23-Feb-2018  
  
Dear Dr. Curry:  
  
I write to you regarding manuscript # ECE-2018-01-00114 entitled "SPATIALLY EXPLICIT ENSEMBLE MODELS DO NOT ALWAYS IMPROVE SPECIES DISTRIBUTION MODEL ACCURACY" which you submitted to Ecology and Evolution.  
  
The manuscript, in its current form, is not suitable for Ecology and Evolution.  However, we invite you to consider substantially revising your manuscript and re-submit it as a new paper.  Given the scope of changes we are suggesting, this new manuscript will be treated as a new submission. You have 120 days to make the changes.  If you resubmit, please quote the original reference number and title in your covering letter (please note that if you have received previous reviews or editor comments for this manuscript, please include a detailed response in your resubmission cover letter).  
  
Thank you for considering Ecology and Evolution for the publication of your research.  
  
Sincerely,  
Dr. Chris Foote  
Editor in Chief, Ecology and Evolution  
[cfoote@wiley.com](mailto:cfoote@wiley.com)  
  
Associate Editor Comments to Author:  
  
Associate Editor  
Comments to the Author:  
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 the manuscript to better emphasize this point, as described below.

Reviewer Comments to Author:  
  
Reviewer: 1  
  
Comments to the Author  
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.  
  
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.  
  
  
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.  
  
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?**  
  
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.  
  
**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?**  
  
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?**  
  
**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 difference**s. 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.**  
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 this statement 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!  
> Added some wording to further elucidate our ideas here.

L. 47: Describe what you mean by 'base model' and how STEM can be applied with/to these  
  
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 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, open habitats are not homogenous.

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?  
>penciled in  
*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.*  
>penciled in

**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.**>added to to-do list

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 will have eliminated the different year datasets as you point out that the NLCD datasets have not changed.**   
  
L. 141: Insert the reference.  
>added to list

**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.**> need an argument against this. I think we can say that this is comparable to Fink et al. paper, as they did not filter by distance, and also maybe just note we don’t have the resources to re-run. Most sightings are closer. It will also not change comparison between models.

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.

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?  
>penciled in

L. 168: Specify what 'they' are - presumably RF, but it currently refers to the tuning parameters.  
>penciled in

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'.  
> penciled in also added to to-do list

L. 173: which 'later' models? Those described later in the manuscript?  
>This is unclear even to me, delete (penciled in)

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**?  
>penciled in and added to list. That makes a lot of sense. Will re-run, think it is possible with smaller computer.

L. 197: 'bootstrap sampling'

>I am uncertain what is meant here as we did not discuss bootstrapping in this section.  
  
L. 206: delete 'trees failed in these conditions' - any model will fail if the response is uniform.  
>penciled in

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.  
>make a figure to show this better. Penciled in wording suggestion.

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)?  
>this was not the spatial resolution of the prediction. I believe I got it from Fink et al. but not sure. Check.

*L. 219-220: Was this just for model evaluation or did you also repeat the sampling 50 times to build the STEM models?*>penciled in to change wording to specify for 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.  
  
**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.**>penciled in that there is no biological effect or any meaningful different. We could calculate effect sizes if desired, but are unsure that would show any meaning as well. We have removed figure 5D as we have removed “different year” analyses per previous comments. However, we do argue that what is a “significant” amount of difference here is not something that can be defined by eye.

**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.**  
>penciled in to reword this specifically, and to rework discussion to frame paper better comparing coarse vs not.

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  
  
Comments to the Author  
Summary:  
---------------  
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 ?. We feel that as our methodology is sound and we have provided justification, this example will benefit other researchers who may have also thought of applying the method to their systems which may not be at the same scale, temporal or spatial, as the original cases.

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**.  
  
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 a paragraph citing other studies where variation in predictor variables is known from small regional studies of grassland 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 using this method have found similar justifications for use of the method.

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 above. Additionally, we have removed the different year section as suggested by reviewer 1, such that the predictor and response data are matched and of appropriate resolution to capture these variations.

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.  
  
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 is borne out by the factor that several species do in fact, show improvements with the SEEM models. We have reworded the title and the discussion to emphasize that we, in fact, show that they DO improve prediction for some species.

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 did not test for nonstationarity but to our reading neither did Fink et al or any other STEM paper. We have further explained the necessity of STEM in lines

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 now in lines ??-??.   
  
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 ?? that the support set centers are randomized and now cite this article.

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. 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.  
  
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).**  
>This is still not quantitative. Fig. 8D in Fink et al. 2010 also shows blockiness. What is too low, exactly?

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 resampling.  
References:  
------------------  
Fink, D., Auer, T., Ruiz-Gutierrez, V., Hochachka, W., Johnston, A., La Sorte, F, Kelling, S. Modeling Avian Full Annual Cycle Distribution and Population Trends with Citizen Science Data. bioRxiv 251868; doi: <https://doi.org/10.1101/251868>  
  
Johnston, A., Fink, D., Reynolds, M. D., Hochachka, W. M., Sullivan, B. L., Bruns, N. E., Hallstein, E., Merrifield, M. S., Matsumoto, S. and Kelling, S. (2015), Abundance models improve spatial and temporal prioritization of conservation resources. Ecological Applications, 25: 1749–1756. doi:10.1890/14-1826.1  
  
Kuncheva, L., and Whitaker, C. 2003. Measures of Diversity in Classifer Ensembles and Their Relationship with the Ensemble Accuracy. Machine Learning 51(2): 181–207. [dx.doi.org/10.1023/A:1022859003006](http://dx.doi.org/10.1023/A:1022859003006)  
  
Zuckerberg, B., Fink, D., LaSorte, F. A., Hochachka, W. M., Kelling, S.  Novel seasonal land cover associations for eastern North American forest birds identified through dynamic species distribution modeling, Diversity & Distributions, 1-14, DOI: 10.1111/ddi.12428 ADDED IN INTRO

>We appreciate the additional references, particularly the first paper, as it was not available at the time we submitted the manuscript. We have added them as appropriate to lines ?, ?, and ?. (added to to-do list)