Rejoinder for Discussion on "Horseshoe-Based Bayesian Nonparametric Estimation of Effective Population Size Trajectories"

James R. Faulkner[1,2], Andrew F. Magee[3], Beth Shapiro[4,5], Vladimir N. Minin[6,*]

[1]Quantitative Ecology and Resource Management, University of Washington, Seattle, WA 98195, U.S.A.

[2]Fish Ecology Division, Northwest Fisheries Science Center, National Marine Fisheries Service, NOAA, Seattle, WA 98112,

[3]Department of Biology, University of Washington, Seattle, WA 98195, U.S.A.

[4]Ecology and Evolutionary Biology Department and Genomics Institute, University of California Santa Cruz, Santa Cruz, CA 95064, U.S.A.

[5]Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, CA 95064, U.S.A.

[6]Department of Statistics, University of California Irvine, Irvine, CA 92697, U.S.A.

[*]Corresponding author: vminin@uci.edu

This paper has been submitted for consideration for publication in *Biometrics*

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/biom.13273

1. Introduction

We thank the discussants for their thoughtful examination of our paper. Their discussion addresses some important points about assessing model fit with posterior predictive checks. It also investigates the ability of the horseshoe Markov random field (HSMRF) models to detect rapid changes in effective population size. In what follows we provide some commentary on the discussants' results and offer some additional examples that expand on their findings.

2. Posterior Predictive Checks

We agree that posterior predictive checks are an important part of assessing the goodness of fit of particular models. These checks can be used in addition to model selection criteria to probe model adequacy. Posterior predictive checks are based on examining predictive distributions of test statistics (or discrepancy measures more generally). However, we argue that in our phylodynamic context most test statistics and discrepancy measures applied to the entire data set will result in similar goodness of fit among competing models, and we show this is especially true when using statistics based on sequence data. We suggest that locally-targeted test statistics may be more informative for assessing model adequacy when such statistics are available.

The discussants provided examples for fixed coalescent times that used the total tree length and the time to most recent common ancestor (TMRCA) as test statistics. The differences in posterior predictive p-values among the models were small for these metrics. They mention that other meaningful test statistics may offer better ability to measure model adequacy and to discriminate among alternative models. We agree with this statement and think that it would be more informative to use discrepancy measures that target segments of time when population trends display specific features of interest. For example, the timing of the start of the epidemic and the period of increase were of most interest in our analysis of the HCV data.

A set of custom discrepancy measures that target such specific time periods would allow for better probing of the adequacy of the reconstructed effective population size trajectory in those regions of interest.

We demonstrate the effectiveness of such local discrepancy measures on a set of simulated coalescent times. We use a piecewise constant effective population size trajectory that exhibits a rapid change in population size. The scenario has a population size of 1.0 between times 0 and 4 and size 0.1 after time 4, where time is in units before the present. We simulated coalescent times using 500 samples, where 50 were sampled at time 0 and the remaining 450 sampled uniformly between times 0 and 8. We used a grid with 100 grid cells and parameterized the models following methods described in the Web appendices. The relatively large sample size and uniform sampling allowed us to make more distinct comparisons among methods.

For global application to the entire data set, we used the negative log-likelihood (NLL) for coalescent times as a discrepancy measure, which depends on both the parameters and the data. Note that we also investigated the Kolmogorov-Smirnov statistic developed by Karcher et al. (2019) for coalescent times, and the TMRCA, but these gave similar results to the NLL so we did not pursue them further. For the local discrepancy measures, we used the NLL and the proportion of coalescent (PCoal) events among genealogy lineages present during the k consecutive grid cells forming time interval $[x_h, x_{h+k}]$. Here $\text{PCoal}_{x_h,x_{h+k}} = c_{x_h,x_{h+k}}/(n_{x_h} + m_{x_h,x_{h+k}})$, where x_is are grid cell boundaries, n_{x_h} is the number of lineages at time x_h, and $c_{x_h,x_{h+k}}$ and $m_{x_h,x_{h+k}}$ are the number of coalescent events and number of new samples during the time period x_h to x_{h+k}, respectively, for $h \in \{1, \ldots, H+1-k\}$. The discrepancy measure $\text{NLL}_{x_h,x_{h+k}}$ is the coalescent log-likelihood calculated for k consecutive grid cells $[x_h, x_{h+1}], \cdots, [x_{h+k-1}, x_{h+k}]$. These discrepancy measures can be calculated for individual grid cells or for a sequence of grid cells. We applied the NLL to three consecutive grid cells

covering the times from 4.0 to 4.24, which in our simulations is the period right before the population size change. We applied the PCoal statistic to the single grid cell covering the time period of 4.0 to 4.08. We drew 1,000 random samples from the posterior distribution for each model. For each draw, we simulated a posterior predicted set of coalescent times and calculated each discrepancy measure for the observed and replicated data.

The global NLL discrepancy measure showed little difference among the models and in all cases showed satisfactory performance (Figure 1). Recall that posterior predictive p-values close to 0.5 indicate adequate performance. However, both the NLL and PCoal statistics applied to the specific time periods of interest showed that the GMRF models and the HSMRF-2 model did not perform very well. The HSMRF-1 model was satisfactory for the local NLL statistic but was not really adequate for the PCoal statistic, although it was much better than the other methods. These results indicate that posterior predictive checks based on locally-targeted discrepancy measures can provide useful information about model adequacy for specific time periods of interest.

[Figure 1 about here.]

The discussants mentioned the possibility of using posterior predictive checks based on sequence data. We worry that the sequence data will generally be far removed from the effects of the tree prior. Namely, variation in the rate of molecular evolution along branches may obscure differences that exist between the models in inferred divergence times. To investigate the applicability of sequence-based posterior predictive checks, we thus focus on the HCV example where the molecular clock rate is fixed and the coalescent times should be most directly linked to patterns of variation in the alignment. We employ five different posterior predictive checks to this dataset using the posteriors for the HSMRF and GMRF models (of both first and second order). The multinomial test statistic of Goldman (1993), M, summarizes the diversity of site patterns in the alignment. The average number of distinct

nucleotides observed at a site in the alignment, also called the "biochemical diversity," BCD (Lartillot et al., 2007), is a measure of the overall variability in the sequence data and is therefore (somewhat indirectly) related to the overall tree length. The proportion of invariant sites, PI, observed in an alignment, is a summary statistic related to Watterson's estimator of the effective population size. Tajima's D statistic (Tajima, 1989), D, is sensitive to the relative branch lengths leading to tips compared to more deeply nested branches. The number of "singleton" sites, S, in the alignment, sites where a single sampled sequence differs from all others, is a measure of how long the tip branches of the tree are (as these sites are most likely the result of a single mutation along a terminal branch).

Applied to the HCV dataset, all five statistics produce very similar model adequacy results across all four models (Figure 2). Unfortunately, with our sequence-based posterior predictive checks it is hard to target specific regions of the coalescent trajectory. Therefore, we remain skeptical about usefulness of posterior predictive checks based on sequence alignment summary statistics.

[Figure 2 about here.]

3. Detection of Rapid and Gradual Changes in Population Size

In Section 3, the discussants investigated the ability of the GMRF and HSMRF models to detect rapid changes in population size. They did this by testing whether the models could distinguish between the true rapidly-changing trajectory and ones shifted in time by varying amounts. The HSMRF-1 model did the best at capturing the true trajectory for the scenarios they tested. Their results support our finding from simulated data that HSMRF-1 is more effective at recovering rapid changes in effective population size than the other methods we investigated.

We expand on the discussants' results with simulated examples that allow us to see

behaviors of different models under periods of gradual change and periods of rapid change. We use a set of effective population size trajectories similar to the birth-rate trajectories used by Magee et al. (2019) for investigating the ability of HSMRF birth-death models to capture slow and fast changes in rates. These are 1) a constant population size of 0.55, 2) a piecewise constant population size of 1.0 between times 0 and 4 and 0.1 after time 4, 3) a combination of constant population size and log-linearly changing population size, where population size is 1.0 between time 0 and 2.5, log-linearly decreasing between times 2.5 and 5.5, and 0.1 after that, and 4) a log-linearly decreasing population size from 1.0 to 0.1 between times 0 and 8 and constant at 0.1 after that. We simulated coalescent times using 500 samples, where 50 were sampled at time 0 and the remaining 450 sampled uniformly between times 0 and 8. We used a grid with 100 grid cells and parameterized the models following methods described in the Web appendices.

We show results from a single set of simulated data for each scenario that produced behaviors representative of those seen across multiple simulations (Figure 3). For the constant trajectory, we see that the GMRF models of both orders resulted in less variable estimated trajectories and narrower credible intervals than the corresponding HSMRF models of the same order. This is expected behavior because the HSMRF models pay the price for their additional flexibility. The piecewise constant trajectory has an instantaneous jump at time 4 that was best captured by the HSMRF-1 model and next best by the HSMRF-2 model. The GMRF models with their single variance parameter controlling the magnitude of the increments of log effective population size have to increase this variance to make the jump at the expense of increasing variance everywhere else along the trajectory. In contrast, the local flexibility of the HSMRF models allows for large jumps in some places and small variation in others. For the log-linear-constant trajectory and the log-linear trajectory, the models of the same order perform similarly, with the order-2 models best capturing the period of log-linear

population change. This is expected because the order-2 models produce piecewise-linear trajectories. One feature to note is the stair-stepping behavior of the order-1 models over the periods of log-linear change in population size, which is more pronounced with the HSMRF model. This behavior follows from the piecewise-constant trajectories generated by the order-1 models. The more pronounced stair-stepping of the HSMRF-1 models is a result of the shrinkage properties of the horseshoe priors. The HSMRF-1 priors will favor periods of small change followed by periods of rapid change when population change is relatively rapid, but this behavior will be less pronounced in periods of gradual population change. However, we have found that this stair-stepping behavior is typically not problematic when one considers the coverage of the credible intervals. These simulated examples provide evidence that rapid changes in effective population sizes are best captured by HSMRF models, and while they may be unnecessarily flexible in situations with constant or gradually changing population sizes, they can still provide reliable estimates when uncertainty measures are considered.

[Figure 3 about here.]

References

Goldman, N. (1993). Statistical tests of models of DNA substitution. *Journal of Molecular Evolution* **36,** 182–198.

Karcher, M. D., Suchard, M. A., Dudas, G., and Minin, V. N. (2019). Estimating effective population size changes from preferentially sampled genetic sequences. *arXiv preprint arXiv:1903.11797* .

Lartillot, N., Brinkmann, H., and Philippe, H. (2007). Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evolutionary Biology* **7,** S4.

Magee, A. F., Höhna, S., Vasylyeva, T. I., Leaché, A. D., and Minin, V. N. (2019). Locally

adaptive Bayesian birth-death model successfully detects slow and rapid rate shifts. *bioRxiv 853960, doi: 10.1101/853960* .

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123,** 585–595.

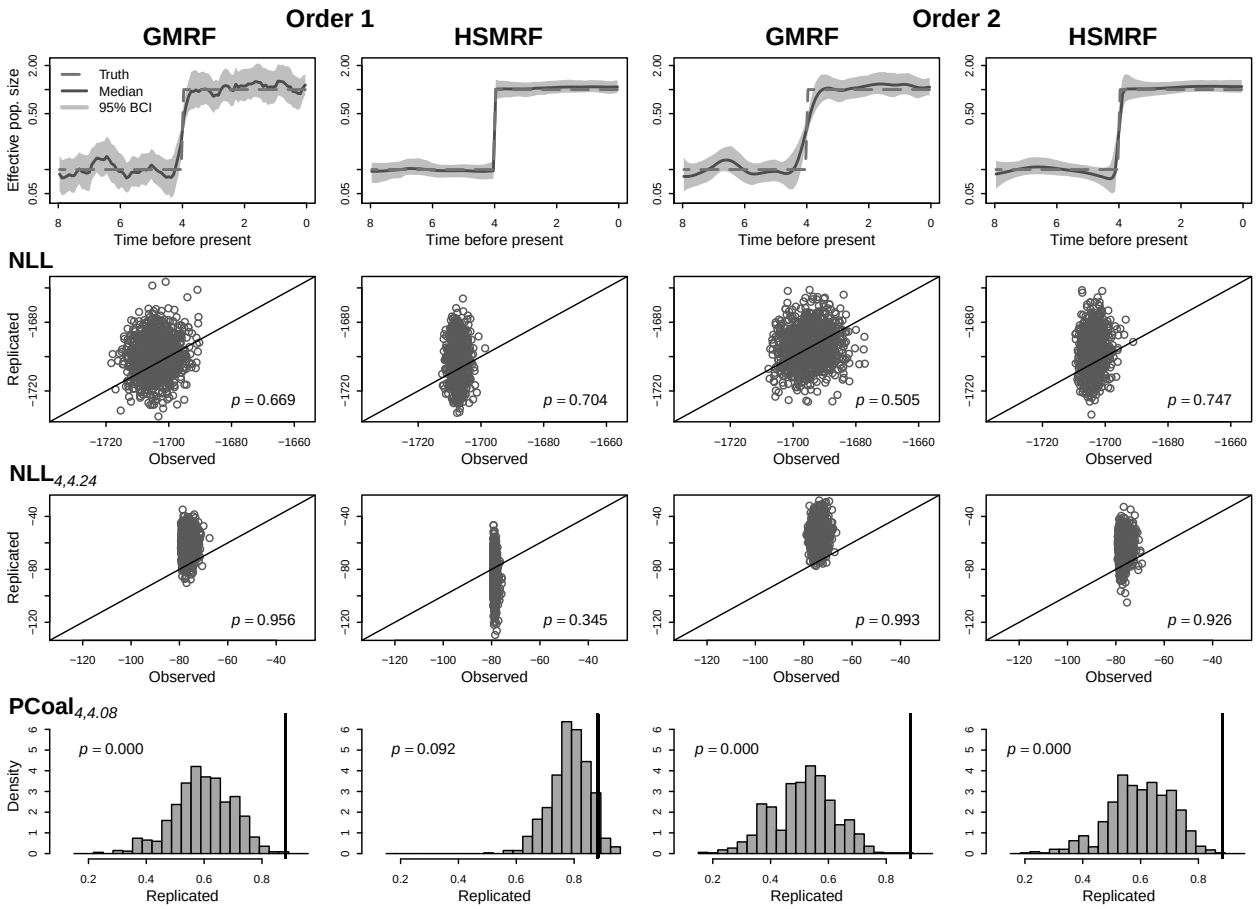

Figure 1. Posterior predictive checks for first- and second-order Gaussian Markov random field (GMRF) and horseshoe Markov random field (HSMRF) models for a piecewise-constant effective population size trajectory with simulated coalescent times. Top row shows the true effective population size trajectories that generated the data (dashed line), posterior medians of estimated trajectories (solid line) and associated 95% Bayesian credible intervals (shaded band). Remaining rows show posterior predictive test values for observed and replicated data and associated one-sided posterior predictive p-values for overall negative log-likelihood (NLL), negative log-likelihood for grid cells between time 4.0 and 4.24 ($\text{NLL}_{4,4.24}$), and the proportion of lineages resulting in coalescent events in the grid cell covering time 4.0 to 4.08 ($\text{PCoal}_{4,4.08}$). The vertical bars in the plots for $\text{PCoal}_{4,4.08}$ show the test statistic values for the observed data.

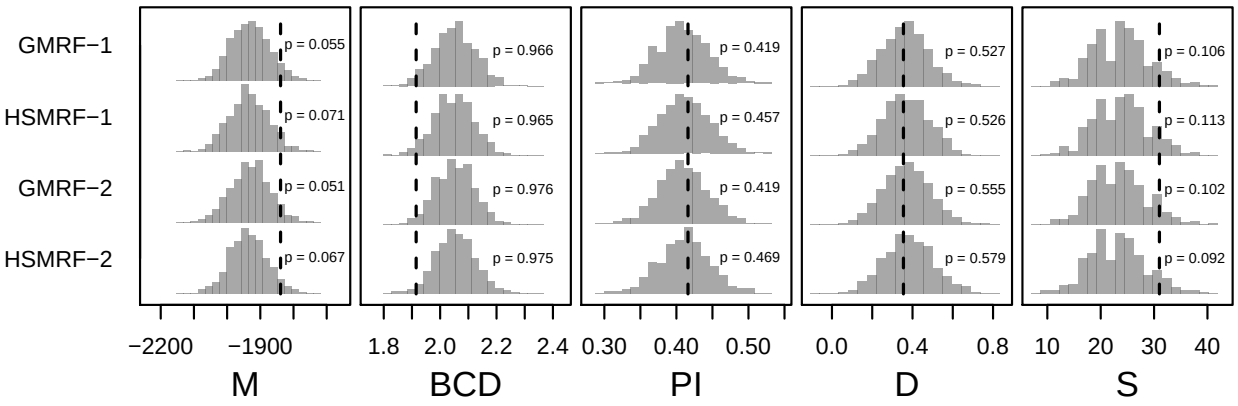

Figure 2. Posterior predictive model checks using sequence data for first- and second-order Gaussian Markov random field (GMRF) and horseshoe Markov random field (HSMRF) models. Predictive distributions are plotted as histograms and the true values of the test statistics are shown as vertical lines. The test statistics are the multinomial test statistic (M), the biochemical diversity (BCD), the proportion of invariant sites (PI), Tajima's D (D), and the number of singleton sites (S). Posterior predictive p-values are shown for each model and test.

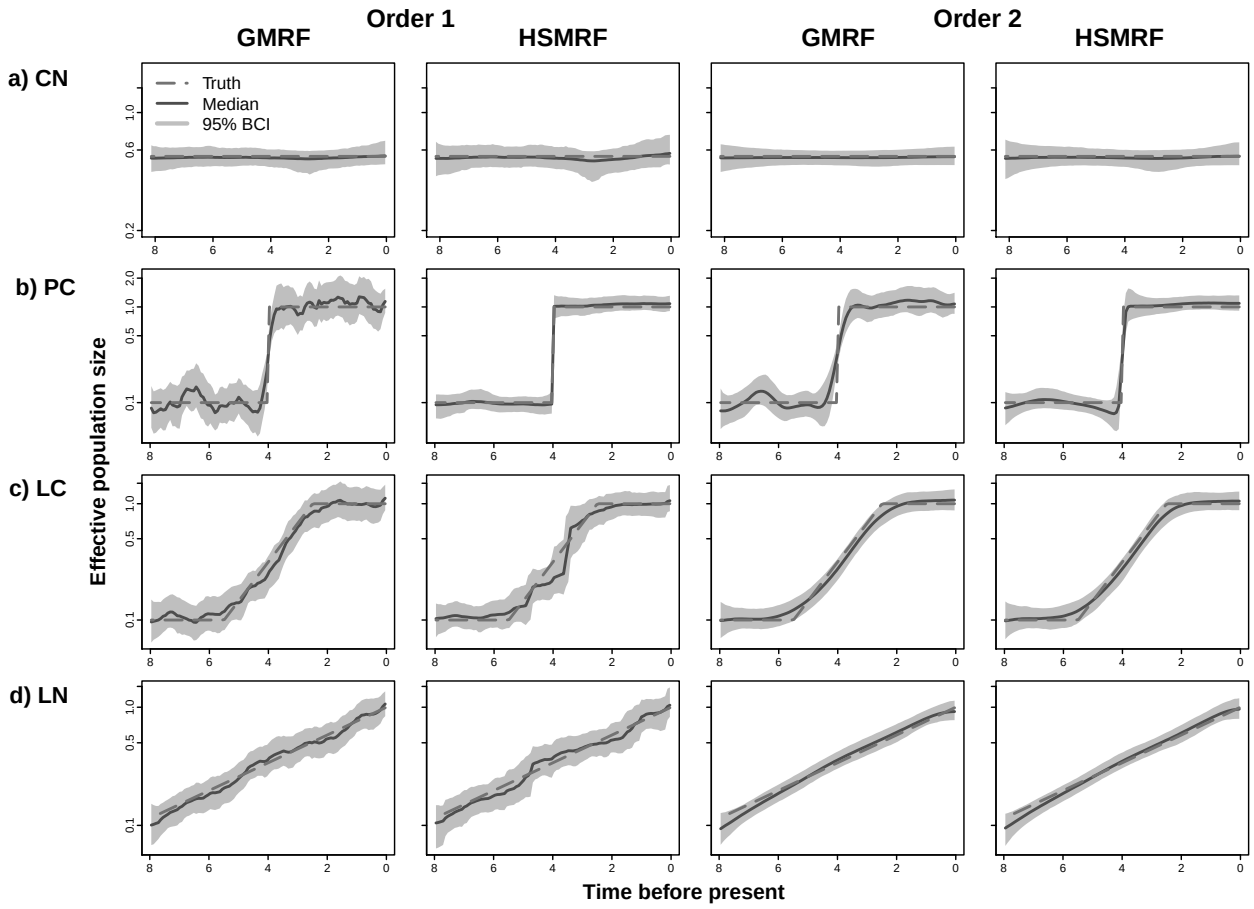

Figure 3. Estimated trajectory examples produced by the first- and second-order Gaussian Markov random field (GMRF) and horseshoe Markov random field (HSMRF) models for four different simulation scenarios demonstrating various rates of change in effective population size. Scenarios are a) Constant (CN), b) Piecewise-Constant (PC), c) Linear-Constant (LC), and d) Linear (LN), where linearity is on the natural log scale. Results for all models within a particular scenario are for the same set of simulated data. Shown are the true effective population size trajectories that generated the data (dashed lines), posterior medians of estimated trajectories (solid lines) and associated 95% Bayesian credible intervals (shaded bands).