General questions

- **Title and abstract**
  - Does the title clearly reflect the content of the article? Yes
  - Does the abstract present the main findings of the study? Yes
- **Introduction**
  - Are the research questions/hypotheses/predictions clearly presented? Yes
  - Does the introduction build on relevant research in the field? Yes
- **Materials and methods**
  - Are the methods and analyses sufficiently detailed to allow replication by other researchers? Yes
  - Are the methods and statistical analyses appropriate and well described? No, in my opinion analyses are not appropriate, see my comments below.
- **Results**
  - In the case of negative results, is there a statistical power analysis (or an adequate Bayesian analysis or equivalence testing)? NA
  - Are the results described and interpreted correctly? I don't know
- **Discussion**
  - Have the authors appropriately emphasized the strengths and limitations of their study/theory/methods/argument? No, see my comments below. (Note that this come from my concerns on the methods so it is a redundant disagreement).
  - Are the conclusions adequately supported by the results (without overstating the implications of the findings)? No, see my comments below. (Note that this come from my concerns on the methods so it is a redundant disagreement).

General comments

Rieger *et al.* tackle the problem of estimating bird population trends when sampling has been done over heterogeneous habitats that do not reflect the landscape composition and by different observers. This is for sure an important challenge for ecology and conservation, for which we do not have clear recipes or workflow. The introduction is well written, and challenges are well explained. However, I am not convinced by the way authors solved these challenges. They used unusual way to overcome this challenge, which is not a problem in itself, but they did not bring any analytical or numerical (simulations) proofs of the relevance of the solutions they proposed. In my opinion the solutions they proposed are not elegant statistically and likely to introduce non-desired bias while poorly correcting for what authors want to correct. I might be wrong, but in the absence of analytical or numerical analyses which show that their corrections achieve what they want to achieve, I am not convinced. Below I detailed few of my concerns.

Otherwise, aside of my concerns about the statistical analyses, the manuscript is well written and clear, with well organized supplementary materials.

About correction for spatial bias:

Authors corrected for sample bias (representativity in habitats) in a very special way that raise lot of questions and possible problems. I do not understand why they decided to weight the model according to the representativity of the sampling for each site and year, and I am not sure this does what the authors want to do. Weighting the model in the way authors do will affect the fit and the uncertainty associated with that fit, artificially decreasing it. I think adding the natural regions as a random effect to the model would allow to account for heterogeneous sampling across habitats in a much proper

way, although not perfect. Below I develop some ideas and example showing why I think weighting the models as authors do is a problem.

Let's imagine a situation with two habitats equally abundant in the landscape. A perfectly balanced sample will be 50% of the sites in one habitat, and 50% in the other one. If 20 sites are sampled every year, 10 should be sampled in each site. Now let's imagine that the first year the sampling is perfectly balanced but get unbalanced on the second year in the following way:

	HABITAT	NUMBER OF SITES SAMPLED	WEIGHT PER SITE
YEAR 1	Habitat 1	10	1
	Habitat 2	10	1
YEAR 2	Habitat 1	15	0.66667
	Habitat 2	5	2

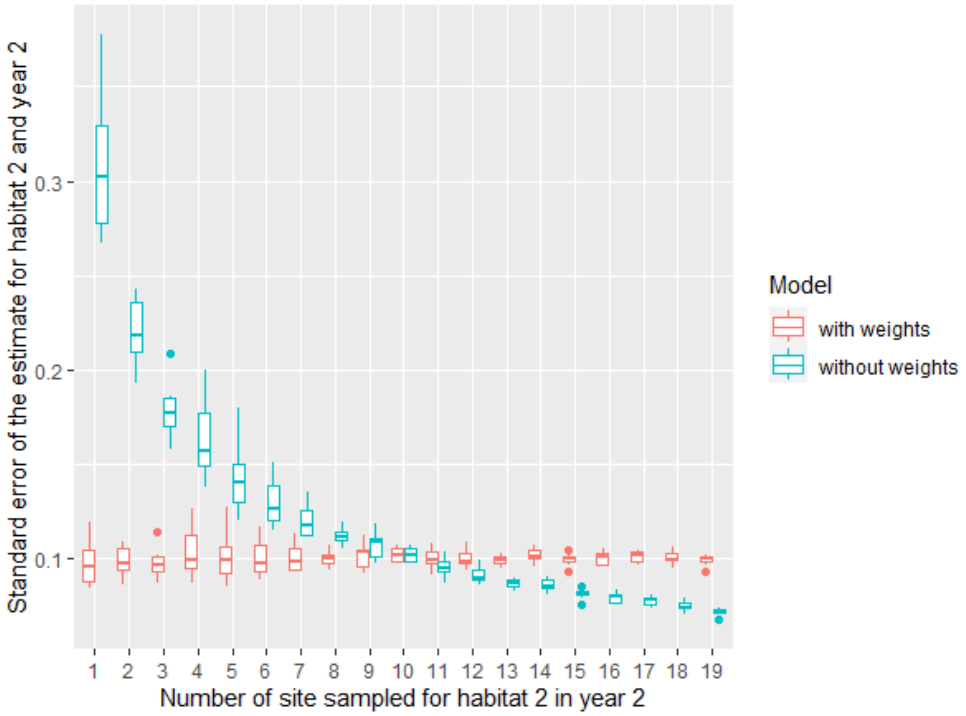
This means that any error of count in the year 2 in the habitat 2 will have a disproportionate importance in the model. Moreover, compensating the undersampling of habitat 2 by increasing the weight of few sites sampled is likely to lead to an overfit of the observed values and to artificially decrease the statistical uncertainty.

I simulated count data for a fictive species for two habitats over two years. As in the previous table, first year the sampling was perfectly balanced (10 sites sampled in each habitat), while in the second year it was unbalanced (20 – *n* sites in habitat 1 and *n* sites in habitat 2). I investigated the consequences of the weighting on the statistical uncertainty of count estimation for the second year for the second habitat, across 10 different random simulations.

To model this 2-year dataset I just used a GLM that can be described as follow in R formalism: *count ~ year * habitat, family= poisson*

where year and habitat are categorical predictors.

When using a classic GLM, without weights (equal weights for all points), the standard error decreased when increasing sample size, which was expected. However, when weighting the model as authors



suggested, I found similar standard error, regardless the sample size. This suggests that the weighting method proposed by the authors artificially decrease statistical uncertainty, making it independent of sample size, thus producing unreliable results.

Also, in the case of a habitat is not sampled on a given year, the sum of the weights would not be constant across years. This will give more importance to some years, which can strongly bias trends if by chance these years exhibit specific climatic conditions or whatever else that affects bird population. This correction is likely to bias the model but does not provide any information on the missing habitats and thus does not allow a proper correction. In contrast, although it is not perfect, including the habitat as a random effect allows the model to account for unbalanced sampling and missing habitats for some years, while properly estimating statistical uncertainty.

About the observer effects:

Again, I find the way authors account for observer effect very special and not appropriate. The observer effect is an outlier effect rather than an observer effect. They do not correct neither for the identity of the observer neither for its experience. In the introduction, authors described precise mechanisms explaining observer effects, but then use an "observer effects" that is disconnected from any mechanism and do not model what they want to account for.

If a species exhibits a decline/increase in population size on site, "observer effects" will be detected at the beginning and at the end of the time series, just because of the way they are detected. The measure used by authors relate to outliers but not to observer effects.

Moreover, even as an outlier effect, the used by the authors measure seems badly build because it does not account for the inter-annual variability in abundance. to detect outliers, authors used only deviation from the mean and not the standard deviation associated with the mean. Some species are more likely to exhibit inter-annual variations and thus will exhibit higher percentage of outliers. Similarly, some habitat, are more likely to exhibit strong inter-annual variations, and thus will exhibit higher percentage of outliers.

I do not understand why authors do not use a random observer effect to account for difference in the baseline detection levels among observers. To that random effects they could add an effect that model if the observer is naïve to that site or not to account for site-experience.

Specific comments

Lines 44-53: the organisation of the challenges is not super clear for me. I am not sure I see well what distinguished the prime and the second sources of challenges. For example, from what I understand, the estimation errors due to variation in experience and the imperfect detection seems to be overlapping sources of errors. More experienced observers are likely to provide better estimation because they detect better the birds.

Lines 189-194: using a zero-inflated model has the advantage of allowing to model rare species, so I struggle a bit to understand why authors perform this pre-model filtering. It is not really a concern but I think this choice could be explained to the reader.

Lines 208-227: In addition to what I detailed in the general comments, multiple modelling choices are obscure for me. Authors used a GAM, that in contrast to GLM, can estimate nonlinear effects without fixing the number of degree of the polynomial. Instead of using this advantage for the effects of the site-specific environmental attributes, they used a fixed polynomial effect of degree 2 to estimate non-linear effects.

Also, in the description of the data collection, authors said that the number of visits per year and time of these visits varied across sites, but there is no variables related to sampling time/pressure included in the model. I think this point is key and authors should explain why.

Table 1:

The main model formula described in the caption of table 1 is:

~ s(survey year, by = R) + survey year + poly(PC1, 2) + poly(PC2, 2) + poly(PC3, 2) + OE + (1|ID)

which suggests that authors included survey year as a linear effect and as a smoothed effect, but this is not explained in the main text. Is that a typo?

Lines 243-244: eight chains is a lot, why did authors used so many? How did authors checked the convergence of the model?

Lines 251-256: If I got it well authors used this selection value to select the most appropriate model? Why not using the metrics that are classically used to perform model selection (BIC for example)? This choice is again unusual and not properly argued. Would authors find the same best model if using the Bayesian information Criterion (BIC)?

Lines 257-258: I did not get what authors called posterior probability. Probability of observing each predicted values? Or posterior probability at the model level?

Line 273: what is the meaning of those fine-scale changes? It looks like authors were interested in first derivative of the function describing temporal changes, but considering that the sampling is discrete (every year) I struggle to see which kind additional information it brings, because interannual changes already relates to the first derivative of the temporal changes.