Published in final edited form as:
Nat Ecol Evol. 2020 November 01; 4(11): 1502–1509. doi:10.1038/s41559-020-1256-9.

Thresholds for ecological responses to global change do not emerge from empirical data

Helmut Hillebrand[1,2,3,✉], **Ian Donohue**[4], **W. Stanley Harpole**[5,6,7], **Dorothee Hodapp**[2,3], **Michal Kucera**[8], **Aleksandra M. Lewandowska**[9], **Julian Merder**[10], **Jose M. Montoya**[11], **Jan A. Freund**[12]

[1]Plankton Ecology Lab, Institute for Chemistry and Biology of the Marine Environment, Carl von Ossietzky University Oldenburg, Oldenburg, Germany

[2]Helmholtz-Institute for Functional Marine Biodiversity at the University of Oldenburg (HIFMB), Oldenburg, Germany

[3]Alfred-Wegener-Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany

[4]School of Natural Sciences, Department of Zoology, Trinity College Dublin, Dublin, Ireland

[5]Department of Physiological Diversity, Helmholtz Center for Environmental Research – UFZ, Leipzig, Germany

[6]German Centre for Integrative Biodiversity Research, Leipzig, Germany

[7]Martin Luther University Halle-Wittenberg, Halle, Germany

[8]MARUM – Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany

[9]Tvärminne Zoological Station, University of Helsinki, Hanko, Finland

[10]Marine Geochemistry, Institute for Chemistry and Biology of the Marine Environment, Carl von Ossietzky University Oldenburg, Oldenburg, Germany

[11]Centre for Biodiversity Theory and Modelling, Theoretical and Empirical Ecology Station, CNRS and Paul Sabatier University, Moulis, France

[12]Theoretical Physics/Complex Systems, Institute for Chemistry and Biology of the Marine Environment, Carl von Ossietzky University Oldenburg, Oldenburg, Germany

✉**Correspondence and requests for materials** should be addressed to H.H. helmut.hillebrand@uni-oldenburg.de.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Author contributions

H.H. designed the analysis and discussed the framework with I.D. J.M.M., J.A.F. and J.M. developed the statistical approach with input from H.H. and D.H. H.H. assembled the effect size information. J.A.F. and J.M. performed the statistical analysis. M.K., A.M.L. and W.S.H. provided input on palaeoecological and experimental constraints, respectively. H.H. wrote the manuscript together with W.S.H. and J.M.M. as well as input from all co-authors.

Competing interests

The authors declare no competing interests.

Peer review information Peer reviewer reports are available.

Europe PMC Funders Author Manuscripts

Abstract

To understand ecosystem responses to anthropogenic global change, a prevailing framework is the definition of threshold levels of pressure, above which response magnitudes and their variances increase disproportionately. However, we lack systematic quantitative evidence as to whether empirical data allow definition of such thresholds. Here, we summarize 36 meta-analyses measuring more than 4,600 global change impacts on natural communities. We find that threshold transgressions were rarely detectable, either within or across meta-analyses. Instead, ecological responses were characterized mostly by progressively increasing magnitude and variance when pressure increased. Sensitivity analyses with modelled data revealed that minor variances in the response are sufficient to preclude the detection of thresholds from data, even if they are present. The simulations reinforced our contention that global change biology needs to abandon the general expectation that system properties allow defining thresholds as a way to manage nature under global change. Rather, highly variable responses, even under weak pressures, suggest that 'safe-operating spaces' are unlikely to be quantifiable.

Concepts of thresholds, tipping points and regime shifts dominate current ecological frameworks aiming to understand ecosystem responses to anthropogenic global change[1–4]. A threshold corresponds to a level of environmental pressure that creates a discontinuity in the ecosystem response to this pressure. Thresholds and tipping points pervade environmental policy documents[5,6] as they allow definition of levels of pressure below which ecosystem responses remain within 'safe ecological limits'[6] and above which response magnitudes and their variances increase disproportionately[7,8]. Anticipating when and under what conditions such threshold transgression might occur is important for sustainable environmental management.

Threshold-related concepts and their implementation in policy hinge upon the assumption that the presence of thresholds can be detected in data or—even better—predicted. Testing this assumption requires knowledge of the ecosystem response to an environmental pressure for present-day and potential future pressure magnitudes. Ecological meta-analysis has led to the publication of thousands of effect sizes in response to in-situ trends or experimental manipulations of key pressures of global change such as eutrophication, warming, land-use change, fisheries and ocean acidification. Each study in a meta-analysis quantifies the magnitude of the response of an ecosystem variable to the strength of an applied environmental pressure (Fig. 1a). The entire set of studies in the meta-analysis then represents a wide range of pressure strengths, which often exceed the conditions observed in nature but might be expected in future ecosystems. We capitalize on this richness of data by combining available information from 36 meta-analyses, providing 4,601 effect sizes across ecosystems and pressures of global change into multiple tests of whether these datasets—individually or aggregated—reveal a response pattern that indicates transgression of a threshold (Fig. 1b). We first tested whether and how ecosystems respond to increased environmental pressures by simply exploring whether ecosystems show a directional change in response to a pressure, regardless of the presence of a threshold (Fig. 1c). Second, we quantified discontinuities in the variance of responses, which would be a way to define the existence of a threshold. Finally, we tested for existence of multimodality of responses,

Europe PMC Funders Author Manuscripts

which would be stronger evidence for alternative states under different environmental pressures.

Results

To test for general changes of systems along gradients of environmental pressures, we used an averaged Kullback–Leibler (KL) divergence method (see Methods) to quantify the overall deviation between the response distribution for a given stressor value and the marginal response distribution; that is, the response distribution when collapsing all response data onto a single axis ignoring the magnitude of the stressor variable. Most meta-analyses (23 of 36) showed changes in the response magnitude along the gradient of pressure strengths (KL; Supplementary Table 1). This provides strong evidence that direction and increasing magnitude of global environmental pressures have significant effects on ecosystem variables. While necessary, this evidence is not sufficient to support the general prevalence of threshold-type responses across ecosystems.

If thresholds are common, then we expect to see increased variance in response variables as the pressure strength crosses the threshold value[7,8] (see Fig. 1c). To test for discontinuities in the variance of effect size responses, we used a weighted quantile ratio (QR) of interquantile range (95–5%) to quantify substantial inhomogeneity in the width of the response distribution across the range of observed stressors (see Methods). Significant changes in the variance of effect sizes were present in only eight out of 36 cases (QR; Supplementary Table 1), challenging the widespread expectation of rising variance as a signal of threshold transgression. Moreover, in those cases with a significant QR test, the increase in variance occurred frequently only at the most extreme pressure level observed in the respective meta-analysis (see later for further details).

Stronger evidence for threshold-type ecosystem responses to increasing environmental pressure would be provided by the existence of multimodal distributional patterns, reflecting a state transition. We used Hartigan's dip test method (HD; see Methods) to assess the multimodality of effect sizes[9], which provides a narrow test for the case of bi-(multi)-stability of responses. We found no support for widespread existence of alternative states in ecological responses to increasing pressure intensities. None of the 36 meta-analyses revealed any sign of bimodality in the frequency distribution of effect sizes (HD, $P > 0.3$ in every case; Supplementary Table 1).

Comparing these empirical results (Supplementary Table 1) to model data (Fig. 2 and Extended Data Fig. 1) with known presence or absence of thresholds shows that our three approaches are suitable to detect threshold transgression. For idealized data, the three tests provide a clear differentiation between gradual and threshold-associated disproportional changes in response magnitudes. However, empirical observations will be affected by different sources of variance, both systematic (cases with different locations of thresholds and magnitudes of response shift) and stochastic. With increasing noise-to-signal ratios, thresholds—although present—quickly become undetectable, as the power of QR and HD declines rapidly. The exponential decline in detection probability for QR shows that thresholds can only be identified reliably for nearly ideal data without random variation

around the response magnitude (scenarios g–i in Fig. 2), with the exception of the unlikely case that all systems are characterized by the same threshold (scenario f in Fig. 2). For HD, the power collapses completely with only moderate noise levels (Fig. 2). Only KL is still able to detect changes in response magnitude with increasing pressure with increasing variance, either around gradual shifts in response magnitude (scenarios c and d in Fig. 2) or around thresholds (scenarios e–i). The simulations corroborate our general empirical finding across the 36 datasets that thresholds are rarely detectable in data even if using statistical methods developed for threshold detection.

Even when thresholds were empirically detected, limited inference can be made as shown by highlighting several individual meta-analysis datasets to illustrate specific ecosystem responses to particular environmental pressures. The first meta-analysis in our dataset (MA1.1) exemplifies the general results. The overall response of biomass production to biodiversity loss tended to be negative and became more negative for larger proportions of species lost without changes in the variational range of effect sizes (Fig. 3). This gradual response type was also found in the analysis of fertilization effects on biomass production (MA2.1) and in soil responses to changes in precipitation (MA8) and land-use change (MA9) as well as prey responses to predator loss (MA16.1). Ten additional examples of this type of response involving other drivers of environmental change are provided in the supporting material (Extended Data Fig. 2 and Supplementary Table 1). In all of these cases, the magnitude of the environmental change altered the magnitude of the response—as expected—but the variance around this relationship did not indicate the emergence of a ‘novel’ ecosystem response beyond a pressure threshold. Eight cases showed significant QR tests, of which three showed an increase in response variance only at highest pressure strength and two cases showed a reduction in response variance with increasing pressure. Thus, only three out of 36 cases showed a shifting distribution of effect sizes with increasing pressure that was consistent with the emergence of new types of responses above a threshold. These comprise land-use change effects on mammal abundance (MA6.5), warming effects on corals (MA10) and fertilization effects on microbial respiration (MA17.2; all Extended Data Fig. 2). By contrast, in 12 of the 36 meta-analyses, neither KL nor QR were significant (exemplified by MA23.1 in Fig. 3; for others see Extended Data Fig. 2), indicating that no increases in response magnitudes or threshold trangressions were observed.

These results are relevant for across-system analyses of single pressure gradients but in many cases management might not have a-priori knowledge of which pressure gradient leads to transgressions. To analyse this situation, we further aggregated our analysis across drivers, organism groups and ecosystems, by standardizing and normalizing the pressure gradient to a median of 0 and a range of –1 to 1 (Fig. 4). The range of responses was impressive, the effect sizes in cases indicated >200-fold increase or decrease in the measured ecosystem variable (Fig. 4a). Both KL and QR tests were highly significant for the aggregated data, indicating a strong impact of pressure intensity on the strength and variance of the ecological response (Supplementary Table 1). However, this increase in the variance of effect sizes was found for studies with normalized pressures >0.5, which comprised the top 3.5% of the manipulated range of potential impacts (Fig. 4b). This observation resembles a

‘sledge-hammer effect’; that is, system transformation by huge impact, which is a trivial consequence of the large pressure magnitude and the complete transformation of the system.

As the sign of the effect size depends upon the specific association of driver and effect in each meta-analysis, we also analysed the absolute magnitude of response (|LRR|) independent of sign for the aggregated dataset (Fig. 4c). We found that the median |LRR| increased with increasing environmental pressure, as did the variance, particularly so at the highest pressure magnitudes (significant KL and QR tests; Supplementary Table 1). The median |LRR| corresponded to 1.5–2-fold increases or decreases in process rates or properties, whereas the range of responses (the 5–95% quantiles of |LRR|) exceeded fivefold changes even at the smallest pressure strengths. Thus, even at very small pressures, very large responses can occur.

Discussion

Analysis of the 4,601 experiments that we assembled here, potentially the most comprehensive data available, did not enable us to estimate where thresholds might have been crossed. Instead, the data suggest that the ecosystem impacts of human-induced changes in environmental drivers are better characterized by gradual shifts in response magnitudes with increasing pressure coupled with broad variations around this trend. While our analyses do not rule out the existence of tipping points, they bring into question the utility of threshold-based concepts in management and policy if we cannot detect thresholds in nature[10,11]. Expectation of threshold responses ultimately leads to an underestimation of the large consequences of small environmental pressures[12]. Moreover, it marginalizes the importance of other, more complex, nonlinear dynamics under global change, which may underlie the considerable variance around gradually increasing response magnitudes.

Our use of field and seminatural experiments has the advantage that these often involve pressures that are larger than observed environmental conditions, as they commonly incorporate future scenarios of severe environmental change[13]. This counters the argument that thresholds exist but have not yet been reached. Still, some caveats to our approach need to be acknowledged. First, the absence of evidence is obviously not the evidence of absence: as shown by our explicit analysis of test power, the existence of thresholds can be masked by high interstudy variance (especially for HD). However, this also questions the usefulness of thresholds if their occurrence is dependent on the complex interaction of multiple pressures and their detection is only possible under very high signal-to-noise ratios. Without a-priori knowledge across specific systems of when thresholds might appear, any definition of thresholds—even if precautionary principles are used—must remain arbitrary. Second, we focused on functional, not compositional, aspects of ecosystems and do not make conclusions about threshold pressures for changes in composition. However, compositional and functional stability often show interdependencies[14] because compensatory dynamics between species may dampen the response in ecosystem functions[15] or allow for rapid recovery from a phase shift[16–18]. Given that the functions addressed here often are aggregate properties of the communities investigated, we thus consider it unlikely that thresholds are more prevalent for compositional responses. Third, the temporal extent of the experimental studies in our database is limited; it rarely exceeds the scale of tens of generations of

organisms. However, there is no strong support for why threshold transgressions should increase through time. Threshold-related concepts thus would be untestable in ecology, as their absence could always be ascribed to insufficiently long observation periods.

The lack of clearly defined and generally applicable thresholds distinguishing between tolerable and non-tolerable responses has obvious implications for environmental policies. The use of thresholds has been critically discussed in ecosystem management, conservation and restoration[19–21] to establish precautionary principles for environmental policy. Using such threshold arguments in a world where changes are too case-specific and variable to allow prediction of tipping points undermines this precautionary argument. It leads to the anticipation of major system transformation as thresholds are passed, whereas most observed responses to environmental change represent progressively shifting baselines on timescales of human perceptions[22,23]. Consequently, environmental concerns might appear overstated if thresholds are taken for the general case but critical transitions associated with transgressing thresholds are not observed[24,25]. The frequently major and highly variable responses we observed even at low pressure magnitudes indicate that safe-operating spaces are unlikely to be definable from data. The data resonate well with the fact that, conceptually, thresholds occur under special and limiting conditions. Our results thus question the pervasive presence of threshold concepts in management and policy.

Methods

Data

We searched the ISI Web of Science (WoS) using a search string targeted towards detecting meta-analyses in a global change context (Topic: [‘metaanalysis’ or ‘meta-analysis’ or ‘metaanalyses’ or ‘meta-analyses’] AND Topic: [‘global change’ or ‘fertili*’ or ‘land-use’ or ‘acidification’ or ‘warming’ or ‘temperature’ or‘eutrophication’ or ‘disturbance’ or ‘invasion’ or ‘extinction’ or ‘drought’ or ‘ultraviolet’] AND Topic: [‘chang*’ or ‘manipulation*’ or ‘experim*’ or ‘treatm*’]). We refined the results by focusing on the WoS research area ‘Environmental sciences and ecology’. This search (done on 11 September 2016) yielded 979 studies from which most did not fit all of our inclusion criteria (upon request, we can provide a list of all studies with the study-specific criteria to include or exclude), which were as follows:

- The paper provided a formal meta-analysis with effect sizes, which quantified the responses to a factor that represented a global change impact. The factor was either an experimental treatment or an in-situ change. This excluded numerous studies that were verbal/vote-counting reviews or provided effect sizes as a response to non-global-change factors (for example, mitigation efforts).
- The response was measured at the level of ecological communities or ecosystems. This excluded studies where responses were measured at the level of single species, as these were deemed inappropriate to detect regime shifts, or at the level of human societies (for example, health aspects and economy). We also excluded fossil data as not being affected by anthropogenic global change and

non-biological response variables (for example, the effect of CO_2-enrichment on water pH).

- Given that effect sizes on species richness have recently been criticized strongly for being statistically biased[26], we decided not to use biodiversity response variables but only functional processes or properties at the community or ecosystem level (details below). As we explicitly address the statistical distribution of effect sizes (see below), this statistical bias was considered to be potentially misleading in the context of our analysis. However, we used cases where biodiversity loss was the manipulated component of global change and a functional response was measured.

From the remaining 162 meta-analyses that fulfilled these criteria, we extracted the information needed to perform our analyses. This included a measure of the magnitude of the stressor (impact, driver) and the effect size as well as its sampling variance or weight (response). When the information was not given in an online appendix or associated data table, we contacted the authors to ask for data access. Still, we had to exclude further meta-analyses, as they: did not quantify the stressor magnitude (this was especially common in meta-analyses addressing the response to invasive species); did not contain enough cases to perform analysis (we set the critical number of effect sizes to 35 as a minimum to detect variance shifts); overlapped with other meta-analyses on the same subject (this was especially found for analyses on eutrophication and biodiversity loss, where we always opted for the most consistent and information-dense alternative); did not provide available data.

The final database contained 24 meta-analyses (information derived from 29 papers[27–55]), which were divided into 36 cases (Supplementary Table 1). Subsetting multiple cases from a meta-analysis was done if different drivers were tested or different response categories were used in a single meta-analysis. We followed the authors in defining response categories and stressor variables. We excluded laboratory experiments and focused our study solely on field experiments and observational studies. The resulting dataset reflects ecological responses in the form of ecosystem processes (primary or secondary production, feeding rates and element fluxes) to the most pervasive anthropogenic alterations of our planet (Supplementary Table 1).

Statistical approach

For each meta-analysis dataset containing a measure of the stressor magnitude (X), the response variable (log response ratio, LRR) and its sampling variance (var.LRR), we assessed whether the dataset reflects any statistically significant influence of the stressor variable on the response. As the data basis of each meta-analysis (and thus the sources of variation of LRR within each dataset) is unknown to us, we devised three robust non-parametric test statistics and assessed their statistical significance by permutation tests.

An averaged KL divergence quantified the overall deviation between the response distribution for a given stressor value and the marginal response distribution (that is, the response distribution when collapsing all response data onto a single axis ignoring the stressor variable). Second, a quantile ratio (QR) of interquantile range (95–5%) was then

used to quantify substantial variability of the response distribution width across the range of observed stressors. Finally, we used the HD test to assess the multimodality of effect sizes[9]. Based on simulation-based *P* values, HD provides a narrow test for the case of bi-(multi)-stability of responses, analogous to the bimodality test proposed by Scheffer and Carpenter[56]. A significant HD indicates that the responses along the pressure gradient fall into two (or more) clearly separated categories, which indicates the presence of two (or more) alternative ecosystem states. Essentially, strict bimodality across a wide range of studies is a rather narrow expectation but we include this test as the bifurcation case is the one most often discussed in considerations of thresholds, tipping points and regime shifts[56,57].

For both KL and QR, the assessment of statistical significance was done by a permutation test: the null hypothesis (NH) that the response distribution is unrelated to the stressor is simulated by breaking up paired variables (X, LRR, var.LRR) and recombining them in the form (X', LRR, var.LRR), where X' is a permutation of recorded stressor values. If the NH were valid, this permutation should induce no substantial difference. Computing the two test statistics (KL and QR) for the permuted dataset (X', LRR, var.LRR) and repeating these steps 10,000 times generates the distribution of the test statistics under validity of the NH and allows extraction of a *P* value as the fraction of permutations that yielded a similar or larger value for the test statistic (KL or QR) as the original dataset (X, LRR, var.LRR).

In comparison to alternative approaches, our methods are robust and non-parametric—they do not rely on functional assumptions and use only the supposed smoothness of a possible connection between stressor and response. Reconstructing the NH by simulating surrogate data guarantees perfect control of errors of the first kind (false positive statements) and even would handle a constant bias of estimators. Given the breadth of underlying meta-analyses, we also consider our analysis highly conservative with regard to publication bias and study selection. Finally, using a weighted approach downgrades the influence of studies with very high internal variance and thus decreases the chance of missing threshold-like responses because of too-noisy data (false negative statements).

It should be noted that neither the single experiments summarized in each meta-analysis nor the meta-analyses themselves were designed to detect thresholds. The inclusion of studies not necessarily looking for thresholds actually reduces the risk of publication bias towards positive results. However, even if the underlying experiments were not planned to detect thresholds, our statistical approach should reveal these if they fall into the covered range of stressors, which can be expected as this range encompasses stressor magnitudes not yet experienced under realistic conditions.

Statistical analyses

For each effect size in each meta-analysis, a statistical weight is assigned to each data point as the log-transformed inverse sampling variance of the effect size

$$\log\left(1 + \frac{1}{\text{var.LRR}}\right) \quad (1)$$

As described above, surrogate datasets (reflecting the NH) are created by permuting the list of stressor values in X (yielding $X' = X$ shuffled). From the list of stressor values, a smooth probability distribution $P_X(gx)$ is computed via weighted (with statistical weights calculated following equation (1)) kernel density estimation (with a Gaussian kernel and an optimized bandwidth, compare Simulations) for grid points gx that span the range of observed stressor values (Extended Data Fig. 4). A smooth density surface over the grid (gx, gy) in the (X, LRR) plane is computed from the dataset (and the surrogates) via a two-dimensional weighted (with statistical weights calculated following equation (1)) kernel density estimation (bivariate Gaussian D-class kernel with optimized bandwidth) (Extended Data Fig. 5). For each grid point gx, the density profile along gy is converted to a conditional probability distribution $P_{\mathrm{LRR}|X}(gy|gx)$ by normalization (Extended Data Fig. 6, with results for the original data and the surrogate data). Based on the conditional cumulative distribution function,

$$F_{\mathrm{LRR}|X}(gy|gx) = \sum_{gy' \le gy} P_{\mathrm{LRR}|X} \sum P_{\mathrm{LRR}|X}(gy'|gx) \tag{2}$$

(Extended Data Fig. 7), the 5%, 50% (median) and 95% quantiles can be extracted for each grid point gx (Extended Data Fig. 8). The test statistics that we devised are:

1. the average KL divergence

$$\mathrm{KL} = \sum_{gx} P_X(gx) \sum_{gy} P_{\mathrm{LRR}|X}(gy|gx) \log \frac{P_{\mathrm{LRR}|X}(gy|gx)}{P_{\mathrm{LRR}}(gy)} \tag{3}$$

   that shares the useful property of being non-negative and that vanishes if, and only if, $P_{\mathrm{LRR}|X} \equiv P_{\mathrm{LRR}}$ (almost everywhere). Pronounced differences between the two empirical distributions are thus condensed in values substantially larger than zero.

2. the quantile ratio (QR) of interquantile (5–95%) ranges (IQR)

$$\mathrm{QR} = \frac{\mathrm{IQR}_5^{95}(99\,\%)}{\mathrm{IQR}_5^{95}(1\,\%)} \tag{4}$$

   where $\mathrm{IQR}_5^{95}(qx)$ denotes the qx-quantile of the 5–95% interquantile range of the conditional probability distribution $P_{\mathrm{LRR}|X}(gy|gx)$ and the subsequent percentage in brackets indicates the related weighted quantile across the stressor grid points. We choose this latter definition for robustness, rather than the maximum/minimum ratio which may be prone to distortions by extremes. This measure was devised to indicate substantial changes of the LRR variance along the stressor axis.

3. the HD statistic tests for multimodality, which, if significant, indicates that a frequency distribution has more than one mode.

Values of all test statistics obtained for the original dataset were assessed for statistical significance. This was done by excessively repeating the permutation strategy to create

surrogate data in accordance with the NH of a non-existent connection between stressor X and response LRR. P values for both test statistics (KL and QR) were obtained as fractions of 10,000 surrogate sets (in the case of HD, 2,000 permutations), leading to test statistics exceeding related values of the original dataset.

In addition to the employed kernel density estimates generating cumulative distribution functions and derived quantiles, we used a nonlinear quantile regression supplied by the R package qgam[58]. This package is based on general additive models and returns quantiles instead of standard mean response. With qgam we estimated the following quantiles: 0.001, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 0.999. Because these quantile curves were computed sequentially, independently resultant lines could intersect. To resolve this problem, we used the R package cobs to perform a penalized B-spline regression of obtained quantiles (separately for every grid point gx), bound to the constraint of a monotonic increase, thus yielding a smooth cumulative distribution function. As for the kernel density estimation, exemplary 5%, 50% (median) and 95% quantiles are shown in Extended Data Fig. 8.

Under default settings, the qgam routine was very time consuming and, in comparison with the bandwidth optimized kernel density method, had inferior test power (Fig. 2 and Supplementary Table 1). This may be due to the fact that, because of excessive run time an optimization of qgam parameters was not feasible. We therefore constrain reporting of our results to those obtained with the optimized kernel density method.

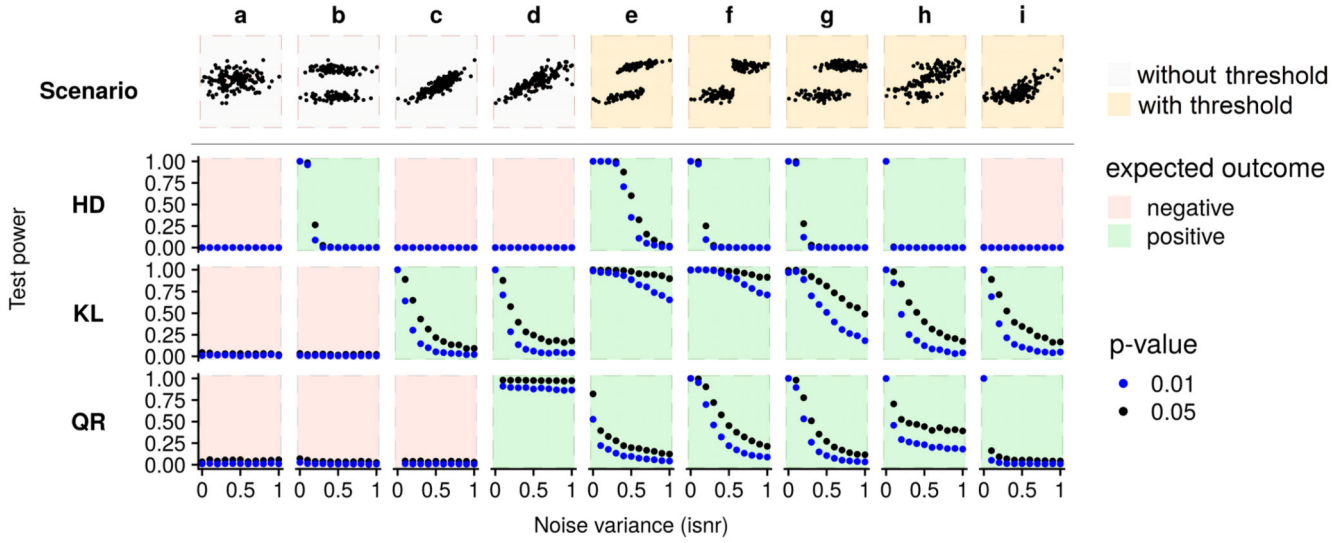
Simulations

We examined the performance of our tests by simulating artificial datasets that combined nine deterministic backbone structures with additive noise (normally distributed random fluctuations) of controlled intensity. The deterministic backbone structures were chosen to reflect a broad range of scenarios. The noise intensity is quantified via the inverse signal-to-noise ratio (ISNR), that is the size ratio of fluctuations and backbone structure. The nine cases are depicted in Extended Data Fig. 3, each for small (ISNR = 0.05) and large (ISNR = 0.95) noise intensity. In Fig. 2 and Supplementary Table 1, we list the expected outcome of the three designed tests for the noise-free case. To assess the performance of the tests under various noise conditions, we simulated, for each ISNR value (in the range [0–1]), 1,000 artificial datasets and collected related test decisions (for two decision criteria $P = 0.05$ and 0.01 and all three tests). In the case of an expected positive test, the fraction of positive test decisions thus estimates the test power (1-error of the second kind). We note that simulations of the test power were also underlying the optimization of the kernel bandwidth, where bandwidth selection was based on the ‘solve-the-equation’ method of Sheather and Jones[59].

In all simulated cases with small to moderate noise (ISNR < 0.5), threshold structures in simulated response–stressor relations could be detected with high reliability (at least for the KL and QR tests). Of course, for strong noise (ISNR ≥ 1), thresholds may be masked by random fluctuations reflecting natural variability. In such situations, the underlying threshold structure, although present, will no longer be ecologically relevant because it is overridden by natural variability.

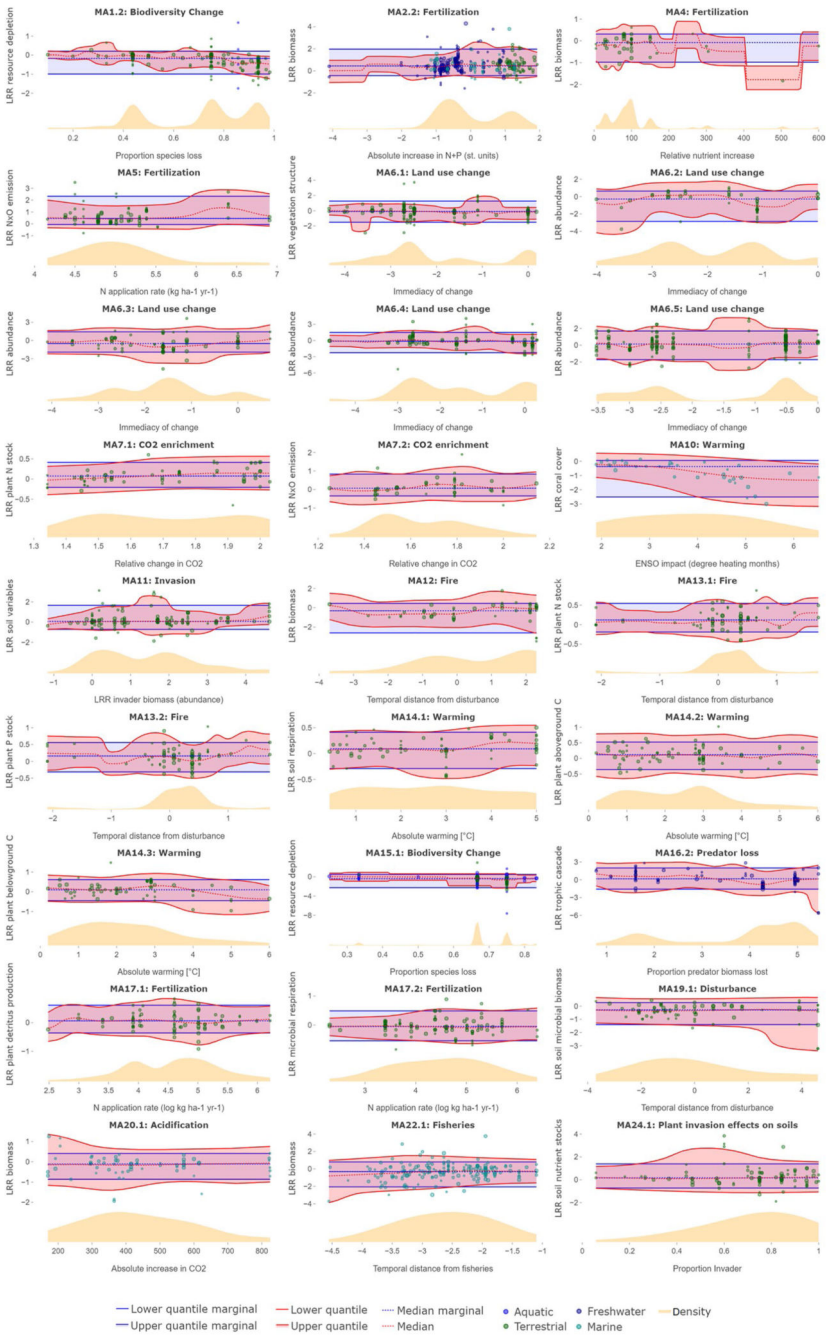
Europe PMC Funders Author Manuscripts

Extended Data



Extended Data Fig. 1. Test power as in Fig. 2, but for the "qgam" approach.

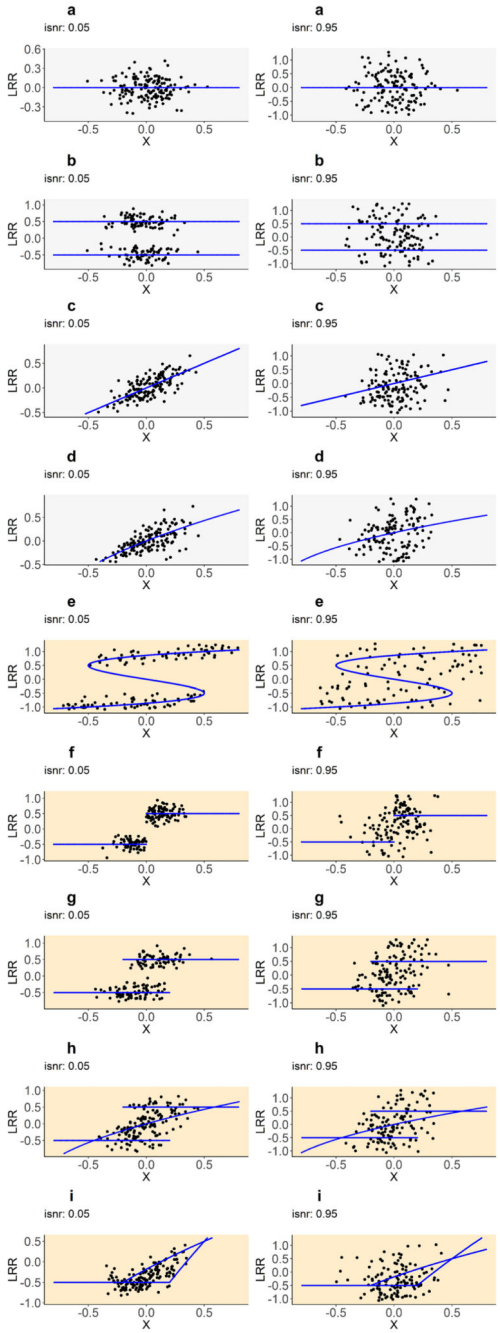
Fractions of positive test results (equals test power when test should be positive) for simulated test cases. We analysed the test power for 9 scenarios of responses to pressure in meta-analyses, the derivation of each scenario is described in the supplementary online material, Extended Data Fig. 7. Scenarios a–d do not comprise a threshold, where scenario a is the null model without an effect of the pressure on the response. Scenarios e–i do comprise a threshold, for the latter two combined with intermediate responses. For the three statistical test used in our analyses, the expected outcome is colour-coded, with green representing that the test should be significant. We then tested the proportion of 1000 simulated datasets for which the tests were significant with a probability p = 0.05 (black) and p = 0.01 (blue). We did for increasing noise variance (= inverse signal-to-noise ratio). The three tests together allow perfect detection of thresholds at the absence of noise (scenarios e–h), only if threshold-type and gradual responses are mixed (scenario i), the analysis of multimodality (HD) fails, giving the same output as a gradual increase in mean and variance of the response (scenario d). With increasing noise variance, however, the detection probability for thresholds via HD and QR rapidly decreases.We used default settings for the "qgam" approach due to high runtimes and computational effort, thus settings are not optimized as for the test power calculations based on kernel density estimation. Note: HD is equal to kernel method, because it is not based on different quantile estimations.



Extended Data Fig. 2. Changes in the response magnitude along increasing pressure strength.
Further meta-analyses testing for changes in the response magnitude along increasing pressure strength. Red and blue shaded regions indicate the (5%-95%) interquantile ranges for the bivariate data (including the pressure gradient) and the univariate LRR data (ignoring the pressure gradient = homogeneous marginal probability), respectively. Solid red and dashed blue thick lines trace the related median (50% quantile). Overlain are the data points and at the bottom the yellow shaded area indicates the distribution px(gx) resultant

Europe PMC Funders Author Manuscripts

from a weighted kernel density estimation. Colour codes for habitat (dark blue: freshwater, aquamarine: marine, green: terrestrial), circle size reflect statistical weight.



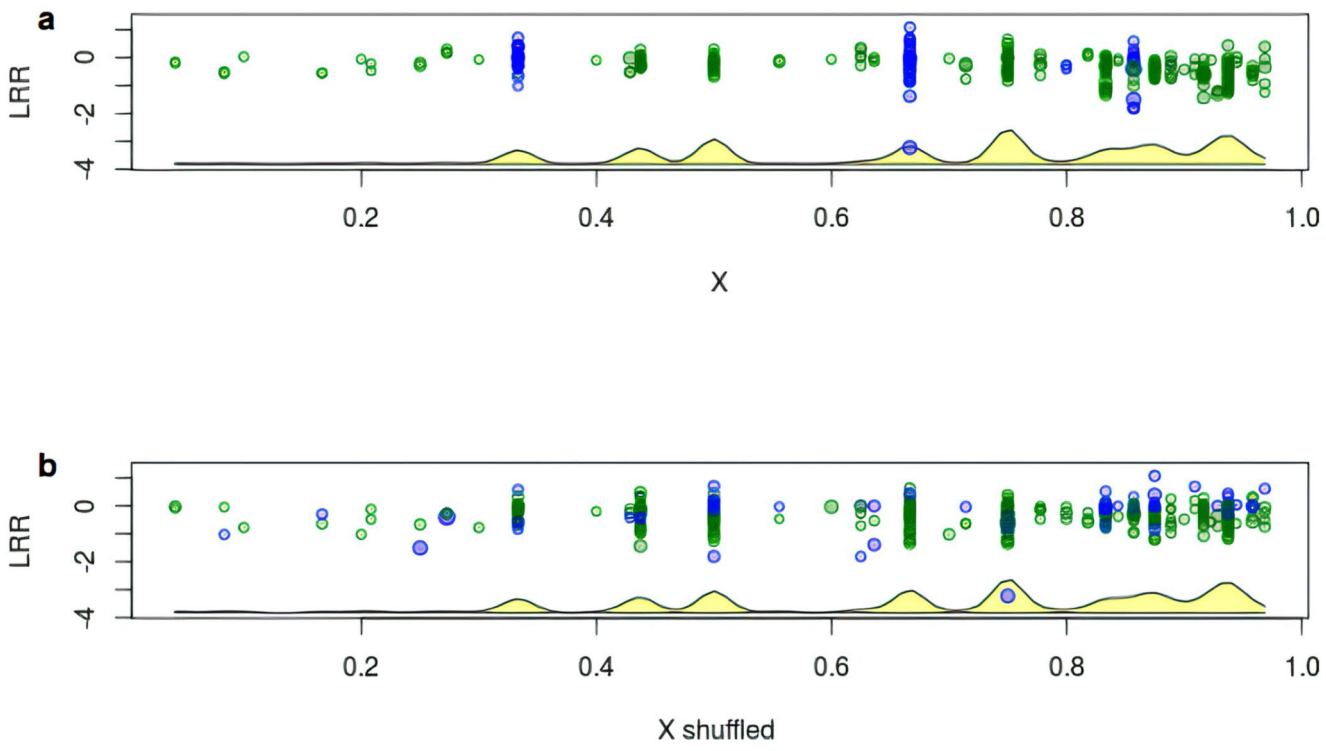
Extended Data Fig. 3. Test cases at different noise levels.

In order to assess the power of our statistical tests, we simulated artificial meta-analyses combining prototypical response~stressor relationships with (normally distributed) random fluctuations reflecting natural variability, and compared related statistical test results with expectations. Stressor range (along horizontal range) and deterministic effect sizes (along

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

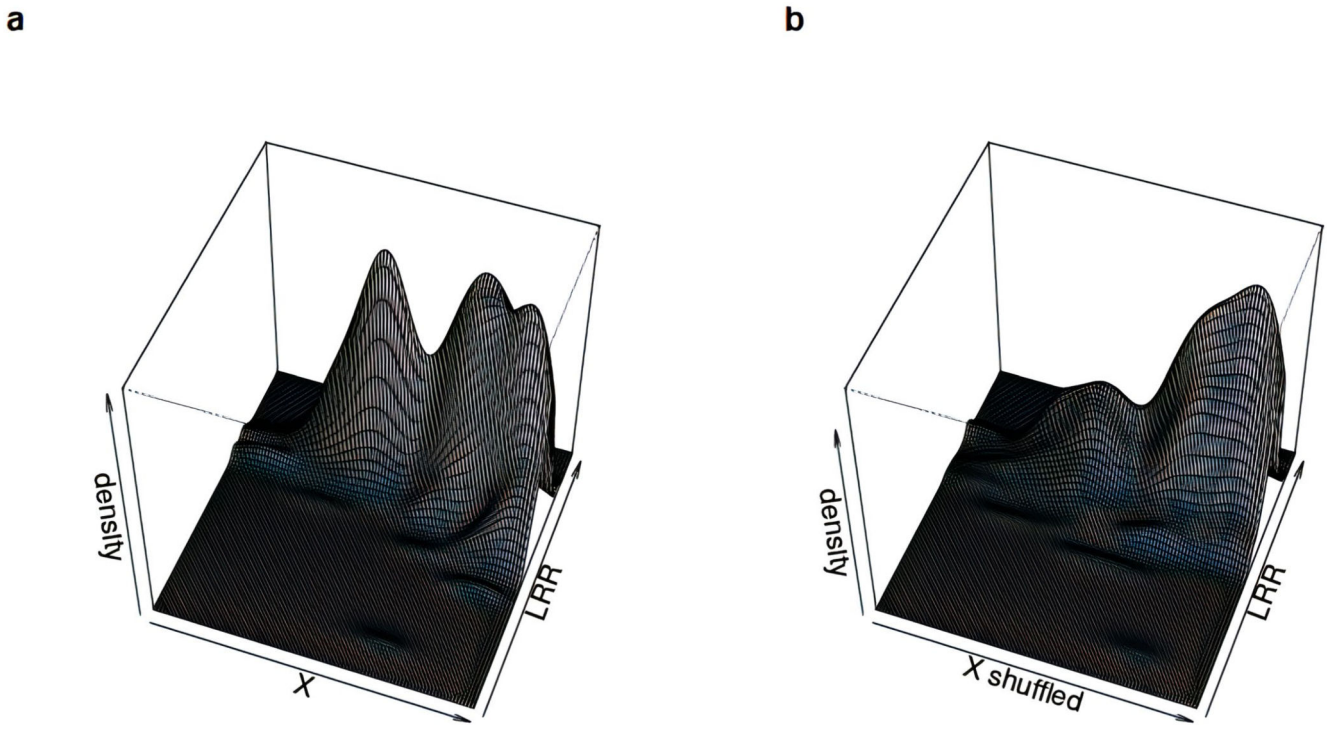
vertical axis) are normalized to [-0.5,0.5] x [0.5,0.5]. Stressor values are normally distributed with mean zero. The relative intensity of random fluctuations is quantified by inverse signal-to-noise ratio (isnr). A grey background indicates absence of thresholds, yellow background threshold presence. **a**, (neutral -simple-): Here pressure strength has no impact on the response, which falls into a single response. Thus, we assume that across all "studies" in this "meta-analysis", there is one main response type and no threshold. **b**, (neutral -bimodal-): Here pressure strength has no impact on the response, which falls into either of two alternative attractors: a weak and a strong response. Thus, we assume that across all "studies" in this "meta-analysis", there are two main response types and no threshold. **c**, (plain trend, proportionate response): A gradual response with no change in variability revealing a trend but no threshold. **d**, (gradual, no threshold): A nonlinear but smooth increase with smoothly increasing variability. Here we assume that the responses increase with some normally distributed error with the pressure without transgressing any threshold. **e**, (saddle-node bifurcation): A widely discussed model situation in the context of 'tipping points' and 'catastrophic regime shifts'. **f**, (strict threshold): Here we assume that across all studies in a meta-analysis, the response switches from weak to strong (as defined in case **a**) at exactly the same threshold for each study. This assumption is very unrealistic (see below) but makes the case when there are two main response types and a global threshold holding for any single study in the meta-analysis. **g**, (variable threshold): Here we assume that all studies in a meta-analysis potentially transgress a threshold, but the position of the threshold differs. Thus, the probability that the response switches from weak to strong increases with increasing pressure. Response similar to Case **a**. **h**, (variable threshold with intermediates): Here we assumed that not all studies in a meta-analysis potentially transgresses a threshold, but some of the studies show gradual responses. As in Case **f**, the position of the threshold differs between studies and the probability that the response switches from weak to strong increases with increasing pressure. As for cases **a**,**b**,**e** and **f**, we assume there are two main response types. This scenario can be distinguished from case d by the abrupt change in variance along the pressure gradient. **i**, (variable threshold and variable effect sizes below and above threshold): Here we assumed that the position of the threshold differs between studies (as in Case **f**) and any experiment in the study had a 50% chance that the threshold was crossed, independent of the pressure magnitude. By contrast to cases **a,b** and **e-h**, we relax the assumption that there are two main response types, but transgressing the thresholds leads to an increase in effect size, which depended on the position on the pressure gradient. Thus, if a study with a large pressure magnitude transgressed the threshold, the increase in response magnitude was larger than if a study with an overall small pressure did so.



Extended Data Fig. 4. Permutation example.

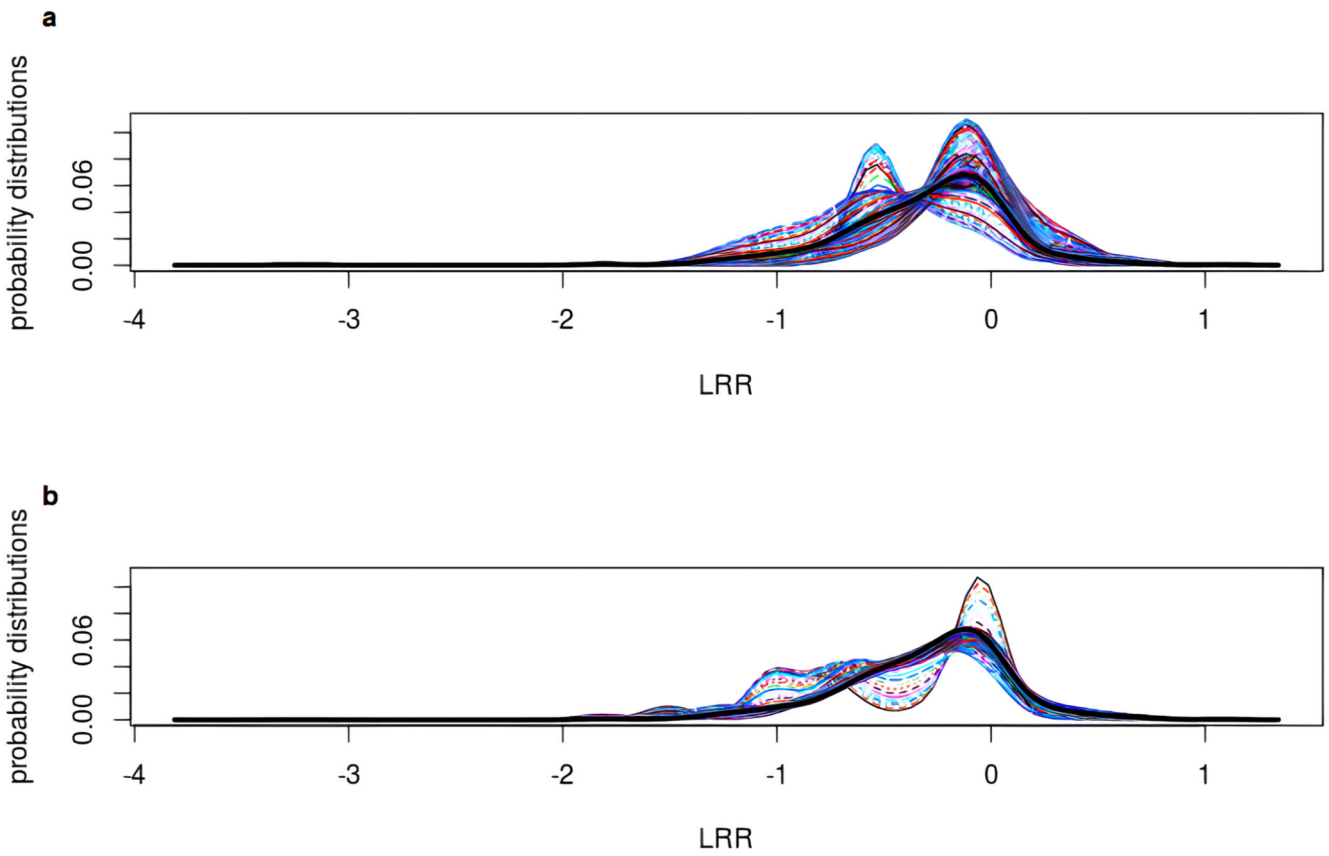
An example dataset (**a**) together with a surrogate dataset based on permuted X values (**b**); as in Fig. 2 of the main text, colour codes habitat (blue: marine, green: terrestrial), circle size reflects statistical weight, and the yellow shaded area indicates the distribution $p_X(gx)$ resultant from a weighted kernel density estimation.

Europe PMC Funders Author Manuscripts



Extended Data Fig. 5. Two-dimensional probability densities.
Densities are calculated over a grid (gx,gy) for the original dataset (**a**) and the surrogate dataset (**b**).

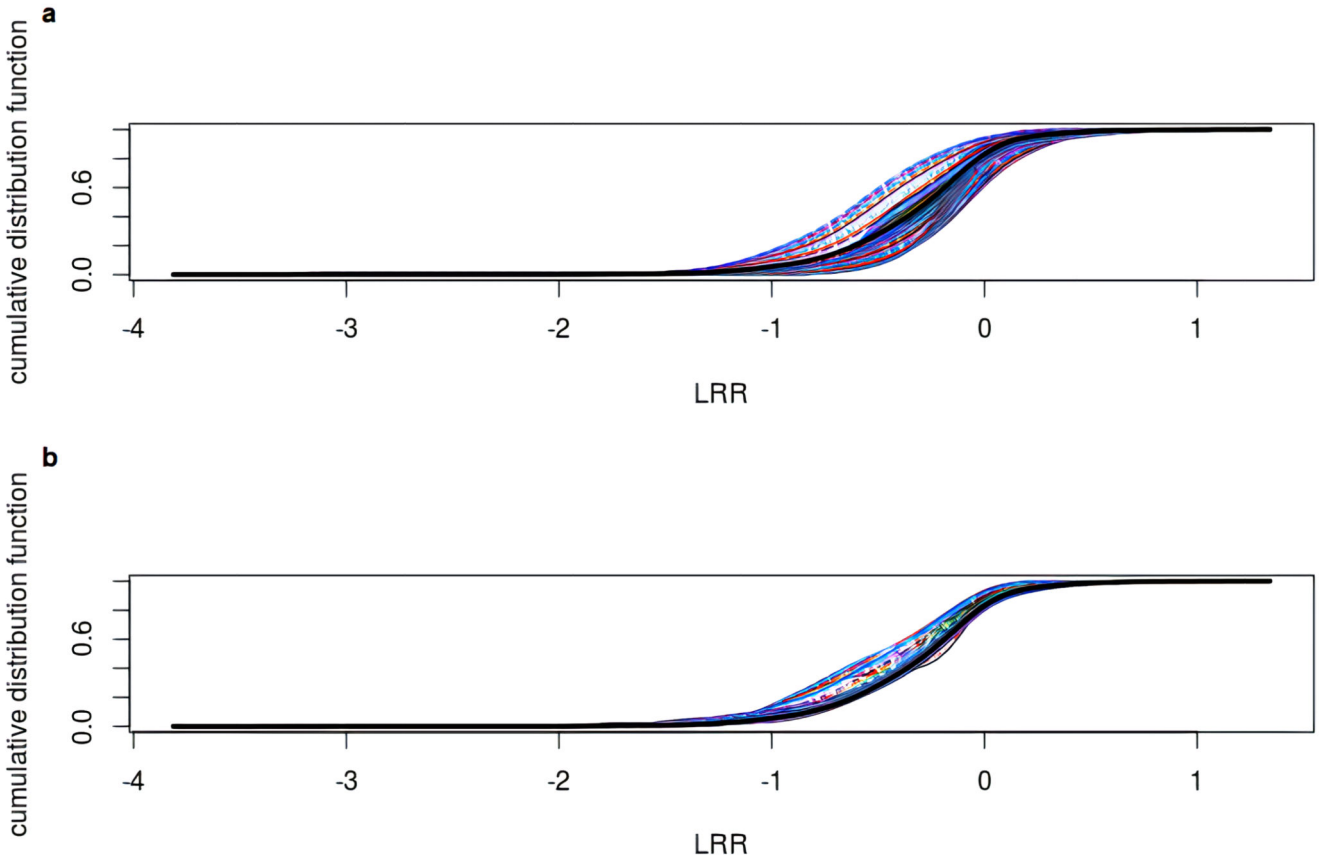
Europe PMC Funders Author Manuscripts



Extended Data Fig. 6. Conditional probability distribution example.

The conditional probability distribution $p_{LRR|X}(gy \mid gx)$ for each grid point gx together with the marginal distribution $p_{LRR}(gy)$ (thick black line). **a**, original dataset, **b**, surrogate dataset.

Europe PMC Funders Author Manuscripts

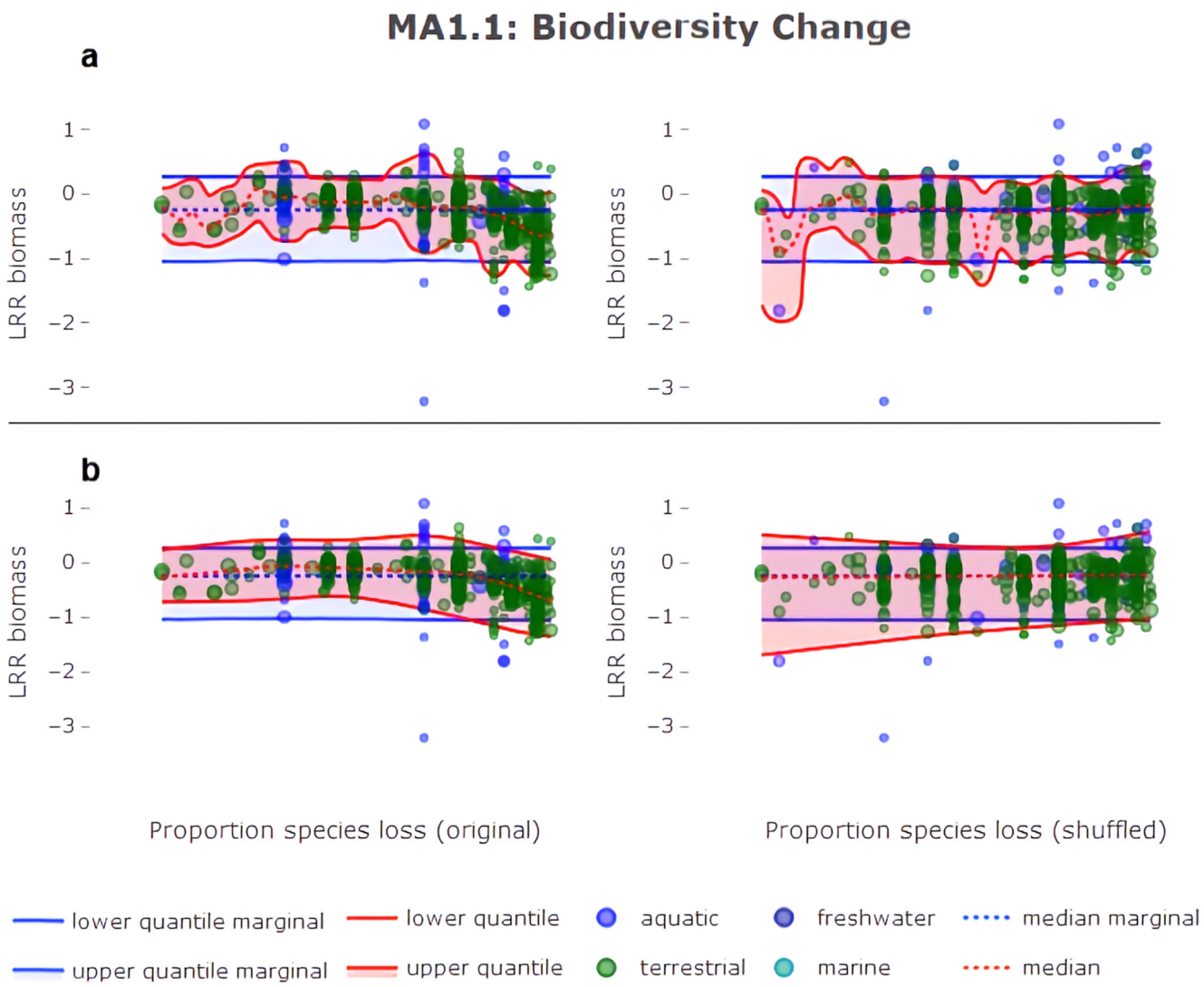


Extended Data Fig. 7. Cumulative distribution example.

The cumulative distribution functions $F_{\mathrm{LRR}|X}(gy \mid gx)$ and $F_{LRR}(gy)$ (thick black line) for the probability profiles shown in Supplementary Fig. 3. **a**, original, **b**, surrogate.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts



Extended Data Fig. 8. Comparison of kernel density estimation and "qgam".

Images of the reconstructed statistical structures for an original dataset (MA1.1) and one of its surrogate datasets. **a**, Quantiles estimated by optimized kernel density estimation; **b**, Quantiles estimated by "qgam".

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The data reported in this paper are presented and derived from 36 different meta-analyses; they are archived and available from each of these as indicated in the Supplementary Text. The concept of this paper emerged during scientific discussions with T. Blenckner at Stockholm University, at the UK NERC/BESS Tansley Working Group on ecological stability and the TippingPond EU Biodiversa project. The actual work was funded by the Lower Saxony Ministry of Science and Culture through the MARBAS project to H.H. and the HIFMB, a collaboration between the Alfred-Wegener-Institute, Helmholtz-Center for Polar and Marine Research and the Carl von Ossietzky University Oldenburg, initially funded by the Ministry for Science and Culture of Lower Saxony and

the Volkswagen Foundation through the 'Niedersächsisches Vorab' grant programme (grant no. ZN3285). The work was finalized with support by Deutsche Forschungsgemeinschaft grant no. HI848/26–1. L. Toaspern helped with gathering data from invasion meta-analyses. P. Ruckdeschel helped with the statistical approach. M. Vilà provided additional information on their published meta-analyses. We acknowledge the comments by U. Feudel, G. Gerlach and the members of the Plankton Ecology Lab at the Carl von Ossietzky University Oldenburg on the manuscript which helped with our argumentation.

Data availability

All data are available at https://zenodo.org/record/3828869#.XsI4ZmgzaUk.

Code availability

All code are available at https://zenodo.org/record/3828869#.XsI4ZmgzaUk.

References

1. Scheffer M, Carpenter S, Foley JA, Folke C, Walker B. Catastrophic shifts in ecosystems. Nature. 2001; 413: 591–596. [PubMed: 11595939]
2. Scheffer, M. Critical Transitions in Nature and Society. Princeton Univ. Press; 2009.
3. Rockström J, et al. A safe operating space for humanity. Nature. 2009; 461: 472–475. [PubMed: 19779433]
4. Folke C, et al. Regime shifts, resilience and biodiversity in ecosystem management. Annu Rev Ecol Evol Syst. 2004; 35: 557–581.
5. Donohue I, et al. Navigating the complexity of ecological stability. Ecol Lett. 2016; 19: 1172–1185. [PubMed: 27432641]
6. Aichi Biodiversity Targets. UN; 2010. https://www.cbd.int/sp/targets/
7. Carpenter SR, Brock WA. Rising variance: a leading indicator of ecological transition. Ecol Lett. 2006; 9: 308–315.
8. Scheffer M, et al. Early-warning signals for critical transitions. Nature. 2009; 461: 53–59. [PubMed: 19727193]
9. Hartigan JA, Hartigan PM. The dip test of unimodality. Ann Stat. 1985; 13: 70–84.
10. Montoya JM, Donohue I, Pimm SL. Planetary boundaries for biodiversity: implausible science, pernicious policies. Trends Ecol Evol. 2018; 33: 71–73. [PubMed: 29126565]
11. Pimm SL, Donohue I, Montoya JM, Loreau M. Measuring resilience is essential to understand it. Nat Sustain. 2019; 2: 895–897. [PubMed: 31858022]
12. Clark CM, Tilman D. Loss of plant species after chronic low-level nitrogen deposition to prairie grasslands. Nature. 2008; 451: 712–715. [PubMed: 18256670]
13. Korell L, Auge H, Chase JM, Harpole WS, Knight TM. We need more realistic climate change experiments for understanding ecosystems of the future. Glob Change Biol. 2020; 26: 325–327.
14. Hillebrand H, et al. Decomposing multiple dimensions of stability in global change experiments. Ecol Lett. 2018; 21: 21–30. [PubMed: 29106075]
15. Connell SD, Ghedini G. Resisting regime-shifts: the stabilising effect of compensatory processes. Trends Ecol Evol. 2015; 30: 513–515. [PubMed: 26190138]
16. Bruno JF, Sweatman H, Precht WF, Selig ER, Schutte VGW. Assessing evidence of phase shifts from coral to macroalgal dominance on coral reefs. Ecology. 2009; 90: 1478–1484. [PubMed: 19569362]
17. Diaz-Pulido G, et al. Doom and boom on a resilient reef: climate change, algal overgrowth and coral recovery. PLoS ONE. 2009; 4 e5239 [PubMed: 19384423]
18. Carpenter SR, et al. Early warnings of regime shifts: a whole-ecosystem experiment. Science. 2011; 332: 1079–1082. [PubMed: 21527677]
19. Suding KN, Hobbs RJ. Threshold models in restoration and conservation: a developing framework. Trends Ecol Evol. 2009; 24: 271–279. [PubMed: 19269057]

20. Vaquer-Sunyer R, Duarte CM. Thresholds of hypoxia for marine biodiversity. Proc Natl Acad Sci USA. 2008; 105: 15452–15457. [PubMed: 18824689]
21. Groffman PM, et al. Ecological thresholds: the key to successful environmental management or an important concept with no practical application? Ecosystems. 2006; 9: 1–13.
22. Hughes TP, Carpenter S, Rockstrom J, Scheffer M, Walker B. Multiscale regime shifts and planetary boundaries. Trends Ecol Evol. 2013; 28: 389–395. [PubMed: 23769417]
23. Papworth SK, Rist J, Coad L, Milner-Gulland EJ. Evidence for shifting baseline syndrome in conservation. Conserv Lett. 2009; 2: 93–100.
24. Schlesinger WH. Planetary boundaries: thresholds risk prolonged degradation. Nat Clim Change. 2009; 1: 112–113.
25. Duarte CM, et al. Reconsidering ocean calamities. BioScience. 2015; 65: 130–139.
26. Chase JM, Knight TM. Scale-dependent effect sizes of ecological drivers on biodiversity: why standardised sampling is not enough. Ecol Lett. 2013; 16: 17–26. [PubMed: 23679009]
27. Cardinale BJ, et al. Effects of biodiversity on the functioning of trophic groups and ecosystems. Nature. 2006; 443: 989–992. [PubMed: 17066035]
28. Gruner DS, et al. A cross-system synthesis of consumer and nutrient resource control on producer biomass. Ecol Lett. 2008; 11: 740–755. [PubMed: 18445030]
29. Elser JJ, et al. Global analysis of nitrogen and phosphorus limitation of primary producers in freshwater, marine and terrestrial ecosystems. Ecol Lett. 2007; 10: 1135–1142. [PubMed: 17922835]
30. Lin D, Xia J, Wan S. Climate warming and biomass accumulation of terrestrial plants: a meta-analysis. New Phytol. 2010; 188: 187–198. [PubMed: 20609113]
31. Treseder KK. Nitrogen additions and microbial biomass: a meta-analysis of ecosystem studies. Ecol Lett. 2008; 11: 1111–1120. [PubMed: 18673384]
32. Akiyama H, Yan X, Yagi K. Evaluation of effectiveness of enhanced-efficiency fertilizers as mitigation options for N_2O and NO emissions from agricultural soils: meta-analysis. Glob Change Biol. 2010; 16: 1837–1846.
33. Gibson L, et al. Primary forests are irreplaceable for sustaining tropical biodiversity. Nature. 2011; 478: 378–381. [PubMed: 21918513]
34. Liang JY, Qi X, Souza L, Luo YQ. Processes regulating progressive nitrogen limitation under elevated carbon dioxide: a meta-analysis. Biogeosciences. 2016; 13: 2689–2699.
35. Liu LL, et al. A cross-biome synthesis of soil respiration and its determinants under simulated precipitation changes. Glob Change Biol. 2016; 22: 1394–1405.
36. van Lent J, Hergoualc'h K, Verchot LV. Reviews and syntheses: soil N_2O and NO emissions from land use and land-use change in the tropics and subtropics: a meta-analysis. Biogeosciences. 2015; 12: 7299–7313.
37. Ateweberhan M, McClanahan TR. Relationship between historical sea-surface temperature variability and climate change-induced coral mortality in the western Indian Ocean. Mar Pollut Bull. 2010; 60: 964–970. [PubMed: 20447661]
38. Gärtner M, et al. Invasive plants as drivers of regime shifts: identifying high-priority invaders that alter feedback relationships. Divers Distrib. 2014; 20: 733–744.
39. Dooley SR, Treseder KK. The effect of fire on microbial biomass: a meta-analysis of field studies. Biogeochemistry. 2012; 109: 49–61.
40. Dijkstra FA, Adams MA. Fire eases imbalances of nitrogen and phosphorus in woody plants. Ecosystems. 2015; 18: 769–779.
41. Lu M, et al. Responses of ecosystem carbon cycle to experimental warming: a meta-analysis. Ecology. 2013; 94: 726–738. [PubMed: 23687898]
42. Griffin JN, Byrnes JEK, Cardinale BJ. Effects of predator richness on prey suppression: a meta-analysis. Ecology. 2013; 94: 2180–2187. [PubMed: 24358704]
43. Srivastava DS, et al. Diversity has stronger top-down than bottom-up effects on decomposition. Ecology. 2009; 90: 1073–1083. [PubMed: 19449701]
44. Östman Ö, et al. Top-down control as important as nutrient enrichment for eutrophication effects in North Atlantic coastal ecosystems. J Appl Ecol. 2016; 53: 1138–1147.

45. Katano I, Doi H, Eriksson BK, Hillebrand H. A cross-system meta-analysis reveals coupled predation effects on prey biomass and diversity. Oikos. 2015; 124: 1427–1435.
46. Borer ET, et al. What determines the strength of a trophic cascade? Ecology. 2005; 86: 528–537.
47. Hodapp D, Hillebrand H. Effect of consumer loss on resource removal depends on species-specific traits. Ecosphere. 2017; 8 e01742
48. Liu LL, Greaver TL. A global perspective on belowground carbon dynamics under nitrogen enrichment. Ecol Lett. 2010; 13: 819–828. [PubMed: 20482580]
49. Martinson HM, Fagan WF. Trophic disruption: a meta-analysis of how habitat fragmentation affects resource consumption in terrestrial arthropod systems. Ecol Lett. 2014; 17: 1178–1189. [PubMed: 24866984]
50. Holden S, Treseder K. A meta-analysis of soil microbial biomass responses to forest disturbances. Front Microbiol. 2013; 4: 163. [PubMed: 23801985]
51. Nagelkerken I, Connell SD. Global alteration of ocean ecosystem functioning due to increasing human CO_2 emissions. Proc Natl Acad Sci USA. 2015; 112: 13272–13277. [PubMed: 26460052]
52. Kaiser MJ, et al. Global analysis of response and recovery of benthic biota to fishing. Mar Ecol Prog Ser. 2006; 311: 1–14.
53. Gill DA, et al. Capacity shortfalls hinder the performance of marine protected areas globally. Nature. 2017; 534: 665–669.
54. Gallardo B, Clavero M, Sánchez MI, Vilà M. Global ecological impacts of invasive species in aquatic ecosystems. Glob Change Biol. 2016; 22: 151–163.
55. Vila M, et al. Ecological impacts of invasive alien plants: a meta-analysis of their effects on species, communities and ecosystems. Ecol Lett. 2011; 14: 702–708. [PubMed: 21592274]
56. Scheffer M, Carpenter SR. Catastrophic regime shifts in ecosystems: linking theory to observation. Trends Ecol Evol. 2003; 18: 648–656.
57. Andersen T, Carstensen J, Hernandez-Garcia E, Duarte CM. Ecological thresholds and regime shifts: approaches to identification. Trends Ecol Evol. 2009; 24: 49–57. [PubMed: 18952317]
58. Fasiolo M, Goude Y, Nedellec R, Wood SN. Fast calibrated additive quantile regression. J Am Stat Assoc. 2020; doi: 10.1080/01621459.2020.1725521
59. Sheather SJ, Jones MC. A reliable data-based bandwidth selection method for kernel density estimation. J Royal Stat Soc B. 1991; 53: 683–690.

Europe PMC Funders Author Manuscripts

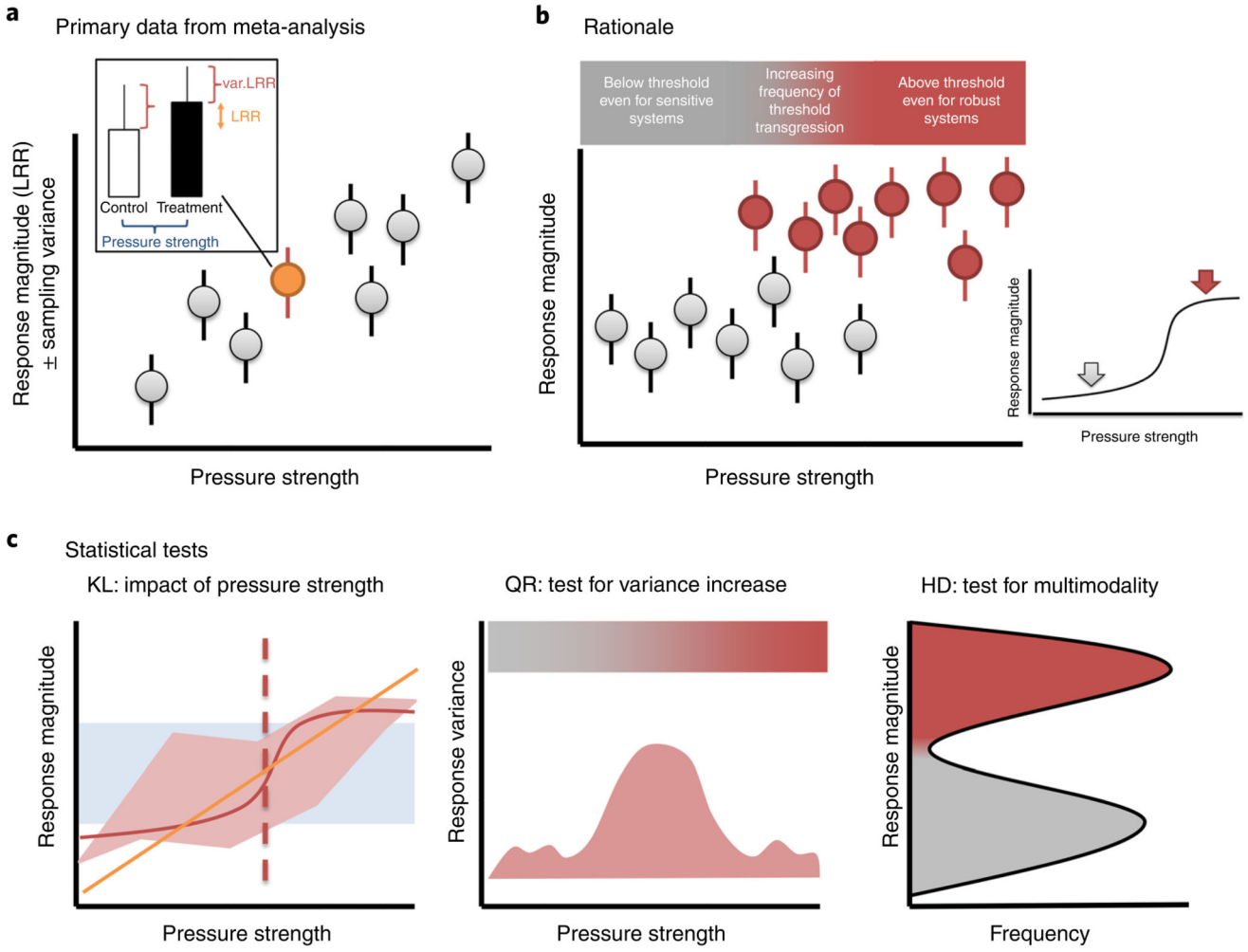


Fig. 1. Detecting thresholds in response to environmental change.

a, Classically, the approach to detecting thresholds is to address the discontinuity of responses to an environmental driver over time. Instead of a temporal axis, our analyses use the multitude of experiments or observations testing the same driver in independent studies. Each meta-analysis summarizes the results of multiple experiments characterized by different magnitudes of the same pressure and response magnitudes ± sampling variance. The basis of each meta-analysis is represented by single experiments (or observational studies) measuring the response in a variable of interest in control and disturbed environments (insert). The distance in the environmental variable (for example, temperature in warming experiments) between control and treatment gives the intensity of the pressure, the LRR measure the relative change in the response variable (for example, plant biomass) based on treatment and control means, whereas the pooled standard deviations result in an estimate of sampling variance per study (var.LRR). **b**, If the response shows discontinuity, we expect a tendency towards a new category of responses (red cases reflecting critical transitions) at higher pressure strengths. **c**, We developed two robust non-parametric test statistics and assessed their statistical significance using permutation tests: KL divergence to test for general changes in the response magnitude along the pressure gradient and

the weighted QR of interquantile (5–95%) ranges to test for changes in the variability of effect sizes. We tested for multimodal frequency distribution of effect sizes, reflecting alternative responses to a common driver using the HD test. To visualize the KL approach, we indicated a potential realized distribution of responses by a red area, compared to a randomized distribution (blue area; see Methods). The significant deviation between realized and randomized responses can occur if there is gradual increase in response with increasing pressure (orange line) or if shifts in the response (red solid line) occur at a threshold (vertical dashed line).

Europe PMC Funders Author Manuscripts

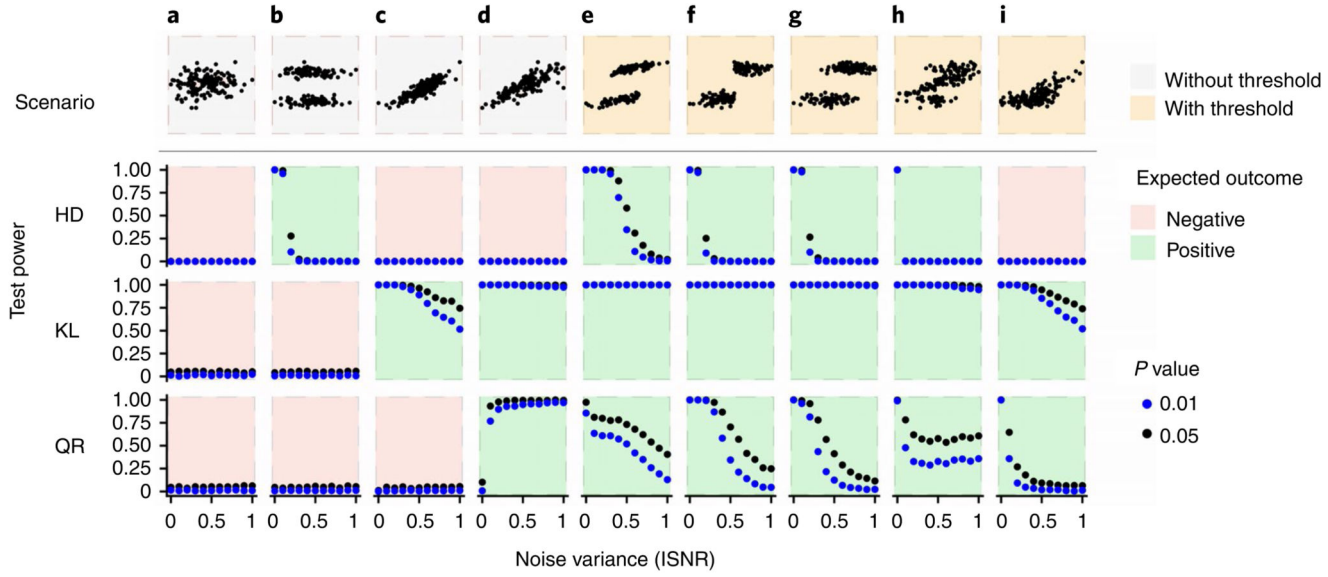


Fig. 2. Detection probability for thresholds in global change experiments using kernel density estimation.

We analysed the test power for nine scenarios of responses to pressure in meta-analyses. The derivation of each scenario is described in the Supplementary Text and Extended Data Fig. 3, the top panel presents an illustration of each scenario (full description in Extended Data Fig. 3). Each scenario is represented in a column, the test power is then given for each of the three statistical tests used in our analyses (HD, Hartigan's dip test; KL, Kullback–Leibler; QR, quantile ratio). **a-d**, These scenarios do not comprise a threshold, where the scenario shown in **a** is the null model without an effect of the pressure on the response. **e-i**, These scenarios do comprise a threshold, for the last two combined with intermediate responses. For the three statistical tests used in our analyses, the expected outcome is colour-coded, with green representing that the test should be significant. Test power is given as the proportion of 1,000 simulated datasets for which the tests were significant with a probability $P = 0.05$ (black) and $P = 0.01$ (blue) along a gradient of increasing noise variance (inverse signal-to-noise ratio, ISNR). Bandwidth selection was based on the 'solve-the-equation' method of Sheather and Jones[59]. The estimated bandwidth was adjusted by a factor of 2.5 in each case because this optimized test power for all cases. The three tests together allow perfect detection of thresholds in the absence of noise (**e-h**). If threshold-type and gradual responses are mixed (**i**), the analysis of multimodality (HD) is no longer able to pick up the threshold embedded in the data, as the simultaneous increase in mean and variance of the response (as in **d**) masks modes in the response distribution. With increasing noise variance, however, the detection probability for thresholds via HD and QR rapidly decrease.

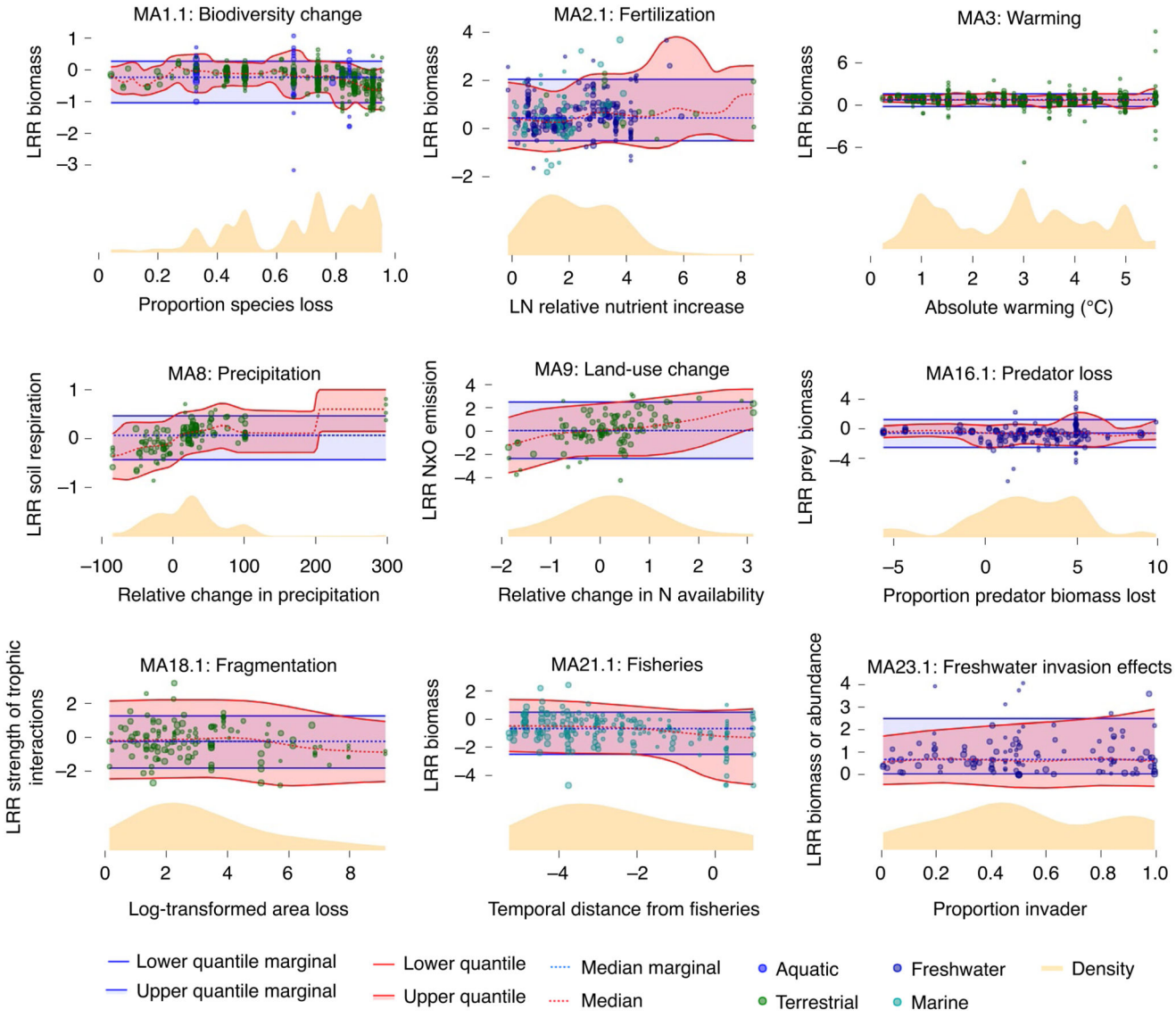


Fig. 3. Example meta-analyses testing for changes in the response magnitude along with increasing pressure intensity.

Red- and blue-shaded regions indicate the (5-95%) interquantile ranges for, respectively, the bivariate data (including the pressure gradient) and the marginal distribution of LRR (integrating out the pressure gradient). Dashed red and thick blue lines trace the related median (50% quantile). Overlain are the data points and, at the bottom, the yellow-shaded area indicates the distribution of stressor variables resulting from a weighted kernel density estimation. Circle size reflects statistical weight. The suggestive break in the responses in MA8 is induced by a lack of data covering intermediate pressure magnitude. LN, natural log.

Europe PMC Funders Author Manuscripts

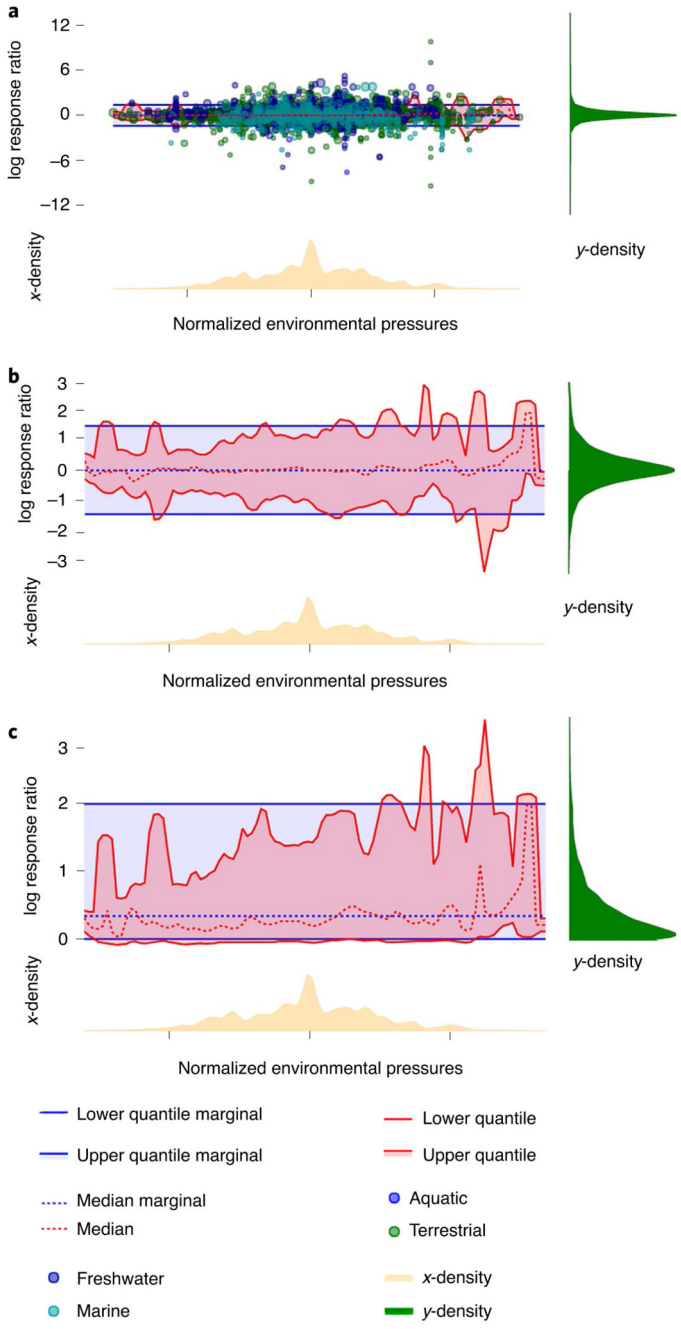


Fig. 4. Analysis of aggregate data across meta-analyses.

a, LRR of ecological processes across a gradient of environmental change, where the different pressures were normalized to a median of 0 and a range of -1 to 1. Circle size reflects statistical weight. Shaded regions indicate the interquantile (5–95%) ranges for the marginal distribution (blue) and the bivariate distribution (red). Density of values along the stressor and the response axis are given below (yellow) or at the right margin (green), respectively. **b**, Same as **a** but without single effect sizes, focusing on the distribution of

response magnitudes over the normalized pressure gradient. **c**, Same as **b** but for absolute response magnitudes. Note the change in scale of the *y* axis in the three panels.

Europe PMC Funders Author Manuscripts