Referee: 1  
  
Comments for the Authors   
General comments:  
  
This manuscript reports the findings of a meta-analysis investigating the influence of thermal variability on the capacity for phenotypic plasticity in ectothermic vertebrates and invertebrates. When reading through this paper, it was very difficult to find anything to suggest improving. The manuscript is very clear and well written, the meta-analysis was carried out according to the latest methods and accounting for a range of potential confounds (e.g. I appreciated the checking for potential publication bias), I appreciate that the data and code have been made publicly available, the supplementary materials are very comprehensive and provide all necessary information for the reader to understand how the study was conducted and how to interpret the results, and the conclusions drawn are fully supported by the data.   
  
All in all, it is my opinion that this manuscript covers an important and timely topic, and is a good fit for the broad readership of Ecology Letters. I have provided line-by-line feedback below that I hope is helpful but it is all very minor.  
  
Line-by-line feedback:  
  
Page 2, line 43: ‘Most studies used’ is an understatement, given that it was 98% of studies that used diurnal temperature fluctuations—right? Would be good to strengthen this wording to make it more accurate.

**RESPONSE:** We now state that 98% of studies used diel fluctuations;

Page 2, line 47: It is not mentioned in the Abstract that the dataset is dominated by invertebrates. The proportion of studies on invertebrates is probably worth mentioning as a mini-limitation (perhaps, in the same sentence where you mention that most studies used diurnal temperature fluctuations?)

**RESPONSE**: we changed the text as suggested and now state that "most data were derived from invertebrates" in the same sentence as the comment on diel fluctuations;

Page 3, line 71: This sentence needs supporting citations: ‘The capacity for phenotypic plasticity has been tested mainly in response to changes in constant temperatures.’

**RESPONSE**: we now added Schulte et al 2011 to support this statement

Page 3, line 81: Comma needed after ‘cost’.

**RESPONSE**: corrected;

Page 4, line 104: There is a problem with the formatting of the Foo citation (brackets).

**RESPONSE**: corrected;

Page 5, line 113: Comma needed after ‘treatments’.

**RESPONSE**: corrected;  
  
Page 5, line 119: How many papers were extracted collectively to ensure consistency?

**RESPONSE**: we now added that we extracted data collectively from five papers;

Page 8, lines 174–175: Is there a paper that you could cite to support this decision to add 0.5 to all means and SDs?

**RESPONSE**: Dan, do you know of one?

Page 9, line 204: I very much appreciate that all of the data and code has been made publicly available.  
  
Page 9, line 209: I also appreciate that the various statistical packages used have been properly cited.  
  
Page 11, line 262: This is a fantastic section and an equivalent is sorely needed in many meta-analyses published.  
  
Page 12, line 292: ‘between study’ should be hyphenated.

**RESPONSE**: corrected.

Page 13, line 294: ‘between species’ should be hyphenated.

**RESPONSE**: corrected;  
  
Page 14, line 327: There is a grammatical issue with this sentence: ‘…development time exhibit a non-linear exponential…’

**RESPONSE**: we edited this sentence to improve the grammar.

Referee: 2  
  
Comments for the Authors   
This study evaluates how constant vs fluctuating temperature treatments affect the capacity for plasticity across 40 ectothermic species. Overall, the study finds that constant and fluctuating temperatures do not differently affect plastic responses, and other considered factors (type of plasticity, trait, fluctuation, etc) also had limited impacts. Evaluating how temperatures may impact organisms, especially ectotherms, in face of global climate change seems an important line of research, and I thought the study’s meta-analysis seemed overall well executed. However, I felt the study could have made more effort to establish how it builds on previous work, expanded explanation of the PRRD metric, and could potentially soften conclusions in places. Below I have a few larger comments followed by specific comments for the authors to consider:

**RESPONSE**: we provide responses to these points below.

-I think the introduction would benefit from illustrating ideas and hypotheses with empirical examples or previous meta-analyses. Given this is a synthesis, readers may benefit from being more aware of studies that provide support for introduced ideas in ectothermic species before authors outline the aim of this meta-analysis (especially if current state of knowledge is equivocal). After reading Stocker et al. 2024 Ecol Lett, I felt the introduction overlapped in content a bit with this manuscript so perhaps expanding on some of the hypotheses and predictions unique to this manuscript using examples would help reduce overlap and increase novelty/interest.  
  
- Following this point, there seem to be a few recently published meta-analyses that broadly look at how temperature affects phenotypic responses in reptiles or ectotherms more generally (Noble et al. 2018 Biol Rev; Raynal et al. 2022 JEB; Stocker et al. 2024 Ecol Lett; Noble et al. 2025 Func Ecol); with the focal questions differing by examining a specific type of trait (e.g., physiology) or phenotypic response (mean, variance, plasticity). The studies seem to address different questions to the current manuscript but, given the subject area, I wondered why they were not acknowledged more explicitly in the manuscript. Do findings from previous studies shape predictions in this study? How is this study building on previous ones to fill remaining knowledge gaps? Is the dataset used in this manuscript a subset of datasets in these previous studies? How do the results fit in with this broader state-of-knowledge? I think the manuscript would benefit from explicitly linking previous related works to the studies’ aims/hypotheses/results.

**RESPONSE** (to both comments above): we have now re-written sections of the Introduction to more explicitly summarise findings from our earlier analyses and point out how the present analysis makes a conceptual advance. We also updated the text in the Methods section, clarifying searches and the data used in this analysis (i.e., we used data from a previous search that was not published earlier, and we updated the earlier search to cover the years 2022-2026). Please note that the ms is considered as a Letter so that we have to adhere to strict word limits, so that we had to be relatively brief with the added text;

- The PRRD effect size was new to me and it was unclear whether the manuscript was introducing it for the first time? Since the manuscript heavily relies on this (potentially new) meta-analysis effect size, I think the explanation of this could be expanded. I have included quite a large comment on this below with several points that came to mind, and I think addressing at least some of these would help improve clarity on PRRD and how to interpret it under different contexts. I thought that Figure 1 could also be modified to incorporate PRRD as well.

**RESPONSE**:

- I also felt that some conclusions could be softened or caveated further. There are limited studies that examine plastic responses, especially under the specific experimental design required by this meta-analysis’ criteria, so the number of effect sizes and studies in the final dataset are comparatively low to other meta-analytical studies. I acknowledge that it would be difficult to obtain a large dataset on this topic, but I think it could be important to highlight some more of the study’s limitations and how this affects interpretation. For example, results will mainly apply to the species in the dataset: how phylogenetically and geographically diverse, which ones have replication? Also, see my comment below on the model’s ability to separate variation explained by study and species effects, with implications for some of the authors’ conclusions.

**RESPONSE**: We now added a section in the Discussion (plus references) to point out taxonomic and geographic limitations of the data set

Specific comments:   
L61: Are there empirical examples of this in ectotherms that would help illustrate this point for readers?

**RESPONSE**: we now provide an empirical example (acclimation of heart rate in fish);

L65: I am not sure about the term “within-individual phenotypic plasticity” here... are the authors wanting to use this term here for a specific reason, e.g., to describe context where the same individual shows variation in plastic responses, and maybe then should elaborate? Or would using phenotypic plasticity suffice here and then broaden this statement so that phenotypic responses include those estimated at the population level (which comprises both among- and within-individual variation)? For me there seemed to be a bit of a mismatch or disconnect between the biological levels described, so maybe clarify.

**RESPONSE**: We edited this sentence and removed (within-individual), which indeed was superfluous here.

L70: I think including examples here on how plasticity at different constant temperatures affects specific phenotypes could help establish why this question is important to readers?

**RESPONSE**: We removed this sentence, and we provide an example of acclimation earlier in the text (please see also above).

L88/99/93: I wonder if the authors should define (maybe later in the methods?) what they mean by the “capacity” for plasticity. I suspect there is a reason why this term is being used (i.e., population-level phenotypic difference between contexts could be driven by other processes?) and I think the reasoning for this should be explicitly stated.

**RESPONSE**: We used "capacity" in the sense of the extent to which different phenotypes are expressed. It is a somewhat confusing word and we have now circumscribed it throughout the text to clarify our meaning.

L74-89: The authors set up the hypothesis that plasticity should be reduced and/or costly in (short-term) fluctuating thermal environments. I think this is a valid hypothesis, but I wondered whether it would be worth setting up other hypotheses where plasticity increases or shows no change in (short-term) fluctuating thermal environments. The authors acknowledge that current empirical evidence is equivocal so perhaps it would be more informative to set up the state of knowledge to illustrate this statement? I.e., the reference you cite here seems to suggest that impacts of temperature fluctuations may depend on species characteristics and type of fluctuations.

**RESPONSE**: We edited the Introduction to improve arguments (+literature) of how temperature fluctuations could promote or constrain plasticity.

L92: The study currently includes no predictions on how fluctuating vs. constant temperatures will impact phenotypic responses overall, and I feel that the basis for the study would be clearer if the current state of knowledge and follow up predictions were better introduced. Are there clear directional predictions? What about for the different moderators? If not, I feel that this is an important point to highlight.

**RESPONSE**: Please see also our responses above: we now provide brief summaries of the findings from our earlier meta-analyses that looked at how fluctuations affect phenotypic responses relative to constant temperatures. These meta-analyses also present extended discussion of why and how fluctuations could affect phenotypic trait values. Given the space restrictions of a Letter we will not be able to repeat the details of these arguments here;

L162: Provide number of effect sizes requiring this transformation? 

**RESPONSE**: We now give the information that 23 effect sizes were transformed.

L181: The use of the PRRD effect size in a meta-analysis was new for me and I found Figure 1 helpful in at least working out how lnRRs were calculated. I was not able to find the PRRD in the associated reference in this section (Noble et al. 2022), so I was wondering if this is the first use of this metric? The PRRD seems to be a two-step process where lnRR is subtracted: the first is the difference in lnRR between treatments (H minus L) within contexts (C&F) and the second is the difference in (difference of) lnRR between contexts (F minus C). Would it be worth modifying Figure 1 in a way to show both steps somehow as the second is currently absent (i.e., this scenario yields a PRRD < 0)? As the authors state on L180, this seems to be interpretable similarly to an effect size of an interaction. Were there H/L\*C/F interaction effect sizes reported in these papers and how does the PRRD metric relate? I also thought it would be helpful for the authors to outline to readers how PRRD captures something biologically new compared to lnRR and lnCVR. Since the paper relies on the effect size PRRD, perhaps the explanation could be expanded here.

**RESPONSE**:

Figure 1: what happens if there are cases where: i) within contexts (C and F) the phenotypic effect size is higher at TL than TH (negative slope),

**RESPONSE**: In this case the effect size will be negative; there is no *a priori* assumption that phenotypic responses are higher at TL - we mention this now in the caption.

ii) the TL/TH slope is stronger in the F than C context,

**RESPONSE**: As above, there is no *a priori* assumption that one slope is steeper than the other.

or iii) the TL/TH slope have different signs between F and C contexts… how does this affect the calculation and interpretation of lnRR(F/C) and PRRD?

**RESPONSE**: Any variation will change the sign and magnitude of the effect size and there are no constraints on the actual phenotypic values or slopes-----Dan, could you add something clever about PRRD?

I wondered whether an expanded example(s?) to illustrate the calculation of PRRD in Figure 1 be helpful. If the paper does not include scenarios of i), ii) or iii), then I think it is still worth mentioning and explaining if there are any limitations based on which treatment/context is subtracted from which so that readers know how to apply this effect size to their own work. I may have missed something important about this or a related reference, but I think it could be important to add these points more clearly to the manuscript.

**RESPONSE**:

L210: What important information does effect heterogeneity tell us? This terminology may be less well known by more general readers, so it could be helpful to explain.   
  
**RESPONSE**:

L221: Given you have 44 studies and 40 species, I assume Study ID has levels = 44 and Species ID has levels = 40? Trait ID had 4 or 6 levels? And what about shared animal ID? I didn’t see information on the data structure in the supp (e.g., in Table S3), so perhaps consider adding this somewhere?

**RESPONSE**: these data are shown in Tables S3-S8 ---Dan, we'll have to update the TAbles but this is essentially the info they ask for isn't it

L258: I thought it was interesting that the study examines amplitude and number of fluctuations as continuous moderators. I wondered whether it would be possible or if the authors considered also examining something similar to wavelength (i.e., amount of time between temperature peaks or troughs). Is this something that differed between studies? Depending on the organism or trait, this could have really different biological impacts on whether it pays to “filter out” short-term signals and reduce plastic responses.

RESPONSE: we recorded the fluctuation period, and it turned out that 98% of all papers used diel fluctuations so that there are not enough degrees of freedom to analyse period any further. This information is given in the first paragraph of the Results section;

L269: What does average effect decreasing over time mean in plain words?

**RESPONSE**: It simply means that as more effects accumulate overtime they begin to converge on their most likely average value across the species included in the meta-analysis.   
  
L294: I wondered how well the model can separate i) variation between studies and species and ii) variation between species and phylogeny. I had a quick look at the models and when removing Study ID: i) it seemed that most variation moved to the Scientific\_Name effect (species; estim = 0.0005) suggesting that these effects explain similar variation and ii) when removing also Scientific\_Name then phylogeny starts to explain a bit more variation (estim = 0.0003). I would suggest the authors consider how this may affect their conclusions and interpretations. I wondered if the dataset limits how much they can conclude about random effects and heterogeneity here.

**RESPONSE**: It’s a good question, and fair point. In theory, this is possible, but we agree that the ability to disentangle sources of variability from a practical perspective is often challenging when groups overlap a lot causing identifiability issues. This is likely a common problem. A recent simulation study by Cinar et al. 2022 looked at this issue in the context of estimating between species and phylogenetic variance components (in this case these are nearly perfectly confounded), and they still conclude that a model accounting for both species and phylogeny should still be fit. As quoted from their paper:

“*Based on our results, we suggest that meta- analyses in ecology and evolution should use the model that accounts for both the nonphylogenetic and phylogenetic species- level variance in addition to the multilevel structure of the data. Any attempts to simplify this model, such as using only the phylogenetic variance component, may lead to erroneous inferences from the data.*

*In addition, most meta-analyses include study as a random effect, and it would be odd to not report this obvious source of variance (even if it may not be estimated well).”*

Whichever way we look at this, not including random effects because of extensive overlap or including them and having them mis-estimated, it will impact our conclusions. So, there are no easy solutions. Considering the simulations from Cinar et al. 2022, we agree with their conclusions and believe it’s better to include all our *a priori* random effects (capturing known hierarchical structure) and simply acknowledging the challenges in estimating them. We have done this now.

Dan, I suppose it may be limiting that we have just about as many species as studies so that we cannot discriminate between the two - I added something on species diversity to the Discussion but we may have to be a bit more specific - also comment below, what do you think?

L304: I guess the biological significance of the uncertainty around the predicted percent change (10%) would depend on the trait and metric. Perhaps a 10% change could be a lot in some contexts? But I agree that the predicted effect size seems overall small.  
  
L319: I think this conclusion hinges on whether you can successfully separate variation between studies and species in the dataset. You have a low number of species (N = ?) that are from multiple studies in the dataset, so I would suggest caution with this conclusion. I wondered what the (descended) geographical and phylogenetic extent was for the species in the study, could this be discussed below?

**RESPONSE**: We now added a paragraph here to note the limitation of relatively few species that were unevenly distributed phylogenetically.

L324: I noted that fluctuating thermal environments reduced both variance (from Stocker et al. 2024) and plasticity in body mass, but this maybe depended on context? How do results in this study compare to previous meta-analyses?

**RESPONSE**: We now comment here on the parallel between our results regarding body mass here and our earlier meta-analysis.

L327: I like the added example here to explain, but could you add citations or a specific study? You mention “differently impact traits” above so what about another trait?

**RESPONSE**: We now provided an additional example and references.

L340: Empirical examples that provide support for the theory in this section?

**RESPONSE**: we added a reference with an empirical example.

L344: I also wondered if this conclusion was too strongly phrased. This conclusion is not actually tested in the study, so this is more of a hypothesis that stems from the results rather than a conclusion of the study no? Also maybe the authors could be more precise here with what they mean by “longer-term”.

**RESPONSE**: We toned down this conclusion and now state that "we suggest that plastic responses are driven by longer-term (days to weeks) mean temperatures."

L346-357: Something that came to mind when reading this paragraph was the nuance of spatial and temporal scale (diel, seasonal, annual fluctuations over micro-, meso-, macro-climates) and how different scales can have very context-dependent biological effects depending on the trait, fitness consequences, species, habitat, etc. Maybe this makes it difficult to meaningfully generalize how fluctuations in temperature across one scale (diel) will impact ectotherms more broadly? I would be interested to hear if the authors have thoughts on this and whether the experiments in the dataset use really different time periods (i.e., over the study period or between temperature fluctuations)?

**RESPONSE**: The data in our dataset came from laboratory experiments so that the period and amplitude of fluctuations experienced by study animals was well defined, and we are comfortable with the conclusion that diel variation in temperature has limited effects on plasticity. However, the question raised by the reviewer and up to a point in our paragraph is very interesting beyond the conclusions of this analysis: what is the effect of concurrent variation at different temporal and spatial scales. As far as we know, this question is unresolved (and also un-addressed) in the literature and would be important to consider in future work.

L361: How does this conclusion fit in with this sentence from the abstract of the previous paper: “Fluctuating temperatures also decreased longevity, and increased amplitudes had negative effects on population responses in aquatic organisms.” (Stocker et al. 204)

**RESPONSE**: Stocker et al analysed effects of fluctuations on mean trait values, while here we investigated effects of fluctuations on plasticity of traits. Hence, there is no contradiction between these statements. As we pointed out earlier, damaging conditions could also constrain plasticity if particular traits become "uninducible" but (as we point out) from the data in the literature, we have no indication that damage limited plasticity.

L369: Since the study only uses experiments, how do these results suggest that these findings are transferable to natural environments? Maybe expand this explanation.

**RESPONSE**: We now edited this concluding paragraph to better place data from diel variations into a natural context.

Figure 7: Is there some uncertainty that could be shown around these effects?

RESPONSE: each circle represents a single effect size (rather than the mean effect size for a given category) so that there are no associated error bars;  
  
Table S1: Small formatting thing. 100% associated with Meta-analysis and Meta-regression have different alignment.

RESPONSE: I can't quite see what this refers to  
  
Figure S2: This figure is very clear and organized. I wondered whether a nice addition would be to add number of studies (N = X) under “Exclude” since this decision tree has several inclusion criteria steps. This could be informative to readers to know which steps filtered out the most studies and would expand the break down shown in Figure S1.

**RESPONSE**: Unfortunately, we did not keep track of this level of detail at each stage of this process.

Table S4-13: I noticed that most subset and meta-regression analyses do not include the full dataset (e.g., Table S8 = 97 effect sizes?), perhaps you should be more explicit about this by adding N = effect sizes in the table legends and pointing this out in the text methods/results? This does seem to be illustrated in the figures, but could be more consistent. Further, maybe it is worthwhile and more thorough to report random effect variance and CIs in these results tables? I see Table S14 has the prop variance explained, but showing the actual estimates and uncertainty somewhere may also be nice, especially if certain moderators could change the variation explained for these effects.

**RESPONSE**: do this after the final analysis?

Referee: 3  
  
Comments for the Authors   
Stocker et al., present a well-written manuscript of a meta-analysis comparing constant and fluctuating temperatures on phenotypic plasticity. After a rigorous exploration into the data, including thoroughly assessing heterogeneity and publication bias, the authors show compelling evidence that phenotypic plasticity does not vary between constant and fluctuating temperatures.

However, it is unclear to me how this work builds on the author's previous manuscript using, in essence, the same dataset – Stocker et al (2024), Ecol. Letters. The authors cite this manuscript but do not explicitly state how this work differs or builds on what they previously found. In its present form, the current manuscript appears to largely be a re-analysis of the 2024 manuscript with a focus on phenotypic plasticity rather than biological responses more broadly. The authors should revise the manuscript to describe how this work differs and builds upon their 2024 manuscript.

**RESPONSE**: We have now revised the Introduction to state more clearly how the present work builds on our previous analyses.

Otherwise, the content of the manuscript is of a high standard and I only have a few further minor comments and suggestions, detailed below.  
  
Suggestions:  
  
I wonder if it is possible to account for the variation in temperature a species could experience in the wild. It could be species that do not experience as much temperature variation in the wild are less likely to exhibit phenotypic plasticity and mask any responses of species that would experience high temperature fluctuations in the wild. Perhaps a moderator with natural temperature range could be added.

**RESPONSE**: This is a really interesting question. While we agree it would be good to quantify temperature variation, practically, estimating biologically relevant temperature ranges for a population or species is very challenging for meta-analyses with diverse traits (as we have here). Wide-ranging versus range-restricted species will result in thermal ranges being widely different. However, even if we are to focus on a given population, there are major challenges obtaining biologically relevant thermal data because these data are usually derived from broad climatic data that is not representative of the microclimates a population experiences. Spatial resolution aside, when we are discussing plasticity, temporal resolution is also of critical importance. More specifically, we need to know what the biologically relevant temperature window is that should be obtained for a given population. Is it at the temporal resolution of weeks? Months? These factors combined, along with the fact that all these studies are conducted in the lab make mapping thermal variability to effect dubious (see a discussion of this challenge in Noble et al. 2025).

More detail about the moderators would be beneficial. In particular, the phenotypic trait categories are not all present in the results (behavioural and gene expression). Is this because some didn’t have enough effect sizes to be included? Please specify. Perhaps a table showing the moderators and what kind of traits fall within these. For example, the category “life-history” could cover a very broad range of traits, and it would be useful if this was detailed.

**RESPONSE**:

Specific comments:  
  
Line 1: “negative changes” sounds like it’s a reduction in temperature. Perhaps “adverse changes”?

**RESPONSE**: We changed this to "adverse effects".

Line 39-43: This sentence is quite long, and “type of phenotypic plasticity” doesn’t seem to fit in the sentence structure. I’d suggest breaking this down.

**RESPONSE**: We edited the text to split this sentence into two.

Line 70: “days to weeks” seems oddly specific. Not months? “Short term” is probably more suitable.

**RESPONSE**: This sentence was deleted in the revised version of the ms.

Line 140: What is the rationale for only selecting one temperature treatment if a given study had more than one? Why not include all data available and have study as a random effect in the model?

**RESPONSE**: We now added that we did this to minimise inclusion of data from potentially damaging temperature ranges when studies included exposures to more extreme temperatures and provide a reference.

Line 148: What “associated data”? Taxanomic information, preferred ecosystems & life history stage mentioned in the next sentence? Or were there others? Please make this clear.

**RESPONSE**: We now combined these two sentences in one and deleted mention of "associated data".

Lines 112-114: I do not see the relevance of adding references for your search terms.

**RESPONSE**: We removed the references.  
  
Line 344: I am unsure that the manuscript presents evidence that plastic responses are driven by longer-term mean temperatures. Please rephrase.

**RESPONSE**: we edited this sentence to now state that:"...plastic responses are driven by mean temperatures across more than one day".  
  
Line 593: Given this figure is referred to at the start of the materials and methods. Please state what all acronyms stand for, specifically PRDDs.

**RESPONSE**: We edited the Figure caption to include explanations of acronyms.

Figures: In their current form, the figures are quite small and difficult to see the CIs/whiskers. Perhaps arrange vertically rather than horizontally.

**RESPONSE**: As suggested, we have now revised figures so panels are organised vertically rather than horizontally.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
Editor  
Editors Comments for the Author(s):  
Based on the thorough and constructive reviews from three experts in the field, along with my own detailed assessment, which I found especially necessary to do for this manuscript, I share many of the reviewers’ concerns and have identified several additional significant issues that warrant substantial revision. Given these considerations, I recommend that the manuscript be rejected in its current form, with the possibility of resubmission after major revisions. I regret that I am unable to offer a more favourable decision at this time.  
I would like to briefly highlight what I consider to be the most pressing issues. However, should the authors choose to resubmit their manuscript to Ecology Letters, they will be expected to provide a detailed, point-by-point response to all comments raised by the three reviewers, as well as the additional points outlined in my own review (appended below).  
  
In agreement with two of the reviewers, I too feel that this manuscript should better integrate the findings of the (at least) five meta-analyses published to date on a “similar topic”: Noble et al. 2018, Bio Rev; O’Dea et al. 2019, Fish & Fisheries; Raynal et al. 2022, JExB; Stocker et al. 2024, Ecol Lett; and Noble et al. 2025, Func Ecol. Although three of those meta-analyses are cited, I can appreciate that, for readers who are not specialists in this specific area, the distinction between the previous studies and the present work may not be immediately apparent. I say this after having spent some time going through the methodology, results, and conclusions of all five previous meta-analyses, prompted by the reviewers' comments. Understandably, this can raise questions about the unique contribution of the current manuscript and how its findings fit into the broader body of work. While these differences may be evident to the authors, I strongly encourage them to clearly articulate how their study both differs from and builds upon the earlier meta-analyses in endotherms and ectotherms. Doing so will significantly enhance the clarity of the manuscript and help ensure its relevance and impact are fully recognised by a broader audience.

**RESPONSE**: We have now added more clarity about how our work here is quite different from those listed by the Editor. While they all involve plastic responses to temperature, none of these have been focused on how these plastic responses vary by a second factor.

In addition, there are two main analytical decisions that would require further explanation and confirmation. First, as highlighted by one of the reviewers, it is important to provide more detailed information about the effect size developed for this meta-analysis: PRRD. From what I understand, this effect size may behave similarly to an interaction effect size, and its associated sampling variance appears to be larger than that of a standard log response ratio (lnRR). This suggests that detecting a statistically significant effect – which the authors are using for concluding whether temperature variability influences phenotypic plasticity – with this metric may require more data compared to when one uses a classical lnRR. Importantly, my comment does not imply that the new effect size is incorrect; rather, it means that drawing reliable conclusions, especially avoiding false negatives, may necessitate a larger dataset than the current one (n = 40 studies, k = 212 effect sizes). Since the authors developed this effect size, and as far as I know, it has not been previously validated, I would expect a more thorough discussion and evidence supporting its properties before we can confidently interpret the statistically non-significant results as not being simply due to increased data demands.

**RESPONSE**: We agree that using PRRD will require larger sample sizes (as will be the case for any interaction-based effect size), however, we haven’t just focused on significance, as implied here. We also report all mean effect sizes and associated uncertainty (PIs, relative and magnitude hterogenity). The average effects are very small so we are confident that there is little evidence, when combined with significance thresholds, that there are meaningful differences between plastic responses under constant and fluctuating environmental conditions.

In terms of PRRD, its properties are well understood given it’s a derivation of log response ratios, which have very well-known properties (see Lajeunesse 2011, Hedges et al. 1999). Similar derivations have been used in other interaction-based effect studies (See for example, Pottier et al. 2021; Gurevitch et al. 2000; Macartney et al. 2022; Noble et al. 2022). However, we agree that some more detail would be useful. As such, we have expanded our section to describe it in more detail in accordance with the Editors and Reviewer 2’s suggestions.

**References**

M. J. Lajeunesse, On the meta-analysis of response ratios for studies with correlated and multi-group designs. *Ecology* **92**, 2049–2055 (2011).

L. V. Hedges, J. Gurevitch, P. S. Curtis, The meta-analysis of response ratios in experimental ecology. *Ecology* **80**, 1150–1156 (1999).

E. L. Macartney, M. Lagisz, S. Nakagawa, The relative benefits of environmental enrichment on learning and memory are greater when stressed: A meta-analysis of interactions in rodents. *Neuroscience & Biobehavioral Reviews* **135**, 104554 (2022).

J. Gurevitch, J. A. Morrison, L. V. Hedges, The interaction between competition and predation: A meta-analysis of field experiments. *Am. Nat.* **155**, 435–453 (2000).

P. Pottier, S. Burke, S. M. Drobniak, M. Lagisz, S. Nakagawa, Sexual (in)equality? A meta‐analysis of sex differences in thermal acclimation capacity across ectotherms. *Funct. Ecol.* **35**, 2663–2678 (2021).   
  
D. Noble, P. Pottier, M. Lagisz, S. Burke, S. Drobniak, R. E. O’Dea, S. Nakagawa, Meta-analytic approaches and effect sizes to account for “nuisance heterogeneity” in comparative physiology. *J. Exp. Biol.* **225 Suppl\_1**, jeb243225 (2022).

Second, the authors should provide more information regarding the decision to add 0.5 to all means and standard deviations (lines 174–175). Based on the equations for both lnRR and PRRD, it appears that this adjustment could have non-trivial consequences for the resulting effect sizes. Specifically, adding 0.5 to all means and SDs is not expected to have a constant or linear effect across all scales because when ratios such as lnRR are calculated, this transformation could introduce scale-dependent bias. For example, for a group with a mean of 10 and 20, adding 0.5 has a negligible effect on the ratio (10/20 = 0.5 vs. 10.5/20.5 ≈ 0.51). However, for smaller values (e.g., 1/2 = 0.5 vs. 1.5/2.5 = 0.6), the impact is more substantial. This issue may be further amplified when calculating differences between lnRRs within the same study, as is done for PRRD. In such cases, the distortion introduced by adding a constant may not cancel out; rather, it could compound the bias in unpredictable ways. I am not entirely sure how large this issue might be in practice for this dataset, but I would appreciate clarification and, if possible, evidence that the procedure does not materially affect the overall conclusions. It would also be helpful if the authors could clarify whether the addition of 0.5 was applied before or after equations 3 and 4 were calculated, as this distinction could introduce further complications, particularly given the exponential terms involved. Finally, while I appreciate the authors’ intention to retain potentially valid data, it is important to acknowledge that lnRR requires ratio-scale data, which, by definition, should not contain true zeros (i.e., a non-arbitrary zero point where zero truly means absence of the quantity). The presence of zeros challenges the interpretability of the effect sizes and raises questions about the suitability of this imputation approach.

**RESPONSE**: Originally, there were not many effect sizes that required this modification, however, we agree that this is not the best approach. We have removed these effects from our data and updated all our results, given how few there were.  
  
Last, although I often avoid suggesting this, I think this study would very much benefit from updating the search, which was conducted about three years ago (2022), and adding any new studies that may have been published since then. The number of studies included in this meta-analysis is right around the average number of studies included in meta-analyses in Ecology and Evolution (median = 41 in plant ecology, Koricheva and Gurevitch 2014; 44 in evolutionary ecology, Pollo et al. 2024), so it is not particularly small in size, but given that the comments regarding the potential need of additional data for the PRRD, as well as the difficulty to disentangle between-study (n = 40 studies) from between-species (n = 44 species) effects, which has strong consequences for interpreting whether the overall effect could be considered to be representative of all studies or not (see comments for lines 317-323 below), the study would benefit from a larger dataset, which might be achieved by updating the search.

**RESPONSE**: As suggested, we have now updated the search for the years 2022-2025. This involved screening an additional 1475 studies, however, as expected there were few additional studies. Nonetheless, we added 38 effect sizes from 3 studies as described in the methods. Dan, not sure whether this addition addresses the problem (if there is a problem!) with using PRRD or interpreting studies vs species

Best wishes, Alfredo Sánchez-Tójar  
  
Here are additional comments from my own review of the manuscript:  
•       The authors could consider introducing the term canalisation and citing relevant literature, particularly when discussing reductions in phenotypic plasticity in the Introduction. Including this concept, along with providing a clearer explanation of the term capacity for plasticity, as suggested by the reviewer, would likely improve the accessibility and clarity of the manuscript

**RESPONSE**:

RE canalisation: we are not sure that this term captures what we are showing in this analysis. We are not analysing the extent to which traits are plastic across environmental gradients (where stable mean trait values across environments could indicate canalisation), but we analyse whether the expression of plasticity differs between constant and fluctuating environments regardless of the extent to which traits are plastic.

RE capacity: we now have removed the term "capacity for plasticity" from the ms because it was misleading, i.e. we did not mean to refer to individual capacity for plasticity (cf. Loughland and Seebacher 2020 Funct Ecol 34, 1380-1390) but simply to the manifestation or expression of plasticity.

•       Please consider acknowledging the limitations of not including studies in non-English as well as the limitations of not having conducted a risk of bias assessment of the included studies.

**RESPONSE**:

•       Please report what databases (and their coverage) were included in the search performed in Web of Science.

**RESPONSE**: We now provde the information that we used the WoS Core Collection including including Current Contents, BIOSIS Previews, CAB Abstracts, Medline, Agricola and Pubmed.

•       Please provide all the versions of the R packages cited in the text.

**RESPONSE**: We now provide versions for all R packages used.

•       Line 126-127: What species? Did the authors check if this may have been resolved since the last check?

**RESPONSE**: We deleted this sentence because it did not pertain to the dataset presented here to test for plasticity.

•       Line 146-147: Please expand the explanation for “We used only biological replicates as the sample size”

**RESPONSE**: We clarified that: "We used only biological replicates as the sample size, rather than technical replicates from repeated measures using the same samples or animals if they were performed.".

•       Line 155-157: It is unclear why the authors are citing Macartney et al. 2022 here

**RESPONSE**: As discussed in our response above, Maccartney et al 2022 developed interaction-based effect sizes along very similar lines to what we do here with PRRD. There are few papers that explicitly formulate these effect sizes and it’s important that the source is cited.

We removed the citation here.

, also in line 164. Were those approaches developed by Macartney et al. 2022? If not, I would suggest citing the original sources instead.

**RESPONSE**: Yes, they were. We have retained the citation to Macartney et al. 2022. We are not aware of other ‘original’ sources that develop interaction-based effect sizes other than Gurevicth et al. 2000, which was specifically designed around Hedge’s ***g***. There are no other sources we are aware of for response ratio type effect sizes, but if the Editor knows of one we are happy to cite that paper.

Also check citation in line 209-210,

**RESPONSE**: Noble et al 2022 is an appropriate citation here as they discuss and present effect sizes that capture interaction-based effects.

I find this to be a recurrent problem in this manuscript. The authors seem to be citing studies that have used certain methodologies rather than (original) studies that have developed or suggested such methodologies. I would recommend revising that thoroughly and throughout, not only to acknowledge “original developers” (e.g., for 95% PI the authors could cite instead: IntHout et al. 2016, BMJ Open, 6, e010247) but also to avoid adding non-fully substantiated citations/shortcut citations, as they can be problematic and have cascading effects.

**RESPONSE**: We now cite IntHout et al here; We also checked references throughout to ensure that they are appropriate.

•       Line 198-199: “Note that Equation 8 assumes that constant and fluctuating treatments are independent.” Are they? Can this be assumed?

**RESPONSE**: Yes, they are in so far as we stipulated a fully factorial design where treatments are conducted concurrently with different study animals.

•       Line 205-207: Please, provide information about how that transformation was performed (e.g., exp()?)

**RESPONSE**:

•       Line 239: The latest preprint update is 2024, and I recommend using CVH2 and M2, as I believe the authors are doing for the final version of the study.

**RESPONSE**: Agreed. That is what these should have been originally. Thanks for catching this mistake. We’ve now updated the table.

•       Line 257-258: Were the effect sizes associated with these levels then excluded from the analysis or lumped in a single category?

**RESPONSE**: Dan, did we actually do this analysis?

•       Line 263-264: Visually inspecting funnel plots is not a valid method for understanding if evidence of funnel plot asymmetry exists, particularly when heterogeneity is present. Please, remove this sentence to avoid other readers thinking this technique is at all valid or useful.

**RESPONSE**: Agreed. Removed.

•       Line 264-267: Please specify how this was performed. In principle, it should not be precision but the inverse of an equivalent effective sample size and (if needed) the square root of it. The reason for not using precision is stated in section 4.3 in Nakagawa et al. 2022, MEE: “lnRR's sampling variance contains both the treatment and control means that are also contained in the point estimate. [...] This can lead to a correlation between point estimates (i.e. lnRR and SMD) and their sampling SE, resulting in ‘artefactual’ funnel asymmetry (Section 3.2; note that this issue is widespread, and also found in other standardized effect sizes, such as odds ratio and risk difference; Peters et al., 2006). Furthermore, we also notice that in Equation 4 (i.e. lnRR's variance), when sample sizes (n1 and n2) are small, the sample mean (x) and particularly, the sample standard deviation (SD) will be poorly estimated. This will result in an unreliable estimate of sampling variance (this is also the case for Equation 2).”

**RESPONSE**: We agree. However, it is important to recognise that it’s not clear exactly how to compute the effective sample size (as per eqn, 25) in our context which is why we used precision Nakagawa et al. 2022’s formula for effective sample size is for a two-group design, but here we have four. In fact, the intuition and logic behind eqn 25 is not entirely clear even for a 2-group design, it’s more obvious for eqn. 26, which assumes sample sizes across groups is the same.

We have consulted the original sources for effective sample size and explored the variation in sample sizes across the four groups more closely because, if all group sample sizes are the same, then we should be able to apply a simple summation of inverse sample size to approximate the effective sample size as is done in a two-group design (eqn, 26, Nakagawa et al. 2022).

•       Line 317-320: Even if I would generally agree with: “However, our findings were remarkably consistent across species with low species-level heterogeneity in effects, which suggests that the small overall effects we observed can be generalised across species when between- and within-study sources of variation have been controlled (Yang et al. 2023).”, I also agree with the reviewer that your dataset is not ideal for disentagling between-study (n = 40 studies) vs between-species (n = 44 species), and in fact, since I2Study explains 38% of heterogeneity, this suggests that on average, studies differ in the overall support to the hypothesis, which further suggest high context dependency beyond within-study variability. That is, temperature variability seems to influence phenotypic plasticity in some contexts.

**RESPONSE**: We re-wrote this section to take these comment (as well as the reviewer's comments) on board, and we tone down our conclusion, and we point out explicitly that "...number of species and number of studies were very similar in our dataset (i.e., each study tended to use a unique species) so that we may not be able to disentangle study- and species-specific effects very robustly.

•       Line 322-323: Explaining 30% of heterogeneity is a noteworthy achievement, especially when this is done using a moderator with only four levels. This suggests that the explained variance is unlikely to be a statistical artefact, which can occasionally happen when moderators have many levels. Looking at Figure 3b, I can't help but wonder whether the near-zero overall effect observed in the intercept-only models could be, at least in part, due to opposing directions of effects within the dataset (i.e., essentially, effects cancelling each other out). One possible source of this could be inconsistent coding of effect direction across traits, for example, mixing survival and mortality effects without adjusting the sign accordingly. I would appreciate it if the authors could clarify whether steps were taken to ensure that the meaining of the direction (sign) of effect sizes is consistent across the dataset. If such steps were unnecessary for this analysis, a brief explanation of why that is the case would also be helpful in clarifying this.

**RESPONSE**: We have explored this

•       Line 343-345: I agree with the reviewer that this conclusion — and potentially others — requires substantial revision. Based on the current analyses, including the results shown in Figure 3b and the high levels of heterogeneity, of which more than one-third is attributable to differences between studies, the statement that “diel fluctuations do not impact plasticity” is not fully supported. In particular, interpreting a statistically non-significant effect as evidence of no effect is problematic, especially when heterogeneity is high. More nuance is needed when interpreting non-significant results, particularly in the presence of substantial heterogeneity. For example, a p-value of 0.18 combined with considerable unexplained variance suggests uncertainty, not confirmation of the absence of effect. I encourage the authors to adjust the wording of these conclusions to better reflect the limitations and uncertainty in the current evidence.

**RESPONSE**: We edited the text to read: "The findings of our study overall indicate that there is little support in the current literature that phenotypic plasticity is influenced by regular diel fluctuations around mean temperatures.".

•       Figure S2: It would be important for the authors to clarify how they handled cases where studies did not report all the necessary information for inclusion in the meta-analysis. For example, missing sample sizes or standard deviations. Did the authors exclude such studies, attempt to contact the original authors, or apply imputation methods for missing values? I would generally advise against excluding studies outright in these cases, as this can introduce bias and reduce the comprehensiveness of the analysis. Providing a brief explanation of the approach taken would help ensure transparency and allow readers to assess the robustness of the dataset.

**RESPONSE**: We now added the following information in the Data extraction section: "None of the screened studies were excluded because of missing information stipulated in our inclusion criteria (Fig. S2), but for one study (Breitenbach *et al.* 2020)  we accessed the published dataset to calculate standard deviations.".