a pair of parameters that describe the per-trial probability and a dispersion parameter Morris (1997); larger values of the dispersion parameter δ correspond to less variability. We use a slight modification of this parameterization (see figure 1)

We extend the Reed-Frost model by allowing the infectious period to last longer than one step, and the infectivity to vary based on how long an individual has been infected; we do this by parameterizing a transmission kernel that describes the force of infection coming from individuals who were infected ℓ time steps ago. For convenience, we assumed a fixed maximum window length ($\ell = 5$). We then based our transmission kernel on a negative binomial distribution, truncated to fit this window:

$$\tilde{k}(i) = i^{(G_S - 1)} \times \exp\left(\frac{-i}{G_P \times \ell}\right), i = 1, ..., \ell \tag{7}$$

$$k(i) = \frac{\mathcal{R}_0}{N_{\text{eff}}} \times \frac{\tilde{k}(i)}{\sum_{i=1}^{\ell} \tilde{k}(i)}, i = 1, ..., \ell \tag{8}$$

Here, $\mathcal{R}_0$ represents the basic reproductive number and G_S and G_P are shape and position parameters, respectively.

2.2 Fitting Model

2.2.1 Transmission and Observational Process Errors

The transmission (eq. 4) and observation (eq. 6) processes in the simulation model are both defined as beta-binomial (BB) processes. In fitting, we used the BB to match the simulation model, but also tried several simpler alternatives: binomial (B), Poisson (P), and negative-binomial (NB) processes. Process B does not allow for overdispersion, while NB does not incorporate the size of the pool from which a value is chosen; that is, it is theoretically possible for a NB sample of the number of infections to be larger than the current susceptible population (although this is extremely unlikely when the *per capita* infection probability is small). Process P neglects both of these phenomena. Figure 1 illustrates the relationship of the four discrete distributions.

2.2.2 Multiple Scale Decorrelation

The proportion of the population assumed to be effectively susceptible (P_{eff}) and the reporting proportion (P_{rep}) have very similar effects on observed incidence. We therefore reparameterized the model so that it uses a single parameter P_{effrep} for their

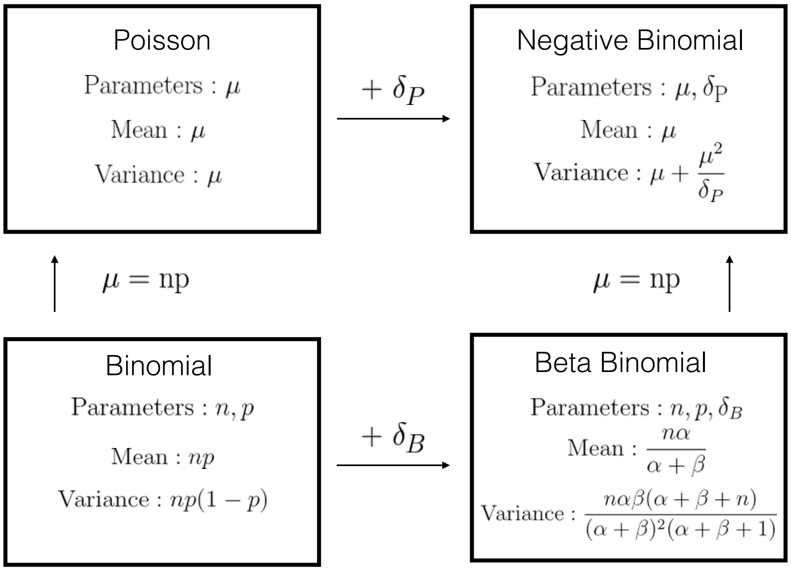

Poisson
Parameters : μ
Mean : μ
Variance : μ
$+ \delta_P$
Negative Binomial
Parameters : μ, δ_P
Mean : μ
Variance : $\mu + \frac{\mu^2}{\delta_P}$
$\mu = \text{np}$
$\mu = \text{np}$
Binomial
Parameters : n, p
Mean : np
Variance : $np(1-p)$
$+ \delta_B$
Beta Binomial
Parameters : n, p, δ_B
Mean : $\frac{n\alpha}{\alpha+\beta}$
Variance : $\frac{n\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Figure 1: Discrete distribution relationships. For beta-binomial distribution (bottom right panel), we used an alternative parameterization α and β, where $\alpha = \frac{\delta_B}{1-p}$ and $\beta = \frac{\delta_B}{p}$. Moving from the top to bottom row adds a size parameter (replacing μ with np). Moving from left to right adds a dispersion parameter δ_P and δ_B for Poisson and Binomial distribution respectively.

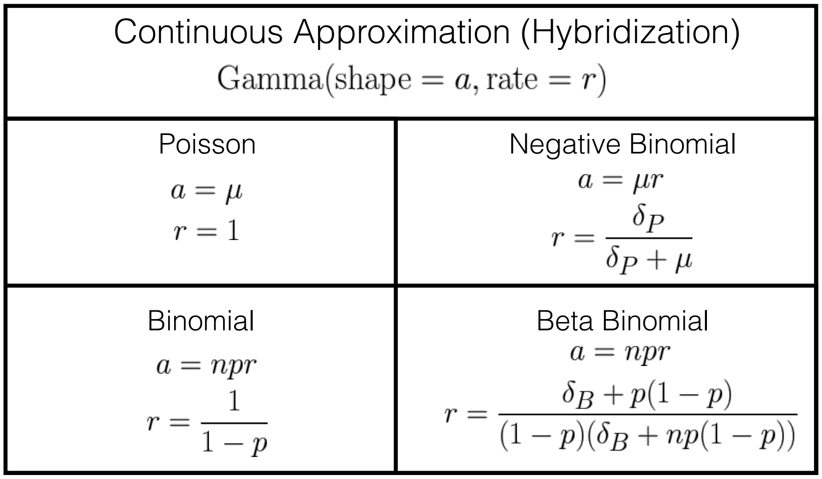
product, and a second to govern how the product is apportioned between the two quantities:

$$\widehat{P}_{\text{eff}} = P_{\text{effrep}}^{1-\rho} \tag{9}$$

$$\widehat{P}_{\text{rep}} = P_{\text{effrep}}^{\rho} \tag{10}$$

We expected *a priori* that this parameterization would improve statistical convergence, since it makes it possible to sample different values of the poorly constrained value of ρ without changing P_{effrep}. It is straightforward to back-calculate P_{eff} and P_{rep} once the model is fitted. For similar reasons, we experimented with measuring infected individuals on a "reporting" scale in our continuous-variable models (see below).

2.2.3 Continuous latent variables



Continuous Approximation (Hybridization) $\text{Gamma}(\text{shape} = a, \text{rate} = r)$	
Poisson $a = \mu$ $r = 1$	Negative Binomial $a = \mu r$ $r = \dfrac{\delta_P}{\delta_P + \mu}$
Binomial $a = npr$ $r = \dfrac{1}{1-p}$	Beta Binomial $a = npr$ $r = \dfrac{\delta_B + p(1-p)}{(1-p)(\delta_B + np(1-p))}$

Figure 2: Continuous approximation of discrete distributions via moment matching. Distributions in Figure 1 were matched to a Gamma distribution with equivalent first and second moments.

Another simplification we considered was treating the unobserved number of underlying cases as a continuous variable. To do this, we matched the first two moments of the discrete distribution to a Gamma distribution (Figure 2).

Eq. 4 and 6 can be rewritten as:

$$\widehat{I}_{t+1} \sim \text{Gamma}(a, \frac{r}{P_{\text{rep}}}) \tag{11}$$

$$\text{Obs}_t \sim \text{NB}(\widehat{I}_t, \delta_{\text{obs}}). \tag{12}$$

One advantage of this continuous approximation approach is that it allows us to scale our latent variable to help with model convergence, so that infected individuals are measured on the reporting scale. Another advantage is that it allows us to use

MCMC sampling procedures such as Hamiltonian Monte Carlo (HMC), which cannot easily use discrete latent variables.

2.3 Bayesian Markov Chain Monte Carlo

In Bayesian MCMC, model parameters are sampled from the posterior distribution by a reversible Markov chain whose stationary distribution is the target posterior distribution. Classical MCMC techniques include the Metropolis-Hasting algorithm (Hastings, 1970), Gibbs sampling (Geman and Geman, 1984), and slice sampling (Neal, 2003). Recently, convenient implementations of a powerful MCMC technique called Hamiltonian Monte Carlo (HMC: also called hybrid MC) (Duane et al., 1987) have become available. HMC uses the concept of Hamiltonian dynamics to create a proposal distribution for the M-H algorithm, together with the leap-frog algorithm and the No U-Turn sampler (Hoffman and Gelman, 2014). HMC requires more computational effort per sample step compared to other MCMC techniques, but because subsequent steps are less strongly correlated it also produces more effective samples per sample step (Carpenter et al., 2016; Hoffman and Gelman, 2014).

2.3.1 Platforms

Many software platforms implement the automatic construction of MCMC samplers for user-defined models. One of the most widely used platforms is JAGS (Just a Gibbs Sampler); despite its name, it combines a variety of MCMC techniques to fit models. NIMBLE (Numerical Inference for Statistical Models for Bayesian and Likelihood Estimation) is a more recent platform that allows users to flexibly model and customize different algorithms and sample techniques for MCMC. Neither JAGS nor NIMBLE has yet implemented HMC. One of the relatively few platforms that currently implements HMC is Stan, which provides full Bayesian inference for continuous-variable models based on the No-U-Turn sampler, an adaptive form of HMC.

2.3.2 Simulation and Evaluations

A typical frequentist statistical simulation scheme fits multiple realizations to data generated from a fixed set of parameters that is determined *a priori* and evaluates the match of the parameter estimates to the true values. When we fit our Bayesian model using informative priors, frequentist coverage is generally higher than nominal values (i.e. 90% posterior intervals will contain the true parameter values with > 90% probability). For validation, we therefore used a Bayesian simulation scheme where we first draw parameters from their prior distribution, generate data given the drawn parameters, and then fit the Bayesian model with the same prior distributions; by construction, this scheme should match the nominal coverage if the model fits are correct under their own assumptions (Cook et al., 2006). We sampled 100 sets of the parameters from the same prior distribution that was used in the fitting process; for each parameter set, we simulated one realization of 15 time steps (10 for fitting and 5 to compare to forecasts). All model variants were used to fit each realization (Table 1 and 2 in the appendix give more detail about parameters and priors). We added two convergence criteria to assess convergence for the main parameters ($\mathcal{R}_0$, P_{eff}, P_{rep}): we required a value of the Gelman and Rubin statistic $\hat{R} < 1.1$ and an effective sample size (ESS) greater than 400 for each replication. For each replication we sample four chains starting with 4000 iterations; we repeatedly double the number of iterations (with a upper threshold of one million iterations) until the convergence criteria are met. Forecasts were made by simulating incidence 5 time steps forward using parameters sampled from the fitted posterior distributions.

We evaluated our estimates of (1) total cases predicted over the forecast window (disaggregated forecasts are analyzed in the supplementary material) and (2) key model parameters (including the estimated mean generation interval (MGI: defined as $\frac{\sum_{i=1}^{\ell} i\hat{k}(i)}{\sum_{i=1}^{\ell} \hat{k}(i)}$)). We used bias, root mean square error (RMSE), and coverage to assess model fit. Bias and RMSE are based on proportional errors, defined as the log ratio of our estimate (taken as the median of the posterior sample) to the known true parameter value from our simulations. Errors were compared on the log scale in order to allow comparison of the accuracy of estimation of different parameters that may be on very different scales. The median is a scale-invariant, robust, summary statistic for the location parameter of a Bayesian posterior distribution Minsker et al. (2014). Thus in order to compare different parameters in a consistent, unitless fashion, the errors were calculated as $\epsilon_i = \log(\text{med}(\hat{\theta}_i)/\theta_i)$. We then calculated bias (median(ϵ)) and RMSE ($\sqrt{\text{mean}(\epsilon_i^2)}$).

3 Results

The full model (which matches the simulation model) provides generally good forecasts and parameter estimates as assessed by bias (Figure 3) or RMSE (Figure 4), except for estimates of P_{eff} using JAGS.

In general, models with any kind of dispersion in the transmission process, or with negative binomial dispersion in the observation process, did well. The exception is that models that combined negative binomial transmission dispersal with beta binomial observation dispersal produced biased forecasts and estimates of P_{rep}.

There are no clear differences in the quality of model fit due to multi-scale decorrelation, latent continuous transmission process or platform.

Figure 5 shows the statistical coverage of our estimates. Similar to the results for bias and RMSE (Figure 3 and 4), we find generally good coverage (i.e., close to the nominal value of 0.9) for models with dispersion in the transmission process, except that the negative-binomial transmission process model undercovers across the board

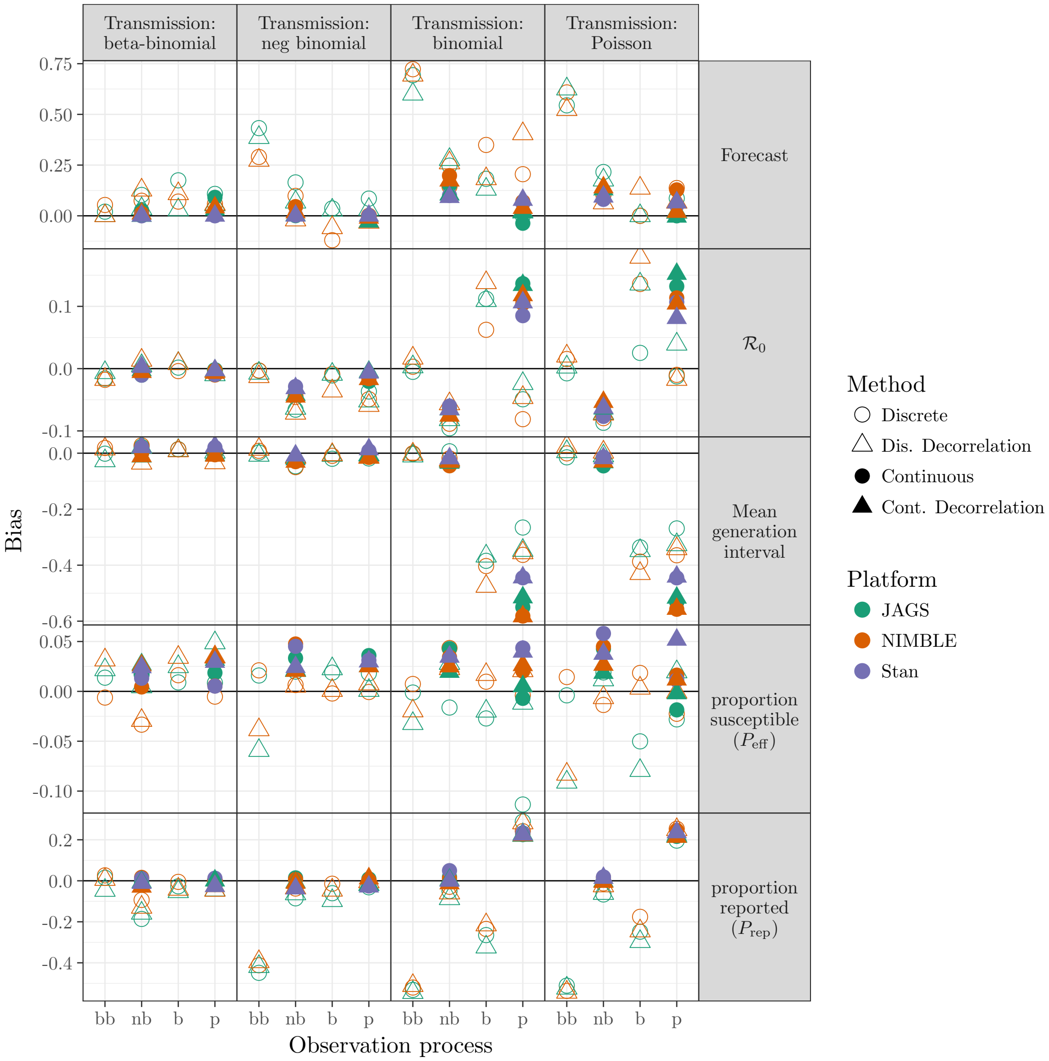


Figure 3: Comparison of bias (based on proportional errors) for forecasts and parameters using models described in Sect. 2.2 across different platforms described in Sect. 2.3.1. Models with overdispersion in the transmission process (BB and NB, leftmost and second-left columns of panels) and models with overdispersion in the observation process (BB and NB, leftmost and second-left x-axis ticks within each panel) have generally low bias. Continuous latent-state models (solid points) are only implemented for negative binomial and Poisson observational processes.

(coverage ≈ 0.8 for all observation process models and platforms) for forecasts and P_{rep}. For models without dispersion in transmission, models with dispersion in the observation process have low coverage (≈ 0.8) for most parameters, while the beta-binomial process model has low coverage (≈ 0.4) for P_{rep} and models without any dispersion have uniformly low coverage.

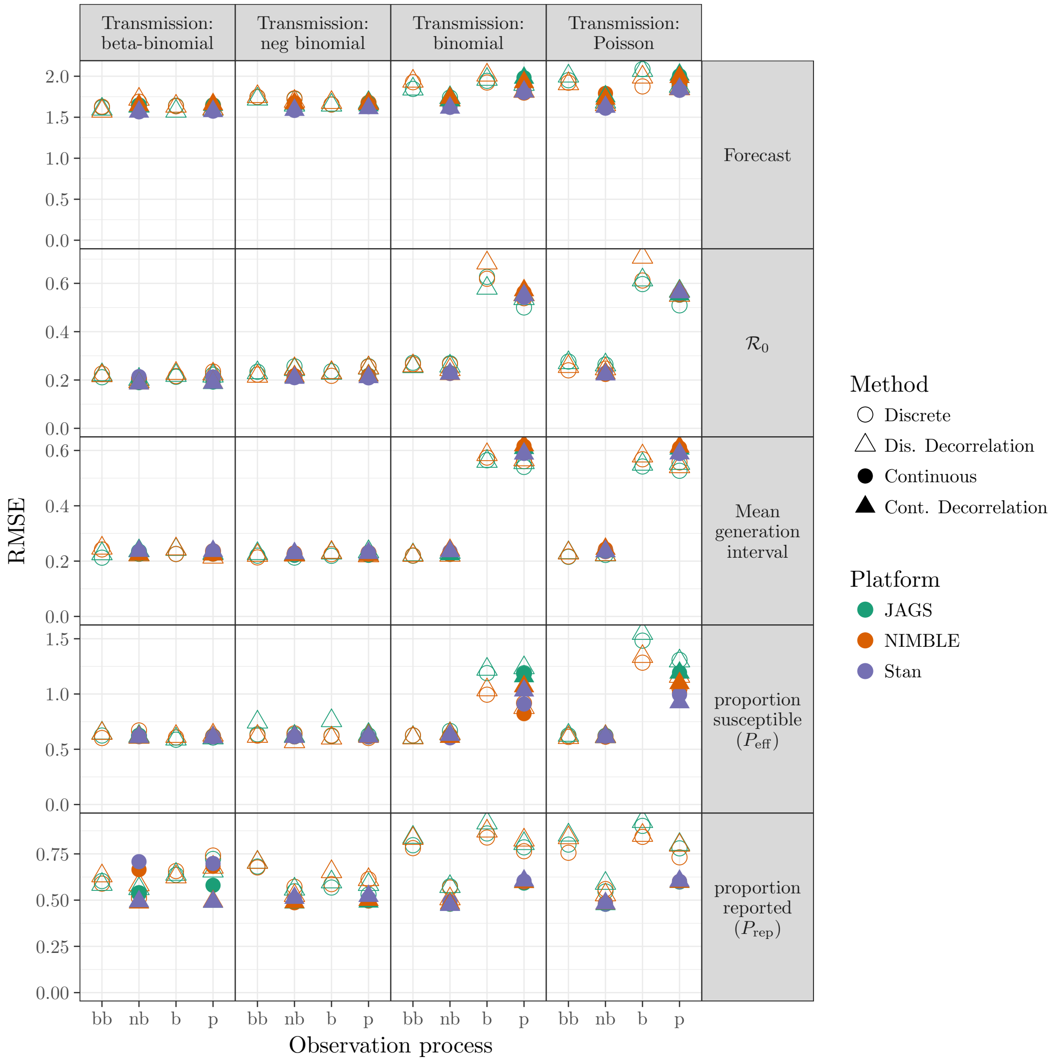


Figure 4: Comparison of RMSE (based on proportional errors) for all fitting model variants. The layout matches that of Figure 3. Patterns across models and platforms are similar to those seen in Figure 3. Short-term forecasts have generally high error, even when bias is low, reflecting inherent uncertainty in the system. The highly correlated parameters P_{eff} and P_{effrep} also show high error but not high bias.

There are substantial efficiency differences between transmission-process approaches (continuous vs. discrete), as measured by time per effective sample size, shown in Figure 6. For a given platform, models using continuous latent variables are generally more efficient than discrete latent processes. Comparing models with continuous latent variables between platforms (Figure 5, second and fourth column of every panel),

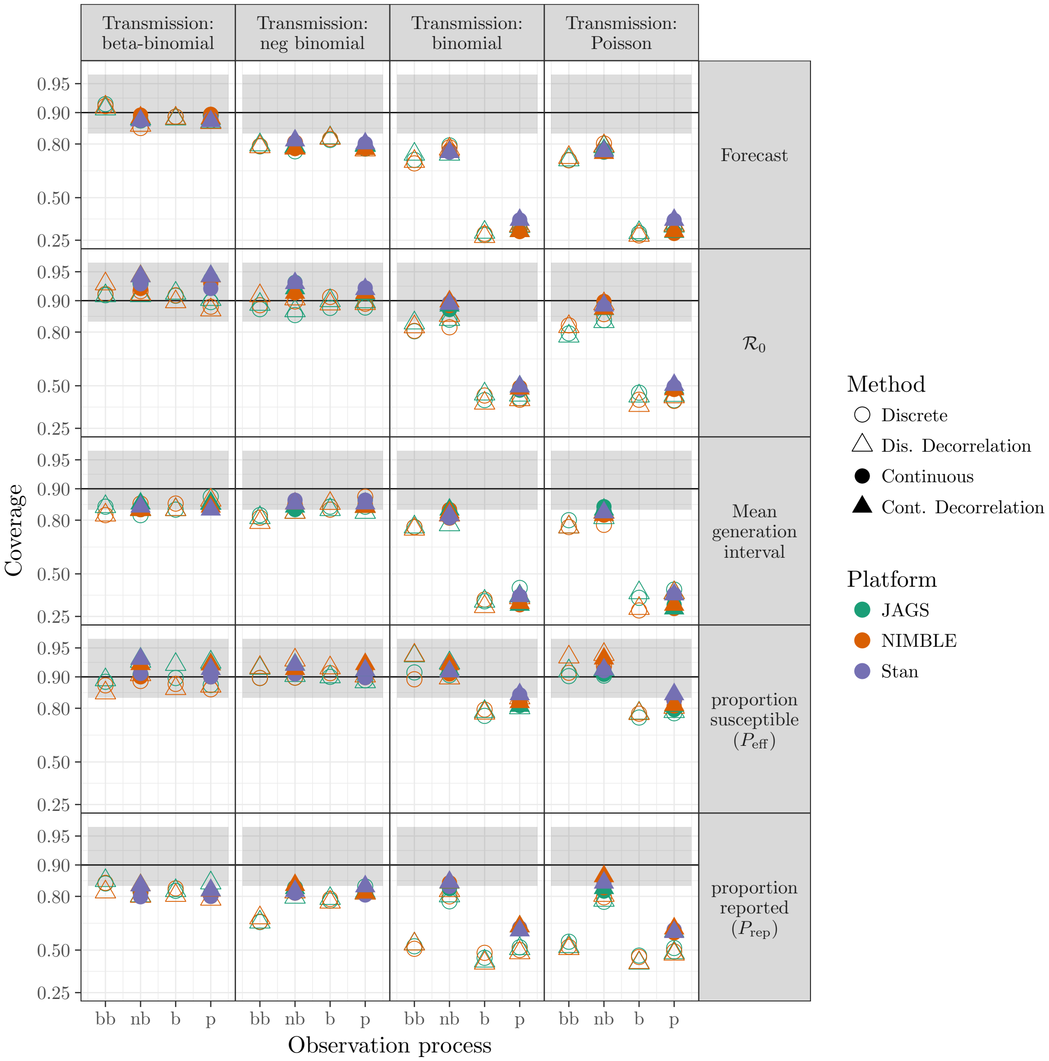


Figure 5: Comparison of coverage probability for forecast and parameters. Models with overdispersion in the transmission process (BB and NB, leftmost and second-left columns of panels) and models with overdispersion in the observation process (BB and NB, leftmost and second-left x-axis ticks within each panel) have coverage near the nominal value of 0.9 for all parameters and model variants. The black line shows the nominal coverage, and the grey ribbon the 95% binomial confidence interval based on 100 simulated fits. Vertical axis is plotted on a logit scale.

Stan (using HMC) is sightly more efficient for majority of the parameters, followed by NIMBLE and JAGS. Furthermore, continuous latent-variable models (especially using HMC in STAN) use fewer iterations (when meeting all convergence criterion described in section 2.3.2) than discrete latent-variable models.

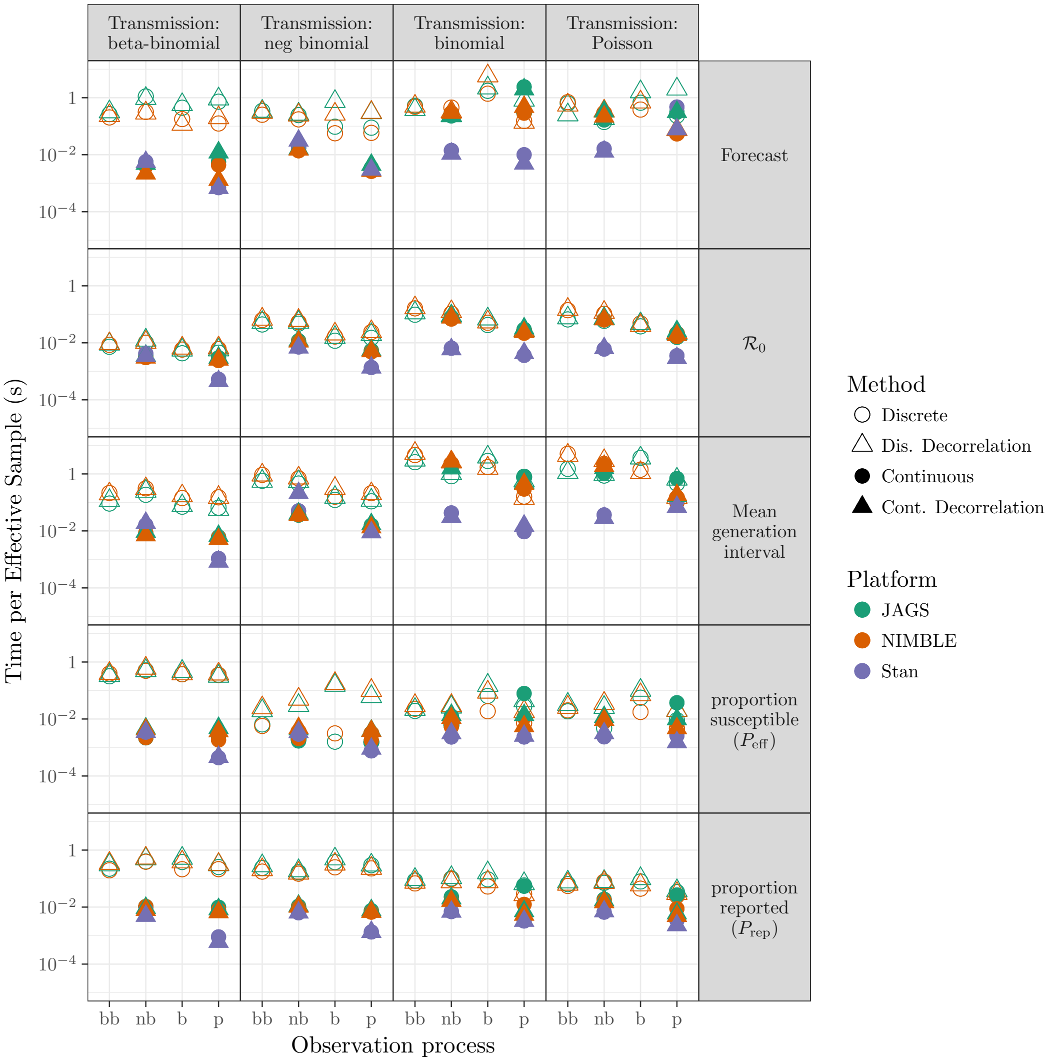


Figure 6: Comparison of efficiency for all fitting model variants: layout of models and platforms as in Figure 3.

4 Discussion

We have fitted models varying in complexity to simulated epidemic data with multiple sources of heterogeneity, using several different platforms. Using models that include some form of overdispersion is necessary for robust fits, but models that include overdispersion only in the transmission process can work as well as or better than the full model. Including overdispersion only in the observation process (if implemented as a negative binomial distribution) also provides relatively robust fits to these data. Simplifying the models by using continuous rather than discrete latent variables increased efficiency with little effect on the quality of the fits.

4.1 Ceilings

The effects of using distributions with ceilings (i.e. binomial and beta-binomial distributions) instead of their less realistic counterparts without ceilings (Poisson and negative binomial) was relatively small. In our framework, ceilings only apply in models with discrete latent variables; the primary effect of such ceilings is to reduce variance as probabilities (of infection or of sampling) become large. (Reporting-process models without ceilings also allow for false positives or over-reporting, which may be important in some contexts.)

4.2 Overdispersion

Accounting for overdispersion had more impact on our fits than the presence or absence of ceilings. In particular, models with no overdispersion in either process lacked flexibility and tended to be over-confident (that is, they showed low coverage). However, models that account for overdispersion in only one process (either transmission or observation) tended to be reliable for estimating parameters such as $\mathcal{R}_0$, mean generation interval, and short-term forecasts, particularly when overdispersion was implemented through negative binomial (a less constrained distribution than the beta binomial). However, parameters that are closely tied to the details of a particular model structure (such as the overdispersion parameters for the observation and transmission processes) must change when the overdispersion model changes, in order to compensate for missing sources of variability.

Several authors (e.g., (King et al., 2015; Taylor et al., 2016)) suggest that accounting for process as well as observation error in estimates of $\mathcal{R}_0$ and in forecasts is necessary in order to avoid over-confident estimates. Our exploration does not include any cases where process error is completely absent — even our “dispersion-free” processes incorporate sampling error in the process. However, we find that neglecting overdispersion can still lead to over-confident and unreliable estimates.

4.3 Reporting

In classic infectious disease models, reducing reporting rate and reducing the total effective population size have similar effects: reducing the observed size of the epidemic. While we want to make as few assumptions as possible about unobservable aspects of the epidemic, underreporting is of huge practical importance. Additionally, modeling observation error explicitly is required for reliable estimates of uncertainty (King et al., 2015). If reporting error is modeled with a ceiling, then underreporting is a necessary component of reporting error (i.e., reporting is always biased downward as well as noisy). Allowing overdispersion decouples the variance from the mean of the reporting process (i.e. the extra overdispersion parameter means that the variance is not determined by the mean).

Because reporting rate and effective population size play similar roles in epidemic dynamics, incorporating them both in a model may make their parameter estimates strongly correlated and hence difficult to identify: we may be very uncertain whether low observed epidemic incidence is driven by a small effective population size or a low reporting rate. We have addressed convergence problems arising from this issue by reparameterizing the model (Section 2.2.2). From a conceptual point of view, joint unidentifiability is not necessarily a serious problem, as long as the quantities we are most interested (such as $\mathcal{R}_0$) are identifiable. In practice, however, weak identifiability can cause hard-to-detect convergence problems; known-parameter simulations like those implemented here are useful for validation in such cases.

4.4 Extensions and alternative approaches

Our analysis covers classical MC (i.e. conditional updating of parameters via conjugate, slice, and Metropolis-Hastings samplers) and HMC approaches. Even within this scope there is additional room for analysis, both in terms of exploring important heterogeneities that we have neglected here (such as spatial, age and social structure), and in improving sampling techniques (e.g. by adjusting the choice of samplers in JAGS or NIMBLE or by redundant parameterization Gelman et al. (2014)).

More broadly, a plethora of other model-fitting tools is available to researchers, from deterministic optimization tools based on the Laplace approximation Illian et al. (2012); Kristensen et al. (2016) to sequential techniques such as iterated filtering and particle MC Del Moral et al. (2012); He et al. (2009); Yang et al. (2014). Ionides et al. (2006). These techniques can in principle be combined flexibly with the methods we explore here, e.g. using HMC to sample top-level parameters while implementing a sequential MC technique for the latent states. It will be interesting to see how the single-technique methods here compete with hybrid approaches, and how flexible toolboxes such as NIMBLE will fare against more focused platforms like Stan.

4.5 Prior distributions

This paper focuses on evaluating Bayesian methods for fitting and forecasting epidemics. For the purposes of evaluation we use parameter distributions for simulation that exactly match our Bayesian priors. We are assuming that researchers have a reasonable method of choosing appropriate Bayesian priors; in real applications this will be an important challenge.

5 Conclusion

We have presented a comparison of simple MCMC approaches to fit epidemic data. We learned two things about fitting epidemic data. First, modeling different processes with dispersion (BB and NB) is a naive but effective way to add uncertainty in the model; models that neglect such uncertainty are likely to be over-confident and less accurate at forecasting. Second, approximating discrete latent state process with continuous processes can aid efficiency without losing robustness of fit. This allows more efficient fitting in the classic framework (e.g., JAGS and NIMBLE), and also allows us to use the more advanced HMC technique (which we implemented via Stan).

6 Acknowledgments

We would like to thank the Ebola challenge organizers for organizing the Ebola model challenge that sparked our interest in this project, and Fred Adler and Michael Betancourt for thoughtful comments. McMaster University, NSERC Discovery grants and a CIHR Ebola grant provided funding.

References

Carpenter, B., A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell (2016). Stan: A probabilistic programming language. *J Stat Softw*.

Cook, S. R., A. Gelman, and D. B. Rubin (2006). Validation of software for bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics 15*(3), 675–692.

de Valpine, P., D. Turek, C. J. Paciorek, C. Anderson-Bergman, D. T. Lang, and

R. Bodik (2016). Programming with models: writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics* (just-accepted), 1–28.

Del Moral, P., A. Doucet, and A. Jasra (2012). An adaptive sequential monte carlo method for approximate bayesian computation. *Statistics and Computing 22*(5), 1009–1020.

Duane, S., A. D. Kennedy, B. J. Pendleton, and D. Roweth (1987). Hybrid monte carlo. *Physics letters B 195*(2), 216–222.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2014). *Bayesian data analysis*, Volume 2. Chapman & Hall/CRC Boca Raton, FL, USA.

Geman, S. and D. Geman (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence* (6), 721–741.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika 57*(1), 97–109.

He, D., E. L. Ionides, and A. A. King (2009). Plug-and-play inference for disease dynamics: measles in large and small populations as a case study. *Journal of the Royal Society Interface*.

Hoffman, M. D. and A. Gelman (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research 15*(1), 1593–1623.

Illian, J. B., S. H. Sørbye, and H. Rue (2012). A toolbox for fitting complex spatial point process models using integrated nested laplace approximation (inla). *The Annals of Applied Statistics*, 1499–1530.

Ionides, E. L., C. Bretó, and A. King (2006). Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences 103*(49), 18438–18443.

King, A. A., M. D. de Cellès, F. M. G. Magpantay, and P. Rohani (2015, May). Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola. *Proc. R. Soc. B 282*(1806), 20150347.

Kristensen, K., A. Nielsen, C. W. Berg, H. Skaug, and B. M. Bell (2016). TMB : Automatic differentiation and Laplace approximation. *Journal of Statistical Software 70*(5).

Ludwig, D. (1973). Mathematical models for the spread of epidemics. *Computers in biology and medicine 3*(2), 137–139.

Ludwig, D. and C. J. Walters (1985, June). Are Age-Structured Models Appropriate for Catch-Effort Data? *Canadian Journal of Fisheries and Aquatic Sciences 42*(6), 1066–1072.

Minsker, S., S. Srivastava, L. Lin, and D. Dunson (2014). Scalable and robust bayesian inference via the median posterior. In *International Conference on Machine Learning*, pp. 1656–1664.

Morris, W. F. (1997). Disentangling effects of induced plant defenses and food quantity on herbivores by fitting nonlinear models. *The American Naturalist 150*(3), 299–327.

Morton, A. and B. F. Finkenstädt (2005). Discrete time modelling of disease incidence time series by using markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 54*(3), 575–594.

Neal, R. M. (2003). Slice sampling. *Annals of statistics*, 705–741.

O'Neill, P. D., D. J. Balding, N. G. Becker, M. Eerola, and D. Mollison (2000). Analyses of infectious disease data from household outbreaks by markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 49*(4), 517–542.

O'Neill, P. D. (2002). A tutorial introduction to bayesian inference for stochastic epidemic models using markov chain monte carlo methods. *Mathematical biosciences 180*(1), 103–114.

Plummer, M. et al. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, Volume 124, pp. 125. Vienna.

Taylor, B. P., J. Dushoff, and J. S. Weitz (2016, November). Stochasticity and the limits to confidence when estimating of Ebola and other emerging infectious diseases. *Journal of Theoretical Biology 408*, 145–154.

Yang, W., A. Karspeck, and J. Shaman (2014). Comparison of filtering methods for the modeling and retrospective forecasting of influenza epidemics. *PLoS Comput Biol 10*(4), e1003583.

Supplementary Material

In the main text, we present the bias, RMSE, coverage and efficiency plots for aggregated forecast, $\mathcal{R}_0$, MGI, P_{eff}, and P_{rep}. Here, we present plots showing the other parameters (shape G_S and position G_P of the transmission kernel and process and observation overdispersion parameters δ_P and δ_{obs}) and disaggregated forecasts (five forecast steps) that are excluded in the main text. We also add some representative plots of the simulated cases and forecast.

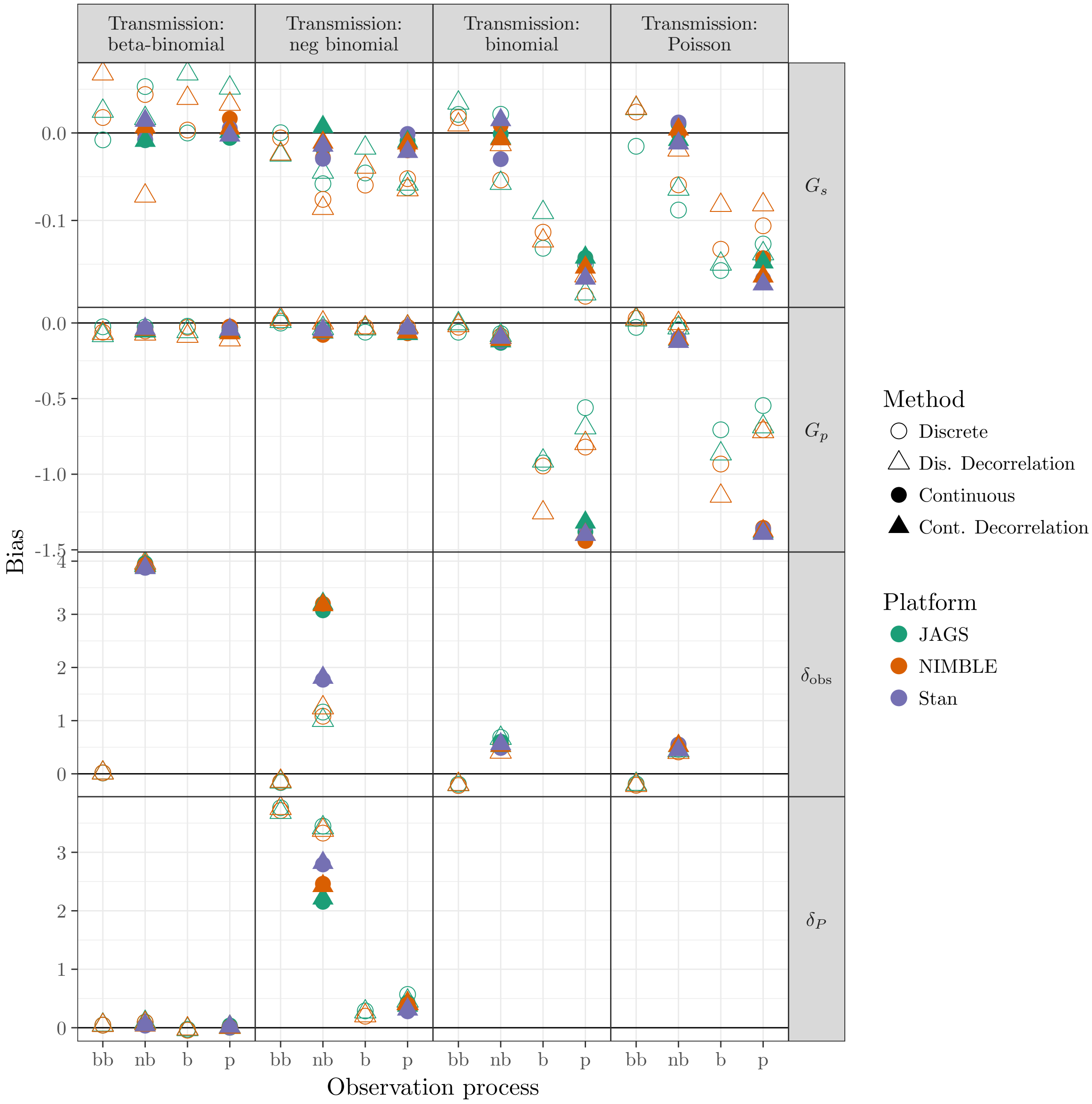


Figure S1: Comparison of bias for G_S (transmission shape), G_P (transmission position), δ_{obs} (observation overdispersion), and δ_P (process overdispersion: more detail given in Sect. 2.2) across different platforms (described in Sect. 2.3.1). Overdispersion parameter δ_P is only applicable in models with dispersion in the transmission process (first and second left column panel) and overdispersion parameter δ_{obs} is only applicable in models with dispersion in the observation process (first and second column within each column panel).

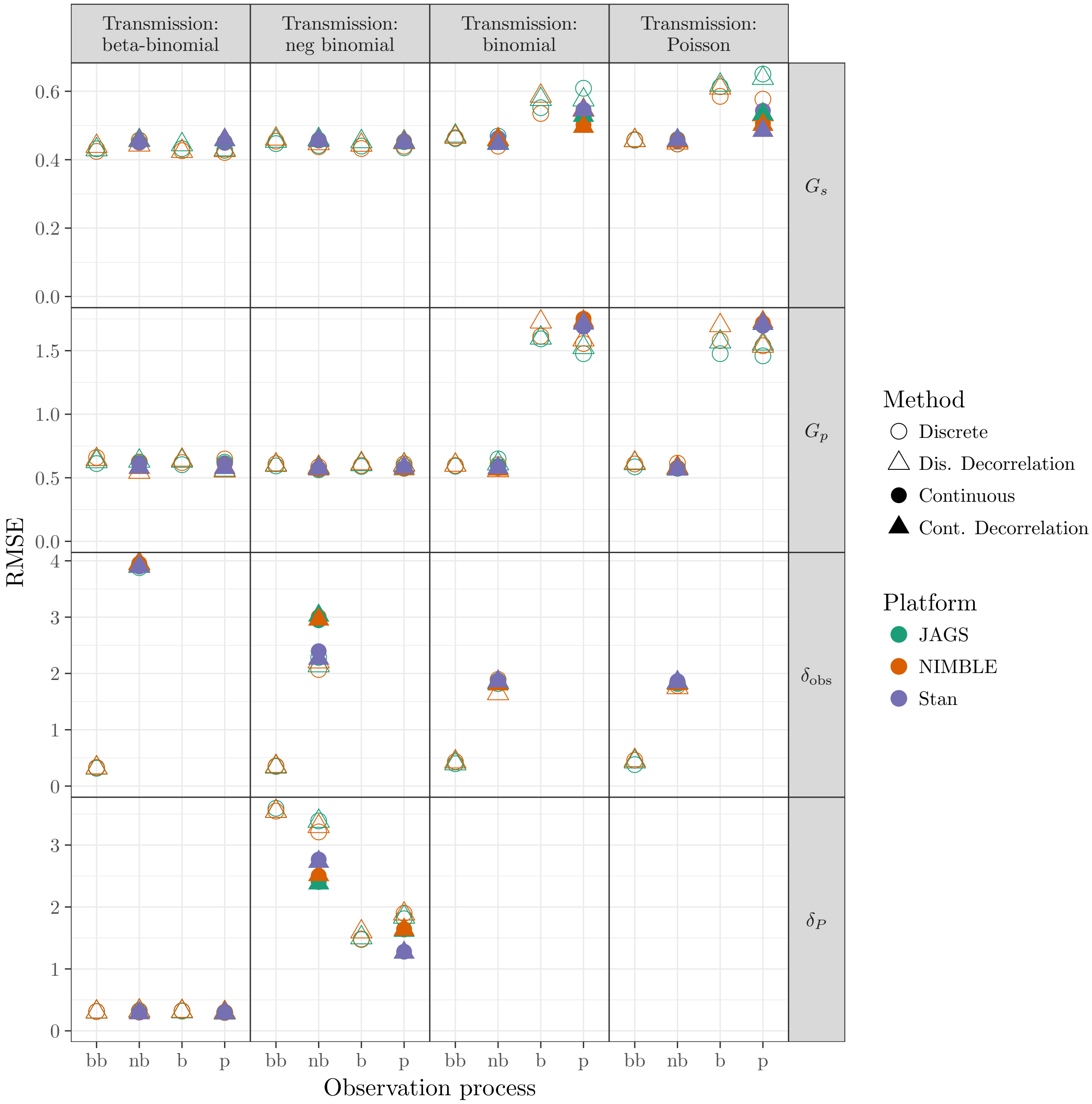


Figure S2: Comparison of RMSE for G_S, G_P, δ_{obs}, and δ_P. See Figure 4 in main text and Figure S1 in appendix for details.

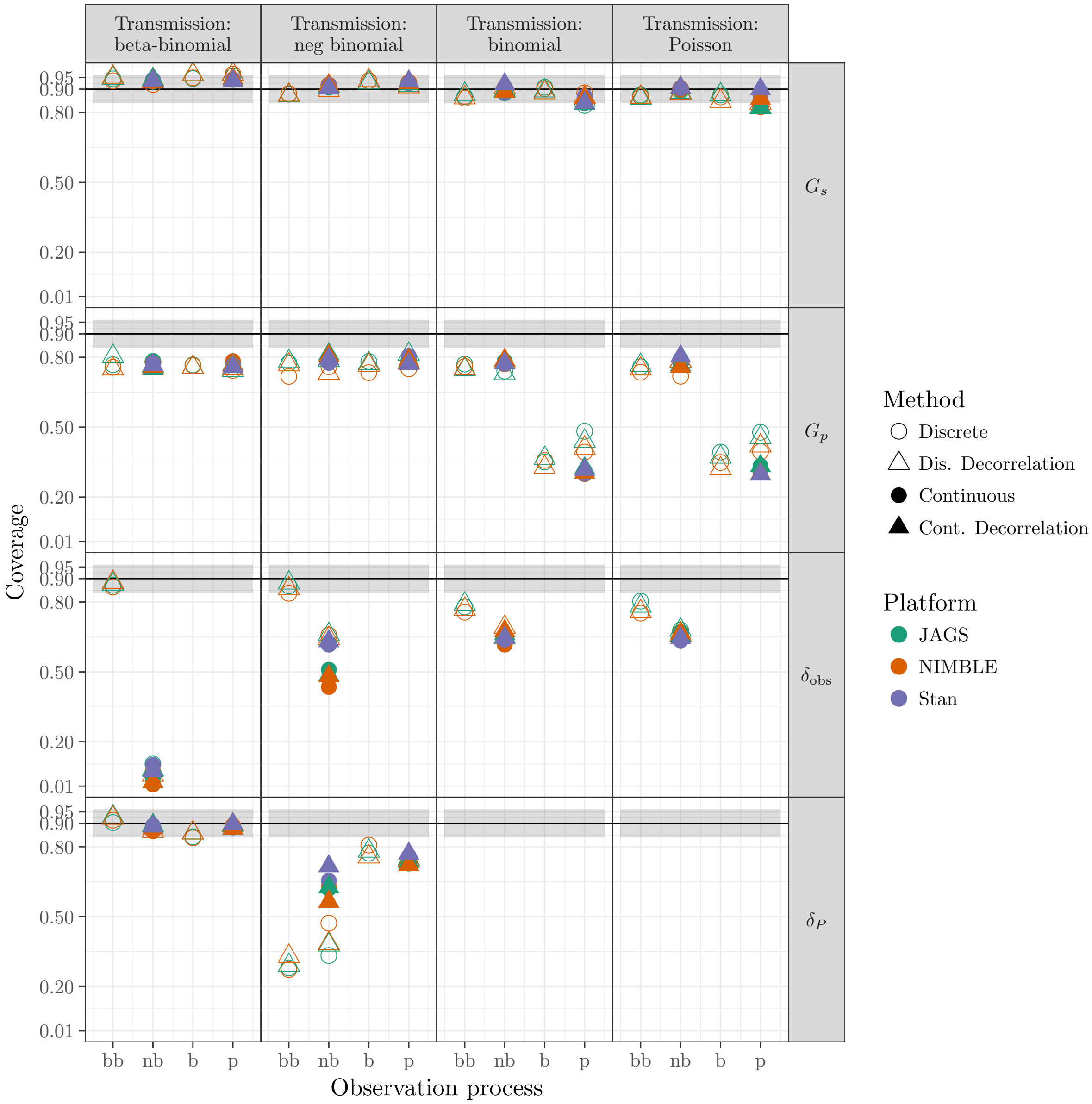


Figure S3: Comparison of coverage for G_S, G_P, δ_{obs}, and δ_P. See Figure 5 in main text and Figure S1 in appendix for details.

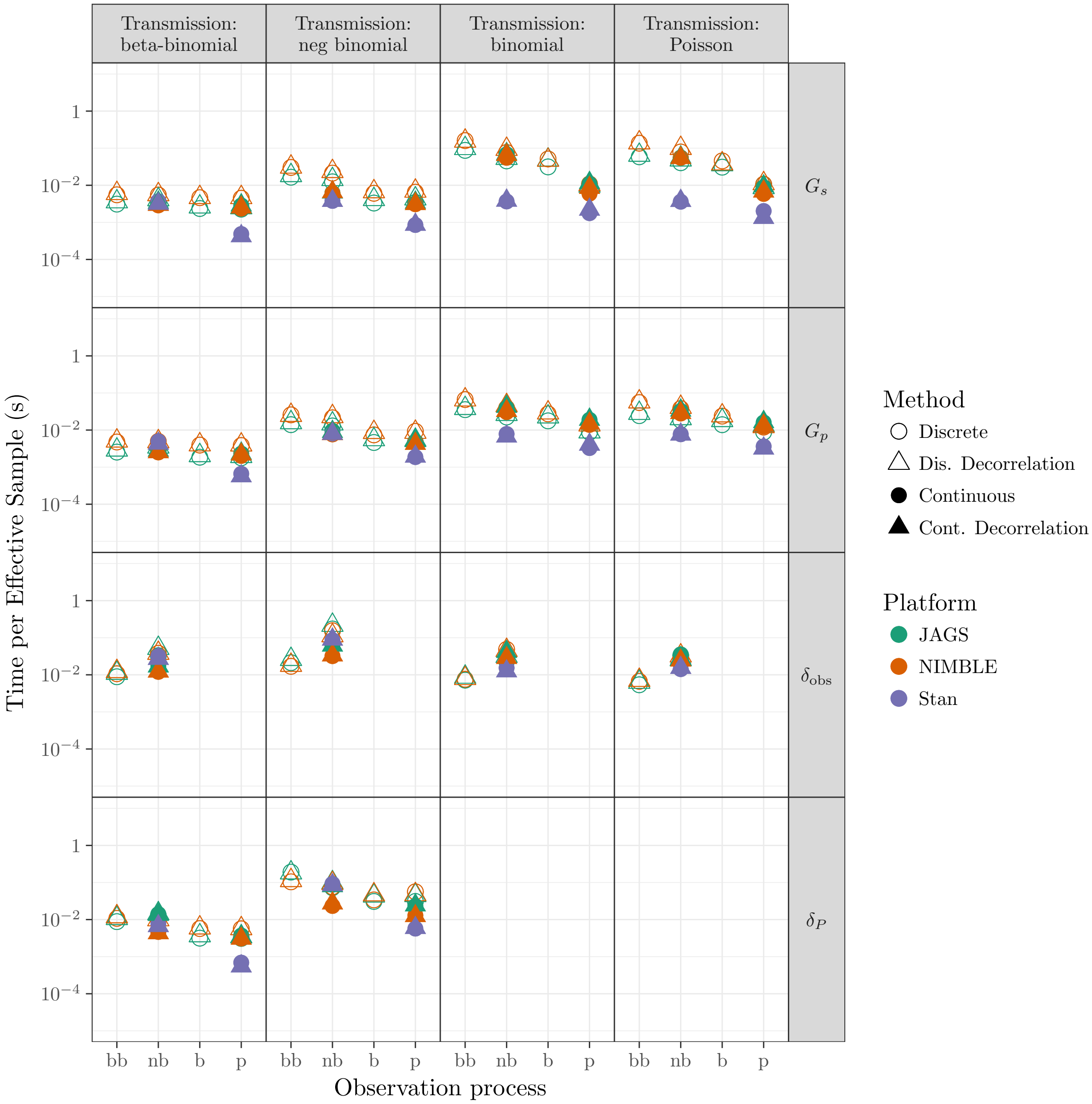


Figure S4: Comparison of coverage for G_S, G_P, δ_{obs}, and δ_P. See Figure 6 in main text and Figure S1 in appendix for details.

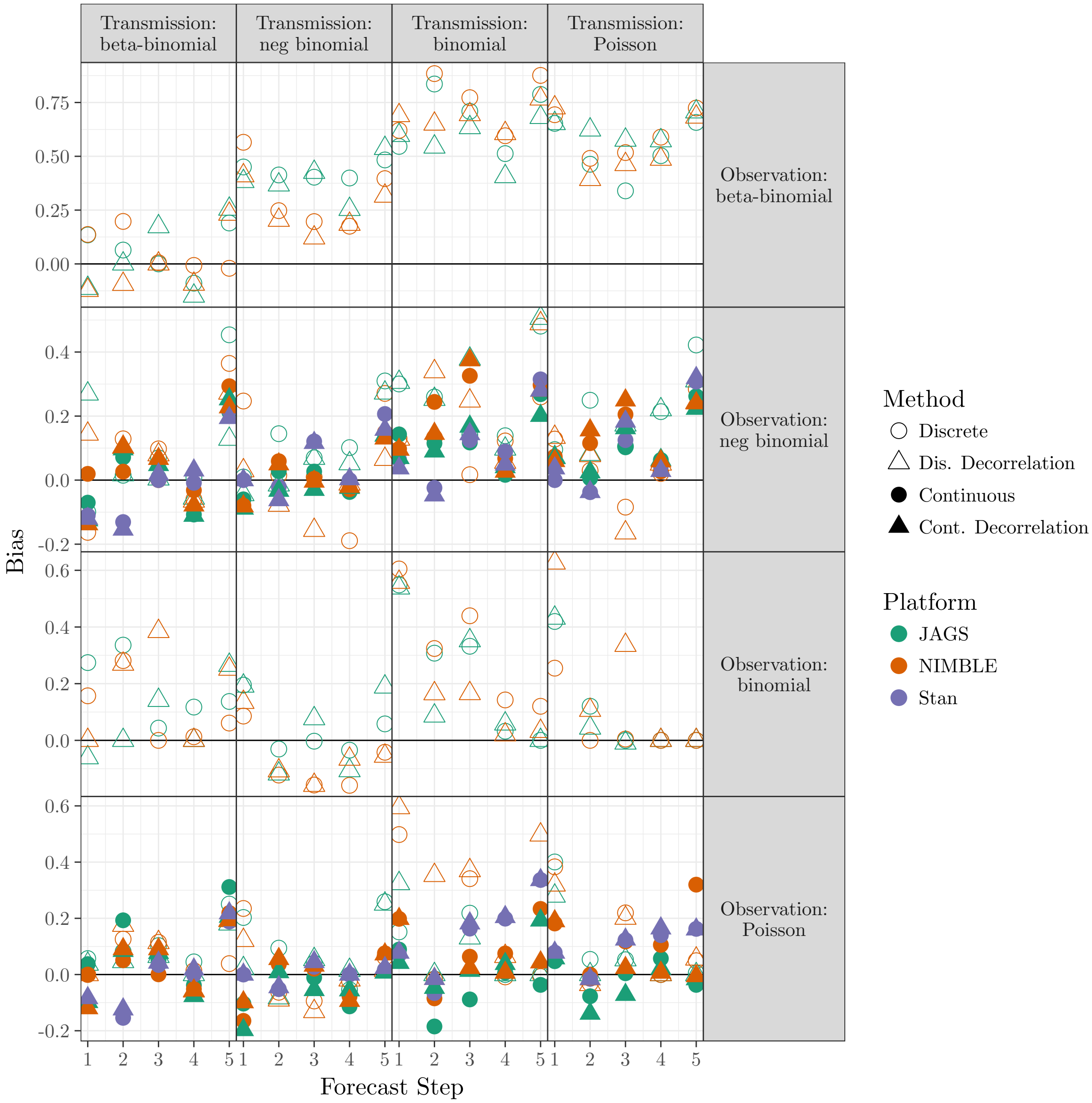


Figure S5: Comparison of bias for five forecast steps (described in Sect. 2.2) across different platforms (described in Sect. 2.3.1).

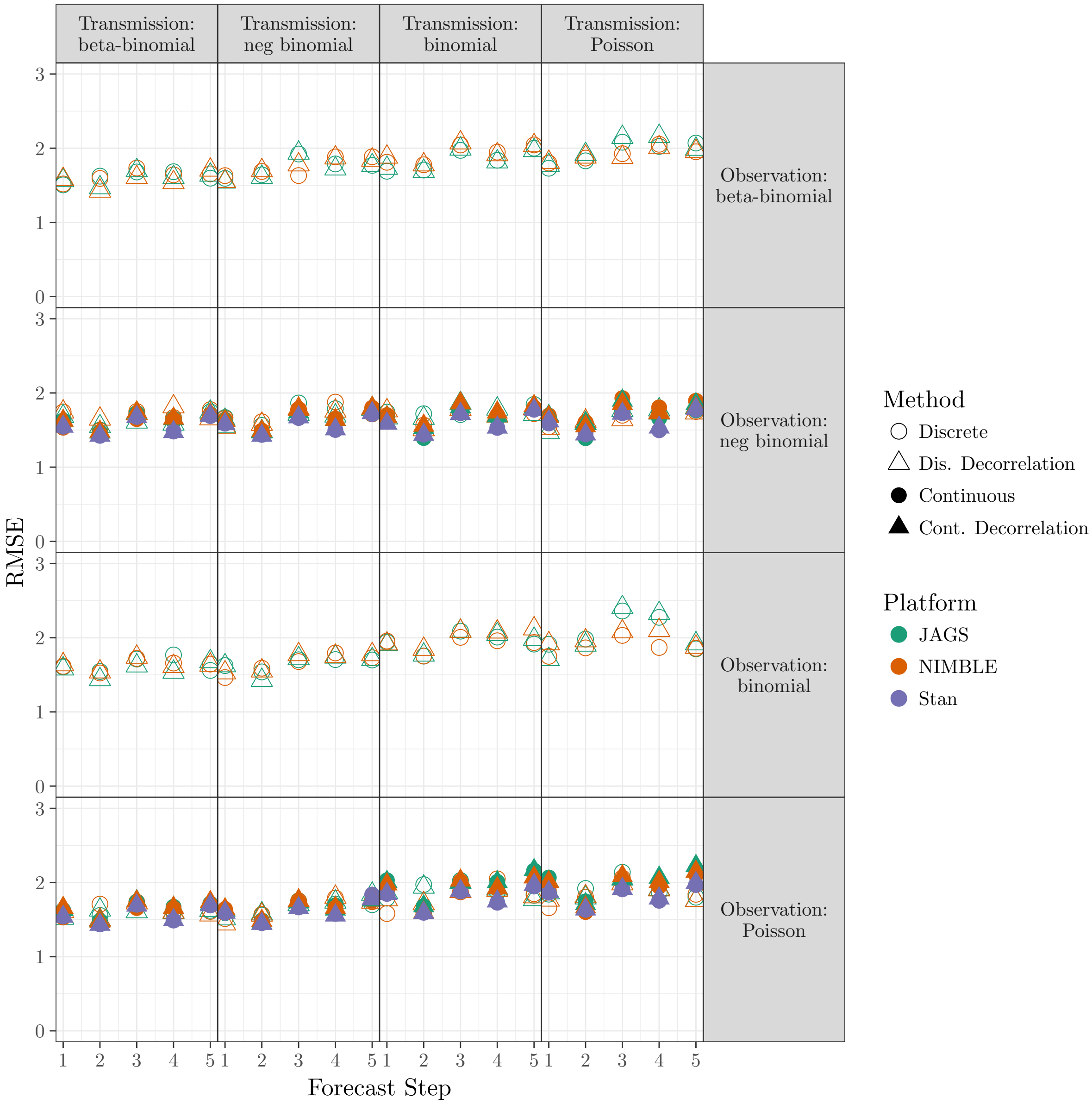


Figure S6: Comparison of RMSE for five forecast steps described in Sect. 2.2 across different platforms described in Sect. 2.3.1. See Figure 4 in the main text for details.

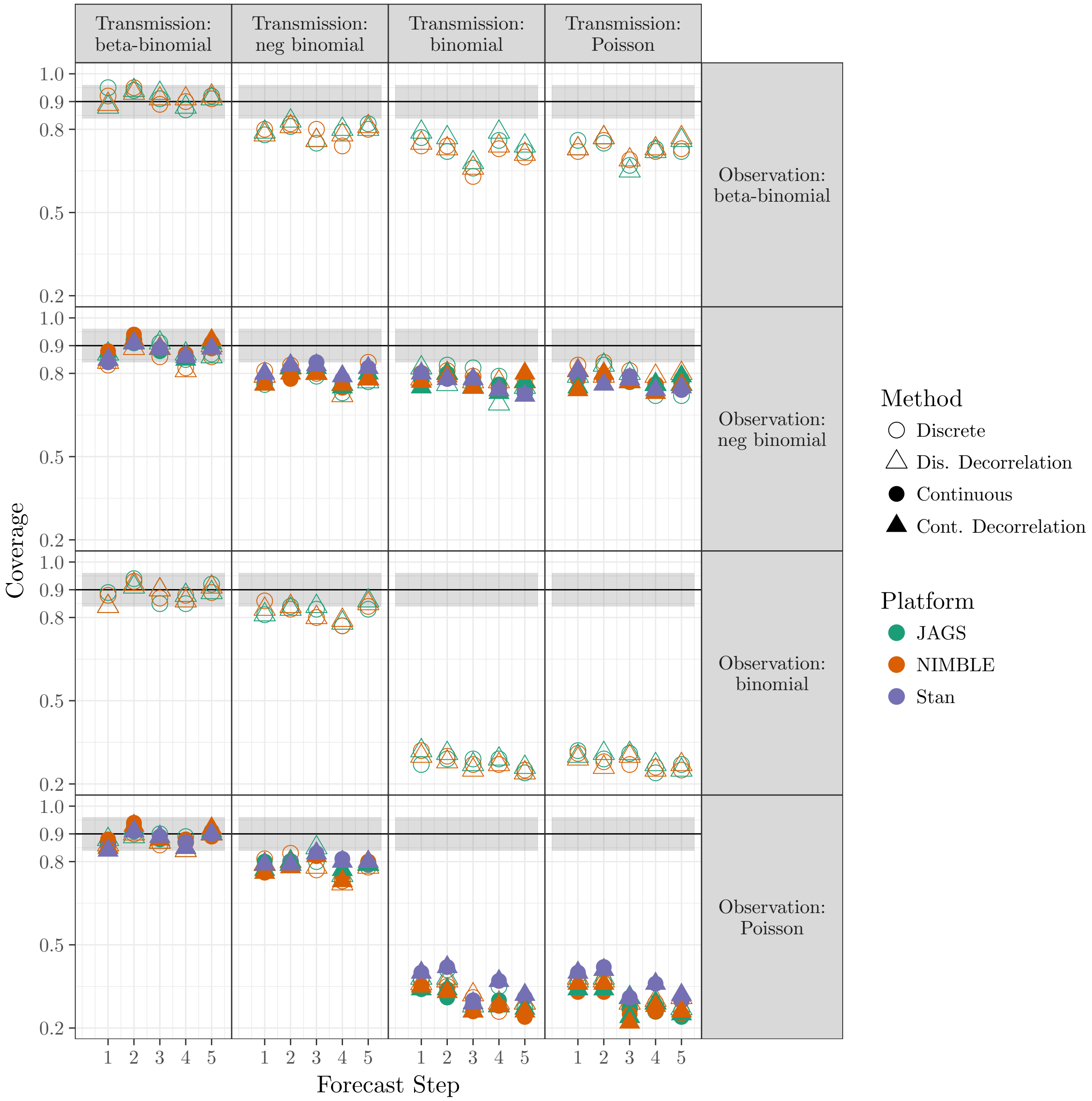


Figure S7: Comparison of coverage for five forecast steps described in Sect. 2.2 across different platforms described in Sect. 2.3.1. See Figure 5 in the main text for details.

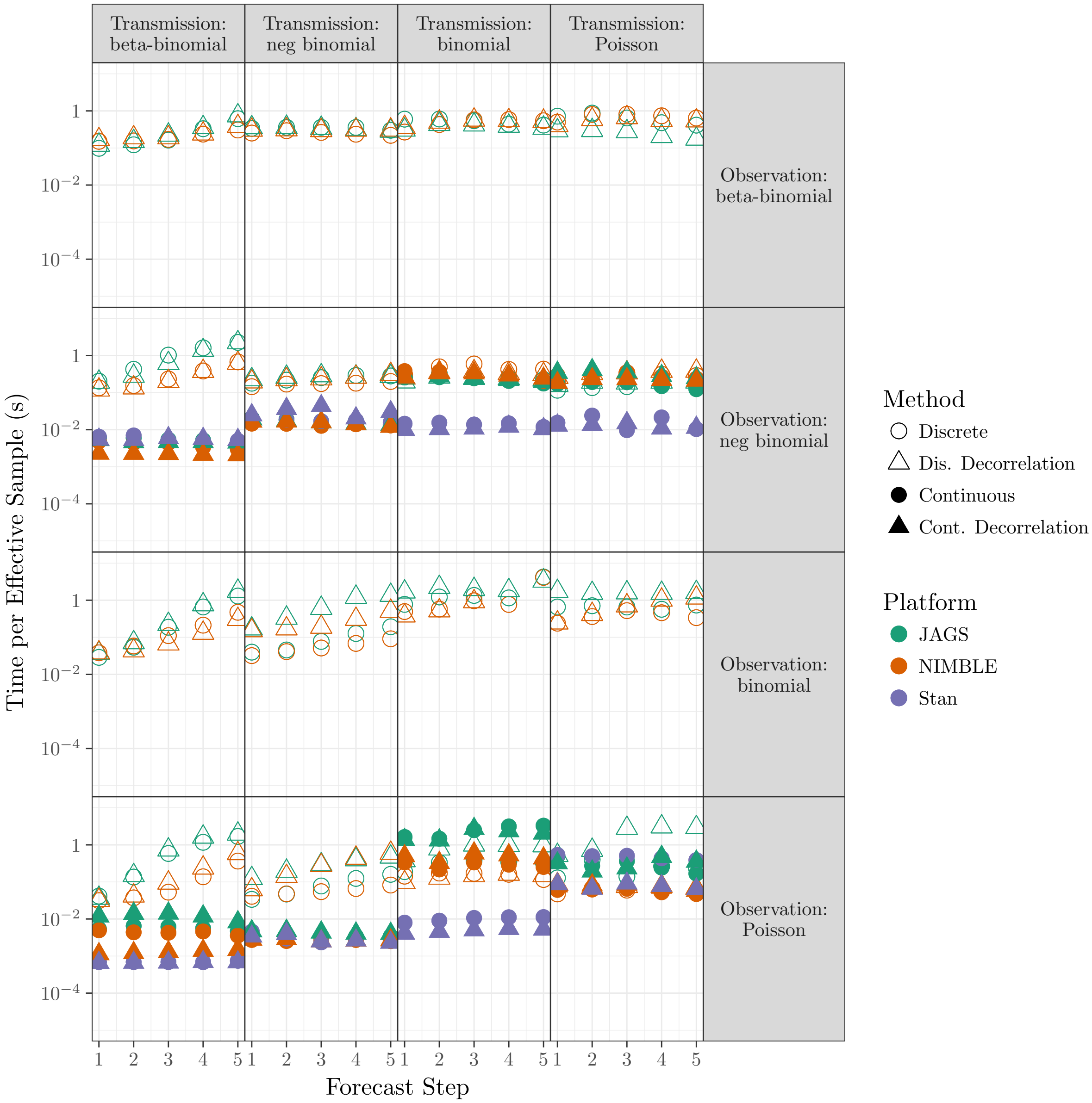


Figure S8: Comparison of sampling efficiency for five forecast steps described in Sect. 2.2 across different platforms described in Sect. 2.3.1. See Figure 6 in the main text for details.

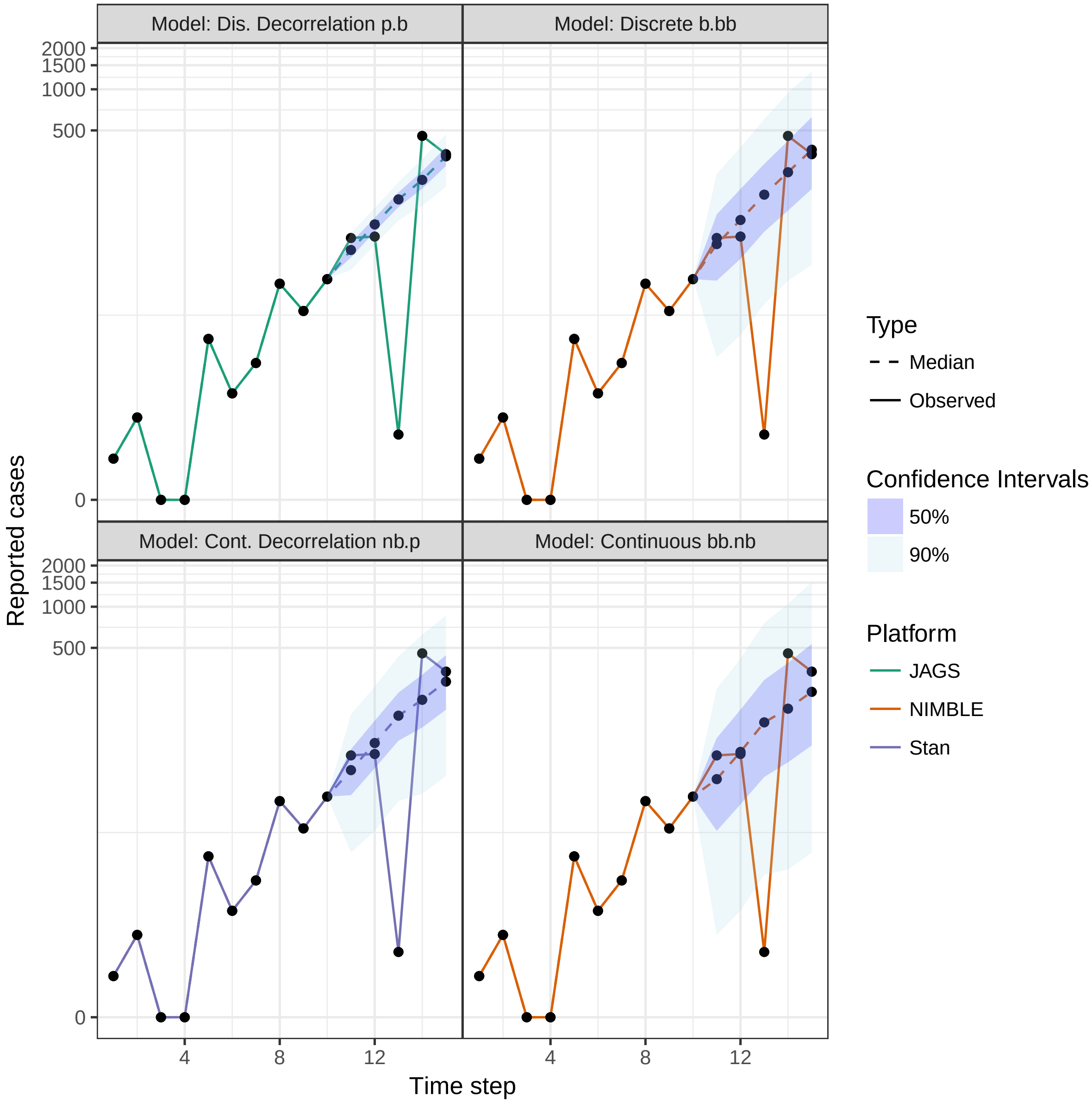


Figure S9: Comparison of forecast using combinations of transmission process, observation process, decorrelation, latent state variables, and platforms described in Sect 2.2 and 2.3.1. Moving from the top to bottom row adds overdispersion in the transmission process (binomial (b) and Poisson (p) to negative-binomial (nb) and beta-binomial (bb)). Moving from left to right adds overdispersion in the observations. Solid line shows the simulated observed cases (15 time steps); dashed line shows the median of the posterior forecast sample with 50% (dark ribbon) and 90% (light ribbon) confidence intervals (last 5 time steps).

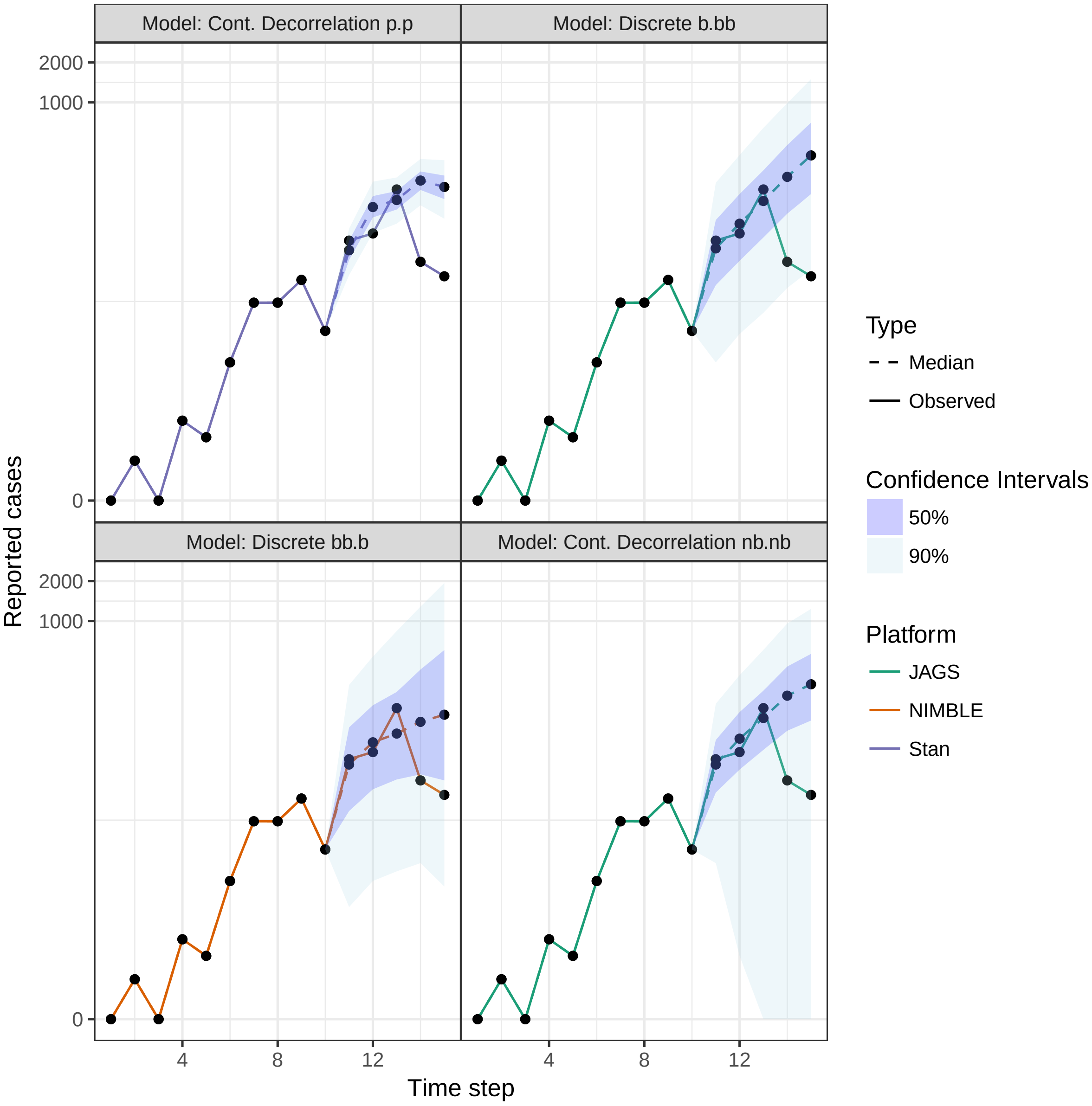


Figure S10: Comparison of forecast using a different set of parameters. See Figure S9 for details.

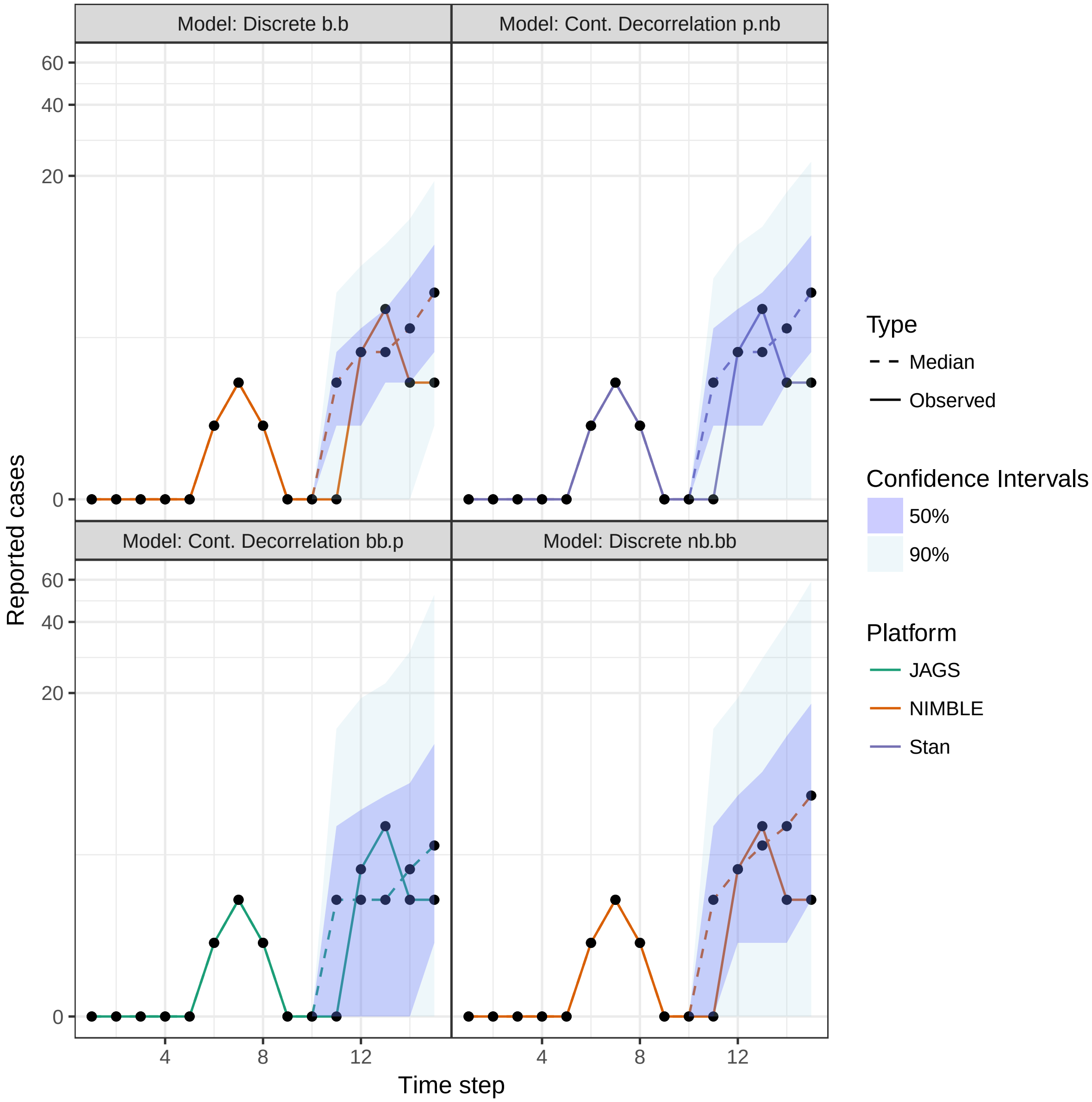


Figure S11: Comparison of forecast of low observed cases. See Figure S9 for details.

Tables

Table 1: Simulation model parameters

Parameter	Description	True	Prior
N	Total population size	Fixed at 100,000	NA
ℓ	Maximum length of the generation interval	Fixed at 5 time steps	NA
$\mathcal{R}_0$	Basic reproductive number	3	Gamma(shape=15,rate=5)
P_{eff}	Effective susceptible proportion of the population	0.5	Beta($\frac{B_{size}}{1-P_{\text{eff}}}, \frac{B_{size}}{P_{\text{eff}}}$)
P_{rep}	Reporting proportion	0.5	Beta($\frac{B_{size}}{1-P_{\text{rep}}}, \frac{B_{size}}{P_{\text{rep}}}$)
G_p	Position parameter for generation interval	0.5	Beta($\frac{2B_{size}}{1-G_p}, \frac{2B_{size}}{G_p}$)
G_s	Shape parameter for generation interval	1	Gamma(shape=5,rate=5)
δ_P	Beta Binomial transmission process dispersion	1	Gamma(shape=10,rate=10)
δ_{obs}	Beta-Binomial Observation process dispersion	1	Gamma(shape=10,rate=10)

Table 2: Fitting model parameters

Parameter	Description	True	Prior
N	Total population size	Fixed at 100,000	NA
ℓ	Maximum length of the generation interval	Fixed at 5 time steps	NA
B_{size}	Beta prior size factor	Fixed at 1	NA
$\mathcal{R}_0$	Basic reproductive number	3	Gamma(shape=15,rate=5)
P_{eff}	Effective susceptible proportion of the population	0.5	Beta($\frac{B_{size}}{1-P_{\text{eff}}}, \frac{B_{size}}{P_{\text{eff}}}$)
P_{rep}	Reporting proportion	0.5	Beta($\frac{B_{size}}{1-P_{\text{rep}}}, \frac{B_{size}}{P_{\text{rep}}}$)
P_{effrep}	Proportion of effective S to I that are observed	$P_{\text{eff}} \times P_{\text{rep}}$	Beta($\frac{B_{size}}{1-P_{\text{effrep}}}, \frac{B_{size}}{P_{\text{effrep}}}$)
ρ	Scale splitting factor	0.5	Beta($\frac{B_{size}}{1-\rho}, \frac{B_{size}}{\rho}$)
G_p	Position parameter for generation interval	0.5	Beta($\frac{2B_{size}}{1-G_p}, \frac{2B_{size}}{G_p}$)
G_s	Shape parameter for generation interval	1	Gamma(shape=5,rate=5)
δ_P	Beta Binomial transmission process dispersion	1	Gamma(shape=10,rate=10)
δ_P (Neg-Binom)	Negative-Binomial Transmission process dispersion	NA	Uniform(min=0,max=100)
δ_{obs}	Beta-Binomial Observation process dispersion	1	Gamma(shape=10,rate=10)
δ_{obs} (Neg-Binom)	Negative-Binomial Transmission process dispersion	NA	Uniform(min=0,max=100)