To replicate, or not to replicate – that is the question: how to tackle nonlinear responses in ecological experiments

Juergen Kreyling,[1*]† Andreas H. Schweiger,[2†] Michael Bahn,[3] Phil Ineson,[4] Mirco Migliavacca,[5] Thibaut Morel-Journel,[6] Jesper Riis Christiansen,[7] Nicolas Schtickzelle[6] and Klaus Steenberg Larsen[7]

[1] Experimental Plant Ecology, Institute for Botany and Landscape Ecology, Greifswald University, Soldmannstraße 15, D-17487 Greifswald, Germany

[2] Plant Ecology, Bayreuth,Center of Ecology and Environmental Research (BayCEER), University of Bayreuth, Universitaetsstr. 30, 95447 Bayreuth,Germany

[3] Institute of Ecology, University of Innsbruck, Sternwartestr. 15, 6020Innsbruck, Austria

[4] Department of Biology, University of York, Heslington, York YO10 5DD, UK

[5] Max Planck Institute for Biogeochemistry, Biogeochemical Integration Department, Hans Knöll Straße, 10, Jena, Germany

[6] Earth and Life Institute, Universit_e catholique de Louvain, Croix du sud 4/ L7.07.04, 1348 Louvain-la-Neuve, Belgium

[7] Department of Geosciences and Natural Resource Management, University of Copenhagen, Rolighedsvej 23, 1958 Frederiksberg C, Denmark

*Correspondence: E-mail: juergen.kreyling@uni-greifswald.de
†These authors contributed equal to this work

Abstract

A fundamental challenge in experimental ecology is to capture nonlinearities of ecological responses to interacting environmental drivers. Here, we demonstrate that gradient designs outperform replicated designs for detecting and quantifying nonlinear responses. We report the results of (1) multiple computer simulations and (2) two purpose-designed empirical experiments. The findings consistently revealed that unreplicated sampling at a maximum number of sampling locations maximised prediction success (i.e. the R^2 to the known truth) irrespective of the amount of stochasticity and the underlying response surfaces, including combinations of two linear, unimodal or saturating drivers. For the two empirical experiments, the same pattern was found, with gradient designs outperforming replicated designs in revealing the response surfaces of underlying drivers. Our findings suggest that a move to gradient designs in ecological experiments could be a major step towards unravelling underlying response patterns to continuous and interacting environmental drivers in a feasible and statistically powerful way.

Introduction

Classically, many of the major questions being raised about ecological responses to environmental drivers (e.g. pollution, climate change, loss of diversity) have involved the experimental testing of a simple null hypothesis, that is, that there is no significant impact of the driver on a particular ecological response variable. One example has been the investigation of the field impacts of elevated CO_2 on ecosystem processes, which initially exclusively involved experiments with CO_2 concentrations set at two levels, typically current ambient (ca. 380 ppm) vs. future elevated (e.g. 650 ppm). The high costs of such investigations has resulted in typically conservative experimental designs, with both limited replication and levels of the driver, CO_2 (see Pugh *et al* . 2016); this simple approach has also been widely applied for assessing potential ecological impacts of global change, for example, pollution, climate change, grazing pressure and species invasion. However, once a significant impact has been identified it then becomes important in developing mitigation or control strategies to determine the shape of any response pattern to determine if observed impacts are linear or follow nonlinear shapes.

Focusing on response patterns along driver gradients rather than exploring differences among treatment groups has stimulated major advances in ecology (Curtis & McIntosh 1951; Whittaker 1967; ter Braak & Prentice 1988) and in many other disciplines such as, physics and chemistry (Stejskal & Tanner 1965; Grier 2003), socio-economics (Moffitt *et al* . 2011), medicine (Helmlinger *et al* . 1997) and psychology (Hare 1965; Matthews & Power 2002). Controlled, manipulative experiments constitute a major tool to determine causality between observable ecological responses (e.g. species richness, biomass production, CO_2 fluxes, etc.) and the explanatory variables of interest, that is, the putative environmental drivers (e.g. temperature, water availability, soil pH or nutrient content, etc.) (Hurlbert 1984; Beier *et al* . 2012). However, ecological experiments still predominantly rely on replicated designs, with few sampling locations along the environmental drivers, although the underlying questions are often more about the nature of the response shapes, which are commonly nonlinear in ecology and evolution (Levin 1998; Gill *et al* . 2002; Scheffer & Carpenter 2003; Liu *et al* . 2007) and frequently comprise non-additive interactions of multiple environmental drivers (Shaw *et al* . 2002; Larsen *et al* . 2011; Dieleman *et al* . 2012; but see Yue *et al* . 2017). These questions may be more effectively achieved by devoting resources to sampling many levels along gradients of the environmental drivers, while accepting little to no replication, (see Table 1 for use of terms).

Table 1. Glossary of terms used

Term used	Explanation
Experimental units	The independent samples, that is, plots or other sampling units such as individual plants, pots, mesocosms, etc.
Environmental driver	The focal gradient, factor, independent variable or explanatory variable
Locations	Sampling positions along the gradient of environmental drivers. In the two-factorial design we are dealing with the coordinates of each sampling location along those two gradients
Replicates	Experimental units at identical locations
Experimental design	The sampling scheme, that is, how the total number of experimental units is allocated to locations in the environmental driver space and replicates (with total number of experimental units = locations × replicates)
Replicated design	Classical replicated design with at least three true replicates at few locations
Gradient design	Experimental design with maximal number of locations without replication (n = 1 per location)
Hybrid design	Designs with a low level of replication at a maximised number sampling locations

Only if potential nonlinearity in ecological responses and interacting environmental drivers are considered, are experiments optimal for improving ecological models such as earth system models, or the terrestrial biosphere models (IPCC 2013) and predicting ecosystem responses and biosphere-atmosphere feedbacks to environmental variability and extreme events (e.g. Zscheischler *et al* . 2014). To date, such complexity has rarely been acknowledged in the design of ecological experiments (Kreyling & Beier 2013). This means that we need to rethink experimental designs and impose more complex, multifactorial and multilevel experiments while still being limited by the total number of experimental units to designs which are logistically and financially feasible, while maintaining statistically sound analyses. This is a major challenge requiring analysis of the potential trade-offs between scientific

output and the investments in experimental units (e.g. number of replicates, number of environmental drivers, total number of sampling units) and the supporting infrastructure (e.g. labour, instrumentation, technology and consumables).

Gradient designs are posited as potentially overcoming this challenge while simultaneously detecting ecologically important thresholds and nonlinear relationships (Kreyling *et al* . 2014; see Figure 1), echoing early suggestions that sequential, unreplicated designs are necessary to locate maximum response values in multivariate, interactive systems (Box & Wilson 1951). However, despite the discussions of gradient designs (Cottingham *et al* . 2005; Steury & Murray 2005; Thompson *et al* . 2013; Kreyling *et al* . 2014) experiment designers still have very real and important questions about whether such designs are better suited for the detection of nonlinearities and interactions of environmental drivers than classical replicated designs, or if hybrids of both approaches (e.g. with minimal replication and maximal sampling locations; Cottingham *et al* . 2005; Steury & Murray 2005; Schweiger *et al* . 2016) are the superior solution. There is also the question of how such gradient designs should be analysed statistically, particularly if the responses are truly nonlinear, that is, they cannot be analysed by linear models (Cottingham *et al* . 2005).

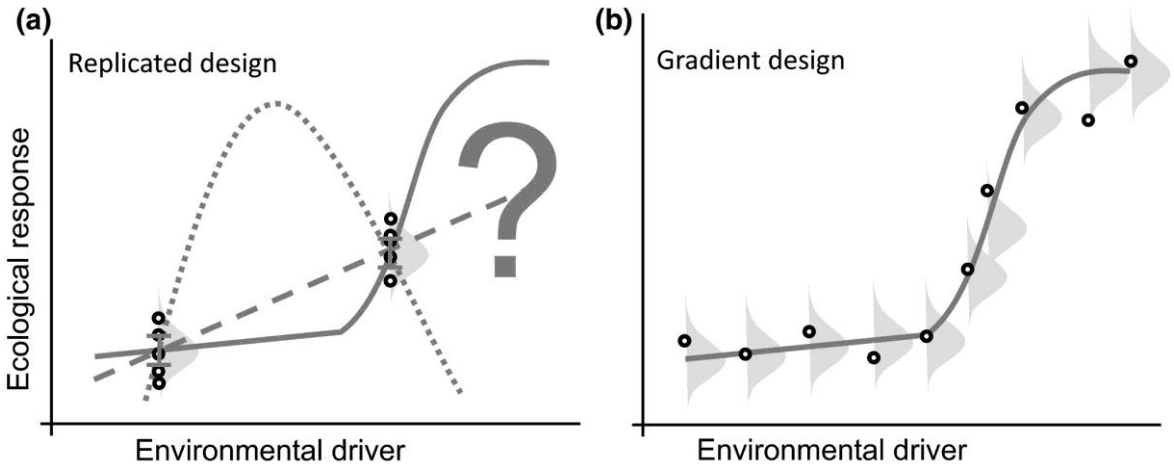

Figure 1: Conceptual visualisation of (a) replicated experiments with high local precision due to replication (indicated by dark grey error bars) but a low number of sampling locations (levels of the environmental driver) and (b) gradient experiments with lower local precision due to no replication but with maximum number of locations (light grey shapes represent probability distributions for random draws in both diagrams). While replicated experiments offer higher confidence in response estimates at the applied locations, the gradient experiments are better suited for characterising underlying nonlinear responses for a given total number of experimental units (10 in this example).

In this study, we specifically addressed the question of how to best allocate a finite total number of experimental units to uncover two-factorial, nonlinear response surfaces; a decision which has to be made frequently when designing any ecological experiment. We

hypothesised that, when constrained to a maximum total number of experimental units, gradient designs would outperform replicated experimental designs with regard to detecting the shape and interaction of underlying response surfaces of two environmental drivers. We tested this hypothesis in two ways: first, we performed a simulation study using artificial data containing stochastic variation in which were embedded known and interacting underlying response surfaces. We compared the ability of a number of different experimental approaches to reveal these known and embedded linear and nonlinear responses, as well as the interactions between the environmental drivers. In these simulations, we deliberately varied the number of replicates and number of sampling locations along each environmental driver, as well as the underlying response shapes (i.e. linear vs. nonlinear) and the amount of stochasticity (noise) in the data. In other words, we allocated a defined number of experimental units to the alternatives of being used to increase replicates or to provide additional points on the response surface to inform how best one should optimise the allocation of limited resources when designing ecological/environmental experiments.

The second approach was to design and execute two very different 'real-world' experiments, specifically constructed with unusually high experimental units and sufficient numbers of replicates and sampling locations, to enable comparison of various resampling strategies from the resulting data. Based on these findings, we hoped to be able to suggest rigorous options for optimising experimental design, together with sound analytical protocols for such designs. The aim was that our findings would help experimentalists optimise towards more appropriate and efficient designs of ecological experiments, providing broader coverage of the environmental driver space, improving potential for interpolation while enabling clearer detection and description of nonlinearities and interactive effects in, and between, different experimental drivers.

Material and Methods

Simulations and analyses of artificial data

We performed simulations based on artificial data to validate the potential of gradient experiments to detect nonlinearities in ecological responses, and interactions, between underlying environmental drivers. We simulated several different response surfaces of an ecological response variable (y) resulting from the interactive effect of different combinations of two underlying drivers (x_1 and x_2). The response variable (y) could represent any kind of measured univariate biotic variable (e.g. biomass production, species richness or microbial activity) in response to the variation in two numeric environmental drivers (e.g. temperature, water availability, soil pH or nutrient availability) which were interactively affecting the studied response variable. For each of the two environmental drivers, we assumed a specific

response of y, formulated as a linear or nonlinear function $y_i = f_i(x_i)$. The final response surface of y, resulting from the interacting effect of the two environmental drivers was modelled as a combined response function $y_{total} = f_1(x_1)+f_2(x_2)$, where the parameters of each of the two individual response functions were dependent on the other environmental driver, thus mimicking the interaction between the two drivers.

The artificial data for our analyses was created (1) completely independent from any kind of empirically derived data and, in addition, (2) based on information taken from empirical data in order to bound the simulations within realistic data ranges. These two approaches yielded the same conclusions; methodological details and results for the second approach are presented in Supporting Information S4. Here, the first approach is described and its results are shown in the results section. Different response surfaces resulting from combinations of three individual response functions were tested: (1) the saturating Michaelis-Menten equation (M), (2) linear (L) and (3) unimodal (U). These individual response functions are frequently used in ecological studies to describe empirical data and differ considerably in form and, hence, provide a good representation of typical and varied biological/ecological responses to environmental drivers. These individual response functions were then combined to provide two-factorial response surfaces.

Here, we present three completely artificial response surfaces with the different combinations of nonlinear and linear response functions (LL, LU and UU; see codes above) and another two response surfaces bounded by empirical data (LL and ML; see Supporting Information S1 for all response surfaces and S4 for further details about the empirically bounded datasets). Completely artificial response surfaces allow for general conclusions, while empirically bounded response surfaces ensure realism. We furthermore tested one response surface (LL) with two different parameter settings, resulting in different absolute response values and, thus, very different levels of total variation in the response variable y (Supporting Information S1). As the simulation results and the consequent conclusions were very robust against this variation in parameter settings/differences in total variation in the response variable for the LL response surface, we did not conduct these additional simulations for the other response surfaces. Due to analytical reasoning, total variation in response values by different parameter settings can be assumed to have similar effects for all response surfaces.

For each artificial response surface, we applied sub-sampling strategies in order to generate artificial test data sets and to explore the ability of each generated sub-data set to reveal the true and known embedded underlying response surface. The sampling strategies thereby varied in the number of total samples drawn from the underlying response surface (total number of experimental units), number of locations sampled along the gradients of the two environmental drivers (locations) and the number of replicates per location. We use the term replicates here for the number of samples taken at a single sampling location, that is, one

point in the driver combinations (cf. Schweiger *et al* . 2016). To reflect different levels of stochasticity (white noise) in the sampled artificial data, we allowed the sampled values of the response variable y to scatter around the 'true' value of y at its specific location on the response surface with a normal distribution corresponding to 20% (i.e. c. 85% percentile) or 100% (100% percentile) of the absolute value of y, at this specific location. Detailed investigations into stochastic variation in empirical environmental data is extremely rare, yet the few studies which explicitly quantify random (i.e. not mechanistically explainable variation) in ecological data estimated the amount of stochastic noise to a maximum of 23% of total variation in the response variable (Richardson *et al* . 2012; based on eddy flux measurements and Kelly *et al* . 2009 for the community composition of diatoms). Based on these empirical estimates, we are confident that the amount of random variation covered in our study covers the majority of situations regarding stochasticity in ecological data (cf. Schweiger *et al* . 2016).

For each response surface, we varied the total number of experimental units sampled from the artificial data sets between 6 and 960, the number of sampled locations between 3 and 960 and the number of sampled replicates per sampling location between 1 (no replication) and 240, with the number of locations multiplied by the number of replicates equalling the total number of experimental units. Sampling for each combination of total number of experimental units, number of locations and number of replicates was repeated 100 times for each level of stochasticity in order to evaluate the success of revealing the known underlying response surface.

Prediction success was quantified as the deviation between the response surface obtained from the response values observed using the different sampling strategies (total number of experimental units, number of locations, number of replicates and amount of stochasticity) and the 'true', known underlying response surface. We therefore predicted response values for a given number of locations for both underlying drivers (x_1 and x_2) ranging from 1 to 100, with an interval of 1, using a local polynomial regression fitting based on the sampled response values obtained for the different sampling strategies by the 'loess' function implemented in R (v. 3.3.2, R Core Team 2016), setting span = 10. For the ML, LU and UU settings, we let the model choose the polynomial degree used for fitting, whereas we fixed degree = 1 for the LL setting. The predicted response values were subsequently tested using linear regression analysis against the 'true' response values at the same given number of locations (100) for all tested sampling strategies. Prediction success was then quantified by multiple R^2 derived from this regression analysis. Prediction success for varying sampling strategies was visualised by plotting multiple R^2 against total number of experimental units for different levels of replication by smoothing the curves using nonlinear/generalised additive and linear modelling.

Collection and analyses of 'real-world' data

Two experiments, the soil nitrous oxide (N_2O) flux experiment and the ciliate experiment, were used to test the validity of the results which had been obtained by artificial simulation. These data sets stem from very different ecological disciplines and were specifically purpose-built for this study to uniquely contain high levels of replication at a large number of two-way experimental and interacting driver gradients. The experiments had to be specifically established because no published data sets were found that allowed for adequate sub-sampling from high to low replication while maximising the number of levels of two interacting ecological drivers; invariably, high numbers of replicates mean fewer levels of interacting ecological drivers in any actual controlled experimental setups.

The soil study provided data for N_2O fluxes from bare soil in response to five levels of carbon (C) and five levels of nitrogen (N) addition that were fully crossed (i.e. a total of 25 treatments, replicated 5 times with a total $n = 125$). The flux data were derived from field measurements conducted at the University of York *SkyLine3D* experimental site using an automated chamber system (*SkyGas3D)* ; N_2O is a highly potent greenhouse gas and its response to changing environmental drivers is important for climate change projections. Experimental details are provided in Supporting Information S2A. All data from this experiment that were used to constrain the artificial response surfaces and conduct the 'real-world' validation of the artificial simulations are provided in Supporting Information S3.

The ciliate experiment focused on *Tetrahymena thermophila* , an actively moving ciliate (unicellular eukaryotic protist), inhabiting fresh water in North America (Collins 2012) and often used as a model organism in laboratory microcosms to address ecological and evolutionary questions (e.g. Jacob *et al* . 2017). We submitted *T. thermophila* cells to a full factorial design of 9 temperatures × 9 levels of nutrients, each combination being replicated 9 times, yielding an exceptionally high total of 729 experimental units. Using well-established protocols based on spectrophotometry and image analysis (Pennekamp & Schtickzelle 2013; Pennekamp *et al* . 2015), we quantified a series of traits on each experimental culture: cell biomass produced after 44 h, average cell size, shape and speed of movement. Cells used were all from the same clonal strain, making this experimental design adequate to study the phenotypic plasticity according to two interacting environmental drivers, a topic of special interest in ecology and conservation biology as plasticity is expected to favour adaptation to changing environmental conditions due to human-induced global change. Experimental details are provided in Supporting Information S2B. All data from this experiment used here are provided in Supporting Information S3.

We applied the same procedures of resampling the data from these 'real-world' experiments as described above and compared the results to those obtained from the artificial simulations. Accordingly, we randomly resampled from the whole 'real-world' (empirical) data sets using

different sampling strategies (total number of experimental units, number of locations in the driver space, number of replicates). For the soil N_2O flux experiment, the total number of sampled experimental units thereby varied between 9 and 25, with an interval of 1, and with the number of replicates varying between 1 and 5, while the number of sampled locations varied between 3 and 25. For the ciliate experiment, the total number of sampled experimental units varied between 9 and 81, with an interval of 1, and with the number of replicates varying between 1 and 9, while the number of sampled locations varied between 3 and 81. These limits stem from sub-sampling the total number of experimental units at locations to provide replicated and gradient designs contrasts and to result in numbers of total experimental units more commonly applied in ecological field and laboratory experiments. Again, the number of locations multiplied by the number of replicates equals the total number of experimental units for each sampled design. Similar to the procedure we applied on the artificial data sets, we quantified prediction success based on modelled, nonlinear reference response surfaces in comparison to the response surface derived from a local polynomial regression fitting based on the resampled data. Each sampling strategy was replicated 100 times for each of the response surfaces. The nonlinear reference response surfaces were constructed based on the full empirical data set of the two experiments by combining linear, saturation, unimodal and skewed unimodal relationships between the response variable (soil N_2O flux and ciliate traits) and the two respective, interacting drivers (C and N addition for soil N_2O flux study, while using respiration and temperature and nutrients variation for the ciliate experiment respectively). Detailed information on the mathematical description of the different reference response surfaces is given in Table S2.1.

We furthermore compared the results obtained for prediction success based on these nonlinear response surfaces for soil N_2O flux and ciliate shape to the prediction success obtained for linear (plane) reference response surfaces constructed from the same, respective empirical data. In addition, we quantified skewness and excess kurtosis of the residuals (absolute differences between predicted and 'true' response values). We did not conduct these additional tests for the other ciliate trait responses (biomass, size and movement) as very similar results were expected (see Fig. S2.2). Skewness and excess kurtosis provide information about the deviation of the observed residuals from a normal distribution, that is, the colour (reddening) of noise in the response variable, which cannot be explained by the mechanisms assumed to underlie the observed ecological response to interacting, environmental drivers. Positive values of skewness thereby indicate that the response values (N_2O soil fluxes, ciliate shape) which were estimated based on the sampled data using local polynomial regression fitting were generally smaller than the 'true', modelled response values, whereas negative values for skewness indicate a general overestimation of the response values from the resampled data. Positive values for excess kurtosis indicate that the residuals were more narrowly distributed around the mean than expected for a normal distribution, whereas negative values of excess kurtosis indicated a flatter distribution. Thus,

the observed higher variability in the observed residuals when compared to normally distributed residuals.

We performed all simulations and calculations in R (v. 3.3.2, R Core Team 2016) with the add-on packages minpack.lm (v. 1.2-1), moments (v. 0.14) and MASS (v. 7.3-45). Statistical relationships were tested with simple linear models with a level of significance of $\alpha = 0.05$. The R script we used for the simulations can be found in the Supporting Information S3.

Results

At any given total number of experimental units, the prediction success in the simulated data generally increased with increasing number of sampling locations per environmental driver against decreasing number of replicates. The gradient designs were superior in detecting the underlying response surface as compared to replicated designs and were robust across all the different underlying response surfaces tested (Fig. 2, see also Supporting Information S4 for empirically bound response surfaces). Even in the absence of nonlinearity, that is, when the response surface along the two environmental drivers was a plane (LL, Fig. 2 a and b and Fig. S4.1 c and d), the gradient design generally outperformed the replicated designs for any given number of experimental units. This finding was also robust across the different levels of stochasticity (white noise) implemented in the sampled data (Fig. 2 and Fig. S4.1).

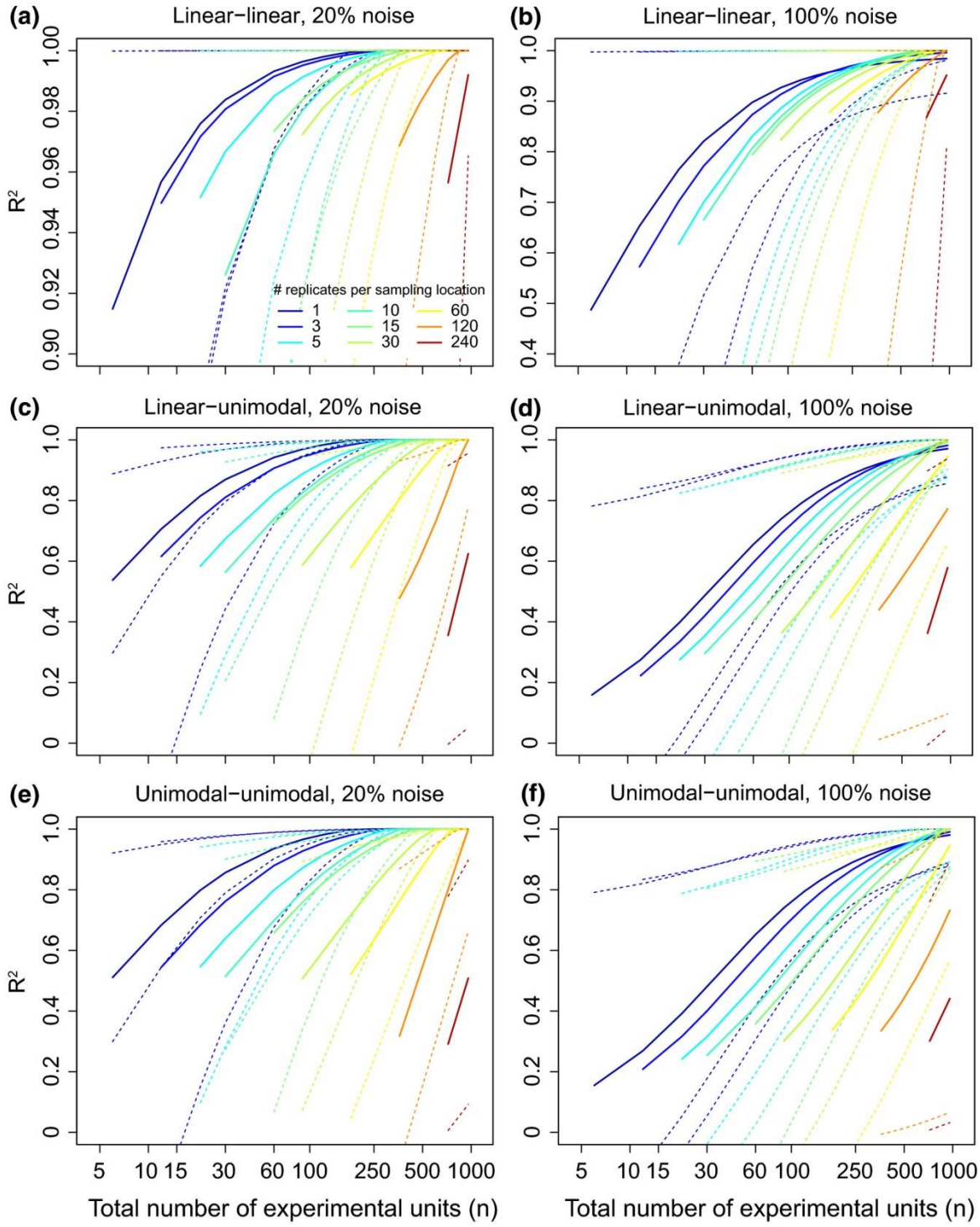

Figure 2: Prediction success at any given total number of experimental units (n = 6–960) but with varying numbers of locations vs. local replicates (n = locations × replicates). Results are based on artificial data sampled from three types of inferred response surfaces, that is, linear-linear (panels a, b, with an arbitrarily chosen parameter setting and thus response values different to the empirically-bound linear – linear response surface in Supporting Information), a linear-unimodal (panels c, d), and a unimodal-unimodal (panels e, f), and with 20% (panels a, c, e) or 100% (panels b, d, f) stochasticity levels (white noise). The replicate = 1 lines (dark blue) represent unreplicated sampling design with maximum number of locations, whereas the replicates = 240 lines (brown) represent the most extreme replicated design (maximum replication with minimum number of locations). Solid lines show mean values whereas the dashed lines represent the 95% confidence intervals from 100 simulation runs.

For the empirical experiments (the soil N_2O flux and ciliate experiments), we reached the same conclusion as for the simulated response surfaces (Fig. 3 and Figure S2.2), namely, the gradient sampling showed the strongest increase in R^2 with total number of experimental units, while also generally outperforming the replicated designs at each given total number of experimental units. Interestingly, differences among the nonlinear reference response surfaces (Fig. 3) and fitting a plane reference response surface (Fig. S2.3) were negligible for soil N_2O flux but were more pronounced for ciliate shape, with higher predictive success for the nonlinear reference surface.

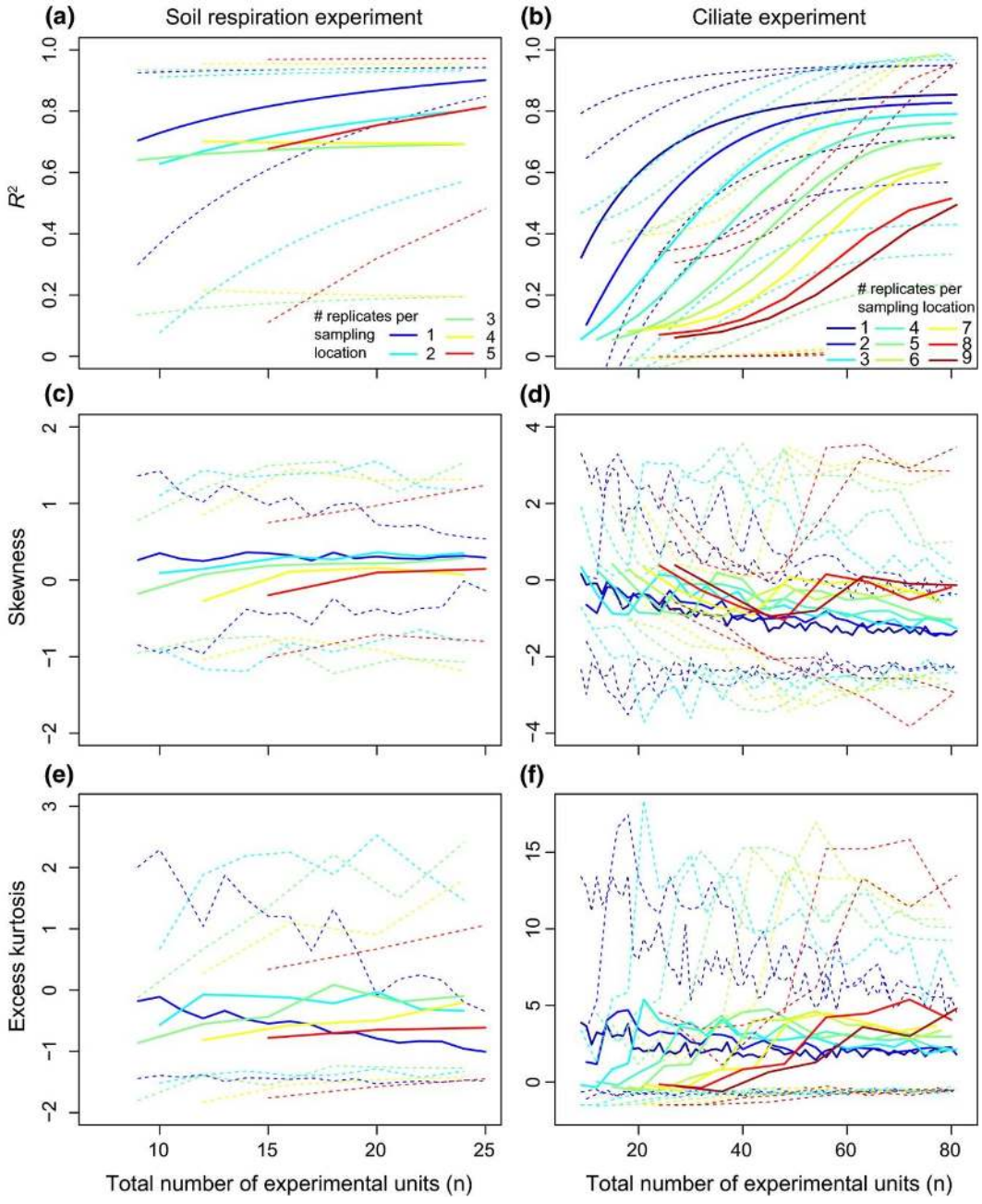

Figure 3: Prediction success for empirical data on soil N_2O fluxes (a, c, e) and *Tetrahymena thermophila* ciliate cell shape (b, d, f) at any given number of experimental units (n = 9–25 and 9–81

respectively) but with varying numbers of locations vs. local replicates (*n* = locations × replicates). (panels a, b), skewness (panels c and d), and excess kurtosis of the residuals (panels e and f). Results are based on a comparison of response surfaces created from 'real world' experimental data on soil N_2O fluxes to carbon and nitrogen additions and on *Tetrahymena thermophila* ciliate cell shape in response to temperature and nutrients variation assuming nonlinear, underlying reference response surfaces created from well-known mechanistic relationships between microbial activity and resource addition (see Table S2.1 for details). The replicate =1 lines (dark blue) represent the gradient design with maximum number of locations, whereas the replicates = 5 lines (red) represent the replicated design (maximum replication with minimum number of locations). Solid lines show mean values whereas the dashed lines represent the 95% confidence intervals from 100 model runs.

The analysis of the residuals between the modelled 'true' reference response surface and the response surface obtained from resampling the empirical data sets by each respective experimental design resulted in similar findings: despite generally smaller differentiation among sampling designs, the variation in skewness (shown as the confidence lines in Fig. 3c and d) decreased with increasing number of locations for any given total number of experimental units. For the gradient design (rep = 1), the number of locations was most strongly increased with increasing number of total samples, thus the decrease in variation in skewness was most pronounced here. This holds true for the nonlinear reference response surfaces (Fig. 3c and d) and to a lesser degree also for the plane reference surfaces (Fig. S2.3 c and d). These findings indicate that increasing the number of locations sampled at the cost of replication robustly decreases the skewness of the residuals and, thus, decreases the risk of systematic over- or under-estimation of the chosen model response surface.

Variance of excess kurtosis decreased with increasing number of locations for both empirical data sets, although there seemed to be a trend towards an increasingly flattened distribution of the residuals with increasing number of locations for soil N_2O flux (Fig. 3e and f). Accordingly, the number of locations along the environmental drivers, and not the level of replication, drove the amount of mechanistically explainable variance, thereby also leaving less non-stochastic pattern in the unexplained variance, resulting in a whitening (i.e. more normally distribution) of the unexplainable noise (Fig. 3 and S2.3).

Discussion

Analysing patterns along natural and experimental gradients has advanced science in the past and presence across disciplines as disparate as physics and chemistry (Stejskal & Tanner 1965; Grier 2003), socio-economics (Moffitt *et al* . 2011), medicine (Helmlinger *et al* . 1997) and psychology (Hare 1965; Matthews & Power 2002). Likewise, ecological theory has benefitted immensely from the analysis of natural gradients as exemplified by the niche concept (Grinnell 1917), the Intermediate Disturbance Hypothesis (Connell 1978), or the Stress-Gradient Hypothesis (Bertness & Callaway 1994). Ecological experiments, however,

predominantly test for differences among groups in replicated experiments rather than exploiting response patterns along experimental gradients of the environmental drivers of interest.

The gradient design consistently outperformed replicated designs with respect to the prediction success of the underlying response surface for any given total number of experimental units in our analyses. The generality of this finding is emphasised because it was found to consistently hold true across (1) a variety of different response surface structures of two interacting environmental drivers in (2) simulated data and very different empirical examples. Our analyses of model residuals furthermore implied that increasing the number of sampling locations, and not the number of replicates per location, was the primary way to reduce unexplained, non-stochastic variance (i.e. skewness and excess kurtosis of the variation). In other words, increased coverage of the underlying environmental gradients generally caused a whitening of the unexplained variance and resulted in models which were more robust to noise. Noise whitening was strongest for the gradient design and weakest for the highly replicated designs. The assumption of normally distributed residuals underlies all parametric statistical tests (e.g. regression or ANOVA) but is often violated when analysing empirical ecological data. We here showed that this basic statistical assumption about white-noise residuals is best met when maximising number of sampling locations for any given number of experimental units available.

Interestingly, we found gradient designs to outperform replicated designs even for linear relationships. This appears counter-intuitive as one would assume the differences among designs to level out in such an underlying pattern and explain this as a probable trade-off between the local precision of prediction and the overall precision of prediction. Higher replication at a certain point on the response surface increases the local precision of predicting the 'true' value. However, if noise in the data is not completely white, sampling a larger area with the same number of total samples instead of local replication reduces the risk of estimating wrong response values from the drawn samples. A structured noise in the response surface space was observed especially for our empirical data set on soil N_2O flux as the residuals got closer to normality with increasing number of sampling locations, being strongest for the unreplicated gradient design.

Sampling locations were randomly assigned in our simulations, based on the logic that the shape of the response surface and the position of nonlinearities and thresholds is usually unknown in ecological experiments. If knowledge on the assumed shape were available, for example, based on literature or pre-trials, one could easily design the sampling strategy in a way to ideally cover the response surface by concentrating sampling in regions with stronger gradients or nonlinearities between drivers and response or identify such regions by sequential

experiments (Box & Wilson 1951). Still, our results imply that gradient designs would allow for a more precise detection of such thresholds than replicated designs.

Hybrid designs, which are a compromise between gradient and replicated designs with some minimal replication and maximal locations have been suggested (Cottingham *et al* . 2005; Steury & Murray 2005; Schweiger *et al* . 2016) or have been applied (Piepho & Bahn 2017). Our data consistently suggest that even such designs are outperformed by gradient designs in terms of both prediction success and whitening of residuals.

Replicated experiments remain the method of choice whenever testing binary environmental drivers such as presence or absence of specific species or functional groups (Table 2). Replicated designs further make sense if a study aims at testing differences among groups (Table 2), for example, comparing sites or management schemes, which differ non-continuously or along unknown gradients. For such replicated experimental designs, analysis of variance (ANOVA) is routinely applied to detect statistical differences between the mean values across treatment groups, but ANOVA does not allow for inter- or extrapolation or the characterisation of the actual shape of response along the considered environmental gradient (response curve, Fig. 1), since experimental units are used to optimise replication and the detection of significant effects rather than determination of response shapes. Note that regression-techniques have been suggested as the statistically more powerful approach even in replicates designs (if the ecological driver of interest was continuous), because they estimate fewer parameters than ANOVA (Cottingham *et al* . 2005). Another example calling for replicated designs is if one is specifically interested in high local precision at one or few sampling locations or in stochastic variation (white noise) rather than means (Table 2).

Table 2. When to use gradient and replicated designs in ecological experiments

Interest in	Statistical method	Recommended experimental design
Differences in a numeric response among treatment groups (mainly for factorial drivers)	Analysis of variance	Replicated
Patterns of a numeric response variable along continuous drivers (e.g. detection of nonlinearity, thresholds, interpolation, extrapolation)	Regression	Gradient
High local precision	Mean	Replicated
Quantification of local variation	Variance, Standard deviation, Confidence Intervals	Replicated

Based on our results, we recommend gradient designs for experiments dealing with ecological responses to continuous environmental drivers (e.g. temperature, precipitation, nutrients, species richness, CO_2-level, physical disturbance, etc.), because the response surface is what is really needed here (Table 2). For example, in the case of research into ecosystem responses to elevated concentrations of atmospheric CO_2, two treatment levels are invariably used, even though the changes in atmospheric CO_2 levels are increasing in a steady upward fashion; we know little of the intermediate parts of the response curves, even though these shapes are needed for developing realistic coupled General Circulation Models (see Pugh *et al* . 2016). The need for independent devices at each sampling locations such as climate chambers in temperature manipulation studies is basically the same as in classical replicated designs if avoidance of pseudo-replication (Hurlbert 1984) is taken seriously. For continuous environmental drivers, gradient designs further allow for better extrapolation, characterisation of (nonlinear) response functions, and, consequently, quantitative outputs better suited for ecological models than replicated designs (Cottingham *et al* . 2005). Confidence bands of regressions further provide sound estimation of uncertainties.

Unreplicated experimental designs to uncover response functions and response surfaces require different analytical strategies to those traditionally used in experimental ecology. Regression focuses directly on the response function (James *et al* . 2017). Linear and multiple

linear regression may serve as null models, while *a priori* knowledge and hypotheses on the likely relationships from existing empirical or theoretical models may be used to set up potentially better, nonlinear candidate models to be tested against the linear model(s). Model indices, such as corrected AIC, BIC, R^2 and Root Mean Square Error and others, may be used to infer the overall better model (Janssen & Heuberger 1995; Burnham & Anderson 2010). A case example testing a range of different models on the empirical soil N_2O flux data set used here is shown in the Supporting Information S5.

Power of regression analyses depends upon the gradient length considered, suggesting that gradient length of the focal environmental drivers in such a gradient design should be maximised. Incorporating very broad ranges in the environmental drivers, potentially including also highly rare levels (Kreyling *et al* . 2014) (i.e. extreme events with low return-time), will add to the power of regression and help advancing our understanding of important ecological processes, specifically those that show a nonlinear response behaviour. Effects of extreme events, which are considered to be of disproportionate ecological importance (Jentsch *et al* . 2007) and not well captured in monitoring networks (Mahecha *et al* . 2017), may therefore also be better captured by gradient experiments.

In ecological experiments, experimental units are rarely identical (differences between individuals, spatial variability, legacies, etc.). Interspersion and replication are basic principles to deal with such random or spatially structured variation (Hurlbert 1984). Gradient designs do not allow for such a quantification of variance, as they lack groups of identical treatments. Of course, interspersion of treatment level combinations, that is, random assignment of experimental units to treatment levels, remains a prerequisite in order to avoid natural variation getting confounded with the focal environmental drivers. Still, pre-treatment observations can inform about spatial variability (compare to Before-After Control-Impact (BACI) designs, e.g. Green 1979). In case of high spatial variability among experimental units, our results concerning different noise levels imply that increasing total number of experimental units rather than increasing replication will improve overall prediction success. Generally, the basis of controlled experiments to keep all conditions constant except for the environmental driver(s) of interest, ideally resulting in only white noise left for true replicates, is also crucial in gradient designs. Here, any other driver potentially affecting the response in the given setting should therefore be kept constant, or specifically be analysed as another gradient.

Beyond the allocation of sampling units discussed here, sound detection of nonlinearities will require higher total sample sizes due to increased degrees of freedom than needed for simpler, linear relationships (see e.g. Babyak 2004). While gradient experiments outperformed replicated experiments in our simulations, prediction success and confidence always increased with total sample size. Consequently, a shift towards gradient designs will generally not lead

to reduced total numbers of samples needed, but an optimised use of available experimental units.

We conclude that gradient designs are a powerful tool for detecting patterns in ecological responses to continuous and interacting environmental drivers as they generally outperform replicated designs in terms of prediction success of known response surfaces independently of noise level and shape of the underlying response surface (linear, saturating, hump-shaped) for given, realistic total numbers of experimental units (total sample size). Improved mechanistic understanding of the underlying processes appear to be a reasonable expectation, because such designs allow for testing for the most probable relationship among the response variable and the environmental drivers by (multiple) hierarchical regression analysis. Gradient experiments should thereby improve the value of the experimental output for process-based modelling. Replicated designs remain the solution to test for differences among (categorical) groups, optimising local precision, or quantification of (stochastic) variation. However, for approaching nonlinearity and interacting environmental drivers in ecological responses, a paradigm shift to unreplicated multilevel gradient designs of experiments would be a major step forward.

Acknowledgements

This manuscript was developed and discussed during multiple workshops funded by the EU-COST-Action ES1308 'Climate Change Manipulation Experiments in Terrestrial Ecosystems (ClimMani)'. The SkyGas3D system used to provide data for the soil N_2O flux experiment was funded by NERC under the Macronutrients Consortium and we acknowledge assistance from Dr. Sylvia Toet. The laboratory microcosm system used to provide data for the ciliate experiment has been funded by F.R.S.-FNRS (grants 1.5.135.09F, 1.5065.11F, U.N035.16) and UCL (grants ARC 10-15/031, FSR, Move-In Louvain). JK acknowledges support by the research training group RESPONSE funded by the German Research Council (DFG Fi 846/8-1, DFG GRK2010). MB acknowledges support from the Austrian Science Fund (FWF). TMJ is a Move-In Louvain Marie Curie postdoc at UCL, and NS is Research Associate of the F.R.S.-FNRS.

Authorship

JK and KSL conceived the ideas; AS and JK designed the simulation experiment; PI and JRC designed the soil trace gas experiment and collected the data; NS and TM-J designed the ciliate experiment and collected the data; AS, JK and KSL analysed the data; JK, AS, and PI led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Data Accessibility Statement

The R-scripts for simulating the data used are found in the supporting information.

References

Babyak, M.A. (2004). What you see may not be what you get. A brief, nontechnical introduction to overfitting in regression-type models. Psychosom. Med., 66, 411– 421.

Beier, C., Beierkuhnlein, C., Wohlgemuth, T., Penuelas, J., Emmett, B., Körner, C. et al . (2012). Precipitation manipulation experiments – challenges and recommendations for the future. Ecol. Lett., 15, 899– 911.

Bertness, M.D. & Callaway, R. (1994). Positive interactions in communities. Trends Ecol. Evol., 9, 191– 193.

Box, G.E.P. & Wilson, K.B. (1951). On the experimental attainment of optimum conditions. J. R. Stat. Soc. Series B, 13, 1– 45.

ter Braak, C.J.F. & Prentice, I.C. (1988). A theory of gradient analysis. Adv. Ecol. Res., 18, 271– 317.

Burnham, K.P. & Anderson, D.R. (2010). Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, 2nd edn. Springer, New York, NY.

Collins, K. (2012) Tetrahymena thermophila. Methods in Cell Biology, pp. 452. Elsevier.

Connell, J.H. (1978). Diversity in tropical rain forests and coral reefs - high diversity of trees and corals is maintained only in a non-equilibrium state. Science, 199, 1302– 1310.

Cottingham, K.L., Lennon, J.T. & Brown, B.L. (2005). Knowing when to draw the line: designing more informative ecological experiments. Front. Ecol. Environ., 3, 145– 152.

Curtis, J.T. & McIntosh, R.P. (1951). An upland forest continuum in the prairie-forest border region of Wisconsin. Ecology, 32, 476– 496.

Dieleman, W.I.J., Vicca, S., Dijkstra, F.A., Hagedorn, F., Hovenden, M.J., Larsen, K.S. et al . (2012). Simple additive effects are rare: a quantitative review of plant biomass and soil process responses to combined manipulations of CO2 and temperature. Glob. Change Biol., 18, 2681– 2693.

Gill, R.A., Polley, H.W., Johnson, H.B., Anderson, L.J., Maherali, H. & Jackson, R.B. (2002). Nonlinear grassland responses to past and future atmospheric CO2. Nature, 417, 279– 282.

Green, R.H. (1979). Sampling Design and Statistical Methods for Environmental Biologists. Wiley, New York.

Grier, D.G. (2003). A revolution in optical manipulation. Nature, 424, 810– 816.

Grinnell, J. (1917). The niche-relationships of the California thrasher. Auk, 34, 427– 433.

Hare, R.D. (1965). Temporal gradient of fear arousal in psychopaths. J. Abnorm. Psychol., 70, 442– 445.

Helmlinger, G., Yuan, F., Dellian, M. & Jain, R.K. (1997). Interstitial pH and pO(2) gradients in solid tumors in vivo. High-resolution measurements reveal a lack of correlation. Nat. Med., 3, 177– 182.

Hurlbert, S.H. (1984). Pseudoreplication and the design of ecological field experiments. Ecol. Monogr., 54, 178– 211.

IPCC (2013) Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Jacob, S., Legrand, D., Chaine, A., Bonte, D., Schtickzelle, N., Huet, M. et al . (2017). Gene flow favours local adaptation under habitat choice in ciliate microcosms. Nat. Ecol. And Evol., 1, 1407– 1410.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2017). An introduction to statistical learning. With applications in R. 8th edn. Springer, New York, Heidelberg, Dordrecht, London.

Janssen, P.H.M. & Heuberger, P.S.C. (1995). Calibration of process-oriented models. Ecol. Mod., 83, 55– 66.

Jentsch, A., Kreyling, J. & Beierkuhnlein, C. (2007). A new generation of climate change experiments: events, not trends. Front. Ecol. Environ., 5, 365– 374.

Kelly, M., Bennion, H., Burgess, A., Ellis, J., Juggins, S., Guthrie, R. et al . (2009). Uncertainty in ecological status assessments of lakes and rivers using diatoms. Hydrobiologica, 633, 5– 15.

Kreyling, J. & Beier, C. (2013). Complexity in climate change manipulation experiments. Bioscience, 63, 763– 767.

Kreyling, J., Jentsch, A. & Beier, C. (2014). Beyond realism in climate change experiments: gradient approaches identify thresholds and tipping points. Ecol. Lett., 17, 125.

Larsen, K.S., Andresen, L., Beier, C., Jonasson, S., Albert, K., Ambus, P. et al . (2011). Reduced N cycling in response to elevated CO2, warming, and drought in a Danish heathland: Synthesizing results of the CLIMAITE project after two years of treatments. Glob. Change Biol., 17, 1884– 1899.

Levin, S.A. (1998). Ecosystems and the biosphere as complex adaptive systems. Ecosystems, 1, 431– 436.

Liu, J., Dietz, T., Carpenter, S.R., Alberti, M., Folke, C., Moran, E. et al . (2007). Complexity of coupled human and natural systems. Science, 317, 1513– 1516.

Mahecha, M.D., Gans, F., Sippel, S., Donges, J.F., Kaminski, T., Metzger, S. et al . (2017) Detecting impacts of extreme events with ecological in-situ monitoring networks. Biogeosciences, 14, 4255– 4277.

Matthews, S. & Power, C. (2002). Socio-economic gradients in psychological distress. A focus on women, social roles and work-home characteristics. Soc. Sci. Med., 54, 799– 810.

Moffitt, T.E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R.J., Harrington, H. et al . (2011). A gradient of childhood self-control predicts health, wealth, and public safety. Proc. Natl Acad. Sci. USA, 108, 2693– 2698.

Pennekamp, F. & Schtickzelle, N. (2013). Implementing image analysis in laboratory-based experimental systems for ecology and evolution: a hands-on guide. Methods Ecol. Evol., 4, 483– 492.

Pennekamp, F., Schtickzelle, N. & Petchey, O.L. (2015). BEMOVI, software for extracting behavior and morphology from videos, illustrated with analyses of microbes. Ecol. Evol., 5, 2584– 2595.

Piepho, H.P. & Bahn, M. (2017) Designing an experiment with quantitative treatment factors to study the effects of climate change. J. Agron. Crop Sci., 203, 584– 592. https://doi.org/10.1111/jac.12225.

Pugh, T.A.M., Muller, C., Arneth, A., Haverd, V. & Smith, B. (2016). Key knowledge and data gaps in modelling the influence of CO2 concentration on the terrestrial carbon sink. J. Plant Physiol., 203, 3– 15.

Richardson, A., Aubinet, M., Barr, A., Hollinger, D., Ibrom, A., Lasslop, G. et al . (2012). Uncertainty quantification. In Eddy Covariance: A Practical Guide to Measurement and Data Analysis. (eds M. Aubinet, T. Vesala, D. Papale). Springer, Netherlands.

Scheffer, M. & Carpenter, S.R. (2003). Catastrophic regime shifts in ecosystems: linking theory to observation. Trends Ecol. Evol., 18, 648– 656.

Schweiger, A.H., Irl, S.D.H., Steinbauer, M.J., Dengler, J. & Beierkuhnlein, C. (2016). Optimizing sampling approaches along ecological gradients. Methods Ecol. Evol., 7, 463– 471.

Shaw, M.R., Zavaleta, E.S., Chiariello, N.R., Cleland, E.E., Mooney, H.A. & Field, C.B. (2002). Grassland responses to global environmental changes suppressed by elevated CO_2. Science, 298, 1987– 1990.

Stejskal, E.O. & Tanner, J.E. (1965). Spin diffusion measurements. Spin echoes in the presence of a time-dependent field gradient. J. Chem. Phy., 42, 288.

Steury, T.D. & Murray, D.L. (2005). Regression versus ANOVA. Front. Ecol. Environ., 3, 356– 357.

Thompson, R.M., Beardall, J., Beringer, J., Grace, M. & Sardina, P. (2013). Means and extremes: building variability into community-level climate change experiments. Ecol. Lett., 16, 799– 806.

Whittaker, R.H. (1967). Gradient analysis of vegetation. Biol. Rev., 42, 207.

Yue, K., Fornara, D.A., Yang, W., Peng, Y., Peng, C., Liu, Z. et al . (2017). Influence of multiple global change drivers on terrestrial carbon storage: additive effects are common. Ecol. Lett., 20, 663– 672.

Zscheischler, J., Reichstein, M., Harmeling, S., Rammig, A., Tomelleri, E. & Mahecha, M.D. (2014). Extreme events in gross primary production: a characterization across continents. Biogeosciences, 11, 2909– 2924.