Dear Dr. Foote:

Please find below responses as requested for our resubmission ECE-2018-01-00114.

Best,

Claire M. Curry

Associate Editor  
Both reviews raise the same fundamental concern about your work - that you are applying STEM models to a situation they were not designed for and would not seem to be necessary for, therefore making your finding that they do not add anything to 'standard' species distribution models unsurprising. While I am sure the other concerns raised by the reviewers could be easily enough addressed, this does seem to be a somewhat fundamental issue. In order to consider a revised version of your manuscript I would really need to see some very convincing arguments as to why STEM models would in principle be useful in the situation presented by your data; otherwise it would seem that your conclusions lack real use to other ecologists.  
>We thank you and the reviewers for your careful and constructive comments. We believe we can address your concerns and describe in more detail below why STEM models should be useful in the cases we studied. We have reorganized and retitled the manuscript to better emphasize this point, as described in detail below.

Reviewer: 1  
This manuscript aims to evaluate the trade-off between computational efficiency and accuracy of two different species distribution modelling approaches, namely Spatio-temporal Exploratory Models (STEM), which can model spatio-temporally varying responses of species to environmental variables, compared to much simpler models that fit a single model over the extent of the study area. The authors test this for 11 grassland bird species in a single state (Oklahoma) and find that the STEM models do not improve the accuracy of a standard SDM (RandomForest), but are computationally much more demanding. The manuscript is fairly clearly written, but the description of the methods is too brief and insufficient to understand or replicate the modelling approach.  
  
With the consent of the editor we discussed this manuscript in a journal club at the University of Cambridge. The following review was compiled by me (Steffen Oppel) on behalf of the five attendants of the discussion. The manuscript was anonymised before being shared with other participants of the discussion, and all participants agreed to standard reviewing code of practice.  
>We appreciate the extensive and helpful comments.

The key concern about this manuscript is that the STEM approach was developed to overcome regionally varying species responses to environmental variables over large (continental) scales. This approach will therefore perform best in situations where it is realistic to expect that non-stationarity exists in the data, i.e. species respond to the same environmental variable differently in different parts of the study region. Attempting to use these models at a relatively small spatial scale (a single state, in a fixed time period, with only open-country species) seems curious and highly unlikely to require a complicated STEM approach. From that perspective it is unsurprising that the STEM models do not perform better than the basic SDM. In addition, the choice of RandomForest, which is a very flexible algorithm that may account for some non-stationarity (more than for example a GLM), further reduces the chance that the STEM models may out-perform the basic SDM. While the comparison in this manuscript is not scientifically flawed, it is unlikely to yield any great insights and the authors should emphasise in the Intro and Discussion that this comparison should ideally be performed at a larger scale than a single state (or with a group of species that can realistically be expected to exhibit non-stationarity at that scale). Expanding the spatial, temporal, or species composition scale would therefore increase the value of the manuscript.  
>We now further explain our reasoning for the use of this study system in lines 57-71 and 99-103, by detailing Oklahoma’s diversity in species richness, grassland habitat types, and how climate can affect related physiological processes. We also note that similar studies have used similarly scaled and diverse regions, and that grassland birds in smaller areas have shown turnover spatially and temporally in variable importance (Ethier et al. 2017). Finally, in the discussion we emphasize that these models are still useful and scale dependent by making that point the final sentence of the paper. We hope that these changes address the concern towards the relevance of this approach.  
  
Apart from the poorly chosen data set, the manuscript also suffers from a lack of methodological detail that is essential to understand the modelling process, while at the same time there are too many different questions/analyses/examinations that are only tangential to the key goal of comparing the performance of STEM and rangewide models. We were therefore not able to judge whether the model evaluation was technically valid, and much more detail will need to be provided to allow a proper assessment of the validity of the Results.  
>We hope that our clarifications described below and in response to reviewer 2 will address your concerns.

The key questions with respect to the modelling were:  
  
What is the response variable in models? Presumably detection/non-detection as a classification tree is used. How is this aggregated over temporal replicates or are all temporal replicate point/transect counts used as input in the model?  
>We have clarified our tree approach in line 194 as a random forest regression tree. We now specify the temporal window of April-July in line 146 and how replicates are used in lines 132-136 by noting that we did presence/absence based on date/time combinations for a given location.

The STEM approach derives its strength from repeatedly fitting base models to random spatial subsets. There was no description how many iterations were performed and how spatial subsets were jittered/shifted among iterations from the random starting point. The description in L. 199-206 is insufficient to understand how the STEM models were fitted.  
>We have added a summary in lines 232-234 (“Building a SEEM consists of creating random support sets, generating trees and predictions for each support set, and then combining each support set model predictions into the final overall prediction. “) to clarify what we refer to later in the paragraph. We used the randomization process in the R package sp (line 221) and the number of randomized subsets (i.e. the subset number of iterations) is given in lines 223-224, resulting in pixel coverages as given in lines 225-226 (means were 6.9, 6.3, and 6.6 for small, med, and large).

What environmental and response variables are used for evaluation? Is any detection record in a grid cell assumed as indicating presence? Or are detection and non-detection records averaged to a %-value? Are the real time and effort variables of the validation data used for prediction or are predictions based on averages (as described in L. 180-184)? For the 'different years' evaluation, were land cover variables also used from different years, given that agricultural and grassland areas can change quite rapidly and could therefore account for poor transferability?

>The evaluations were used with the same variables as in the original models. The presence/absences are for a given pixel and the grid cells are to ensure an evenness in spatial sampling, which we now clarify in lines 251-256.  
  
How are temporal changes in bird distribution accounted for? The STEM models allow for temporal variation (e.g. such as migration), but this component is omitted from this manuscript. How can the sighting of migrating individuals (which may lead to low predictive 'accuracy' if migrants are spotted in unsuitable breeding habitat) be excluded, or how is it accounted for in model evaluation?  
>We now specify the timespan of the models in lines 136-137 to specify that we surveyed during the breeding season only (April-July) and that we only used eBird data from that same time period.

The use of LMMs to compare model performance was unnecessary and distracting. Presenting the differences in AUC and/or RMSE would be sufficient to demonstrate that there are no differences. The only 'statistically significant' difference/relationship (Fig. 5d) is meaningless and the large amount of beta and p-values distract from the main message. Tables 1 and 2 should contain numbers and not symbols.  
>We have removed these tables as we eliminated LMMs per your critiques and replaced them with a visual comparison of medians via notched box plots (per Chambers et al. 1983, in Figs. 7 and 8). This does change our results quantitatively in which species are best with each model, likely because medians are now being compared instead of means, but we feel that the main result (that data resolution changes which models are useful) is qualitatively similar.

Besides the main question of the paper (model performance evaluation) the authors also attempted to project models into the future to evaluate habitat loss under climate change for the focal grassland bird species. We found this side-story slightly deficient and distracting, and would recommend to remove it from the paper altogether. The key reasons why this section is unlikely to be very informative are that (1) such a climate change projection really only makes sense if the species show very strong climatic niches at the spatial scale of the model (unlikely at such small state scale); and (2) over the next 50 years it is highly likely that most (if not all) of the current grassland will be converted or altered, either due to agriculture, re-forestation, urbanisation or water level changes, and these habitat changes will likely have a much stronger impact than climatic changes on your focal bird community.  
>We have removed all references to future climate projections, as your critiques have persuaded us this is a weak point.  
  
Specific comments:  
L. 18-19: This statement is a bit misleading as it implies that you optimised the level of computational effort to increase accuracy - but instead you simply quantified the level of effort for two types of modelling approach and report their respective accuracy  
>We have reworded lines 18-20 to reflect your accurate reading of it.

L. 28-31: These are not really 'policy' implications, but rather suggestions for future research. Very few policy makers can be bothered with the technical details of species distribution models!  
> We have added a more specific caution in lines 25-31 that more research is necessary while not specifying that policy makers would be examining these models.

L. 47: Describe what you mean by 'base model' and how STEM can be applied with/to these  
>Wording has been expanded in lines 45-48 to clarify these points (“This averaging of overlapping smaller models (the model type used here is referred to as the base model”)

L. 55-80/Introduction: The beginning of the Introduction makes a valid case why you might want to test STEM models at a smaller spatial scale and assess the trade-off in computational efficiency vs. accuracy, but this section does not provide a compelling argument why Oklahoma grasslands would be a suitable test scenario. Presumably these models will only be useful in areas where you can expect bird responses to differ spatially to environmental variables - which may not be the case if you are looking at birds that prefer open habitats.

>We have reframed the introduction in lines 57-82 to emphasize why Oklahoma grasslands should show the requisite variation. First, we have organized it so that the variation comes before explanations of why grasslands are important. We now cite Ethier et al. 2017 and Fink et al. 2015 about how birds that inhabit single ecotypes and in smaller areas (Ethier et al.’s study is in an area about half the size of Oklahoma for a grassland specialist, the Bobolink) show spatial and temporal variation in variable importance. In other words, we believe we have made an improved case that these open habitats are not homogenous and that other ensemble studies have used similarly small regions with useful results.

L. 97-99: Please use SI units in an international journal  
>done

L. 105: How were temporal replicates of the same point/transect used in the analysis?  
>We now describe this in lines 135-136: “A zero (absence) or 1 (presence) was assigned for each combination of date and time and species”.

L. 108: Distance to what? Your analysis makes no attempt to correct for imperfect detection by either using the repeated counts or the distance information, so there is no need to describe that here. Somewhere in the manuscript (Methods or Discussion) you should acknowledge that bird sightings are a stochastic process and that you do not account for the imperfect detection process - but given that this error is similar in your two model types the comparison is not affected by this issue.  
>We have added this qualifier in lines 132-134: “We only used sightings within 500 m of the observer to preserve identification accuracy and recognize that detection is imperfect; however, all models compared use similar data and as such it should not impact our comparison of models.”

L. 127-128: from how many discrete locations? It would be useful to present a map with the location of raw data, which can be combined with a conceptual map of the STEM grids to visualise the approach.  
>We have added Fig 1 which shows the locations used and Fig 2 to show the support set grids of all sizes.

L. 134: This is true, but only if the predictor variables are also from the different time period. With many grasslands, you might expect habitat conversion due to agriculture at relatively short intervals, so land that may have been suitable in one year may not necessarily be suitable in a different year. Were the predictors for the external validation also used from 2011-2012, or were the ones from 2013/2014 used? If you use the latter, then you assume that no habitat has changed, which may explain the poor performance of the models (Fig. 5b - AUC is ~0.5!!).

>We eliminated the different year datasets as you point out that the NLCD datasets have not changed. We hope that this removal also lets the overall paper have a clearer focus.  
L. 141: Insert the reference.  
>fixed, to Table S1.

L. 145: If you record birds within 500 m (L. 109) then it is unlikely that habitat within a fraction of this distance (150 m) would be informative. Remove the 150m pixel to eliminate one confusing dimension of your models.  
> We added a justification to line 132-134 as to why we did this filter and why it should still be used in the models in lines 172-174 (“Although the 15 x 15 pixel unit is smaller than our 500 m cutoff, most sightings are from even larger areas with the maximum length being under 4.3 km, an area comparable to Fink et al. 2010.”). We would also like note that we unfortunately lack funds to run the models again without that variable.

L. 151-158; 211-215; 322-328: We would recommend to remove these sections as they are not well incorporated into the key question of the manuscript and are too poorly developed to provide credible or informative results.  
>We have removed all reference to climate change models as we agree with you.

L. 163: Please state what your response variable was - a classification tree suggests you modelled detection/non-detection data, but when was a detection confirmed? If you had 4 replicate visits to a given point and only one recorded the focal species as present, was that scored as a 1 or a 0?  
>Line 125 now explains that each presence or absence was counted for a given date/time/observer/location.

L. 168: Specify what 'they' are - presumably RF, but it currently refers to the tuning parameters.  
>You are correct that it referred to RF and we have altered wording to clarify (now line 185).

L. 173: The majority vote of all trees is used for the ensemble prediction, not the average. In classification, the response is categorical, so there can be no 'averaging'.  
>As we now clarify in line 177 we used regression.

L. 173: which 'later' models? Those described later in the manuscript?  
>We deleted this wording as we agree it was confusing.

L. 179: This statement is false as written - we assume that you mean you cannot use the 'predict' function in R unless your new data frame has the same predictor variables as the model, so if date and time were in your model then your predictions must specify a date and time. But you do not need to use all variables in the model, only in a prediction based on that model. For evaluation, however, it would be unfortunate to use averages, because surely the time and effort of the validation data should be used for prediction?  
>You are quite correct. We have re-run the evaluations with the actual data and time of the validation data. We have updated the wording to match, now in lines 211, 252-254.

L. 197: 'bootstrap sampling'

>I am uncertain what is meant here as we did not discuss bootstrapping in this section; we will be happy to work on this with further clarification.  
  
L. 206: delete 'trees failed in these conditions' - any model will fail if the response is uniform.  
>Reworded in line 242-243 (“models cannot run with uniform values”).

L. 207-208: Given that your output is a categorical value (classification tree - yes or no), how were these values 'averaged'? Did you score a predicted presence as '1' and a predicted absence as '0' and then calculate averages? This could be formulated as the number of cells with a presence prediction over the number of cells that were aggregated. Are the 'support set rasters' the iterations of different STEM pixels? This is poorly described and difficult to understand.  
> We ran the model as a regression tree (now clarified in line 194), and as such had numerical values to average. We now illustrate the support set extents in Fig. 2.

L. 218: Was the 10 x 10 km grid also the spatial resolution of your prediction? If so, how did you treat the 10 observations per grid cell when validating your model output (if model output is yes/no, and observations are a series of numbers)?  
>We have clarified wording (now line 254-256 (“These sampling grid cells are larger than either predictor size and are used to ensure that we do not weight the accuracy of the models towards regions with more reports or surveys.”) that this was not the spatial resolution of the prediction. We followed Fink et al. (2010) in this technique (lines 250-252).

L. 219-220: Was this just for model evaluation or did you also repeat the sampling 50 times to build the STEM models?  
>We have clarified in line 256 that this is for the evaluation only.

L. 231: Given that the model outputs are not the result of some random sampling process, there is no need to test for a 'statistical significance' of predictive accuracy. It would be perfectly ok to use common sense - if the AUC only differs by 0.01 then it doesn't matter whether that is statistically significantly different or not, because the effect is so small that nobody cares. We would recommend to remove all the confusing beta and p-values from the Results section and simply provide the actual AUC/RMSE values (or differences) in the text or in Tables 1 and 2.  
>We have removed the LMM analysis and used the more visual approach of comparing notched box plots of the actual resampled AUC and RMSE values (Figs 7 and 8), allowing both a comparison of significant difference and the reader to judge for oneself the biological significance of the change in AUC/RMSE values.

L. 252-253: If the increase was <0.001 then this result is negligible, regardless of what an LMM may say. Looking at Fig. 5d confirms that for all practical purposes there is no relationship between runtime and RMSE.>We have removed the direct runtime/accuracy analysis and left only comparisons of relative ratios of runtime.

L. 288-289: This sentence was confusing - what performance varied? The key conclusion from your study is that STEM models are neither better nor worse than a state-wide RandomForest model at a spatial scale where you cannot expect STEM models to improve accuracy, especially when the temporal dimension is omitted from the models.  
>As described above we have reframed the paper and added more justification for use of this method in our system in the introduction. We have also revised the sentence itself (lines 286-288): “their performance differed by species and predictor resolution even in a state with variable climate and diverse ecoregions”.

L. 335: insert 'to' elucidate  
>done

Please contact me if you have any questions about this review.  
Steffen Oppel  
[steffen.oppel@rspb.org.uk](mailto:steffen.oppel@rspb.org.uk)  
  
Reviewer: 2  
The primary goal of this study was to test the predictive accuracy of a spatially explicit ensemble model based on the SpatioTemoral Exploratory Model (STEM) of Fink et al. (2010) for 11 grassland bird species during the breeding season in the state of Oklahoma. This is a novel test of STEM because of the relatively small spatial and temporal extent of the analysis. However, it is not an appropriate application of the STEM methodology.

>We respectfully disagree that it is not an appropriate application despite it not being case envisioned in the original application. We have further explained our reasoning and clarify this in the introduction in lines 57-82 and 99-103. We feel that as our methodology is sound and we have provided justification, this example will benefit other researchers who may have also considered applying the method to their systems which may not be at the same scale, temporal or spatial, as the original cases (lines 109-114).

STEM is an ensemble regression model designed to adapt to non-stationary predictor-response relationships, i.e. relationships between species occurrence (the response) and environmental characteristics like land cover and climate (predictors) that vary over the spatial and/or temporal extent of the study area. The paper does not make the case that non-stationarity is a modeling concern for this application. Solid a priori ecological reasons why one would expect spatially varying species occurrence-environment relationships are lacking. Empirical tests demonstrating substantial spatial variation in species occurrence-environment relationships (at the scale of the available predictors) are also lacking.  
>We have expanded our case that non-stationarity is a concern in this system in lines 57-82 and 99-103. We also note here that none of the STEM papers we have cited test for non-stationarity before presenting their models and that we are following this precedent. As this is how the method has been presented, we feel we are replicating the principles of it in a new system in a way that will be useful to others who have considered applying these useful models (lines 109-114).

The empirical tests (described below) provide a simple way and computationally efficient way to determine if the additional computational effort of STEM is necessary to deal with non-stationarity.  
  
When is spatial &/or temporal non-stationarity in predictor-response relationships a concern?  
In general, non-stationarity in predictor-response relationships (e.g. environment-occurrence or environment–abundance relationships) become a concern when the:  
1) The ratio of the size of the study extent to the size of the predictor resolution or grain is relatively large, and  
2) When the underlying processes driving species occurrence or abundance are expected to vary at spatial &/or temporal scales between the size of the study extent and the resolution of the predictors.  
The study fails to motivate the need for a regression modeling approach that deals with non-stationarity. Regional variation in the predictors (e.g. land cover and climate) by itself (as suggested by the topic sentence of the paragraph on line 71) is not enough to warrant a concern over non-stationarity.  
>We have expanded our argument from line 71 into lines 70-74 citing other studies where variation in predictor variables is known from small, regional studies of birds. We also argue that “relatively large” is too subjective to determine a priori from intuition. A county scale would be “obviously” too small relative to even a 30 m scale for birds, but why is a large, ecologically diverse state too small? We also expand on the climatic variation present in Oklahoma. In addition to the study species occurring across the state in several cases, the climatic variables included in the study do vary strongly from east to west. We also note that other studies we cite using this method have found similar justifications for use of the method (variation in local climate and ecology).

In addition to regional variation in the available predictors, we need to know - Would one expect any of these 11 grassland species to use their local environments in different ways in different parts of the study area or different times during the study period? Is there an ecological process that you expect to vary across the region? Or through time? Do you have predictor and response data with enough spatial and temporal resolution to capture these sources of variation?  
>We have expanded our argument on this point as described in the previous response. Additionally, we have removed the different year section as suggested by reviewer 1, such that the predictor and response data are matched temporally to address these variations and have clarified that we only used breeding season sightings in lines 146-147.

At a state or regional level, seasonal variation in land cover associations can vary substantially with predictors at similar scales to those used in this study. Zuckerberg et al. (2016) studied the seasonal variation species’ occurrence–land cover relationships from spring to autumn for resident and migratory species within individual Bird Conservation Regions (these regions are not too dissimilar in size compared to the state of Oklahoma). Temporal variation in species occurrence – environment relationships was expected among migrant species because of the varying demands on these birds as they advanced through different stages of the annual cycle – spring arrival, establishing territories, breeding, fledging, autumn migration.

>We have clarified in lines 146-147 that we used only breeding season sightings and in lines 70-74 why we expect changes in variable importance across the state.  
  
Based on my personal experience fitting species distribution models (using eBird observations and predictors at similar scales to this paper), the relatively small spatial and temporal extent of the study (breeding season in the state of Oklahoma) combined with the relatively coarse grain of the predictors makes it very unlikely that there would be any need to account for spatial or temporal non-stationarity in the species occurrence – land cover & climate relationships. However, my personal experience is not something that can be generalized.  
>The grain of the predictors is similar to that used in Fink et al and all studies cited, including in relatively small regions. We argue that our prediction that it should matter to these species is borne out by the factor that two species do in fact, show improvements with the spatially explicit ensemble models. We have reworded the title, results (275-276 in particularly), and discussion to emphasize that we, in fact, show that they do improve prediction for some species (288-289), but that it is not consistent among predictor resolutions or species (297-299).

How do you test for non-stationarity?  
If the analyst suspects non-stationarity is an issue, it can be tested by simply running models in different regions and then testing to see if the predictor-response relationships vary substantially between the regional models. Did the authors test for non-stationarity?  
> We appreciate the detailed explanations on how we might test for non-stationarity in the future. In this case, we did not test for non-stationarity but to our reading neither did Fink et al or other STEM papers before conducting the analyses.

When is STEM needed to deal with non-stationarity?  
Even if there is non-stationarity, STEM may not be the best model for all jobs and inferences. STEM is especially useful when trying to characterize changes comprehensively over space and time (e.g. Fink et al. 2017). However, if you know the region or season of interest, it can be more efficient to simply fit local models in the regions of interest (and compare them if needed).

>We agree with the reviewer’s point and say this is actually why our study is needed. We aim to provide further clarification in the literature on when such a model might be needed. It was not evident to us from the current literature that it should not apply to our use case and we further expand on this in lines 109-114.   
  
Implementing STEM  
An important part of STEM is the creation of a randomized ensemble of overlapping support sets. Randomization of the spatiotemporal location of individual support sets serves several important purposes. Randomization allows the mixture to adapt to unknown spatial &/or temporal patterns of variation in predictor-response relationships, it helps induce variation among base-models (which is a useful strategy for generating good ensemble models (Kuncheva, L., and Whitaker, C. 2003)), and it is useful for generating uncertainty estimates (essentially, spatiotemoral-block resampling).  
>We further emphasize in line 232 and 234-236 that the support set centers are randomized and now cite this.

Each location should be covered or supported by a sufficient number of base models for that the averaging to effectively control for inter-model variation. The number of base models covering a location is the “ensemble support”. In my experience, ensemble support should be at least 50, however this may vary depending on the specific study.

>We appreciate this concern and now provide mean and quartiles for pixel coverage in the methods for each support set type in lines 225-226. We are unclear where the threshold of 50 is determined or how one would otherwise determine the appropriate support number. We argue that our evaluation procedure should determine the effectiveness of the ensembling coverage. The models we have presented generally perform well above AUC = 0.5 and often above other thresholds chosen in other papers such as Elith (2000) of 0.75 or 0.85.

It is not clear how the ensemble of support sets were created and randomized. Lines 199-211 provide too few details for this important part of the model. Moreover, the paper did not communicate what the ensemble support was? Was there a maximum number of base models generated to cover each location? What was the minimum ensemble support used to generate an estimated occurrence? These have been important details when I have implemented STEM (see Johnston et al. 2015 and Fink et al. 2017 including supplemental materials)  
>We appreciate these suggestions on what details to include to improve the clarity of our methods. Based on these comments we now provide the support set 1st quartile, mean, and 3rd quartile coverage of a given pixel in lines 238-241 and Fig. 1 shows support sets.   
  
Based on the distribution estimates shown in the figures, it looks like ensemble support was too low. Figures 1-4 and Supplemental Figures S1-S7 all show some degree of blocky-ness (most apparent in the small and medium support set sizes). This blocky-ness indicates insufficient ensemble support to average out spurious inter-model variation (i.e. bias).  
>Fig. 8D in Fink et al. 2010 also shows blockiness. We welcome suggestions on how to quantify an allowable level of blockiness.

STEM and Uncertainty Estimates:  
While this paper focused on predictive accuracy, the STEM approach, combined with data resampling during ensemble randomization can be used to generate uncertainty estimates for predictions and inferences (see Zuckerberg et al. 2016 and Fink et al. 2017).

>Thank you for this point. We will use it in future but at the moment do not feel it detracts from our current point that model usefulness varies due to factors as yet unknown but related to variability. Other effective presentations of ensemble models did not generate uncertainty estimates for predictions (Johnston et al. 2015) and we do present uncertainty around evaluation with bootstrapping (new Figs. 7 and 8).

References:  
>Removed for char. limit.

>We appreciate the additional references, particularly the first paper, as it was not available when we originally submitted. We have added them to lines 50, 60-61, 72, and 236.