We would like to thank the editor and the two reviewers for the productive feedback on our manuscript. We have addressed the comments related to our choice of data, clarified the results, added more description to our data and methods, and included more language on future directions of this research. More detailed responses to individual comments are included below.

Editors Comments

The reviewers and I agree that this manuscript is interesting, timely and well-written. I will add a few of my own observations here. First, and least important - normally submitted manuscripts do not hyphenate and do not right justify.

Thank you. We have removed the right justification and hyphenation.

Next, in your figure 1, which forms the basis for comparison of model accuracy over time, I really expected you to choose data WITH a temporal pattern. After all, it looks to me like you have a time series in which species richness simply varies around a more or less stable average value, and the variation can be due to nothing (random), observer bias, and any number of species population trends (and combinations thereof). Thus, you have a situation of years of no trend, and use those data to predict no trend into the future? I would have thought you would choose data with a definite time trend that was already potentially correlated with, say, climate change, and then use hindcasting and forecasting to test whether the models accurately predict a trend into the "future"! To me, it is not surprising that models do not "...successfully turn spatial data into useful temporal predictions about biodiversity at decadal time-scales." But, I wonder if you chose data WITH a trend, you would have gotten other results.

> We chose the North American Breeding Bird Survey data because it is one of the largest scale most long running spatiotemporal datasets available, thus making this kind of quantitative assessment of both spatial and temporally focused forecasting methods possible. We agree that the relative stability of these time-series is important, which is why we dedicated a large discussion paragraph to this point in the initial submission. To further emphasize this point and to highlight the idea that assessing these methods on selectively chosen data exhibiting significant trends is interesting we have added an additional paragraph to the discussion addressing this point directly and recommending it for future research.

Finally, your figure two seems to reverse panels A and B, neither of which is marked in the text.

We have switched the panels and labeled them. Thank you for noticing this error.

I will leave the remainder of the comments to the two reviewers.

Reviewer 1: Eric Ward

Basic reporting

I found the paper very clear and easy to follow, with a good intro/discussion that tied the current work into previous studies. All data shared - in addition to code. More minor comments in General Comments section.

Experimental design

This question and forecasting competition was well designed. I had minor comments to improve clarity with models. The analysis is all documented/replicable on the Github repo. More minor comments in General Comments section.

Validity of the findings

All data is robust, and analyses are sound. Conclusions and recommendations are well stated. The authors have really done a great job with transparency - this is a model of what all papers should strive for. See additional comments in General Comments section

Comments for the author

Review of Harris et al. Forecasting biodiversity in breeding birds using best practices

In general, I thought this paper was interesting and well written. These results add to the increasing literature supporting the idea that ecological forecasting in time/space is difficult. I think the recommendations/best practice bullet points described by the authors are essential for anyone interested in forecasting this kind of data. All of my points below can be considered minor issues, which I think will improve the paper's clarity.

Thank you.

Introduction:

Line 41: I think this is generally more common in terrestrial ecology – in fisheries these SDM models with more of a theoretical background don't seem to be used as much

This is a good point. We have added the word "terrestrial" to this line.

Line 62: I really like the recommendations / clarity in Box 1, and think these will be accessible to most ecologists

Thank you.

Line 71-73: I'm re-reading this after commenting on data, and am now a little confused. The bird responses are annual snapshots during the breeding season? And I thought the

environmental data you used was annual or quarterly, not monthly?
https://www.mbr-pwrc.usgs.gov/bbs/grass/genintro.htm

> Bird data is collected once each year during the breeding season. Climate data is available monthly. We want to use the chunk of the climate data that is mostly likely to predict the bird response. We used climate values from the 12 months leading up to the bird sampling since this is assumed to influence the habitat and other resources the birds are responding to (We have also clarified this in the methods). Other choices are possible, but rarely explored b/c what is most commonly used is long-term averages (e.g., 30 year climate means). Future work could explore shorter time scale environmental data such as monthly NDVI trends (instead of just means), migration indicators for migratory species, or more localized weather around each sampling site that could be incorporated into an observer model.

Methods:

Line 92: For those not familiar with the BBS protocol, maybe describe in a sentence. I could be mis-remembering, but isn't each site made up of 50 stops, and presumably you aggregated richness across these stops?

> This is correct, we've added clarification to the Methods.

Line 105: I assume all these species have been similarly ID'd over the course of the dataset – e.g. birds split into sub-species haven't been included.

> This is correct, see the section "Richness Data" in the Methods.

Line 107: For the environmental variables, do you have any that match up with the temporal resolution of the response data (monthly?). It seems like all the variables are annual / quarterly / etc.

> See comment above on using shorter time scale environmental variables.

Line 124: I'm not as familiar with terrestrial projections, but are the CMIP5 projections relatively unbiased / precise when compared to recent data?

> The accuracy of CMIP5 models varies by the variable considered. Temperature is represented very well over North America. Precipitation is challenging, especially over mountainous terrain, but the CMIP5 ensemble represents precipitation across North America relatively well ("Evaluation of precipitation and temperature simulation performance of the CMIP3 and CMIP5 historical experiments", Koutroulis et al. 2016).

Line 138: Maybe also cite any of the interfaces you used, like RStan:
http://mc-stan.org/users/citations/

> RStan is cited in the last paragraph of the Methods.

Line 140: I'm curious why you didn't consider spatial structure in the random effects here, using GMRF models or similar? One hand these models might be overkill if the

habitat/environmental variables (which are spatially correlated) explain variability in richness, but the spatial correlation may pick up other sources of variability not accounted for by those variables.

> The goal of this model is to control for observer effects, not to provide a focal model of the system (though we do use it for the average model as a convenience). Because observers typically observe the same site year after year it was necessary to include site in this model to avoid classifying variation among sites as differences among observers and incorrectly removing it before fitting other models. Given the use of this model, it is unclear what including spatial structure would accomplish for the broader goals of the paper. Moreover, it seems possible that it could end up correcting out meaningful ecological patterns and therefore confuse the comparison of more complex models to the baselines.

Line 141-143: I think the paper thus far is pretty clear. I was hoping to see a few equations here – it helps me especially when components start to become hierarchical!

> We have added the equations and Stan code as a supplemental file.

Line 163-164: again, I think equations could really help here to describe the baseline models. I think it's also important to use these to clarify key details – for example, by training models on a site-by-site basis, you're estimating separate parameters by site. For example with the uncorrelated noise model, you're estimating unique variances by site right? Maybe say something like this, in addition to including equations.

> We now clarify this, using an equation for each baseline model. In particular, we identify which parameters are estimated on a site-by-site basis and which ones are constant across all sites, as suggested.

Line 180 – Maybe also mention this is in the forecast package – as you do with gbm below?

> We have added this.

Line 183: Because these defaults may change over time, it'd be good to specify things like the ranges of AR / MA order that you searched over, which criteria was used to compare models, etc. And whether or not you included a drift parameter?

> We agree that software can evolve over time and make methods non-repricable. We have added details of the auto.arima model in the supplement.

Table 1 caption: without getting to more methods (below), I don't understand the sentence 'Environmental models were trained using all sites together, without information regarding which transects occurred at which site or during which year'.

> We have clarified the Table 1 caption to be more clear about model covariates.

Table 1: Just my personal preference, but I'd include the check marks and replace the 'Xs' with blank cells

> We have made this change

Line 194-221: It seems for each approach you only considered the full models with all environmental covariates? I'm curious whether any of the covariates appeared to not improve

explanatory power, and whether you considered models that dropped these uninformative predictors?

> We considered dropping uninformative predictors, but decided not to because of the complexity it would add to our analyses. In particular, it wasn't clear how to assess variable importance in a consistent way across three different model types and 500 Monte Carlo samples, and we did not want to bias the results by choosing a variable selection method that worked better for one method than another.

Line 216: Cite randomForest / caret/ party R package – or whatever else you used?

> The randomForest package was cited in the final methods paragraph. We've cited it where it is first mentioned as well.

Line 242: Again, I think including the equation here would be useful

> We have added an in-line equation and clarified the wording

Line 245: In our 2014 forecasting paper, we used MASE over RMSE because of rationale in Hyndmand and Koehler (2006). Using RMSE here is probably ok because you probably don't have zeros/huge values/etc

> We agree.

Line 249: Can you elaborate on what you did with the deviance? By itself mean deviance is just a measure of average fit (I think most of the models are giving you this). Var(deviance) should incorporate precision – but I've only seen this extracted from Bayesian models

> We were unclear here. We had only intended to convey that mean deviance was lowest for predictions that were made with appropriate uncertainty. We did not use Var(deviance). The new sentence should be clearer on this point.

Line 266: I think you guys have done a great job putting together this public repository + made the analysis totally transparent. My one hang up with the repo was that it'd be awesome to include a short vignette with example data from this project that is more easily accessible to readers without having to install Jupyter/R packages/etc. For example – I was travelling when I started this review and couldn't install one of the R packages on my government computer without making an appointment with our IT staff (so it took ~ a day). This isn't your fault of course, but much harder for me to dig through the repo.

> Thanks! We tried to handle this using binder (https://mybinder.org/) to allow anyone to quickly engage with the full environment, but unfortunately the R support isn't quite ready yet. Fortunately it is under active development and as soon as it is available we plan to set it up and link to it from the repository to support this use case.

Line 269: See my above comments about spatially correlated random effects / GMRF models. One way to confirm these other approaches are overkill would be to make some maps of the estimated random effects, and fit some simple models to estimate variogram parameters – e.g. even gls() in R can do this

> See our comments about spatial autocorrelation above

Line 276: Is one of the potential reasons that richness is stationary that presence/absence on the BBS varies less than other measures – maybe abundance, or other measures of diversity are more variable?

> We agree that this is likely. We make this point on lines 423-429 of the revised manuscript.

Line 294: Is the narrow predictive distribution for the stacked SDM also why this exhibits the strongest trend in decreasing coverge as a function of forecast length (Fig 5)?

> Yes, exactly. More and more observations fall outside the narrow prediction intervals.

Line 329 – maybe reword, you use forecast 3x in this sentence

> Thanks for pointing this out. We have reduced this to 1x.

Line 414: Agreed – doing this same kind of forecasting competition with compositional data, or other species diversity measures (Simpson's) would be interesting

> We are currently working on a few projects along these lines.

Line 461: Another thread with conservation implications is that back in the 1990s, it was pretty common for single species PVA models to be forecasted 100-200 years. This seems less common, in part because some predictions may have not performed well, and others are on a time horizon that can't be evaluated. In Steve Beissinger's PVA book he cites Goldwasser (2000) "Variability and measurement error in extinction risk analysis..." for this point (p 133).

> We agree that this is an apt comparison. However, we are unfamiliar with this literature and UF does not have access to the paper you cited. The existing citations should prove sufficient to justify these points.

Reviewer 2

Basic reporting

see general comments

Experimental design

see general comments

Validity of the findings

see general comments

Comments for the author

In this paper, the authors use the incredibly rich data from the Breeding Bird Survey to examine how well various models perform at predicting bird richness at sites throughout North America. They train models on data from 1982-2003 and attempt to predict data from 2004-2013. They examine the performance of six different model types, two of which are null or baseline models.

I find it a bit surprising that the authors think that any of these richness models could predict a high level of accuracy for theses ten years. I think of richness estimations as a very coarse technique for looking at broad shifts in richness over relatively long time periods rather than these 5-species (~5-7%) shifts in richness on annual bases indicated in Figure 2. The authors don't provide a lot of their raw data on sites other than as examples, but my impression is that they are pursuing an unrealistic level of accuracy and detail. That being said, given what the paper does, it's okay. The models seem to be performed well. The authors give all of the keywords to indicate that they know what they are doing.

> We agree that the relative shortness of the forecasting window and the lack of major directional shifts in richness (on average) are important to consider in interpretting our results. This is why we included long paragraphs on both topics in the initial submission. To further emphasize this point and to highlight the importance of longer time scales analyses and assessments on data showing larger changes in richness we have added an additional paragraph to the discussion addressing these points directly and recommending them for future research.

Here are some comments:

Birds are not the best system for doing any sort of distribution analyses. They are incredibly vagile, hard to detect, and the distributions of breeding birds are affected by many components besides environmental variables (particularly interspecific interactions). No assessment of detectability is included here other than accounting for observer identity. However, I do realize that BBS data are detailed and compelling data to use given what a large amount of detailed time-series it provides.

> As the reviewer notes the North American Breeding Bird Survey data is one of the largest scale most long running spatiotemporal datasets available. We chose it because these features make this kind of quantitative assessment of both spatial and temporally focused forecasting methods possible. While there are downsides to any dataset we do not agree that birds uniquely problematic for this type of analysis. In addition to large amount of available data the vagility of birds potentially allows them to respond more quickly to changes in environmental conditions due to their low dispersal limitation. Imperfect detection and non-environmental drivers are common challenges in most animal species.

Abstract: The authors forecast the data and then test it using known data. Hindcast only refers to back-in-time predictions.

> The definition of hindcasting varies by field. We use it here as it's used in meteorology, where any prediction of observations already available is generally a hindcast, and forecasts are only for predictions which are actually in the future.

See Jolliffe & Stephenson 2003 (in manuscript) section 1.4.2 for a longer discussion. We've clarified this in the text.

25-29: rather than relay the type of information that the paper will provide, the authors should summarize their results and findings

We have added a summary of our results and findings.

57-58: this is not the definition of hindcasting. This is the definition of testing. Also, the authors exclude the many deeper-time hindcasting studies that have been performed on these types of models

See comment above on hindcasting. We agree that studies evaluating models using fossil records are an important aspect of this field of research and have included several references for them.

Methods:

the authors perform their analyses on data that are collected from an incredibly dynamic time period. I wouldn't expect species to have settled into a richness pattern given the incredibly disrupted nature of the last 32 years. Some strategies to try to handle these issues might be to train and predict at different temporal scales and to both forecast and hindcast the predictions. (e.g. train on 2004-2013 and try to predict 1982-1992 data + 1993-2003 data). If the authors were to make splits based on expected periods of dynamism, they could better tell whether the drivers that are included in the models are accurate predictors.

We agree that understanding the conditions under which forecasts are likely to perform well is an important step for ecological forecasting. We have added a sentence to the discussion making this point. That said, most ecological forecasts are intended to be applied to the future, which at least at the moment we would expect to remain "incredibly dynamic". As such the dynamics of the last decade are likely a reasonable representation of the kinds of dynamics we are trying to forecast.

124-130: I didn't see where future projections were performed/used in this paper. Why is this here?

Thank you for pointing out that we were unclear here. The methods and purpose for the future predictions are now explained in more detail on lines 136-139 and 502-505 of the revised manuscript.

Fig. 2: Panels not labeled. No key for colors. Somewhat confusing.

Thank you for pointing this out. As the Editor noted, we had inadvertently reversed the two panels. We have fixed this, and also labeled them as you suggested.

Fig. 6 caption: "compoonents"

Thank you for pointing out this error. We have fixed it

466: only other mention of future forecasts... not sure why this is included in this paper.

Thank you again for pointing out that we were unclear here. The purpose of the future forecasts is now explained on lines 136-139 and 502-505.