Strapline: Correspondence

No empirical evidence to support the hypothesis that daily climate variation has an effect on species' elevational range size: Reply to Chan et al.

Richard Field[1] and Hong Qian[2,*]

[1]School of Geography, University of Nottingham, Nottingham, UK.
[2]Research and Collections Center, Illinois State Museum, Springfield, IL, USA.

*To whom correspondence should be addressed. E-mail: hqian@museum.state.il.us

ORCID
Richard Field https://orcid.org/0000-0003-2613-2688
Hong Qian http://orcid.org/0000-0002-1381-7496

Keywords: Diurnal temperature range, Gilchrist's hypothesis, seasonal temperature range, species elevational range, structure equation modelling.

Running head: Climate variation and elevational range of species

Abstract

Following our critique, Chan et al. defend the approach used in their original paper. They reveal their “iterative strategy of SEM” (structural equation modelling), which they claim is “standard” (we show otherwise) and “required for the proper and most effective use of SEM” for hypothesis testing. However, publishing their detailed procedure exposes fundamental flaws: capitalizing on chance and violating important assumptions and principles of SEM. They used the same data to first explore numerous correlations, then fit 29 candidate models (all failed) using the best correlates, then fix model parameters to gain degrees of freedom, then evaluate the ‘best’ model. In producing the ‘best’ model, they fixed five parameters using estimates from regression on the same dataset. They further argue that their stationary bootstrap cures the problems of bias and pseudoreplication; we disagree. At best, Chan et al. developed a hypothesis; they did not perform a valid test of one.

In our critique (Qian et al., 2017), we demonstrated serious flaws and errors in the paper by Chan et al. (2016). We showed that key results they reported could not be repeated using their data and the methods described in their paper; their model should have been rejected. Even overlooking that, we also showed, using very simple techniques to exemplify some of the problems as clearly as possible, that correcting some of the biases and pseudoreplication in their data (geographical and taxonomic) overturned the main conclusion of their study each time. In reply, Chan et al. (2018) argue that we did not understand their modelling approach, and state that most—but far from all—of their results had originally been reported correctly. We did not understand their modelling approach because they failed to even mention the ‘model modification’ they performed, let alone provide any details. They now admit that they did not specify the details of their analyses in their original paper, and express regret about it.

Importantly, Chan et al. (2018) state that in our critique we “did not follow the standard, iterative process of SEM”, and even say that this “suggests a misunderstanding regarding the iterative nature of SEM analysis”. They strongly imply that we should somehow have figured out what they did and thus be able to repeat their analyses without any of the details they have now provided. We challenge anyone to read the original paper (Chan et al., 2016) without the new Additional Supplementary Materials that have now been added to it and be able to repeat the analyses with no further information! Certainly none of the (many) people we have asked about it have even got close to guessing what they actually did. Further, as we demonstrate below, the results presented in the original paper seem to contradict the modelling procedure that they now describe. We conclude from this that Chan et al. (2016) failed to make their science repeatable. This has now been largely rectified. Crucially, however, this failure also meant that referees of the original paper were unable to evaluate the modelling approach that was actually used. Here we provide criticism that was thus not possible previously.

Chan et al. (2018) argue that their analyses were appropriate and that the methods that we used to reanalyze their data were inappropriate. We reject both these points. Now that we know more about what they did, a new set of very serious concerns arises, but our original conclusion remains correct, and is strengthened: their data and analyses do not support their central conclusion of a significant negative effect of daily temperature variation on elevational ranges of species. We expand on these issues below.

Modelling approach

Now that we know about it, we here argue that the modelling approach of Chan et al. (2016) is the most fundamentally flawed aspect of their analyses. The new Additional Supplementary Materials of Chan et al. (2016) give the details; here we try to capture the essence, highlighting what is wrong with their modelling procedures. First, a quick recap: Chan et al. (2016) presented an SEM model as an empirical validation of a 'novel macrophysiological principle'. Actually, they ran 29 SEM analyses, after using hierarchical partitioning to determine which were the best correlates of the response—which seems incongruous with their claim that they began by "building models that represent a group of theoretically-meaningful, candidate hypotheses, based on the literature and our *a priori* subjective judgment about causes and processes among geographic factors, climatic factors, and elevational range sizes". They only presented the 'best' model in their main text; to keep focus on the key issues, we refer almost exclusively to that 'best' model here, but the problems apply to the modelling approach as a whole.

The “novel macrophysiological principle” was supposedly validated by one of the paths within the ‘best’ model: a direct effect of daily temperature range on mean (elevational) range size (hereafter ‘DTR→MRS’). However, when we ran the model, as presented in their paper, with the same data, we found that it failed the χ^2 test, and failed the authors’ own criteria for model-fit statistics (specifically root mean square error of approximation). In both cases, the results presented in their original paper differed from those we obtained: their reported results suggested an acceptable model while our results indicated that the model should have been rejected. This was the most fundamental issue in our critique.

In their response, Chan et al. reveal that they actually got the same results as we did: their model did indeed fail the two tests. However, rather than rejecting the model, they instead modified it by (if we understand their description correctly) sequentially fixing (‘specifying’) five of the parameters. They did this using values obtained from regression and summary statistics of the **same data**, thereby gaining degrees of freedom, until the SEM model passed the tests. The key outcome was to change the degrees of freedom in the ‘best’ SEM model by 5, so that the χ^2 test changed from 2 degrees of freedom to 7 and this changed the model from failing the χ^2 test to passing it ($\chi^2 = 10.6$; with 2 d.f. $P = 0.005$ but with 7 d.f. $P = 0.157$). Fixing these 5 parameters also made the problems with the model-fit statistics go away.

We were particularly surprised to learn that one of the fixed parameters was the one that supposedly demonstrates the novel macrophysiological principle: DTR→MRS. Given that this parameter was fixed, we do not understand why it was presented by Chan et al. (2016) with *P*-values (e.g. in the SEM in their Fig. 1A they showed it as $P < 0.05$ and in their Table S3 they listed it as a “parameter estimate for the structural equation model” with $P = 0.012$). Similarly, it is not clear to us why, when the approach was to specify particular parameter values (and the same set across the models, where possible), the values specified for the same parameter were different in the different SEMs (e.g. three different values for DTR→MRS in their Table S3). Surely DTR→MRS is either a fixed parameter or a parameter to be estimated by the model, but not both! This appears to be a contradiction, with important ramifications for how readers interpret the results—the path representing the ‘novel macrophysiological principle’ is presented as having been specifically tested statistically, when actually it was a fixed parameter (despite being estimated from the data).

To put it as simply as we can, Chan et al. (2016) took an SEM model that was rejected by the data. They then used the same data to estimate parameters, and used these parameter estimates as exact, fixed parameters in the SEM, which they then reran with the same data. Their stated reason is “to achieve acceptable degrees of freedom”. They sequentially fixed parameters in this way until there were enough degrees of freedom for the model to pass the tests and be deemed acceptable. They then presented this model as empirical evidence for the ‘novel macrophysiological principle’, but failed to mention that any parameters had been fixed—thus referees would have had no opportunity to question or assess the validity of the procedure. Crucially, the modelling reported in their 2016 paper implies a confirmatory analysis, which must have been how editors and referees would have judged it, but actually the analysis was, at best, a preliminary exploratory exercise. Such an exercise should, at most, be used to establish hypotheses for testing on independent data (Grace, 2006), not for claiming to have found a ‘novel macrophysiological principle’.

Chan et al., in their response paper, suggest that it is fine, and even good practice, to proceed in the way just described, and reveal that they fixed parameters in **all** of their 29 SEM models. They state that this approach is ‘standard’, in ecology and beyond. We sampled 30 papers recently published in ecological journals that have used SEM (specifically, we used the phrase “ecolog* and structural equation model*” to search ISI Web of Science and randomly picked one from every 10 resulting records for checking until we checked 30 papers in which SEM was used). None of these papers used the process described by Chan et al. We specifically checked SEM-based papers published by SEM experts, including James B. Grace, who was cited by Chan et al. in support of their approach (but see next paragraph), and was a co-author of the following SEM-based papers: Chen et al. (2013), Eldridge et al. (2013), Miller et al. (2015), and Knick et al. (2017). We did not find the supposedly ‘standard’ process described by Chan et al. (2018) in any of these papers. Although we have not attempted to sample the literature beyond ecology, we did check the non-ecological SEM book that Chan et al. cite to support their approach (Hoyle, 2012), and found no evidence that their modelling approach is either common or recommended (see next paragraph). Thus, counter to the claim of Chan et al., we conclude that their approach is not standard.

Much worse than not being standard, the approach is seriously flawed. Chan et al. cite two SEM books to support the legitimacy of their approach (Grace, 2006; Hoyle, 2012 [especially the chapter by Chou & Huh, 2012]). On the contrary, however, these books repeatedly warn about the dangers of capitalization on chance. Grace (2006) most clearly described various fundamental problems that are crucial here, in his short section on model modification (pp. 133–134), including this passage in which the italics are his:

> "When a model is found to have an inadequate fit to a set of data, we must recognize that our statistical model has failed and also that our estimates of structural coefficients are deemed invalid. At this point, we may either simply reject the hypothesized model or, as is typically the case, we may seek to discover a model that is consistent with the data. *It is important to be aware that any subsequent models developed or results achieved are exploratory until an independent data set is obtained to evaluate the adequacy of the revised model.* This fundamental tenet arises from the fact that structural equation model evaluation is a *confirmatory process*. By this we mean that the χ^2 test is designed to confirm or reject our proposed model, not to inform us of what model might fit the data if we had been clever enough to guess it in the first place. When one explores a data set using SEM in order to generate a hypothesis, one violates certain assumptions of the method."

We suggest that the procedure Chan et al. (2016) followed is so fundamentally flawed as to be almost circular. They effectively used degrees of freedom to estimate parameters, and then fixed those parameters so they could re-use the degrees of freedom. At best this uses the data to generate a model (or hypothesis) that then needs to be tested using an independent dataset, but Chan et al. did not attempt any test with an independent dataset. Thus we maintain our original conclusion that their main advance "should be regarded as currently remaining unsupported by empirical evidence".

Other issues

Before we address further points newly raised by Chan et al. (2018), two other SEM-related issues deserve a mention. First, we note from the new supplementary material appended to

Chan et al. (2016) that when they repeated their own SEM analyses, they found that 8 of the 29 models reported in their original paper were incorrect. Second, they question whether we did actually try repeating their analysis using the same software as they used (AMOS), which we said we did. The basis for their accusation is that we reported code and results in our supplementary material that used LAVAAN in R. Yet it is clearly explained on the second page of Qian et al. (2017): “corrected values (highlighted in blue) which resulted from when we ran the same SEM using the same data, model and software as in Chan et al. We also used the LAVAAN package to repeat the analysis (see Appendix S1 in Supporting Information for code and results when repeating the analyses using the LAVAAN package”. This clearly indicates that we (i.e. Qian et al., 2017) used both AMOS and LAVAAN in R to re-analyze their data. In our supplementary material we chose to provide the R code and its results because this allows anyone with the dataset to exactly repeat our re-analysis in commonly used, freely accessible software. Indeed, a referee ran this code on the data and successfully repeated our results. In contrast, reproducing the results of Chan et al. (2016) using their new supplementary materials requires expensive commercial software.

We do not wish to enter a turgid debate on details of biases and pseudoreplication issues; the problems addressed above are more fundamental. However, in their response to our critique, Chan et al. (2018) devote considerable space to these issues, so we cannot completely ignore them. They state that “Qian et al. argued against conducting a single, taxonomically comprehensive analysis”. This misinterprets our view. We are not “against conducting” such an analysis per se, but consider that it should be subjected to the sort of scrutiny that should be applied to any model before publication. Chan et al. argue that the problems of bias and pseudoreplication that we demonstrated are not problems, mainly because their stationary bootstrap procedure deals with all such issues. We do not think this is valid, though we do not profess to be experts on stationary bootstrapping. Our understanding is that this method is a type of block bootstrap, and as such the length of the block becomes important, affecting whether it reduces (or even increases) biases. Chan et al. (2016) failed to report any information on the block lengths (including whether or not they were random); their response to our critique now provides some clarification. Further, research suggests that block bootstrap methods (including the stationary bootstrap), which were developed for time-series and gridded spatial data, may not work for irregularly spaced spatial data (Lahiri & Zhu, 2006). Given these considerations, we are not convinced that the stationary bootstrap (including fitting locations as random effects) reliably rebuts our criticisms of bias and pseudoreplication.

Chan et al., in response to our criticisms of biases and pseudoreplication, repeatedly state that setting up "arbitrary dichotomies" is not as good as using techniques that use all the data and try to control the problems statistically. This misses the point that the very simple analyses we presented were merely to demonstrate the problems as simply and clearly as possible; we were not claiming that data splitting is 'the solution'. Chan et al. criticize our use of simple regression to analyze data for dry and humid sites separately, and also McCain's (2009) use of simple regression. To take the use of simple regression first, this and correlation analysis may indeed miss important and significant relationships when confounding variables are not considered in the analysis, as Chan et al. point out. This is why we did not restrict our reanalysis to these methods. However, a more complex model such as SEM may produce spurious relationships when different variables in the model are not appropriately related and some important variables are not included in the model. This is more likely for a weak relationship such as the DRT→MRS relationship representing the main 'advance' of Chan et al. (2016).

Second, Chan et al. (2018) criticize us for using a dichotomous approach to divide sites into two groups, in contrast to their approach: "to deal with these sampling issues, we conducted formal statistical analyses that use all the data and do not depend on arbitrary dichotomies." Contradicting this, however, in Chan et al. (2016), they used a dichotomous approach to divide the 180 sites of McCain (2009) into two groups: those with mountain gradients spanning >2000m and ≤2000 m, respectively. The 2000-m cut-off point is just as arbitrary as any other cut-off point. They excluded mountain gradients spanning ≤2000 m from their analysis, which seems to have influenced their results and conclusions (biased them, arguably), as follows. When they analysed all 180 sites, the relative importance of DTR→MRS was smaller than that of STR→MRS (seasonal temperature range; Figure S2a of Chan et al., 2016). However, when mountain gradients spanning ≤2000 m were excluded, the relative importance of DTR→MSR increased nearly threefold, and exceeded that of STR→MRS (Figure S2b of Chan et al., 2016). Following their logic, this suggests that the greater importance of DTR→MSR than STR→MRS, which they stress in their response to our critique, may simply result from their arbitrary data splitting and exclusion. Our purpose in dividing the 137 sites into two groups based on an unbiased humidity index (following McCain, 2009) was to select an appropriate proportion of dry sites from which to repeatedly sample, to correct the over-representation of dry sites in Chan et al. (2016). Thus our aim was to test the effect of reducing the bias in the analysis in Chan et al. (2016). That we provided analysis in which SEMs were built separately for dry and humid sites was simply to demonstrate that the SEM results differ significantly between the two groups of sites, which supports our point that the over-representation of dry sites would have biased the results of Chan et al. (2016). We did not suggest that dry sites and humid sites should be analyzed separately for other purposes. In fact, our final analysis included both dry sites and humid sites (Figure 2c of Qian et al., 2017).

Chan et al. (2018) agree that the sample sites used in their study were not evenly distributed across Earth’s continents or among taxonomic groups. We demonstrated (in Figure 2c of Qian et al. 2017), using the very simple analysis just described, that accounting for just one of many sampling biases in Chan et al. (2016) overturned the main conclusion of their study. They have not provided evidence that their stationary bootstrap corrected biases in the path coefficients of their SEM. Instead, they present an analysis (Figure 1 of Chan et al., 2018) to argue that dry sites were actually not over-represented in their original analysis. Specifically, they compare the area frequency distribution of their study sites with that of global terrestrial area above 2000 m in elevation along a precipitation gradient. This is flawed because their study sites (used in SEM) covered elevations both below and above 2000 m. Comparing the area of full elevational gradients with that of >2000 m is similar to comparing oranges with apples; importantly, precipitation typically increases with elevation, including in regions with dry climates.

We take this opportunity to stress that throwing everything into a mixed-effects model is not a panacea. Such models can certainly be helpful, if used with sufficient care and thought, but they should not replace model checking. In their response to our critique, Chan et al. appear to confuse “subjective data manipulation” with model checking. They also ask why the two points with lowest DTR are “not just as ‘influential’” as the two with the highest DTR (which we highlighted in our critique). These four points are labelled 3 and 4, and 1 and 2, respectively, in Figure 1. The specific answer is simple: a point's influence is a function of both how different its value for the predictor variable(s) is from the mean of that predictor (i.e. its leverage) and how different its value for the response is from the value predicted by all the other points (i.e. its distance). Points 3 and 4 in Figure 1 have similar leverage to points 1 and 2, but their distance is less. The highly influential nature of points 1 and 2 is also clear visually, as illustrated in Figure 1 by a grey circle centred on 0,0 that also highlights the main pattern (no relationship between the variables) in the rest of the points. Nonetheless, it is worth noting that points 3 and 4 are also pseudoreplicated (see Figure 1 legend), and they are also quite influential, again illustrating the sensitivity of the key result of Chan et al. (2016) to a very small number of non-independent points. (This is all in the context of a model that should have been rejected in the first place, as explained above.)

Finally, in response to the criticism that, even if one ignores all the problems with the analyses of Chan et al. (2016), their 'novel macroevolutionary principle' accounts for only a very small proportion of the variation in elevational range size (as does the entire 'best' model), Chan et al. (2018) say "Qian et al. (2017) criticized our decision not to report R^2 in our SEM analysis. Because a large number of variables and relationships were included in the SEM analysis (i.e., MAP, Latitude, DTR, STR, mean range size), modest values of R^2 are expected for the overall fit of the model. It is precisely for this reason that many researchers emphasize AIC over R^2 in model comparisons." This response is baffling! First, larger numbers of parameters fit (or fit then fixed) would increase R^2. Second, AIC is usually used as an alternative to *P*-values (on which Chan et al. (2016) relied heavily), and not as an alternative to R^2, which remains very important because it measures effect size. Most, if not all, ecological studies based on SEM have reported values of R^2 for their models (e.g. Hawkins & Porter, 2003; Hawkins et al., 2007; Jetz et al., 2009; Spitale et al., 2009; Jonsson et al., 2011), including those by SEM experts (e.g. James B. Grace: Oberle et al., 2009; Chen et al., 2013; Eldridge et al., 2013; Miller et al., 2015; Knick et al., 2017). In the case of Chan et al. (2016), the effect size, even if present at all, was very small ($R^2 = 0.11$ for their SEM; see Figure 1c of Qian et al., 2017)—a fact they failed to mention.

In summary, the analyses of Chan et al. are biased and suffer pseudoreplication; each of the simple illustrative analyses that we reported to address small parts of these problems resulted in the evidence for Chan et al.'s (2016) main finding disappearing. Most importantly, their main modelling procedure was fundamentally flawed, confusing exploratory analysis with confirmatory analysis, and capitalizing on chance. At best, Chan et al (2016) only developed a hypothesis (which they attributed to Gilchrist, 1995) that still requires empirical testing. We conclude exactly as in our original critique: the "novel macrophysiological principle" that represents the main advance in their paper should be regarded as currently remaining unsupported by empirical evidence.

ACKNOWLEDGMENTS

We wish to thank two anonymous referees for very helpful comments on our manuscript. One of them, in particular, must have devoted a lot of time to reading and scrutinizing all four manuscripts (the original Chan et al. paper, our response, and the two follow-up papers published here). We consider it very important for science that debates like these are allowed, and for them to be properly overseen. The comments of this referee were very insightful and we have paraphrased some of them herein. We also thank Richard Pearson, Michael Dawson and Peter Linder for their oversight and constructive handling of our manuscript, under unusual circumstances.

REFERENCES

Chan, W.-P., Chen, I.-C., Colwell, R.K., Liu, W.-C., Huang, C.-Y., & Shen, S.-F. (2016) Seasonal and daily climate variation have opposite effects on species elevational range size. *Science*, **351**, 1437–1439.

Chan, W.-P., Chen, I.-C., Colwell, R.K., Liu, W.-C., Huang, C.-Y., & Shen, S.-F. (2018) Response to Qian et al. (2017): Daily and seasonal climate variations are both critical in the evolution of species' elevational range size. *Journal of Biogeography* DOI-XXX

Chen, D., Lan, Z., Bai, X., Grace, J.B., & Bai, Y. (2013) Evidence that acidification-induced declines in plant diversity and productivity are mediated by changes in below-ground communities and soil properties in a semi-arid steppe. *Journal of Ecology*, **101**, 1322–1334.

Chou, C. & Huh, J. (2012) Model modification in structural equation modeling. *Handbook of structural equation modeling* (ed. by R.H. Hoyle), pp. 232–246. Guilford Press, New York, USA.

Eldridge, D.J., Beecham, G., & Grace, J.B. (2013) Do shrubs reduce the adverse effects of grazing on soil properties? *Ecohydrology*, **51**, 371–381.

Gilchrist, G.W. (1995). Specialists and generalists in changing environments. I. Fitness landscapes of thermal sensitivity. *The American Naturalist*, **146**, 252–270.

Grace, J.B. (2006) *Structural equation modeling and natural systems*. Cambridge University Press.

Hawkins, B.A. & Porter, E.E. (2003) Does herbivore diversity depend on plant diversity? the case of California butterflies. *The American Naturalist*, **161**, 40–49.

Hawkins, B.A., Diniz-Filho, J.A.F., Jaramillo, C.A., & Soeller, S.A. (2007) Climate, niche conservatism, and the global bird diversity gradient. *The American Naturalist*, **170**, S16–S27.

Hoyle, R.H. (2012) *Handbook of Structural Equation Modeling*. Guilford Press, New York, USA.

Jetz, W., Kreft, H., Ceballos, G., & Mutke, J. (2009) Global associations between terrestrial producer and vertebrate consumer diversity. *Proceedings of the Royal Society of London B*, **276**, 269–278.

Jonsson, M., Englund, G., & Wardle, D.A. (2011) Direct and indirect effects of area, energy and habitat heterogeneity on breeding bird communities. Journal of Biogeography, 38, 1186–1196.Knick, S.T., Hanser, S.E., Grace, J.B., Hollenbeck, J.P., & Leu, M. (2017) Response of bird community structure to habitat management in piñon–juniper woodland-sagebrush ecotones. *Forest Ecology and Management*, **400**, 256–268.

Knick, S.T., Hanser, S.E., Grace, J.B., Hollenbeck, J.P., & Leu, M. (2017) Response of bird community structure to habitat management in piñon-juniper woodland-sagebrush ecotones. *Forest Ecology and Manage*ment, **400**, 256–268.

Lahiri, S.N. & Zhu, J. (2006) Resampling methods for spatial regression models under a class of stochastic designs. *The Annals of Statistics*, **34**, 1774–1813.

McCain, C. (2009) Vertebrate range sizes indicate that mountains may be ‘higher’ in the tropics. *Ecology Letters*, **12**, 550–560.

Miller, J.E., Damschen, E.I., Harrison, S.P., & Grace, J.B. (2015) Landscape structure affects specialists but not generalists in naturally fragmented grasslands. *Ecology*, **96**, 3323–3331.

Oberle, B., Grace, J.B., & Chase, J.M. (2009) Beneath the veil: plant growth form influences the strength of species richness–productivity relationships in forests. *Global Ecology and Biogeography*, **18**, 416–425.

Qian, H., Field, R., Zhang, J., & Zhang, Y. (2017) Does daily climate variation have an effect on species' elevational range size? *Journal of Biogeography*, **44**, 2432–2436.

Spitale, D., Petraglia, A., & Tomaselli, M. (2009) Structural equation modelling detects unexpected differences between bryophyte and vascular plant richness along multiple environmental gradients. *Journal of Biogeography*, **36**, 745–755.

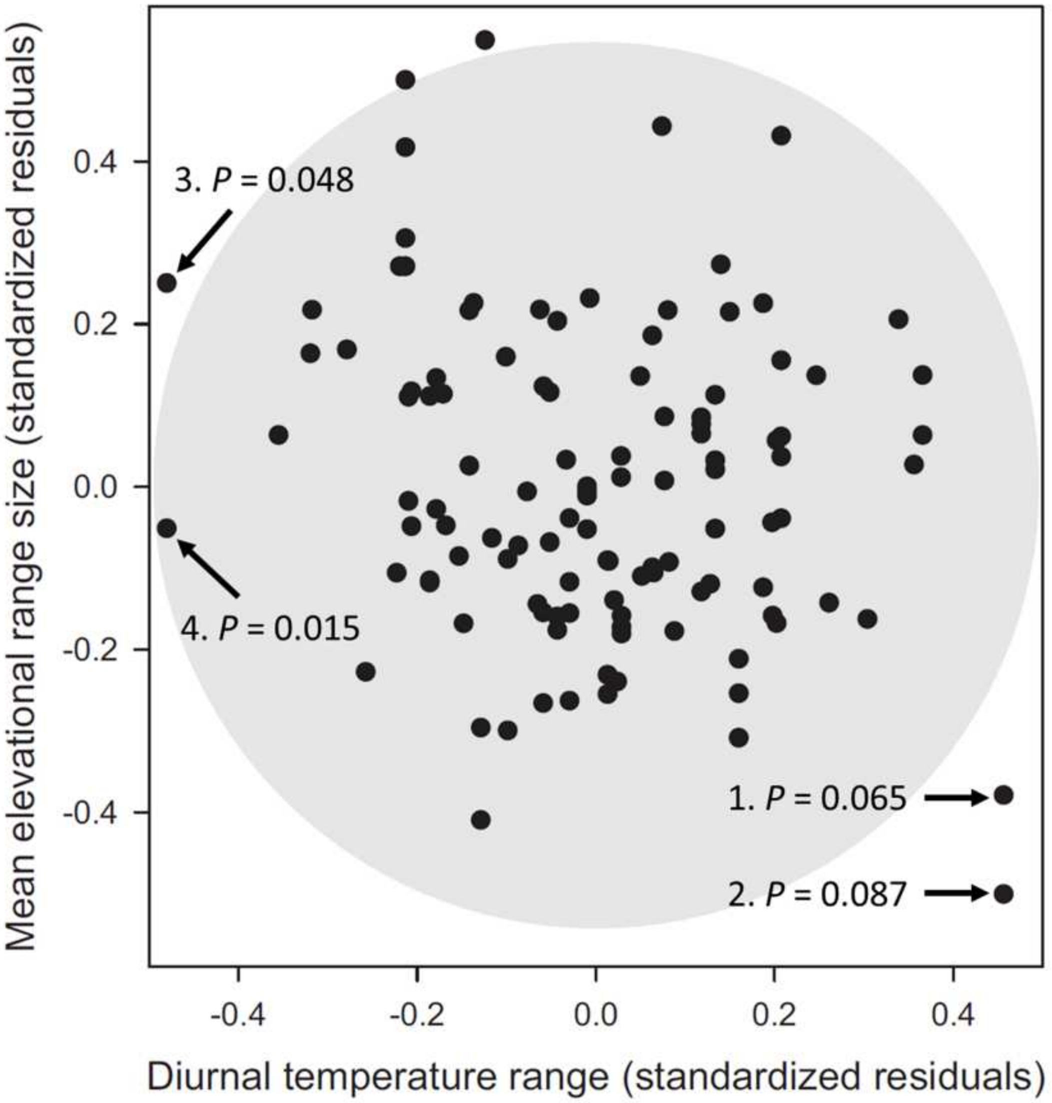

Figure 1 Partial residual plot of the modelled relationship between diurnal temperature range and mean elevational range size, illustrating the sensitivity of the key finding of Chan et al. (2016) to a few pseudoreplicated points. This is the same plot as Figure 3 in Qian et al (2017), except for the labelling of pseudoreplicated points with high leverage. The *P*-values shown refer to the key path (‘DTR→MRS’) in the ‘best’ model of Chan et al. (2016) when the point in question is removed. Compare with *P* = 0.024 with none of the points removed. Points 1 and 2 represent two samples of reptile taxa from the same study in the same study site, with identical values for all the environmental variables. Points 3 and 4 represent two samples of amphibian taxa from the same study in the same study site, with identical values for all the environmental variables. When two of the points, one from each pair, are removed, the resulting *P*-values for DTR→MRS range from 0.044 (points 1 and 4 removed) to 0.156 (points 2 and 3 removed). Note that this significance testing is done, as in Chan et al. (2016), within a model that fails the diagnostic tests (see text). The grey shading is referred to in the text.