Associate Editor Comments to Author:  
  
Associate Editor  
Comments to the Author:  
We have received a review on this revised version of your manuscript from one of the original reviewers. They find your re-worked version to be much improved, an opinion I share. I think addressing the remaining comments they have should be relatively straightforward and I would anticipate accepting a revised version of your paper.

Response: We appreciate the positive remarks upon our revision and hope that the additional changes will address the remaining comments.  
  
Reviewer(s)' Comments to Author:  
  
Reviewer: 2  
  
Comments to the Author  
Ecology and Evolution   
ECE-2018-01-00114, Revision September 2018  
Title: SPATIALLY EXPLICIT ENSEMBLE MODELS DO NOT ALWAYS IMPROVE SPECIES DISTRIBUTION MODEL ACCURACY  
  
Summary:   
In this revision the authors have improved several technical aspects of the study including the test-set sampling and model evaluation. They have also improved the description of the ensemble of support set and how they are created.   
Response: We thank you for the positive comments on our revised manuscript.

When do you need SEEM / STEM?   
The manuscript does not accurately and completely communicate when these models are needed and how to test for this. This is important so that readers can assess these methods in other use cases.   
  
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 introduction needs to be edited to communicate this information. Also, regional variation in the predictors (e.g. land cover and climate) by itself is not enough to create non-stationarity in predictor-response relationships. These lines in the introduction and discussion suggest that variation in predictors alone is sufficient (e.g. Lines 59 -60, 324-25) and need to be edited.

Response: We have added wording in lines 47, 60-61, 101, and 329-330 to clarify that these relationships and that variation in predictors alone is not sufficient.  
  
Test for non-stationarity:   
The authors are correct that fitting SEEM / STEM type ensembles requires additional computational cost. So, it is natural to ask when is that extra expense worth it. It is easy to test for spatial and temporal non-stationarity in predictor-response relationships: If the analyst suspects non-stationarity is an issue, it can be tested by a small number of base models, each fit different regions &/or seasons and then comparing the predictor-response relationships to see if they vary. The discussion or introduction should describe this test to help equip readers with the tools to evaluate when & where they need to account for non-stationarity in predictor response relationships. This test does not require any expertise, beyond that necessary to run a species distribution model, and it will only cost relatively little in terms of extra computations. Lines 309-10 need to be revised accordingly. 

Response: We have added wording in that paragraph to suggest this test in lines 318-320.

Creating Random Support Sets:   
It’s good to see more information about the ensemble of randomized support sets and their creation.  
•       While randomization of the support set center is an important part of creating the ensemble of support sets, the goal for constructing a good ensemble of base models is to fit low bias, high variance models (Kuncheva & Whitaker 2003). This should be stated as the goal in line 235.

Response: We have edited now-lines 237-238 accordingly.

•       Bootstrapping or subsampling is used to induce sampling variation among base models in the ensemble. Thus, simply relying on the bootstrapping that is done within the base model will not induce variation across base models as is implied in lines 230-231. These lines need to be edited accordingly.

Response: We have edited line 232 to reflect this.

•       Ensemble averaging Is used to reduce the sampling variation across base models. Thus, it is important to average across a sufficient number of base models. If I understand the description, the maximum number of randomized support sets covering any single prediction was 10 or 11 (depending on the size of the support set) and on average closer to 7. This is an extremely small number for this kind of averaging and can be seen in the blocky edges of the estimated distribution maps. The number of base models should be at least 25, if not greater. 

Response: We have added wording in lines 243-245 to suggest that future users of these models repeat the sets more frequently to reduce the blockiness. We lack the funding to repeat the models with larger sets of base models and would appreciate any suggestions on citations to direct readers for how to choose an appropriate sampling size threshold to reach a suitably low level of blockiness.

Figures – Blank areas should be shown in a different (non-black) color so that they can be distinguished from predictions of 0.

Response: They are currently shown in white (blank) and black (0), and we have clarified all captions so that this is clearer.  
  
Line 218 - “SEE Models”

Response: Done.

Line 259 - Remove “1983)”

Response: Done.