**Handling editor: Holger Kreft**

The reviewers raise a number of points that you need to address before we can consider publication. Most of their comments circle around clarifying concepts, improving parts of the writing, and finding a more balanced discussion. Reviewer #2 is concerned about the reproducibility of the study - a point I also think is extremely important.

**Reviewer #1**  
The authors made a big piece of work with this review, but I would like to encourage them to clarify some points and the presentation of the results and to go further in their recommendations. I do not think this last objective (L42) is totally fulfilled in the manuscript.  
  
First, the context and explanations about temporal scaling could be more developed. The theory of temporal scale effects is mentioned without explaining the implications when calculating trends beyond the good practices of providing the grain (L274-276). What are the implications of having a different temporal grain from the sampling to the metric? What are the implications of 2.5 min versus 10min? The selection of the grain of the sampling depends on what is measured (e.g. Bonthoux & Balent 2012 Journal of Ornithology https://doi.org/10.1007/s10336-011-0766-2). Should it be discussed? Here, you focused on diversity trends based on occurrences, but among these studies, many measured their trends using data sampled to assess the relative abundances of birds.  
  
Also, I do not understand how you extracted or measured the temporal grain of the studies. I think I would not be able to reproduce the calculations from the description and examples in the methods section.  
In the example of Wretenberg et al. (2010) you provided L279-280, when I read the paper, I think that I would say that the temporal grain is provided: 10 min. In Monnet et al., even if the change in protocol from 5 to 10 min can make the calculations difficult, I have no idea how you got to 0.09 hour as a temporal grain.

>>> The temporal grains in our manuscript are provided in decimal hours, which was not specified in the original version (but is now corrected). They are given for the smallest unit of area at which the biodiversity metric is computed. For instance, concerning Monnet *et al.* (2014), the indicated temporal grain is 5 minutes: 5/60 = 0.08333 rounded up to 0.09 h. Other example: articles using the North American Breeding Birds Survey are summarizing the data at the scale of the routes. Each route is divided into 50 census points surveyed for 3 minutes each. Thus: (3\*50)/60 = 2.5 h.

Finally, concerning Wretenberg *et al.* (2010), the only information about the temporal grain we found was: *“Birds were censused with point counts (Bibby et al., 1992). All sites were visited six times in 1994 and five times in 2004 during early morning (mainly from sunrise to 10 am), once in each of the periods 15–30 April (only 1994), 1–10 May, 11–20 May, 21–31 May, 1–10 June and 11–20 June.”* (part 2.2). It would be helpful, however, if the reviewer indicates where this information (*i.e.* temporal grain = 10 min) is provided.

Overall, despite the examples and the figure, I didn't find the explanation of the temporal grain of the metric very clear.

>>> Clarified line XXX as follow: *“[…] i.e.* the duration of a single census session at which biodiversity is assessed […]*”*. Also, the lack of clarity and consensus about the concept of temporal grain in the scientific literature is one point of our manuscript.

L289 What if the authors keep the maximum between the two samples?

>>> In this case, authors should keep the original 5 minutes temporal grain as they use the data from only one census

L291: Fig 1b?

>>> Corrected

L363 "this need to be addressed' I think you should go a step further and provide clear guidelines regarding the good practices on reporting the temporal grain of the metrics and of the sampling plan, with examples of good practices from the literature if they exist.

>>> “[…] *and this needs to be addressed.”* is a reference to the last paragraph of the part “*Issues of temporal grain*” in which we give guidelines about specification of temporal grains as follow:

“*If one wants to study the temporal scaling of biodiversity trends, a clear assessment of the temporal grain needs to be done systematically. That is: if a biodiversity metric is computed by combining samples together, the temporal grain (i.e. time span) of samples is summed, and this sum should be considered as the final temporal grain of the metric (i.e. Fig. 1, sum of the temporal grains of the black dots in the red boxes). Likewise, if the metric is averaged over several samples, the mean temporal grain of the samples should be specified (i.e. Fig. 1, mean of the temporal grains of the black dots in the red boxes). Finally, if the metric is predicted by a statistical model, the temporal grain of the model should be specified. These recommendations also apply to spatial grain.*”

For the reader to understand that we are referring to this part, we modified the sentence as follow: “[…] *and this needs to be addressed* *by clearly specifying the summed, averaged or modelled temporal grain of the metric”*.

2) I also have comments and suggestions to improve the figures:  
  
Figure 1b: I understand the definitions, but I found the figure did not help to understand them.  
I would rather represent the temporal grain as a zoom from a point in the timeline. Are the y-axis and the continuous variable needed in the figure?  
  
Figure 2: Even if the legend explains it, the varying size of birds used for aesthetic reasons is very confusing. I would suggest changing it. Are they too many European-wide studies to add them to the map?  
  
Figure 3: In the text, panel c in Figure 3 is used to compare only tBetadiv, sR and Fdiv between local, regional, and global scales. I would consider panel c by a barplot with only these three indices grouped by spatial extents. Also, I think this graph is not colour-blind friendly.  
  
3) The authors resume the trends into the three categories: increase, decrease, and stable. It must have been not easy in some papers. From there, could you also provide some recommendations of good practices when reporting results on temporal trends in scientific papers?  
  
4) Finally, a so appealing title and this big work of review call for many expectations that cannot be all adressed. I understand that some questions might have been discarded by the authors when delimitating their work of review. Still, I think some are in the scope of the question of spatio-temporal scaling and a couple of them should be at least mentioned or discussed if not addressed. Among them, I would like to know what the authors thought about the implications on the analyses of the following points:  
- The effect of diachronic versus continuous data to calculate trends  
- The implications of the trade-off between long extent (L351) and huge time lags  
- The implications of the different groups of birds that the authors may have considered in their studies (all species, common birds, some guilds…) to calculate the trends and compare them  
- Did you notice an effect of the studies' starting point on the trends (from the studies starting in the 1960s vs in the 2000s)?  
- Did the studies performed with the same data and same grain find the same trends?

>>> Indeed, we noticed that studies using the same data, temporal extent and spatial grain were displaying the same trends. That is, we decided to account for pseudo-replication and to only consider the contributions of the similar studies as one in Fig. 3. The similar studies can be found in the Table 1.

Line-by-line comment:  
I noticed throughout the manuscript many sentences that are not very informative:  
- L59-61 ("critical", "we know", "affects macroecological patterns", "great interest")

>>> Clarified  
- L71

>>> Reworded  
- L339 Which kind of study is Fraixedas'?

>>> Clarified  
  
L149 Does the threshold of 50 x50 km² discriminate well the categories local or regional, or are some studies close to this threshold?

>>> The higher and lower closest area to 50x50= 2,500 km² are 11,000 km² and 800 km² respectively. There is only one study for which area = 2,500 km²: Keller *et al.* (2020).

L228: The first sentence of the discussion seems to show that the main result of your review (i.e. negative trends on a global scale and none on smaller scales) is already known. You may need to explain the specificities of your contribution on this point compared to the existing literature.  
  
L305 on how many?  
  
L309-311 reference missing  
  
L317-320 and so, what did they bring?  
  
330-333 Explain what developments in EBBA and GBIF make "hope to see trends with spatial replicates at national grain soon"?  
  
L364 unclear - what are your recommendations in modelling?  
  
  
  
**Reviewer #2:**

This papers investigates the extent to which temporal trends of biodiversity measured in birds are affected by different aspects of the data used : their temporal resolution, location, and the specific metrics analyzed. The authors use a classical query to construct and analyse a corpus of academic references. They summarize the major sources of uncertainties and confusions, highlight the scale dependency for the direction of the trends, and indicate why and how space and time should be considered together to derive meaningful trends in community metrics in birds (of richness, structure and composition).  
  
The study has several strengths: it's well written and easy to read, the topic is timely and well explained in the introduction, the temporal coverage is ambitious, the outputs are useful for many other types of research in ecology and conservation. Most of the results and claims are well supported.  
  
I however have some critical concerns to express, needing adjustments. Some are already in the paper but do not receive an appropriate consideration.  
  
  
1) Defining biodiversity and related metrics. While there is a general agreement according to which "global biodiversity is undoubtedly decreasing" (L228), there is a debate in the scientific community regarding the nature of the biodiversity crisis and the need to rely on multiple metrics. Even the term "decline" is challenging to define when simultaneous metrics are considered. So overall what is meant by biodiversity? Is it only species richness? Is it all species-based metrics? Or all existing metrics? Giving a formal definition of the 12 retained metrics (lines 117-121) in a glossary or in a supplementary item would be a good start as well as a broader approach to define biodiversity.  
  
2) Global vs local scale. What is meant by "global biodiversity" also needs clarification. Some analyses consider local biodiversity but include it in meta-analyses whose objective is to characterise trends on a global scale [1,2]. Can a collection of local studies scale up to a global scale?

>>> This is a point which indeed needs to be more emphasized. We considered trend assessment as global when studies were assessing the trends for the entire globe, and not as a cumulation of several sparse locations. However, assessing biodiversity trends in this way is challenging, hence the very few numbers of global trends found in the scientific literature. The only manuscript claiming to assess global decline is Jarzyna and Jetz (2018) who used the work by Szabo *et al.* (2012). The latter used a set of extinctions lists in order to assess global biodiversity decline since 1500.

3) The selection process for the papers considered would be difficult to reproduce. What and why some papers are excluded would be useful in a supplementary material.  
How the 156 references are processed to select the 24 final papers? This is key to the understanding of the corpus and to potential future replication.  
  
4) The limits of community metrics. It's surprising that pure and simple abundance metrics (population size) are not considered. We understand the choice of focusing on species richness and related metrics. But by ignoring multi-species average trends in abundance, part of the story is missed. The trends in LPI type of metrics or those targeting birds in particular (Indicators produced by European or National Bird Census) are essential to capture biodiversity responses to global changes and missed by community metrics. This should be more emphasized.  
  
For instance, L228 again, "there is still no evidence of this negative trend at local and regional scales" is somewhat misleading. First, isn't it twisted to expect local trends to reflect the global ones? Indeed, I do not expect the global trend to "propagate" locally but rather the opposite. Second, the query adopted in this paper does not capture the local evidences of loss in habitats and populations or species massively documented. And this is a methodological decision, not a lack of evidences. I take one example at random following a rapid search on google scholar about individual species extinctions: Ford et al. 2009. Extinction debt or habitat change? - Ongoing losses of woodland birds in north-eastern New South Wales, Australia. Biological Conservation. This is a case (among tons) of evidences for a regional extinction of individual species. It's of course filtered out in the query conducted in the paper because it's not using species richness. But isn't it contributing to biodiversity dynamics? Isn't it a possible evidence that global changes have local consequences ? This is mentioned briefly L270-272. But too briefly and without the due attention.  
  
I suggest that this issue to be also proposed as a necessary angle to adopt when approaching biodiversity trends: not focusing on aggregated metrics, whatever the scale and resolution issues, but also on other aspects (population sizes, local extirpations of individual species, habitat destruction...). The authors seem aware of this, but I think it should be made clearer to the reader.

>>> Emphasize on the fact that species based metrics have their weaknesses

5) The discussion about the meaning and implications of any given trend needs more cautious. For instance, take a metric of beta-diversity reflecting homogenization. An increase means less diversity and is possibly interpreted as a deterioration from conservation perspectives. It can result from an increase in local species richness. So simply counting increases or decreases (as proposed in the section "overall trends" p 10) is hazardous. At least some kind of preferable direction for each metric should be agreed on I suppose. After all, 20 increases + 1 decrease is similar than 1 increase and 20 decreases if one ignores the increase/decrease of what? Why an increase "per se" should be better? I missed introduction/discussion about the lack of meaning of any trend if no objective or expectation is associated to the trend. Again a table with the definition and meaning of each metric would be useful.  
  
6) The study challenges the spatial/temporal resolution of the ~60 studies collected. But the resolution of this approach is also a potential problem. For instance, in figure 3c when the conclusion for metrics x grain rests on very few numbers ( 3 papers for 3 metrics at global scale), it's unclear what can really be concluded. One point is not really considered is that we still miss data for such a review. With 300 studies analyzed instead of 50, maybe the conclusions would be different? So in other words, could regional and local scale declines in biodiversity simply still be under sampled? I think that concluding anything general from Fig. 3c is really risky.  
  
Other comments:  
  
L105-113. Another justification to work on birds is that they cover a large spectrum of strategies, diet, habitat etc… They are not all similarly responsive to a given disturbance (natural or anthropogenic)…So this variability offers the possibility to formulate and test specific predictions.  
  
L131. With these replications…You mean with Keller compared to Barnagaud ? "More general and reliable"… than what scenario? Please explain to avoid confusion.  
  
Fig. 2. Why 24 studies ? I cannot find this number above (Why not 20 as in L180 ? I am confused)

>>> Corrected

L251. This is really hazardous to equate the "quality of biodiversity" to ecosystem services unless you want to endorse a hard anthropocentric and instrumental lens. Some services need very few species but many individuals (eg game species for food), others would need very specific functions but not superior any "quality"…So I would not use the ecosystem services concept here, it's really not needed. That losing beta-diversity is bad for the ecology of species and systems because it is a loss of potential interactions is I think a much stronger justification.  
  
L255. Is increasing functional diversity always wanted? Some habitats harbor poor functional diversity. Increasing "per se", the number of species or functions is meaningless from a conservation perspective. It is really context dependent I think. Is maximizing functional diversity everywhere a desirable goal? It would be meaningless and lead to maximum functional homogenization. That some ecosystems have less diversity (whatever the metric used) seem perfectly fine. What matters is their trajectory, the cause of the changes…I suggest more cautious is used when a drop/increase in diversity is taken as a problem/improvement per se.  
  
L259-269 and L335-352. Very good points. Some of this could I think come earlier and introduce the necessary nuances to interpret the results (see concern expressed above regarding the difficulty to interpret a local increase in species richness or function as a synonym of no biodiversity decline or as a good sign for the "quality" of biodiversity).  
  
L275. Explain what is the "species-time" relationship.  
  
L273. Issues of temporal grain and extent. It's unclear what introduce uncertainty (sampling error) and what can produce real and more problematic bias (estimating false negative/positive trends because of this discrepancy in temporal grain). Moreover, I felt that more attention should be given to the variability in temporal "extent". Is it better to rely on 2 time series of 10 years or 1 of 20? Most of the time, only the extent of the longest time series is displayed by a given study, even though it does not reflect the overall abundance and quality of the data. For instance, in Dornelas et al., 2014 ([1] below), not all time series extend from 1960 to 2000, only the longest ones. So what is the extent in this case? It might be interesting to consider the impact of length by looking at the time series one by one for example, or by considering the average length of the time series in a given study. Several solutions can be discussed, but I think it is important to raise this point as well, because otherwise the temporal extent of the data is overestimated.  
  
L329. Why national standardization per se is a problem here? Because it's different among countries? I guess standardization would be a good thing for comparing trends among countries, right?

>>> Our phrasing was maybe unclear: we meant that the standardization is made at the national scale (which indeed is a good thing) but that those standardization criteria are different for each nation, making the national datasets spatio-temporal features different and inter-national comparisons challenging. We proposed a better phrasing.

Fig. 2. Varying the size of the bird for aesthetic reason does not work. The norm is that different sizes in a symbol represent different sample sizes. Maybe use different shape of different species for aesthetic reason and given true information related to the spatial grain or size on that figure.  
  
[1] Dornelas M, Gotelli NJ, McGill B, Shimadzu H, Moyes F, Sievers C, et al. Assemblage Time Series Reveal Biodiversity Change but Not Systematic Loss. Science 2014;344:296-9. https://doi.org/10.1126/science.1248484.  
[2] Vellend M, Baeten L, Myers-Smith IH, Elmendorf SC, Beauséjour R, Brown CD, et al. Global meta-analysis reveals no net change in local-scale plant biodiversity over time. Proc Natl Acad Sci 2013;110:19456-9. https://doi.org/10.1073/pnas.1312779110.