Multiple stressor null models frequently fail to detect most interactions due to low statistical power

Benjamin J. Burgess[1]

Michelle C. Jackson[2]

David J. Murrell[1]*

[1] **Centre for Biodiversity and Environment Research, Department of Genetics, Evolution and Environment, University College London, Gower Street, London, WC1E 6BT, United Kingdom**

[2] **Department of Zoology, University of Oxford, Oxford, OX1 3SZ, United Kingdom**

***corresponding author (d.murrell@ucl.ac.uk)**

Temporary Running Head: Limited power in multiple stressor null models

Word count: 6933

Abstract

1. Most ecosystems are subject to co-occurring, anthropogenically driven changes and understanding how these multiple stressors interact is a pressing concern. Stressor interactions are typically studied using null models, with the additive and multiplicative null expectation being those most widely applied. Such approaches classify interactions as being synergistic, antagonistic, reversal, or indistinguishable from the null expectation. Despite their wide-spread use, there has been no thorough analysis of these null models, nor a systematic test of the robustness of their results to sample size or sampling error in the estimates of the responses to stressors.
2. We use data simulated from food web models where the true stressor interactions are known, and analytical results based on the null model equations to uncover how (i) sample size, (ii) variation in biological responses to the stressors and (iii) statistical significance, affect the ability to detect non-null interactions.
3. Our analyses lead to three main results. Firstly, it is clear the additive and multiplicative null models are not directly comparable, and over one third of all simulated interactions had classifications that were model dependent. Secondly, both null models have weak power to correctly classify interactions at commonly implemented sample sizes (i.e., ≤6 replicates), unless data uncertainty is unrealistically low. This means all but the most extreme interactions are indistinguishable from the null model expectation. Thirdly, we show that increasing sample size increases the power to detect the true interactions but only very slowly. However, the biggest gains come from increasing replicates from 3 up to 25 and we provide an R function for users to determine sample sizes required to detect a critical effect size of biological interest for the additive model.
4. Our results will aid researchers in the design of their experiments and the subsequent interpretation of results. We find no clear statistical advantage of using one null model over the other and argue null model choice should be based on biological relevance rather than

statistical properties. However, there is a pressing need to increase experiment sample sizes otherwise many biologically important synergistic and antagonistic stressor interactions will continue to be missed.

Keywords

Ecosystem drivers; Experimental design; Sample size; Critical effect size; Synergism; Antagonism; Power analysis.

54 ***Introduction***

55 Most, if not all ecosystems are currently being impacted by multiple co-occurring, anthropogenically-
56 driven, stressors including land-use change, invasive species, climate change, and pollution (**Brook et**
57 **al., 2008**; **Dirzo et al., 2014; Hillebrand et al., 2020; Jackson et al., 2020**), and these stressors (also
58 known as disturbances, drivers, factors or pressures, **Orr et al. 2020**) can have a wide range of
59 effects on biodiversity and ecosystem services (**Newbold et al., 2015; Sala et al., 2021; Tittensor et**
60 **al., 2014; Vinebrooke et al., 2004**). A key goal of ecology is to determine whether the combined
61 effect of multiple stressors is more than (synergistic) or less than (antagonistic) that expected by
62 their individual effects (**Orr et al. 2020**). Such an understanding is crucial to designing management
63 strategies to mitigate against the effects of the stressors (e.g. **Lindenmayer et al. 2020**), and indeed
64 may prevent these strategies from making the situation worse (**Brown et al. 2013**).

65

66 The effects of multiple interacting stressors are predominately determined through the
67 implementation of null models where the observed response is compared to an expectation that the
68 stressors are non-interacting (**De Laender, 2018**). Of the range of available null models, the additive
69 null model (**Gurevitch et al., 2000**) is the most widely applied (**e.g., Crain et al., 2008; Jackson et al.,**
70 **2016**) and has the expectation that the overall effect of the multiple interacting stressors is simply
71 equal to the sum of the effects of the stressors acting individually. The statistical test is therefore
72 whether this null model can be rejected in favour of an alternative hypothesis that interactions are:
73 i) greater than anticipated by the additive null model (*Synergistic interactions*); ii) less than the sum
74 of the individual stressor effects (*Antagonistic interactions*); or iii) opposite to that suggested by the
75 additive null model (*Reversal interactions*) (see **Figure 1**).

76

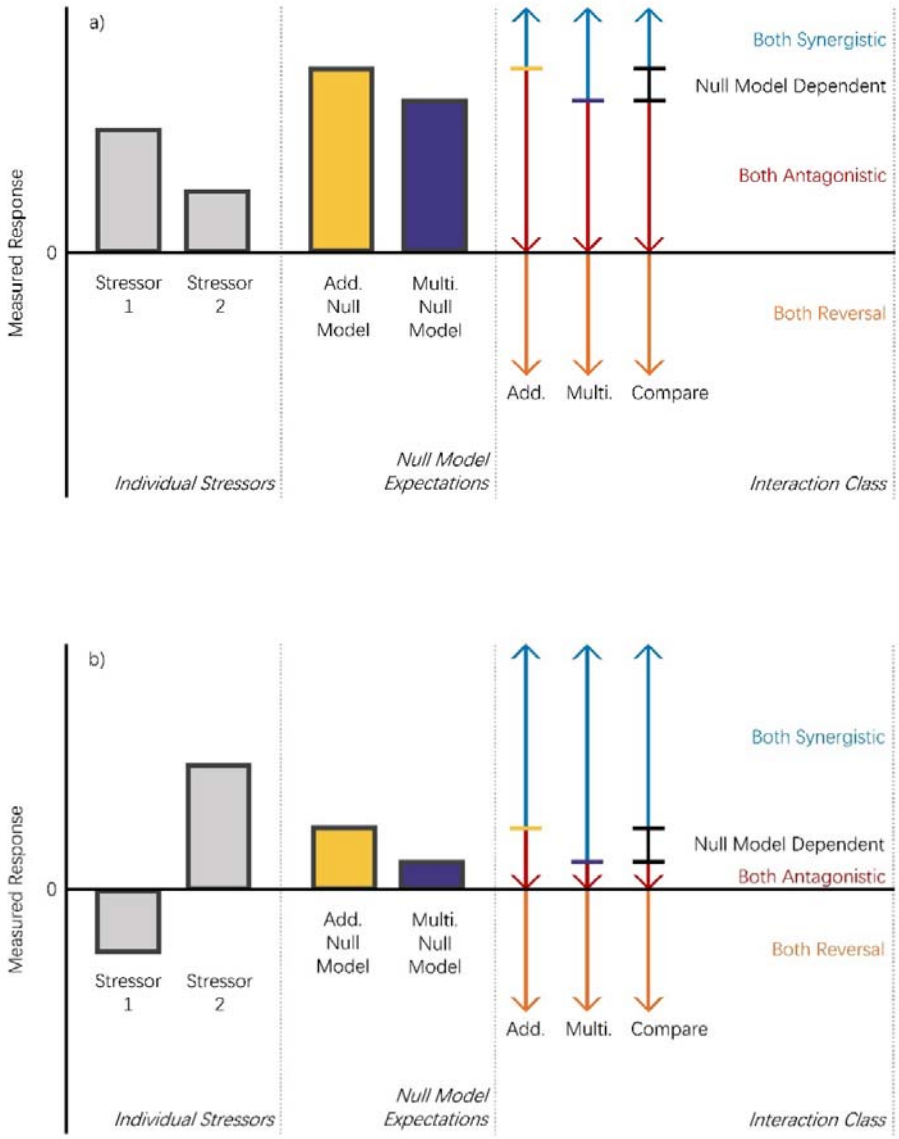

Figure 1: Comparison of the additive and multiplicative null models, illustrating the conditions under which the different interaction classifications are observed. Examples are for both Stressor 1 and Stressor 2 act upon the response in (a) the same direction, and (b) contrasting directions. Individual Stressors: Grey bars represent the change in a measured response due to a given stressor (Stressor 1 or Stressor 2). Null Model Expectations: Yellow bar represents the expected response under an additive null model; Purple bar represents the expected response under a multiplicative null model. Interaction Class: Arrows denote how an interaction would be classified for a given observed

response. Blue arrows denote synergistic interactions; Pink arrows denote antagonistic interactions; Orange arrows denote reversal interactions. For the additive and multiplicative null models, a null interaction is shown by the purple and yellow lines respectively. When comparing the additive and multiplicative null models, the black line denotes the range of observed responses where the additive and multiplicative null models classify interactions differently (i.e., the interaction class is null model dependent).

While the majority of studies currently apply the additive null model there is a growing body of literature suggesting alternative null models (**e.g., Tekin et al., 2020; Thompson et al., 2018**) that may potentially be more appropriate when investigating the impacts of multiple stressors (**Dey & Koops, 2021; Schäfer & Piggott, 2018**). The greatest issue with the additive null model is that it can lead to biologically meaningless expectations. For example, if stressor A (perhaps a pollutant) reduces biomass of our target species by 60%; and stressors B (perhaps an invasive species) reduces it by 50%, then the expected joint effects under the additive model would be a reduction in biomass of 110%. Similar problems arise for other responses that may have bounded values (e.g. survival and/or mortality).

The most common alternative is the multiplicative null model (**Hawkes & Sullivan, 2001; Lajeunesse, 2011)**, where the expected interaction is simply the product of the responses to the two individual stressors acting alone. Under our example above the null expectation for the stressor interaction would be for a reduction in biomass of 80% (i.e. (0.6+0.5) – (0.5*0.6)). As such, it has been suggested that a multiplicative null model may better reflect biological and ecological systems because it can prevent biologically implausible expectations (**Stephens et al. 2013**). Deviations away from the null expectation can follow the same nomenclature as for the additive model, but the same interaction can result in a different classification under the two null models (**Figure 1**). Very few analyses have

compared interactions in the same data using both null models (see **Stephens et al., 2013**), but model dependent interaction types means (i) management decisions can vary greatly depending on which null model is used, with the potential for remedial action to be ineffective or even detrimental if the 'wrong' null model is chosen, and (ii) comparisons of experiments or reviews that use different null models are very difficult to interpret as we do not yet know the circumstances where the model inferences will differ. It is therefore imperative to understand how the additive and multiplicative null models inherently relate to one another.

An arguably greater problem is that there is very little understanding of null model robustness to data uncertainty (variation), or consequently, the sample sizes required to detect non-null interactions. This means there is no guidance of how many replicates are required to guarantee the detection of a non-null interaction of a given strength should it exist in the data, and researchers have been left to base the crucial decisions on experimental design solely on heuristics or economic constraints (**Lakens, 2021**). Without this guidance there is a risk that many experiments will be statistically underpowered and lack the ability to reject the null model for all but the most extreme cases. Indeed, several reviews have reported >70% incidences of null interactions (**Stephens et al., 2013; Lange et al., 2018; Burgess et al., 2021**); but the failure to reject the null model prediction is not proof it is true, and we cannot be sure how many biologically important ecological surprises remain hidden from view due to data constraints. As a result, discussions regarding the relative commonness or rarity of ecological surprises (**e.g., Darling and Côté, 2008; Côté et al., 2016**) cannot reach any firm conclusions until the ability of the null models to detect the interactions has been quantified.

Recently, **Burgess et al. (2021)** simulated data using food web models where parameters were altered under pairs of 'ecological stressors' to investigate the ability of the additive null model to

uncover the true classification in the face of various levels of measurement/observation error. They found that it can be very difficult to correctly reject the additive null model at even very low levels of observation error, and this means many strong synergies and antagonisms may be left classified as indistinguishable from additive. However, the analyses of **Burgess et al. (2021)** fall short of a power analysis because they did not investigate the relationship between detection of true stressor-pair interactions and the sample size of the data for the treatment means. The number of replicates in experiments to investigate stressor interactions rarely reach double figures. As examples, two recent meta-analyses (**Gomez Isaza et al., 2020; Seifert et al., 2020**) included no experiments with more than six replicates per treatment, while a third (**Burgess et al., 2021**) found <1% of experiments used more than eight replicates per treatment (**Supplementary Information**). It remains unknown whether increasing current sample sizes would greatly improve the statistical power of the additive null model and potentially lead to fundamental changes in our understanding of how pair of ecosystem stressors interact. A further open question is to how general the results of **Burgess et al. (2021)** are to other null models, or whether the additive model is statistically a poor choice to detect stressor interactions.

Here, we focus on closing these important gaps in knowledge by focussing on the additive and multiplicative null models and answering the following questions: (1) What are the conditions under which the additive and multiplicative null models agree, and disagree? (2) To what degree does treatment uncertainty (in the form of observation error) reduce the ability of both the additive and multiplicative null models to correctly infer the 'true' interaction classification? (3) What is the relationship between the power to correctly reject the null models and sample size, and are current ranges of sample sizes adequate to detect most non-null interactions?

Methods

The null model approach requires a factorial experiment design with four treatments that each measure the same metric of interest (e.g., individual survival; population density or biomass; species richness). Each measure X_x, is the mean value of this metric taken over N_x replicates, where $x \in \{C, A, B, I\}$. The first treatment, *C*, is the control which is simply the system (i.e., population, community) of interest with no stressor. There are two treatments (*A*, *B*) that account for the individual effects of each of the stressors of interest in isolation. The final treatment, I, is the estimate of the effect of both stressors acting simultaneously. Associated with each treatment is an estimate of the standard deviation of the response to the treatment, and these are denoted by SD_X, where $x \in \{C, A, B, I\}$. All three elements, X_x, SD_X, and N_x are required for the additive and multiplicative null models which are defined below, and from this input each null model computes an effect size, with associated confidence intervals from which the interaction type is inferred.

Additive null model

The additive null model compares the observed interaction effect $X_O = X_I - X_C$, with $X_E = X_A + X_B - 2X_C$, the response expected under the assumption the joint effect is equal to the sum of the individual effects of stressors *A* and *B*. To compute effect sizes we use Hedges' d which is unbiased by small sample sizes (**Hedges & Olkin, 1985**). The calculation of the additive effect size, (ES_{Add}), is given by Equation 1.1, and this requires a pooled sampling standard deviation, denoted by *s* (Equation 1.2); while *J(m)* (Equation 1.3) is the small sample bias correction factor (**Borenstein et al., 2009**). For each additive effect size a corresponding variance, (V_{Add}), standard error, (SE_{Add}), and confidence intervals, (CI_{Add}), are calculated by Equations 1.4 - 1.6. Here, we set the significance level to 0.05 ($Z_{\alpha/2}$ = 1.96), hence 95% confidence intervals are calculated.

$$ES_{Add} = \frac{X_I - X}{} \qquad 1.1)$$

$$\cdot J(m)$$

1.2) $$s = \sqrt{\frac{(N_I - 1)\cdot(SD_I)^2 + (N_A - 1)\cdot(SD_A)^2 + (N_B - 1)\cdot(SD_B)^2 + (N_C - 1)}{N_I + N_A + N_B + N_C - 4}}$$

1.3) $$J(m) = 1 - \frac{3}{4\cdot(N_I + N_A + N_B + N_C - 4) - 1}$$

1.4) $$V_{Add} = (J(m))^2 \cdot \left(\frac{1}{N_I} + \frac{1}{N_A} + \frac{1}{N_B} + \frac{1}{N_C} + \frac{(ES_{Add})^2}{2(N_I + N_A + N_B + N_C)}\right)$$

1.5) $$SE_{Add} = \sqrt{V_{Add}}$$

1.6) $$CI_{Add} = Z_{\alpha/2}\cdot SE_{Add}$$

Multiplicative null model

The effect size for the multiplicative null model is derived from the response ratio $RR = \ln\left(\frac{X_O}{X_E}\right) = \ln(X_O) - \ln(X_E)$ **Lajeunesse (2011)**, where the observed effect $X_O = \frac{X_I}{X_C}$ (the joint effect of the stressors), and the expected effect $X_E = \frac{X_A \cdot X_B}{X_C^2}$. The multiplicative effect size, (ES_{Mul}), is then given by Equation 2.1; while the corresponding variance, (V_{Mul}), standard error, (SE_{Mul}), and 95% confidence intervals, (CI_{Mul}), are determined using Equations 2.2 – 2.4 (**Lajeunesse 2011**).

2.1) $$ES_{Mul} = \ln(X_I) - \ln(X_A) - \ln(X_B) + \ln(X_C)$$

2.2) $$V_{Mul} = \frac{(SD_I)^2}{(X_I)^2\cdot N_I} + \frac{(SD_A)^2}{(X_A)^2\cdot N_A} + \frac{(SD_B)^2}{(X_B)^2\cdot N_B} + \frac{(SD_C)^2}{(X_C)^2\cdot N_C}$$

2.3) $$SE_{Mul} = \sqrt{V_{Mul}}$$

2.4) $$CI_{Mul} = Z_{\alpha/2}\cdot SE_{Mul}$$

Interaction classifications

Using the naming convention of **Orr et al., (2020**) (see also Figure 1), an interaction is classified as null if the 95% confidence intervals for that interaction overlapped zero. In this case the interpretation is not that the null model is true, rather that we have insufficient data to reject the null hypothesis. Interactions were classed as synergistic if effect size was positive and the 95% confidence intervals for that interaction did not overlap zero. Antagonistic and reversal interactions were classified whenever the effect size is negative, and the 95% confidence intervals do not overlap zero. For an interaction to be classed as antagonistic, the expected and observed interaction effects had to both act in the same direction (e.g., for an additive interaction: $X_I - X_C > 0$ & $X_A + X_B - 2X_C > 0$). In contrast, for an interaction to be classed as reversal, the expected and observed interaction effects had to both act in differing directions (e.g., for an additive interaction: $X_I - X_C < 0$ & $X_A + X_B - 2X_C > 0$). The method for classifying interactions is explored more in the **Supplementary Information**.

Data simulation

We use the same model as **Burgess et al. (2021)** to simulate data from food chains where population densities are the response metric to 'stressors' that alter key parameters in the models. We refer the reader to the **Supplementary Information** for full details, including formulae and details of distributions. The food chains are based on Lotka-Volterra consumer-resource equations enabling us to establish food chains of three, four, or five trophic levels and determine the densities of populations at equilibrium. Regardless of the length of food chain considered, the models contain three distinct types of trophic level. These levels are: i) basal trophic level, equivalent to a pool of a limiting nutrient such as silica; ii) consumer trophic level (including primary producers that require the nutrient), which only consumes the trophic level directly below, and is only consumed by the trophic level directly above; iii) apex consumer trophic level, which only consumes the trophic level directly below, but is not itself consumed by another level.

The equilibrium densities are governed by five key parameters (α_i, ε_i, δ_i, δ_i, ω) which stay constant over time. The attack rate, α_i, denotes the rate at which trophic level *I* feeds upon a resource (trophic level *i-1*). The conversion efficiency, ε_i, represents the proportion of the consumed resource that is converted into new consumers within trophic level *i*. The density independent mortality rate, δ_i, represents the background mortality rate of trophic level *i*. The density dependent mortality rate, λ_i, increases the mortality rate of a trophic level with increasing population density. Finally, the basal input rate, ω, represents the constant rate at which the basal trophic level increases in density.

Stressors are simulated by altering the values of the key parameters relative to a control (baseline) set of values and stressor interactions are computed using a factorial experiment design. Key parameter values for each trophic level are drawn from pre-determined uniform distributions. Hence for example, the value for the conversion efficiency parameter at trophic level *i* is unlikely to be the same as that for trophic level *j*. Food chains are subjected to a stressor by altering the value for a single parameter (α, ε, δ, or ω) chosen at random for a single trophic level, and keeping all other parameters constant under the control values. All parameters are equally likely to be selected to be stressed, although a single parameter is only ever impacted by one stressor (i.e., two stressors cannot act on the same parameter). For the interaction treatment all parameters were under control conditions, except for those two which were stressed in the individual treatments which were instead assigned their stressed values. We then calculate the equilibrium densities of each trophic level within the food chain for all four treatments using Mathematica v10.4 (**Wolfram Research Inc., 2016)**. We retained datasets where all four treatments resulted in biological feasible (i.e. all densities non-zero) and locally stable equilibria **(Nattrass et al. 2012)**. All retained datasets showed equilibria that were either stable nodes or stable foci. We then selected one trophic level at random (excluding the basal level) to be the focal response group and used its equilibrium densities across the four

treatments to classify the stressor interactions using the null models (Equations 1 and 2, *Interaction classifications section*). Using this approach, we generated a dataset comprising of 100,000 interactions, with approximately equal numbers of equilibrium densities for populations from the three different lengths of food chain.

Treatment means, treatment uncertainty, and sample sizes

Given that the food chain models are deterministic, there is an absence of data uncertainty at this stage. Interactions estimated in the absence of any data uncertainty are referred to as the 'true' stressor interaction. To mirror empirical experiments, we modelled treatment uncertainty, in the form of observation error for the equilibrium densities recorded at each treatment within each of the 100,000 interactions. In brief, we follow the method of **Burgess et al., (2021)**, with each equilibrium density being multiplied by a number drawn from a Gaussian distribution with a mean of 1.00 and a standard deviation σ, with this process being repeated N times per treatment. Treatment means are then estimated by taking the mean of the values from these replicates. Smaller values of σ will result in a tighter distribution of observed mean densities around the true population mean compared to larger values of σ. Our approach allows for an intuitive understanding of σ, as we can expect 99.7% of all observed treatment densities to fall within 3σ of the true value. We use 250 levels of σ, hence referred to as levels of observation error, ranging from 1.0×10^{-3} to 2.5×10^{-1} (see **Supplementary Information** for more details).

By incorporating 'noise' in this way we were able to test the ability of the null models to recover the true stressor interaction in the face of model uncertainty, and for different sample sizes. We then compute the probability of assigning the true interaction classification to each interaction under every level of observation error, thereby investigating the ability of each null model to correctly

266 classify interactions (i.e., true positives) in the face of uncertainty (see Supplementary Information
267 for more details).

268

269 ***Results***

270 *Stressor-pair classification across the null models*

271 To compare null model relationships, each of the 100,000 simulated interactions in the absence of
272 uncertainty was assigned an interaction using both models. Overall, both the additive and
273 multiplicative null models report similar frequencies of the different interaction classifications, with
274 both reporting that antagonistic interactions are the most prevalent (additive 0.493; multiplicative
275 0.519), followed by a slightly reduced number of synergistic interactions (0.485; 0.462), minimal
276 frequencies of reversal interactions (0.023; 0.020), and a complete absence of null interactions
277 (0.000; 0.000). However, this overview masks the important fact that one third of individual
278 interactions change classification when moving from one null model to the other (Table 1). Null
279 model dependent classifications are predominately interactions assigned as synergistic by one null
280 model and antagonistic by the other (Table 1, see also Figure 1).

281

Table 1: Proportions of interactions classed differently, or the same, across additive and multiplicative null models. Shaded boxes indicated the proportion of interactions where both null models agree.

		Additive Null Model			
		Synergistic	Antagonistic	Reversal	Null
Multiplicative Null Model	Synergistic	0.299	0.158	0.005	0.000
	Antagonistic	0.181	0.328	0.010	0.000
	Reversal	0.005	0.007	0.007	0.000
	Null	0.000	0.000	0.000	0.000

Robust of the null models to data uncertainty (observation error)

Both null models are sensitive to data uncertainty in the form of observation error, with the correct classification of synergistic and antagonistic interactions rapidly declining as error increases (Figure 2). In contrast, classifications of reversal interactions, although quite rare, are much less affected by observation error. The frequency of null interactions rapidly increases with error, accounting for over 50% of all interactions for both null models by $\sigma \sim 0.005$ (i.e., when nearly all sampled treatment means are within 1.5% of the true values). This shows how even low levels of observation error lead to the failure to correctly reject the null model (i.e., low statistical power).

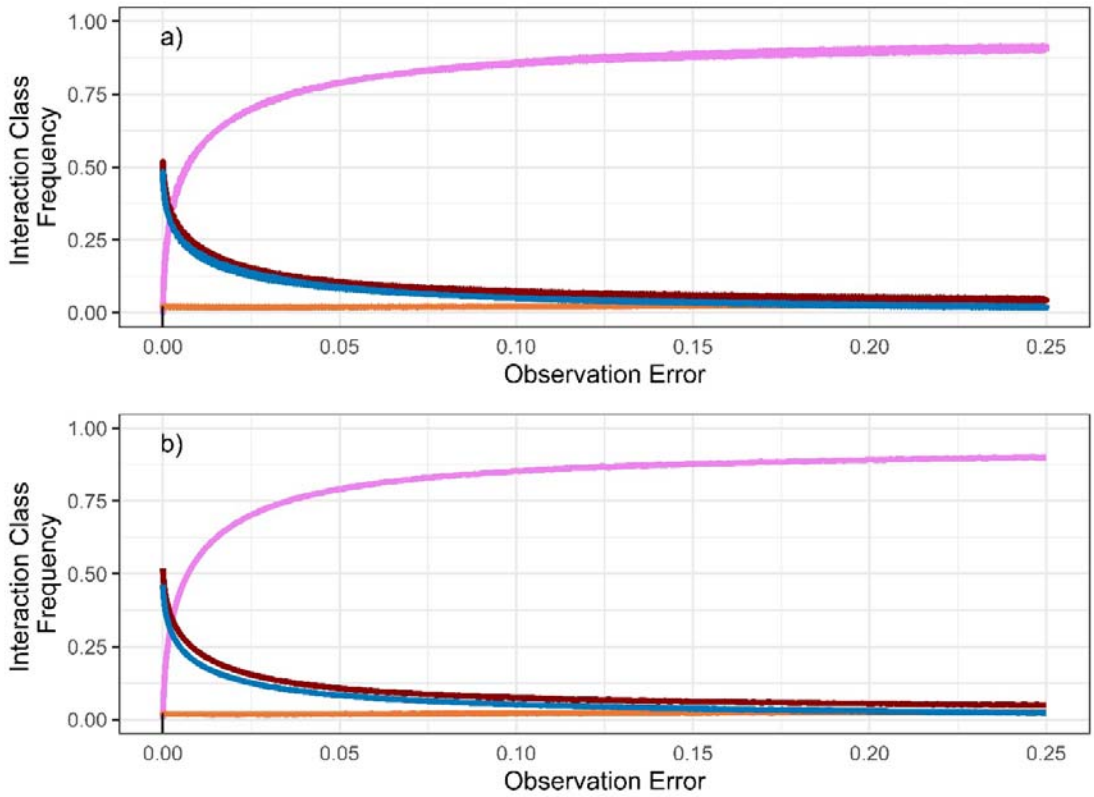

Figure 2: The effect of observation error on the frequency of interaction classifications, showing how the proportions differed from those in the absence of observation error, using an additive null model (a) and multiplicative null model (b). Frequencies of the different interaction classes for the 100,000 simulated interactions at each level of observation error using either null model. Pink line indicates null interactions. Orange line denotes reversal interactions. Blue line indicates synergistic interactions. Red line denotes antagonistic interactions.

Sample size and true positive rate

Increasing sample size naturally leads to a higher rate of correctly rejecting both the additive and multiplicative null models (i.e., an increase in the true positive rate correctly detecting an interaction) across all observation error levels tested (Figure 3). However, after an initial increase in true positive rates by moving from N = 3 to approximately N = 15 replications, further improvement

is slow, and even for large sample sizes, it can be difficult to detect the 'true' interaction if observation error is anything other than minimal. Moreover, there is no clear superiority in performance for either model with the multiplicative model showing higher statistical power for all sample sizes at the lowest observation errors (i.e., $\sigma = 0.001$ or $\sigma = 0.01$), whereas the additive null model shows higher power for the highest level ($\sigma = 0.1$) of observation error.

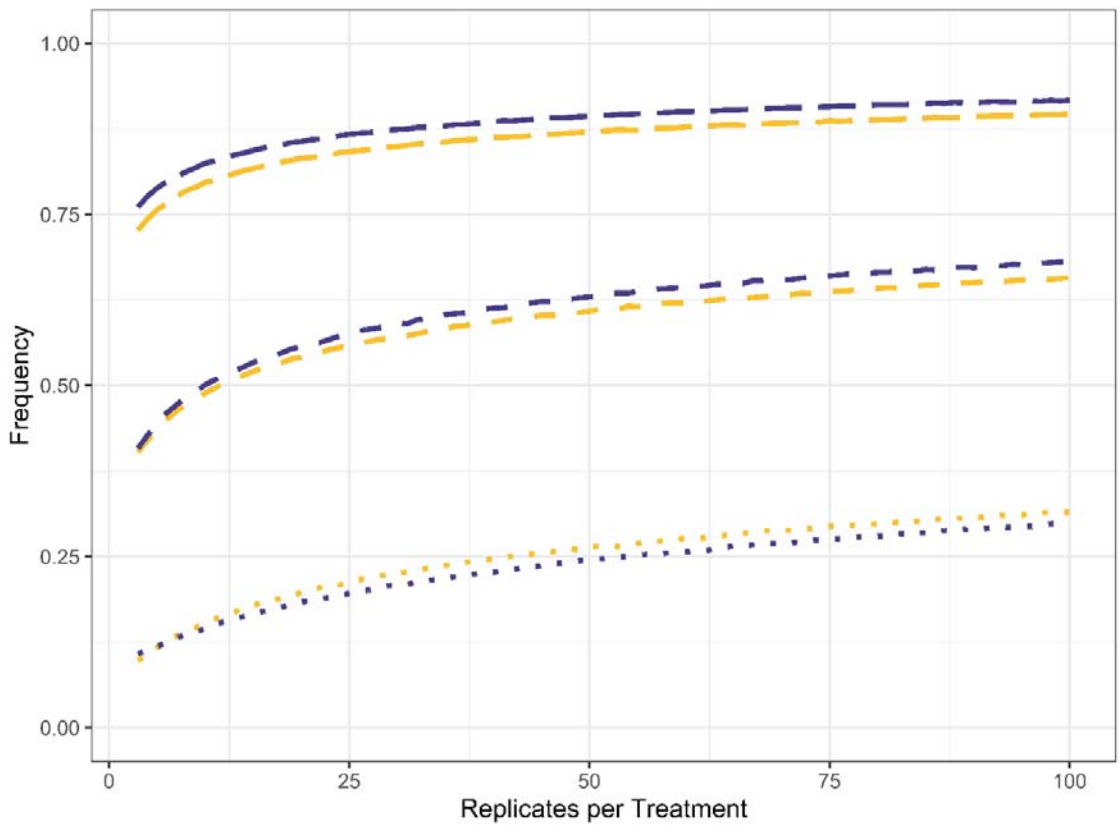

Figure 3: Frequency of the 100,000 simulated interactions that are correctly classified (i.e., true positive rate) for a range of different treatment replicates. Three different levels of observation error are considered (0.001, 0.01, 0.1), with these being indicated by longdashed, dashed, and dotted lines respectively. Yellow lines denote use of an additive null model, purple lines denote use of a multiplicative null model.

Critical effect sizes

By combining and rearranging Equations 1.1 – 1.6, we can show that for a given sample size (N), there is a minimum additive effect size value, also referred to as the critical effect size Mudge et al. (2012), which must be exceeded before the interaction can be correctly classified as being non-null:

$$|ES_{Add}| > \sqrt{\left(\frac{1}{N_C} + \frac{1}{N_A} + \frac{1}{N_B} + \frac{1}{N_I}\right) \cdot \left(\frac{2 \cdot (Z_{\alpha/2})^2 \cdot (N_C + N_A + N_B + N_I) \cdot \left(J(m)\right)^2}{2 \cdot (N_C + N_A + N_B + N_I) - (Z_{\alpha/2})^2 \cdot \left(J(m)\right)^2}\right)}$$

(Inequality 1)

Unfortunately, the multiplicative null model is too complex to yield a solution. Figure 4 illustrates the relationship between critical effect sizes and samples sizes for both the additive (Figure 4a) and multiplicative (Figure 4b) null models using our simulated data. For the additive null model, Inequality 1 is confirmed as the exact criterion for whether an effect size leads to a rejection of the null model or not (Figure 4a), whereas the broken nature of the threshold for the multiplicative model (Figure 4b) confirms its complex relationship with sample size. Importantly, both null models show a rapid decline in critical effect size as sample sizes increases from N = 3 to approximately N = 25, and in both models the critical effect size more than halves in this range (albeit on different scales).

As an aid to experimental design we provide an R code function (Supplementary Information) that takes the choice of significance level (alpha) and the potential sample size (N) for all treatments and returns the critical effect size ($|ES_{Add}|$) for the additive model. For example, setting alpha = 0.05 and N = 3, leads to $|ES_{Add}|$ = 2.192, whereas increasing sample size to N = 16 leads to $|ES_{Add}|$= 0.982. Only observed effect sizes greater than $|ES_{Add}|$ will be classified as non-additive. To judge how useful an increase in N might be, we can see that in the freshwater stressor interactions analysed in Burgess et al. (2021) approximately 20% of interactions are classified as non-null for N = 3, and this rises to approximately 60% if N = 15 (Figure S4, Supplementary Information).

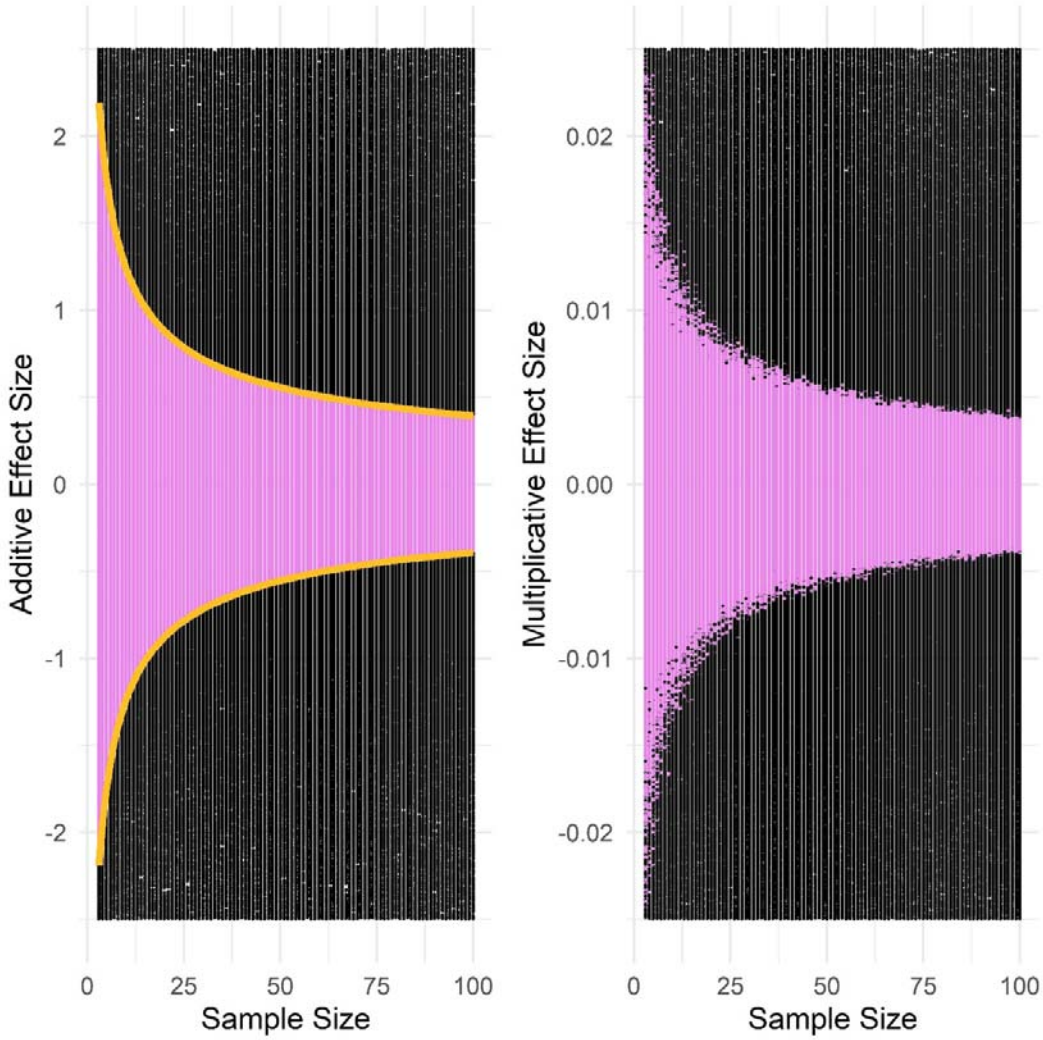

Figure 4: The relationship between treatment sample size and effect size for the additive (a) and multiplicative (b) null models. For each treatment sample size (3-100), points are plotted from simulations where an observation error of 0.01 was used. Pink points illustrate interactions that are classified as null, while black points denote interactions classified as antagonistic, reversal, or synergistic. The yellow line (Figure 4a) denotes the critical effect size threshold, computed from Inequality 1, that interaction effect sizes must exceed in order for an interaction to be significantly different to zero (and hence classified as being non-null) under the additive null model. As in the methods, 95% confidence intervals were calculated (i.e., for Inequality 1, $Z_{\alpha/2} = 1.96$).

356 ***Discussion***

357 Our explicit aims were to better understand the statistical properties of the two most frequently
358 used null models for multiple stressor interactions. Our analyses lead to three main conclusions that
359 have important consequences for both experimental design and interpretation of results. Firstly, the
360 additive and multiplicative null models make different assumptions that often lead to different
361 interpretations about the stressor interactions, rendering direct comparisons between the model
362 meaningless. Secondly, both models are very sensitive to uncertainty in the estimation of the
363 stressor effects (i.e., the biological responses), and it can be very hard to reject the null interaction
364 even with very small observation errors. Consequently, only interactions that very strongly deviate
365 from the null expectation can be uncovered, meaning it is likely that many non-null interactions may
366 have been missed in previous experiments. Finally, our results show how sample sizes need to be
367 unrealistically large to uncover these 'hidden' synergies and antagonisms, but that increasing sample
368 sizes from N = 3 to N = 25 yields the biggest increase in statistical power. Overall, we conclude that
369 from an analytical perspective there is not an obviously *better* null model, and that model choice can
370 therefore be based upon biological arguments without undue concerns over loss of statistical power.

371

372 The two null models we have focussed on are based on different assumptions, hence it should be no
373 surprise that they lead to different interaction classifications, but it does highlight the importance of
374 model choice. Differences in the results of null models are well-documented within fields such as
375 pharmacology (**Tang et al., 2015**), where the class of observed interaction may be prefixed by the
376 null model implemented (e.g., "Loewe antagonism", **Greco et al., 1992**), but it has been rarely
377 addressed in multiple stressor literature (but see **Stephens et al. 2013** for a direct comparison on a
378 single dataset). Currently, there is no equivalent framework for interpreting or reporting results for
379 multiple stressor interactions. This discrepancy should be remedied given that the ambiguous usage
380 of terms (such as synergistic or antagonistic interactions) are potentially hindering interpretation of

multiple stressor interactions both within, and across, disciplines (**Dey & Koops, 2021; Orr et al., 2020**). Ultimately, we echo previous calls (**Schäfer & Piggott, 2018**) to thoroughly consider the choice of null model to be implemented, and we argue that this decision needs to be based largely on biological grounds since our results show there is little to choose between the null models when it comes to statistical power. The decision should however be made very carefully, since different choices may well lead to different types of interactions being assigned with the consequences of very different management strategies being required to mitigate the stressor-pair effects. Choosing the null model without reference to the relevance of the biological assumptions behind it could therefore lead to remedial action that makes the situation worse.

The additive and multiplicative null models have many similarities regarding their robustness to data uncertainty and their responses to sample size. In particular we show that, for likely realistic levels of observation error, both models have poor statistical power (i.e., ability to correctly reject the null model) at commonly used sample sizes of $3 \leq \mathrm{N} \leq 8$. Several meta-analyses have highlighted a dominance for null interactions (e.g., **Burgess et al., 2021; Lange et al., 2018; Stephens et al., 2013**), and our results suggest this should be expected on purely statistical grounds. The important corollary is that many non-null interactions are expected to be hidden from view, masked by data uncertainty in the form of not only observation error as we model here, but also other forms such as spatial, environmental, or genetic variation (**Higdon, 2013**).

'Will my study answer my research question?' is the most fundamental question a researcher can ask when designing a study **(Johnson et al., 2015**), and its answer requires an understanding of the statistical power of the tools at hand. Unfortunately, our results show that large sample sizes are required to uncover even modest effect sizes (Figure 4). For example, a sample size of around N = 100 is required to detect an effect size of 0.5 as being non-additive (Inequality 1; Figure 4). As is the

case for interaction detection in epidemiological studies (**Greenland, 1983**), we find a slow convergence of statistical power with sample size (Figure 4). Despite these unwelcome findings we do observe a rapid initial increase in the minimum detectable effect size for both models (Inequality 1, Figure 4). This means increasing sample sizes from N = 3 up to around N = 25 can have important consequences for uncovering non-null interactions. Indeed, we estimate that increasing sample sizes to N = 15 would likely result in an increase of around 40% of non-null interactions in the dataset of Burgess et al. (2021) (Supplementary Information). However, increasing sample sizes is not a trivial consideration, and we recommend researchers consider what they believe constitutes a biologically important effect size *before* the experiment is started. This so-called critical effect size (**Lakens, 2021**) is likely to change from study to study as the question, response metric, and taxonomic group(s) change. Having determined the critical effect size means our results (Inequality 1, Figure 4, and our R function *AddCritES* in the SI) can then be used to determine what sample size is required to uncover it, whether this is even possible given budget and or practical constraints, or whether a different approach might be necessary. However, we add that there is still high value in collecting the data if it is then made available for future meta-analytical studies (**e.g., Jackson et al., 2016; Burgess et al., 2021**) where the pooling of data overcomes the low power of small individual study sample sizes.

A final consideration should be given to the choice of significance level. There has been much discussion (e.g. **Mudge et al., 2012**) about how the significance level should be a function of the sample size of the study, and not a fixed entity (e.g., the canonical level of 5%). For cases of low power, Inequality 1 shows that a smaller alpha level (i.e., smaller value for Z) will lead to a smaller critical effect size and therefore higher power to detect non-null stressor interactions. However, raising the significance level comes at the cost of increased Type-1 errors (i.e., false rejection of the null model). We argue that in multiple stressor research the cost of Type-1 errors is outweighed by

the increase in power because (i) we expect most interactions to be truly non-null even if only weakly so (e.g. Table 1); (ii) most studies will not be able to have the large sample sizes required to uncover even relatively strong effect sizes; and (iii) the ecological cost of failing to uncover the true stressor interaction could be high. As shown by **Figure S2** (Supplementary Information), increasing the significance level results in a modest increase in statistical power, across all levels of observation error and number of replicates. Crucially, the altered level of significance must be justified, and not arbitrarily chosen (**Lakens, 2021**)**.** Such justification may be provided by either a cost-benefit analysis (**see Field et al., 2004, Mudge et al., 2012**) or a compromise power analysis (**see Lakens, 2021**), and extensions to our work could consider building a framework to encourage an informed choice of significance level.

Finally, a word on the generality of our findings. The results based upon algebraic relationships (i.e., regarding critical effect size) are general and can be applied to any dataset, whereas results that depend upon data simulation might be dependent upon the distribution of the effect sizes generated by the underlying food chain model (e.g., Table 1; Figures 2, 3). Whilst the food chain model used here has been shown to generate ecologically plausible results (**Burgess et al., 2021**), the relationships revealed here may differ with other datasets (for instance, where a response other than density is considered) which contain differing distributions of effect sizes. Ultimately, we anticipate that many trends will remain across different datasets (e.g., the dominance of null interactions at commonly implemented sample sizes), though others (e.g., the frequency of antagonistic, reversal, and synergistic interactions in the absence of uncertainty) may vary between datasets.

Conclusions

Robust conclusions about the prevalence of non-null interactions can only be reached once there is a strong understanding of the abilities of the statistical tools to uncover them in the face of small sample sizes and the underlying variation in the estimates of the biological responses. Our analyses have confirmed that both additive and multiplicative null models have low power to detect non-null interactions at commonly used sample sizes. However, we have provided guidance on the sample sizes required to uncover specific effect sizes and we hope these results will be used to guide future experimental designs, as well as lead to an increase in sample sizes beyond the typically small numbers (<10). An important next step will be to investigate other null models (e.g. **Schäfer & Piggott, 2018; Tekin et al., 2020**) and assess their abilities to uncover non-null interactions. Finally, our work highlights the importance of null model choice and its effect on interaction classification, and although we do not see any statistical advantage to either the additive or multiplicative null model, deciding which model is biologically appropriate for the study question at hand is an important choice that has consequences for potential downstream management decisions.

470 **References**

471 Borenstein, M., Cooper, H., Hedges, L. and Valentine, J., 2009. Effect sizes for continuous data.
472 *The Handbook of Research Synthesis and Meta-Analysis*, 2, pp.221-235.

473 Brook, B.W., Sodhi, N.S. and Bradshaw, C.J., 2008. Synergies among extinction drivers under
474 global change. *Trends in Ecology & Evolution*, 23(8), pp.453-460.

475 Brown, C.J., Saunders, M.I., Possingham, H.P. and Richardson, A.J., 2013. Managing for
476 interactions between local and global stressors of ecosystems. *PLoS One*, 8(6), p.e65765.

477 Burgess, B.J., Purves, D., Mace, G. and Murrell, D.J. (2021), Classifying ecosystem stressor
478 interactions: Theory highlights the data limitations of the additive null model and the
479 difficulty in revealing ecological surprises. *Global Change Biology*, 27, pp. 3052-3065.

480 Christensen, M.R., Graham, M.D., Vinebrooke, R.D., Findlay, D.L., Paterson, M.J. and Turner, M.A.,
481 2006. Multiple anthropogenic stressors cause ecological surprises in boreal lakes. *Global*
482 *Change Biology*, 12(12), pp.2316-2322.

483 Côté, I.M., Darling, E.S. and Brown, C.J., 2016. Interactions among ecosystem stressors and their
484 importance in conservation. *Proceedings of the Royal Society B: Biological Sciences*,
485 *283*(1824), p.20152592.

486 Crain, C.M., Kroeker, K. and Halpern, B.S., 2008. Interactive and cumulative effects of multiple
487 human stressors in marine systems. *Ecology Letters*, 11(12), pp.1304-1315.

488 Darling, E.S. and Côté, I.M., 2008. Quantifying the evidence for ecological synergies. *Ecology*
489 *Letters*, 11(12), pp.1278-1286.

490 De Laender, F., 2018. Community-and ecosystem-level effects of multiple environmental change
491 drivers: Beyond null model testing. *Global Change Biology*, 24(11), pp.5021-5030.

492 Dey, C.J. and Koops, M.A., 2021. The consequences of null model selection for predicting
493 mortality from multiple stressors. *Proceedings of the Royal Society B*, 288(1948),
494 p.20203126.

495 Dirzo, R., Young, H.S., Galetti, M., Ceballos, G., Isaac, N.J. and Collen, B., 2014. Defaunation in the
496 Anthropocene. *Science*, 345(6195), pp.401-406.

497 Field, S.A., Tyre, A.J., Jonzén, N., Rhodes, J.R. and Possingham, H.P., 2004. Minimizing the cost of
498 environmental management decisions by optimizing statistical thresholds. *Ecology*
499 *Letters*, 7(8), pp.669-675.

500 Gomez Isaza, D.F., Cramp, R.L. and Franklin, C.E., 2020. Living in polluted waters: a meta-analysis
501 of the effects of nitrate and interactions with other environmental stressors on
502 freshwater taxa. *Environmental Pollution*, p.114091.

503 Greco, W., Unkelbach, H.D., Pöch, G., Sühnel, J., Kundi, M. and Bödeker, W., 1992. Consensus on
504 concepts and terminology for combined-action assessment: the Saariselkä agreement.
505 *Arch Complex Environ Stud*, 4(3), pp.65-9.

506 Greenland, S., 1983. Tests for interaction in epidemiologic studies: a review and a study of power.
507 *Statistics in Medicine*, 2(2), pp.243-251.

508 Gurevitch, J., Morrison, J.A. and Hedges, L.V., 2000. The interaction between competition and
509 predation: a meta-analysis of field experiments. *The American Naturalist*, 155(4), pp.435-
510 453.

511 Hawkes, C.V. and Sullivan, J.J., 2001. The impact of herbivory on plants in different resource
512 conditions: a meta-analysis. *Ecology*, 82(7), pp.2045-2058.

513 Hedges, L. V., and I. Olkin., 1985. Statistical methods for meta-analysis. *Academic Press*, New
514 York.

515 Higdon R., 2013. Experimental Design, Variability. In: Dubitzky W., Wolkenhauer O., Cho KH.,
516 Yokota H. (eds) Encyclopaedia of Systems Biology. *Springer*, New York.

517 Hillebrand, H., Donohue, I., Harpole, W.S., Hodapp, D., Kucera, M., Lewandowska, A.M., Merder,
518 J., Montoya, J.M. and Freund, J.A., 2020. Thresholds for ecological responses to global
519 change do not emerge from empirical data. *Nature Ecology & Evolution*, pp.1-8.

520 Jackson, M.C., Loewen, C.J., Vinebrooke, R.D. and Chimimba, C.T., 2016. Net effects of multiple
521 stressors in freshwater ecosystems: a meta-analysis. *Global Change Biology*, 22(1),
522 pp.180-189.

523 Jackson, M.C., Fourie, H.E., Dalu, T., Woodford, D.J., Wasserman, R.J., Zengeya, T.A., Ellender,
524 B.R., Kimberg, P.K., Jordaan, M.S., Chimimba, C.T. and Weyl, O.L., 2020. Food web
525 properties vary with climate and land use in South African streams. *Functional Ecology*,
526 34(8), pp.1653-1665.

527 Johnson, P.C., Barry, S.J., Ferguson, H.M. and Müller, P., 2015. Power analysis for generalized
528 linear mixed models in ecology and evolution. *Methods in Ecology and Evolution*, 6(2),
529 pp.133-142.

530 Lajeunesse, M.J., 2011. On the meta-analysis of response ratios for studies with correlated and
531 multi-group designs. *Ecology*, 92(11), pp.2049-2055.

532 Lakens, D., 2021. Sample Size Justification. *PsyArXiv*. https://doi.org/10.31234/osf.io/9d3yf.

533 Lange, K., Bruder, A., Matthaei, C.D., Brodersen, J. and Paterson, R.A., 2018. Multiple-stressor
534 effects on freshwater fish: Importance of taxonomy and life stage. *Fish and Fisheries*,
535 19(6), pp.974-983.

536 Lindenmayer, D.B., Foster, C.N., Westgate, M.J., Scheele, B.C. and Blanchard, W., Managing
537 interacting disturbances: lessons from a case study in Australian forests. *Journal of*
538 *Applied Ecology*, 57, pp1711– 1716.

539 Mudge, J.F., Baker, L.F., Edge, C.B. and Houlahan, J.E., 2012. Setting an optimal α that minimizes
540 errors in null hypothesis significance tests. *PLoS One*, 7(2), p.e32734.

541 Nattrass, S., Baigent, S. & Murrell, D.J. (2012) Quantifying the Likelihood of Co-existence for
542 Communities with Asymmetric Competition. *Bulletin of Mathematical Biol*ogy 74,
543 pp2315–2338.

544 Newbold, T., Hudson, L.N., Hill, S.L., Contu, S., Lysenko, I., Senior, R.A., Börger, L., Bennett, D.J.,
545 Choimes, A., Collen, B. and Day, J., 2015. Global effects of land use on local terrestrial
546 biodiversity. *Nature*, 520(7545), pp45-50.

547 Orr, J.A., Vinebrooke, R.D., Jackson, M.C., Kroeker, K.J., Kordas, R.L., Mantyka-Pringle, C., Van den
548 Brink, P.J., De Laender, F., Stoks, R., Holmstrup, M. and Matthaei, C.D., 2020. Towards a
549 unified study of multiple stressors: divisions and common goals across research
550 disciplines. *Proceedings of the Royal Society B*, 287(1926), p.20200421.

551 Sala, E., Mayorga, J., Bradley, D., Cabral, R.B., Atwood, T.B., Auber, A., Cheung, W., Costello, C.,
552 Ferretti, F., Friedlander, A.M. and Gaines, S.D., 2021. Protecting the global ocean for
553 biodiversity, food and climate. *Nature*, pp.1-6.

554 Schäfer, R.B. and Piggott, J.J., 2018. Advancing understanding and prediction in multiple stressor
555 research through a mechanistic basis for null models. *Global Change Biology*, 24(5),
556 pp.1817-1826.

557 Seifert, M., Rost, B., Trimborn, S., Hauck, J., 2020. Meta-analysis of multiple driver effects on
558 marine phytoplankton highlights modulating role of pCO2. *Global Change Biology*, 26, pp.
559 6787– 6804.

560 Stephens, A.E., Srivastava, D.S. and Myers, J.H., 2013. Strength in numbers? Effects of multiple
561 natural enemy species on plant performance. *Proceedings of the Royal Society B:*
562 *Biological Sciences*, 280(1760), p.20122756.

563 Tang, J., Wennerberg, K. and Aittokallio, T., 2015. What is synergy? The Saariselkä agreement
564 revisited. *Frontiers in Pharmacology*, 6, p.181.

565 Tekin, E., Diamant, E.S., Cruz-Loya, M., Enriquez, V., Singh, N., Savage, V.M. and Yeh, P.J., 2020.
566 Using a newly introduced framework to measure ecological stressor interactions. *Ecology*
567 *Letters*, 23, pp1391-1403.

568 Tittensor, D.P., Walpole, M., Hill, S.L., Boyce, D.G., Britten, G.L., Burgess, N.D., Butchart, S.H.,
569 Leadley, P.W., Regan, E.C., Alkemade, R. and Baumung, R., 2014. A mid-term analysis of
570 progress toward international biodiversity targets. *Science*, 346(6206), pp.241-244.

571 Thompson, P.L., MacLennan, M.M. and Vinebrooke, R.D., 2018. An improved null model for
572 assessing the net effects of multiple stressors on communities. *Global Change Biology*,
573 24(1), pp.517-525.

574 Vinebrooke, R.D., Cottingham, K.L., Norberg, M., Scheffer, J., Dodson, S.I., Maberly, S.C. and
575 Sommer, U., 2004. Impacts of multiple stressors on biodiversity and ecosystem
576 functioning: The role of species co-tolerance. *Oikos*, 104(3), pp.451-457.

577 Wolfram Research. Inc., 2016. Mathematica (Version 10.4). *Wolfram Research, Inc., Champaign,*
578 *Illinois*.

579

580 **Acknowledgments**

581 This research was supported by the United Kingdom Natural Environment Research Council
582 studentship grant NE/M010481/1. The authors declare no conflicts of interest.

583 **Author contributions**

584 BB and DM conceived the ideas and designed methodology; BB generated the data; BB and DM
585 analysed the data; BB and DM led the writing of the manuscript. All authors contributed critically to
586 the drafts and gave final approval for publication.

Data availability

All data used in this work is publicly available via the supporting information of Burgess et al. (2021) https://doi.org/10.1111/gcb.15630**.** Code for all results will be made available on Github.