We would like to deeply thank Prof. Yoccoz, for his interest in our study and his editorial work, as well as the two anonymous reviewers who provide invaluable insights and comments on the manuscript which greatly improve our initial work. We responded to all their general remarks and addressed their comments separately below.

Editor

Mr. Nabias

Two reviewers have carefully evaluated your paper, and the reviews are positive. They have provided many constructive comments, that will help you revise the paper. I concur with the reviewers that your analyses represent a nice case of hierarchical modelling.

Your paper is also a useful reminder that there can be a large uncertainty in population size estimates of common species in a country that is large and relatively heterogenous. As reviewer #2 wrote, the difference between the two estimates cannot be simply interpreted as a measure of bias, as your estimates using citizen science data do not represent the truth, they are likely to be biased even if less than the original method, so the real bias when comparing to the first estimates might be lower or higher than the one you refer to when writing "over/under-estimation". It is enough to mention the large differences in estimates between the two methods.

I was also wondering if there is any information available from other countries with bird monitoring schemes that could inform the estimates you discuss, particularly when there is a large discrepancy. I understand that it is hard to compare France to e.g. Germany, Spain or the UK, but it might be worthwhile for at least a few species to compare your estimates to those available elsewhere.

Best regards

Nigel Yoccoz

- Thank you for considering our work. In regards to your comment, we added a section in the discussion entitled “Comparison to other European countries”, including also a discussion on the potential limits of such comparison. For instance, not accounting for the respective habitat availability within each country might make such comparison difficult to achieve because bird population estimates are generally habitat-specific for most species. We further suggested in this section either to employ Integrated Models to account for such country-specific sampling differences while still modelling both ecological and observation processes, or alternatively to promote a standard data recording protocol across countries, such as the Pan-European Common Bird Monitoring Scheme (PECBMS) to improve data harmonisation among European countries. **(L459 - 482 and table S6.4)**

Reviewer #1

This paper is at the interface between methodological development and applied ecology. While threatened or restricted-range species often benefit from exhaustive counts and surveys, monitoring common and widespread species is often neglected or poses significant challenges. In this context, citizen science (CS) emerges as a powerful tool for long-term monitoring, engaging volunteers to collect data across vast spatial scales. Despite the substantial increase in sampling efforts facilitated by CS, issues of detectability persist, potentially leading to the oversight of even common species in suitable habitats. This study draws on data from a large French CS program, EPOC-ODF, which amassed over 27,000 complete checklists across nearly 4,000 sites during the 2021-2023 breeding seasons. Using Hierarchical Distance Sampling (HDS), population size estimates were derived for 63 common bird species. Comparing these population size estimates to those from a previous expert-based atlas reveals significant underestimations in the atlas, likely due to conservative estimates. Some species with long-range songs were overestimated. The findings stress the importance of employing robust statistical methodologies to ensure unbiased ecological inferences and advocate for increased use of structured CS for biodiversity monitoring.

The manuscript is well-written, and the authors have managed to make it flow smoothly despite the multitude of in-depth analyses presented. The research question is quite clearly presented, and the different components of the article well linked to this question. The study's context is well-defined, elucidating the novelty the authors aim to underscore and the associated challenges. Note that I'm not an expert on the methods used in this paper nor citizen data analysis. Hence while some of my comments may be a bit naive, I believe they can be useful as other readers may share similar misunderstandings.

Although the methods used are quite innovative and present several advantages, the manuscript could highlight better a number of points regarding concerning the assumptions underlying the models and the associated methodological choices, that seem crucial and are not covered sufficiently at this stage in my opinion. The authors could also spell out in greater detail the limitations of their approach, which might lead them to be more cautious about their overall conclusions. With a view to improving our knowledge of abundances, which as the authors explain is a non-trivial issue, it seems important to make the reader aware that the approach can be further improved and the potential consequences.

General comments:

Results and discussion focus heavily on comparison with the old method, and perhaps too little on the current method. In particular, the probability of detection, which seems to me to be at the heart of this new approach, is not sufficiently discussed. Among other things, the influence of methodological choices and selected covariates is only very slightly addressed. In particular, I suggest expanding the bibliography on the subject. This is an example, but other could be used: Joshua H. Schmidt, Carol L. McIntyre, Margaret C. MacCluskie, Accounting for incomplete detection: What are we estimating and how might it affect long-term passerine monitoring programs?,Biological Conservation,Volume 160,2013, Pages 130-139,ISSN 0006-320.

Besides, I didn't quite understand how the different intra-annual replicates were incorporated into the model. I think I understood that they were linked to the probability of detection via the date and time of the survey, but I think this could be explained more clearly. If that's the case, why didn't the results section deal with the phenology of detection over the course of the season and the day, in order to suggest potential improvements to the protocol in the future, for example?

- We have now clarified the modelling approach section using tables (now Table 1, in the main text) and figures (now Figures 2 and 3, in the main text) previously presented in the supplementary information section. These figures and tables allow better describing the modelling process used and how we tested multiple hypotheses to better capture bird species phenology. We relied upon a model averaging approach across all tested hypotheses.

- Given the nature of EPOC-ODF scheme, clarified in the new version of the section "sampling protocol", the scheme does not provide inter-year replicates and collects data solely over a three year-long timeframe. As such, we suggest that schemes such as the FBBS could better describe and inform ornithologists on species detection variation across time and years with its robust design spanning from 1989 to 2023, with revision in 2001 and 2015. However, specific trends from these schemes rely on TRIM (TRends and Indices for Monitoring data) that does not account for the observation process, which is precisely why we weighted our $\delta_{????}$variation in the PGLMM model.

- Your relevant comment was also raised by the reviewer #2, so that we assessed the model assumption of the N-mixture model on detection probabilities (species availability) and highlighted the divergence between the framework used in our study, based upon the conceptual framework of Chandler et al., 2011. We also present now the limits of the framework used, more specifically on the assumption of a homogeneous distribution of individual within-samples (Mizel et al., 2018). See section "Discussion – Study limitations" **(L483 - 528), specifically on that matter : L508 - 519)**

- *Chandler, R.B., Royle, J.A., King, D.I., 2011. Inference about density and temporary emigration in unmarked populations. Ecology 92, 1429–1435. https://doi.org/10.1890/10-2433.1*

- *Mizel, J.D., Schmidt, J.H., Lindberg, M.S., 2018. Accommodating temporary emigration in spatial distance sampling models. Journal of Applied Ecology 55, 1456–1464. https://doi.org/10.1111/1365-2664.13053*

The authors go back and forth between the concepts of abundances and trends (and sometimes distribution) throughout the manuscript, leading to confusion on the part of the reader as to which specific question is being addressed by which part. Among other things, this led me to wonder about the differences between the EPOC-ODF program and the FBBS (that is quickly mentioned in the method part because used to extrapolate current abundances from ArGeom). I think that a sentence explaining all this might help the reader to understand why this approach is not directly compared in the article with the FBBS results.

- We fully agree with your comment. As the main aim of our study was to compare population size estimates obtained from data collected during the period of the Breeding Bird Atlas work (*ArGeom* and HDS), we solely used the species trends resulting from FBBS data to provide comparable population sizes timewise. To tackle the confusion in the main text, we removed the prior comparison between the two at the end of the introduction and sub-sectioned the results to differentiate them. **(L311 - 314; 316-323)**

The, I was a bit confused as for why the use of covariates is only mentioned from line 156 onwards. The latter seem important in the approach considered, and perhaps their use and what it implies should be mentioned earlier, particularly in terms of the precision they bring, or not, to detection modeling. In addition, their influence may deserve to be discussed in the discussion part: to what extent do these choices influence the estimates?

- To assess these comments, we include a paragraph in a new section of the discussion, “Study limitations”, about the consequences of our choice, such as the assumption that species responses follow a linear pattern towards the studied gradients, and provide interesting alternatives tackling this issue. **(L483 - 528), specifically on that matter : L497 - 507)**

Finally, I believe results could be better organized and benefit from subsectioning. There are a lot of models and methods, and it was difficult for me to know what results were linked to what method. The result part is short, maybe some results about detection probabilities, and notably the different covariates relations could be added to then enrich the discussion. I also was wondering why the IC for detection probability is so small on Figure 5 and it means.

- Thank you for pointing out this major issue. After a careful revision of our R code, we realised that we did not properly scale the value prior to the PGLMM model fitting. After applying this correction, we updated IC and the effect of covariate over δ_{mean} responses on (now figure 7.A-C) and in the results section. **(L352 - 356)**

Specific comments:

I suggest rephrasing the subtitles to make it more explicit to help make reading easier (L156, 184, 220, 248, 276).

- [A1] We reformulated the highlighted subtitle as suggested.

  - Bird species selection and environmental covariates -> PCA reduction of environmental covariates

  - Modelling approach -> Modelling framework

  - Population estimation and extrapolation assessment -> Trimming of HDS population size estimate: assessment of model extrapolation

  - Estimates comparison -> Comparison of ArGeom and HDS estimated population sizes

  - Phylogenetic model -> Study of variation of estimated population sizes between the two approaches

L67-69: Consider rephrase the sentence which as it stands is too vague. The second half is not entirely clear. Additionally, the link between "agricultural and planning policies" and bird abundances/trends has not been explained before.

- [A2] We now give an example earlier in the introduction, using one previously sourced article to better highlight the link between wild bird population declines and agricultural practices. **(L60-62)**

L 106: What concrete criterion does "Medium" refer to?

- [A3] Quality criterion of estimated population size for the prior French breeding bird atlas are based upon work from the CEPO (Committee for the Estimation of Bird population sizes) and particularly from the work of Comolet-Tirman et al., 2015. Quality assessment is composed of 4 levels; “none”: lack of information; “bad”: assessed poor state of knowledge and lack of semi-quantitative data; “medium”: knowledge is more

or less good, but semi-quantitative data can be incomplete or outdated; and “good”: reliable updated semi-quantitative data available.

- The lack of semi-quantitative data is linked to a lack of sampled 10x10 km grids regarding species known distribution in France. We added more information in the introduction. **(L105-110)**

Comolet-Tirman, J., Siblet, J.-P., Witté, I., Cadiou, B., Czajkowski, M.-A., Deceuninck, B., Jiguet, F., Landry, P., Quaintenne, G., Roché, J.-E., Sarasa, M., Touroult, J., 2015. STATUTS ET TENDANCES DES POPULATIONS D’OISEAUX NICHEURS DE FRANCE Bilan simplifié du premier rapportage national au titre de la Directive Oiseaux. Alauda 83.

L 125: quality of which aspect of inferences?

- [A4] In this part, we meant the inferences of the ecological process (abundance variation related to habitat cover) as well as those of the observation process (reliance upon replicated counts and distance sampling).

- We rephrased this part making it more concise while taking account of your next suggestion while removing mention of the FBBS in this part. **(L127 - 128)**

L 124: The objectives could be rephrased and further detailed in separate sentences. I wonder if it might be helpful to flip the sentence, starting with the objective "we propose an estimation method..." and then coming to the methods.

- [A5] Thank you for the suggestion, we have flipped the sentence to make it more concise and accounted for comments from reviewer #2, see [A35]. **(L127 - 138)**

L 312: Breeding bird populations abundances and/or trends? I propose to clarify this point throughout the whole manuscript. I feel that the authors go back and forth between these notions, which can sometimes be confusing.

- [A6] As FBBS aims to track breeding birds population trends, EPOC-ODF aims to estimate their population status. They both rely on volunteer birders, but EPOC-ODF is specifically dedicated toward the production of a new Breeding Bird Atlas. The present work represents the first iteration of the EPOC-ODF sampling scheme and its potential future applications. We are currently working on complementary methods such as IM (Integrated Models) that could allow the joint use of such data for the upcoming new atlas or novel monitoring schemes.

L142: « encountered » visually and/or singing?

- [A7] Yes we considered both types of detection, we clarified that sentence in the main text. **(L142 - 143)**

L145: What is the surface of the square of the grid? Does it correspond to the point counts or is it used to set a round buffer?

- [A8] We reformulated this part to clarify. The sampling location corresponds to the centroid of the 2x2km grid; we integrated this information in the main text. **(L145 - 151)**

L145: Perhaps add some information on why this choice of 5min, in light of the literature on the subject (5min sufficient for all species?).

- [A9] We added more information on the matters, the main reason being the nature of sampling schemes based solely on benevolent birders without financial compensation. See [A8] **(L146 - 148)**

L162: To this stage, we don't know what the covariates were chosen for? Is this to model p or N. I feel this whole section is a bit confusing.

- [A10] We included the figure (now fig 2) depicting covariate usage in the main text, as proposed by one of your next suggestions.

L 166: I didn't understand this sentence when first reading the manuscript, only later when reading the part on modelling. I wonder if it could be rephrased somehow.

- [A11] In this sentence, we explained the variable use of the Habitat cover extracted from OSO rasters. We used two buffer radii (100m and 500m around observers locations). The first buffer aims to define direct habitat composition that could hinder species detectability, and the second buffer is used to define species habitat. For better clarity, we added the figure (fig 2), as suggested in another comment see [A15]

L168: Why chose to group water bodies and mineral surfaces? What ecological meaning justify this choice?

- [A12] The primary goal of such a grouping is related to the main objective of our study. As we aimed at estimating species abundance across a large spatial extent using a hierarchical model, we relied on a PCA reduction to maximise environmental information while reducing the number of used environmental covariates (Tredennick et al., 2021).
- Due to the total habitat covers of our prediction extent, both of these covariates were underrepresented. As we also used bioclimatic PCA axes in the abundance state of the hierarchical models, we estimated species abundance according to the main collected/sampled habitats. We thought that a combination of both these information (habitat and bioclimatic) could better define these habitats, i.e., low variation of temperature and precipitation to water bodies, in contrast to mineral surfaces.
- Another reason for this choice is linked to species specifically targeted by the sampling scheme, as species linked to this habitat type are already targeted by other, complementary national sampling schemes (Quaintenne et al., 2020) or by institutions such as Wetland International.

- *Quaintenne, G., Gaudard, C., Béchet, A., Benmergui, M., Boutteaux, J.-J., Cadiou, B., Camberlein, P., Chapalain, F., Croset, F., Culioli, J.-M., Dalloyau, S., Debout, G., Dubois, P., Dulac, P., Flitti, A., Gallien, F., Gendre, N., Girard, O., Havet, S., Vincent-Martin, N., 2020. Les oiseaux nicheurs rares et menacés en France en 2016 et 2017. Rare and endangered breeding bird survey in France in 2016-2017 Ornithos 27-2, 73–111.*

- *Tredennick, A.T., Hooker, G., Ellner, S.P., Adler, P.B., 2021. A practical guide to selecting models for exploration, inference, and prediction in ecology. Ecology 102, e03336. https://doi.org/10.1002/ecy.3336*

L172 and S3: It seems to me that Axis 1 is probably highly correlated with elevation, and maybe moisture as well. In a general way, I feel that covariate choices are not discussed enough. Could you discuss to what extent these choices influence results? What are the ecological hypotheses behind these choices? Perhaps you could discuss whether some bioclimatic covariates (wind?) could also be used to model detection?

- [A13] We dedicated a new paragraph in the discussion about this issue thanks to your suggestion **(L533-543)**

- Early analyses performed originally included meteorological data such as information of rain, wind or cloud cover during the visit on the species detectability. The hypothesis was that bad meteorological conditions could negatively affect species detectability, but we faced two issues: the first one was the data availability, so that we relied upon coarse meteorological data (~10x10km grids) and low number of timestamps during the day: and the second one was that observers simply tended to avoid bad weather conditions to perform the field work.

Fig S3.1: This figure is extremely useful. Consider adding it to the main text?

- [A14] We included the figure in the main text (now fig 2).

L 174: Why 500m buffer radii? Scales and sampled surfaces are quite confusing in a general way. Maybe a figure illustrating/summarizing this could help?

- [A15] We depicted buffers radii in figure 2, previously figure S3.1

- We used 500m radii to depict habitat covers sampled by observers, as a standard measure of landscape-scale habitat availability for birds in the landscape ecology literature. An increase in buffer radius might lead to the inclusion of large proportions of non-used habitat and add noise rather than information to the modelling process.

L 180-182: I'm not entirely sure what is done here and how it is linked with previous paragraphs.

- [A16] In this paragraph, we explain the methodology used for the acquisition of PCA values at the site-level. **(L180 - 183)**

- We first computed a PCA over the prediction grid corresponding to the global geographical range over which we aimed at predicting abundance. We used results from this first PCA to determine PCA axes values at the sampled locations according to their extracted habitat cover.

L189: Why 5 bin classes?

- [A17] The number of classes is arbitrary, it was shown in a recent study, on simulated and case studies, that the number of distance classes does not affect estimates of bird abundance (Neubauer et Sikora, 2020).

- *Neubauer, G., Sikora, A., 2020. Abundance estimation from point counts when replication is spatially intensive but temporally limited: comparing binomial N-mixture and hierarchical distance sampling models 97, 18.*

L191: I'm not sure if Julian date and hour are the effort covariates. I'm not quite sure which parameters they are incorporated into. Is it detection probability or also abundances?

- [A18] We added the table depicting the sub-model approach to show which covariates were linked to which states of the model. See Table 1 **(L219)**

L200-204: It might be helpful to consider cutting this sentence. Perhaps we could have one with the general case and then a second for the exceptions.

- [A19] Thank you for pointing this out, we cut the sentence and slightly reformulate the second half about the exceptions. **(L223 - 227)**

L213-218: Could you please clarify what these data from 2022 are and why they were not included in the whole analysis? It might be helpful to rephrase this section to make it more explicit.

- [A20] We changed this section according to the suggestion from reviewer #2 (see [A42]). We rephrase this part to inform readers that we assessed robustness of population size estimates to the exclusion of one year of data for calibration. **(L239 - 245)**

L 272: Could you explain how you took account for it?

- [A21] To account for the *ArGeom* uncertainty, we analysed the variation of the percentage of difference between *ArGeom* and HDS, used as a fixed covariate of the phylogenetic model. We used the decimal logarithm of the difference (*ArgeomHigh – ArGeomLow*) instead of raw count, to take account of the different magnitude in population size across studied species.

- We removed this sentence from this section to reformulate the information and include it later in the ms. **(L304-306)**

L 278-288: This section is quite complex to follow. It would be helpful to have more clarity on how you implemented this in your models.

- [A22] We reformulated this section accordingly. **(L297-309)**

L291: Could you please clarify what 14.84 refers to?

- [A23] We added the missing +/- symbol. **(L312)**

Figure 4: Point (2) is not very clear, what does extrapolation sign refer to?

- [A24] We clarified this sentence, which refers to NT1 extrapolation (model prediction outside of the environmental domain collected by the sampling) **(L329-332)**

L 356-358: Perhaps it would be helpful to be more explicit about what your hypothesis is about where these differences come from? If not species detection probabilities.

- [A25] We changed the formulation of the beginning of the discussion according to suggestion from reviewer #2. **(L381-383)**

- We rephrase this section to present a clearer link between our hypothesis and our results. It is followed by two sections giving more details about (i) effects of the detection process **(L384-392)** and (ii) in depth context of data acquisition from the ArGeom method. **(L392-404)**

L365: Did the previous method also use covariates?

- [A26] The previous method did not use any covariate, since breeding bird pairs were simply estimated over a 10x10km grid. **(L395-400)**

- Population size estimates corresponded to an extrapolation from the number of grids prospected (1953 out of 5879 10x10 grids) to the number of grids where species were considered breeding. Moreover, the breeding classification of 10x10km grids did not rely on semi-quantitative methods, but on opportunistic data using breeding status categorical codes.

- *Issa, N., Muller, Y., 2015. Atlas des oiseaux de France métropolitaine: Nidification et présence hivernale, Illustrated édition. ed. DELACHAUX, Paris.*

L362-368: This is an interesting point, but it might be clearer if it were rephrased slightly. In particular, it would be helpful to have more information about the effect of time and date on detection probabilities in the results section. It's not clear to me to what extent you're modelling the phenology of detection during spring.

- [A27] We now clarify the description of our modelling approach by adding a table and afore-mentioned figure S3 (now table 1 and figure 2). For bird phenology, we tested diverse effects of time and date, as well as their quadratic effects, on species probabilities of being exposed to sampling during visits, i.e. being considered available. For species detectability, we used covariates such as distance to roads and proportion of habitat near observer locations to seek if this could affect observers during surveying.

L368-376: This paragraph could be in the introduction instead, because it presents the « old » method?

- [A28] We aimed at giving more context about the *ArGeom* approach and directly link it to results from the $\delta_{????}$(previously delta_methods, but changed according to suggestion from reviewer #2)

- We rephrased the section giving more context details about data acquisition and its potential effect over the positive effect of *ArGeom* uncertainty to $\delta_{????}$. **(L393-404)**

L381: I'm not quite sure what this means, but it sounds interesting. Perhaps you could try rephrasing it?

- [A29] We reworded the paragraph containing this sentence to elaborate more on that issue. **(L405-417; specifically L409-417)**

L 398: « Actually, community … » instead?

- [A30] Thank you for pointing that out, we modified this part accordingly. **(L432)**

L421: This seems like an important point. It's not entirely clear how many species are concerned by this in your study. Could you please elaborate on how this affects your results?

- [A31] In our study, we estimated population size for species that are not social during the breeding season. We decided to remove social species from the available pool of species (i.e. considered target of the sampling scheme), as the modelling of species occurring in flocks could differ from non-flocking species (Schmidt and Rattendury, 2018) **(L486 - 489)**

- These social species include 6 species, such as the Common starling (*Sturnus vulgaris*), Western jackdaw (*Corvus monedula*) or the Western house martin (*Delichon urbicum*).

- *Schmidt, J.H., Rattenbury, K.L., 2018. An open-population distance sampling framework for assessing population dynamics in group-dwelling species. Methods in Ecology and Evolution 9, 936–945. https://doi.org/10.1111/2041-210X.12932*

Reviewer #2

The paper compares estimates of abundance of common birds across France using two different atlas data sets. The first data set is a survey from 2012 for which abundance estimates were derived without statistical modelling. These estimates are then compared to estimates from a new atlas scheme for which the authors suggest hierarchical distance sampling (HDS) models to estimate abundance.

Considerable effort and thought have been put into the modelling process for the more recent data, with seemingly well considered choices regarding which covariates enter the different components of the HDS model.

The new atlas survey and how the hierarchical distance sampling model is used to estimate abundance is described in detail (although descriptions could sometimes be clearer, see below). However, given that the comparison between estimates from the previous survey and the new survey is the main focus, I'm missing sufficient detail about the previous scheme, especially as the main reference provided for it is in French. What was the statistical design of the previous survey, how were counts conducted, estimates derived etc? Estimates from the old survey are also described as expert based, but as currently described in the text it is just a quantitative computation from "measured abundance", with no expert knowledge used in the process.

- We would like to thank the reviewer for pointing out this important issue. We now provide more in depth information than previously in the Methods and Discussion sections. Previous population size estimates were obtained through extrapolation from a subset of 10x10 km grid sampled (corresponding to a third of the total area) to a 10x10 km grid containing cues of breeding per each bird species (probable and confirmed breeding evidence). The sampling was, later on, qualified as uneven (Issa and Muller, 2015; Roché et al., 2013) with variable participation, resulting in the reliance of expert-based opinion in the number of breeding pairs (prior to the extrapolation). See answer [A3] **(L105-110 & L395 - 404)**.

Issa, N., Muller, Y., 2015. Atlas des oiseaux de France métropolitaine: Nidification et présence hivernale, Illustrated édition. ed. DELACHAUX, Paris.

Roché, J.-E., Muller, Y., Siblet, J.-P., 2013. Une méthode simple pour estimer les populations d'oiseaux communs nicheurs en France. Alauda 81, 241–268.

My main concern is otherwise that the authors claim without evidence that their estimates from the new scheme are better than those of the old one. For example, if estimates from the old survey are lower than those from the new survey, they are referred to as underestimating abundance (and vice versa). The assumption is that the modelling provides more accurate inference than the previous ad-hoc approach. This may seem reasonable, especially since the modelling is largely based on sound reasoning. But the fact is that since we don't know the true abundances we do not known which estimates are closest to truth. A more nuanced discussion of the differences, not taking for granted that the HDS modelling will automatically provide better estimates is therefore necessary.

- We fully agree with these comments, and therefore nuanced our former wording throughout the main text as well as the supplementary materials. We also added more details on model assumptions and limitations in a new section of the discussion "Study limitations" **(L483 - 528)**

One concern, for instance, is that the N-mixture model used as one component of the HDS is not a very robust approach because essential information to estimate detection (availability in the HDS model) is missing (Barker et al. 2018). The N-mixture model can underestimate or overestimate abundance, it is not necessarily unbiased (e.g. Duarte 2018).

- We also agree and talk in more detail about N-mixture model bias due to unaccounted variation in species availability. We discuss how the use of distance data can allow to alleviate some of the issue arises (Chandler et al., 2011), can still induced biassed estimations (Mizel et al., 2018). **(L520 - 528)**
- *- Chandler, R.B., Royle, J.A., King, D.I., 2011. Inference about density and temporary emigration in unmarked populations. Ecology 92, 1429–1435. https://doi.org/10.1890/10-2433.1*
- *- Mizel, J.D., Schmidt, J.H., Lindberg, M.S., 2018. Accommodating temporary emigration in spatial distance sampling models. Journal of Applied Ecology 55, 1456–1464. https://doi.org/10.1111/1365-2664.13053*

In addition to the above, I would suggest the authors to take another careful pass with the text. There is missing text in some places, new paragraphs where they are not needed etc. The Methods section could be improved for better clarity, and the Discussion better structured and more focused on the central questions.

- We rewrote the Methods section, adding figures and tables previously presented as supplementary material in the main text. We also changed Methods sub-section names, according to comments from reviewer #1, and also added a new subsection to the Discussion.
- Please refer to [A1] for the subsections names changes of the Methods. We added subsections to the results and discussion parts.
    - Results :
        - Species trends over 2012-2023
        - HDS population size estimations
        - Population size comparison between ArGeom and HDS
    - Discussion
        - Potential consequences for community-level assessments
        - Conservation implications
        - Comparison to other European countries
            - (New section) accounting for editor's suggestion
        - Study limitations
            - (New section) accounting for reviewers #1 and #2 general comments

Barker, R. J., Schofield, M. R., Link, W. A., & Sauer, J. R. (2018). On the reliability of N-mixture models for count data. Biometrics, 74(1), 369–377. https://doi.org/10.1111/biom.12734

Duarte, A., Adams, M. J., & Peterson, J. T. (2018). Fitting N-mixture models to count data with unmodeled heterogeneity: Bias, diagnostics, and alternative approaches. Ecological Modelling, 374, 51–59. https://doi.org/10.1016/j.ecolmodel.2018.02.007

Detailed comments ;

Line 49-52. This is an example of a sentence that need to be more carefully worded. You have not shown that your estimates are unbiased.

- [A32] We changed our wording accordingly, and also accounted for your general comment to nuance the comparison between ArGeom and HDS population size estimates throughout the ms. **(L43-46)**

L99. In Europe there are monitoring schemes specifically targeting common species though, it would be a bit of a stretch to say that they are neglected or overlooked.

- [A33] We agree and nuanced our views on this part, we meant that there were generally no fundings specifically dedicated to the monitoring of common species other than the general Bird Breeding Survey protocol. **(L98-101)**

L112-115. Should be the other way around? Geometric means are smaller than arithmetic means.

- [A34] Thank you for pointing that out, we actually inverted the two parts of the sentence. **(L114 - 117)**

L130-130. There is in fact no test of whether the new data set provides estimate closer to the truth.

- [A35] We agree and consequently removed the sentence. **(L131 - 133)**

L174-175. Revise wording.

- [A36] We rephrased this part and added a figure (fig 2) from the supplementary in the main text. **(L168 - 171)**

L187-188. Not quite clear what is meant here. Do you mean that you truncated distances above the 95% quantile to the 95% quantile?

- [A37] We clarified this part to make it more understandable. **(L195-196)**

L185-204. I would not be able to repeat the modelling process or model selection strategy from these explanations. Please try to revise the method description to improve clarity.

- [A38] We modified the method description as suggested and moved the figures and tables from appendix S5 et S3 to better explain the secondary candidate set approach from Morin et al., 2020 (fig 3), as well as, the use of covariates in each state of the model (fig 2 and table 1)

Morin, D.J., Yackulic, C.B., Diffendorfer, J.E., Lesmeister, D.B., Nielsen, C.K., Reid, J., Schauber, E.M., 2020. Is your ad hoc model selection strategy affecting your multimodel inference? Ecosphere 11, e02997. https://doi.org/10.1002/ecs2.2997

L185- Was the year of survey included in model somehow? Why, why not?

- [A39] We did not formerly include the year of the survey because of the sampling design, as once a location is prospected (3 visits during the breeding season with 3 completion of 5-min point count for each visits) it is considered complete and removed from the pool of available locations to sample for the next year. Considering this sampling design, there is no inter-year variation at the site level (in contrast to a robust design).
- We primarily aimed at estimating the superpopulation (i.e. quantity of individuals over the three years) rather than accounting for temporal variations across years.
- We did not account for the effect of population dynamics on the number of individuals or on species detection probabilities, as we thought it would make the model too complex. As the number of years surveyed are not long enough to soundly estimate a trend (White 2018), we would have to resort to integrating inter-year variation as a

categorical covariate of three levels (2021/2022/2023) potentially increasing the risk of model non convergence.

White, E.R., 2019. Minimum Time Required to Detect Population Trends: The Need for Long-Term Monitoring Programs. BioScience 69, 40–46. https://doi.org/10.1093/biosci/biy144

L205-208. Not clear how C-hat was defined or calculated.

- [A40] We gave more details for C-hat computation. **(L228 - 232)**

L213. Clarify that you are assessing robustness of estimates to exclusion of one year of data (rather than general robustness)

- [A41] We added a sentence to clarify this point. **(L239 - 240)**

L216-218. How do you draw the conclusion that estimates are robust to exclusion of one year when confidence intervals for 9 out of 30 species don't overlap?

- [A42] We changed this part since, in the previous version we assessed if the mean population size estimated from 2021-2023 data was in the upper/lower limit of the estimated population size of the 2021-2022 dataset. We changed the text, code and table (S2.2) to reflect overlap and non-overlap between upper and lower CI between population size estimated from the two datasets.
- This results in the following changes in the main text.
- Out of the 7 species whose population size estimates do not overlap, only three Eurasian Blackcap, Common Chaffinch and Common Cuckoo displayed important changes between the two estimates. The other four, *i.e.*, Blackbird, Short-toed Treecreper, Song Thrush and Eurasian Nuthatch showed only slights changes between the two estimates, see table S2.2 for population size estimates

L233-234. Define "coefficient of variation of the range uncertainty between pre- and post-treatment estimates".

- [A43] We elaborated on this part while taking account of the change suggested from your comment see [A45]. **(L258-259)**

L233-234. What about NT2 extrapolation?

- [A44] We did not account for this type of extrapolation in the trimming procedure, as it was negligible for the majority of species, for instance for the Eurasian Hoopoe in fig 5.B, it corresponds to 1.2% of the total predicted area.

L234. The pre- and post-treatment labels do not accurately convey what is done. Something like 'outlier-trimmed' and 'untrimmed' seems more appropriate.

- [A45] We took account of your comment and changed the wording across the main text, supplementary materials and code. **(L258-261)**

L239. Which "comparison analysis"?

- [A46] We ought to mention the comparison between *ArGeom* and HDS population size estimates, we changed the sentence to be clearer. **(L263-245)**

L249. → "comparable estimates between the old and the new survey, we restricted…"

- [A47] We changed the wording. **(L273)**

L251-256. Remind the reader here that ArGeom estimates the number of breeding pairs.

- [A48] We added the information. **(L302)**

L255-256. This could lead to errors though, since lack of sexual dimorphisms does not imply that males and females are equally likely to be detected.

- [A49] Yes, we implied a sex ratio of 1:1 when we applied this ad-hoc filter, we nuanced this bias in the discussion where we added a section about Integrated Population Models (IPM) taking account of the population structure (sex ratio and age structure). **(L526-532)**

L260. As the conclusion is that ArGeom provides lower estimates, it might be of interest to compare estimates using the upper bound in addition to the midpoint (just a suggestion).

- [A50] We added a supplementary material comparing estimates using upper bounds, we see the same response pattern of $\delta_{????}$across both covariates. See supplementary S7.

L268. The notation “delta_methods” is somewhat unfortunate as the delta-method is a standard statistical approach not related to the use here.

- [A51] We changed the notation to $\delta_{????}$in the main text, figures and code

L272-274. Not clear in what way you “took account of ArGeom uncertainty”.

- [A52] We revised our wording for this part. We implied that (i) we accounted for the uncertainty as a fixed variable in the study of differences between estimates of the two approaches and (ii) as we compared multiple species with variable magnitudes of population sizes, we applied the decimal logarithm of the range (*ArGeomHigh – ArGeomLow*). **(L304-306)**

L284-285. Revise wording in “using weighted means final candidate sets models in regards to AICc scores”.

- [A53] We changed the wording. **(L303-304)**

L283-288. This analysis does not account for uncertainty in delta_methods. i.e. error in the estimate of delta is not accounted for.

- [A54] In the PGLMM, we included a random effect taking account of species phylogenetic relatedness, as we supposed that related species could have similar (i) relations to habitats determining their abundance and (ii) behaviors and morphometric characteristics affecting their availability and detectability. We also incorporated response weights to take account of the uncertainty extracted from FBBS (French Breeding Bird Survey) species trends, given more importance to $\delta_{????}$of species with certain species trends estimated.

L297. I suggest providing estimates of average estimated availability and detection probability in an appendix. This would be useful for understanding to what extent the N-mixture part inflates abundance, for example.

- [A55] We provided the averaged estimated availability in table S6.1. We did not provide the detection probability (intercept of the sigma parameter) but the maximal observation distance instead, after 95% truncation, to reflect better species

detectability. This table was primarily created to give a quick assessment to field ornithologists participating in the scheme.

- We specifically created a new supplementary table to display species parameters in HDS model (table S6.3)

L308. “sits” → “its”.

- [A56] Typo corrected. **(L328)**

L317 and elsewhere. Avoid qualifiers like “under” and “overestimation” and use something neutral like “estimated lower compared to HDS”.

- [A57] We took account of these comments and changed our wording throughout the main text and supplementary materials.

L318. Why is habitat specialist/generalist a relevant variable for the difference between the two approaches? Was this based on a formal analysis?

- [A58] We classified species as specialist/generalist to generalised $\delta_{????}$variation on the functional level. We supposed that specialist species use less habitat, and are more prone to have localised ranges (sensu Rabinowitz 1981). In the discussion section, we linked species specialist characteristics to misrepresentation of used habitat, potentially leading to biassed estimation from approaches like *ArGeom* that do not account either for the detection process or for habitat covariates.

L350-352. Here you are assuming that the HDS estimates are correctly representing truth.

- [A59] We rephrased this part according to your general comments. **(L373-375)**

L352. “presumed known uncertainties ranges” ?

- [A60] We referred to *ArGeom* uncertainty, we revised our wording throughout. **(L375-376)**

L359-362. You found no association between detection probabilities and delta_methods, but still draw the conclusion that detection causes the difference? This needs further elaboration.

- [A61] It is correct that we did not find any associations between detection probabilities and $\delta_{????}$in our studies, but in the conclusion we attempted to expand from the estimated parameter of the availability state of the model. We concluded that these differences may not be due specifically to species availability but may be caused by the overall modelling framework accounting for the observation process and inferring abundances based on species-habitats relations.**(L381-383)**

L369. “deviating from expert opinion reliance” - do you mean “derived from export opinion”?

- [A62] Indeed, we meant “derived from expert opinion”, thank you for pointing that out. We took account of it in the rephrasing of the paragraph suggested in your next comment. **(L395-400)**

L370-372. Difficult sentence.

- [A63] We attempted to rephrase this whole paragraph giving more context information regarding the *ArGeom* approach from French sources, see [A26]. **(L423-434)**

L377-384. I had a hard time following the argument in this paragraph. Consider rephrasing.

- [A64] We rephrased the paragraph giving more details over the second part. **(L405-417)**

L407-414. If conservation status was an important question, why is it not mentioned in the methods/results?

- [A65] As for the specialist/generalist dichotomy, we aimed to expand the result toward a conservation issue. We presented conservation status in the appendix presenting results S6.1, but we did not formally test for its effect on modelled estimates. **(L438-458)**

L421. What does “inferences of clustered individuals” mean?

- [A65] We meant species with gregarious behaviour, such as the Common Starling, that required specific consideration. As the number of individuals could directly affect species detectability. We reword the paragraph according to a suggested comments from reviewer #1, see [A31]. **(L486 - 489)**

Fig 5B-C. Explain what the figure shows. “Marginal” can mean many different things in a statistical context. I think the figure shows the predicted response across different values of detection probabilities while the other covariate is held constant (perhaps at its mean?)?

- [A66] The figure 5B-C (now 7B-C) actually shows model predictions of $\delta_{????}$ according to the gradient of species detection probabilities (7B) and *ArGeom* uncertainty (7C) while other covariates are averaged. We rephrased the figure legend in this regard. **(L364-367)**