Answers to reviewer

Review by anonymous reviewer 1, 14 Aug 2023 14:45

Major comments:

My main difficulty with the text was understanding the different classification levels used for the databases. This made reading and interpreting the results challenging, and it was hard for me to connect the objectives with the workflow. I would strongly recommend standardizing terminology throughout the text and potentially creating a table or glossary with the terminologies, their definition and what they refer to specifically. For example, I am still unsure of what “local”, “regional” and “national” mean in this context. Are those terminology used to describe the databases and their level of coverage, or is it referring to how the dataset was processed and analyzed?

We have standardized the terminology:
- With regards to spatial scale terminology, we use “local”, “regional” and “country” (“country” replacing the previous term “national”) in the revised version and we have defined these scales in section 2.3.1 “The three spatial scales used for data collection are depicted in Figure 1: the local scale (i.e. study site with a 10km buffer around), the regional scale (i.e. French administrative regions); and the country scale (i.e. continental France).” Line 139-141
- the term “study site” is now used exclusively for the three institutional areas (T1, T2, T3) and we have deleted the term “territory”.
- The term “national” is exclusively used to name the combination of the two databases available at to country scale, i.e. combination of GBIF and Vigie Nature
- The combination of three databases (i.e. GBIF, Vigie Nature and SINP) is named exclusively in quotes, i.e. “All databases”

For example, you first introduce SINP as a database structured at “regional scale” (line 133), but then in the workflow diagram it is highlighted as “local” (and there is no “regional” scale present also).

We have improved the understanding of the paragraph: the SINP database is structured at the regional administrative scale but we collected the data at local scale due to the limited extent of the requests.

However, in table 2, I understood these terms to be used as parameters for subsetting the data. And while there is a definition for “local” and “France”, there is none for “regional” (and this is not introduced in the text either—at least I could not find it). Another example is in the use of the term “national database”, which in line 151-152 is introduced as being both GBIF and Vigie Nature (which I assumed is related to their level of coverage); then in line 176 is introduced as something else “(i.e. GBIF, Vigie Nature and National databases)”, which seems to be referring to a third category (!?). It looks like there is some ambiguity with the terms that makes it hard to know what exactly is being tested and analyzed.

This point has been treated above.

Besides the table/glossary of terms, another suggestion I think would be helpful would be to align the objectives of the study with the workflow diagram. So, for example, point out on the Analysis section: “Step 1.2 – Objective 1”. This structure could also be replicated in the text when describing the methods and results.

This is done in the methods section where we describe in detail the methods. We feel that it is not necessary to repeat the steps in the introduction / objectives.

Because I had some difficulty to understand clearly the objectives and methods of this study, I do not want to devote too much time with minor comments at this point. Once these major issues are clarified, I think I will be in a better position to provide more feedback. However, I present here some specific notes I think are worthwhile addressing at this stage.

Minor comments

Line 104-106: This phrase is repeated.

This has been removed.

Line 158: Did you mean "spatial" or "sampling"? In your appendix it looks like you are referring to sampling bias, and I think it merits a replacement of word in the main text. It's more specific and just for this one word is not worth it to make your reader consult the appendix.

Corrected. "limited their sampling biases" Line 172

Line 168-172: These terminologies need to be clarified. Are they referring to the spatial extent of the observations (e.g. the territories versus the country)? Or are they related to the coverage of the different databases for the same territories (shown in Fig. 1)? I think the second is what makes the most sense in this comparison, am I right?

We deleted the term "territories" with its systematic replacement by the term "study sites".

Line 179: Replace "seems to be sufficient" with a more specific term. Phrased like this sounds like you were going on a hunch, but I imagine that it was more methodical than that and they are following the thresholds mentioned above, right?

This has been corrected and we used the thresholds mentioned to assess the possibility of using these databases. "The data available in France for these groups is sufficient in quantity to realize SDM with each database." Line 195-196

Line 190-193: Again, the terminology needs some clarification here and throughout the document to reduce ambiguity.

We modified the terminologies, see previous answer.

Line 204: Replace "biogeographic" with "geographic". I do not think this is the proper term here, since biogeography relates to the interplay between biology and geography, and you are listing environmental variables.

Corrected in the text and in the Appendix. "Three types of environmental variables were used for SDM: geographic, human occupancy and pollution and fragmentation (Appendix A.2)." Line 223-224

Review by anonymous reviewer 2, 25 Aug 2023 16:35

L 27. I found the term "stakeholder" quite large. I would suggest a better explanation to whom this paper is adressed, and consequently, define more precisely the stakeholders, among scientists, public authorithies and managers. If i write this remark at L.27, a more accurate term or definition would have a better place in the Introduction. I suggest to reduce the use of verbing at the passive voice.

L94-95 : The role of stakeholders is not well defined -> how do they manage data are they enough competent to assess it? REgarding stakeholders, it depends also of conflict of interest e.g. between stakeholders and promotors/investors that require a less constraining mitigation measure. Some word about it would be welcome.

We now state in the methods that “This study thus directly addresses SEA stakeholders (i.e. decision makers, environmental consultants and conservation managers)” Line143-144. The term “stakeholders” has been systematically changed to “SEA stakeholders” and we describe the range of such stakeholders in the introduction (Line 80-81), methods (143-144) and their role is mentioned in the discussion, where we present how land-use planning staff can use workflow as a step towards the testing of the quality of data and the rationalization of data in order to reduce biases (Line 462-481).

L.28. "was done" please use a more accurate word (I may suggest "performed")

Corrected. “The study used data for nine taxonomic groups” Line 28

L35-36 : I would rephrase for: "Second, the collection of individual databases at the national 35 scale is necessary to complete local data and ensure the suitability of SDMs in a local context"

Corrected. The text is now : “Second, the collection of individual databases at the country scale is necessary to complete local data and ensure the suitability of SDM in a local context” Line 35-37

L50 : You present the biotic homogenization. I also understand the invasive species under this term.

We have specified the term biotic homogenization:
“as well as biotic homogenization, i.e. mostly the extinction of specialist species and the introduction of exotic species, which involves an increase in genetic, taxonomic and functional similarity (Olden and Rooney, 2006; Zambrano et al., 2019).” Line 49-51

Although presented, it underlines one of major lack of the study regarding the SCP. Within each taxonomic group, some species have a particular importance, like the invasive species and the protected/rare/threatened species. As far as I know, stakeholders give more prioritization to areas with rare or threatened species. Conversely, if the biodiversity of an area included a large proportion of invasive species it can mismatch with an adequate choice of SCP. I would suggest two ways to deal with that : 1/ You can redo SDMs taking account the status of the species and/or 2/ you can present the % of observations including invasive or threatened/protected species, therefore if the % of species and the % of the observations are low, it should not have an impact on the global conclusion of your analysis. If they are high, i would highly recommend to perform the 1/.

Thank you for the highlighting this point. There may be such biases in a database for the reasons the reviewer mentions. We have however no reasons to predict a sampling bias, that differs among databases, neither in relation to invasive or threatened species.

The problem and importance of incorporating different dimensions of biodiversity to prioritization spatial his is mentioned in the perspectives " In future studies, species conservation issues for spatial prioritization could be considered by focusing on (for example) the issues associated with threatened and/or invasive species. The multiple dimensions of biodiversity could be analyzed within a context of limited data access and the complementarity of different facets (functional and phylogenetic) in addition to a classical species-based approach (Brumm et al., 2021; Cadotte and Tucker, 2018)” Line 500-505

More importantly and in response to the reviewer’s comments, we provided a new table (cited in the results section for priority areas) in Appendix B.7 which details the list of species of Aves and Papilionidae and their status for each study site (presence status in France and IUCN regional Red List). This list complements the supplementary data already available. As expected, there are very few exotic/invasive species, and thus no difficulties with respect to invasive species presence.

L78-79. I would insist on a potential lack of knowledge to manage such tools

We have included this.
“The databases available for SEA stakeholders (i.e. decision makers, environmental consultants and conservation managers) are often limited because of data sensitivity or ownership issues, although

more and more programs contain data that are publicly available and use of them can be made without any particular attention to their quality (Costello and Wieczorek, 2014; Tittensor et al., 2014) and they are generally unfamiliar to SEA stakeholders." Line 79-84

L81-85-+87 : About the check the suitability of models. A lot of data available was not gathered in order to answer a particular question. Therefore, analysis performed on data that was not designed to may induce several problems. ++ Some protocols are not completely transferable in each context, therefore add some words about the stat assesment seem to me crucial.

We have added a point about the need to evaluate SDM.
"SDM studies generally use data that has not been designed specifically for this type of analysis, and is often comprised of presence-only data, hence the need for a rigorous assessment of sampling biases (Beck et al., 2014; Botella et al., 2018; Guisan et al., 2017)." Line 89-92
"Confidence in the models must be assessed through the use of metrics adapted to the data (Guisan et al., 2017; Leroy et al., 2018)" Line 98-99

L.97-106: I had some difficulties to understand from where to where you are dealing your point. I would suggest to rephrase some sentences to be more precise and to clearly see each point. For each point, please indicate a more concise problematic. L.100: taxonomic groups. Which one are chosen? Why? What hypothesis do you suggest on these tax group?

Clarified.
"The overall goal of this study is to test the influence of different database sources that can be used by SEA stakeholders to map priority conservation areas in SEAs based on SCP. To do so, we studied three local administrative territories that occur in different socio-ecological contexts in France. The study has three main objectives. First, we assess the content of three open- access databases for nine taxonomic groups commonly used in naturalist inventories in environmental assessment studies. We evaluate their suitability in terms of data quantity for SDM application, at three scales (local, regional and country). SDM and SCP analyses were performed for two taxonomic groups (Aves and Papilionidae) to test the hypothesis that sampling bias and differences in ecological response scales of species may influence the identification priority conservation areas. Second, we explore the influence of databases on the application of SDM to assess priority conservation areas. Third, we analyse the influence of this data-driven approach on the composition of species communities that are ultimately used in the identification of priority conservation areas relative to the actual communities in the original databases." L102-113

L.104: A second third point is indicated. Maybe a fourth one?

It was a duplicated sentence, now deleted.

L.110, Figure 1 and Table 1 : there is a mismatch between T2 and T3. Please correct it

Corrected.
"Figure 1 - Localization of the study sites in French administrative regions: T1 is Lodévois-Larzac, T2 is La Rochelle, T3 is Brocéliande. Source: IGN, Google, 2023." and in first column of Table 1 : "Study site | Lodévois-Larzac (T1) | La Rochelle (T2) | Brocéliande (T3)"

Table 1 : Urbanization : A quick check on a satellite map show me that the T2 area is surrounded by two big cities (La Rochelle and Niort), whereas the T1 area seems quite far from an urban area. Is the city Lodève? Millau or Montpellier? [*] I also found an highway (N11 and a main railroad between the two main cities), whereas the fragmentation context as you describe it seems higher in T1 and T3.

We give the coordinates of main city to identify which city it is (for main city to T1 is Lodève). We specified the land-use category "Artificial" by "Artificial-urban areas" and "Urbanization" by "Urbanization context". In "Urbanization context", we specified name of major cities and we clarified the "across by a highway" with "linked by a highway".

[*] Dou you have any hypothesis on the urbanization context of each zone, since it is highly detailed, but not discussed

"Finally, despite important differences among the study sites in terms of the proportion of artificial land cover and protected areas we found no particular differences between the three study sites. Clearly, the data sources are the most important factor influencing the results" Line 396-398

L.152: for test -> to test

Corrected.
"(i.e. to test the effect of database sources on SDM performance)" Line 166

L.156: mismatch between the period and the time span you present. I think you should indicate "11 years"

Corrected.
"The databases were collected over a period of 11 years (i.e. from 01/01/2010 to 31/12/2020)" Line 168-169 and in Appendix A Line 882

L.158: Which ref did you choose? TaxRef? Did you use infra/supra specific taxa?

It is described in appendix A.1 and we added it to the text.
"We made a series of operations to standardize, correct and homogenize taxa names at the specific taxonomic level using the French taxonomic reference "TAXREF.V14" (Gargominy et al., 2021)." Line 170-172

L.172: All databases, you mean all 3 databases combined, right?

Yes, we standardized the terminology using quote "All databases".

L.180-182: IT should be more explained as hypothesis, therefore maybe a list of species may be useful? Does it include migrating species? if not, please move or discuss it in discussion.

As mentioned in section 2.3.1, Aves selected are only "nesting Aves". We have provided the list of species in a supplementary table (Appendix B7).

L185 : Do you mean a different dispersal zone

We do not understand this remark. We have supposed that the reviewer is referring to line 180-182
"The use of these two groups allows for a comparison between one group of highly mobile taxa with a large home range (Aves) and another group with a smaller home range and whose movement closely tracks local environmental variation (Papilionidae)." We have now added :
"These two taxonomic groups thus have different biological traits associated with their dispersal and function, hence we predict differences in in terms of the spatial resolution of their distribution." Line 198-200

We trust this is sufficient.

L.189 : How do you deal with the buffer zone in the sea?

For data collection only terrestrial data were considered and for SDM calibration the sea were not integrated in the models. Nevertheless, for T2, part of the "Island of Ré" was included in the analyses, as it falls within the 10km buffer zone. The extent of study area influences priorities, so as mentioned in section 2.4.2, we restricted to administrative boundaries with a buffer zone of 1km that exclude "Island of Ré" to prioritization analysis.

L.174-182: If you discuss about the migrating species, I am wondering why did you not include plants as 'not migratory' (at least at individual level), and bird as migratory. They have also a higher observational data, which could improve the SDM. Moreover, Papilionidae distribution is highly dependant on plant distribution as hosts.

In section 2.4 we focused on Aves and Papilionidae, and as mentioned in section 2.3.1 we focus on “nesting aves” which includes migratory species. Using plants would be highly complicated by the large number of species and a totally different mode of dispersal. We feel that the use of these two groups with different dispersal characteristics, biological traits and expectations in terms of spatial resolution of their responses provides a sufficient test for studies of the fauna.

L.198: I find it very discutable since the observer of a point is maybe not able to identify other taxa. (I am wondering if the bird data is often more accurate and with a better sampling effort than plants or insects) Therefore i think that a large inter-taxa identification variation occurs.

In this part, the method to generate pseudo-absences for data protocoled (i.e. Vigie Nature) is presented. As mentioned in section 2.3.1, for Vigie Nature, "homogeneity in identification criteria and compliance with the protocol are ensured by offering training to volunteers". There is one program by taxonomic group, thus we hypothesize that there is no problem of identification within the same group. Thus, the generation of pseudo-absences was carried out separately for each taxonomic group, as we specified more clearly in the text.
“For model calibration, pseudo-absences were generated with two methods, separately for each taxonomic group.” Line 213-214

L.216,L.224: Please do not repeat the value of the threshold.

Corrected.
“The features were only the good quality models defined previously.” Line 243

L.221. Please rephrase to : "The aim of SEA biodiversity conservation strategies.."

Corrected.
“The aim of SEA biodiversity conservation strategies is to establish priorities for the whole territory and all the cells have the same cost value of 1.” Line 240-241

L.227,L.229 please change for active voice

Corrected.
“We used the package “prioritizr” (Hanson et al., 2021) with the open-source solver SYMPHONY (Kim et al., 2023).” Line 245-246
“We analysed the influence of database sources on SDM predictions and priority conservation areas (**Erreur ! Source du renvoi introuvable.**, step 2.2 and 3.2).” Line 248-249

L.236. Does the original community included deleted species? ex Nobs<15?

Yes, we specified this in the text.
“We assessed the influence of the complete data driven workflow on the composition of species communities, i.e. differences between the original community (i.e. all species observed in study site) in the database and the final community used to identify priority conservation areas (Figure 2, step 4).” Line 254-256

L.251: Please mention directly that Aves follows a Cauchy distribution

Corrected.
“We adapted the link function to the data distribution, using a “cauchit” link for Aves and a “logit” link for Papilionidae.” Line 270-271

L.254: How did you specify the null model

As specified in the available scripts, we performed a classic null model (Residence ~ 1) with the link function corresponding to the taxon.

L.258: Should not it be written as (1| Studysite + Database) ?

“Studysite_Database” is a variable combining the Study site and Database, we change by "/" for clarity.

L.274: I am wondering if a comparison with another contry would be complementary (eg. comparing a "european database")

It is a very interesting suggestion that we added to the discussion. In France, there is a real conflict of ownership of data and the data context is probably different in other European countries. To our knowledge, no observation database exists on a European scale, and GBIF accomplishes its role on a global scale. On the other hand, we believe that complementary analysis would be of greater interest for the various country databases. We trust our study may stimulate such studies in other countries.

L.283 : I think that the number of species observed comparatively to toal species number would be useful.

This is in the table 2 and for more details see appendix B.1.

L.325 : Did you interpret some beahviour or group like moth and butterflies? Could we expect different conclusions between these two groups?

We only studied the traits mentioned in this study and we did not have behavioral traits. We observed a sampling bias that likely influences other taxonomic groups in the same way.

L.374 : What are the advantages and disadvantages of the use of expert data vs the use of non-expert data, and the implication of sharing data with the scientific communiy and also with the public
L405-406 : (in regard to my comment on L.374). Could you provide some discussion about an exepcted model performance. ie discuss about the performance of the model about only but numerous opportunist data? To complete though it is partly discussed L.423. Maybe consider reordering?

We have included discussion of this issue.
“Furthermore, there is a dilemma between protocolized and opportunistic data. Although protocolized data are recommended for SDM (Guisan et al., 2017; Guillera-Arroita et al., 2015), very often the amount of such data is low, which can be detrimental at the local scale, particularly for model evaluation with data having the same sampling bias. For opportunistic data, their large number is of course a positive point, however the estimating their sampling bias can be a real challenge (Botella et al., 2018; Fithian et al., 2015; Matutini et al., 2021) to ensure the reliability of the results.” Line 420-4425

L.396-397: Please rephrase the sentence as it is not very understandable.

Corrected.
“Indeed, the high overlap in species distribution between data sources, as indicated by Schoener's D index, indicates that, regardless of data source, species are predicted in similar environments (Warren et al., 2008)” Line 427-429

L.402-404 : This sentence seems to generalize the conclusions though I understand that this fact can be only extended to certain conditions (taxa, data...)

We are not sure what the reviewer asks us to do.

L.409: ... therefore how to deal with absences (or pseudo-absence)

Our suggestion is just above “Models using opportunist data with a target-group approach to generate pseudo-absences provides a sufficient quality of information on species distribution (Phillips et al., 2009; Barber et al., 2022) and can be correctly used in SCP (Sofaer et al., 2019; Baker et al., 2021)” Line 434-437. In this sentence, we discuss possible reasons for our contradictory results with Hermoso et al. 2015.

L.414-428: I am very curious about the consequences of your study to define conservation priorities regarding protected, rare or threatened species. Since these areas are the one to prioritize.

We have added this perspective, as mentioned in our response to reviewer 2, line 50.

L.427-428: Opportunistic data tends to provide more observation of rare species. Indeed, people (naturalists) focus more on target rare or beautiful species whereas common species, though often observed, are not always reported in databases. I would be very please to have a short discussion about more biases like this one

This is mentioned in the perspectives. “Furthermore, their use is particularly interesting to help strategically direct inventory campaigns (especially for under sampled taxa and areas) that go beyond the emphasis on rare, threatened and emblematic species.” Line 487-489

L.433-434: I think this is one of the most important point, with the hardest remaining issue is found l.439 about sharing data. A deeper discussion about this issue and even some solutions to deal with it could be useful.

We added suggestions of solutions.
“To overcome this data sharing problem, the structuring of networks of different contributors of data and users of the databases and ambitious regional policies is necessary.” Line472-473

L.460: I would include a discussion/conclusion about the integration of site managers that can provide some tools about the suitability. However, these tools are not systematically shared, or are often published elsewhere. Solutions to deal with that should be underlined.

We added a discussion about that in perspective.
“The integration of local experts may help limit any misjudgements in the workflow procedure. Indeed, the integration of "expert" knowledge and local studies is valuable information, which is important to share, and which it is important to consider in order to complete our proposal.” Line 498-500

Appendix A : Some errors should be corrected, the same as explained above about the site names (T2, T3), and the time span (11 years). Some helpful information is provided in this appendix that, I think, should be move into the main text. (LL. 819,825, 828,829, 831)

This has now been added in section 2.3.1, and the entire process is detailed in appendix A.1.

In a general way, I can suggest to clearly detail the hypothesis and the expectations at the end of the introduction. The discussions could also be improve detailing some unused information (about the site and the species). Some concepts are not discussed and it could very interesting to give further details and discussion.

I look forward this paper soon published.

We thank the reviewers very much for their help and the relevance of their comments, which we believe have significantly improved the manuscript.