nature portfolio

Peer Review File

Pesticides have negative effects on non-target organisms

Corresponding Author: Professor Nian-Feng Wan

This manuscript has been previously reviewed at another journal. This document only contains information relating to versions considered at Nature Communications. Mentions of the other journal have been redacted.

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The authors have put significant effort into revising their work, which is substantially improved now. However, there are some remaining concerns.

1. The GitHub repor now contains the code but not the data to re-run the code. Also, there are multiple R files and no description how to use them to recreate study results – this should be described in detail in the README file. Meta-data should be also provided for all used variables.
2. Poor coverage of grey and non-English literature should be acknowledged as limitations in the main manuscript.
3. Trim-and-fill is a publication bias test not sensitivity analysis.
4. The issue of percentage and proportion data is not in the normality of the residuals, but in the fact that they are bounded between 0 and 1 (100). As such they should be arcsin-transformed to make them unbounded. All transformations can be documented in R code, so there is no issue of transparency.
5. The issue with folded normal distribution when using absolute values has not been resolved. See Nakagawa, S. & Lagisz, M. (2016) Visualizing unbiased and biased unweighted meta-analyses. 334. Journal of Evolutionary Biology, 29, 1914-1916, for a visual clarification of the issue.
6. A synthetic phylogenetic tree can be retrieved from Open Tree of Life, which is accessible via R package rotl. As, such phylogenetic analyses can and should be conducted beyond plants.
7. Publication year should be included in the analyses (meta-regression) – it is not confounded with study identity, and it is potentially important and interesting moderator of the effect.
8. Funding and conflict of interest statements can be easily extracted from the included papers, even if they are not consistently included. Then, presence of potential links with the relevant industry should be analysed in a meta-regression. This is a too important issue to dismiss.

Reviewer #3

(Remarks to the Author)
The current ms derives from a previous version submitted to [Redacted] that has been commented by three reviewers. I now have checked for the implementation of the comments of Reviewers 2 and 3 in the revised version.

To my opinion, almost all comments of Reviewer 2 have been adequately addressed, except for

(1) Comment: 216-217 were vertebrates and invertebrates equally affected, or one more than the other? Does that depend on the intended target of an pesticide, e.g. distinguishing between pesticides aimed at invertebrates or otherwise?

Response: We have removed the sentence due to space limited, but from the statistical values in Supplementary Table 4, we can see that vertebrates and invertebrates were equally affected (see Supplementary Table 4).

and

(2) Comment: 223 ‘affected different taxonomic groups’ how?
Response: We have removed the sentence due to space limited, but from the statistical values in Supplementary Tables 32–36, we can see that “insecticides, fungicides and herbicides affected different taxonomic groups with decreased growth, reproduction or behavior and with perturbed biomarkers” (see Supplementary Tables
32–36).

I feel that both informations are important and should be kept within the main body of text rather than being implied to be taken from the supplementary material.

All comments of Reviewer 3 have been perfectly addressed in this version of the ms.

In the Competing Interests paragraph, "B.W." should be replaced by "B.A.W.".

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Reviewer comments:
The authors addressed my concerns partially. However some new concerns emerged in the process.

Response to the referee comments
Replies to Reviewer #1:
Comment: The authors have put significant effort into revising their work, which is substantially improved now.
Response: We thank the reviewer for this positive evaluation (see this revision).

Comment: However, there are some remaining concerns. 1. The GitHub report now contains the code but not the data to re-run the code.
Response: We have uploaded both code and data to GitHub (https://github.com/Liwan-Fu/Impact-of-pesticides).
New reviewer comment - Round3: Thank you, for providing a link to the files on GitHub. I note that now the data and code are also provided on Zenodo and the contents of these two repositories do not match. Specifically, some code files that are archived on Zenodo are absent from GitHub version. Please remove any special (Chinese) characters from your R code.

Comment: Also, there are multiple R files and no description how to use them to recreate study results-this should be described in detail in the README file. Meta-data should be also provided for all used variables.
Response: We have provided a README file for all used variables and uploaded the file to GitHub (https://github.com/Liwan-Fu/Impact-of-pesticides).
New reviewer comment - Round3: Thank you for providing a README file. Unfortunately, the meta-data provided applies to only one of the data files, “Meta combined data.csv” (there seems to be 16 data files in total). The meta-data for this one key data table used to run the key models is still too rudimentary to allow cross-checking of the data or replication of the data extraction process. For example, the variable described as “Control (value): the value in control group;” does not explain which value should extracted be there, apart that it should relate to the control group. Another, example – “Control-n: the number in control group;”, which is supposed to be the sample size, it also needs explaining the units of counting sample size, as these can be different for different taxa and experiment types – for animals it might be number of individuals, for plants could be individuals or plots, for microorganisms it could be a colony in a tube/plate, or number of natural plots, etc. One column in the data table called “LnR” is not described – it seems to be an effect size that was not be used in the analyses, as the main effect size used was SMD. However, I cannot find the code the authors used for calculating SMD. Also, I cannot find the code for Figure1 (the map), and instead there is code for Figure 1b, a histogram which is not in the main manuscript. Further, the study codes in the main data file “Meta combined data.csv” do not match the study numbers in the table characterizing included studies (with their references) – “447069_2_data_set_8320286_s47vd1.xls”. Please include full study references as a column in the data file “Meta combined data.csv”, so that each data point can be linked to its original source. Overall, lack of consistent and comprehensive documentation is still a major limitation to transparency and reproducibility of this meta-analysis, which should be addressed before this work is accepted.

Comment: 2. Poor coverage of grey and non-English literature should be acknowledged as limitations in the main manuscript.
Response: We have now made it explicit that our study did not include grey or non-English literature (lines 420-421).
New reviewer comment - Round3: I can see it in the methods section now. I suggest adding this also to the header of Figure1 (the global) map, because it is highly relevant to the pattern presented there.

Comment: 3. Trim-and-fill is a publication bias test not sensitivity analysis.
Response: We have removed text referring to this as a sensitivity analysis (see lines 631-632 in the main text; line 326 in the Supplementary Results).
New reviewer comment - Round3: thank you.

Comment: 4. The issue of percentage and proportion data is not in the normality of the residuals, but in the fact that they are bounded between 0 and 1 (100). As such they should be arcsin-transformed to make them unbounded. All transformations can be documented in R code, so there is no issue of transparency.
Response: We have consulted with our institutional statistician (Dr Pete Henrys - CStat, Royal Statistical Society Chartered statistician, CSci, The Science Council Chartered Scientist) who states that a transformation is there to ensure model assumptions are met. Transformation purely on the basis that data is of a particular type is not appropriate, and rather transformation should only be applied on a case by case basis to address underlying issues that may affect model assumptions. He confirms that focusing on model assumption checks (as we have done) including assessing residuals represents the critical stage in the process for assessing the need for underlying processing of input data. We strongly argue that percentage and proportion data in the context of this analysis does not warrant to be arcsin-transformed without a reason relating to model distributional fit. In the context of this analysis this is not required.
New c reviewer omment - Round3: it is great that you consulted the statistician and conducted the model assumptions checks. Please provide the results of the model assumptions test in the supplementary materials. Conducting sensitivity analysis for the main model that show that applying the recommended transformation does not change the results should be also presented in the supplementary materials to support your claims.

Comment: 5. The issue with folded normal distribution when using absolute values has not been resolved. See Nakagawa, S. & Lagisz, M. (2016) Visualizing unbiased and biased unweighted meta-analyses. 334. Journal of Evolutionary Biology, 29, 1914-1916, for a visual clarification of the issue.
Response: This relates to our standardizing of biomarkers (e.g. protein regulation, gene expression) which may respond positively or negatively to pesticides. Our approach has been to use absolute values and focus on a deviation from zero in the meta-analysis, essentially quantifying departure from the norm. Again, we emphasize that checks of model assumptions suggest that the current use of a normal distribution was robust. From a practical perspective it is not possible within our modeling framework to specify a folded distribution (which we do not think is required). However, we would be prepared to simply no longer use the absolute values thus negating the suggested need for a folded distribution. We do feel that this would potentially lose important information about a general impact of pesticides on either up- or down-regulation of biomarkers - which we think is hugely significant from the context of the impacts of pesticides on non-obvious species level effects. The author team has discussed this issue extensively and we still believe that taking absolute values biologically makes more sense here.
New reviewer comment - Round3: Please provide the results of the model assumptions test in the supplementary materials. Conducting sensitivity analysis for the key model that show that applying the recommended transformation does not change the results for biomarkers data subset should be also presented in the supplementary materials to support your claim. Also, please flip the sign on the absolute values to make it easier to compare magnitude of estimated effects in all forest plots.

Comment: 6. A synthetic phylogenetic tree can be retrieved from Open Tree of Life, which is accessible via R package rotl. As, such phylogenetic analyses can and should be conducted beyond plants.
Response: We have used rotl, a freely available package for R designed to takes advantage of the Open Tree of Life's Application Programming Interfaces (APIs) to access subtrees from the synthetic Open Tree, as well as the published source trees that contribute to the synthesis to generate comprehensive phylogenies for animals and microorganisms (see lines 296-321 in the Supplementary Methods). In this revision, we have now conducted all the phylogenetic analyses for plants, animals and microorganisms (see detailed results in Supplementary Tables 1.3-1.5, 2.3-2.5, 45.3-45.5 and 46.3-46.5). We still note that an overall agreed-upon phylogeny for the tree of life will likely never arrive, but we did our best to address the referee´s comment.
New reviewer comment - Round3: Thank you and I agree regarding the conditionality of any phylogeny used – but we will never arrive there.

Comment: 7. Publication year should be included in the analyses (meta-regression)-it is not confounded with study identity, and it is potentially important and interesting moderator of the effect.
Response: We have clarified this now as “Publication year: a continuous metric according to the year when the articles were published” (lines 528-529 in the main text). We have included publication year in the meta-regression model (see detailed results in Supplementary Tables 1.1-1.5, 2.1-2.5, 45.1-45.5 and 46.1-46.5; Extended Data Figs. 2-7 and other associated supplementary tables).
New reviewer comment - Round3: thank you.

Comment: 8. Funding and conflict of interest statements can be easily extracted from the included papers, even if they are not consistently included. Then, presence of potential links with the relevant industry should be analysed in a meta-regression. This is a too important issue to dismiss.
Response: We now extracted conflict-of-interest statements where possible. Among the 1705 papers, we could confirm that 1,411 ones did not have conflict of interest while 25 self-identified as having a conflict of interest. The remainder made no statement about conflicts of interest (lines 244-248 in the main text; Supplementary Table 58). In the meta-regression analysis, we consider “Conflict of interest status” as one binary variable (i.e., “1” denotes that a certain article has conflict of interest and “0” represents no conflict) (lines 526-528 in the main text). Detailed results for “Conflict of interest status” were presented in Supplementary Tables 1.1-1.5, 2.1-2.5, 45.1-45.5, 46.1-46.5 and 58, Extended Data Figs. 2-7 and other associated supplementary tables).

New comment - Round3: thank you for doing this.

New comment - Round3:
By looking at the raw data, “Meta combined data.csv”, I can see “Life expectancy” and “Longevity” are in “insecticide-animal reproduction” category. It seems like all (and there are many) measures related to survival are classified as “insecticide-animal reproduction” (similar issue is also present for plants and microorganisms in the data set). Please provide justification for this choice. Also animal feeding rates could represent both growth and behavior – having them solely as a measure of growth may require additional sensitivity analyses and clarification in the methods section.

New reviewer comment - Round3:
In the “Meta combined data.csv”, there are some clear mistakes too, e.g. the column for describing measurements is named “Animal growth indicator”, and sometimes is empty, “nummber of eggs/female” is classified as “insecticide-animal behavior”, this raises concerns for the consistency of data extractions and the quality control procedures in this work.

New reviewer comment - Round3:
Lines 463-466: “When a study included different levels of pesticide application rates, measurements for the control groups without pesticides versus different pesticide application rates were considered as independent paired observations” - This is a very bold assumption given “26,096 estimates of pesticide effects reported from 1,705”. Now looking at the actual data, I can see that although sometimes multiple pesticides or species were used in a single study, in most cases multiple doses were used, which means data non-independence related to the repeated use of the same control group to compare it with different dose levels is a serious issue. Specifically, by not controlling for repeated use of the same individuals, the precision of the overall estimate is very likely inflated. One very simple way is to divide sample size of the control group by the number of comparisons it is used in, thus avoiding “double-counting” of the same replicated units.

New reviewer comment - Round3:
In the main text, please acknowledge study limitations due to sample sizes in the included experiments (from the raw data I can see that the median sample size is likely 3) – having a histogram of the distribution of sample sizes by type of organism and/or treatment would be very useful too.

Version 3:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Reviewer comments:
The authors addressed some of my concerns but there are still outstanding issues related to transparency and reproducibility of this work, as well as some methodological concerns.

My previous comments and authors responses are quoted below when they are linked to new issues identified. My new comments are provided as “New reviewer comment – Round4:”

“New reviewer comment - Round3: Thank you, for providing a link to the files on GitHub. I note that now the data and code are also provided on Zenodo and the contents of these two repositories do not match. Specifically, some code files that are archived on Zenodo are absent from GitHub version. Please remove any special (Chinese) characters from your R code.
Response: We thank the referee for pointing us at this. To avoid confusion and duplication we have now uploaded all raw data and code to Zenodo and removed them from GitHub (see this revision). We have checked that all relevant files are present. We have now also removed all Chinese characters as suggested. All data used in this analysis are available on Zenodo
(https://zenodo.org/records/10495263). All code for this analysis is available on both Zenodo
(https://zenodo.org/records/10509420).”
New reviewer comment – Round4: Thank you for providing your data and code on Zenodo. It seems like putting the contents of these two repositories together still does not allow reproducing all the analyses and they are really hard to peer-review. There are three reasons: first the code is very poorly annotated, and thus it is hard to figure out what different sections of code are supposed to produce. Second, it seems like pieces of code are missing, e.g. file “Figures 2-3.R” starts from loading the file 'Meta combined data.csv' and then immediately runds a meta analytic model on the data from it, however the file does not contain effect sizes and there is no line of code that calculates effect sizes before running the model. Third, the code refers to loading many additional data files, that look like pre-processed data, but they are not included in the Zenodo repository, e.g. “Supplementary Tables 1(1-3).R” calls “Old_New_pesticide <-
read_excel(paste("/lustre/huyueqing/ssy/Wan2023/DataPre/","Supplementary Data 2-Classification of old and new pesticides-2023-0825.xls",sep = ""),sheet = 1)”, but there is no such file or no explanation where it originates from. Overall, I think all the analytic pipeline should be made fully transparent reproducible and then subjected to independent peer-review.

“New reviewer comment - Round3: Thank you for providing a README file. Unfortunately, the meta-data provided applies to only one of the data files, “Meta combined data.csv” (there seems to be 16 data files in total). The meta-data for this one key data table used to run the key models is still too rudimentary to allow cross-checking of the data or replication of the data extraction process. For example, the variable described as “Control (value): the value in control group;” does not explain

which value should extracted be there, apart that it should relate to the control group. Another, example – “Control-n: the number in control group;”, which is supposed to be the sample size, it also needs explaining the units of counting sample size, as these can be different for different taxa and experiment types – for animals it might be number of individuals, for plants could be individuals or plots, for microorganisms it could be a colony in a tube/plate, or number of natural plots, etc.
Response: We have now revised the README file to improve clarity. We hope this addresses the problems.”
New reviewer comment - Round 4: Thank you. The README file only describes data in the main data file, but not in any of the additional data files and does not provide any information about the project in general and how the code and files should be used to reproduce the analyses. Please provide more comprehensive description of all your files, starting from overview and then providing meta-data for all data that is needed to reproduce your work. Consider using Quarto with R to produce computationally reproducible documents that contain the R code, verbal descriptions, and outputs.

“Comment: One column in the data table called “LnR” is not described – it seems to be an effect size that was not be used in the analyses, as the main effect size used was SMD. However, I cannot find the code the authors used for calculating SMD.
Response: The column in the data table called “LnR” had not been used in the analyses, and we have now deleted it. The code used for calculating SMD has been added in the README file.”
New reviewer comment - Round 4: The code used for calculating SMD should not be placed in the README file, but in the code where it is needed to calculate effect sizes to reproduce the results.

“Comment: Further, the study codes in the main data file “Meta combined data.csv” do not match the study numbers in the table characterizing included studies (with their references) – “447069_2_data_set_8320286_s47vd1.xls”. Please include full study references as a column in the data file “Meta combined data.csv”, so that each data point can be linked to its original source.
Overall, lack of consistent and comprehensive documentation is still a major limitation to transparency and reproducibility of this meta-analysis, which should be addressed before this work is accepted.
Response: We thank the referee for pointing us at this. We have now made sure that all study codes match up. We have listed all the 1705 study references in the file “Meta combined data” with a new sheet (Full study references). The paper code order in “Meta combined data” is based on the time order that we extracted data, and the paper order in Supplementary Data 1 is based on the authors' name in alphabetical order (see Meta combined data). In addition, we have added another data file “ExtendData” for other supplementary files (e.g., for Supplementary Tables 59-61, Supplementary Figures 24-27).”
New reviewer comment - Round 4: Thank you for adding the study references next to study codes in the data file. I was now able to compare the extracted data with the information reported in the original papers. I randomly selected 4 papers I have full-text access to and found the following issues:

- Paper 16. (Colin et al. 2004. A Method to Quantify and Analyze the Foraging Activity of Honey Bees: Relevance to the Sublethal Effects Induced by Systemic Insecticides): rows 16-21: the concentration of Fipronil is wrong - the study reports using 2 micrograms per kilogram while the extracted data is 6 g/kg. For Imidacloprid the unit is also wrong – it should be microgram per kilogram not gram per kilogram. For the outcome “number of daily attendance of active bees-nuclei ” which was extracted for nuclei A, B, C, it is and somehow from control nuclei, but I cannot find the matching exact numbers in the text, and it is not possible to extract them from the figures in the text. Given the complexity of the experimental design, raw data would need to be used to calculate such values. There are no comments in the data file how the numbers were obtained, and thus hey are not reproducible.
- Paper 1628. (Zhao X, Li YL, Zhang LX, Dorna H, 2003. Effects of priming and fungicide treatment on germination of china aster (Callistephus chinensis L.) seeds. Seed Science & Technology, 32: 451-457.): rows 1943-1944: it looks like only germination capacity data from Sample 1 and only from 30C temperatire treatement were extracted and the standard deviations were manually imputed by assuming they are 10% of the measured value. This is only a small fraction of available data in this paper (multiple control and exposure groups, all samples should be extracted and used for calculating mean and SD) and is incorrectly extracted.
- Paper 1008. (Gabaston J, Khawand TE, Wafo-Teguo P, Decendit A, Richard T, Mérillon JM, Pavela R, 2018. Stilbenes from grapevine root: a promising natural insecticide against Leptinotarsa decemlineata. Journal of Pest Science, 91: 897-906.) rows 5808-5809: looks like data was extracted from Table 6 for 3rd day of exposure only and two lower doses out of 3 used (this is not reported on the data extraction sheet). The mean mortality % was converted to survival %, but the reported standard deviation values were somehow replaced with 10% of the men value. The sample size is claimed to be 4, but I cannot find such information in the article, and it seems improbable given the typeof the organism, and reported values.

Overall, this small exercise clearly shows that 3/3 papers had some issues with data extractions. This indicates a very high rate of errors in the whole data set. The data lacks information on the exact sources of the extracted values for the outcomes at least (table/figure number, text page, raw data, etc.), any extractions decisions made, calculations, imputations, etc. – this makes assessing trustworthiness of the data set very hard and makes data extractions irreproducible. Because of this I request that all extracted data is annotated in detail. Further it has to be cross-checked by a researcher who was not involved in original extractions, re-extracted where issues are found and clearly described to make it fully transparent and trustworthy.

New reviewer comment - Round 4: When checking the extracted data, I noticed that many of the extracted standard deviation values are exactly 10% of the mean value. This cannot be coincidence and the manuscript does not mention doing manual imputations for missing standard deviation values (or any other missing values). Please explain how these values originated and the detailed procedures used to deal with any missing data.

“New reviewer comment - Round 3: it is great that you consulted the statistician and conducted the model assumptions checks. Please provide the results of the model assumptions test in the supplementary materials. Conducting sensitivity analysis for the main model that show that applying the recommended transformation does not change the results should be also presented in the supplementary materials to support your claims.
Response: As suggested, we now include a sensitivity analysis where we apply arcsine transformation for percentage and proportion data (as you suggest) and compare the results to our approach (no transformation or proportion/percentage data) (see Supplementary Table 59). This shows no qualitative difference in the results. We also provide a supplementary figure in the supporting information for the model checks of the original analysis (see Supplementary Fig. 24).
Original comment for context: 5. The issue with folded normal distribution when using absolute values has not been resolved. See Nakagawa, S. & Lagisz, M. (2016) Visualizing unbiased and biased unweighted meta-analyses. 334. Journal of Evolutionary Biology, 29, 1914-1916, for a visual clarification of the issue.”
New reviewer comment - Round 4:
The code for reproducing this should be in the file “Supplementary Tables 59-61.R”, and as such cannot be rerun. In the provided code I cannot find any line that would execute arcsine transformation for proportion data. And the code starts from loading a mysterious data file that is not archived on Zenodo, so cant tell what is there nor how it got there: AllData <- read.table(paste(DataPre_path,"SuppTab59Data.txt",sep=""),header = T). This seems to be ongoing issues across the provided code which is compounded by almost complete lack of code comments and descriptions.

“New reviewer comment - Round 3: Lines 463-466: “When a study included different levels of pesticide application rates, measurements for the control groups without pesticides versus different pesticide application rates were considered as independent paired observations” - This is a very bold assumption given “26,096 estimates of pesticide effects reported from 1,705”. Now looking at the actual data, I can see that although sometimes multiple pesticides or species were used in a single study, in most cases multiple doses were used, which means data non-independence related to the repeated use of the same control group to compare it with different dose levels is a serious issue.5 Specifically, by not controlling for repeated use of the same individuals, the precision of the overall estimate is very likely inflated. One very simple way is to divide sample size of the control group by the number of comparisons it is used in, thus avoiding “double-counting” of the same replicated units.
Response: Observations with no pesticide had been considered as control groups, while those with different pesticide application rates were considered as the treatment groups. When an article included different levels of pesticide application rates, measurements for the control groups (no pesticide) were compared to all other treatment levels of pesticide application rates and treated as independent paired observations. Within the context of these analysis, this reflects typical experimental designs for studies considered in ecotoxicology that have a single control which is then compared to multiple test treatment residue levels. The random allocation of treatments to replicates, combined with study as a random effect in the meta-analytical models, means that these comparisons are appropriate and do not represent double accounting as suggested. In response to the referee´s comment, we have now re-run our analyses with the correction suggested, i.e. dividing the sample size of the control group by the number of comparisons it had been used in.
The results of this additional analysis can now be found in a new Supplementary Table 62. When comparing these results to our former approach (see Supplementary Table 3), there were no qualitative differences in the results (see Supplementary Table 3 vs. Supplementary Table 62). In particular, all negative effect sizes remained negative (and significant), and all positive effect sizes remained positive and significant.”
New reviewer comment - Round 4: The authors are wrong when stating that “The random allocation of treatments to replicates, combined with study as a random effect in the meta-analytical models, means that these comparisons are appropriate and do not represent double accounting as suggested.” The measures from the same control individual are used repeatedly in comparisons of effects of different doses of the same pesticides and they are such effect sizes are not independent and are a form of ”double-counting”. The analyses that account for such non-independence should be used as main models not as supplementary. The author used sample-size corrections their sensitivity analyses for a few models. I suggest testing their all models using cluster-rubust approach which is already implemented in metafor package (https://wviechtb.github.io/metafor/reference/robust.html) as their main approach. Please also run and report an overall (global) meta-analytic model on all data with organism, trait and pesticide type as random effects.

“New reviewer comment - Round 3: In the main text, please acknowledge study limitations due to sample sizes in the included experiments (from the raw data I can see that the median sample size is likely 3) – having a histogram of the distribution of sample sizes by type of organism and/or treatment would be very useful too.
Response: The median value of sample size was 4 and 4 for the control and treatment, respectively (see lines 122-123 in the main text). In addition, we have added a histogram to show the distribution of sample sizes by types of organisms and separately by control and treatment (see Supplementary Fig. 27).”
New reviewer comment - Round 4: Thank you for adding information about the median sample sizes. However, this raises new issue since Line 562 states “A large-sample approximation was used to compute the sampling variances46.” This approximation is not suitable for such small sample sizes, and a small-size-corrected version of the SMD (Hedges’ g) should be used as an effect size in this meta-analysis.

(Remarks on code availability)
code incomplete and poorly annotated

Version 4:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Thank you for your patience and hard work on this manuscript. I am pleased to note that it is much more transparent and robust now (it seems to have some new interesting findings as well). I really hope that in your next meta-analytic project you will take all steps from its inception to make sure it is trustworthy and well-documented, so I dont need to do 5 rounds of review again... good luck

Response to the referee comments

Replies to Reviewer #1:

Comment: The authors have put significant effort into revising their work, which is substantially improved now.

Response: We thank the reviewer for this positive evaluation (see this revision).

Comment: However, there are some remaining concerns. 1. The GitHub report now contains the code but not the data to re-run the code.

Response: We have uploaded both code and data to GitHub (https://github.com/Liwan-Fu/Impact-of-pesticides).

Comment: Also, there are multiple R files and no description how to use them to recreate study results-this should be described in detail in the README file. Meta-data should be also provided for all used variables.

Response: We have provided a README file for all used variables and uploaded the file to GitHub (https://github.com/Liwan-Fu/Impact-of-pesticides).

Comment: 2. Poor coverage of grey and non-English literature should be acknowledged as limitations in the main manuscript.

Response: We have now made it explicit that our study did not include grey or non-English literature (lines 420-421).

Comment: 3. Trim-and-fill is a publication bias test not sensitivity analysis.

Response: We have removed text referring to this as a sensitivity analysis (see lines 631-632 in the main text; line 326 in the Supplementary Results).

Comment: 4. The issue of percentage and proportion data is not in the normality of the residuals, but in the fact that they are bounded between 0 and 1 (100). As such they should be arcsin-transformed to make them unbounded. All transformations can be documented in R code, so there is no issue of transparency.

Response: We have consulted with our institutional statistician (Dr Pete Henrys - CStat, Royal Statistical Society Chartered statistician, CSci, The Science Council Chartered Scientist) who states that a transformation is there to ensure model assumptions are met. Transformation purely on the basis that data is of a particular type is not appropriate, and rather transformation should only be applied on a case by case basis to address underlying issues that may affect model assumptions. He confirms that focusing on model assumption checks (as we have done) including assessing residuals represents the critical stage in the process for assessing the need for underlying processing of input data. We strongly argue that percentage and proportion data in the context of this analysis does not

warrant to be arcsin-transformed without a reason relating to model distributional fit. In the context of this analysis this is not required.

Comment: 5. The issue with folded normal distribution when using absolute values has not been resolved. See Nakagawa, S. & Lagisz, M. (2016) Visualizing unbiased and biased unweighted meta-analyses. 334. Journal of Evolutionary Biology, 29, 1914-1916, for a visual clarification of the issue.
Response: This relates to our standardizing of biomarkers (e.g. protein regulation, gene expression) which may respond positively or negatively to pesticides. Our approach has been to use absolute values and focus on a deviation from zero in the meta-analysis, essentially quantifying departure from the norm. Again, we emphasize that checks of model assumptions suggest that the current use of a normal distribution was robust. From a practical perspective it is not possible within our modeling framework to specify a folded distribution (which we do not think is required). However, we would be prepared to simply no longer use the absolute values thus negating the suggested need for a folded distribution. We do feel that this would potentially lose important information about a general impact of pesticides on either up- or down-regulation of biomarkers - which we think is hugely significant from the context of the impacts of pesticides on non-obvious species level effects. The author team has discussed this issue extensively and we still believe that taking absolute values biologically makes more sense here.

Comment: 6. A synthetic phylogenetic tree can be retrieved from Open Tree of Life, which is accessible via R package rotl. As, such phylogenetic analyses can and should be conducted beyond plants.
Response: We have used rotl, a freely available package for R designed to takes advantage of the Open Tree of Life's Application Programming Interfaces (APIs) to access subtrees from the synthetic Open Tree, as well as the published source trees that contribute to the synthesis to generate comprehensive phylogenies for animals and microorganisms (see lines 296-321 in the Supplementary Methods). In this revision, we have now conducted all the phylogenetic analyses for plants, animals and microorganisms (see detailed results in Supplementary Tables 1.3-1.5, 2.3-2.5, 45.3-45.5 and 46.3-46.5). We still note that an overall agreed-upon phylogeny for the tree of life will likely never arrive, but we did our best to address the referee´s comment.

Comment: 7. Publication year should be included in the analyses (meta-regression)-it is not confounded with study identity, and it is potentially important and interesting moderator of the effect.
Response: We have clarified this now as "*Publication year: a continuous metric according to the year when the articles were published*" (lines 528-529 in the main text). We have included publication year in the meta-regression model (see detailed results in Supplementary Tables 1.1-1.5, 2.1-2.5, 45.1-45.5 and 46.1-46.5; Extended Data Figs. 2-7 and other associated supplementary tables).

Comment: 8. Funding and conflict of interest statements can be easily extracted from the included papers, even if they are not consistently included. Then, presence of potential links with the relevant industry should be analysed in a meta-regression. This is a too important issue to dismiss.

Response: We now extracted conflict-of-interest statements where possible. Among the 1705 papers, we could confirm that 1,411 ones did not have conflict of interest while 25 self-identified as having a conflict of interest. The remainder made no statement about conflicts of interest (lines 244-248 in the main text; Supplementary Table 58). In the meta-regression analysis, we consider "Conflict of interest status" as one binary variable (i.e., "1" denotes that a certain article has conflict of interest and "0" represents no conflict) (lines 526-528 in the main text). Detailed results for "Conflict of interest status" were presented in Supplementary Tables 1.1-1.5, 2.1-2.5, 45.1-45.5, 46.1-46.5 and 58, Extended Data Figs. 2-7 and other associated supplementary tables).

Replies to Reviewer #3:

Comment: The current ms derives from a previous version submitted to [Redacted] that has been commented by three reviewers. I now have checked for the implementation of the comments of Reviewers 2 and 3 in the revised version.

Response: We thank the reviewer for this positive evaluation. We carefully addressed all comments in revising the manuscript. We are grateful for the suggestions, which have improved the quality of our study (see this revision).

Comment: To my opinion, almost all comments of Reviewer 2 have been adequately addressed, except for "(1) *Comment: 216-217 were vertebrates and invertebrates equally affected, or one more than the other? Does that depend on the intended target of an pesticide, e.g. distinguishing between pesticides aimed at invertebrates or otherwise?*

Response: We have removed the sentence due to space limited, but from the statistical values in Supplementary Table 4, we can see that vertebrates and invertebrates were equally affected (see Supplementary Table 4)."

Response: We have added this information in the main text, and we have clarified it as — "*vertebrates and invertebrates were equally affected by pesticides*" (lines 258-259 in the main text).

Comment: and "(2) *Comment: 223 'affected different taxonomic groups' how?*

Response: We have removed the sentence due to space limited, but from the statistical values in Supplementary Tables 32–36, we can see that "insecticides, fungicides and herbicides affected different taxonomic groups with decreased growth, reproduction or behavior and with perturbed biomarkers" (see Supplementary Tables 32–36)."

Response: We have added this information in the main text, and we have clarified it as —

"*Insecticides, fungicides and herbicides affected different taxonomic groups though decreased growth, reproduction or behavioural responses, and biomarkers being perturbed from baseline conditions (see Supplementary Tables 32–36)*" (lines 216-218 in the main text).

Comment: I feel that both informations are important and should be kept within the main body of text rather than being implied to be taken from the supplementary material.
Response: Please see above two responses.

Comment: All comments of Reviewer 3 have been perfectly addressed in this version of the ms.
Response: Thanks for your positive comments.

Comment: In the Competing Interests paragraph, "B.W." should be replaced by "B.A.W.".
Response: Done as suggested (lines 717, 720 and 721).

Response to the referee comments

Replies to Reviewer #1:

Note we only show responses here for outstanding issues you identified in your previous review. Issues you identified as being addressed to your satisfaction we do not refer to here.

New reviewer comment - Round3: Thank you, for providing a link to the files on GitHub. I note that now the data and code are also provided on Zenodo and the contents of these two repositories do not match. Specifically, some code files that are archived on Zenodo are absent from GitHub version. Please remove any special (Chinese) characters from your R code.

Response: We thank the referee for pointing us at this. To avoid confusion and duplication we have now uploaded all raw data and code to Zenodo and removed them from GitHub (see this revision). We have checked that all relevant files are present. We have now also removed all Chinese characters as suggested. All data used in this analysis are available on Zenodo (https://zenodo.org/records/10495263). All code for this analysis is available on both Zenodo (https://zenodo.org/records/10509420).

New reviewer comment - Round3: Thank you for providing a README file. Unfortunately, the meta-data provided applies to only one of the data files, "Meta combined data.csv" (there seems to be 16 data files in total). The meta-data for this one key data table used to run the key models is still too rudimentary to allow cross-checking of the data or replication of the data extraction process. For example, the variable described as "Control (value): the value in control group;" does not explain which value should extracted be there, apart that it should relate to the control group. Another, example – "Control-n: the number in control group;", which is supposed to be the sample size, it also needs explaining the units of counting sample size, as these can be different for different taxa and experiment types – for animals it might be number of individuals, for plants could be individuals or plots, for microorganisms it could be a colony in a tube/plate, or number of natural plots, etc.

Response: We have now revised the README file to improve clarity. We hope this addresses the problems.

Comment: One column in the data table called "LnR" is not described – it seems to be an effect size that was not be used in the analyses, as the main effect size used was SMD. However, I cannot find the code the authors used for calculating SMD.

Response: The column in the data table called "LnR" had not been used in the analyses, and we have now deleted it. The code used for calculating SMD has been added in the README file.

Comment: Also, I cannot find the code for Figure 1 (the map), and instead there is code for Figure 1b, a histogram which is not in the main manuscript.

Response: Figure 1 had been prepared in ArcGIS, so no R code can be provided for that. We have

now deleted code for “Figure 1b” which was part of a prior version of the manuscript.

Comment: Further, the study codes in the main data file “Meta combined data.csv” do not match the study numbers in the table characterizing included studies (with their references) – “447069_2_data_set_8320286_s47vd1.xls”. Please include full study references as a column in the data file “Meta combined data.csv”, so that each data point can be linked to its original source. Overall, lack of consistent and comprehensive documentation is still a major limitation to transparency and reproducibility of this meta-analysis, which should be addressed before this work is accepted.

Response: We thank the referee for pointing us at this. We have now made sure that all study codes match up. We have listed all the 1705 study references in the file “Meta combined data” with a new sheet (Full study references). The paper code order in “Meta combined data” is based on the time order that we extracted data, and the paper order in Supplementary Data 1 is based on the authors' name in alphabetical order (see Meta combined data). In addition, we have added another data file “ExtendData” for other supplementary files (e.g., for Supplementary Tables 59-61, Supplementary Figures 24-27).

New reviewer comment - Round 3: I can see it in the methods section now. I suggest adding this [grey and non-English literature] also to the header of Figure 1 (the global) map, because it is highly relevant to the pattern presented there.

Response: We have added “Grey and non-English literature was not included in this meta-analysis” in the caption of Figure 1.

New reviewer comment - Round 3: it is great that you consulted the statistician and conducted the model assumptions checks. Please provide the results of the model assumptions test in the supplementary materials. Conducting sensitivity analysis for the main model that show that applying the recommended transformation does not change the results should be also presented in the supplementary materials to support your claims.

Response: As suggested, we now include a sensitivity analysis where we apply arcsine transformation for percentage and proportion data (as you suggest) and compare the results to our approach (no transformation or proportion/percentage data) (see Supplementary Table 59). This shows no qualitative difference in the results. We also provide a supplementary figure in the supporting information for the model checks of the original analysis (see Supplementary Fig. 24).

Original comment for context: 5. The issue with folded normal distribution when using absolute values has not been resolved. See Nakagawa, S. & Lagisz, M. (2016) Visualizing unbiased and biased unweighted meta-analyses. 334. Journal of Evolutionary Biology, 29, 1914-1916, for a visual clarification of the issue. **New reviewer comment - Round 3:** Please provide the results of the

model assumptions test in the supplementary materials. Conducting sensitivity analysis for the key model that show that applying the recommended transformation does not change the results for biomarkers data subset should be also presented in the supplementary materials to support your claim. Also, please flip the sign on the absolute values to make it easier to compare magnitude of estimated effects in all forest plots.

Response: As suggested, we now present additiojnal model checks in Supplementary Fig. 25 that show that the distribution used was appropriate (see Supplementary Fig. 25). However, we acknowledge the referee´s wider concern about the use of absolute values for biomarker responses. As already pointed out, biomarkers include metrics that may potentially be both be up- and downregulated as a response to exposure to pesticides. In both cases, this would represent biologically meaningful responses that may have longer term fitness consequences for individual populations. To account for the potential of both up- and down-regulation of biomarkers, we have treated these as absolute values within the main analyses. To address the referee´s comment, and to support a broader interpretation of the data, we now provide a comparative sensitivity analysis for the response of biomarkers to overall pesticide exposure for animals, plants, and microorganisms using absolute values of 'biomarker' (as described in the methods) *and the original raw values* of 'biomarker', i.e. with directionality of the effect (see Supplementary Table 60). As can be seen, even when the biomarkers are not treated as absolute values, they are still on average characterized by negative effects in response to pesticide exposure so that the results remain qualitatively equivalent to those presented in the main paper. This would suggest that while biomarkers may in principal be both up and down regulated following pesticide exposure, across the studies considered this effect remains overwhelmingly down regulation. We hope that this addresses your wider concerns about the biomarker responses.

New comment - Round 3: By looking at the raw data, "Meta combined data.csv", I can see "Life expectancy" and "Longevity" are in "insecticide-animal reproduction" category. It seems like all (and there are many) measures related to survival are classified as "insecticide-animal reproduction" (similar issue is also present for plants and microorganisms in the data set). Please provide justification for this choice.

Response: We acknowledge that to some extent the categories we had used for reproduction, growth, behaviour and biomarkers are synthetic, but we believe they still allow the most meaningful interpretation of the plethora of metrics measured across the 1705 studies as responses to pesticide exposure. While further sub-divisions of these categories could have been undertaken, this would not have added to the clarity of the manuscript. As indicated, certain responses to pesticides were allocated to these four categories based on the authors´ combined opinion, as well as though solicitation of opinion for colleagues. As such, "Life expectancy" and "Longevity" were considered as "animal reproduction" as premature death may limit lifetime potential reproductive output. Similarly, we think that animal survival determines population reproduction of animals, and thus

should be considered as “animal reproduction”. There are no “Life expectancy” and “Longevity” measures for plants or microorganisms. The overwhelming negative effects (or positive in the case of the use of absolute values for biomarkers) across reproduction, growth, behavior and biomarkers suggests that, while further sub-categorization of these responses could have been undertaken, this would not have changed the results.

New comment - Round 3: Also animal feeding rates could represent both growth and behavior – having them solely as a measure of growth may require additional sensitivity analyses and clarification in the methods section.

Response: The suggestion of the reviewer is correct as animal feeding rates could represent both growth and feeding behavior. Here, we assumed that animal feeding is primarily for animal growth, and thus we put “animal feeding” indicator into “animal growth” in this meta analysis. We do however follow the referee´s suggestion and apply a sensitivity analysis to assess the effect of allocation of feeding to either growth or behavior in our meta analysis. We have added a supplementary table and a supplementary figure to measure the sensitivity analyses of animal growth and animal behavior when “animal feeding” indicator was included in animal growth or in animal behaviour (see Supplementary Table 61 and Supplementary Fig. 26).

New reviewer comment - Round 3: In the “Meta combined data.csv”, there are some clear mistakes too, e.g. the column for describing measurements is named “Animal growth indicator”, and sometimes is empty, “nummber of eggs/female” is classified as “insecticide-animal behavior”, this raises concerns for the consistency of data extractions and the quality control procedures in this work.

Response: We have addressed these minor errors and checked that this has not affected data extraction. In “Meta combined data”, we have replaced the column “Animal growth indicator” with a new column (“Non-target organism indicator”), added the missing experimental year (i.e., 2017) in rows 7849-7853, and have added the missing unit (i.e., Individuals/g) in row 14201. In “Meta combined data.csv”, we have replaced “number of eggs/female” with “Walking distance (cm)” in row 8051, and have replaced “number of eggs/female” with “Walking speed (10^{-2} cm s^{-1})” in row 8053. Both “Walking distance” and “Walking speed” belong to “animal behavior”.

New reviewer comment - Round 3: Lines 463-466: “When a study included different levels of pesticide application rates, measurements for the control groups without pesticides versus different pesticide application rates were considered as independent paired observations” - This is a very bold assumption given “26,096 estimates of pesticide effects reported from 1,705”. Now looking at the actual data, I can see that although sometimes multiple pesticides or species were used in a single study, in most cases multiple doses were used, which means data non-independence related to the repeated use of the same control group to compare it with different dose levels is a serious issue.

Specifically, by not controlling for repeated use of the same individuals, the precision of the overall estimate is very likely inflated. One very simple way is to divide sample size of the control group by the number of comparisons it is used in, thus avoiding "double-counting" of the same replicated units.

Response: Observations with no pesticide had been considered as control groups, while those with different pesticide application rates were considered as the treatment groups. When an article included different levels of pesticide application rates, measurements for the control groups (no pesticide) were compared to all other treatment levels of pesticide application rates and treated as independent paired observations. Within the context of these analysis, this reflects typical experimental designs for studies considered in ecotoxicology that have a single control which is then compared to multiple test treatment residue levels. The random allocation of treatments to replicates, combined with study as a random effect in the meta-analytical models, means that these comparisons are appropriate and do not represent double accounting as suggested. In response to the referee´s comment, we have now re-run our analyses with the correction suggested, i.e. dividing the sample size of the control group by the number of comparisons it had been used in.

The results of this additional analysis can now be found in a new Supplementary Table 62. When comparing these results to our former approach (see Supplementary Table 3), there were no qualitative differences in the results (see Supplementary Table 3 vs. Supplementary Table 62). In particular, all negative effect sizes remained negative (and significant), and all positive effect sizes remained positive and significant.

New reviewer comment - Round 3: In the main text, please acknowledge study limitations due to sample sizes in the included experiments (from the raw data I can see that the median sample size is likely 3) – having a histogram of the distribution of sample sizes by type of organism and/or treatment would be very useful too.

Response: The median value of sample size was 4 and 4 for the control and treatment, respectively (see lines 122-123 in the main text). In addition, we have added a histogram to show the distribution of sample sizes by types of organisms and separately by control and treatment (see Supplementary Fig. 27).

Response to the referee comments

Replies to Reviewer #1:

The authors addressed some of my concerns but there are still outstanding issues related to transparency and reproducibility of this work, as well as some methodological concerns. My previous comments and authors responses are quoted below when they are linked to new issues identified. My new comments are provided as "New reviewer comment – Round4:"

Response: We thank you for these suggestions to improve transparency and reproducibility of our manuscript. We have now revised the manuscript according to these suggestions. The responses to the below comments describe the changes that have been implemented in greater detail. In particular, we have made sure that the work is transparent and reproducible. We have also re-done the data extraction from all original studies, which has greatly improved data quality.

New reviewer comment – Round4: Thank you for providing your data and code on Zenodo. It seems like putting the contents of these two repositories together still does not allow reproducing all the analyses and they are really hard to peer-review. There are three reasons: first the code is very poorly annotated, and thus it is hard to figure out what different sections of code are supposed to produce. Second, it seems like pieces of code are missing, e.g. file "Figures 2-3.R" starts from loading the file 'Meta combined data.csv' and then immediately runds a meta analytic model on the data from it, however the file does not contain effect sizes and there is no line of code that calculates effect sizes before running the model. Third, the code refers to loading many additional data files, that look like pre-processed data, but they are not included in the Zenodo repository, e.g. "Supplementary Tables 1(1-3).R" calls "Old_New_pesticide <- read_excel(paste("/lustre/huyueqing/ssy/Wan2023/DataPre/","Supplementary Data 2-Classification of old and new pesticides-2023-0825.xls",sep = ""),sheet = 1)", but there is no such file or no explanation where it originates from. Overall, I think all the analytic pipeline should be made fully transparent reproducible and then subjected to independent peer-review.

Response: Thank you for your comment. We have now completely overhauled the analytical pipeline in response to your comments. 1) We have now gone through the code and improved the annotation so that it allows a full interpretation of what has been done. 2) We have ensured that all pieces of code are present, including those relating to the production of all figures, tables and supplementary materials; 3) All files required for the analysis are included.

New reviewer comment - Round 4: Thank you. The README file only describes data in the main data file, but not in any of the additional data files and does not provide any information about the project in general and how the code and files should be used to reproduce the analyses. Please provide more comprehensive description of all your files, starting from overview and then providing meta-data for all data that is needed to reproduce your work. Consider using Quarto with R to

produce computationally reproducible documents that contain the R code, verbal descriptions, and outputs.

Response: We acknowledge that the previous version lacked reproducibility and had been hard to follow. Addressing this recommendation, we have now re-written the README file using Quarto with R, and extended this to include information on additional data files, as well as our analysis approach in general. We hope that this has now improved transparency of dataset descriptions and relevant code information, allowing clear interpretation of the approach used for all stages of the analysis. We have double-checked all code and made sure it runs on machines across continents.

New reviewer comment - Round 4: The code used for calculating SMD should not be placed in the README file, but in the code where it is needed to calculate effect sizes to reproduce the results.

Response: Thank you for your comment. Following your wider recommendation, all R code and a README file with R annotation developed using Quarto have now been uploaded to Zenodo. The specific code you are referring to for calculating SMD has been also been revised again in this version to avoid problems stemming from our initial use of SMD rather than log response ratios.

In our revised version, we now use the log response ratio (InRR) to calculate the effect sizes (the range of effect sizes in InRR is smaller than when using SMD). We have contacted Shinichi Nakagawa, an expert in meta-analysis and statistics, whose recommendations we follow in our revisions. Briefly, we now use log response ratios rather than SMD, because these do not require standard deviations and are more robust for use in meta-analysis, as demonstrated by Nakagawa et al. (2023; see reference below). For missing value imputation, we have now employed the so-called 'all cases' method, for sensitivity analysis and main analysis. Reporting follows the PRISMA-EcoEvo checklist to enhance quality. Additionally, we used the latest publication bias tests and submitted our fully annotated R code to Zenodo for transparency. We believe that these extensive revisions have considerably increased the robustness of our analysis.

Reference: *Nakagawa et al., 2023. A robust and readily implementable method for the meta-analysis of response ratios with and without missing standard deviations. Ecol. Lett. 26: 232-244.*

New reviewer comment - Round 4: Thank you for adding the study references next to study codes in the data file. I was now able to compare the extracted data with the information reported in the original papers. I randomly selected 4 papers I have full-text access to and found the following issues:

Response: Following your comments about potential errors in the data (and as an explanation for the significant delay in producing this review) we have now re-checked all the original data and studies used in this paper. This involved going back to the original papers and confirming values derived to produce effect sizes as used in the analysis (note this is the reason for the considerable delay in this revision as the complete recheck took a lot of time). Where on reassessment papers were not considered to meet requirements for data derivation, they have now been excluded. In

some cases, new papers were identified from citations within those that were rejected for the above reasons but not identified in the initial scanning process. The process for including these additional papers is detailed within the PRISMA process and diagram. Below, we reply to the specific further comments on individual papers.

- Paper 16. (*Colin et al. 2004. A Method to Quantify and Analyze the Foraging Activity of Honey Bees: Relevance to the Sublethal Effects Induced by Systemic Insecticides*): rows 16-21: the concentration of Fipronil is wrong - the study reports using 2 micrograms per kilogram while the extracted data is 6 g/kg. For Imidacloprid the unit is also wrong–it should be microgram per kilogram not gram per kilogram. For the outcome "number of daily attendance of active bees-nuclei " which was extracted for nuclei A, B, C, it is and somehow from control nuclei, but I cannot find the matching exact numbers in the text, and it is not possible to extract them from the figures in the text. Given the complexity of the experimental design, raw data would need to be used to calculate such values. There are no comments in the data file how the numbers were obtained, and thus they are not reproducible.

Response: In the context of the paper you refer to, the means, SD and sample size (*n*) of the indicator (i.e., "number of daily attendance of active bees-nuclei A") were calculated on days 0, 1, and 4. Because this was not a strictly repeated trial, we have removed this paper.

- Paper 1628. (*Zhao X, Li YL, Zhang LX, Dorna H, 2003. Effects of priming and fungicide treatment on germination of china aster (Callistephus chinensis L.) seeds. Seed Science & Technology, 32: 451-457.*): rows 1943-1944: it looks like only germination capacity data from Sample 1 and only from 30C temperature treatment were extracted and the standard deviations were manually imputed by assuming they are 10% of the measured value. This is only a small fraction of available data in this paper (multiple control and exposure groups, all samples should be extracted and used for calculating mean and SD) and is incorrectly extracted.

Response: We thank you for pointing us at this. To exclude the interactive effects of pesticide doses and temperature levels on non-target organisms, and only to measure the individual effect of pesticides on non-target organisms, our principle is that we extract the data at the highest temperature to conduct our analysis to exclude the interactive effects of pesticides and temperatures. The authors mentioned that "*The germination tests were conducted at 30 ℃, 20 ℃ and 10 ℃ in darkness on 8 replicates of 50 seeds for each treatment*." This is why the sample size was 8, and the SD of the value "Germination capacity (%)" at different temperatures is missing in Table 1. Because this is a comparatively old paper, we did not receive a response from the original corresponding author. To exclude your worries about this paper, we have now also removed this paper.

Seed germination tests

The germination tests were conducted at 30°C, 20°C and 10°C in darkness on 8 replicates of 50 seeds for each treatment. Control seeds were untreated. Fifty seeds were placed in each 9 cm diameter Petri dish containing 6 layers blotting paper wetted with 5ml of distilled water. Seeds were considered as germinating when there was a visible protrusion of the radical through the seed coat and pericarp.

For germination index and mean germination time were recorded at 24h intervals for 14 days. Germination index (G I) = (Gt/Tt)) Where G is the number of seeds germinated on day t and t is the number of days. Mean germination time (MGT)= TiNi/ Ni, where Ni is the number of newly germinated seed at time Ti.

Figure: Screenshot proof showing that these were 8 replicates

(from Zhao et al., Seed Science & Technology, 32: 451-457)

- Paper 1008. (*Gabaston J, Khawand TE, Wafo-Teguo P, Decendit A, Richard T, Mérillon JM, Pavela R, 2018. Stilbenes from grapevine root: a promising natural insecticide against Leptinotarsa decemlineata. Journal of Pest Science, 91: 897-906.*) rows 5808-5809: looks like data was extracted from Table 6 for 3rd day of exposure only and two lower doses out of 3 used (this is not reported on the data extraction sheet). The mean mortality % was converted to survival %, but the reported standard deviation values were somehow replaced with 10% of the men value. The sample size is claimed to be 4, but I cannot find such information in the article, and it seems improbable given the type of the organism, and reported values.

Response: In the data extraction sheet (i.e., Meta combined data-2024-0814), in rows 131-132 in column "N" (i.e., Non-target organism indicator), we have presented it as "survival percent-transformed by mortality-3rd day-Table 6". In the Methods section (i.e., Earthworm acute toxicity test), the authors have clearly shown that the sample size was 4: "*Ten adult earthworms were placed in glass containers (1 L) filled with the test substrate (650 g), and the test containers were enclosed with a polythene sheet with integrated gauze to prevent the worms from escaping and to ensure optimal ventilation. After 3, 7 and 14 days of incubation, living worms were sorted by hand; the test endpoint was mortality. The tests were repeated four times. The glass containers were placed in a growth chamber in an artificial climate (L16:D8, 20 ± 1 °C)*".

Ten adult earthworms were placed in glass containers (1 L) filled with the test substrate (650 g), and the test containers were enclosed with a polythene sheet with integrated gauze to prevent the worms from escaping and to ensure optimal ventilation. After 3, 7 and 14 days of incubation, living worms were sorted by hand; the test endpoint was mortality. The tests were repeated four times. The glass containers were placed in a growth chamber in an artificial climate (L16:D8, 20 ± 1 °C).

Fig. Screenshot proof that this paper had 4 replicates (Gabaston et al., 2018. Journal of Pest

Science, 91: 897-906)

Question: Overall, this small exercise clearly shows that 3/3 papers had some issues with data extractions. This indicates a very high rate of errors in the whole data set. The data lacks information on the exact sources of the extracted values for the outcomes at least (table/figure number, text page, raw data, etc.), any extractions decisions made, calculations, imputations, etc. – this makes assessing trustworthiness of the data set very hard and makes data extractions irreproducible. Because of this I request that all extracted data is annotated in detail. Further it has to be cross-checked by a researcher who was not involved in original extractions, re-extracted where issues are found and clearly described to make it fully transparent and trustworthy.

Response: As an expansion to our initial statement about a complete re-check of the data, we have now added the column "Y" (i.e., Exact sources of extracted values), the column "Z" (i.e, Extraction decisions made) and listed the extracted data information in column "N" (i.e., Non-target organism indicator). In this revision, Dr. Siyuan Shen re-extracted and cross-checked the data. After Siyuan's re-extracting and checking, Dr. Nian-Feng Wan and Dr. Ben A. Woodcock re-checked all the extracted information again, respectively, with additional checks being done by Christoph Scherber. During this re-extraction process, papers considered to have been unclear in their definition of measures of variance or described sample sizes were subjected to a group discussion (mainly composed of Nian-Feng Wan, Ben A. Woodcock, Dave Goulson, Adam J. Vanbergen, David J. Spurgeon and Siyuan Shen) to produce consensus on these derived metrics. At least three group members participated in each discussion. If our understanding was the same, we retained the data. If not, we contacted the authors directly to ask for the data. If the authors did not respond or could not be contacted, we removed the dataset originating from the respective paper. However, to collect more data to test pesticide effects on non-target organisms, we replaced papers from which no valid data could be extracted with a new, matching paper by the same authors or institutions (see lines 16-29 in Supplementary methods). This process has (as we state above) now been clarified within the revised PRISMA diagram and associated processing description for data set derivation.

New reviewer comment - Round 4: When checking the extracted data, I noticed that many of the extracted standard deviation values are exactly 10% of the mean value. This cannot be coincidence and the manuscript does not mention doing manual imputations for missing standard deviation values (or any other missing values). Please explain how these values originated and the detailed procedures used to deal with any missing data.

Response: You are correct that where the values were 10%, this was due to an imputation for missing values (refferring to: Luo et al., 2006. Elevated CO_2 stimulates net accumulations of carbon and nitrogen in land ecosystems: a meta-analysis. Ecology, 87: 53-63). In consultation with Professor Shinichi Nakagawa we have now revised this approach to deal with missing SDs in meta analyses. According to Dr. Nakagawa´s suggestions, we now followed the approach described in his

paper (*Nakagawa et al., 2023. A robust and readily implementable method for the meta-analysis of response ratios with and without missing standard deviations. Ecol. Lett. 26: 232-244*) and used Equations 6 and 7 (see methods section in our manuscript) to calculate effect sizes and sampling variances when SDs were missing by simply imputing the pooled coefficient variation (CV) from the subset of studies that do report SDs. This is now fully described in the new revised version of our manuscript (see lines 30-34 in Supplementary methods).

New reviewer comment - Round 4: The code for reproducing this should be in the file "Supplementary Tables 59-61.R", and as such cannot be rerun. In the provided code I cannot find any line that would execute arcsine transformation for proportion data. And the code starts from loading a mysterious data file that is not archived on Zenodo, so cant tell what is there nor how it got there: AllData <- read.table(paste(DataPre_path,"SuppTab59Data.txt",sep=""),header = T). This seems to be ongoing issues across the provided code which is compounded by almost complete lack of code comments and descriptions.

Response: In this revision, and in response to your other above comments, we have re-written the code for executing the arcsine transformation for proportion data. Code for the arcsine transformation for proportion data is now included and annotated in the README file and all operational R files. We also had an internal discussion on using arcsine vs. logit transformation by hand, but sticked with the arcsine as it is more established in a meta analysis context. The datasets are described in README file and can now all be fully accessed. Annotation has been improved for the code in general using as you suggested using Quarto with R.

New reviewer comment - Round 4: The authors are wrong when stating that "The random allocation of treatments to replicates, combined with study as a random effect in the meta-analytical models, means that these comparisons are appropriate and do not represent double accounting as suggested." The measures from the same control individual are used repeatedly in comparisons of effects of different doses of the same pesticides and they are such effect sizes are not independent and are a form of "double-counting". The analyses that account for such non-independence should be used as main models not as supplementary. The author used sample-size corrections their sensitivity analyses for a few models. I suggest testing their all models using cluster-rubust approach which is already implemented in metafor package (https://wviechtb.github.io/metafor/reference/robust.html) as their main approach.

Response: Thanks for your comment. We have addressed this important issue following the approach described by Shinichi Nakagawa in his paper on "*Quantitative evidence synthesis: a practical guide on meta-analysis, meta-regression, and publication bias tests for environmental sciences.*" (*Nakagawa et al., 2023. Environ. Evid. 12: 8*). In order to adjust for repeated measurements of control values, we assigned the argument "V" in the "rma.mv()" function with the sampling variance-covariance matrix estimated by function "vcalc()" (lines 584-592).

New reviewer comment - Round 4: Please also run and report an overall (global) meta-analytic model on all data with organism, trait and pesticide type as random effects.

Response: We derived responses for 830 different species (560 animals, 192 plants, 78 microorganisms) and 129 non-species-level groups (i.e. 79 non-species-level animals, 8 non-species-level plants and 42 non-species-level microorganisms) to 471 different pesticide active ingredients (243 insecticides, 104 fungicides and 124 herbicides) (Supplementary Tables 14–28; Supplementary Data 1–3) (lines 118-122). In the last version of this paper, pesticide identity (active ingredients pesticide name) and study identity were used as random effects that were recognized by you. In our revision, we have now added organisms and traits as random effects, according to your suggestions. Namely, we run/reported an overall (global) meta-analytic model on all data with organisms (unique species name and unique non-level species groups), traits (growth, reproduction, biomarker and behavior), pesticide identity (active ingredients pesticide names) and study identity as random effects (see Supplementary Table 1.1, Supplementary Table 1.2, Supplementary Table 2.1 and Supplementary Table 2.2), run/reported an overall meta-analytic model on all data with plant species (excluding non-level species groups), traits (growth, reproduction and biomarker), pesticide identity (active ingredients pesticide names), and study identity for plants (see Supplementary Table 1.3 and Supplementary Table 2.3), run/reported an overall meta-analytic model on all data with animal species (excluding non-level species groups), traits (growth, reproduction, biomarker and behavior), pesticide identity (active ingredients pesticide names), and study identity as random effects for animals (see Supplementary Table 1.4 and Supplementary Table 2.4), and run/reported an overall meta-analytic model on all data with microorganism species (excluding non-level species groups), traits (growth, reproduction and biomarker), pesticide identity (active ingredients pesticide names), and study identity as random effects for microorganisms (see Supplementary Table 1.5 and Supplementary Table 2.5).

New reviewer comment - Round 4: Thank you for adding information about the median sample sizes. However, this raises new issue since Line 562 states "A large-sample approximation was used to compute the sampling variances46." This approximation is not suitable for such small sample sizes, and a small-size-corrected version of the SMD (Hedges' g) should be used as an effect size in this meta-analysis.

Response: Thank you for your comment. According to suggestions from Professor Shinichi Nakagawa, we have replaced SMD with InRR to calculate the effect size and sample variance. Professor Shinichi Nakagawa reported four new methods to calculate effect size when missing SDs exist, which was reported in his paper entitled "*A robust and readily implementable method for the meta-analysis of response ratios with and without missing standard deviations*" (*Nakagawa et al., 2023. Ecol. Lett. 26: 232-244*). He suggested that using the average CV to estimate sampling variances for all observations, regardless of missingness, performs with minimal

bias. We have exactly followed his suggestions.

Response to the referee comments

Replies to Reviewer #1:

Comment: Thank you for your patience and hard work on this manuscript. I am pleased to note that it is much more transparent and robust now (it seems to have some new interesting findings as well). I really hope that in your next meta-analytic project you will take all steps from its inception to make sure it is trustworthy and well-documented, so I dont need to do 5 rounds of review again... good luck.

Response: We thank you for your positive evaluation. We have checked the manuscript again.