Dear Dr. Yang,

We are submitting a revised version of PONE-D-15-41521. We thank you and the reviewers for thoughtful comments and suggestions. We've responded to the comments point by point below. Also, please note that we cite a Goring et al. submission to PLOS ONE that we hope will be accepted soon enough that we can cite it as PLOS ONE article.

Editor comments

1. *PLOS style requirements*: We believe we have satisfied the style requirements. Please let me know of anything we have overlooked.

2. *Copyright and figures*: With regard to figure copyright, all three figures were produced by me using open source software and results from our analysis and do not include any Google images. If you have further concerns about them, please let me know.

3. *Maps and proprietary data*: Similarly, the figures are of course maps, but they show results from our analyses and do not contain proprietary data.

4. *Data archiving and PLOS data policy*: We have placed the output results and the two input datasets with NIS, and this is now detailed in the revised manuscript (lines 97, 100-101, 446). With regard specifically to the township data, these are data collated by one of the co-authors, Dr. Cogbill, and he has agreed to release them publicly as one of the datasets archived with NIS. This response also pertains to Reviewer #1's comment that we did not adhere to the PLOS data policy.

Reviewer #1 comments

1. *Page 6, in "Gaussian process approximation" - could you provide a few sentences' justification for why the Q matrix was defined this way?*

   Response: The definition of the Q matrix follows directly from the technical specification of the statistical representation given in [21]. There is not space to go into a lot of detail and we feel it wouldn't be appropriate given the subject of the manuscript, but we've added a clause to make clear that the definition of Q is a special case of the model representation found in [21] (see lines 233-234).

2. *Pages 6-7 - why does alpha_p have different expectations in the two models (0 in CAR, miu_p in SPDE)?*

   We neglected to note that the CAR model mean is taken to be 0 because the model corresponds to an improper prior on the mean, as is well known in the spatial statistics literature (e.g. Section 3.3.1 of the Banerjee et al. [20] reference). We've added a sentence to clarify this (lines 210-211).

3. *Page 10, Lines 340-345 - could you explicitly write out the formulae for RMSPE and MAE?*

   Yes, this is a good idea given the potential confusion in what is meant by weighting here. Please see the added equations at the bottom of page 10.

4. *Could you speculate on why the two types of comparisons (posterior mean of metrics, versus metrics for the posterior mean distributions) do not agree?*

   This is a good question, but it's hard to say anything definitive, and we would prefer not to add any text on top of what we already say. First, as we mention in the manuscript, the predictions and overall performance for the two models are quite similar, so one should probably not read too much into what are fairly small differences. Second, the posterior mean of the metric accounts for the full predictive distribution of the model, while the posterior mean predictions do not, as we mention in the manuscript. It's not too surprising that there is some disparity on what are somewhat different metrics given the relatively small differences in prediction performance of the two models.

5. *Tables 1-5 show the relative performance of the two models under different scenarios. What might be more relevant to users, however, is their absolute performance. How capable are the models to accurately predict the empirical pattern, especially when data are scare?*

   To consider absolute performance, the MAE, RMSE, coverage and mean/median interval lengths can all be interpreted on an absolute basis for each model, without reference to the other model. MAE indicates the average difference between a prediction and the held-out value, while RMSE is based on the squared deviations between prediction and held-out value. Both of these can be interpreted as errors that lie in the (0,1) interval on which probabilities lie. Coverage can be interpreted relative to the gold standard of 95% for a 95% interval and the lengths of intervals can be interpreted relative to the fact that probabilities are in (0,1). The hold-out experiment (particularly the 95% held-out cells) are intended to assess both absolute and relative performance when data are scarce. We have added a sentence noting that these metrics can be interpreted in an absolute sense (lines 390-392).

Reviewer #2 comments

1. *The manuscript is technically very sound. I have two major suggestions for further improvement. Firstly, the methods are quite technical and many readers will not be able to follow the details. To some extent this is unavoidable, as most data users will be unfamiliar with Bayesian spatial modelling. However, it could help to provide additional plain-language descriptions of how the model works. The authors have done this in some places, and below I've indicated areas where more could be added.*

   We've responded to the specific suggestions below for adding plain-language descriptions (see point 4 below), as well as additional such description in lines 181-184, 186-187, 196-198, 199-200, 3 lines between lines 223 and 224, 251-254. In addition we've added a paragraph near the start of the Statistical Model section (lines 148-157) that gives a high-level plain language description of the Bayesian model.

2. *Secondly, the discussion is very short and provides only a cursory interpretation of the results. I don't expect the authors to describe the map patterns in depth, but the discussion should provide further interpretation of the estimated uncertainty. For example: What was the overall magnitude of the uncertainty? How significant (meaningful) is this? How is uncertainty affected by the choice of spatial model? How did it depend on data density and*

format (gridded vs township level)? How well can we estimate composition in grid cells with missing data? Does the uncertainty vary among species?

We've added several paragraphs to the beginning of the discussion to interpret results with an emphasis on considering the uncertainty in the estimated composition.

3. *L65 This discontinuous Quebec township appears in Fig 1 but not Fig 2 or 3. Most of the language in the paper refers to data across the "northeastern United States", so it is unclear whether the Quebec data are part of the analysis and results or not.*

   We do not make predictions outside our core states, but since the Quebec township is near to our core domain (in particular the northern border of Vermont, which has incomplete data in that area), we included the data as it can help inform composition in that area of Vermont. The one township in northern Delaware is analogous. We've added some language to this effect in the text (lines 66-69).

4. *L156 (spatial GP), L173 (MRF), start of p.6 (CAR), L193 (Matern correlation function), L201 (Lindgren approximation), L223 (multinomial probit). Adding plain-language explanations for these various terms would make the methods more accessible to readers, especially those unfamiliar with Bayesian spatial modelling.*

   We've added clauses to add a plain-language interpretation of these (lines 186-187 [spatial GP], 199-200 [MRF], below line 223 [Matern], 181-184 [multinomial probit]), with the exception of the Lindgren approximation – this is simply a mathematical approximation to a GP, as we already say.

5. *L286 The description given does not sound like cross-validation, unless the fitting/testing process was performed repeatedly on different folds. The current description reads like a basic hold-out validation test (i.e., random data held out only once). Also, it seems that these validation experiments were separate from the main data product, which I presume included all data. Some clarification is required.*

   Yes, good point, that was poorly-worded. This was a train/test split (i.e., evaluation by holding out a fixed subset of the data), not cross-validation. We've reworded as appropriate (lines 321, 323, 328, 403). And in the description of the data product we have added a clause to indicate all the data were used for the main product (lines 433-434).

6. *L390 It would be helpful to see a visual example of these edge effects. Might also be useful for showing fine details for a portion of the maps in Figs. 2-3.*

   The edge effects only occur in the SPDE model, which we do not use for the data product. Given that this is an alternative model, we are reluctant to add additional figures to illustrate this point, but if the editor would like us to, we can do so. We note that we provide the maps to the publisher as EPS/PDF vectorized graphics, so viewers can always zoom in to see fine details in figures 2-3. Also the Shiny application that we reference allows one to zoom in without losing resolution.

7. *L396-399 Repeats info from methods at L362.*

   Good catch. We have removed the language from what was lines 396-399 (now lines 433-434), simply referring back to what was line 362 (now lines 399-402).

8. *Fig 1 Why are WV and MD plotted if they have no data?*

   This was simply a default in our plotting code, with all state boundaries within our rectangular window plotted. We've removed state boundaries for states not in our core domain.

9. *You might consider a panel showing the relative error as well (sd/mean). Currently, the third column of panels reflects that uncertainty increases with the estimated/observed mean.*

   This is an interesting idea and we looked at adding a fourth column to Figure 2. However, we have large areas in which the mean for a given taxon is very close to zero, which blows up the coefficient of variation. The result is that the areas with low mean are the ones that stand out. One possibility is to mask out these areas, but this would seemingly involve some somewhat arbitrary choices so we've chosen to stay with the plot as is.