Response to the associate editor and the referees regarding *Looking for compensation at multiple scales in a wetland bird community* by F. Barraquand et al., submitted to *J Anim Ecol*

Associate Editor Comments for Authors:

Dear authors,

Two reviewers and myself have read the MS in detail. The paper makes an interesting analysis about compensation in avian communities using abundance time series data. The taxonomic and functional group comparisons makes it particularly interesting, and it is good to see a study on vertebrates on this topic. Both reveiwers and myself think it is topic wise suitable for JAE. However, both reviewers are also very critical about a diversity of aspects, for example the analysis on functional groups (though for different reasons), and in addition reviewer 1 has many other concerns about the approach taken, the interpretation of results and how they are presented. My biggest worry is that one can perform sophisticated analysis on many long timeseries, but detecting compensation is going to be always very challenging using such data, without a good mechanistic understanding of the system and the biology of the species involved. In that respect I found it disconcerting to see that the authors did not make strong attempts to look more in detail into the species involved (see also comment by reviewer 2); at times the MS feels too remote from the biology and species involved. This leaves me with a rather difficult decision. Given the strong competition for publications space, and the fact that this MS needs a lot of work in various aspects (rewriting the Introduction, presentation of results, reanalysis and reinterpretaton of results), I recommend to reject this paper. However, I can see that there is a case to be made for allowing a resubmission as well, as the topic is of broad interest and both reviewers clearly see also very positive points, and thereby the potential of the MS.

Dear Associate Editor,

We thank you for these comments and for this opportunity to resubmit our work. We have now considerably edited the manuscript in response to reviewer feedback, as well as additional opportunities for improvement that we stumbled upon. Among the main changes:

- The Introduction has been shortened as requested. We keep a pedagogical exposition of the compensation concept, given that some readers of J Anim Ecol may not be familiar with it.
- The species list, as requested by yourself and both referees, is now provided in Appendix S1 with the frequency of occurrence and abundance of each species. A full justification for the groups of species studied is then given in the main text in a dedicated subsection of the Methods. We carefully checked the taxonomy as well, and updated accordingly (some species changed genus, the previous *Anas* duck genus is now the *Anatini* tribe, following Gonzalez *et al.* (2009) - this does not change any analysis, and the *Anatini* group is of comparable size to the *Calidris* genus, both having 7 species).
- We previously used the “Duck” functional group as a shorthand for “Ducks, geese and swans”, as most observed species in this guild are ducks. Although this was clearly stated in the manuscript, and easy to catch for non-birders, we feared that it might not be precise enough for ornithologists. Hence, we now describe this group as “Waterfowl”, although it also includes one species of coot, which is often considered not part of waterfowl. This is duly explained l. 133-149.

- We revised and simplified Fig. 2 and Fig. 5, which were misleading, as noted by the referees. We hope that the revised versions are much clearer.

- A small (but perceptive!) comment by referee 2 about our use of the word biomass in the Discussion in spite of there being no biomasses in our analyses prompted us to evaluate the robustness of our analyses to using biomass instead of abundance. We did so using mean body mass for all species, and now present these complementary analyses in SI Appendix S5; our results do hold. We now mention this at the beginning of the Discussion l. 286-287.

- The Results and Discussion sections have been divided into subsections for the sake of clarity.

We provide below a response to referee comments, together with additional analyses whenever needed. We appreciate the many thoughtful remarks of both referees and have considerably rewritten the manuscript to match their recommendations. That said, there are two technical points where we respectfully differ with referee 1 and for which we provide a detailed rebuttal in the response letter:

1. R1 considers that Gross *et al.* (2013)'s synchrony metric is not valid (in spite of this method being not only widely used but also validated on simulated community data in the original paper). Unfortunately, this viewpoint seems to originate from viewing the mean Pearson correlation coefficient (averaged over all species pairs) as a meaningful measure of covariation between species in many-species system. We show, using simulations in the response letter, that this is not true; and that in fact the Gross *et al.* (2013) synchrony metric correctly outputs compensation or synchrony instead, based on four carefully crafted simple mathematical examples. As a sidenote, Appendix S4 (now S6) demonstrated this already to some degree, but the Fig. 1 of the response letter should make the superiority of the Gross *et al.* (2013) synchrony metric obvious, since the mean Pearson correlation coefficient shrinks to zero as the number of species increases. We can incorporate these details in the SI of the article for the next version, should you find it relevant.

2. Much of the other criticisms of R1 are based on the fact that statistical significance is at times weak or unclear, so that some of the results could have been generated as well by the null hypothesis of zero correlation between species. We agree with some of these criticisms, as our wording when describing compensation between waders and ducks was obviously too strong in places, and we have edited the text accordingly. But R1's remarks go a little further, and his interpretation seems to border on finding the null hypothesis more likely than the alternative. Although we share the concerns of R1 (the null hypothesis should always be kept in mind), we believe, as do many applied statisticians (Amrhein *et al.*, 2019), that one should not extrapolate from statistical significance to statements about the probability of the null. Statistical significance is an aide in interpreting statistical metrics and should not be the alpha and omega of ecological interpretation, especially for small sample sizes. Based on our simulations in Appendix S6 (under the alternative hypothesis) and new simulations under the null in this response, we show that for realistic ecological time series like ours (35 years, very long by ecological standards but somewhat short by statistical ones), weak statistical significance is inevitable when only two groups are considered, even when compensation is very much present in the simulations. In other words, the tests have low power, which we now mention explicitly several times in the text. Thus it is still quite likely that compensation is present between waders and ducks in winter in our data, which has clear ecological implications, and not publishing it as such on the ground of no clear *statistical* significance would only increase publication bias on abundance compensation in the literature.

We hope that the revised manuscript is now much clearer and matches the standards of Journal of Animal Ecology. Again, we thank the referees for this excellent feedback.

On behalf of all co-authors,

Best regards,

Frederic Barraquand

REFEREES' COMMENTS TO AUTHORS

Reviewer: 1
CONFIDENTIAL COMMENTS TO AUTHORS

Review of JAE-2019-00289 Barraquand et al "Looking for compensation at multiple scales in a wetland bird community"

This manuscript deals with broad themes that are certain to be of interest to readers of the Journal of Animal Ecology - how we can interpret temporal patterns in community dynamics to infer how (animal) species interact with each other and their environments. The authors focus on a relatively recent metric for summarising covariance patterns among species or functional groups, applying Gross *et al.* (2013) 'overall synchrony' index to time series of multiple waterbird species sampled from a wildlife reserve in France, considering taxonomic and functional groupings to search for evidence of compensatory dynamics. The authors seem to rely heavily on visual inspection of plots to support their main interpretations. The main conclusions are not too clear - but appear to be that it's difficult to find evidence of compensatory dynamics at any temporal scale in this data-set (though my paraphrasing contradicts some of the authors own conclusions, which I don't feel are well supported/justified). This general outcome matches some previous findings, while also falling into some of the same traps as previous research. The comparison of taxonomic vs functional is one of the most interesting aspects of the manuscript, even if there was little evidence of clear differences in outcomes at these two levels (again, my interpretation differs from the authors' claims about this), while applying this general framework to vertebrate populations is also fairly novel.

My main concerns are based on the technical assumptions that underpin the analyses, that led me to different conclusions than the authors arrived at. There are some nice ideas here - but more important foundations need to be laid before the results and interpretations can be put into their proper context. If that can be done, I think there is potential for a useful contribution here.

We thank the referee for this appraisal of our manuscript, and have made considerable edits to make sure that the manuscript is clearer and does not overstate its results. However, for reasons that we demonstrate below, we do think that our main conclusions still hold. We would like to stress two general points:

- Mean Pearson correlation is a poor indicator of covariation in large dimensional systems, so that it is in fact perfectly possible to find a near-zero *mean* Pearson coefficient in systems where most pairs of species are negatively correlated (hence showing compensation), corresponding to a quite negative Gross index η even when the species richness gets large. This is a deficiency of the mean Pearson coefficient as an indicator of negative covariation (and not a deficiency of η). We illustrate this with several numerical examples where we control the degree of compensation (i.e., the distribution of correlation coefficients between individual time series). Unlike what the referee suggests, η does work very well in many cases of ecological relevance, some already demonstrated in Gross *et al.* (2013) with simulated data in their SI. But we re-demonstrate this again here, in this response letter, using four model scenarios.
- In the opinion of most applied statisticians, statistical significance is but one facet of the many threads of evidence that one has to weight to choose between competing hypotheses (see recently Amrhein *et al.*, 2019). We agree with the referee that we used a quite unfortunate wording when reporting compensation: our words were too strong given the weak or absent statistical significance *sensu* NHST. This has now been corrected (see below). However, neither p-values nor confidence intervals are direct measures of the evidence for the null (Amrhein *et al.*, 2019), and our findings may well be genuine. Although we are convinced that of course the referee is well aware of the theoretical interpretation of p-values, many seasoned statisticians have at times been fooled by p-values and statistical significance in applied cases. We feel that in this particular case, the conclusions drawn by the referee hinge towards interpreting our failure to convincingly reject the null as *evidence for the null*, which they are not. As we already showed in Appendix S4 (now S6), $t = 35$ constitutes a very long annual time series for animal ecology but a small dataset for clear-cut statistics: even when there are known compensation effects

(i.e., simulated ones), significance at 95% or even 90% as used here might not be assured. This relationship between the size of the dataset and the level of significance required must, in our opinion, always be kept in mind (also one of the reasons why we use wavelets as well, as these can make use of the longer, monthly time series). We now state explicitly that compensation tests can have relatively low power l. 207-209 in the Methods, and mention this again in the Discussion.

(1) The Gross index is based on averaging 'point estimates' of multiple correlation coefficients - without considering the evidence for whether these individual estimates (or their means) differ from 0, or how the lack of independence of re-using the same series repeatedly (in the $\sum_j Y(j)$ terms) affects these interpretations.

Let us first remind the formulas for the sake of clarity. The Gross index is defined as

$$\eta = \frac{1}{n} \sum_{i=1}^{n} \mathrm{corr}(X_i, \sum_{j \neq i} X_j) \tag{1}$$

where the correlation is computed in our case using different times. In the case where the time series have all unit variance (e.g., though standardization), this formulation is itself equivalent to

$$\eta = \frac{1}{n} \sum_{i=1}^{n} \sum_{j \neq i} \mathrm{corr}(X_i, X_j) \tag{2}$$

because the covariance is a bilinear form. Therefore, η essentially measures *how, on average, species correlate with the rest of their community.*

The first issue is partially mitigated by the permutation/randomisation tests, which seem to bring the results for positive mean overall synchrony values in line with what would be expected from Pearson's correlation critical values for a given number of data points/length of time-series.

While we appreciate the referee's effort to simplify the exposition, we believe that it is inexact to state that the randomisation tests correct for the "first issue" of some potentially zero coefficients: the randomisation considers an $H0$ where all individual time series are independent while being identically distributed to the real data and with the exact same autocorrelation structure as the real data. Because we assume that this may be unclear to reader as well, we now write this in full l. 180-187. Moreover, we do not think that the "first issue" is really an issue: many statistics sum over many components with some that can be zero: the χ^2 and other summed statistics are excellent examples of this. We now provide an expanded description in the Methods following the request of the referee.

E.g., in Fig. 2, Pre-2006 Anas ($\eta \approx 0.35$, based on 25 years of data; Pearson's critical value = 0.38, which is reasonably close) are significantly different from the null, while Post-2006 Calidris ($\eta \approx 0.45$, based on 10 years of data, Pearson's critical value = 0.58) is not.

The general point of the referee – statistical significance is unclear at times – is indeed correct. We hope that this will be much clearer with the revised Fig. 2, that includes a boxplot of the values generated under the null. We do stress again, however, that statistical significance should not be over-interpreted.

Also, we would like to remind that the analogy with a simple test of significance of the Pearson correlation coefficient only works when:

- η is computed for two groups only ("between" cases in Fig. 2)
- The abundance values are themselves serially independent (i.e., no autocorrelation in the time series), which is not true here (we checked the ACFs).

However, given that the mean minimum Pearson's correlation values - those we are particularly interested in when searching for evidence of compensation - are limited to rho = -1/(n-1), yet Gross's index 'expands' to allow a minimum value of -1 for any community size (n), how can we interpret this inevitable expansion of parameter space with respect to rejecting a null hypothesis in different sized communities and over different lengths of time series? E.g., simulating stochastic series with a (mean) Pearon's correlation of rho = -0.056 for 3 to 20 coupled series (this value is chosen to match -1/(n-1) for n = 20) shows a decline in Gross's index from -0.1 to -0.7; i.e., coupled systems with exactly the same Pearson's correlation coefficients show considerable variation in the Gross index as the number of players increases.

Here, we have to (respectfully) point out a logical error in the above reasoning.

The referee initially correctly pointed out that – under equal correlation for all species pair – the minimum mean Pearson correlation coefficient shrinks as $\propto \frac{1}{n-1}$ where n is the number of species.

Unfortunately, the referee then suggests that a mean Pearson correlation coefficient across all species pairs could remain the same when the number of species goes from 3 to 20 (and that changes in value of the Gross index are instead problematic). This is incorrect. As we demonstrate below with several carefully constructed examples, one just cannot compare the mean Pearson correlation coefficient for vastly different numbers of species, these comparisons are meaningless.

In a large dimensional system, a negative average correlation, across all species pairs, is both mathematically and biologically improbable, because the "enemy of my enemy is my friend", i.e., strong negative correlation between some pairs will make some other positive correlations appear. Thus, averaging everything will necessarily make the mean Pearson correlation shrink.

Comparing the somewhat implausible scenario of constant Pearson correlation across all species to much more realistic ones puts the Gross index η in an even more favorable light. We consider four scenarios. In the first one, we consider pairs of highly negatively correlated species with pairwise Pearson correlation $\rho = -0.8$ and unit variance, that are not correlated to other pairs. That is, the variance-covariance matrix of the vector $\mathbf{x}_t = (X_{1,t}, ..., X_{n,t})^T$ is

$$\Sigma = \begin{pmatrix} 1 & \rho & 0 & 0 & \cdots & \cdots & 0 & 0 \\ \rho & 1 & 0 & 0 & \cdots & \cdots & 0 & 0 \\ 0 & 0 & 1 & \rho & \cdots & \cdots & 0 & 0 \\ 0 & 0 & \rho & 1 & \cdots & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & 0 & 0 \\ 0 & 0 & \ddots & \ddots & \vdots & \vdots & 1 & \rho \\ 0 & 0 & \ddots & \ddots & \vdots & \vdots & \rho & 1 \end{pmatrix} \tag{3}$$

We have the formula for η

$$\begin{aligned} \eta &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{corr}(X_i, \sum_{j \neq i} X_j) \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j \neq i} \mathrm{corr}(X_i, X_j) \end{aligned} \tag{4}$$

In the above pairwise case, only one of the correlations in the inside summation is non-zero so that $\eta = \frac{1}{n}\sum_{i=1}^{n} \rho = \rho = -0.8$ in theory (for infinite sample sizes). This makes sense that η is fairly negative because as half of the species go up, the other half go down.

The average Pearson correlation coefficient, however, is computed over all non-diagonal entries of the matrix ($n(n-1)$), so that $\bar{\rho} = \frac{1}{n}\frac{1}{n-1}(\rho + (n-2) \times 0) = \frac{\rho}{n(n-1)}$. For 20 species, $\bar{\rho} = -0.0021$ which does not reflect at all the strong negative pairwise covariation occurring in this system, which can be construed as a pathological property of the mean Pearson coefficient to quantify synchrony/compensation.

Let us consider now three other examples using Beta distributions to model the distribution of correlations coefficients ρ_{ij}.

The second scenario assumes a quasi-normal distribution. We set $\mathbf{x}_t \sim \mathcal{N}(0, \Sigma)$ with $\Sigma_{ij} = \rho_{ij}$ and $\Sigma_{ii} = 1$ (unit variance), where $\rho_{ij} = 2B_{ij} - 1$ with $B_{ij} \sim \text{Beta}(15, 15)$. In this case, we have $\eta \to 0$ and $\bar{\rho} \to 0$ as n gets large. We show in Fig. 1 the behaviour of both η and $\bar{\rho}$ as n increases.

We now rig that lottery towards compensation (scenario 3), assuming that $\rho_{ij} = 2B_{ij} - 1$ with $B_{ij} \sim \text{Beta}(2, 4)$ so that on average most correlations are negative, but we allow for some large positive ones to arise. Finally, do the same for temporal synchrony (scenario 4), with $B_{ij} \sim \text{Beta}(4, 2)$, most correlations now being positive.

We only keep those matrices that allow for positive-definiteness. We now examine in Fig. 1, for the four abovementioned scenarios:

- How η changes with the number of species n
- How the mean correlation coefficient $\bar{\rho}$ changes with n

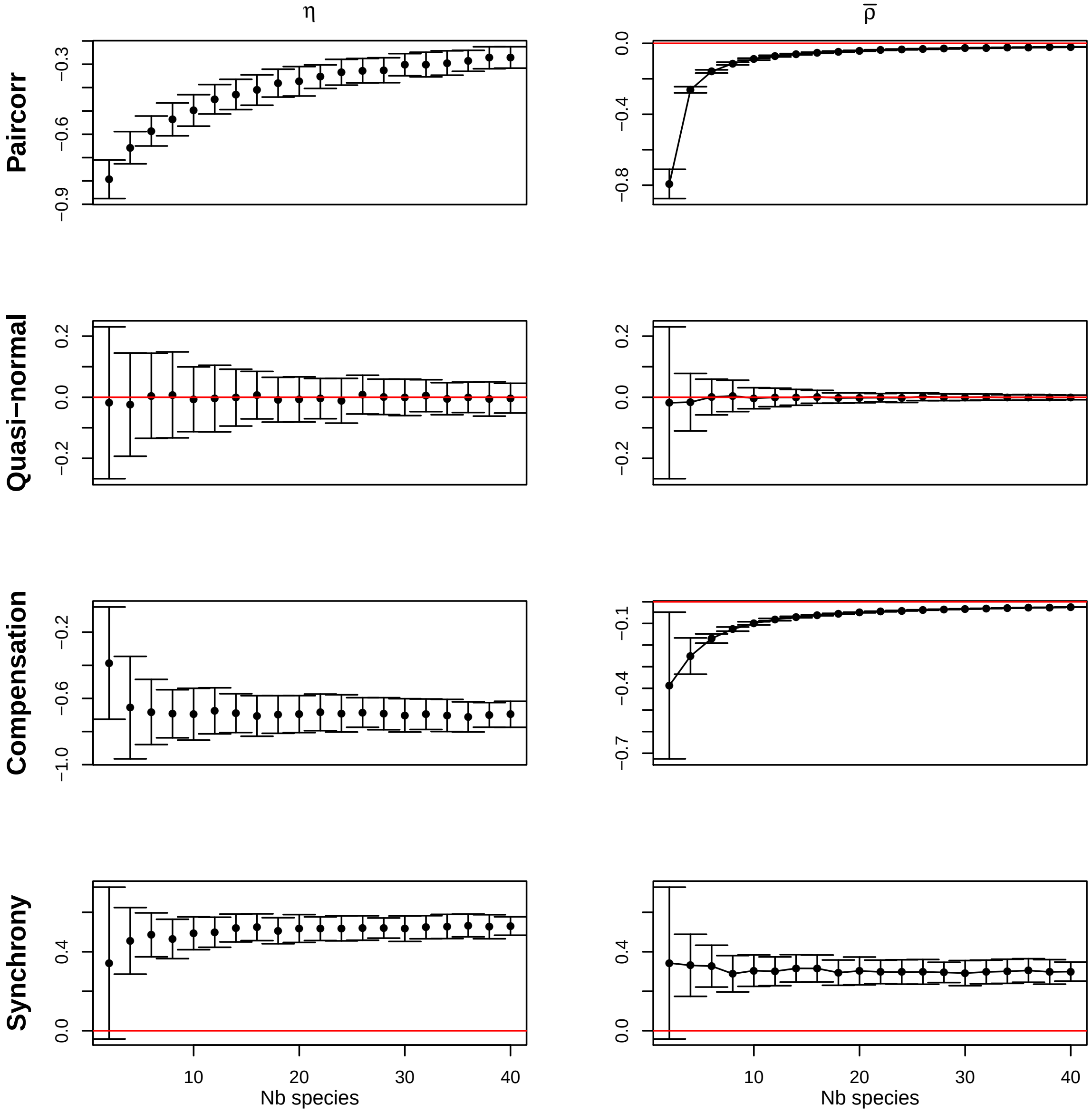

Figure 1: How the Gross index η and the mean Pearson coefficient $\bar{\rho}$ change with number of species in four cases. Paircorr: otherwise independent pairs of species with negative covariation of 0.8; Quasi-normal: Gaussian-looking, symmetric Beta distribution; Compensation: compensation generated by a correlation coefficient distribution skewed towards negative values; Synchrony: temporal synchrony generated by a correlation coefficient distribution skewed towards positive values. The first value in the x-axis starts at two species. The red line indicates a zero y-value.

We show that while η is negative whenever there is compensation ("Paircorr" and "Compensation" in Fig. 1), $\bar{\rho}$ shrinks fast with the number of species in all the compensation cases, because any negative correlations in a many-species system will inevitably generate some other indirect positive associations. And in the case of synchrony only, η outputs the same information as $\bar{\rho}$.

This is a major issue around the use and interpretation of this synchrony index (which the original index's authors didn't deal with; similar issues arise with most synchrony indices). While it doesn't affect all of the results presented here, it does affect some important results/interpretations and needs to be dealt with, in my opinion. What sort of evidence would be required to reject the null hypothesis in favour of an alternative hypothesis of compensation? Is that sort of value likely to arise in any time-series data (which tends to be whitened/less correlated by measurement/observer error anyway)? This is related to carrying out an appropriate power analysis for this study, which is complicated by the choice of metric used in this study.

As we explain above, there is no major mathematical/conceptual issue with η, only with the mean Pearson correlation across all pairs of species. Furthermore, Gross *et al.* (2013) did consider what η means to a large degree and whether its interpretation is robust to deviations, cf. their online appendices C and D (mathematics in C and simulations, including with measurement error in D).

Regarding statistical significance, in the case where compensation actually exists - the "alternative" hypothesis - we already investigate in our Appendix S6 what within-guild η values can be expected for various community sizes, length of time series and sampling design (we have just corrected that Appendix to include the case with two compartements, which is useful for between-guild comparisons). The alternative hypothesis considered realistic for our study was that two guilds react in opposite ways to an abiotic driver. We showed that in most cases, η was indeed negative whenever there was compensation generated by the joint action of competition and weak opposite reaction of the growth rate to shared abiotic drivers.

Regarding comparisons to the null, to allow for a clearer test, we have generated here (in this response letter) values of η with the randomisation techniques of the main text and compared those to the observed values of η for the simulated examples (the four abovementioned scenarios). The results in Fig. 2 show that while significance can be attained for large number of species/groups, two groups is typically a scenario where even with substantial compensation, the null is hard to reject.

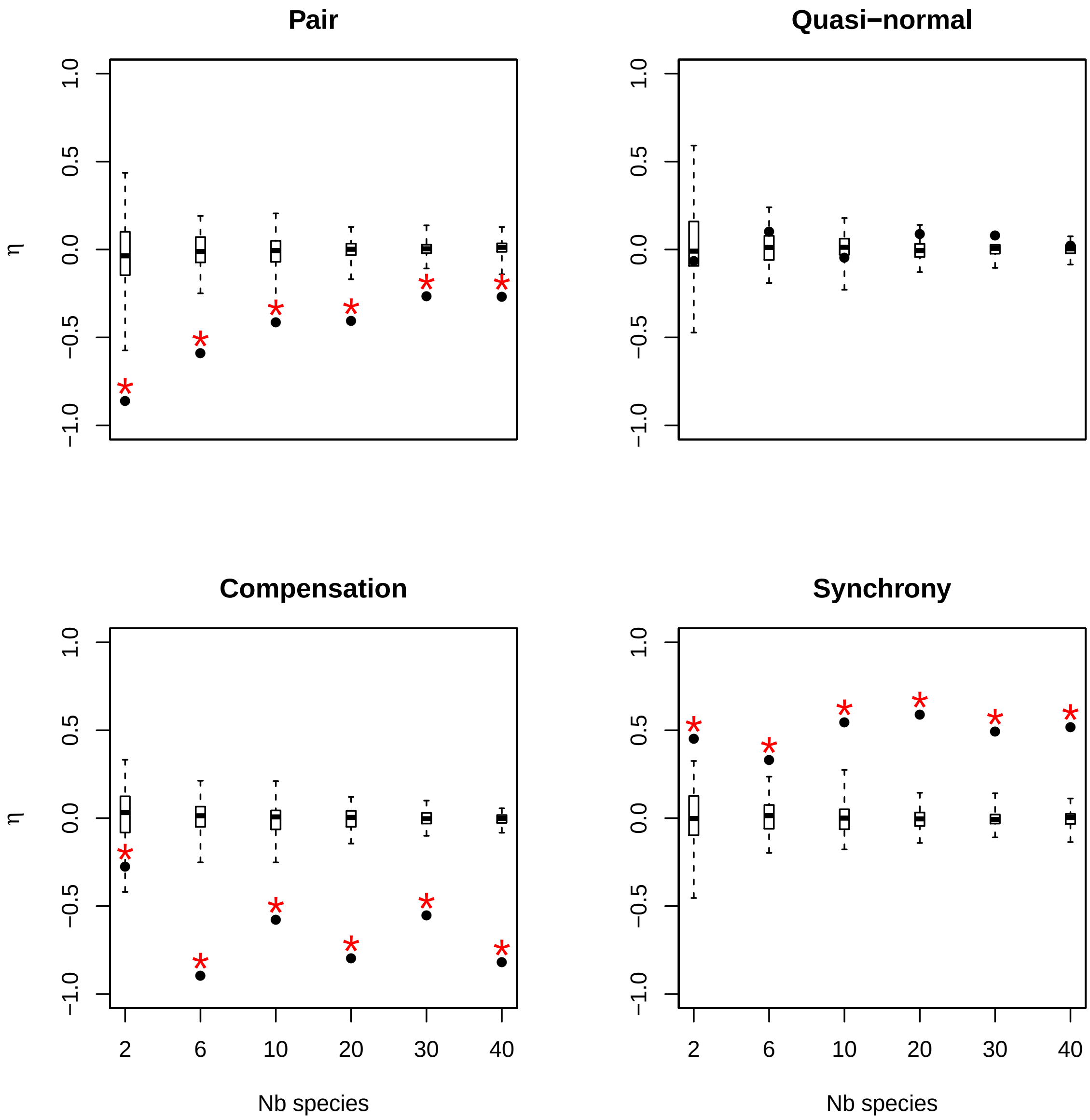

Figure 2: Values of the Gross synchrony index η as a function of the number of species in four different scenarios. Pair: otherwise independent pairs of species with negative covariation of 0.8; Quasi-normal: Gaussian-looking, symmetric Beta distribution; Compensation: compensation generated by a correlation coefficient distribution skewed towards negative values; Synchrony: temporal synchrony generated by a correlation coefficient distribution skewed towards positive values. Red stars indicate that the results are significant at the 10% α threshold.

This answers the referees' concern regarding power: power is high when there are many groups but low when there are two groups, where not reaching significance, even at the 10% level, does not invalidate true effects. We mention this in the discussion now. Should that be needed, we can also include these additional simulations of the response letter in the SI of the article. Note that this is a qualitative statement on power: we might be able to calculate power as a function of sample sizes (with some assumptions on true effect sizes), but in our opinion, as we cannot change the observational design, quantitative estimates of power = f(sample size) would be of little use here.

(2) Related to this point are the conclusions the authors have drawn in the manuscript, that I don't feel are supported by the results presented. E.g., Abstract (lines 18-19), Results (lines 239 - 243) and Discussion (262-264, 276-277).

For clarity, let us remind the content of those sentences:

- Abstract: "We also found that compensation varies with taxonomic and functional scale: compensation appeared more frequently *between* rather than *within* guilds."
- Results: "More clear-cut results can be found when we examine the synchrony vs compensation between functional groups (Fig. 2d) [...] Waders and ducks are negatively correlated during the cold season and positively during the warm season. These patterns are in contrast unclear when using a taxonomic classification (no compensation, Fig. 2c)"
- Discussion: "Yet, summing species abundances within a guild and comparing the "biomass sums" of contrasted guilds, community composition did change in frequency in the long run; in other words, there was some compensation between guilds."
"The functional group classification produced much more clearly compensation between guilds than the taxonomic classification."

My interpretation of the main results in Fig. 2 (& 5) is that there is no statistical evidence for compensation in any of the configurations considered, nor is there justification for claiming differences in sync values between seasons. Even in Fig 2d, where there is a strong, negative synchrony value for the Cold Waders/Ducks Post-2006 result (eta ≈ -0.7), there is no evidence that this is significantly different from a null sync value ≈ 0 (no red asterisk shown in the panel) nor between cold and warm seasons Post-2006.

We are partly in agreement with the referee here, in the sense that our use of the word "clear" is misleading – and very unfortunate, given the recent call by Dushoff *et al.* (2019) to replace "non-significant" by "unclear": our results on compensation are *present* but *unclear* according to Dushoff *et al.* (2019)'s terminology. We have now corrected these sentences considerably in order to tone them down:

- Abstract: "We also found that synchrony varies with taxonomic and functional scale: the rare cases where compensation appeared consistently at the annual scale were between rather than within guilds."
- Results: "More interpretable results can be found when we examine the synchrony vs compensation between functional groups (Fig. 2d). [...] In contrast, waders and waterfowl are negatively correlated during the cold season and positively correlated during the warm season. Although the negative correlation is not statistically significant, it is consistent for both pre- and post-2006 periods. When using instead a taxonomic classification with the two phylogenetic groups, we find no compensation, with either zero correlation or synchrony (Fig. 2c)."
- Discussion: "Yet, summing species abundances within a guild and comparing the total abundance of contrasted guilds, it was possible to find compensation (although the null hypothesis of no correlation could not be rejected); in other words, there was some compensation between guilds."
"The functional group classification produced some compensation between guilds while the taxonomic classification did not, despite the contrasted habitat preferences of these two phylogenetic groups. Using functional groups produced more logical results, although as we stressed above, due to the low power of the tests, the null hypothesis of no compensation at the yearly scale is still plausible as well."

The absence of statistical significance for compensation at the annual scale should now be apparent to the reader. However, we believe that one should not fall into the other extreme: interpreting compatibility of the null with the data with evidence in favour of the null, which it is not. It is also quite possible that true compensation is present between waders and waterfowl throughout the cold season, especially as periods pre- and post-2006 considered indicate the same thing.

There may be differences between some specific comparators, but the authors haven't presented a convincing way of demonstrating/supporting these. I think more biological (and statistical) motivation for the null/alternate hypotheses should be presented, that could help clarify these issues. Could a factorial (LM/GLM) analysis help to disentangle some of the post-hoc comparisons that probably need to be made to support the claims being made?

As we have highlighted above, there are no real issues with η, but we now more explicitly mention the null hypothesis. We reckon that power is low with two groups and mention this issue l. 207-209 in the Methods: "This was found satisfactory based on simulated data, although power is low for detecting compensation (i.e., the null cannot always be rejected) when only two groups are compared."

and in the Discussion l. 308.

(3) There seems to be conflation of intrinsic (biological) and extrinsic (environmental) factors at places in the manuscript. A clearer focus on the biologically relevant processes that occur at the different time-scales considered (annual vs monthly, births, deaths, immigration, emigration; community species turnover, etc) would help distinguish these and put the work in better context.

We have strived to present more clearly the biological processes throughout (Introduction and Discussion in particular). However, it is also important to remember that some time-scales are indeed conflated in real life, for instance, while birth is usually a pulse, death is continuous through time, just like the local movements of birds.

(4) More information about the composition of the functional groups compared to the taxonomic groups is required. How much overlap is there between these groupings? What does this lack of independence mean for interpretation?

We thank the referee for this suggestion (referee 2 made similar remarks) which will no doubt clarify the manuscript. The information about the full species composition has been added to SI Appendix S1, with the frequency of occurrence along the time series and the relative abundance of the species in the community. The taxonomic groups are both *included* within the functional groups. Within each taxonomic group, not only habitat preferences and diet are more constant than within functional groups, but also body sizes and other trait descriptors. Or if you will, phylogeny makes in that case the species more alike than function.

The lack of independence between the two groupings was the reason why we Bonferroni-corrected in Fig. 2 in the first place. Now we use the Benjamini-Hochberg FDR-correction, following an excellent suggestion of reviewer 2.

(5) More clarity is required in describing how you generated the 100 independent series for the randomisation tests (lines 190 - 198). I can guess how you might have done this, but I shouldn't have to guess (or view your online code, as I may prefer another coding language) and even then, my guess may be wrong! Were periodic 'boundary' conditions used? Was each pair of (taxonomic/functional) time series shifted more than once? There's a maximum number of times a pair of series can be shifted, which may mean you don't really carry out 100 independent randomisation tests, particularly for those cases where you start with relatively few time-series (e.g., Fig. 5).

We understand the reviewer's point and we have therefore added much more details about the algorithms used (see quote below).

With many species, there are always many more possible combinations of shifted time series than actual surrogate time series that are used, so we kept this simple method. However, with only two time series, the referee is correct. To avoid having several times the same shift for correlations between the summed abundances of two groups (as mentioned by the referee), we have used the IAAFT, a standard method (albeit more complex) to construct surrogate time series (Detto *et al.*, 2012).

The full explanation now reads:

"We computed the statistical significance of the synchrony index by comparing the observed values to the distribution of η under the null hypothesis (Gouhier & Guichard, 2014), which amounts to zero cross-correlations between species abundances (or guild-level abundances when considering functional groups). The challenge, in order to construct such null hypothesis, is to remove all cross-correlations while keeping the exact same autocorrelation in each individual time series. Therefore, for each set of time series (each combination year × season for a given community), we constructed 100 "surrogates" in which we kept auto-correlations but removed the cross-correlations between time series. There are multiple ways to erase cross-correlations depending on the resolution of the considered community. Within guilds, we shifted the time-series (Purves & Law, 2002) while between guilds (two groups only), we used a frequency-based approach (Iterated Amplitude-Adjusted Fourier Transform or IAAFT, see Schreiber & Schmitz, 2000). We first explain the shift-based approach: the suite of abundance values (after seasonal averaging) is displaced by a random temporal lag τ, so that a value y_t is now found at $y_{t+\tau}$. At the boundary (the end of the time series), remaining points are displaced towards the beginning of the time-series, which implements a toroidal shift. This method works well when comparing many times series corresponding to the multiple species. However, when computing synchrony across only two groups (between guilds), spurious cross-correlations could emerge with a shift-based approach as the number of possible combinations is more limited. Therefore, to test for synchrony between the summed abundances of two guilds or taxonomic units, we used the more sophisticated IAAFT method (Schreiber & Schmitz, 2000), which retains the frequency spectrum of the time series while randomising its values. We obtained 100 sets of randomised time series for each computed synchrony index. We then compared the number of η_{H0} values which exceeded or were inferior to the observed value to compute the p-value North *et al.* (2002): we use the ratio $(r+1)/(n+1)$ where r is the number of surrogate values that are $\geq \eta_{obs}$ (respectively $\leq \eta_{obs}$) and n is the number of surrogates. Independence of species was rejected at the 10% threshold with a Benjamini-Hochberg correction, as we compare across 2 seasons and 3 periods, with partially overlapping data. This was found satisfactory based on simulated data, although power is low for detecting compensation (i.e., the null cannot always be rejected) when only two groups are compared."

Minor Comments: The first 3 paragraphs of the Introduction are rambling and don't do a great job of succinctly introducing readers to the relevant background on the topic. I'd recommend a briefer, much more focussed paragraph or two summarising this material - and including some of the other work that has dealt with the problems inherent in summarising complex emergent systems behaviour in a single number. In my opinion, ecological systems are enormously complex - and it's unlikely that this can sort of synchrony/compensation summary be done meaningfully with a single value. In general, I think it will be much more productive to embrace the complexity and build understanding from there!

Unfortunately, we had in initial drafts a concise Introduction but rewrote a longer one for the manuscript that you read, in order to take into account previous referee comments (reviewers that were not specialists wanted more background). We therefore rewrote once again the Introduction to a crisper format of ≈ 850 words. However, we did keep some pedagogical exposition of what compensation is, based on this previous feedback of non-specialists – in our opinion, we have to think of readers unfamiliar with the theory-driven synchrony literature in a general journal such as J Anim Ecol. Regarding summarizing compensation with a single number:
- we use analyses at multiple scales (not a single number then)
- the abovementioned simulations on the Gross index show that it can capture the main features of compensation whenever it is present at a given temporal scale.

Figure 1 gave me a headache! Particularly, panel (c) contains too much information to be useful - it really doesn't illustrate anything useful to me, beyond the presence of a bunch of messy, fluctuating species. Same general comment for many of the time-series plots in the Appendices.

While we understand that some of the information in Fig. 1 might not be essential for all readers, we believe that readers with an empirical background would like to see the data before the analyses (and the seasonally averaged data are already "simpler" than the monthly ones presented in the SI). We would therefore like to keep that figure, unless referee 2 also feels that it is useless (in which case we would of course remove or simplify it). The data may be messy because of rare species, but they are the data - it would feel awkward to us (ethically speaking) to remove data plots because they don't "look neat". We hope that the referee understands our wishes to show the data as it is.

I found Fig. 2 very difficult to get my head around - I'd encourage the authors to consider a different way of presenting these key results that readers will be able to extract information from more easily. Also, they could consider other colour, and symbol choices sand ensure the lines and error bars are described in the Fig. 2 caption. Including the number of taxonomic/functional groups in the panel legends would also be a useful guide to interpretation.

Regarding Fig. 2, we completely agree with the referee. We have actually completely reworked that figure following comments from reviewer 2 as it had some conceptual problems as well. We hope that the new design will be easier to read and interpret. We now only output 1/ the observed η values and 2/ a boxplot that shows a corresponding distribution of η values under the null.

I recommend reading Royama (1981, Ecological Monographs), as it could provide useful pointers for some of the approaches used here. Per-capita growth rates may be a more appropriate measure for working with than raw population densities.

We thank the referee for this suggestion. We are actually familiar with Royama's work (the book as well) and multivariate autoregressive models which we use extensively for other purposes.

However, the objective of the present work is not to partition density-dependence by working on growth rates (which here partially represent movements in/out), but to quantify the amount of synchrony in *abundance*. We would like to stress that an analysis on per capita growth rates would essentially answer a different question, since maximum per capita growth rates can be synchronous without abundances being synchronous (as the referee surely knows). For instance, Loreau & de Mazancourt (2008) demonstrate that per capita growth rates ($\gamma_i(t)$ in their paper, Appendix A) co-vary negatively even in cases where the population abundances demonstrate synchrony
https://www.journals.uchicago.edu/doi/10.1086/589746

Signed: Mike Fowler

Many thanks for this thorough and well-argued review.

Reviewer: 2

CONFIDENTIAL COMMENTS TO AUTHORS

This is very well-written manuscript about synchrony in population dynamics both within and between functional groups. The paper makes use of an excellent dataset, and is creative and thoughtful. The statistical analysis is well chosen and the conclusion support the question and conclusion. Regarding the equations and derivation of the analysis in Appendix 4, I am not qualified to judge its correctness.

We thank the referee for these kind words, general overview of our manuscript and constructive suggestions.

I have two major issues that would require some work 1) The introduction is presently very long and seems to contain a lot of information that is unnecessary for the story. This is most evident in the part about zero-sum dynamics, which is a theoretical construct that does not occur in real nature, and is especially unlikely in this bird community. I suggest cutting down the introduction to 800 words. This will probably naturally lead to the exclusion of non-essential parts

The Introduction has now been cut down to 850 words. The previous version of the Introduction, which we submitted to J Anim Ecol, had unfortunately been lengthened to answer previous feeback by non-specialists who wanted a more detailed introduction. Despite the simplification, we kept the mention of zero-sum dynamics which, we think, helps naive readers to get an impression of what we are talking about – we use such a limit case for pedagogical purposes. Please note also that we refer to “close to zero-sum”(l. 45), not exactly zero-sum

2) The authors make comparisons among different functional and taxonomic groups. This is very interesting, but the choice of functional groups to be studied is poorly justified.

It seems that we have unfortunately assumed that our readers would immediately realize why waders and ducks or waterfowl (herbivorous divers) would trade off. We now better justify these choices, as of course the groups have been chosen based on their ecological differences. Waders prefer low water levels and ducks prefer to dive down, preferring more depth close to their foraging grounds. A whole subsection (Bird taxonomic and functional groups, l. 132-154) now explain why we focused on waders vs duck-like birds.

The functional groups presently studied are: waders vs ducks and herons vs cormorants. Since no species list is provided, the reader does not know which other groups were present. I would imagine also songbirds and raptors were observed and counted, as well as geese, swans, gulls and terns.

That is a fair comment. A species list has now been added in SI Appendix S1, which we refer to at the beginning of the Methods, with a breakdown of the percentages covered by each species. Songbirds actually comprise 0.2% of the dataset, these are clearly not relevant here. Waders and waterfowl, as now explicitly mentioned in the text l. 133-135, represents in abundance about 70% of all species. Herons, egrets and cormorants represent a fair share of the species that are neither waders nor waterfowl. The other major players are gulls. Other than that, we have few raptors, but as the reserve is really on the water, most of the birds seen there are waterbirds.

We took this opportunity to rename the “Duck” group as “Waterfowl”. As we already highlighted p. 10 l. 181-184 of the previous manuscript, this guild actually comprised mostly ducks *stricto sensu* and some other anatids, including the much larger geese and swans. The guild is therefore composed of the herbivorous divers (NB: we cannot rename it “divers”, it would induce confusion since we have other diving birds such as cormorants). Although “waterfowl” may be less transparent to readers that are not birders, we have opted for this designation to avoid confusion. We recall in several places that waterfowl = ducks, geese and swans (+1 other species in our case). Indeed, the Eurasian coot (*Fulica atra*), which is a rail that has a duck-like morphology (like most coots), is usually not included within waterfowl. We did include it in this case as we use waterfowl as a functional group and not a phylogenetic unit.

At present the authors do not mention these and only focus on the aforementioned groups. I agree that the groups tested can be expected to show any kind of compensation more than a comparison between let's say waders and song-birds, but this would be especially interesting to show! This would also diminish the danger of reporting spurious results, especially if the expectations of synchrony/compensation for the different comparisons are well justified.

On this particular issue, we do not fully agree with the referee. Increasing the number of comparisons between groups actually generates opportunities for some groups to covary negatively by chance. Even if corrections for multiple testing that contain the FDR are implemented, we are wary of multiplying tests which do not have a clearly associated, logical biological hypothesis. Moreover, as highlighted above, our groups have been chosen so as to be the main groups present in the reserve which should be evident now with the species list (see Tables 1& 2 in SI S1).

We recently published a paper on a similar topic, but measured on ground beetles (Carabidae) in the Netherlands (https://esajournals.onlinelibrary.wiley.com/doi/10.1002/ecy.2748). Here we use related methods to show that functional differences decrease community synchrony, and that functionally less similar species show compensation more often. This might be a useful reference.

We thank the referee for this useful reference, now cited in the Introduction (l. 55, l. 71, l. 85 and l. 89), notably in support of the idea that functional diversity might lead to more compensation. The study is again cited in the Discussion l. 334 and l. 352.

I suggest that the authors also calculate the synchrony within and between these functional/taxonomic groups where no effect of the change in water management would be expected (e.g. waders vs song birds and gulls vs ducks) in this way we can better understand the significance of the compensation found for the presently studied taxa/functional groups.

Again, as nothing prevents two groups from covarying negatively or positively in absence of any true mechanism, we are wary of post-hoc interpretations for species that do not have reasons to covary and we prefer to avoid this. Referee 1 is already wary of potential misinterpretations in cases where we have good reason to believe that the two species abundances will trade-off (waders vs ducks), hence we prefer not to add any results that might lead to speculation in the MS.

If such a broader comparison of functional groups is made, this will naturally change the manuscript severely. Roscher et al (2011 J ecol) also made a comparison in synchrony between different functional groups of plants, if I remember well, and this may also be a useful reference, although they use a different synchrony metric.

We now cite Roscher et al as well l. 89.

I have only few minor textual comments:

L68-74: Even small deviations from perfect synchrony have a stabilizing effect
We now mention this with a reference to Loreau and De Mazancourt (2013).

L141 a random '.'

Corrected.

L158 and further: n as species number is very counterintuitive and against conventions. Normally, species number is designated with 'S' and 'n' means number of individuals (see Anne Magurran's work). I did not check if and how Gross / Loreau formulate their equations, but I would promote consistency with conventions.

Gross et al. used n, Loreau and de Mazancourt used S; we used Gross et al.'s notation, notably because in Appendix S4 (now S6) we used the same model as they did. We have therefore kept that notation. It is therefore consistent with Gross et al.'s notation, but if that generates any confusion, we could also change it.

L195: 10% level – do I understand correctly that this means you consider p=0.1 (after correction) significant? I find this a rather high threshold.

On the contrary, given the relatively low power for two groups (see the response to referee 1) and limited length of ecological time series, we find the 10% threshold quite appropriate (note that a dataset of $t = 35$ *with* correlation between points has a lower effective sample size). The tests are two-sided too, so there is 5% on both sides.

Additionally, Bonferroni is known to be conservative and tends to lead to false negatives. I would rather suggest to use FDR (False discovery rate) and a threshold of $p = 0.05$) to correct for multiple testing. I cannot predict how this will affect the outcome.
This is indeed an excellent suggestion for which we are very thankful. We implemented the Benjamini-Hochberg FDR correction (but kept the overall 10% level, in accordance to the time series length).

L209 'biomass'- I assume you mean community abundance? Since you didn't measure biomass

Yes – we meant abundance and corrected accordingly l. 220. That said, your remark prompted us to redo those analyses with biomasses (using average body masses for all species), these are now presented in SI S5. The results are similar, we now mention this at the beginning of the discussion l. 286-287.

225 . . . which consisted OF lowering. . .

Corrected.

233 Not sure what overcompensation means

This is a typo, it should be compensation – corrected.

Fig 2: print the numbers on the y-axis horizontally (las = 1).

This has been corrected.

These errorbars are extremely small, and don't seem to reflect well the deviance from 0. I wonder how they were calculated. If they are calculated using a permutation approach, it makes more sense to report a measure of error that is independent of sample-size, since sample-size is determined by the number of permutations

Upon closer inspection, this presentation was misleading (and erroneous, given a division by the number of replicates). This was indeed not the proper way to present the uncertainties, which lead us to revise completely the way we present the significance of η. We now present the full distribution of η under the null (as a boxplot) in Fig. 2 (see also the simulations in the detailed response to referee #1, where we compare the null to carefully constructed simulated examples).

Fig 2 and 5: I have some difficulties understanding what the cold/warm season comparison means and how I should interpret this. Please provide some clarity.

The figures have been redone, we hope that it is now clearer. The seasons are defined l. 122-124, the cold season is "end of fall and winter" (i.e., wintering season for birds), the warm season is "spring and summer", i.e., reproduction season.
L235 contrastING

Actually here we believe that it should be "contrasted".

L238 not a very elegant sentence.

It has been rewritten l. 254-256

L242-243 move ‘in contrast’ to start of sentence

The sentence has been changed in ‘in contrast’ deleted.

Fig 4: T is missing in Cormorant. Both in figure and in legend

Thank you for noticing, this has been corrected.

L246 we see IN fig. 5 – this is a rather surprising result, looking at Fig 4. I would expect clear compensation. It actually makes me a bit doubtful of the method (which I never used myself).

Indeed, this is surprising, but true – we checked with additional representation of the time series in Appendix S4. This is because compensation only occurs at long temporal scales ≈ 6 years, but not at the yearly scale.

L277 – much more clearly – please rephrase

We have rephrased to: “The functional group classification produced some compensation between guilds while the taxonomic classification did not, despite the contrasted habitat preferences of these two phylogenetic groups. Using functional groups produced more logical results, although as we stressed above, the null hypothesis of no compensation at the yearly scale is still plausible as well.” .

L287-288 – I guess you mean here waders and ducks respectively – I think it will be better to be specific

This has been corrected.

L318 ‘work nicely’ please rephrase

Corrected to ‘be appropriate’.

Title Appendix 3: has a type-o

Corrected.

Fig S6 and S7: explain what is b

This information has been added to the legend, this is the strength of the effect of the environmental variable on the population growth rate.

Keep up the good work
Roel van Klink

Many thanks for this constructive review.

References

Amrhein, V., Greenland, S. & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, 567, 305–307.

Detto, M., Molini, A., Katul, G., Stoy, P., Palmroth, S. & Baldocchi, D. (2012). Causality and Persistence in ecological systems: A nonparametric spectral Granger causality approach. *The American Naturalist*, 179, 524–535.

Dushoff, J., Kain, M.P. & Bolker, B.M. (2019). I can see clearly now: reinterpreting statistical significance. *Methods in Ecology and Evolution*, 10, 756–759.

Gonzalez, J., Düttmann, H. & Wink, M. (2009). Phylogenetic relationships based on two mitochondrial genes and hybridization patterns in Anatidae. *Journal of Zoology*, 279, 310–318.

Gouhier, T.C. & Guichard, F. (2014). Synchrony: quantifying variability in space and time. *Methods in Ecology and Evolution*, 5, 524–533.

Gross, K., Cardinale, B.J., Fox, J.W., Gonzalez, A., Loreau, M., Wayne Polley, H., Reich, P.B. & van Ruijven, J. (2013). Species richness and the temporal stability of biomass production: a new analysis of recent biodiversity experiments. *The American Naturalist*, 183, 1–12.

Loreau, M. & de Mazancourt, C. (2008). Species synchrony and its drivers: Neutral and nonneutral community dynamics in fluctuating environments. *The American Naturalist*, 172, E48–E66.

North, B.V., Curtis, D. & Sham, P.C. (2002). A note on the calculation of empirical p balues from Monte Carlo procedures. *American Journal of Human Genetics*, 71, 439–441.

Purves, D.W. & Law, R. (2002). Fine-scale spatial structure in a grassland community: quantifying the plant's-eye view. *Journal of Ecology*, 90, 121–129.

Schreiber, T. & Schmitz, A. (2000). Surrogate time series. *Physica D: Nonlinear Phenomena*, 142, 346–382.