Low statistical power and overestimated anthropogenic impacts, exacerbated by publication bias, dominate field studies in global change biology

Yefeng Yang[1,2*] | Helmut Hillebrand[3,4,5] | Malgorzata Lagisz[1] | Ian Cleasby[6] | Shinichi Nakagawa[1*]

[1] Evolution & Ecology Research Centre and School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, NSW 2052, Australia

[2] Department of Infectious Diseases and Public Health, Jockey Club College of Veterinary Medicine and Life Sciences, City University of Hong Kong, Hong Kong, China

[3] Institute for Chemistry and Biology of Marine Environments (ICBM), Plankton Ecology Lab, Carl-von-Ossietzky University Oldenburg, Schleusenstrasse 1, 26382 Wilhelmshaven, Germany

[4] Helmholtz-Institute for Functional Marine Biodiversity at the University of Oldenburg (HIFMB), Ammerländer Heerstrasse 231, 26129 Oldenburg, Germany

[5] Alfred Wegener Institute, Helmholtz-Centre for Polar and Marine Research (AWI), Bremerhaven, German

[6] RSPB Centre for Conservation Science, North Scotland Regional Office, Inverness, UK

* Correspondence:

Y. Yang, e-mail: yefeng.yang1@unsw.edu.au

S. Nakagawa, e-mail: s.nakagawa@unsw.edu.au

Running title: Low power and bias widespread global change studies

Abstract

Field studies are essential to reliably quantify ecological responses to global change because they are exposed to realistic climate manipulations. Yet such studies are limited in replicates, resulting in less power and, therefore, unreliable effect estimates. Further, while manipulative field experiments are assumed to be more powerful than non-manipulative observations, it has rarely been scrutinised using extensive data. Here, using 3,847 field experiments that were designed to estimate the effect of environmental stressors on ecosystems, we systematically quantified their statistical power and magnitude (Type M) and sign (Type S) errors. Our investigations focused upon the reliability of field experiments to assess the effect of stressors on both ecosystem's response magnitude and variability. When controlling for publication bias, single experiments were underpowered to detect response magnitude (median power: 18% – 38% depending on mean difference metrics). Single experiments also had much lower power to detect response variability (6% – 12% depending on variance difference metrics) than response magnitude. Such underpowered studies could exaggerate estimates of response magnitude by 2 – 3 times (Type M errors) and variability by 4 – 10 times. Type S errors were comparatively rare. These observations indicate that low power, coupled with publication bias, inflates the estimates of anthropogenic impacts. Importantly, we found that meta-analyses largely mitigated the issues of low power and exaggerated effect size estimates. Rather surprisingly, manipulative experiments and non-manipulative observations had very similar results in terms of their power, Type M and S errors. This suggests that the previous assumption about the superiority of manipulative experiments is overstated. These results call for highly powered field studies to reliably inform theory building and policymaking, via more collaboration and team science, and large-scale ecosystem facilities. Future studies also require transparent reporting and open science practices to approach reproducible and reliable empirical work and evidence synthesis.

KEYWORDS

Climate change, experimentation, exaggerated effect size, reproducibility, transparency, small-effect effect, selective reporting bias, meta-science, meta-research, second-order meta-analysis

1 | INTRODUCTION

As human-induced environmental changes accelerate, it is more important than ever that we can reliably quantify ecological responses to a range of environmental stressors (Hanson & Walker, 2020; Sage, 2020; Way, 2021). Although laboratory experiments could elucidate the underlying mechanisms of such ecological responses, they are often too small, too short-lived, and too artificial to reflect naturally occurring responses accurately (Rineau et al., 2019). Therefore, field experiments (both manipulations and non-manipulative observations) are essential to understand how an ecosystem responds to global change (Elmendorf et al., 2015; Sternberg & Yakir, 2015; Wolkovich et al., 2012). In particular, field experimental manipulations are paramount because they could quantify the effect of stressor magnitudes that go well beyond currently observed levels (Hillebrand et al., 2020; Rineau et al., 2019). Accordingly, thousands of field experiments have been conducted in the field to investigate ecological responses to a wide range of different anthropologic environmental impacts such as climate change, biodiversity loss, and agricultural intensification (Hanson & Walker, 2020; Scheffer, Carpenter, Foley, Folke, & Walker, 2001). Yet, few researchers seem to have asked whether these thousands of global change experiments could provide statistically reliable results to advance our understanding of ecosystems of the future (Korell, Auge, Chase, Harpole, & Knight, 2020). While field experiments offer the possibility to work with realistic abundances and naturally-occurring environmental conditions (and their variation), their replications often are limited by logistical constraints (Fraser, Barnett, Parker, & Fidler, 2020; Nakagawa & Parker, 2015). Therefore, it is essential to know whether these field experiments are adequately powered and reliable.

Earlier work suggests that ecological studies seem to be underpowered in some subfields (Fidler et al., 2017; Jennions & Møller, 2003; T. H. Parker et al., 2016). That is, a study usually has a sample size too small to detect a 'true' effect size as statistically significant (for a given alpha level 0.05). An important yet often underappreciated consequence of underpowered studies is that empirical studies with small sample sizes often present distorted estimates of true effects (Button et al., 2013; Nakagawa & Foster, 2004). This is because, given an underpowered study, the observed effect often fails to achieve statistical significance (i.e., two-tailed p-value < 0.05), unless the effect is overestimated. In other words, when an observed effect reaches statistical significance in an underpowered or small-sample study, the observed effect will be always higher than the corresponding 'true' effect in magnitude (Lemoine et al., 2016; Young, Ioannidis, & Al-Ubaydli, 2008; also see a simulated example in Figure S1). Then, due to preferential publications of statistically significant effects (i.e., publication bias), such overestimated effects would dominate the literature. The inflation of magnitude concerning a 'true' effect is known as exaggeration ratio or Type M (magnitude) error. A related concept is the Type S (sign) error that is the probability of obtaining a statistically significant effect in the opposite direction to the true effect (Gelman & Carlin, 2014).

Recently, a few papers have pointed out the importance of quantifying the Type M and S error rates (Cleasby et al., 2021; Lemoine et al., 2016; T. H. Parker et al., 2018). For example, Lemoine et al. (2016) showed that reported effect sizes of warming on plant growth were, on average, three times larger than a 'true' effect that was approximated by an overall meta-analytic mean (Type M error rate: 3). In animal tracking studies, Cleasby et al. (2021) demonstrated that researchers could be overestimating the effect of bio-logging devices on animal behaviour by 10-fold (Type M error rate) and estimating the direction of the effect incorrectly 20% of the time (Type S error rate), using effect sizes derived from a previous meta-analysis (Cohen's d = 0.1; Bodey et al., 2018). Given these, both studies argued that understanding Type M (and S) error rates, along with statistical power, would lead to better interpretation of results and improve the experimental design in a field of study (cf. Button et al., 2013; Ioannidis, Stanley, & Doucouliagos, 2017; T. Stanley, Carter, & Doucouliagos, 2018).

However, no previous publications have *systematically* quantified statistical power, Type M and S error rates across global change studies (but see Lemoine et al., 2016). Importantly, although earlier work often used meta-analytic means as a surrogate of the true effect to quantify the statistical power and error rates (e.g., Cleasby et al., 2021; Lemoine et al., 2016), large-scale power analyses from other fields have shown that meta-analytic means often suffer from publication bias (Button et al., 2013; Ioannidis et al., 2017; T. Stanley et al., 2018). This can lead to an overestimation of statistical power unless the bias is corrected (Button et al., 2013; Ioannidis et al., 2017; T. Stanley et al., 2018). Further, environmental stressors are likely to influence not only ecological responses in magnitude (mean value) but also the variance of ecological responses (i.e., heteroscedasticity; Figure 1A; for examples of biological explanations of heteroscedasticity see Cleasby & Nakagawa, 2011; De Villemereuil, Morrissey, Nakagawa, & Schielzeth, 2018; Seekell, Carpenter, & Pace, 2011). Therefore, it is important to quantify these three statistical parameters not only for response magnitude but also for response variability. As far as we know, no such investigations for response variability exist in the entire scientific literature so far.

To this end, we conduct the first large-scale quantification of statistical power, Type M and S error rates, using manipulative field experiments and non-manipulative observations covering the dominant stressors in global change biology (cf. Sage, 2020). More specifically, we quantify these three parameters at two different levels, a single experiment, and meta-analysis (e.g., the statistical power of a field experiment *vs*. meta-analysis), for ecological response magnitude and variability (i.e., mean and variance differences between an environmental stressor and a benign or control environment). In addition, we estimate true effects with and without correcting for publication bias because, as mentioned, failing to correct for publication bias can lead to the overestimation of statistical power, and also of type M and S errors. We hypothesize that global change studies are generally underpowered with high exaggeration ratios, although Type S error rates are relatively low. We also predict that manipulative field experiments will have greater statistical power and lower type M and S errors than non-manipulative field observations because manipulative experiments would often involve stressor levels beyond currently observed levels so that ecological responses (i.e., effect size) should be higher both in magnitude and variation (Hillebrand et al., 2020; Kreyling & Beier, 2013).

2 | MATERIALS AND METHODS

2.1 | An overview of the methodology

To address our main aims above, we chose to use a database of global change biology, containing 30 meta-analyses (3,847 field experiments/observations) over a multitude of environmental stressors (see Section 2.2 below; Hillebrand et al., 2020). Using this database, we calculated five standardised effect-size statistics to quantify response magnitude (mean difference) and variability (variance difference) to environmental stressors in global change studies. For response magnitude, we used (1) the natural logarithm of response ratio, (lnRR; Hedges, Gurevitch, & Curtis, 1999), (2) standardised mean difference, SMD (also known as Hedges' *g* or Cohen's *d*; Hedges, 1982), and (3) standardized mean difference with heteroscedastic population variances in the two groups, SMDH (see formulas in Table 1). Note that SMD assumes homoscedasticity (i.e., equal variances; Hedges, 1982) whereas SMDH allows for heteroscedasticity (Bonett, 2008, 2009). Also, heteroscedasticity only affects the sampling variance of lnRR, not the point estimate (Sánchez-Tójar, Moran, O'Dea, Reinhold, & Nakagawa, 2020). For quantifying response variability, we used (4) the natural logarithm of variability ratio, lnVR (Nakagawa et al., 2015), and (5) the natural logarithm of the coefficients of variation, lnCVR (Nakagawa et al., 2015) which adjusts for changes in mean values (see formulas in Table 1).

We used a three-step modelling procedure to test our main hypotheses (Figure 1C). In the first step, we used a meta-analytic approach to obtain the key quantity for power calculations – an estimate of the 'true' effect size of a phenomenon (Nakagawa & Foster, 2004). To achieve this, we employed the meta-analytic (overall) mean, rather than the 'observed' effect size from a given study, as a proxy of true effect to avoid overestimating statistical power (for examples using this approach, see Button et al., 2013; Cleasby et al., 2021). Therefore, we meta-analysed five effect-size statistics (Table 1) separately to obtain meta-analytic means for each meta-analytic case (Section 2.3). For lnRR, SMD and SMDH, we also estimated bias-corrected versions of corresponding effect sizes to adjust for publication bias (also known as the small-study effect; Vevea & Hedges, 1995) (Section 2.4). Contrastingly, we cannot calculate bias-corrected lnVR and lnCVR because statistical significance, rather than response variability (heteroscedasticity or variance difference), drives publication bias (see Senior, Gosby, Lu, Simpson, & Raubenheimer, 2016). Therefore, we assumed that lnVR and lnCVR were not affected by publication bias in the way lnRR, SMD, and SMDH were.

In the second step, we calculated the statistical power to detect the estimates of true effects and their magnitude (Type M) and sign (Type S) error rates, for each meta-analysis and every single experiment included in the meta-analysis (Section 2.5.1). In the third step, to obtain overall estimates of the three parameters across different meta-analyses (which provided us with comparable summaries of the three parameters), we used a weighted regression to statistically aggregate over the three parameters obtained at the meta-analysis level, whereas we used a mixed-effects model to aggregate these parameters at the experiment level. Both procedures involved aggregating the parameters across meta-analyses (i.e., between-meta-analysis modelling; Section 2.5.2). We also conducted a secondary synthesis of the true effects (which were estimated from the first step) across meta-analyses (that is, conducting a meta-analysis of overall means obtained from the included 30 meta-analyses; also referred to as a second-order meta-analysis or meta-meta-analysis; cf. Nakagawa et al., 2019) (Section 2.6). We conducted all analyses in the R environment v. 4.0.3 (R Development Team, 2020). All relevant data and code can be found at https://zenodo.org/record/5496789#.YTmbiI4zY2w.

2.2 | Global change meta-analyses database

Our global change meta-analyses database reflected a range of the responses of ecosystem processes to the most pervasive anthropogenic global change stressors, including climate warming, fire eutrophication, and nitrogen fertilization (Hillebrand et al., 2020). The database was originally used to quantify how evident thresholds, tipping points, or regime shifts were in ecological responses to anthropogenic global change (at https://zenodo.org/record/5496789#.YTmbiI4zY2w). The dataset did not contain laboratory experiments and only included experimental/manipulative field experiments and non-manipulative observations. It followed strict inclusion and exclusion criteria (as depicted in Hillebrand et al. 2020) and finally contained 36 meta-analyses (providing 4,601 effect sizes in the form of lnRR).

We excluded 6 meta-analyses from the original database because they did not provide sampling variance (S^2_{lnRR}; Table 1), which was required for formal weighted meta-analyses and calculations of statistical power and Type M and S errors. Thus, our final database contained 30 meta-analyses (Figure 1B), which provided 3,850 estimates of lnRR paired with a corresponding estimate of sampling variance (S^2_{lnRR}). For these 30 meta-analyses in the form of lnRR (referred to as dataset lnRR*), the number of studies (*N*) included in meta-analysis ranged from 11 to 186 (mean = 37.3, median = 26.5, SD = 37.1). The number of effect sizes (*k*) of lnRR* ranged from 35 to 562 (mean = 128.2, 85.0 = 26.5, SD = 121). In addition, within dataset lnRR*, 12 out of 30 meta-analysis provided descriptive statistics in included primary studies: mean (m_p or m_c), standard deviation (sd_p^2 or sd_c^2), and sample size (n_p or n_c), which enabled us to calculate SMD, SMDH, lnVR and lnCVR and their sampling errors for these 12 meta-analyses. We also re-calculated lnRR (to distinguish with lnRR*, we referred it to as dataset lnRR) using these 12 meta-analyses so as to compare the statistical power, Type M and S errors for lnRR, SMD, SMDH, lnVR and lnCVR (section 2.5). For the 12 meta-analyses (effect size in the form of lnRR, SMD, SMDH, lnVR, and lnCVR), *N* ranged from 11 to 186 (mean = 42.8, median = 19, SD = 58.2), *k* ranged from 44 to 450 (mean = 164.8, median = 119.5, SD = 119.2). The replicates (*n*; sample size per study) in each study of the 12 datasets ranged from 4 to 10000 (mean = 38.4, median = 12, SD = 83.0).

Of the 30 meta-analyses, 11 meta-analyses used non-manipulative observations and 17 used manipulative experiments, while 2 used both non-manipulative observations and manipulative experiments. We followed the original database in defining the categories of environmental stressors; namely, acidification (Acid, *k* = 62; Nagelkerken & Connell, 2015), biodiversity loss (BD loss, *k* = 942; Cardinale et al., 2006; Griffin, Byrnes, & Cardinale, 2013; Östman et al., 2016), fertilization (Fert, *k* = 811; Akiyama, Yan, & Yagi, 2010; Elser et al., 2007; Liang, Qi, Souza, & Luo, 2016; Treseder, 2008), bush fire (Fire, *k* = 179; Dijkstra & Adams, 2015; Dooley & Treseder, 2012), plant invasion (Inv, *k* = 316; Gaertner et al., 2014; Gallardo, Clavero, Sánchez, & Vilà, 2016; Vilà et al., 2011), land use change (LUC, *k* = 612; Gibson et al., 2011; Van Lent, Hergoualc'h, & Verchot, 2014), precipitation (Precip, *k* = 138; Liu et al., 2016), global warming (Warm, *k* = 790; Ateweberhan & McClanahan, 2010; Lin, Xia, & Wan, 2010; Lu et al., 2013).

2.3 | Meta-analyses and estimating the proxies of 'true' effects

As the first step of our three-step modelling procedure, we estimated various proxies of 'true' effects for each meta-analysis. The proxies of 'true' effects included (1) meta-analytic overall means (MAOMs), which represented a common 'true' effect shared by the multiple experiments within a given meta-analysis (section 2.3.1), (2) effect-size-specific predictions (ESSPs), which represented experiment-dependent effects (i.e., multiple true effects within a given meta-analysis; section 2.3.2), and (3) the publication-bias-corrected versions of MAOMs and ESSPs (section 2.4)

2.3.1 | Meta-analytic overall means (MAOMs)

To estimate 'true' effects for each meta-analysis, we employed a multilevel model to estimate meta-analytic overall means (referred to as MAOMs, hereafter; Nakagawa & Santos, 2012), in which the non-independence in the datasets (i.e., multiple effect sizes per study) were accounted for by incorporating effect-size and study identities as random factors (Noble, Lagisz, O'dea, & Nakagawa, 2017). We used the *rma.mv* function in the *metafor* package (Viechtbauer, 2010) to run the following multilevel meta-analytic model for lnRR*, lnRR, SMD, SMDH, lnVR, or lnCVR, respectively (Nakagawa & Santos, 2012):

$$ES_{ji} = \beta_0 + s_j + w_{ji} + e_{ji}, \qquad (9)$$

where $s_j \sim \mathcal{N}(0, \tau^2), w_{ji} \sim \mathcal{N}(0, \sigma^2), e_{ji} \sim \mathcal{N}(0, v_i)$ with $\mathcal{N}$ being a normal distribution with two parameters, mean and variance. Here ES_{ji} is the observed effect size estimates (i.e.

lnRR, SMD, SMDH, lnVR, or lnCVR), β_0 is the intercept (i.e. meta-analytic overall mean, MAOM), and s_j is the between-study effect for the study j, w_{ji} is the within-study effect for the effect size *i* in the study *j*, e_{ji} is the sampling error for the effect size *i* in the study *j*, τ^2, σ^2 and v_i are associated variance components.

2.3.2 | Effect-size-specific predictions (ESSPs)

Given the high heterogeneities in ecological datasets ($I^2 > 90\%$; Senior, Grueber, et al., 2016), there rarely exists a common effect size between different studies within a meta-analysis. For example, nutrient enrichment has a large effect on plant biomass, whereas lack of light stimuli will largely reduce this effect. Therefore, we used an alternative proxy of true effect to accommodate such an experiment-dependent effect (i.e., multiple true effects within a given meta-analysis): effect-size-specific prediction (ESSP; see Equation 10). ESSPs can be estimated by using the best linear unbiased predictions (BLUPs) in the observation level, which are defined as (conditional) point estimates given a set of random effects in a mixed effect model (Hadfield, Wilson, Garant, Sheldon, & Kruuk, 2010). We defined ESSPs as follows:

$$ES_{ji(ESSP)} = \overline{\beta_0} + \overline{s_j} + \overline{w_{ji}}, \qquad (10)$$

where the notations are the same as Equation 9 (note that β_0, $\overline{s_j}$, and $\overline{w_{ji}}$ are the estimated parameters from Equation 9). Equation 10 shows that ESSPs is the sum of the overall mean (MAOM), the between-study effect s_j, the within-study (effect-size-specific) effect w_{ji}. ESSPs were obtained using the *rma.mv* function in *metafor* (Viechtbauer, 2010).

2.4 | Obtaining bias-corrected meta-analytic estimates

For response magnitude (i.e., lnRR, SDM and SMDH), publication bias can translate into overestimated meta-analytic means, MAOMs (Vevea & Hedges, 1995). To alleviate such a bias, we employed an extended version of Egger's regression approach (multilevel meta-regression, cf. Nakagawa, Lagisz, Jennions, et al., 2021) which resulted in a bias-corrected version of MAOMs. In brief, this approach incorporates uncertainty term as a moderator in a multilevel meta-regression model: the inverse of 'effective sample size' $1/\tilde{n}_i = (n_p + n_c)/n_p n_c$ or its square root $\sqrt{1/\tilde{n}_i}$ (strictly speaking, 'effective sample size' = $4\tilde{n}_i$).

$$ES_{ji} = \beta_0 + \beta_1\sqrt{1/\tilde{n}_i} + s_j + w_{ji} + e_{ji}, \qquad (11)$$

$$ES_{ji} = \beta_0 + \beta_1(1/\tilde{n}_i) + s_j + w_{ji} + e_{ji}, \qquad (12)$$

β_0 is the (conditional) bias-corrected meta-analytic overall mean (cMAOM, hereafter) when assuming no uncertainty exists: $\sqrt{1/\tilde{n}_i} = 0$ in Equation 11 or $1/\tilde{n}_i = 0$ in Equation 12. If β_1 in Equation 11 is statistically non-significant (*p*-value > 0.05), β_0 in Equation 11 (the slope of $1/\tilde{n}_i$) is the best estimate of cMAOM. If β_1 in Equation 11 (the slope of $\sqrt{1/\tilde{n}_i}$) is statistically significant (*p*-value < 0.05), β_0 in Equation 12 is the best estimate of cMAOM (Tom D Stanley & Doucouliagos, 2014; Tom D. Stanley, Doucouliagos, & Ioannidis, 2017).

We note that the slope (β_1) of Equation 11 could be in the opposite direction from what was expected from publication bias (Figure S2); in such a case, we considered the dataset did not suffer from the publication bias and we used MAOMs as their cMAOMs. 18 meta-analyses within lnRR* dataset did not report replicates (*n*; sample size per study) for calculation of 'effective sample size'; we used sampling error (se_i, the square-root of the sampling variance) and sampling variance (v_i) to replace $1/\tilde{n}_i$ in Equation 11 and $\sqrt{1/\tilde{n}_i}$ in Equation 12, respectively. When calculating statistical power, Type M and S error rates, we used unconditional standard error (SE) rather than a conditional standard error (*viz*, using standard error for β_0 in Equation 9 to replace that of Equation 11 or 12). The models in Equations 11 and 12 were implemented by the *rma.mv* function in *metafor*. Further, with cMAOMs, we used Equation 10 to obtain 'bias-corrected effect-size-specific predictions (cESSPs). In our datasets, lnRR*, lnRR, SMD, and SMDH had 20 of 30, 6 of 12, 5 of 12, and 5 of 12 meta-analyses, respectively, which did not show the statistical evidence of the small-study effect (Figure S3).

2.5 | Estimating statistical power, Type M and S error rates

2.5.1 | (Within-)meta-analysis level modelling

We calculated statistical power, Type M and S errors at two levels: the meta-analysis level (i.e., three parameters for each of the meta-analysis identified), and single experiment level (i.e., three parameters for experiments or effect sizes within a given meta-analysis; Figure 1C). We expected that statistical power at the meta-analysis level would be much higher than that at the single experiment level, although it would still be possible that a meta-analysis might not have enough statistical power to detect the estimated overall effect (i.e., non-significant overall effect; Cohn & Becker, 2003). In addition to the proxies of 'true' effects (i.e., MAOMs, ESSPs, cMAOMs, cESSPs), we required standard error (SE) for each effect size estimate to calculate statistical power, Type M and S errors. For the meta-analysis level, we used SEs from the meta-analytic models (i.e., Equations 9, 11, or 12). For the single experiment level, we used the square root of the sampling variance for each effect size (see Table 1) as SEs.

2.5.2 | Between-meta-analysis modelling

Importantly, we also obtained an overall (average) statistical power, Type M and S errors for each effect size statistic across different meta-analyses (i.e., between-meta-analyses estimates; Figure 1C). Such overall estimates provided us with comparable summaries of statistical power, Type M and S errors. For the meta-analysis level, we used a weighted regression, implemented with the *base R* function, *lm*, with the number of effect sizes (k) for each meta-analysis as weight. The weighted regression models allowed us to average over the estimates of meta-analysis level power and Type M and S errors (using MAOMs and cMAOMs). For the single experiment level, we used mixed-effects models employing the *lmer* function in the *R* package, *lme4* (Bates, Mächler, Bolker, & Walker, 2014), with study identities as a random factor. These mixed-effects models allowed us to average over the single-experiment level estimates (using MAOMs, cMAOMs, ESSPs, and cESSPs). Further, to these mixed-effects models, we added study approach (manipulative experiment *versus* non-manipulative observation) as a fixed factor, and stressor categories as a random factor to compare the average statistical power, Type M and S errors between manipulative experiments and non-manipulative observations.

Before constructing the above models using *lm* and *lmer*, we ln-transformed the response variables (estimates of statistical power, Type M and S error rates) to better meet the 'normal residuals' assumption (Figure S4 – S6). For easy interpretation, we back-transformed (i.e. exponentiated) the intercept of *lm* and *lmer* models so that we obtained the median value on the original scale (Nakagawa, Johnson, & Schielzeth, 2017). We also obtained the mean value on the original scale (using Equation 5.8; Nakagawa et al., 2017). Further, for the Type S error rate, we added 0.025 to all the cases because the estimates of Type S error included many zeros and extremely small values, which made ln-transformation impossible or ineffective. Note that when we back-transformed estimates from these models, we adjusted these estimates on the original scale by subtracting a value of 0.025. Further, when back-transformed estimates (statistical power and Type S error) went below or above the boundary values (i.e., 0 or 1, respectively), we constrained the estimates to the boundaries.

2.6 | Response magnitude and variability across environmental stressors

To estimate the overall response magnitude and variability across meta-analyses (i.e., between-meta-analysis synthesis), we conducted a secondary synthesis of the estimates of response magnitude and variability from each meta-analysis. Of note, one meta-analysis represented one specific stressor (e.g., a meta-analysis of acidification, a meta-analysis of global warming; see section 2.2). We also assessed whether there were significant differences in such overall effects between manipulative experiments and non-manipulative observations. To achieve this, first, we obtained the absolute values of (c)MAOMs and their sampling variances (i.e., the variance estimated from a folded normal distribution; see Morrissey, 2016) for each meta-analysis (that is, across stressors). Second, we statistically aggregated these absolute estimates (|MAOM| and |cMAOM|) via a random-effect model using *rma* function in the *R* package *metafor* (Viechtbauer, 2010). Third, we conducted meta-regression with the study approach as a moderator to quantify effects for manipulative experiments and non-manipulative observations (we excluded two meta-analyses that contained both experimental and observational data; see Section 2.2).

3 | RESULTS

3.1 | The effects of stressors on ecosystem response magnitude and variability

Overall, environmental stressors had a statistically significant impact on response magnitude (more than a 33.7% increase; Figure 2A). For the result of each stressor, see Figure S7 – S9 (each meta-analysis was focussed upon a specific stressor, but a given stressor may be covered by multiple different meta-analyses, e.g., Warm 1, Warm 2, Warm 3 were three meta-analyses all concerned with global warming). Bias-corrected estimates of response magnitude declined by 17% to 31% (Figure 2B). Similarly, stressors had a statistically significant effect on response variability (more than a 20% increase; Figure 2C; shown by a stressor in Figure 10).

Further, manipulative experiments had a statistically significant larger response magnitude than that non-manipulative observations for some effect size types (i.e., uncorrected SMD, uncorrected SMDH, corrected SMDH; Table S1). In contrast, the differences in response variability between manipulative experiments and non-manipulative observations were not statistically significant.

3.2 | Statistical power in global change studies

3.2.1 | Statistical power in detecting response magnitude

Across all stressors, single experiments had much lower power to detect bias-corrected response magnitude compared to the nominal 80% power (Table 3): 23.3% for lnRR* (Figure 3A), 38.5% for lnRR (Figure 3A), 19.1% for SMD (Figure 3B), 18.2% for SMDH (Figure 3D). When considering that each experiment has its own true effect (cESSP), the power values were similar to the values estimated from a common true effect (cMAOM; Table 3 and Figure 3). The corresponding power values for uncorrected response magnitude were 19% to 66% higher than that of the bias-corrected version (Table 3 and Figure 3). The median proportion of single experiments that had adequate power to detect bias-corrected lnRR*, lnRR, SMD and SMDH were only 16.3, 33.2, 6.6, and 6.9%, respectively (Figure 3). As expected, the median power for meta-analysis to detect bias-corrected response magnitude was greater than that of single experiments although it fell short of the nominal 80% level: 42.4% – 63.5% (depending on effect-size types; Table 3 and Figure 3). As at the single experiment level, uncorrected meta-analyses overestimated power by ~2% to 33% compared to the bias-corrected version (Table 3 and Figure 3).

3.2.2 | Statistical power in detecting response variability

Overall, at the single experiment level, lnVR and lnCVR showed comparatively low statistical power to detect heteroscedasticity than the nominal 80% level: 11.5% for lnVR and 6.4% for lnCVR (Table 3 and Figure 3C and 3E). The median proportion of experimental lnVR and lnCVR that had adequate power to detect response variability was only 3.7 and 0%, respectively (Figure 3). Meta-analysis increased the overall power to identify response variability roughly by 4 to 6-fold: power was now 43.9% for lnVR and 52.6% for lnCVR (Table 3 and Figure 3). The proportion of single experiments that had adequate power increased to 33.3% and 16.7% when using meta-analysis to detect lnVR and lnCVR, respectively (Figure 4).

3.3 | Type M and S error rates in global change studies

3.3.1 | Type M and S error rates in detecting response magnitude

Single experiments tended to overestimate the effect of the environmental stressors consistently (Type M error rates; Table 4 and Figure 4). Depending on which effect metric was used, single experiments were on average 2 – 3-fold larger than the true effect size estimated as MAOMs. Single experiments rarely had the wrong estimation of the sign of the true effect size (Type S error rate; Table 5 and Figure 5). As expected, meta-analyses largely reduced the magnitude of Type M (1 – 2; see Table 4 and Figure 4). When bias correction was not employed, the overestimation of the true effect was even larger (Type M error rates by 2 – 6 and S error rates by 10% – 30%).

3.3.2 | Type M and S error rates in variance differences

At the single experiment level, lnVR and lnCVR on average showed large Type M error rates (~4 and 10, respectively; Table 4 and Figure 4), but low Type S error rates (5% – 19.9%; Table

5 and Figure 5). By contrast, meta-analyses only overestimated lnVR and lnCVR by 1.6-fold and 1.5-fold, respectively.

3.4 | Contrasting manipulative experiments and non-manipulative observations

Both single manipulative experiments and non-manipulative observations were underpowered to detect the effects of environmental stressors on ecosystem response magnitude and variability (16% – 39% depending on effect metrics; Figure 6A – 6F). With one exception, the differences in power between manipulative experiments and non-manipulative observations were not statistically significant (Figure 6D). When bias correction of ESSPs were employed, manipulative experiments had statistically greater power than non-manipulative observations (32% *vs.* 20%). Similarly, differences between manipulative experiments and non-manipulative observations were not significant in terms of their Type M (with one exception: bias-corrected lnRR*; Figure 6G – 6L). Manipulative experiments had statistically larger Type M error than non-manipulative observations if bias correction of ESSPs were used (2-fold *vs.* 6-fold). A similar pattern was found for Type S errors in manipulative experiments and non-manipulative observations (Figure 6M – 6R).

4 | DISCUSSION

We have conducted the first study to systematically assess the power, type M and type S error rates for global change studies. Concurring with our hypotheses, *global change studies are generally underpowered, resulting in high Type M error rates (overestimating the magnitude of the response) whereas Type S error rates (wrong estimation of sign) are relatively low*. Across different ecosystems and stressors, single experiments were underpowered to detect bias-corrected response magnitude (~18 – 38 % depending on effect-size types; Table 3 and Figure 3). Similarly, single experiments also had a much lower power to detect response variability (heteroscedasticity) than response magnitude (~6 – 12%; Table 3 and Figure 3). Such underpowered field experiments could exaggerate an effect by 2 – 3 times for response magnitude (with bias-correction) and by 4 – 10 times for response variability (Table 4 and Figure 4). Also, single experiments rarely incorrectly estimated the direction of the true anthropogenic impact (Table 5 and Figure 5). Notably, our results were consistent regardless of assuming one 'true' effect per meta-analysis (e.g., cMAOM) or experiment-specific 'true' effects within a meta-analysis (cESSP). In contrast to our expectation, apart from one exception, manipulative field experiments and non-manipulative observations were not statistically different in terms of their statistical power or Type M / S errors. Taken together, we conclude that the low statistical power, coupled with publication bias, may have led to distorted estimates of anthropogenic impacts in the literature. Below, we first extend our discussion on the comparisons between manipulative experiments and non-manipulative observations. Then, we consider three statistical (but biologically relevant) points that emerged from our results and how they can improve future empirical studies (manipulative experiments and non-manipulative observations) and meta-analyses in global change biology in general.

4.1 | Manipulative experiments and non-manipulative observations both lack power

Rather surprisingly, the statistical power of manipulative experiments and non-manipulative observations was similar (e.g., uncorrected SMD and bias-corrected SMD in Table S1). The differences between manipulative experiments and non-manipulative observations have been often assumed because experimental work usually has greater effect magnitude (Palmer, 2000). Yet, as far as we are aware, no work has identified whether such differences empirically occur. The lack of power differences between manipulative experiments and non-manipulative observations may be due to the trade-off between the magnitude of effect sizes and the number of replicates (i.e., sample size). That is, higher experimental effect sizes are offset by smaller sample sizes in manipulative experiments than non-manipulative observations. Indeed, we found that manipulative experiments had larger effects than non-manipulative observations. For example, manipulative experiments had statistically larger estimates of SMD than non-manipulative observations (see Table S1). Contrastingly, non-manipulative observations had 2.5-fold larger replicates (sample sizes), on average, than manipulative experiments (25 *versus* 10; Figure S11 – S12). Although we may tend to think manipulative experiments have greater power and are therefore more reliable, this assumption is not tenable, at least in the field of global change studies.

4.2 | Meta-analysis is not only a powerful tool but maybe the only tool?

As expected, meta-analyses have increased the power to detect response magnitude (both before and after correcting for publication bias) by at least 30% compared to single experiments. For example, the overall power for meta-analyses were 51.2% and 62.1% for lnRR and SMD, respectively, compared to 38.5% and 19.1% for single experiments (Table 3). Indeed, the nominal 80% power is difficult to achieve in many disciplines in a single experiment level, such as Neuroscience (median power = 21%; Button et al., 2013), Clinical medicine (median power = 20%; Lamberink et al., 2018), Psychology (median power = 36%; T. Stanley et al., 2018) and Economics (median power = 18%; Ioannidis et al., 2017). Such low statistical power averages for single experiments highlight the importance of meta-analysing response magnitude (Gurevitch, Koricheva, Nakagawa, & Stewart, 2018). We note that, although single experiments are often underpowered and more prone to type M error, they are essential to global change biology research. Such experiments contribute to evidence accumulation, providing raw materials for systematic reviews and meta-analyses. Perhaps, more importantly, local field experiments are an effective way to reveal the casual mechanisms of ecological responses at a particular ecosystem, and idiosyncrasies among ecosystems from different localities (Rineau et al., 2019; Roy et al., 2021).

Similarly, meta-analysis of variance (i.e., synthesizing lnVR and lnCVR from individual studies; Nakagawa et al., 2015) is a powerful approach to detect response variability (i.e., heteroscedasticity). Indeed, we found meta-analysis of variance increased the statistical power by 4 – 6-fold (meta-analytic lnVR *vs.* individual lnVR: 43.9% *vs.* 11.5%, meta-analytic lnCVR *vs.* individual lnCVR: 52.6% *vs.* 6.4%; Table 3). Further, meta-analysis of variance could mitigate Type M and S error rates compared to single experiments. Ecologists have been aware of difficulties in detecting response variability reliably (Andersen, Carstensen, Hernandez-Garcia, & Duarte, 2009; Carpenter & Brock, 2006; Seekell et al., 2011), and have already discussed the need for a large sample size (Engle, 1982; Seekell et al., 2011). Yet, the number of replicates (*n*; sample size per study) in global change studies was usually too small to detect response variability reliably (medium *n* = 12 in our dataset). Practically speaking, to get an adequate sample size for estimating effects on response variability, we need to organise more global research collaboration network, such as Nutrient Network (NutNet; Harpole et al., 2016; Lekberg et al., 2021), US Long-Term Ecological Research network (LTER; Crossley et al., 2020), and Zostera Experimental Network (ZEN; Wu et al., 2017). Alternatively, we would require heavily instrumented and controlled environmental facilities (e.g., UHasselt Ecotron, see Rineau et al., 2019, Clobert et al., 2018; Roy et al., 2021). Fortunately, meta-analysis of variance provides us an alternative approach for increasing the chance of detecting changing response variability hidden in global change studies.

4.3 | Publication bias may have exacerbated the inflation of anthropologic effects

We have shown that meta-analyses result in a sizeable increase in power over single experiments, although some meta-analyses were generally underpowered relative to a nominal value of 80% power (Table 3 and Figure 3). Furthermore, only half of the meta-analyses (15 of 30) had tested for the existence of publication bias in their datasets. Furthermore, only half of the meta-analyses (15 of 30) had tested for the existence of publication bias in their datasets. The methods used to assess publication bias were: funnel plots (n = 8), rank correlation tests (n = 4), fail-safe N (n = 4), Egger's regression (n = 1), and normal quantile plots (n = 1). Among these, only two meta-analyses have corrected for the potential influence of publication bias (i.e., using the trim-and-fill method; see Gallardo et al., 2016; Liu et al., 2016). This means that meta-analyses in global change biology are likely to be overestimating overall effects. In this study, we have used a recently proposed multilevel meta-regression approach (Nakagawa, Lagisz, Jennions, et al., 2021) to adjust for publication bias in meta-analyses. After adjustment of publication bias, the magnitude of overall effect sizes has declined by 17% – 32% (see Figure 2). The corresponding values for single experiment power decreased by 9% – 66%. Type M error rates increased by 20%, which indicates that publication bias might have exacerbated the overestimation of anthropogenic impacts in global change studies.

Our results indicate that effect sizes in global change studies are severely exaggerated and call into question their 'reproducibility'. Peer-review journals are more likely to publish statistically significant results, perhaps using statistical significance as a gate-keeping tool to maintain their 'prestige' (e.g., inflated impact factors). Under the publish-or-perish research culture, ecologists may intentionally 'pick' significant results or 'hack' p-values (e.g., HARKing) to pursue a more publishable result (Amrhein, Korner-Nievergelt, & Roth, 2017; Fraser, Parker, Nakagawa, Barnett, & Fidler, 2018). However, the gate-keeping policy might not work well (e.g., failing to increase the citation of papers; Wardle, 2012) and more importantly does not equal good science research.

Evidence from other disciplines has also shown that meta-analyses without correcting publication bias subsequently led to a biased assessment of power (see Button et al., 2013; Ioannidis et al., 2017; T. Stanley et al., 2018). However, even our bias-corrected effect sizes may still be biased (overestimating) to some degree. This is because our meta-regression approach could not control for heterogeneities between studies, which may have prevented more accurate adjustments for publication bias (i.e., potentially important moderators not available to incorporate in meta-regression; Nakagawa & Santos, 2012; Noble et al., 2017). Therefore, it is necessary not only to test publication bias and further adjust the influence of publication bias in every meta-analysis, but also, to transparently report all predictors and model information in a publication so that any researchers can implement such adjustments later.

4.4 | The choice of effect sizes for global change studies

Our study provides the first empirical evidence that lnRR is, on average, a more powerful and less biased effect size than SMD and SMDH. Experimental lnRR was twice powerful as SMD and SMDH (lnRR *vs.* SMD *vs* SMDH: 38.5% *vs.* 19.1% *vs.* 18.2%; see Table 3 and Figure 3) and less vulnerable to overestimation; lnRR has been exaggerated by 2-fold, whereas SMD and SMDH have been exaggerated by 3-fold (Table 4 and Figure 4). However, lnRR has a major disadvantage; that is it is only appropriate for ratio scale data (i.e., measurements being bounded at zero; cf. Houle, Pélabon, Wagner, & Hansen, 2011; Nakagawa et al., 2015). Nonetheless, lnRR has many other merits over SMD (Nakagawa et al., 2015), which includes: (1) being more robust with small sample sizes (as SMD is biasedly estimated with small *N*; cf. Hamman, Pappalardo, Bence, Peacor, & Osenberg, 2018), (2) incorporating heteroscedasticity (note that SMDH does assume heteroscedasticity;

cf. Bonett, 2008, 2009; Sánchez-Tójar et al., 2020), and (3) being less affected by scale-dependence (Spake et al., 2021). Incidentally, unlike choosing the mean difference metrics based on the power, the choice between lnCVR and lnVR depends on biological questions, which is described elsewhere (Nakagawa et al., 2015; Senior, Viechtbauer, & Nakagawa, 2020).

5 | CONCLUSIONS AND FUTURE PERSPECTIVES

We have demonstrated that low statistical power and exaggerated effect-size estimates are widespread across the field studies in global change biology, especially when correcting for the influence of publication bias. Manipulative field experiments are not superior to non-manipulative observations in terms of their statistical power and Type M and S errors. Therefore, single experiments whether manipulations or non-manipulations may fail, on average, to provide reliable insights into the anthropogenic impacts of global change by themselves. Likewise, although response variability (heteroscedasticity or variance differences) has important biological and statistical implications in the field, our results have shown single experiments are too underpowered to reliably detect response variability. Therefore, to address questions associated with variance, researchers should use meta-analysis of variation to increase power to reliably detect response variability (we have found 8/12 meta-analyses showing significant response variability – lnCVR, which never have been revealed before; see Figure S10). Such use of meta-analysis of variation can generate new biological hypotheses and inform methodological decisions (i.e., choice of standardized mean effect-size; Nakagawa et al., 2015; Senior et al., 2020). Future global change studies warrant highly powered field studies to reliably inform theory building and policymaking. Such studies are likely to call for more collaboration and team science (Camerer et al., 2016; Collaboration, 2015; O'Dea et al., 2021), and the use of large-scale ecosystem research infrastructures (Roy et al., 2021). Moreover, researchers should strive for open and transparent science practices (Gallagher et al., 2020), such as controlling for magnitude and sign errors when planning field experiments (i.e., extension of power analysis; Lemoine et al., 2016), archiving and sharing data, following the FAIR guideline (i.e., findable, accessible, interoperable, and reusable data; Wilkinson et al., 2016; see also, Crystal-Ornelas et al., 2021), increasing transparent reporting (T. H. Parker et al., 2016), embracing preregistrations and registered reports (T. Parker, Fraser, & Nakagawa, 2019), and implementing more replication projects (Fraser et al., 2020). Adopting these practices will not only aid further meta-analytical syntheses but also make ecological findings more reproducible and reliable in general (Nakagawa & Parker, 2015; O'Dea et al., 2021)."

ACKNOWLEDGEMENTS

YY was supported by the New Research Initiative at the City University of Hong Kong. HH acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG HI 848/26-1) and by 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 program (ZN3285). SN and ML were supported by Australian Research Council (ARC) Discovery Grant (DP210100812).

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

DATA AVAILABILITY SYSTEM

The data and code that support the findings of this study are openly available in https://zenodo.org/record/5496789#.YTmbiI4zY2w.

ORCID

Yefeng Yang https://orcid.org/0000-0002-8610-4016

Helmut Hillebrand https://orcid.org/0000-0001-7449-1613

Malgorzata Lagisz https://orcid.org/0000-0002-3993-6127

Ian Cleasby https://orcid.org/0000-0002-4443-0008

Shinichi Nakagawa https://orcid.org/0000-0002-7765-5182

SUPPORTING INFORMATION

Additional supporting information are available online in the Supporting Information section.

REFERENCES

Akiyama, H., Yan, X., & Yagi, K. (2010). Evaluation of effectiveness of enhanced ‐ efficiency fertilizers as mitigation options for N2O and NO emissions from agricultural soils: meta ‐ analysis. *Global Change Biology, 16*(6), 1837-1846.

Amrhein, V., Korner-Nievergelt, F., & Roth, T. (2017). The earth is flat (p> 0.05): significance thresholds and the crisis of unreplicable research. *PeerJ, 5*, e3544.

Andersen, T., Carstensen, J., Hernandez-Garcia, E., & Duarte, C. M. (2009). Ecological thresholds and regime shifts: approaches to identification. *Trends in ecology & evolution, 24*(1), 49-57.

Ateweberhan, M., & McClanahan, T. R. (2010). Relationship between historical sea-surface temperature variability and climate change-induced coral mortality in the western Indian Ocean. *Marine Pollution Bulletin, 60*(7), 964-970.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal of statistical software, 67*(1), 1-46.

Bodey, T. W., Cleasby, I. R., Bell, F., Parr, N., Schultz, A., Votier, S. C., & Bearhop, S. (2018). A phylogenetically controlled meta ‐ analysis of biologging device effects on birds: Deleterious effects and a call for more standardized reporting of study data. *Methods in Ecology and Evolution, 9*(4), 946-955.

Bonett, D. G. (2008). Confidence intervals for standardized linear contrasts of means. *Psychological methods, 13*(2), 99-109.

Bonett, D. G. (2009). Meta-analytic interval estimation for standardized and unstandardized mean differences. *Psychological methods, 14*(3), 225-238.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365-376.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., . . . Chan, T. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*(6280), 1433-1436.

Cardinale, B. J., Srivastava, D. S., Duffy, J. E., Wright, J. P., Downing, A. L., Sankaran, M., & Jouseau, C. (2006). Effects of biodiversity on the functioning of trophic groups and ecosystems. *Nature, 443*(7114), 989-992.

Carpenter, S. R., & Brock, W. A. (2006). Rising variance: a leading indicator of ecological transition. *Ecology letters, 9*(3), 311-318.

Cleasby, I. R., Morrissey, B. J., Bolton, M., Owen, E., Wilson, L., Wischnewski, S., & Nakagawa, S. (2021). What is our power to detect device effects in animal tracking studies? *Methods in Ecology and Evolution, 0*, 1-12.

Cleasby, I. R., & Nakagawa, S. (2011). Neglected biological patterns in the residuals. *Behavioral Ecology and Sociobiology, 65*(12), 2361-2372.

Clobert, J., Chanzy, A., Le Galliard, J.-F., Chabbi, A., Greiveldinger, L., Caquet, T., . . . Roy, J. (2018). How to integrate experimental research approaches in ecological and environmental studies: AnaEE France as an example. *Frontiers in Ecology and Evolution, 6*, 43.

Cohn, L. D., & Becker, B. J. (2003). How meta-analysis increases statistical power. *Psychological methods, 8*(3), 243-253.

Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251).

Crossley, M. S., Meier, A. R., Baldwin, E. M., Berry, L. L., Crenshaw, L. C., Hartman, G. L., . . . Varriano, S. (2020). No net insect abundance and diversity declines across US Long Term Ecological Research sites. *Nature Ecology & Evolution, 4*(10), 1368-1376.

Crystal-Ornelas, R., Varadharajan, C., Bond‐Lamberty, B., Boye, K., Burrus, M., Cholia, S., . . . Ely, K. S. (2021). A guide to using GitHub for developing and versioning data standards and reporting formats. *Earth and Space Science*, e2021EA001797.

De Villemereuil, P., Morrissey, M. B., Nakagawa, S., & Schielzeth, H. (2018). Fixed‐effect variance and the estimation of repeatabilities and heritabilities: Issues and solutions. *Journal of Evolutionary Biology, 31*(4), 621-632.

Dijkstra, F., & Adams, M. (2015). Fire eases imbalances of nitrogen and phosphorus in woody plants. *Ecosystems, 18*(5), 769-779.

Dooley, S. R., & Treseder, K. K. (2012). The effect of fire on microbial biomass: a meta-analysis of field studies. *Biogeochemistry, 109*(1), 49-61.

Elmendorf, S. C., Henry, G. H., Hollister, R. D., Fosaa, A. M., Gould, W. A., Hermanutz, L., . . . Lévesque, E. (2015). Experiment, monitoring, and gradient methods used to infer climate change effects on plant communities yield consistent patterns. *Proceedings of the National Academy of Sciences, 112*(2), 448-452.

Elser, J. J., Bracken, M. E., Cleland, E. E., Gruner, D. S., Harpole, W. S., Hillebrand, H., . . . Smith, J. E. (2007). Global analysis of nitrogen and phosphorus limitation of primary producers in freshwater, marine and terrestrial ecosystems. *Ecology letters, 10*(12), 1135-1142.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: journal of the Econometric Society*, 987-1007.

Fidler, F., Chee, Y. E., Wintle, B. C., Burgman, M. A., McCarthy, M. A., & Gordon, A. (2017). Metaresearch for evaluating reproducibility in ecology and evolution. *BioScience, 67*(3), 282-289.

Fraser, H., Barnett, A., Parker, T. H., & Fidler, F. (2020). The role of replication studies in ecology. *Ecology and evolution, 10*(12), 5197-5207.

Fraser, H., Parker, T., Nakagawa, S., Barnett, A., & Fidler, F. (2018). Questionable research practices in ecology and evolution. *PloS one, 13*(7), e0200303.

Gaertner, M., Biggs, R., Te Beest, M., Hui, C., Molofsky, J., & Richardson, D. M. (2014). Invasive plants as drivers of regime shifts: identifying high‐priority invaders that alter feedback relationships. *Diversity and Distributions, 20*(7), 733-744.

Gallagher, R. V., Falster, D. S., Maitner, B. S., Salguero-Gómez, R., Vandvik, V., Pearse, W. D., . . . Madin, J. S. (2020). Open Science principles for accelerating trait-based science across the Tree of Life. *Nature Ecology & Evolution, 4*(3), 294-303.

Gallardo, B., Clavero, M., Sánchez, M. I., & Vilà, M. (2016). Global ecological impacts of invasive species in aquatic ecosystems. *Global Change Biology, 22*(1), 151-163.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science, 9*(6), 641-651.

Gibson, L., Lee, T. M., Koh, L. P., Brook, B. W., Gardner, T. A., Barlow, J., . . . Lovejoy, T. E. (2011). Primary forests are irreplaceable for sustaining tropical biodiversity. *Nature, 478*(7369), 378-381.

Griffin, J. N., Byrnes, J. E., & Cardinale, B. J. (2013). Effects of predator richness on prey suppression: a meta‐analysis. *Ecology, 94*(10), 2180-2187.
Gurevitch, J., Koricheva, J., Nakagawa, S., & Stewart, G. (2018). Meta-analysis and the science of research synthesis. *Nature, 555*(7695), 175-182.
Hadfield, J. D., Wilson, A. J., Garant, D., Sheldon, B. C., & Kruuk, L. E. (2010). The misuse of BLUP in ecology and evolution. *The American Naturalist, 175*(1), 116-125.
Hamman, E. A., Pappalardo, P., Bence, J. R., Peacor, S. D., & Osenberg, C. W. (2018). Bias in meta‐analyses using Hedges'd. *Ecosphere, 9*(9), e02419.
Hanson, P. J., & Walker, A. P. (2020). Advancing global change biology through experimental manipulations: Where have we been and where might we go? *Global Change Biology, 26*(1), 287-299.
Harpole, W. S., Sullivan, L. L., Lind, E. M., Firn, J., Adler, P. B., Borer, E. T., . . . Hillebrand, H. (2016). Addition of multiple limiting resources reduces grassland diversity. *Nature, 537*(7618), 93-96.
Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological bulletin, 92*(2), 490–499.
Hedges, L. V., Gurevitch, J., & Curtis, P. S. (1999). The meta‐analysis of response ratios in experimental ecology. *Ecology, 80*(4), 1150-1156.
Hillebrand, H., Donohue, I., Harpole, W. S., Hodapp, D., Kucera, M., Lewandowska, A. M., . . . Freund, J. A. (2020). Thresholds for ecological responses to global change do not emerge from empirical data. *Nature Ecology & Evolution, 4*(11), 1502-1509.
Houle, D., Pélabon, C., Wagner, G. P., & Hansen, T. F. (2011). Measurement and meaning in biology. *The quarterly review of biology, 86*(1), 3-34.
Ioannidis, J. P., Stanley, T. D., & Doucouliagos, H. (2017). The power of bias in economics research. *Economic Journal, 127*, F236–F265.
Jennions, M. D., & Møller, A. P. (2003). A survey of the statistical power of research in behavioral ecology and animal behavior. *Behavioral ecology, 14*(3), 438-445.
Korell, L., Auge, H., Chase, J. M., Harpole, S., & Knight, T. M. (2020). We need more realistic climate change experiments for understanding ecosystems of the future. *Global Change Biology, 26*(2), 325-327.
Kreyling, J., & Beier, C. (2013). Complexity in climate change manipulation experiments. *BioScience, 63*(9), 763-767.
Lamberink, H. J., Otte, W. M., Sinke, M. R., Lakens, D., Glasziou, P. P., Tijdink, J. K., & Vinkers, C. H. (2018). Statistical power of clinical trials increased while effect size remained stable: an empirical analysis of 136,212 clinical trials between 1975 and 2014. *Journal of clinical epidemiology, 102*, 123-128.
Lekberg, Y., Arnillas, C. A., Borer, E. T., Bullington, L. S., Fierer, N., Kennedy, P. G., . . . Henning, J. A. (2021). Nitrogen and phosphorus fertilization consistently favor pathogenic over mutualistic fungi in grassland soils. *Nature communications, 12*(1), 1-8.
Lemoine, N. P., Hoffman, A., Felton, A. J., Baur, L., Chaves, F., Gray, J., . . . Smith, M. D. (2016). Underappreciated problems of low replication in ecological field studies. *Ecology, 97*(10), 2554-2561.
Liang, J., Qi, X., Souza, L., & Luo, Y. (2016). Processes regulating progressive nitrogen limitation under elevated carbon dioxide: a meta-analysis. *Biogeosciences, 13*(9), 2689-2699.
Lin, D., Xia, J., & Wan, S. (2010). Climate warming and biomass accumulation of terrestrial plants: a meta‐analysis. *New Phytologist, 188*(1), 187-198.

Liu, L., Wang, X., Lajeunesse, M. J., Miao, G., Piao, S., Wan, S., . . . Li, P. (2016). A cross‐biome synthesis of soil respiration and its determinants under simulated precipitation changes. *Global Change Biology, 22*(4), 1394-1405.
Lu, M., Zhou, X., Yang, Q., Li, H., Luo, Y., Fang, C., . . . Li, B. (2013). Responses of ecosystem carbon cycle to experimental warming: a meta‐analysis. *Ecology, 94*(3), 726-738.
Morrissey, M. B. (2016). Meta‐analysis of magnitudes, differences and variation in evolutionary parameters. *Journal of Evolutionary Biology, 29*(10), 1882-1904.
Nagelkerken, I., & Connell, S. D. (2015). Global alteration of ocean ecosystem functioning due to increasing human CO2 emissions. *Proceedings of the National Academy of Sciences, 112*(43), 13272-13277.
Nakagawa, S., & Foster, T. M. (2004). The case against retrospective statistical power analyses with an introduction to power analysis. *Acta ethologica, 7*(2), 103-108.
Nakagawa, S., Johnson, P. C., & Schielzeth, H. (2017). The coefficient of determination R 2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface, 14*(134), 20170213.
Nakagawa, S., Lagisz, M., Jennions, M. D., Koricheva, J., Noble, D., Parker, T. H., . . . O'Dea, R. E. (2021). Methods for testing publication bias in ecological and evolutionary meta-analyses. *Methods in Ecology and Evolution, accepted*.
Nakagawa, S., Lagisz, M., O'Dea, R. E., Rutkowska, J., Yang, Y., Noble, D. W., & Senior, A. M. (2021). The orchard plot: Cultivating a forest plot for use in ecology, evolution, and beyond. *Research Synthesis Methods, 12*(1), 4-12.
Nakagawa, S., & Parker, T. H. (2015). Replicating research in ecology and evolution: feasibility, incentives, and the cost-benefit conundrum. *BMC biology, 13*(1), 1-6.
Nakagawa, S., Poulin, R., Mengersen, K., Reinhold, K., Engqvist, L., Lagisz, M., & Senior, A. M. (2015). Meta‐analysis of variation: ecological and evolutionary applications and beyond. *Methods in Ecology and Evolution, 6*(2), 143-152.
Nakagawa, S., Samarasinghe, G., Haddaway, N. R., Westgate, M. J., O'Dea, R. E., Noble, D. W., & Lagisz, M. (2019). Research weaving: visualizing the future of research synthesis. *Trends in ecology & evolution, 34*(3), 224-238.
Nakagawa, S., & Santos, E. S. (2012). Methodological issues and advances in biological meta-analysis. *Evolutionary Ecology, 26*(5), 1253-1274.
Noble, D. W., Lagisz, M., O'dea, R. E., & Nakagawa, S. (2017). Nonindependence and sensitivity analyses in ecological and evolutionary meta‐analyses. *Molecular Ecology, 26*(9), 2410-2425.
O'Dea, R. E., Parker, T. H., Chee, Y. E., Culina, A., Drobniak, S. M., Duncan, D. H., . . . Kelly, C. D. (2021). Towards open, reliable, and transparent ecology and evolutionary biology. *BMC biology, 19*(1), 1-5.
Östman, Ö., Eklöf, J., Eriksson, B. K., Olsson, J., Moksnes, P. O., & Bergström, U. (2016). Top‐down control as important as nutrient enrichment for eutrophication effects in North Atlantic coastal ecosystems. *Journal of Applied Ecology, 53*(4), 1138-1147.
Palmer, A. R. (2000). Quasi-replication and the contract of error: lessons from sex ratios, heritabilities and fluctuating asymmetry. *Annual Review of Ecology and Systematics, 31*(1), 441-480.
Parker, T., Fraser, H., & Nakagawa, S. (2019). Making conservation science more reliable with preregistration and registered reports. In: Wiley Online Library.

Parker, T. H., Forstmeier, W., Koricheva, J., Fidler, F., Hadfield, J. D., Chee, Y. E., . . . Nakagawa, S. (2016). Transparency in ecology and evolution: real problems, real solutions. *Trends in ecology & evolution, 31*(9), 711-719.
Parker, T. H., Griffith, S. C., Bronstein, J. L., Fidler, F., Foster, S., Fraser, H., . . . Seppelt, R. (2018). Empowering peer reviewers with a checklist to improve transparency. *Nature Ecology & Evolution, 2*(6), 929-935.
Rineau, F., Malina, R., Beenaerts, N., Arnauts, N., Bardgett, R. D., Berg, M. P., . . . Davin, E. L. (2019). Towards more predictive and interdisciplinary climate change ecosystem experiments. *Nature Climate Change, 9*(11), 809-816.
Roy, J., Rineau, F., De Boeck, H. J., Nijs, I., Pütz, T., Abiven, S., . . . Brüggemann, N. (2021). Ecotrons: powerful and versatile ecosystem analysers for ecology, agronomy and environmental science. *Global Change Biology, 27*(7), 1387-1407.
Sage, R. F. (2020). Global change biology: A primer. *Global Change Biology, 26*(1), 3-30. doi:https://doi.org/10.1111/gcb.14893
Sánchez‐Tójar, A., Moran, N. P., O'Dea, R. E., Reinhold, K., & Nakagawa, S. (2020). Illustrating the importance of meta‐analysing variances alongside means in ecology and evolution. *Journal of Evolutionary Biology, 33*(9), 1216-1223.
Scheffer, M., Carpenter, S., Foley, J. A., Folke, C., & Walker, B. (2001). Catastrophic shifts in ecosystems. *Nature, 413*(6856), 591-596.
Seekell, D. A., Carpenter, S. R., & Pace, M. L. (2011). Conditional heteroscedasticity as a leading indicator of ecological regime shifts. *The American Naturalist, 178*(4), 442-451.
Senior, A. M., Gosby, A. K., Lu, J., Simpson, S. J., & Raubenheimer, D. (2016). Meta-analysis of variance: an illustration comparing the effects of two dietary interventions on variability in weight. *Evolution, medicine, and public health, 2016*(1), 244-255.
Senior, A. M., Grueber, C. E., Kamiya, T., Lagisz, M., O'dwyer, K., Santos, E. S., & Nakagawa, S. (2016). Heterogeneity in ecological and evolutionary meta‐analyses: its magnitude and implications. *Ecology, 97*(12), 3293-3299.
Senior, A. M., Viechtbauer, W., & Nakagawa, S. (2020). Revisiting and expanding the meta‐analysis of variation: The log coefficient of variation ratio, lnCVR. *Research Synthesis Methods*, e176.
Spake, R., Mori, A. S., Beckmann, M., Martin, P. A., Christie, A. P., Duguid, M. C., & Doncaster, C. P. (2021). Implications of scale dependence for cross‐study syntheses of biodiversity differences. *Ecology letters, 24*(2), 374-390.
Stanley, T., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological bulletin, 144*(12), 1325.
Stanley, T. D., & Doucouliagos, H. (2014). Meta‐regression approximations to reduce publication selection bias. *Research Synthesis Methods, 5*(1), 60-78.
Stanley, T. D., Doucouliagos, H., & Ioannidis, J. P. (2017). Finding the power to reduce publication bias. *Statistics in medicine, 36*(10), 1580-1598.
Sternberg, M., & Yakir, D. (2015). Coordinated approaches for studying long-term ecosystem responses to global change. *Oecologia, 177*(4), 921-924.
Treseder, K. K. (2008). Nitrogen additions and microbial biomass: A meta‐analysis of ecosystem studies. *Ecology letters, 11*(10), 1111-1120.

Van Lent, J., Hergoualc'h, K., & Verchot, L. (2014). Soil N2O and NO emissions from land use and land-use change in the tropics and subtropics: a meta-analysis. *Glob Chang Biol (submitted)*.
Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika, 60*(3), 419-435.
Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of statistical software, 36*(3), 1-48.
Vilà, M., Espinar, J. L., Hejda, M., Hulme, P. E., Jarošík, V., Maron, J. L., . . . Pyšek, P. (2011). Ecological impacts of invasive alien plants: a meta‐analysis of their effects on species, communities and ecosystems. *Ecology letters, 14*(7), 702-708.
Wardle, D. (2012). On plummeting manuscript acceptance rates by the main ecological journals and the progress of ecology. *Ideas in ecology and Evolution, 5*.
Way, D. A. (2021). Announcing GCB reviews–The past, present and future of global change biology at your fingertips. In: Wiley Online Library.
Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., . . . Bourne, P. E. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data, 3*(1), 1-9.
Wolkovich, E. M., Cook, B. I., Allen, J. M., Crimmins, T. M., Betancourt, J. L., Travers, S. E., . . . Kraft, N. J. (2012). Warming experiments underpredict plant phenological responses to climate change. *Nature, 485*(7399), 494-497.
Wu, P. P.-Y., Mengersen, K., McMahon, K., Kendrick, G. A., Chartrand, K., York, P. H., . . . Caley, M. J. (2017). Timing anthropogenic stressors to mitigate their impact on marine ecosystem resilience. *Nature communications, 8*(1), 1-11.
Young, N. S., Ioannidis, J. P. A., & Al-Ubaydli, O. (2008). Why current publication practices may distort science. *PLoS medicine, 5*(10), e201.

TABLES

TABLE 1 The formulas for effect-size statistics used to quantify the effect of environmental stressors on ecosystems response magnitude (mean difference: lnRR, SMD and SMDH) and response variability (variance difference or heteroscedasticity: lnVR, lnCVR). In this paper's context, lnRR, SMD and SMDH represent differences in mean values (magnitude) between a group under a global change stressor and another group under a benign environment, whereas lnVR and lnCVR represent differences in variance around mean between the two groups, without and with adjusting the effect of mean change, respectively

Effect size	Statistics	Annotation
Natural logarithm of response ratio, lnRR (ratio of means)	$\mathrm{lnRR}=\ln\left(\frac{\mathrm{m}_p}{\mathrm{m}_c}\right), (1)$	m_p and m_c denote the average values of measurements from a group with an environmental stressor (p) and a control (c) group.
Sampling variance of lnRR	$S^2_{lnRR}=\frac{\mathrm{sd}_p^2}{n_p\mathrm{m}_p^2}+\frac{\mathrm{sd}_c^2}{n_c\mathrm{m}_c^2}, (2)$	sd_p^2 and sd_c^2 denote corresponding variances of m_p and m_c (standard deviations of the sample), and n_p and n_c denote the sample sizes for environmental stressor (p) and a control (c) group. Other symbols are as with Equation 1.
Standard mean difference, SMD (Hedges' g or Cohen's d)	$\mathrm{SMD}=\frac{\mathrm{m}_p-\mathrm{m}_c}{\sqrt{\frac{(n_p-1)\mathrm{sd}_p^2+(n_c-1)\mathrm{sd}_c^2}{n_p+n_c-2}}}, (3)$	Symbols are as with Equations 1 and 2.
Sampling variance of SMD	$S^2_{SMD}=\frac{n_p+n_c}{n_pn_c}+\frac{\mathrm{SMD}^2}{2(n_p+n_c)}, (4)$	Symbols are as with Equations 1 and 2.

Standardized mean difference with heteroscedasticity, SMDH	$\mathrm{SMDH} = \frac{\mathrm{m}_p - \mathrm{m}_c}{\sqrt{\frac{\mathrm{sd}_p^2 + \mathrm{sd}_c^2}{2}}}, (5)$	Symbols are as with Equations 1 and 2.
Sampling variance of SMDH	$S_{SMDH}^2 = \frac{\mathrm{SMDH}^2\left(\frac{\mathrm{sd}_p^4}{n_p - 1} + \frac{\mathrm{sd}_c^4}{n_c - 1}\right)}{2\left(\mathrm{sd}_p^2 + \mathrm{sd}_c^2\right)^2} + \frac{\frac{\mathrm{sd}_p^2}{n_p - 1} + \frac{\mathrm{sd}_c^2}{n_c - 1}}{\frac{\mathrm{sd}_p^2 + \mathrm{sd}_c^2}{2}}, (6)$	Symbols are as with Equations 1 and 2.
Natural logarithm of variability ratio, lnVR	$\mathrm{lnVR} = \ln\left(\frac{sd_p}{sd_c}\right) + \frac{1}{2}(\frac{1}{n_p - 1} - \frac{1}{n_c - 1}), (7)$	Positive values of lnVR indicate that environmental stressor increases the variance of measurements without adjusting for the effect of mean change (i.e., more variable traits). Symbols are as with Equations 1 and 2.
Sampling variance of lnVR	$S_{lnVR}^2 = \frac{1}{2}(\frac{1}{n_p - 1} - \frac{1}{n_c - 1}), (8)$	Symbols are as with Equation 2.
Natural logarithm of the coefficients of variation, lnCVR	$\mathrm{lnCVR} = \ln\left(\frac{CV_p}{CV_c}\right) + \frac{1}{2}(\frac{1}{n_p - 1} - \frac{1}{n_c - 1}), (9)$	CV_p and CV_c are the coefficient of variation (i.e., standard deviation divided by its mean) for Environmental stressor (p) and control (c) groups. Other symbols are as with Equation 2. Positive values of lnCVR indicate that environmental stressor increases the variance of measurements, while

		adjusting the effect of mean change (i.e., more variable traits). Other symbols are as with Equation 2.
Sampling variance of lnCVR	$S^2_{lnCVR} = \frac{\mathrm{sd}_p^2}{n_p \mathrm{m}_p^2} + \frac{\mathrm{sd}_c^2}{n_c \mathrm{m}_c^2} + \frac{1}{2}(\frac{1}{n_p - 1} + \frac{1}{n_c - 1}), (10)$	Symbols are as with Equations 1 and 2.

TABLE 2 The definitions of statistical power, Type M, and S error rates. For the definitions of lnRR, SMD, SMDH, lnVR and lnCVR, see **Table 1**

Terms	**Definitions**
Statistical power	The probability of detecting a statistically significant effect size: response magnitude (lnRR, SMD) or response variability (lnVR or lnCVR), given that the effect size is non-zero. Given a sample size, the smaller the true effect size (response mangnitude or variability), the lower the statistical power. Also, note that statistical power is 1 – Type 2 error.
Type S error	The probability of a statistically significant effect size having an opposite sign to the true direction (for lnRR, SMD, lnVR or lnCVR), if the true effect size is non-zero. Given a sample size, the smaller the effect size (response mangnitude or variability), the higher the Type S error rate.
Type M error	The multiplicative factor by which the magnitude of an effect size (lnRR, SMD, lnVR, or lnCVR) might be exaggerated when the true effect size is non-zero. Given a sample size, the smaller the effect size (response mangnitude or variability), the higher the Type M error.

TABLE 3 The model estimates of statistical power to detect the effect of environmental stressors on ecosystems response magnitude (lnRR*, lnRR, SMD and SMDH and their publication bias-corrected versions) and response variability (or heteroscedasticity: lnVR and lnCVR). The model estimates of power were reported both on single experiment level and meta-analysis level. We used mixed-effects models and weighted regression models to average over single experiment level statistical power (using MAOMs, cMAOMs, ESSPs and cESSPs), and meta-analysis level statistical power (using MAOMs and cMAOMs), respectively. We noted that (1) the confidence intervals of statistical estimate were asymmetrical due to the back-transformation, (2) statistical power estimates below or above the boundary values (i.e., 0 or 1) were constrained to the boundaries (i.e., $0^{\#}$ or $1^{\#}$). MAOM = meta-analytic overall mean, ESSP = effect-size-specific prediction, cMAOM = bias-corrected meta-analytic overall mean, cESSP = bias-corrected effect-size-specific prediction, *k* = the number of effect sizes, *N* = the number of primary studies

	Effect size	True effect	Model estimates of Statistical power				*k*	*N*
			Median	CI.lb	CI.ub	Mean		
Single experiment								
	lnRR*	cMAOM	0.233	0.218	0.248	0.433	3847	1119
		cESSP	0.279	0.262	0.2887	0.547	3847	1119
		MAOM	0.277	0.260	0.2885	0.515	3847	1119
		ESSP	0.286	0.269	0.304	0.560	3847	1119
	lnRR	cMAOM	0.385	0.353	0.420	0.716	1940	516
		cESSP	0.359	0.331	0.390	0.704	1940	516
		MAOM	0.523	0.486	0.780	0.973	1940	516
		ESSP	0.401	0.370	0.436	0.786	1940	516

	SMD	cMAOM	0.191	0.179	0.205	0.356	1977	516
		cESSP	0.209	0.194	0.225	0.195	1977	516
		MAOM	0.318	0.288	0.343	0.591	1977	516
		ESSP	0.268	0.249	0.288	0.526	1977	516
	SMDH	cMAOM	0.182	0.170	0.195	0.339	1977	516
		cESSP	0.187	0.174	0.201	0.367	1977	516
		MAOM	0.269	0.250	0.2881	0.501	1977	516
		ESSP	0.234	0.217	0.252	0.458	1977	516
	lnVR	MAOM	0.115	0.109	0.122	0.214	1902	514
		ESSP	0.186	0.172	0.201	0.365	1902	514
	lnCVR	MAOM	0.064	0.062	0.067	0.120	1886	513
		ESSP	0.105	0.098	0.112	0.205	1886	513
Meta-analysis								
	lnRR*	cMAOM	0.424	0.286	0.628	0.583	3847	1119
		MAOM	0.567	0.424	0.756	0.780	3847	1119
	lnRR	cMAOM	0.512	0.249	1#	0.704	1940	516
		MAOM	0.665	0.195	1#	0.915	1940	516
	SMD	cMAOM	0.621	0.330	1#	0.855	1977	516
		MAOM	0.645	0.357	1#	0.887	1977	516
	SMDH	cMAOM	0.635	0.352	1#	0.873	1977	516

	MAOM	0.646	0.362	1[#]	0.889	1977	516
lnVR	MAOM	0.439	0.250	0.77	0.604	1902	514
lnCVR	MAOM	0.526	0.315	0.878	0.723	1886	513

TABLE 4 The model estimates of Type M error rate in detecting the effect of environmental stressors on ecosystems response magnitude (lnRR*, lnRR, SMD and SMDH and their publication bias-corrected versions) and response variability (or heteroscedasticity: lnVR and lnCVR). The model estimates of Type M error rate were reported both on single experiment level and meta-analysis level. See more details in **TABLE 3**

	Effect size	True effect	Model estimates of Type M error rate				*k*	*N*
			Median	CI.lb	CI.ub	Mean		
Single experiment								
	lnRR*	cMAOM	3.220	2.960	3.503	6.286	3847	1119
		cESSP	2.900	2.666	3.154	6.947	3847	1119
		MAOM	2.604	2.429	2.793	5.084	3847	1119
		ESSP	2.727	2.539	2.930	6.533	3847	1119
	lnRR	cMAOM	2.004	1.835	2.188	3.911	1940	516
		cESSP	2.100	1.946	2.267	5.031	1940	516
		MAOM	1.526	1.431	1.628	2.980	1940	516
		ESSP	1.968	1.819	2.127	4.714	1940	516
	SMD	cMAOM	2.875	2.680	3.085	5.613	1977	516
		cESSP	3.016	2.778	3.274	7.226	1977	516
		MAOM	2.028	1.902	2.162	3.958	1977	516
		ESSP	2.450	2.272	2.641	5.869	1977	516
	SMDH	cMAOM	2.936	2.748	3.137	5.731	1977	516
		cESSP	3.151	2.912	3.409	7.548	1977	516
		MAOM	2.259	2.116	2.413	4.410	1977	516
		ESSP	2.703	2.498	2.924	6.474	1977	516

	lnVR	MAOM	3.949	3.734	4.176	7.709	1902	514
		ESSP	3.386	3.132	3.660	8.112	1902	514
	lnCVR	MAOM	9.925	9.311	10.58	19.375	1886	513
		ESSP	6.292	5.713	6.929	15.073	1886	513
Meta-analysis								
	lnRR*	cMAOM	1.823	1.252	2.648	2.037	3847	1119
		MAOM	1.345	1.123	1.610	1.504	3847	1119
	lnRR	cMAOM	1.600	0.897	2.839	1.788	1940	516
		MAOM	1.251	0.879	1.776	1.399	1940	516
	SMD	cMAOM	1.379	0.836	2.265	1.542	1977	516
		MAOM	1.292	0.868	1.917	1.445	1977	516
	SMDH	cMAOM	1.305	0.875	1.940	1.459	1977	516
		MAOM	1.286	0.874	1.887	1.438	1977	516
	lnVR	MAOM	1.555	1.081	2.231	1.738	1902	514
	lnCVR	MAOM	1.488	0.911	2.421	1.664	1886	513

916

TABLE 5 The model estimates of Type S error rate in detecting the effect of environmental stressors on ecosystems response magnitude (lnRR*, lnRR, SMD and SMDH and their publication bias-corrected versions) and response variability (or heteroscedasticity: lnVR and lnCVR). The model estimates of Type S error rate were reported both on single experiment level and meta-analysis level. See more details in **TABLE 3**

	Effect size	True effect	Model estimates of Type S error rate				*k*	*N*
			Median	CI.lb	CI.ub	Mean		
Single experiment								
	lnRR*	cMAOM	0.032	0.029	0.036	0.079	3847	1119
		cESSP	0.027	0.024	0.030	0.070	3847	1119
		MAOM	0.025	0.022	0.028	0.060	3847	1119
		ESSP	0.027	0.024	0.03	0.069	3847	1119
	lnRR	cMAOM	0.014	0.011	0.017	0.035	1940	516
		cESSP	0.018	0.015	0.020	0.042	1940	516
		MAOM	0.007	0.005	0.009	0.016	1940	516
		ESSP	0.015	0.012	0.018	0.038	1940	516
	SMD	cMAOM	0.023	0.020	0.027	0.046	1977	516
		cESSP	0.028	0.024	0.032	0.064	1977	516
		MAOM	0.013	0.010	0.015	0.025	1977	516
		ESSP	0.020	0.016	0.023	0.045	1977	516
	SMDH	cMAOM	0.026	0.022	0.029	0.049	1977	516
		cESSP	0.030	0.026	0.034	0.065	1977	516
		MAOM	0.016	0.013	0.019	0.031	1977	516
		ESSP	0.023	0.019	0.026	0.051	1977	516

	lnVR	MAOM	0.050	0.046	0.056	0.077	1902	514
		ESSP	0.037	0.033	0.042	0.083	1902	514
	lnCVR	MAOM	0.199	0.187	0.213	0.260	1886	513
		ESSP	0.087	0.078	0.097	0.171	1886	513
Meta-analysis	lnRR*	cMAOM	0.014	0.003	0.029	0.017	3847	1119
		MAOM	0.004	$0^{\#}$	0.009	0.007	3847	1119
	lnRR	cMAOM	0.014	$0^{\#}$	0.045	0.017	1940	516
		MAOM	0.004	$0^{\#}$	0.017	0.007	1940	516
	SMD	cMAOM	0.009	$0^{\#}$	0.031	0.012	1977	516
		MAOM	0.007	$0^{\#}$	0.022	0.010	1977	516
	SMDH	cMAOM	0.007	$0^{\#}$	0.022	0.010	1977	516
		MAOM	0.006	$0^{\#}$	0.021	0.009	1977	516
	lnVR	MAOM	0.007	$0^{\#}$	0.021	0.010	1902	514
	lnCVR	MAOM	0.005	$0^{\#}$	0.021	0.008	1886	513

FIGURE LEGENDS

FIGURE 1 Conceptual diagrams of effect size calculations from existing field studies and meta-analyses in global change biology, and analytic approaches used to assess the reliability of manipulative experiments and non-manipulative observations to evaluate the effect of stressors on both ecosystem's response magnitude and variability. (A) An overview of the effect sizes used to quantify the ecosystem's response magnitude and variability. Mean differences metrics were utilized to quantify the response magnitude to environmental stressors (i.e., lnRR, SMD, and SMDH), while variance differences metrics were used to characterise the response variability to environmental stressors (i.e., lnVR and lnCVR). In the context of this paper, response variability was an indicator of heteroscedasticity (also known as heterogeneous variances or unequal variance). The detailed definitions and formulas for these effect-size metrics are reported in **TABLE 1**. (B) An overview of the datasets used to quantify statistical power, Type M and Type S errors. The datasets were derived from the work of Hillebrand et al. (2020), compiling 36 meta-analyses. Our lnRR* dataset contained 30 meta-analyses whose effect-size metrics were originally expressed as lnRR. Our lnRR dataset contained recalculated metric of lnRR using descriptive statistics available in 12 out of 30 meta-analyses in the lnRR* dataset. Datasets SMD, SMDH, lnVR and lnCVR contained corresponding metrics also calculated using descriptive statistics available in 12 out of 30 meta-analyses in the lnRR* dataset. n_{MA} represents the number meta-analyses per dataset. (C) The three-step modelling procedure was employed to test our hypotheses.

FIGURE 2 Orchard (forest-like) plots showing the weighted average of response magnitude and variability across all environmental stressors. (A) The effects of environmental stressors on ecosystem response magnitude measured as lnRR*, lnRR, SMD and SMDH. (B) Bias-corrected ecosystem response magnitude. (C) The effects of environmental stressors on ecosystem response variability measured as lnVR and lnCVR. The unfilled circles represent the weighted overall average of response magnitude and variability. The filled circles represent the associated meta-analytic overall mean of each type of environmental stressors (MAOMs or cMAOMs estimated at each meta-analysis). The size of filled circles signifies the estimates of single stressors scaled proportionally to their precisions (precision is the inverse of standard error, SE). Bold whisker line = 95% confidence interval (CI), thin whisker line = 95% prediction interval (PI), k = number of effect sizes (in the context of this figure, it represents the number of MAOM or cMAOM estimates). MAOM = meta-analytic overall mean, cMAOM = bias-corrected meta-analytic overall mean. We used the *R* package *orchaRd* (Nakagawa, Lagisz, O'Dea, et al., 2021) for visualizations.

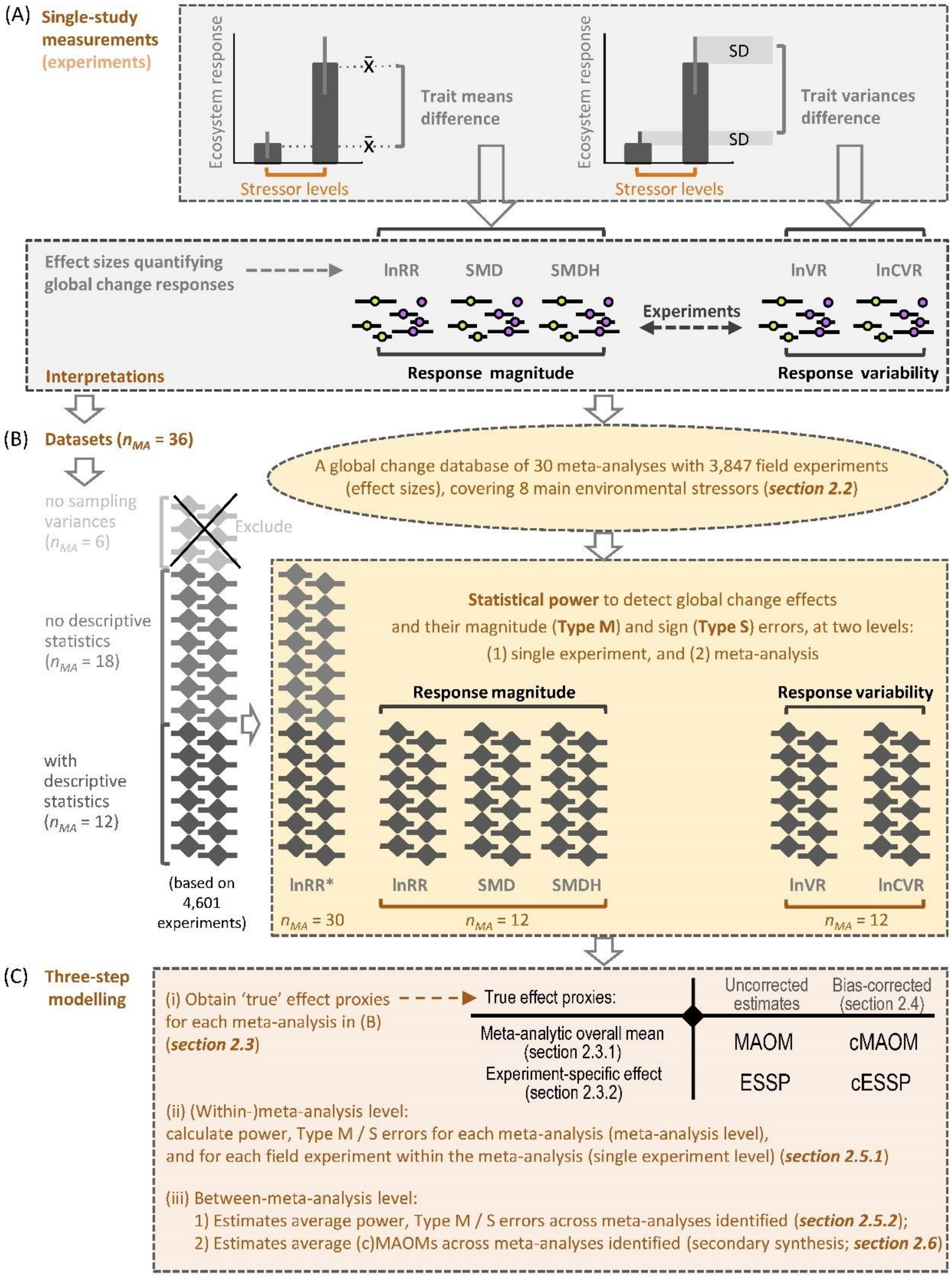
FIGURE 3 Single experiments' median power to detect response magnitude and variability for each category of environmental stressors (on the *y*-axis; stressors with different subscripts denoted that a given stressor may be covered by multiple different meta-analytic cases), assuming one common 'true' effect per stressor (MAOM), experiment-specific 'true' effects within a stressor (ESSP), and their bias-corrected estimates (cMAOM and cESSP) as 'true' effects. The use of meta-analysis increased the statistical power for some environmental stressors (MAOM.MA and cMAOM.MA). (A) the dataset lnRR* (n_{MA} = 30, k = 3,847). (B) the dataset SMD (n_{MA} = 12, k = 1,977). (C) the dataset lnVR (n_{MA} = 12, k = 1,902). (D) the dataset SMDH (n_{MA} = 12, k = 1,977). (E) the dataset lnCVR (n_{MA} = 12, k = 1,886). Warm = global warming, Fire = bush fire, Inv = plant invasion, Fert = fertilization, LUC = land use change, BD loss = biodiversity loss, Acid = acidification, Precip = precipitation. MAOM = meta-analytic overall mean, ESSP = effect-size-specific prediction, cMAOM = bias-corrected meta-analytic overall mean, cESSP = bias-corrected effect-size-specific prediction, n_{MA} = the number meta-analyses per dataset, k = the number of effect sizes.

FIGURE 4 Single experiments' median Type M error rates (i.e., exaggeration ratio) in detecting response magnitude to each category of environmental stressors (on the *y*-axis; stressors with different subscripts denoted that a given stressor may be covered by multiple different meta-analytic cases), assuming one common 'true' effect per stressor (MAOM), experiment-specific 'true' effects within a stressor (ESSP), and their bias-corrected estimates (cMAOM and cESSP) as 'true' effects. The use of meta-analysis reduced the Type M error rates in some environmental stressors (MAOM.MA). (A) the dataset lnRR*. (B) the dataset SMD. (C) the dataset lnVR. (D) the dataset. (E) the dataset lnCVR. The definition of Type M error rate can be found at **TABLE 2**. Grey cells indicate that Type M errors are greater than 3. See more details in the legend of **FIGURE 3**

FIGURE 5 Single experiments' median Type S error rates in detecting response magnitude to each category of environmental stressors (on the *y*-axis; stressors with different subscripts denoted that a given stressor may be covered by multiple different meta-analytic cases), assuming one common 'true' effect per stressor (MAOM), experiment-specific 'true' effects within a stressor (ESSP), and their bias-corrected estimates (cMAOM and cESSP) as 'true' effects. The use of meta-analysis reduced the Type S error rates in some environmental stressors (MAOM.MA). (A) the dataset lnRR*. (B) the dataset SMD. (C) the dataset lnVR. (D) the dataset. (E) the dataset lnCVR. The definition of Type S error rate can be found at **TABLE 2**. See more details in the legend of **FIGURE 3**

FIGURE 6 Forest plots showing the model estimates of statistical power, Type M and S errors. The mixed-effects models were used to compare the statistical power, Type M and S error rates between manipulative experiments and non-manipulative observations. (A) – (F) Statistical power of manipulative experiments and non-manipulative observations to detect response magnitude (lnRR*, lnRR, SMD, and SMDH) and variability (lnVR and lnCVR). (G) – (L) Type M errors in manipulative experiments and non-manipulative observations. (M) – (R) Type S errors in manipulative experiments and non-manipulative observations. * indicates a statistically significant difference between manipulative experiments and non-manipulative observations. See more details in the legend of **FIGURE 3**.

FIGURE 1



(A) Single-study measurements (experiments)
Ecosystem response
x̄
x̄
Trait means difference
Stressor levels
Ecosystem response
SD
SD
Trait variances difference
Stressor levels
Effect sizes quantifying global change responses
lnRR
SMD
SMDH
lnVR
lnCVR
Experiments
Interpretations
Response magnitude
Response variability
(B) Datasets (n_{MA} = 36)
A global change database of 30 meta-analyses with 3,847 field experiments (effect sizes), covering 8 main environmental stressors (*section 2.2*)
no sampling variances (n_{MA} = 6)
Exclude
no descriptive statistics (n_{MA} = 18)
with descriptive statistics (n_{MA} = 12)
(based on 4,601 experiments)
Statistical power to detect global change effects and their magnitude (**Type M**) and sign (**Type S**) errors, at two levels: (1) single experiment, and (2) meta-analysis
Response magnitude
Response variability
lnRR*
lnRR
SMD
SMDH
lnVR
lnCVR
n_{MA} = 30
n_{MA} = 12
n_{MA} = 12
(C) Three-step modelling
(i) Obtain 'true' effect proxies for each meta-analysis in (B) (*section 2.3*)
True effect proxies:
Uncorrected estimates
Bias-corrected (section 2.4)
Meta-analytic overall mean (section 2.3.1)
MAOM
cMAOM
Experiment-specific effect (section 2.3.2)
ESSP
cESSP
(ii) (Within-)meta-analysis level: calculate power, Type M / S errors for each meta-analysis (meta-analysis level), and for each field experiment within the meta-analysis (single experiment level) (*section 2.5.1*)
(iii) Between-meta-analysis level:
1) Estimates average power, Type M / S errors across meta-analyses identified (*section 2.5.2*);
2) Estimates average (c)MAOMs across meta-analyses identified (secondary synthesis; *section 2.6*)

FIGURE 2

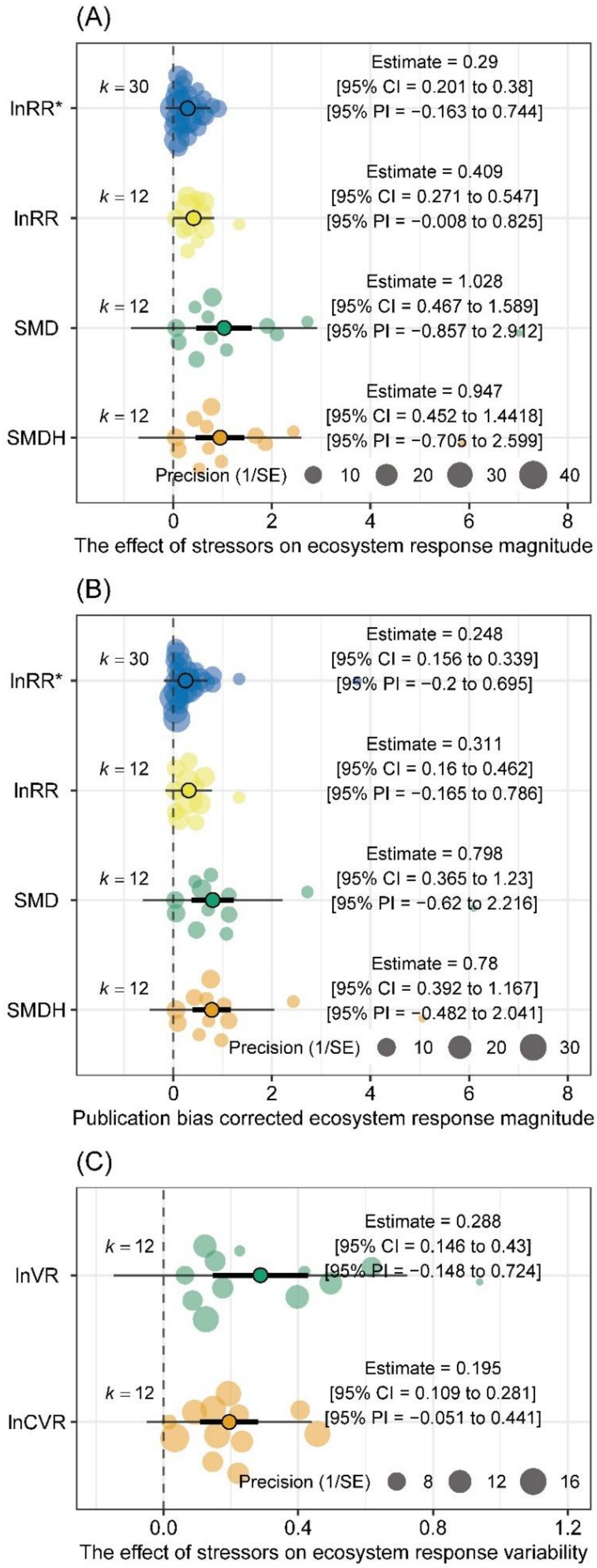
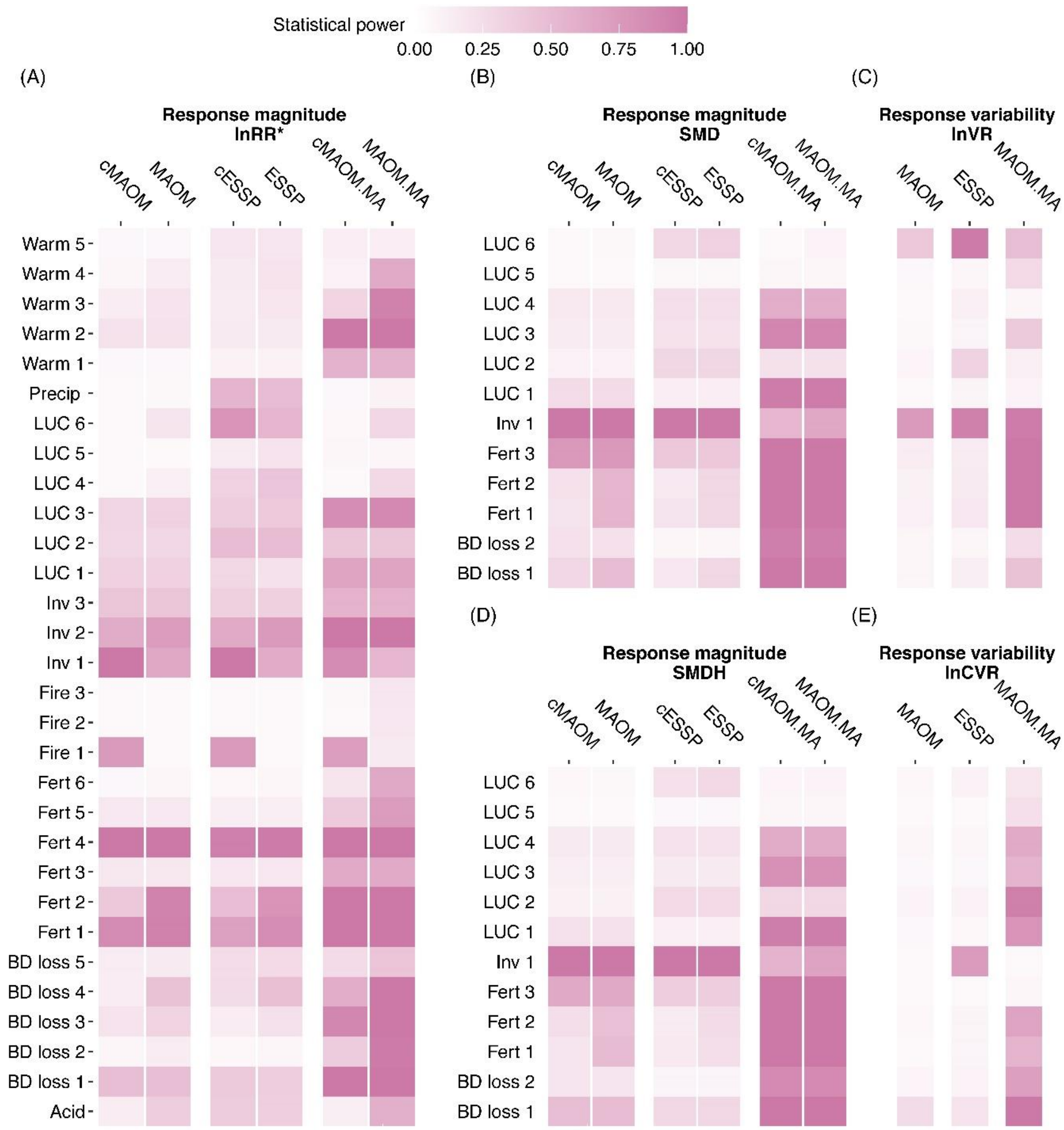
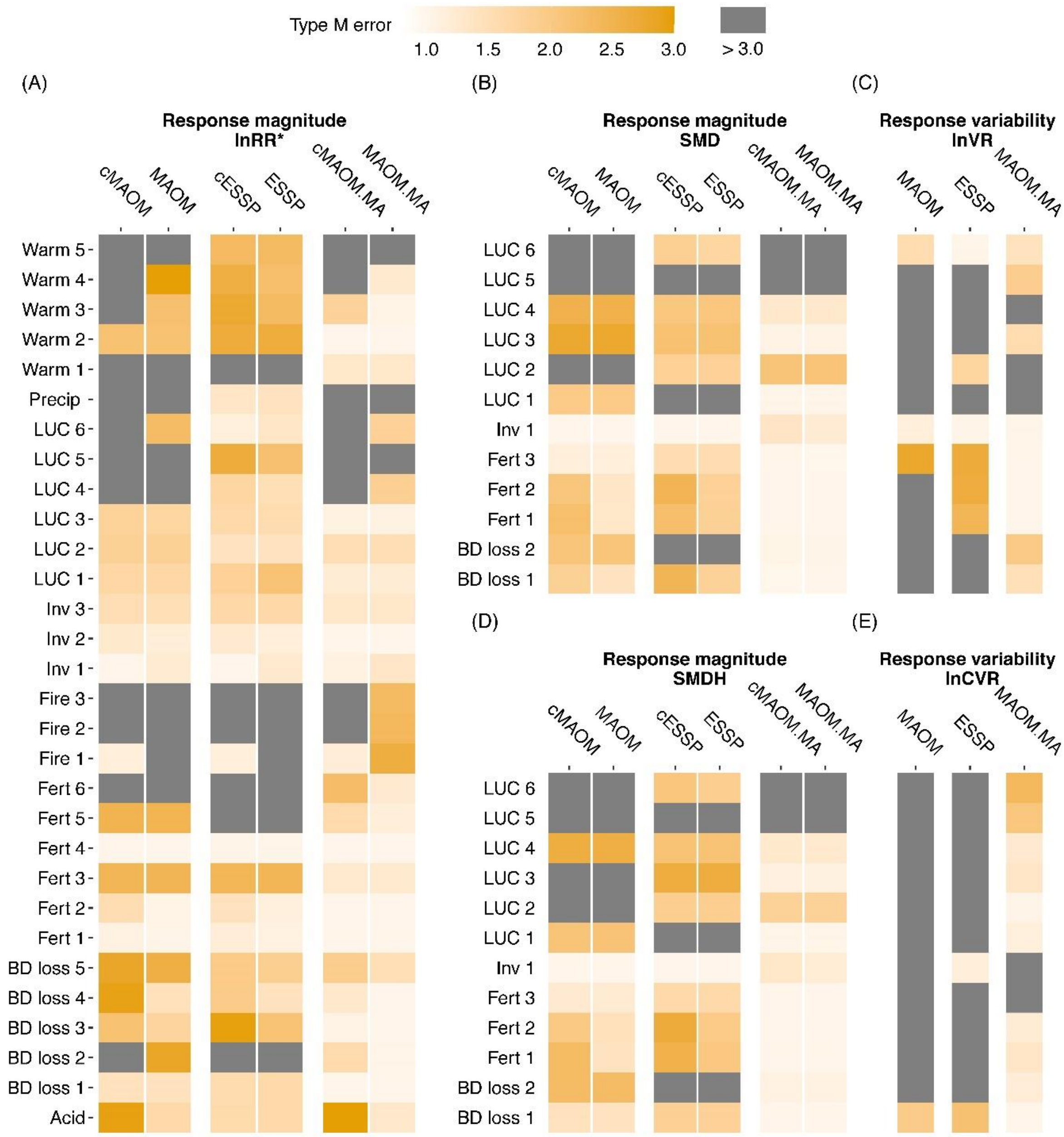

(A)
lnRR*
k = 30
Estimate = 0.29
[95% CI = 0.201 to 0.38]
[95% PI = −0.163 to 0.744]
lnRR
k = 12
Estimate = 0.409
[95% CI = 0.271 to 0.547]
[95% PI = −0.008 to 0.825]
SMD
k = 12
Estimate = 1.028
[95% CI = 0.467 to 1.589]
[95% PI = −0.857 to 2.912]
SMDH
k = 12
Estimate = 0.947
[95% CI = 0.452 to 1.4418]
[95% PI = −0.705 to 2.599]
Precision (1/SE) 10 20 30 40
0 2 4 6 8
The effect of stressors on ecosystem response magnitude
(B)
lnRR*
k = 30
Estimate = 0.248
[95% CI = 0.156 to 0.339]
[95% PI = −0.2 to 0.695]
lnRR
k = 12
Estimate = 0.311
[95% CI = 0.16 to 0.462]
[95% PI = −0.165 to 0.786]
SMD
k = 12
Estimate = 0.798
[95% CI = 0.365 to 1.23]
[95% PI = −0.62 to 2.216]
SMDH
k = 12
Estimate = 0.78
[95% CI = 0.392 to 1.167]
[95% PI = −0.482 to 2.041]
Precision (1/SE) 10 20 30
0 2 4 6 8
Publication bias corrected ecosystem response magnitude
(C)
lnVR
k = 12
Estimate = 0.288
[95% CI = 0.146 to 0.43]
[95% PI = −0.148 to 0.724]
lnCVR
k = 12
Estimate = 0.195
[95% CI = 0.109 to 0.281]
[95% PI = −0.051 to 0.441]
Precision (1/SE) 8 12 16
0.0 0.4 0.8 1.2
The effect of stressors on ecosystem response variability

FIGURE 3



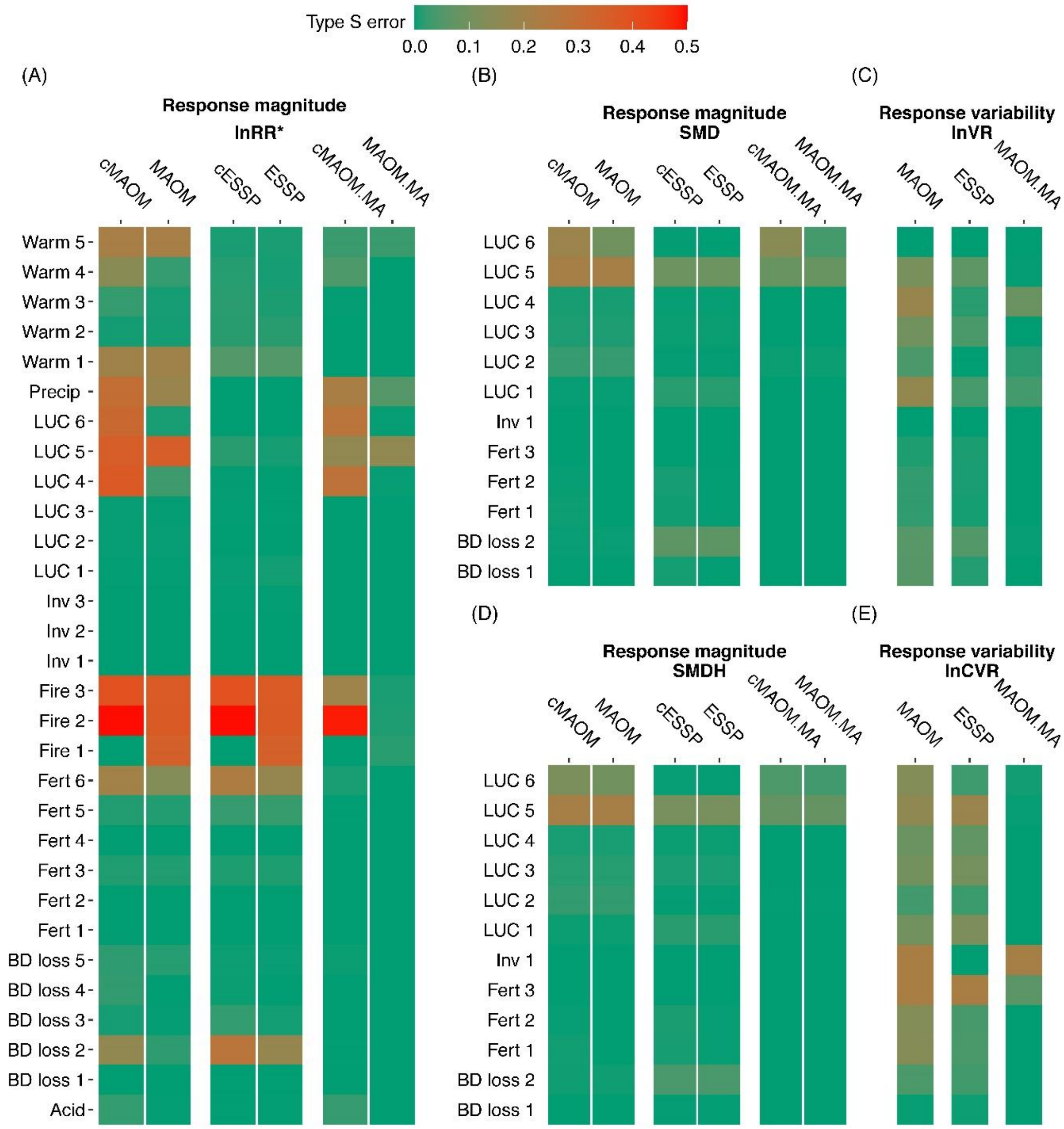
Statistical power
0.00
0.25
0.50
0.75
1.00
(A)
Response magnitude
lnRR*
cMAOM
MAOM
cESSP
ESSP
cMAOM.MA
MAOM.MA
Warm 5
Warm 4
Warm 3
Warm 2
Warm 1
Precip
LUC 6
LUC 5
LUC 4
LUC 3
LUC 2
LUC 1
Inv 3
Inv 2
Inv 1
Fire 3
Fire 2
Fire 1
Fert 6
Fert 5
Fert 4
Fert 3
Fert 2
Fert 1
BD loss 5
BD loss 4
BD loss 3
BD loss 2
BD loss 1
Acid
(B)
Response magnitude
SMD
cMAOM
MAOM
cESSP
ESSP
cMAOM.MA
MAOM.MA
LUC 6
LUC 5
LUC 4
LUC 3
LUC 2
LUC 1
Inv 1
Fert 3
Fert 2
Fert 1
BD loss 2
BD loss 1
(C)
Response variability
lnVR
MAOM
ESSP
MAOM.MA
(D)
Response magnitude
SMDH
cMAOM
MAOM
cESSP
ESSP
cMAOM.MA
MAOM.MA
LUC 6
LUC 5
LUC 4
LUC 3
LUC 2
LUC 1
Inv 1
Fert 3
Fert 2
Fert 1
BD loss 2
BD loss 1
(E)
Response variability
lnCVR
MAOM
ESSP
MAOM.MA

1011 **FIGURE 4**

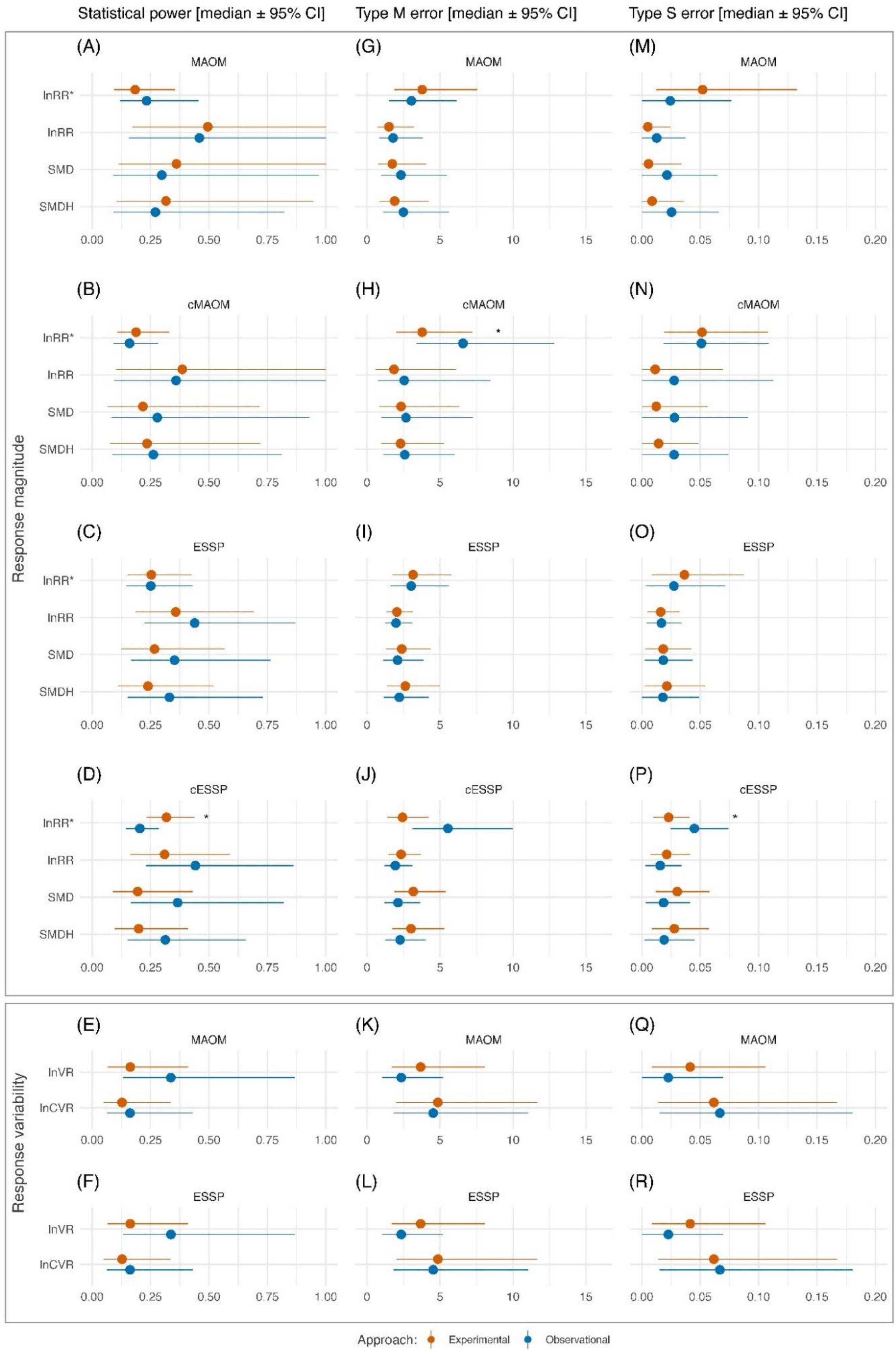


1012

FIGURE 5


Type S error
0.0 0.1 0.2 0.3 0.4 0.5
(A)
Response magnitude
lnRR*
cMAOM MAOM cESSP ESSP cMAOM.MA MAOM.MA
Warm 5
Warm 4
Warm 3
Warm 2
Warm 1
Precip
LUC 6
LUC 5
LUC 4
LUC 3
LUC 2
LUC 1
Inv 3
Inv 2
Inv 1
Fire 3
Fire 2
Fire 1
Fert 6
Fert 5
Fert 4
Fert 3
Fert 2
Fert 1
BD loss 5
BD loss 4
BD loss 3
BD loss 2
BD loss 1
Acid
(B)
Response magnitude
SMD
cMAOM MAOM cESSP ESSP cMAOM.MA MAOM.MA
(C)
Response variability
lnVR
MAOM ESSP MAOM.MA
LUC 6
LUC 5
LUC 4
LUC 3
LUC 2
LUC 1
Inv 1
Fert 3
Fert 2
Fert 1
BD loss 2
BD loss 1
(D)
Response magnitude
SMDH
cMAOM MAOM cESSP ESSP cMAOM.MA MAOM.MA
(E)
Response variability
lnCVR
MAOM ESSP MAOM.MA
LUC 6
LUC 5
LUC 4
LUC 3
LUC 2
LUC 1
Inv 1
Fert 3
Fert 2
Fert 1
BD loss 2
BD loss 1

1015 **FIGURE 6**



1016