Revision round #1

Decision for round #1 : *Revision needed*

Dear Authors,

Thank you for submitting your preprint for review at PCI Ecology. Your manuscript entitled "Analysing bird population trends from monitoring data with highly structured sampling designs" has been thoroughly reviewed by two reviewers. While both reviewers highlight the importance and difficulty of dealing with unbalanced sampling when estimating population trends, they both raise several important issues. Notably, both reviewers pointed at the fact that the proposed weighting method artificially increases statistical precision for underrepresented regions (which is briefly acknowledged in the discussion). More generally, since the aim of the manuscript is to provide a new method that better deals with imbalanced sampling, then I concur with the reviewers that 1) the actual efficiency of the method should be better assessed, e.g., using simulations, and 2) this method should be compared directly with existing methods.

You will find their detailed reviews below.

I look forward to reading a revised version of this preprint.

Best wishes,

Matthieu

*by **Matthieu Paquet**, 18 Oct 2024 12:57*

Manuscript: **https://doi.org/10.1101/2024.06.30.601382**

version: 1

First of all, we wish to express our gratitude to both reviewers and the recommender for their exceptionally thorough and insightful consideration of our manuscript and the associated R scripts. We acknowledge that the raised concerns were well-justified and combined with very useful and constructive suggestions for revision. Not all of them could be implemented in the proposed way as detailed below, but we are confident that our substantially revised analysis finds solid solution to take all raised aspects into account. As for the aspects specifically mentioned by the editor:

(1) Weighting: We have removed the earlier weighting procedure for unbalanced coverage of natural regions, given its unwanted effects on estimated standard errors. Instead, we now model population trends separately per landscape type and combine these trend estimates in a post hoc procedure along the lines proposed by referee 1 and referee 2.

(2) Establishment of a “superior” method to deal with imbalanced sampling: We concur with referee 1 that the emphasis of our work is to develop an analytical protocol that as much as possible reflects key properties of the underlying sampling design and data distributions. This does not imply that our approach is generally superior to alternative trend analysis techniques, but we provide bits and pieces that will likely also be useful to others with similar sampling and data structures. We made this emphasis more transparent in the revised manuscript.

Review by anonymous reviewer 1, 26 Jul 2024 13:10

The paper proposes GAMs combined with weighting as a method for adjusting trend estimation for strong imbalance in population survey data. Unbalanced sampling is a common issue in population surveys and statistical tools and models developed for such data typically deal with unbalanced sampling one way or the other. The TRIM software commonly used for bird surveys in Europe is e.g. designed specifically to address such issues, using a combination of site effects and imputation. Imputation/post-hoc weights (i.e. fitting the model to available data, then weighting model predictions) is probably the most common approach to try to correct for imbalance taken in previous studies.

The authors here instead use a weighted Bayesian likelihood to adjust for imbalance, which in general is an interesting idea. Also, while imbalance has been often discussed and considerable effort is often put into addressing it, I know of few studies that thoroughly investigate strong imbalance in animal survey data. Another interesting idea proposed here is to adjust for survey effort using total counts across species.

Given that the papers is written with a methods focus, I found it somewhat lacking comparisons to previous approaches, particularly regarding the likelihood weights. If the aim of the paper is to provide a method that is better at handling strong imbalance than other methods, which is how I read it, then more effort needs to be put into showing how the approach is better at dealing with imbalance. On the other hand, if the aim was to show how imbalance may be dealt with in the case study, then the "methods" aspects of the paper could be toned down and new material may not be needed (i.e. suggest that this is one way of dealing with imbalance rather than it being a better way).

If the former route is taken, we need to know more about how the suggested methods/models compare to other methods, e.g. imputation methods such as those in TRIM, or perhaps models such as those of Harrison et al 2014 (spatio-temporal GAMs fitted without weights but with predictions over a representative set of points).

Harrison et al. (2014). Assessing trends in biodiversity over space and time using the example of British breeding birds. Journal of Applied Ecology, 51, 1650--1660. https://doi.org/10.1111/1365-2664.12316

Indeed, the emphasis of our work is to develop an analytical protocol that as much as possible reflects key properties of the underlying sampling design and data distributions. This does not imply that our approach is generally superior to alternative trend analysis techniques, but we provide bits and pieces that will likely also be useful to others with similar sampling and data structures. We made this emphasis more transparent in the revised manuscript.

The likelihood weights also need to be discussed more. As far as I understand, this weighting is essentially ad-hoc. The weights are constructed to adjust for imbalance, but there is no particular motivation for the resulting weighted likelihood. Intuitively, the weights are cleverly constructed and make a lot of sense in that sites that are in over-represented bioregions get downweighted, as illustrated in Fig. 7. However, the flip side is that surveys in under-represented regions get upweighted. In effect, if I am understanding correcly, abundances at such sites will be treated as if they were multiple independent data points. A site with a weight of 2 would be treated as 2 data points. Thus positive weights artificially inflate precision provided by corresponding data points, which will lead to underestimation of uncertainty from those points. This issue is briefly mentioned in the Discussion, but needs to be made more explicit and more thoroughly discussed. It may have limited effect in this particular study because most weights are small, but as a general principle it seems problematic and hard to justify.

Another issue with the weights is that they cannot adjust for imbalance if data are missing entirely from one of the categories. This is solved here by smoothing over time. However, the meaning of the weights becomes less clear in such a case. The point of the weights was to get data that are balanced among the classes, but since this cannot be done when a class is missing it could be argued that all weights should be zero (this also makes some mathematical sense as the equation on line 151 suggests infinite weights for missing areas).

The above ad-hoc nature of the weights should be stated up-front, not in the least to motivate future work with more rigorous derivation of weighted likelihoods. Alternatively of course, you could provide more rigorous justifications yourselves.

Thanks a lot for spotting this issue with likelihood weights as implemented in brms, which we weren't aware of. Indeed, upweighting in brms artificially increases sample sizes for the underrepresented groups, leading to an underestimation of standard errors from the respective natural regions. Even though we found – using simulation as proposed by referee 2 – that the effect on our overall trend estimates largely balances out across the natural regions, we decided to abandon this questionable weighting procedure altogether. Instead, and along the lines proposed by referee 2, we now estimate population trends separately per landscape type (so that standard errors are appropriate for the realized sample sizes per landscape = natural region + metropolitan area), and use post hoc trend averaging according to each landscape share within a given biogeographic region or the entire federal state.

L 44. To me this feels like 6 sources rather than 2.

(see similar comment by referee 2): We agree that this section was poorly organized. We streamlined the listed items and integrated them into the description of the three prime challenges that are relevant for the implementation of our own analytical routine.

L79-80. TRIM is indeed widely used in Europe, however, I would not agree that it is the "current analytical standard" (although I'm not sure I correctly understand what this means). E.g. it is not widely used for North American data, and many other types of models have been suggested for this type of data.

We rephrased to point out that TRIM is 'a common analytical tool' (l.79)

L85-86. It is not clear in what sense "TRIM reaches its limits"?

We rephrased to specify that robust imputation and trend estimation is considered to take place unless missing values (i.e.: NAs for year-by-site combinations) exceed 60%. This is a specification of the authors of TRIM (l.85ff).

L89-91. Not fully true as TRIM can estimate annual indices and include breakpoints.

We agree and reformulated to reflect all three options offered by TRIM (l.89ff).

L144-147. If the balance of the subsampling was changed, then it cannot be a strict 6-year rolling scheme as suggested in the first paragraph of the methods. This needs clarification.

That is correct, we have reformulated the first paragraph to clarify that restructuring of the survey protocol added some variation in the alternation cycle (l.113ff).

L149. It is not obvious what "weighted response measures" means. Explain how the weights work. My understanding is that the weights multiply the log-likelihood contribution of each data point.

Please see our response above to the general issue with weighting: The procedure has been abandoned altogether.

L195. Need more information about the response. What is "abundance", is it a count or something else?

The term abundance is explained at the end of the first paragraph in the Methods (l.118ff): "*These repetitive surveys are used to derive the number of territories for each species per km², which represents the response variable abundance in all statistical analyses…*". We briefly repeat this now as a reminder when introducing the model structure.

L201-204. How was this assessment done, i.e. how did you determine if there were excessive zeroes or overdispersion?

If excessive zeroes were not accounted for (no zero-inflated model), model simulated data reveal an underestimation of the proportion of zeroes (propZ), whereas non-accounted overdispersion shows an underestimation of mean abundance, as outlined and illustrated in section 3.1 in the Results. We have tried to more clearly link to the relevant graphs and explanations in the text.

L204. Just 0.5, or also 1.5, 2.5 etc? If so, an alternative is to use 2*abundance as the response with an offset = log(2).

Thank you for this idea, this is indeed a good alternative since we had .5 abundances per se (not only 0.5). We deleted the 0 to make this clear and reformulated the model formula accordingly (l.197ff and l.201ff).

L216-219. The use of weights for natural regions suggests that you believe the trends are not the same, so I don't follow the argument for only estimating separate curves per bioregion?

We focused on bioregions because these are the spatial scale at which the federal agency has to annually document and report population trends. Due to our adjusted procedure to reflect spatial sampling bias (weights were abandoned, see above), we now model smoothed trends per landscape type (here: six national regions plus the metropolitan area as separate 'landscape') and predict trends per biogeographical region by using weighted mean predictions for these landscape types (the weight now corresponds to the respective share of each landscape type per biogeographical region, see l.272ff).

L208-227. The model formulation is not crystal clear to me from the text. I found the model formula in the footnote of Table 1 but it felt somewhat tucked away there. I would suggest to display the formula more prominently, perhaps in the main text or at least with a reference in the main text. Also, the main model formula appears not fully correct as survey_year appears both in s() and as a linear term which would be redundant. Perhaps you meant

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

so that intercepts can vary among regions as suggested in the documentation for the by argument in ?mgcv::s.

What priors did you use?

Thanks a lot (also to referee 2) for spotting our error in the model formulation. Your revision correctly reflects what we did. We corrected the formula accordingly (l.201ff).

We used weakly informative priors (normal (mean = 0, SD = 2.5) for categorical coefficients and normal (mean = 0, SD = 10) for numeric coefficients (polynomials and linear trends)). This was previously 'hidden' in supplement section 4. We now added a short description to the main manuscript (l.242f)

We agree that our description of the model components becomes clearer when presented in conjunction with the model formulation. We therefore now placed the model formulation in the main text (instead of Table 1), and this is followed by brief explanation of each model parameter.

L245. Why were these four statistics chosen? I.e. why do you consider them particularly important for estimating population trends from these data?

We have added a brief justification for the criteria used to assess how well a given model structure captures key characteristics of the abundance distribution (l.245ff). We consider this a necessary (though of course not sufficient) prerequisite for generating reliable population trend estimates. Our model selection is now mainly based on kfold-cross-validation (see comments concerning the selection value and BIC) and these four statistics are now mainly used to visualize how different models result in different distributions.

L247. How did you treat random effects when simulating new data for the p-values? E.g., where site effects regenerated, or did you use their estimated values?

Our predictions did not include specific Group-level effects (= random effects), but we generalized beyond the specific grouping levels using re_formula = NA when fitting predictions via fitted.brmsfit(). This is now specified in the text (l.271f).

L248-253. Bayesian p-values close to 0 and 1 do flag issues as you suggest, but there's no reason as far as I know to expect that they have to be close to 0.5 for a good model. This calls the "selection value" into question. Optimizing to get p-values as close as possible to 0.5 could lead to overfitting. The selection value also seems to suggest that one can somehow trade off a single poor p-value with better ones. E.g., if two p-values for one model are 0.01 and 0.5, the selection value would be 0.24. Another model might have the same two p-values both equal to 0.85 and a selection value of 0.25. This would suggest that the first model with the strongest indication of poor fit should be preferred. I also don't

understand how you are using this value and why it is needed. Unless there is some justification for it that I'm missing, I suggest to omit.

We agree that our 'hand-made' procedure to calculate a selection value would require far better justification and exploration of its behaviour in extreme situations. It may also appear inappropriate given established alternatives for model comparisons – and we had skipped those in our initial manuscript version because many of the 'classic' Bayesian options for model selection like loo (leave-one-out cross-validation) or WAIC ruled out because they often fail to estimate reliable values when models contain random intercepts. After some further testing, for the revised manuscript we now settled on K-fold cross-validation for model selection, where the model is refitted K times, each time leaving out one Kth of the original data and predicting it. As an objective criterion to reject model types we require 95% CI (1.96*standard error) of the elpd-diff to include zero. Among those models that are then still retained, we purse a manual preference for the most parsimonious models (given that there were no convergence issues), i.e. preferring models in the sequence

Poisson > negative binomial > none zip > none nb > PCA zip > PCA zinb > L zip > L zinb > F zip > F zinb.

We outline the procedure and display outcomes in (l.256ff).

L265-267. How were they combined? Were predictions combined at the log scale or at the response scale? If the former, how should we interpret the combined trend estimate in terms of abunance? If the latter, how did you deal with imbalance in covariate values between regions? Explain and motivate the computation.

Thank you for raising this important point! We combine trends at the response scale, but in the previous version did not account for imbalance in covariate values. In the revised version, we now adjust predictions for the mean value of each covariate (PC1, PC2, PC3) in the given landscape (former: biogeographical region). We changed and explicitly explained this accordingly (l.269f).

L269. Notation ab here is slightly confusing (is it a times b, or a single variable ab?). Use a single letter instead (why not N if ab is expected abundance).

This is indeed a single variable, abundance. We changed this as suggested to *N* (l.280 & l.292).

L270. For clarity it would be better to refer to the *i* as posterior draws or posterior samples instead of simulations, if that's what you mean.

We changed this accordingly (l.281 & l.293).

L281. I don't see why you'd want to use the posterior mean in the denominator. I would rather have expected you compute the posterior distribution of the index, i.e. make posterior inference for abij/abiJ, just as you seem to have done for the differences on L269. The uncertainty of the index would then be 0 in 2006 as expected, similarly to e.g. TRIM indices.

You are right, if we want to compute similar indices according to TRIM (with no uncertainty at the base year) this would be the correct computation. However, on purpose, we sought a procedure that displays uncertainty also for the baseline year, also to avoid the problem that the uncertainty of other years is dependent on uncertainty in the baseline year (which is the case for indices computed similarly

to TRIM). We added a remark how users can switch to calculating ‘classical’ indices (no SE for baseline years) within our workflow (l.296f).

L293-295. How does the MhB match temporally to the EAS which was collected during a long period from February to July?

The MhB collects data between 10 March – 20 June, so the core breeding seasons match between the two programmes. We added this information from Kamp et al. 2021 to l.308ff.

L299-301. I can't follow the sentence, rephrase?

We rephrased to make our point clearer: for some species, consistency among trend estimates will be low as a consequence of species biology coupled with the different trend estimate routines (i.e., TRIM calculates yearly indices, which inherently also reflect year-to-year fluctuation in species with strong between-year variation in breeding population sizes. These are inevitably lost when smoothing populations trends as in the EAS-based approach).

L304-305. This 'coincidence' will be large if there is large uncertainty in the EAS estimates. They could also be very different from the 'coincidence' computed by reversing the roles of the two estimates (i.e. the fraction of EAS estimates within 95% Mhb intervals). Would be good with an explanation of what it is supposed to measure.

For the reasons outlined already in the previous manuscript version, the assessment of consistency between (our) GAM-based smoothing trends and (the standard MhB) TRIM indices per year is a big challenge, with probably no entirely satisfying solution. We share the reservations of this referee against our ‘coincidence index’, which is sensitive to the chosen reference trend and has the unwanted property to increase when uncertainty in the reference trend estimate increases. We therefore decided to capitalize on one established measure for trend consistency, namely the ‘median symmetric accuracy’ (MSA), which estimates the median percentage error between both trend estimates. This selection goes along with your comments on line 414-15 below. MSA, combined with correlation coefficients as used previously, now form the basis for our comparison of the MhB and EAS trend analyses.

L320-324. More details about the model comparisons are need: How did you "optimise models"? What does "best-performing model" and "comparably good performance" mean? From the Table caption you seem to be using 0.3 and 0.7 as cutoffs for "good performance", was there any rationale behind this choice? Why do you need to determine a "best-performing model" when you are going forward with the most parsimonious model with "good performance"? How do you determine which model is "most parsimonious"?

See our response above regarding abandonment of the selection value and our revised procedure for model selection (based on k-fold cross-validation and model parsimony).

L338. Since you include a random site effects, your model should be able to capture among-site overdispersion even with a Poisson response. The NB might rather account for overdispersion among

years since the model can only capture smooth inter-annual variation (there are no random year effects as I understood it). Or it might pick up overdispersion in the interaction between year and site.

Thanks for pointing this out. Indeed, overdispersion should primarily arise from within-site and /or between year variation, and we reformulated accordingly (l.369ff).

L365. I'm not sure what the Poisson histograms in Fig 3 represent, do they just show the Poisson distribution with mean equal to the overall mean of the data? If so I don't see the point of including it as this distribution says little about whether the Poisson is a reasonable response for the model (i.e. with covariates, random effects etc histograms could look very different even if the response was Poisson).

Fair point. Indeed, the displayed Poisson distributions were simulated from the observed overall mean. In the revision, we adjusted this so that the Poisson distribution is now derived from the species-specific Poisson model, where we predicted the given dataset and plotted the data as histogram.

L382-388. It would be interesting to see how this would compare to trends derived using the unweighted model for prediction and then weighting the predictions, i.e. post-hoc weighting. Would also be interesting to see how it compares to other approaches, as I suggested above.

See our response concerning the implementation of weights above.

L390-. This is interesting. Do you have any explanation for these strong effects? Is it just a question of observer experience. If so, why are they so large? Could there be other explanations?

In recent years, LANUV often hired students from a nearby university. LANUV assumes that they were highly motivated and put more effort in mapping the site (probably combined with less experience in mapping out territories) and therefore overestimated abundances.

L409-411. It is hard to interpret this result at present because the CIs of the EAS do not measure uncertainty of the relative index, see previous comment.

This became redundant with our revised method concerning trend comparison (MSA) since it is now based on mean values and does not incorporate the calculated CIs.

L412. You could reduce this sensitivity by using the mean across the full time period as the baseline (if you have access to the MhB data), instead of using a single year.

This sounds like a good solution. Unfortunately, we only have yearly MhB-indices and no abundance data, therefore we had to stick to a base year instead of a base period.

L414-415. While the trends in Fig 7b are consistent in their general shape, they suggest very different magnitudes of change. Something that the correlation misses. In other words, you could have perfectly correlated indices that show entirely different magnitudes of change, which may be worth pointing out. Also, you could consider complementing the two measures with something that measures the

difference in values between the two indices, maybe something based on the differences between log-transformed indices.

See our response above regarding the coincidence index, which we replaced MSA following your suggestion of log-ratios.

L483. To strictly evaluate bias one needs to know the truth which we don't here. If you mean that inclusion of observer effects reduced the trends, or that trends came closer to MhB, then say so instead.

Fair point, we adjusted the paragraph accordingly (l.488ff).

L492. Not sure what flagging of extreme years you are referring to, or how that would counter the effects discussed here.

We referred to observer effects classified as 'positive' or 'negative' (extreme values compared to mean total abundance). If one site has indeed an increase or decline in total abundance, mean total abundance will also increase/decrease and thus the threshold of mean total abundance ± 25% increases or decreases, too. We reformulated the sentence to point this out (l.497ff).

L493-496. This can also go the other way. Some species are easy to detect and counts may not be strongly linked to observer skill. The correction could bias trends for such species.

When observer effects are classified as 'positive' or 'negative' even though a species is easy to detect and therefore not affected by these observer effects, the model results in non-directional coefficients for the observer effect levels (e.g., Black Woodpecker, Common Buzzard or Common Kestrel, see appendix section 8). If such species showed a directional trend in abundance, the model would only correct this trend if its strength distributed differently between the three observer effect levels (e.g., a stronger positive trend under positive observer effects, or vice versa). We agree that our approach (but also other approaches that take into account observer ID, or more direct estimates of observer effects such as individual skill or effort) would over-correct in such unlucky cases where – by chance – an increasing trend would mainly occur in sites where more recent surveys are classified as showing a 'positive' observer effects (while no increase occurred in sites where more recent surveys are classified as 'none' or 'negative' observer effect).

L507. What covariates do you mean?

We referred to the covariates described in that paragraph, namely observer effort, age and observer- and species-specific detection rates. We added them in brackets to clarify this (l.514f).

Reviewer 2

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.

Thanks a lot for your thoughts and suggestions on the analysis on which we comment below.

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:

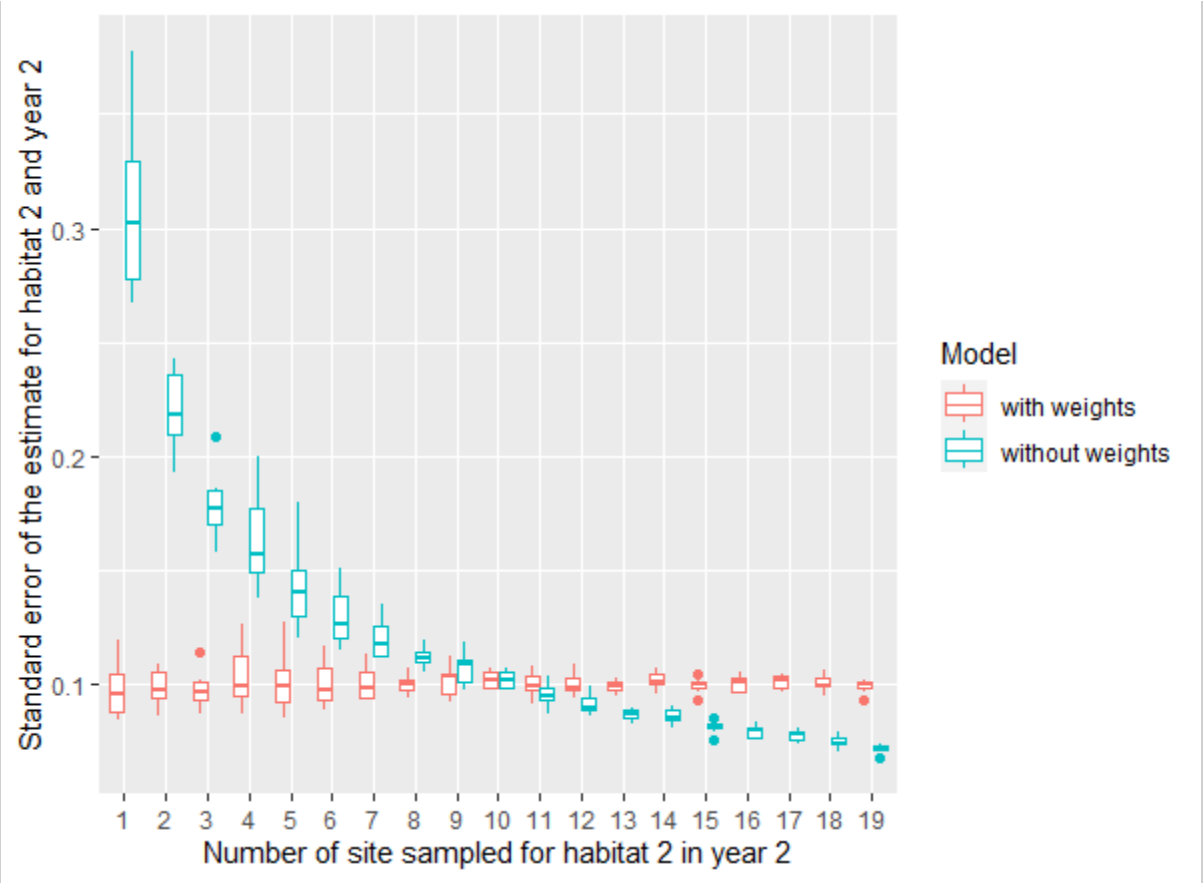
	Habitat	N Sites	Weight per Site
Year 1	H1	10	1
	H2	10	1
Year 2	H1	15	0.6667
	H2	5	2

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.

Thanks a lot for pointing this out – we were not aware that weighting artificially increased (or decreased) our sample size and thus estimates of uncertainty (see also corresponding comment and response by referee 1 above). We tested this effect on some own simulated data within the data range covered in our models. When weights were calculated within one biogeographical region (so mean weights for that region are around 1) this effect essentially levelled out (because increases and decreases are balanced). However, this does not resolve the issue that exceptional high or low years with a high weight can distort our results.

In our revised model approach, we therefore settled on modelling smoothed trend across years per landscape type (the 6 natural regions and one metropolitan area) instead of biogeographical region and predicted trends for these two regions by summing up weighted predictions per landscape type (according to their share of the biogeographical region). We adjusted the manuscript and accompanying R scripts accordingly (most changes were in section 2.3 and 2.5)

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.

The reason why we use this formula to delineate observer effects is that we cannot use the suggested parameters which we list in our introduction and discussion (effort, age, experience) since this data is not available for the EAS monitoring programme. Adding Observer Identity as a random effect is also not feasible because of the high number of different observers (175) with more than 70 observers participating only once and more than 30 observers only twice. Therefore, the number of replicated surveys per observer is (i) very low and (ii) completely unbalanced; adding observer ID as a random effect would thus push our model even further. With our approach we try to correct for this lack of information concerning observers although we know that there are better approaches (which should be preferred when available), which we also state in our discussion.

Concerning the formula itself, we suspect that there was an important misunderstanding: Note that we identify excessive over- or underestimation of bird records from calculations across ALL species (total abundance/mean total abundance), but not separately per species. Hence, an entire survey of a given site in a given year will be classified as 'positive', 'none' or 'negative', without differentiation between species. We intend to identify site-year combinations where total bird counts are clearly more positive or more negative than other years since a total increase (or decrease) in bird abundance summed across species by more (or less) than 25% is quite unlikely. Increases or decreases of single species are usually masked by decreases or increases of other species resulting in an overall ± steady abundance. We reformulated our methods to clarify that these calculations are across species and not species-specific. Even though this procedure will not be able to correct for all variance caused by observer expertise, we are confident that it does solidly mitigate severe over- or underestimates that arise from lack of 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.

(see similar comment by referee 1): We agree that this section was poorly organized. We streamlined the listed items and integrated them into the description of the three prime challenges that are relevant for the implementation of our own analytical routine.

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.

We consider it plausible to restrict the analyses to abundant species that are sufficiently widespread (i.e., represented in a sufficiently large share of sites or surveys) so that we can expect to derive a reasonably meaningful trend estimate. With a 90 % zero count threshold, we are already pushing limits quite heavily. Note that rarer species are typically covered by separate monitoring schemes that do not rest on a random site selection. The 90 % are indeed an arbitrary threshold. We try to make these considerations a bit more transparent (within the text limitations).

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.

We see the main benefit of smoothing in modelling the abundance trends. Using polynomial terms for the environmental variables (PC1-3) was inspired by (i) the expectation that – along environmental gradients such as elevation, temperature etc. – we typically expect linear or hump-shaped (unimodal) relationships that are sufficiently captured in polynomials to degree 2, coupled with (ii) the idea that users may be interested to report and interpret linear and quadratic coefficients for these environmental associations (which is much easier than for smoothers). If users have diverging preferences, we see no problem to adjust the analysis to integrate smoothers also for these variables as long as the models still run properly and don't reach computation or convergence limitations.

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.

The reported values and spans are derived from the handbook of the monitoring programme where they characterize expected survey efforts. The proposed number of visits per year varies for example with site-specific species occurrences and their optimal survey period (e.g. woodpeckers, owls, long-distance migrants with late arrival and therefore late breeding periods, …). Survey duration usually depends on the habitat, e.g. you would need more time to map breeding territories at 1 km² forest than at 1 km² agricultural area.

Unfortunately, the currently available dataset does not contain site- and year-specific information on the realized survey effort underlying each individual abundance value (we added a remark at l.120ff). We consider this a relevant improvement for trend analysis from this specific monitoring program.

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?

Thanks a lot for highlighting this, it is indeed a typo. The corrected formula is now given in the main text (see also corresponding comment of referee 1).

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

Model convergence was checked using Rhat-values, effective posterior samples and Monte Carlo standard errors as stated in the supplement. We now moved this paragraph to the manuscript (l.253ff).

Concerning the chains, we agree and reduced them to 4 for the revised analysis (l.244f).

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)?

Yes, we used this value for model selection. Using BIC for models obtained from brms is not possible since this criterion is not supported by the authors of the package as it assumes flat priors. Calculation of WAIC as well as leave-one-out cross-validation (e.g. LOO-CV or PSIS-LOO via the loo package) for our models resulted in error messages resulting from our hierarchical model structure and suggested that the resulting values were unreliable. We therefore now use K-fold cross-validation where the model is refitted K times, each time leaving out one Kth of the original data and predicting it. We initially refrained from this approach since it is computationally consuming given refitting Bayesian models. Our far more simplistic and easy to calculate selection values essentially resulted in the same models we would choose by visual model assessment. See also our comments concerning model statistics and the selection value for referee 1.

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?

The posterior probability is the proportion of all posterior samples for a given coefficient estimate that exceed zero. This can be used to interpret coefficient estimates in terms of ‘robustness’ (or evidence in favour of directionality. We adjusted the paragraph accordingly to clarify this (l.263ff).

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

We initially implemented these fine-scale changes since smoothing of trends sometimes results in changes in trend direction than fall within two years – but we agree that such minor changes are questionable to be 'real' trends that do not provide relevant information and therefore excluded this part from the revised manuscript. For graphical display, changes in slope are now coloured according to year-wise differences.