Global Change Experiments: Challenges and Opportunities

HANS J. DE BOECK, SARA VICCA, JACQUES ROY, IVAN NIJS, ALEXANDRU MILCU, JUERGEN KREYLING, ANKE JENTSCH, ABAD CHABBI, MATTEO CAMPIOLI, TERRY CALLAGHAN, CARL BEIERKUHNLEIN, AND CLAUS BEIER

Manipulation experiments are invaluable tools in global change ecology because they enable causal and process-based understanding. However, artifacts and inherent limitations can lead to misinterpretations. Across the wide range of approaches to set up such studies, we distill the main challenges associated with the imposed treatment(s), the spatial and time scale, proposing solutions and outlining the limitations in interpreting and extrapolating results. The inherent trade-offs between experimental realism (facilitating extrapolation) and control (facilitating the attribution of observed responses) resonate throughout this review. The focus on realism or control determines which issues become more important and how they should be handled. For example, covarying factors such as temperature and moisture can be explicitly separated to attribute effects more precisely but could also be left uncontrolled to increase realism. Ultimately, combining results across gradients of scale and control, including the use of "natural laboratories," stimulates fundamental understanding, enabling more confident predictions of responses to global change.

Keywords: artifacts, ecosystem, manipulation, methodology, scale

Manipulation experiments have become an increasingly important scientific tool in research focusing on understanding the ecological consequences of global changes such as climate change, nitrogen deposition, and biodiversity loss. The oldest of those is the Park Grass Experiment (Silvertown et al. 2006), which has been running for over 150 years, providing a wealth of data against a background of environmental change. In contrast to traditional field surveys and monitoring in natural and seminatural ecosystems, manipulation experiments enable researchers to specifically test hypotheses (Beierkuhnlein and Nesshöver 2006). Hundreds of ecological experiments are being initiated each year, varying widely in scale and design. Their ultimate goal is to acquire a fundamental understanding of ecological processes to enable the extrapolation of case-specific results. Unfortunately, flawed designs and experimental artifacts can lead to misinterpretations and may severely limit generalizations to wider contexts. Therefore, artifacts should be identified (cf. Englund and Cooper 2003, Hillebrand and Matthiessen 2009, Beier et al. 2012) and then avoided if possible.

Scientists may be aware of the potential problems and eager to improve the methodology but unable to resolve this issue because of financial or logistical constraints. For example, it is well known that passive open-top chambers are an imperfect warming technique (Marion et al. 1997,

Shaver et al. 2000) and that the small size of the chambers may lead to responses to the treatment being outside the plot because of rhizomatous/stoloniferous growth in long-lived clonal plants. Nevertheless, these chambers require no power and hardly any maintenance and are therefore highly practical in warming ecosystems in remote and inaccessible areas. In such cases, the challenge is then to be aware of the drawbacks of the applied method, to formulate hypotheses that reflect these drawbacks, and to extrapolate findings cautiously. On occasion, however, researchers may not realize that the designs and infrastructures they use are reason for concern or that solutions exist to solve or circumvent artifacts. Knowledge on experimental artifacts and workarounds is scattered throughout the literature, often reaching only part of the scientific community and not crossing interdisciplinary boundaries. A molecular biologist studying the model plant *Arabidopsis thaliana* may, for example, learn from ecologists who determine which growth conditions (soil volume, light quantity and quality, etc.) are required for normal plant functioning (Poorter et al. 2012). Or alternatively, results from molecular studies demonstrating that gene expression depends on whether a stress is imposed in an abrupt or a gradual manner (Ambrosone et al. 2011) could explain why experimentally imposed environmental changes may initially act as a disturbance (Wookey 2008) and stimulate global change researchers to be cautious when extrapolating short-term results.
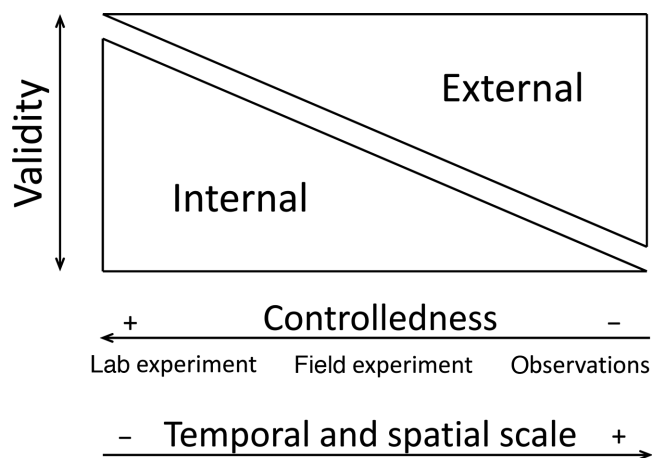
Figure 1. A schematic representation of the trade-off between **internal validity** *(being able to attribute a change to a specific factor) and* **external validity** *(the confidence with which results can be generalized).*

To transcend the very detailed difficulties found in specific domains, we address those issues that are relevant to a wider range of global change ecologists, using specific examples to illustrate and clarify generic problems. The focus is mainly on terrestrial ecosystems, although examples from aquatic systems are used when appropriate. We consider major choices and challenges that experimentalists are confronted with, discuss whether they can significantly affect the studies' conclusions, and if that is the case, we suggest ways forward. Some artifacts may be circumvented; other issues are inherent and cannot be avoided altogether. Even then, choices can be made to ensure that experimental artifacts do not become detrimental to the conclusions of a study. Our aim is therefore to refine the general approach, because fundamentally, the manipulation experiment is an invaluable tool for identifying the processes underlying change and for gaining a process understanding sufficient to drive models of ecosystem response to global change.

The trade-off between internal and external validity

As we will illustrate throughout this article, fundamental trade-offs exist between external and internal validity (figure 1). *External validity* relates to the experimental realism and the applicability of a result to other conditions or systems. However, more stringent control over variables stimulates *internal validity*—that is, the confidence that the treatment is the cause of the observed effects. Indeed, allowing more processes and factors to vary freely in an experiment makes it reflect reality better but also renders attributing changes to specific processes more difficult.

In this regard, it is crucial to recognize that the questions that can be answered by simplified experiments or those involving more complexity will be—and have to be— different. More control by blocking realistic but unwanted sources of variation allows researchers to uncover fine-scale

or specific mechanisms, whereas a better representation of natural complexity allows for a more straightforward extrapolation of results and trends but also less understanding of the underlying mechanisms. The position of the study on the internal-to-external-validity gradient (figure 1) therefore inherently constrains it. Failing to recognize the limitations leads to a number of pitfalls, of which the most important is the inappropriate extrapolation of findings. However, *a priori* awareness of the strengths and weaknesses of an experiment also generates opportunities. It allows for more streamlined hypotheses and well-grounded conclusions and also enables devising explicit combinations with other studies, creating added value by transcending the limitations posed by single experiments.

Treatment choices and resulting challenges

Global change involves the interacting drivers of ecological responses, both environmental and biotic, that can generate nonadditive effects through positive and negative feedback. The Jasper Ridge Global Change Experiment illustrates this well (Shaw et al. 2002). Exposing annual grasslands to elevated carbon dioxide (CO_2), increased temperature, increased precipitation, and nitrogen deposition revealed increases in net primary production in all single-factor treatments. Stimulations were stronger when multiple factors were combined, except when these combinations included elevated CO_2, which substantially suppressed the positive effects of other global changes. In a meta-analysis of 171 studies that manipulated two or more human stressors in marine and coastal systems, Crain and colleagues (2008) found that in the majority of studies, either synergistic (36%) or antagonistic (38%) interactions were observed. Extrapolation from studies in which only one single factor is altered should therefore be done cautiously. Nevertheless, single-factor experiments can improve mechanistic understanding under tightly framed hypotheses because their low complexity renders isolating specific processes easier (figure 2).

Of course, multifactor experiments usually include the single factors, but at the expense of either more total plots (increasing cost and workload) or a limited number of replicates (decreasing statistical power). Testing a complete set of projected future conditions is therefore not feasible. It is important to prioritize experiments that manipulate those factors of ecological relevance that are most likely to change, preferably in full-factorial combination, so that the underlying mechanisms of interactions can be tested for and used to evaluate models. The latter is clearly needed to help parameterize current models that often have difficulties in dealing with interacting global change factors (Luo et al. 2008).

Treatment choices also involve decisions regarding the nature of the changes (e.g., gradual changes versus events; Jentsch et al. 2007); the timing of the treatments, including the possible bias of applying treatments only in some seasons (cf. Bokhorst et al. 2009); the type of systems to consider, with the known underrepresentation of (sub)tropical regions
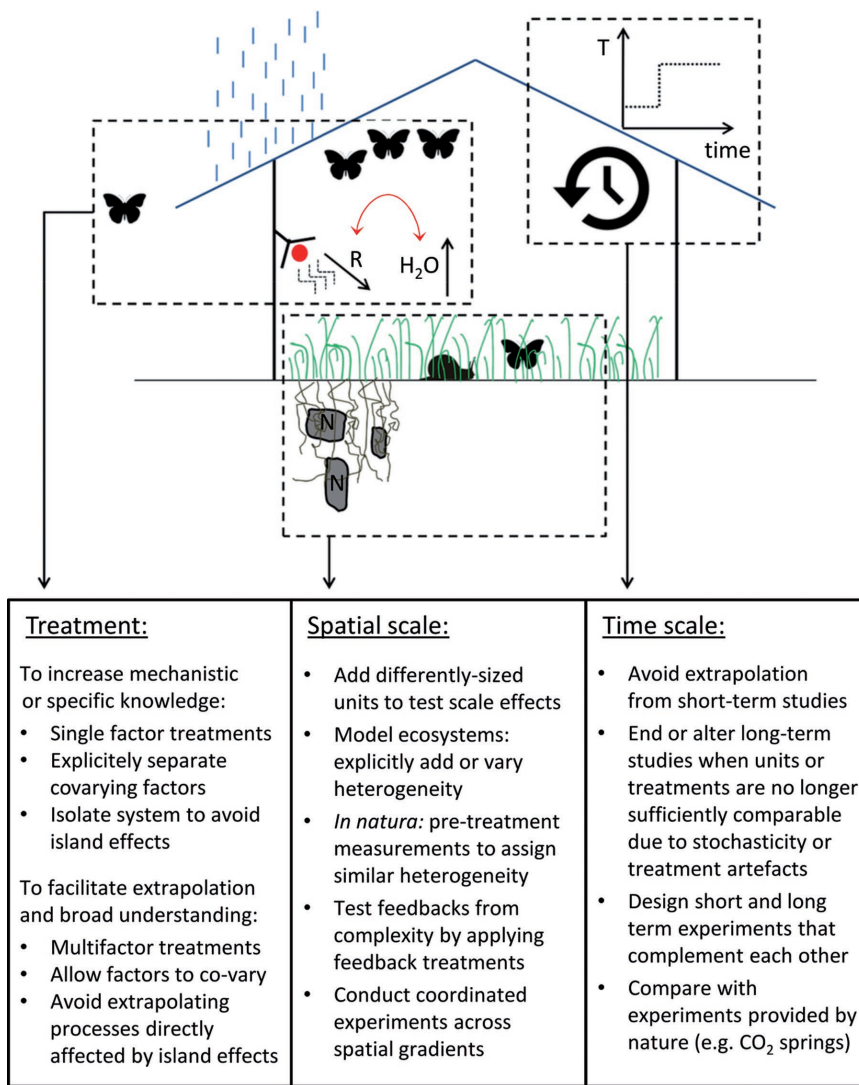
Figure 2. An overview of the main recommendations for issues discussed in our study, visualized by means of an experiment using rainout shelters and infrared heaters. Depicted and discussed are choices and challenges related to the treatment (including covarying factors and island effects), the spatial scale, and the time scale.

New or alternative approaches. Regression/gradient approaches have been suggested as an alternative to the widespread use of replicated analysis-of-variance (ANOVA) designs in order to maximize the number of interactions and levels testable (Kreyling et al. 2014). Such experiments would result in response surfaces to be analyzed by regression techniques and might reveal thresholds and tipping points better than classical replicated setups do, which is particularly important in extreme event research.

Furthermore, promising new methods have been developed to help elucidate the process dynamics and sensitivities related to biotic and/or abiotic feedback. For example, plant–soil feedback treatments can be established by growing a particular plant community long enough to alter the soil characteristics through mutual plant–soil interaction and subsequently using the conditioned soil to grow a different plant community (Manning et al. 2008, Meijer et al. 2011). Such plant–soil feedback studies generate valuable insights into the processes driving ecosystem succession and invasion (Kulmatiski et al. 2008). They can also provide additional explanations for observed biodiversity-functioning relationships (Maron et al. 2011) and sensitivity to drought stress (Meijer et al. 2011). Abiotic feedback treatments, although still rare, could reduce the uncertainties related to the direction and magnitude of biotic feedback to global changes, such as the changing sensitivity of photosynthesis to CO_2 with rising temperatures demonstrated by Milcu and colleagues (2012).

(cf. Beier et al. 2012); and the methodology (e.g., biodiversity experiments using random or nonrandom species extinction; Gross and Cardinale 2005). Discussing all of these choices is beyond the scope of this study and may not be really meaningful because a breadth of sensible approaches has inherent value. Nevertheless, it is only by bringing these experimental approaches and treatments together in syntheses that deeper understanding across the scales can be achieved. Such synthesis work is facilitated if a standardized set of measurements is employed (Vicca et al. 2012) and if enough experimental detail is reported (Hillebrand and Gurevitch 2013). Moreover, combining different areas of expertise within the same experiment (a multidisciplinary approach) allows for integrated syntheses, but these are more complicated to achieve across separate projects.

Finally, making use of "natural laboratories" offers a number of unique possibilities. Geologically active locations harbor areas where CO_2 is released (CO_2 springs) or where soils and streams are warmed. These could be used to test the long-term effects of CO_2 addition or warming in a natural setting. For example, O'Gorman and colleagues (2014) identified stream catchments in several Nordic regions where temperatures differed more than 15 degrees Celsius (°C) regionally, whereas the water chemistry remained stable. The coldest streams in each area could therefore serve as a benchmark to assess warming-induced effects on community structure, biodiversity, and ecosystem processes. Moreover, embedding highly controlled experiments within such settings could create unique possibilities, such as studying the effects of long-term acclimation to warmer

conditions on the sensitivity to temperature pulses (e.g., by heat waves).

Covariation between factors. In full-factorial experiments, isolating the single-factor effects of the imposed treatments is challenging, because manipulating one factor can also change others (Shaver et al. 2000). For example, providing increased levels of atmospheric CO_2 to plants usually increases tissue temperatures, because the lower stomatal conductance decreases latent heat loss (Leuzinger and Körner, 2007). In this case, the temperature increase is mostly limited, but in the case of drought experiments that trigger stomatal regulation, the canopy can be warmed by several degrees (De Boeck et al. 2011). Consequently, the response to such treatments is not solely a direct result of water deprivation but also of responses to co-occurring increases in temperature. Rising temperatures with increasing drought are a logical consequence of atmospheric feedback (De Boeck and Verbeeck, 2011). The same interdependence also causes warming treatments to dry out the soil and vegetation unless precipitation is increased. Similar examples can be found in aquatic systems, where eutrophication affects phytoplankton communities not only directly but also through increased light attenuation and carbon depletion (Reynolds 1998). The consequence is that it becomes more difficult to attribute changes to a specific factor. In other words, interdependences increase the external validity at the expense of the internal validity of the study.

Covarying factors can be separated using explicit designs. One such example is a study by Hautier and colleagues (2009) on nitrogen enrichment and plant species loss, which distinguished between the direct effects of increased nitrogen inputs on belowground competition and the indirect aboveground effects of light limitation in the understory of vegetation as a consequence of the increased growth of dominant plants. To fully separate the two factors, the researchers provided extra light inside the canopy in some treatments (a single-factor light effect or in combination with nitrogen fertilization). Such solutions are warranted if researchers want to keep full control of major factors in order to precisely attribute responses—that is, when very specific hypotheses are being tested (high internal validity required). In other cases, it may be more appropriate to aim for high external validity by allowing naturally occurring interdependence to develop (figure 2). Increased awareness of covarying factors and their quantification through measurements should always be pursued, because it helps to avoid attributing responses to the wrong factor (i.e., being right for the wrong reason) and to understand the direct versus indirect mechanisms of treatment responses.

Island effects. Island effects are another mechanism that can influence imposed treatments. These are a direct result of the inherent limitations regarding scale, making outdoor experiments islands in a wider landscape that remains unchanged. Whenever the conditions in the study area are altered, these may be perceived as more or less attractive to animals, changing the interactions between primary producers and animals. This could be relevant in the aforementioned naturally warmed stream catchments (O'Gorman et al. 2014), because these might attract more birds or other mobile organisms than they would if every catchment in the area were warmed. In terrestrial systems, effects can range from nest building under rainout shelters to preferential grazing in plots, because, for example, nitrogen was added (Latimer and Oetting 1999) or even because a path through the vegetation to reach the plot can facilitate herbivore visits (Cahill et al. 2001). Such island effects can significantly affect the conclusions of a study. For example, Klady and colleagues (2011) observed that floral sexual reproduction increased when arctic plants were exposed to warming applied by means of open-top chambers. Such chambers are, however, warm and calm islands in an ocean of cool and windy surroundings, and flowering plants inside the warmed plots would therefore be inherently more attractive to insects than those in the surrounding landscape. Such an attraction bias would not exist if the whole landscape had been warmer. In this example, island effects have therefore potentially led to an exaggeration of the warming effect on floral sexual reproduction. For cases such as this, in which a variable of interest (pollination) is directly distorted by the experimental manipulation, observational studies using temporal or spatial variation should explicitly be used as a comparison.

Island effects can also lead to a reduction of the effect size by pushing mesocosms closer to the conditions of their surroundings, both in the abiotic and the biotic sense. For example, when testing legacy effects of a drought period, the seed rain from outside would not reflect the quantities and species origins of the seeds (and possibly changes in the expression of drought-related traits; Herman et al. 2012) compared with when the whole landscape would have been subjected to the same drought. This means that the community changes less or in a different way than it would following a landscapewide drought. Similarly, abiotic island effects can also serve to render imposed treatments less extreme than they would be naturally, such as when water reaches drought plots via groundwater or surface-water inflow or because atmospheric feedback is missing. In other words, such effects tend to dilute the experimentally imposed treatments, which is especially important to consider when extrapolating findings.

Some of these issues could be overcome by increased isolation of the communities from the outside (figure 2). In aquatic environments, enclosing systems is almost inevitable, whereas in terrestrial systems, measures can range from fencing to avoid grazing and disturbance to using full enclosures such as greenhouses, which represents different amounts of control (cf. figure 1). However, isolating the ecosystem also implies cutting off a number of natural community processes, such as limiting the number of trophic groups and the exchange of propagules. Isolation may

therefore become an artifact whenever processes such as colonization and herbivory are important (e.g., Kongstad et al. 2012), rendering the potential for making long-term predictions from systems in isolation challenging.

One solution to such issues is to artificially include biotic changes at the larger scales. In the case of drought experiments, this could encompass adding seeds in quantities that reflect drought effects in the entire landscape. Obviously, this requires *a priori* knowledge, which could be gained from observational studies or many small experiments. A solution such as this remains a simplification of natural processes, but at least it acknowledges processes beyond the island that is the experiment in order to increase its realism and potential for extrapolation. Comparisons with natural systems could indicate how important feedback and mechanisms not represented in treated "islands" may be. It is, in any case, necessary to quantify which treatment levels are ultimately realised, because, as we have demonstrated, island effects can alter the intended treatment. Detailed quantification of the realised treatments was also recommended regarding covarying factors. Such quantification would not only allow for a more correct interpretation of single experiments but would also facilitate comparison between experiments (Vicca et al. 2012).

Spatial scale

The spatial scale of the experiment is a fundamental choice that directly influences its internal and external validity (figure 1). This implies that by using an inappropriate scale relative to the studies' aims, either too much or not enough biotic complexity and spatial heterogeneity are taken into account, undermining the conclusions (cf. Bommarco and Banks 2003). Therefore, the spatial scale needs to be carefully matched to the processes and ecosystems under study. For experiments highly focused on specific processes or mechanisms, such as the role of the spatial patterning of plant species in invasion resistance (Yurkonis et al. 2012), the scale can be kept limited, because in such studies, high internal validity (i.e., low "noise" of uncontrolled aspects or processes) is key. When external validity gains importance because we want to understand how given small-scale responses can be generalized or lead to larger-scale responses, the scale of the experiment has to be large enough to feature higher levels of complexity. Challenges are different for *in natura* and assembled (model) ecosystems, however.

Artificially assembled systems. Model systems strive for large internal validity and will decouple many of the interactions between the scale of the mesocosm and the complexity and heterogeneity. In plant ecology, this usually involves even-aged individuals planted in regular designs on homogenized soil (e.g., Jentsch et al. 2011). Enlarging the surface area of plots in such experiments implies multiplying homogeneous blocks and therefore does not increase complexity and heterogeneity as such. In other fields of global change research that use nonnatural systems, such as aquatic mesocosms

(Englund and Cooper 2003), a similar reasoning could be used. Enlarging plots in these highly controlled experiments would only make sense if this significantly affects key response variables. Factors that limit how small plots can be made in terrestrial systems are the well-known edge effect and also soil volume. Indeed, physical soil space can have a direct impact on plant–plant interactions and growth (Dimitrakopoulos and Schmid 2004), meaning that responses and relationships between, for example, biodiversity and ecosystem functioning may be distorted if the soil volume is too small. Detailed information on minimum substrate requirements can be found in Poorter and colleagues (2012). In aquatic systems, both volume and the surface area-to-volume ratio may be relevant (Englund and Cooper 2003, Spivak et al. 2011), especially if more than one trophic level is involved. Petersen and Englund (2005) explored the possibilities of defining the minima for preserving "effective scales"—that is, to maintain the functions of mesocosms in question with respect to natural systems.

Like the minima for plot or mesocosm dimensions, the minimum amounts of resource heterogeneity may also be required. Studies have shown that heterogeneity in the distribution of nutrients can modulate interactions between climatic changes and biodiversity (Garcia-Palacios et al. 2012), with biodiversity affecting functioning more in heterogeneous than in homogeneous systems (Tylianakis et al. 2008, Cardinale 2011). In stream systems, Lancaster and Belyea (1997) showed that site heterogeneity affects the persistence of populations in the face of perturbations. Such results suggest that the extrapolation potential of experiments which are artificially homogeneous is likely limited. In assembled systems, it may therefore generally be advisable to actively include sources of heterogeneity in a controlled manner, such as patchy nutrient and water supply (cf. Mommer et al. 2012), if these are considered important for the processes being studied (figure 2). However, some solutions may surpass practical feasibility. In those cases, applying the same experiments also *in natura* could give information about the generality and extrapolation potential of the findings (cf. figure 1). Alternatively, using intact samples of established ecosystems keeps most heterogeneity near its natural levels. Obviously, this is not possible in all systems, but bringing monoliths of grasslands or other small-statured vegetation into climate-controlled Ecotrons (Milcu et al. 2014) couples an amount of ecosystem naturalness with high levels of environmental control.

In natura experiments. *In natura* experiments are often limited in scale by practical considerations, because many manipulations are difficult or even impossible to carry out across large areas. Even when the manipulation is provided for free by natural laboratories, the large scale can pose challenges regarding workload. It would therefore be useful to establish minimum plot sizes—that is, the minimum of a range in which processes are independent of scale (cf. Roscher et al. 2005 on biodiversity–ecosystem functioning research).

However, such knowledge is scarce and often very specific. Including plots or mesocosms of different spatial scale in the same experiment would help to elucidate whether scale affects the results and where (and whether) scale saturation occurs, facilitating interpretation and increasing comparability (figure 2). This would also be valuable in artificially assembled ecosystems. Such a multiscale approach has mostly been used in aquatic systems so far (see Englund and Cooper 2003 for a review).

Whatever the size of the plots or mesocosms, the heterogeneity of *in natura* systems in question should be captured adequately, avoiding results being skewed by overrepresenting some patches and underrepresenting others. This requires using an adequate number of replicates (i.e., more if heterogeneity is more variable) and a well-reasoned distribution across the landscape and across treatments. Evidently, heterogeneity should be similar across treatments and controls when the experiment is initiated so that response differences are not caused by differences in heterogeneity between treatments. Pretreatment assessment or characterization of physical, chemical and/or biological heterogeneity can be very valuable in this regard (figure 2), which may sometimes be fairly straightforward (such as temperature; cf. Scherrer and Körner 2010) but sometimes more difficult to quantify *a priori* (such as nutrients). Selecting areas in close proximity with similar species composition may in such cases be used as an indication of comparable abiotic conditions. However, this search for relatedness and comparability can lead to the pitfall of spatial autocorrelation and pseudoreplicates and should be monitored.

Biological complexity. *Biological complexity*, defined here as the whole of biotic interactions within an ecosystem—including food-web structure, community composition, and their interactions—changes with spatial scale, as was mentioned before. Like heterogeneity and direct-scale effects, it can interact with ecosystem responses to global changes. For example, chains of interactions can result in feedback loops that can have stabilizing (negative feedback) or destabilizing (positive feedback) effects. The so-called Janzen Connell effect of reduced success of establishment of conspecifics in the proximity to established adult individuals due to increased seed predator/pathogen loads is an example of localized negative feedback (van der Putten et al. 2013). In contrast, the presence of dwarf shrubs can aid the establishment of tree seedlings at alpine and sub-Arctic treelines by diverting herbivore activity (Grau et al. 2013). Including such interactions will provide a less fragmented understanding of the mechanisms behind a wide range of ecological questions.

Complexity is inherently included in experiments *in natura*, although within limits (Kreyling and Beier 2013). Sometimes, these limits result mainly from active measures such as fencing, but usually, they stem from the inability to extend the manipulation to the entire landscape (cf. island effects). A choice of mesocosm size that correctly represents spatial heterogeneity will likely adequately represent smaller-scale complexity such as species richness, evenness, and community age structure. However, complexity also entails interactions between primary producers and consumers, and because herbivores vary widely in size, not all interactions can be captured within a plot or mesocosm in many systems. Some experiments artificially induce such interactions to test specific hypotheses (figure 2), such as those by Stevnbak and colleagues (2012), who applied insect herbivory as a treatment by means of meshbags to test cascadal effects on belowground processes. Including all aspects of complexity can only be approached by means of observation studies on long-time scales, however. For experimentalists, it is important to gain *a priori* knowledge of the potential interactions and feedback affecting the ecological processes or organisms in question (through preliminary trial experiments or literature) to establish to which extent the results can be extrapolated. Consequently, long-term observations that identify change should be integrated with experiments that can explain change and extrapolate change in time and space (Callaghan et al. 2013).

In assembled systems, many aspects of complexity are not included (e.g., colonization or herbivory) in order to safeguard high internal validity. Experiments could gain in realism if aspects such as more natural demography (particularly in regions characterized by individuals with long life spans), colonization rates (e.g., by providing natural propagule influx), or intraspecific variation were introduced, albeit at the cost of internal validity.

An overarching recommendation. Our overarching recommendation regarding scale issues is to increase the scale of the experiment instead of the plot/unit. This can be achieved by combining experiments and gradient studies by conducting the same experimental manipulations along climatic gradients (Beier et al. 2012), with manageable cost, work, and space requirements for each site (figure 2). This reconciles safeguarding internal validity at the local scale with increasing external validity across all sites. The trends and relationships found across study sites would substantially increase confidence that these are not merely a product of local conditions and would therefore improve the potential for extrapolation. At the same time, locations where diverging trends are found can reveal specific mechanisms and place limits on the generality of uncovered trends and processes (e.g., Sowerby et al. 2008). The framework proposed by Dunne and colleagues (2004) offered ways to interpret differences between similar treatments along a spatial gradient. Asking the same research question in studies differing in realism, such as along the gradient in figure 1, could work in a similar way: identifying robust mechanisms when similar results are found, or, when results differ, marking areas where improved process understanding is needed.

Time scale

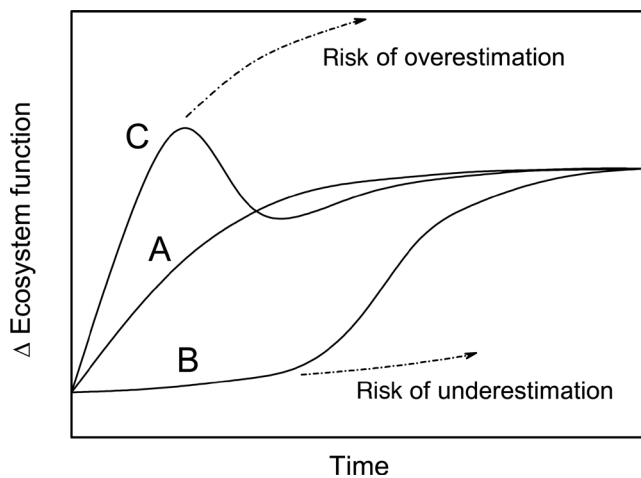Whether the duration of an experiment can be deemed "sufficient" depends on the hypotheses the experimental

Figure 3. An illustration of an incorrect extrapolation derived from short-term responses, because courses of deviation in ecosystem function triggered by an imposed global change can differ: (a) gradual response, (b) delayed response, and (c) overshoot scenario followed by stabilization. All responses ultimately align in this example.

study aims to test and the amount of extrapolation envisaged. Short-term experiments can be important for process understanding and identifying thresholds and tipping points. However, examples from CO_2 research have taught us that extrapolating from short experiments may be misleading, with short-term conclusions on the stimulating effect of CO_2 on growth and productivity (Ainsworth and Long 2005) being contested by observations following a longer period of treatment (Norby and Zak 2011). Part of the problem here may be that many aspects of global change manifest slowly and incrementally in time, whereas experiments are typically imposed in a stepwise manner.

Stepwise treatments. In a modeling study, Luo and Reynolds (1999) showed that ecosystem carbon sequestration and nutrient demand gradually increased with gradually rising CO_2 levels, whereas a stepwise CO_2 increase caused an abrupt yet short-lived increase in both variables, reminiscent of the studies referred to earlier (Ainsworth and Long 2005, Norby and Zak 2011). A similar response pattern has been observed empirically by Klironomos and colleagues (2005), who found that the diversity and functioning of arbuscular mycorrhizal fungi were irresponsive when the plant–soil system was exposed to a gradual CO_2 increase, whereas an abrupt increase in CO_2 levels caused significant yet short-lived changes. These and other results suggest that stepwise changes could be considered a perturbation, to which systems cannot acclimate in the short term (Wookey 2008, Knapp et al. 2012).

Obviously, there are few alternatives to step-change experiments. The challenge is then to correctly assess and more fully exploit their potential. To achieve this, a number of

recommendations can be made. First, experiments can be specifically designed to assess to which extent stepwise treatments differ from treatments imposed more gradually (cf. Hui et al. 2002, Ambrosone et al. 2011). Alternatively, observations from long-term monitoring studies (e.g., tree ring analysis, ecosystem stocks, and phenological records) and studies involving natural gradients could be used as comparison (figure 2). For example, the FORHOT experiment in Iceland uses natural increases in soil temperature caused by volcanic activity to study soil-warming effects (O'Gorman et al. 2014). Because these hot spots are variable in time and emerge in new places, researchers can compare how short-term (approximately 5 years) and longer-term (more than 30 years) effects differ and therefore filter out perturbation-induced transient responses resulting from the stepwise change. The mechanistic understanding gained can then be used to inform and correct existing models to better predict responses to gradual scenarios (Luo et al. 2011).

Short and long term studies. The stepwise nature of imposed treatments is not the only reason to be cautious in extrapolating from short-term studies. For example, the importance of species richness may be underestimated if inferred from short-term data, because the effects of biodiversity loss can escalate through time because important plant–soil feedback develops only slowly (Eisenhauer et al. 2012). Similarly, the negative effects of the abundance of the invasive species *Orconectes rusticus* (rusty crayfish) in a Wisconsin lake were only found after nearly 20 years of monitoring (Dodds et al. 2012). The challenge is therefore to understand how short-term and potentially transient treatment effects may translate into long-term changes or how an apparent lack of responses in the short term may progress or cascade through the system, leading to long-term changes and adaptations (Shaver et al. 2000, Jentsch et al. 2011, Knapp et al. 2012). On the basis of our current knowledge, long-term responses often cannot be derived from short-term effects (figure 3), which therefore warrants experiments on longer time scales (figure 2).

An advice to simply conduct experiments longer is not a miracle solution, however. Apart from logistical constraints, one complicating factor is that the amount of variation will increase the longer the experiment is running. This can occur because random events such as disease, plagues, etc. or stochastic trajectories of succession (Kreyling et al. 2011) may make replicates in a study less comparable (i.e., a loss of internal validity). The degree to which communities are affected by treatments should therefore be quantified, such as by means of similarity measures. The decision can then be made to end the experiment (if communities are no longer sufficiently comparable) or to shift the focus to a more individualistic perspective at the level of communities (figure 2). Furthermore, unavoidable imperfections in the experiment, such as island effects and other artifacts, may be proliferated as the length of the treatment increases (i.e., a loss of external validity). Researchers therefore need to be aware of the

drawbacks of the setup and techniques used, which is often not readily apparent.

Ultimately, short- and long-term experiments are both needed to link short- and long-term responses, and this requires experiments dealing with the same question along the gradients of scales. More specifically, we make the following suggestions: First, if an experiment focuses on short-term responses, the signal from the imposed environmental change must be strong enough to be detected against the natural intra- and interannual variation. If this signal-to-noise ratio is low, short experiments run the risk of concluding that "no changes are detected," leading to erroneous extrapolation. This logically leads to the next recommendation: Design short- and long-term experiments that complement each other (figure 2). For example, extreme events often act like a disturbance, allowing two complementary strategies to be adopted: (1) a short-term experiment focusing on the processes related to the extreme event and its immediate effects on the ecosystem and (2) a long-term experiment to evaluate fundamental changes in the ecosystem structure and functioning, such as through changes in biodiversity and community composition (Kreyling et al. 2011) or food-web complexity (Byrnes et al. 2011). This would help develop the knowledge needed to predict long-term effects based on shorter-term observations and which could be verified in the same setup. Finally, compare the findings with experiments provided by nature, in which acclimation to the local conditions has developed across decades. This could help avoid making unjustified extrapolations, as was illustrated by the discrepancy between the early conclusions of the substantial effects of CO_2 on primary production from Free Atmospheric CO_2 Enrichment (FACE) experiments (cf. Ainsworth and Long 2005) and the long-term responses found in trees located closer and farther from natural CO_2 springs showing weak or insignificant responses to CO_2 (cf. Hättenschwiler et al. 1997). These recommendations again illustrate how important awareness of the study's limitations and possibilities (cf. figure 1) is in avoiding several pitfalls and highlight the opportunities that arise when combining experiments across scales.

Conclusions

To enable an understanding of ecosystem responses to global change and making predictions that are valuable also to local stakeholders confronted with global changes that trigger very site-specific responses (Callaghan et al. 2013), we need experiments to test specific hypotheses (Beierkuhnlein and Nesshöver 2006). As we have argued throughout this article, the answers that single experiments can provide are confined by the trade-off between internal and external validity (figure 1). The experimental challenges often differ for studies on opposite sides of the gradient. Highly controlled studies aiming to uncover specific processes could, for example, gain from explicitly separating covarying factors (to attribute responses more precisely) and sticking to short time frames (to keep maximum control over treatments),

whereas opposite recommendations would benefit studies targeting realism and broad understanding. Some recommendations, such as including additional experimental units of different size to reveal scale dependencies, are relevant to any study, however.

Multiple problems could be remedied by combining studies. First, this could encompass the combination of studies across spatial gradients, using coordinated experiments. This would facilitate the identification of broad trends (when effects are similar across sites) and the factors overruling these trends (when the effects differ on one or more sites) (Dunne et al. 2004, Fraser et al. 2013, Kreyling et al. 2014). Second, studies of different duration can be combined, such as short-term (imposed) warming experiments and studies on locations where natural warming (e.g., in geothermal streams) has been present for decades (O'Gorman et al. 2014). Third, bringing together experiments carried out across the gradient of internal to external validity would reconcile both highly specific and broader understanding.

These recommendations would help identify overarching patterns and strengthen process-based models, improving both our understanding of global change impacts and the confidence of our predictions on ecosystem responses. Such an approach is to be preferred above too simplistic generalizations, such as space-for-time substitution (Johnson and Miyanishi 2012) or universal response functions such as temperature–soil respiration sensitivity equations (Davidson et al. 2006). Finally, process or ecosystem models that are strongly integrated with experiments could provide added value as a means of qualifying the hypotheses behind a given experiment, guiding the scenarios and scales applied, and extrapolating the results (Lee and Mishurov 2013). The complexities and potentially relevant scenarios and scales exceed our experimental capacities, which is why experiments should serve as well-chosen examples providing the understanding necessary to extrapolate the case-specific results. Integration of models and experiments is still lacking, however.

Designing and carrying out experiments are clearly fraught with both challenges and choices. The ways forward suggested here are relevant to most global change experiments, but the importance of each individual issue will depend on the specificities of the experiment and the nature of the system under study (e.g., terrestrial or aquatic). Although challenges remain, applying our recommendations should lower the probability of incorrect extrapolation and overstating conclusions from experiments and should lead to a better understanding and prediction of ecosystem responses to global change.

References cited

Ainsworth EA, Long SP. 2005. What have we learned from 15 years of free-air CO_2 enrichment (FACE)? A meta-analytic review of the responses of photosynthesis, canopy properties, and plant production to rising CO_2. New Phytologist 165: 351–372.

Ambrosone A, Costa A, Martinelli R, Massarelli I, De Simone V, Grillo S, Leone A. 2011. Differential gene regulation in potato cells and plants upon abrupt or gradual exposure to water stress. Acta Physiologiae Plantarum 33: 1157–1171.

Beier C, et al. 2012. Precipitation manipulation experiments: Challenges and recommendations for the future. Ecology Letters 15: 899–911.

Beierkuhnlein C, Nesshöver C. 2006. Biodiversity experiments: Artificial constructions or heuristic tools? Progress in Botany 67: 486–535.

Bommarco R, Banks JE. 2003. Scale as modifier in vegetation diversity experiments: Effects on herbivores and predators. Oikos 102: 440–448.

Bokhorst S, Bjerke JW, Tömmervik H, Callaghan TV, Phoenix GK. 2009. Winter warming events damage sub-Arctic vegetation: Consistent evidence from an experimental manipulation and a natural event. Journal of Ecology 97: 1408–1415.

Byrnes JE, Reed DC, Cardinale BJ, Cavanaugh KC, Holbrook SJ, Schmitt RJ. 2011. Climate-driven increases in storm frequency simplify kelp forest food webs. Global Change Biology 17: 2513–2524.

Cahill JF Jr, Castelli JP, Casper BB. 2001. The herbivory uncertainty principle: Visiting plants can alter herbivory. Ecology 82: 307–312

Callaghan TV, et al. 2013. Ecosystem change and stability over multiple decades in the Swedish sub-Arctic: Complex processes and multiple drivers. Philosophical Transactions of the Royal Society B 368 (art. 20120488).

Cardinale BJ. 2011. Biodiversity improves water quality through niche partitioning. Nature 472: 86–89.

Crain CM, Kroeker K, Halpern BS. 2008. Interactive and cumulative effects of multiple human stressors in marine systems. Ecology Letters 11: 1304–1315.

Davidson EA, Janssens IA, Luo Y. 2006. On the variability of respiration in terrestrial ecosystems: Moving beyond Q10. Global Change Biology 12: 154–164.

De Boeck HJ, Verbeeck H. 2011. Drought-associated changes in climate and their relevance for ecosystem experiments and models. Biogeosciences 8: 1121–1130.

De Boeck HJ, Dreesen FE, Janssens IA, Nijs I. 2011. Whole-system responses of experimental plant communities to climate extremes imposed in different seasons. New Phytologist 189: 806–817.

Dimitrakopoulos PG, Schmid B. 2004. Biodiversity effects increase linearly with biotope space. Ecology Letters 7: 574–583.

Dodds, et al. 2012. Surprises and insights from long-term aquatic data sets and experiments. BioScience 62: 709–721.

Dunne JA, Saleska SR, Fischer ML, Harte J. 2004. Integrating experimental and gradient methods in ecological climate change research. Ecology 85: 904–916.

Eisenhauer N, Reich PB, Scheu S. 2012. Increasing plant diversity effects on productivity with time due to delayed soil biota effects on plants. Basic and Applied Ecology 13: 571–578.

Englund G, Cooper SD. 2003. Scale effects and extrapolation in ecological experiments. Advances in Ecological Research 33: 161–213.

Fraser LH, et al. 2013. Coordinated Distributed Experiments: An emerging tool for testing global hypotheses in ecology and environmental science. Frontiers in Ecology and the Environment 11: 147–155.

Garcia-Palacios P, Maestre FT, Bardgett RD, de Kroon H. 2012. Plant responses to soil heterogeneity and global environmental change. Journal of Ecology 100: 1303–1314.

Gross K, Cardinale BJ. 2005. The functional consequences of random vs. ordered species extinctions. Ecology Letters 8: 409–418.

Grau O, Ninot JM, Cornelissen JHC, Callaghan TV. 2013. Similar tree seedling responses to shrubs and to simulated environmental changes at Pyrenean and subarctic treelines. Plant Ecology and Diversity 6: 329–343.

Hättenschwiler S, Miglietta F, Raschi A, Körner C. 1997. Thirty years of in situ tree growth under elevated CO_2: A model for future forest responses? Global Change Biology 3: 463–471.

Hautier Y, Niklaus PA, Hector A. 2009. Competition for light causes plant biodiversity loss after eutrophication. Science 324: 636–638.

Herman JJ, Sultan SE, Horgan-Kobelski T, Riggs C. 2012. Adaptive trans-generational plasticity in an annual plant: Grandparental and parental drought stress enhance performance of seedlings in dry soil. Integrative and Comparative Biology 52: 77–88.

Hillebrand H, Matthiessen B. 2009. Biodiversity in a complex world: Consolidation and progress in functional biodiversity research. Ecology Letters 12: 1405–1419.

Hillebrand H, Gurevitch J. 2013. Reporting standards in experimental studies. Ecology Letters 16: 1419–1420.

Hui DF, Sims DA, Johnson DW, Cheng WX, Luo YQ. 2002. Effects of gradual versus step increases in carbon dioxide on Plantago photosynthesis and growth in a microcosm study. Environmental and Experimental Botany 47: 51–66.

Jentsch A, Kreyling J, Beierkuhnlein C. 2007. A new generation of climate change experiments: Events not trends. Frontiers in Ecology and the Environment 5: 315–324.

Jentsch A, et al. 2011. Climate extremes initiate ecosystem-regulating functions while maintaining productivity. Journal of Ecology 99: 689–702.

Johnson EA, Miyanishi K. 2008. Testing the assumptions of chronosequences in succession. Ecology Letters 11: 419–431.

Klady RA, Henry GHR, Lemay V. 2011. Changes in high arctic tundra plant reproduction in response to long-term experimental warming. Global Change Biology 17: 1611–1624.

Klironomos JN, Allen MF, Rillig MC, Piotrowski J, Makvandi-Nejad S, Wolfe BE, Powell JR. 2005. Abrupt rise in atmospheric CO_2 overestimates community response in a model plant—soil system. Nature 433: 621–624.

Knapp AK, et al. 2012. Past, present, and future roles of long-term experiments in the LTER Network. BioScience 62: 377–389.

Kongstad J, Schmidt IK, Riis-Nielsen T, Arndal MF, Mikkelsen TN, Beier C. 2012. High resilience in heathland plants to changes in temperature, drought, and CO_2 in combination: Results from the CLIMAITE experiment. Ecosystems 15: 269–283.

Kreyling J, Beier C. 2013. Complexity in climate change manipulation experiments. BioScience 63: 763–767.

Kreyling J, Jentsch A, Beierkuhnlein C. 2011. Stochastic trajectories of succession initiated by extreme climatic events. Ecology Letters 14: 758–764.

Kreyling J, Jentsch A, Beier C. 2014. Beyond realism in climate change experiments: Gradient approaches identify thresholds and tipping points. Ecology Letters 17 (art. 125-e1).

Kulmatiski A, Beard KH, Stevens JR, Cobbold SM. 2008. Plant–soil feedbacks: A meta-analytical review. Ecology Letters 11: 980–992.

Lancaster J, Belyea LR. 1997. Nested hierarchies and scale-dependence of flow refugium use. Journal of the North American Benthological Society 16: 221–238.

Latimer JG, Oetting RD. 1999. Conditioning treatments affect insect and mite populations on bedding plants in the greenhouse. HortScience 34: 235–238.

Lee H, Mishurov M. 2013. Scaling climate change experiments across space and time. New Phytologist 200: 595–597.

Leuzinger S, Koerner C. 2007. Water savings in mature deciduous forest trees under elevated CO_2. Global Change Biology 13: 2498–2508.

Luo YQ, Reynolds JF. 1999. Validity of extrapolating field CO$_2$ experiments to predict carbon sequestration in natural ecosystems. Ecology 80: 1568–1583.

Luo Y, et al. 2008. Modeled interactive effects of precipitation, temperature, and [CO$_2$] on ecosystem carbon and water dynamics in different climatic zones. Global Change Biology 14: 1986–1999.

Luo Y, et al. 2011. Coordinated approaches to quantify long-term ecosystem dynamics in response to global change. Global Change Biology 17: 843–854.

Manning P, Morrison S, Bonkowski M, Bardgett RD. 2008. Nitrogen enrichment modifies plant community structure via changes to plant–soil feedback. Oecologia 157: 661–673.

Marion GM, et al. 1997. Open-top designs for manipulating field temperature in high-latitude ecosystems. Global Change Biology 3: 20–32.

Maron JL, Marler M, Klironomos JN, Cleveland CC. 2011. Soil fungal pathogens and the relationship between plant diversity and productivity. Ecology Letters 14: 36–41.

Meijer SS, Holmgren M, Van der Putten WH. 2011. Effects of plant–soil feedback on tree seedling growth under arid conditions. Journal of Plant Ecology 4: 193–200.

Milcu A, Lukac M, Subke JA, Manning P, Heinemeyer A, Wildman D, Anderson R, Ineson P. 2012. Biotic carbon feedbacks in a materially closed soil-vegetation-atmosphere system. Nature Climate Change 2: 281–284.

Milcu A, et al. 2014. Functional diversity of leaf nitrogen concentrations drives grassland carbon fluxes. Ecology Letters 17: 435–444.

Mommer L, van Ruijven J, Jansen C, van de Steeg HM, de Kroon H. 2012. Interactive effects of nutrient heterogeneity and competition: Implications for root foraging theory? Functional Ecology 26: 66–73.

Norby RJ, Zak DR. 2011. Ecological lessons from free-air CO$_2$ enrichment (FACE) experiments. Annual Review of Ecology, Evolution, and Systematics 42: 181–203.

O'Gorman EJ, Benstead JP, Cross WF, Friberg N, Hood JM, Johnson PW, Sigurdsson BD, Woodward G. 2014. Climate change and geothermal ecosystems: Natural laboratories, sentinel systems, and future refugia. Global Change Biology 20: 3291–3299.

Petersen JE, Englund G. 2005. Dimensional approaches to designing better experimental ecosystems: A practitioner's guide with examples. Oecologia 145: 216–224.

Poorter H, et al. 2012. The art of growing plants for experimental purposes: A practical guide for the plant biologist. Functional Plant Biology 39: 821–838.

Reynolds CS. 1998. What factors influence the species composition of phytoplankton in lakes of different trophic status? Hydrobiologia 369/370: 11–26.

Roscher C, Temperton VM, Scherer-Lorenzen M, Schmitz M, Schumacher J, Schmid B, Buchmann N, Weisser WW, Schulze ED. 2005. Overyielding in experimental grassland communities—irrespective of species pool or spatial scale. Ecology Letters 8: 419–429.

Scherrer D, Körner C. 2010. Infra-red thermometry of alpine landscapes challenges climatic warming projections. Global Change Biology 16: 2602–2613.

Shaver, et al. 2000. Global warming and terrestrial ecosystems: A conceptual framework for analysis. BioScience 50: 871–882.

Silvertown J, Poulton P, Johnston AE, Edwards G, Heard M, Biss PM. 2006. The Park Grass Experiment 1856–2006: Its contribution to ecology. Journal of Ecology 94: 801–814.

Sowerby A, Emmett BA, Tietema A, Beier C. 2008. Contrasting effects of repeated summer drought on soil carbon efflux in hydric and mesic heathland soils. Global Change Biology 14: 2388–2404.

Spivak AC, Vanni MJ, Mette EM. 2011. Moving on up: Can results from simple aquatic mesocosm experiments be applied across broad spatial scales? Freshwater Biology 56: 279–291.

Stevnbak K, Scherber C, Gladbach DJ, Beier C, Mikkelsen TN, Christensen S. 2012. Interactions between above- and belowground organisms modified in climate change experiments. Nature Climate Change 2: 805–808.

Tylianakis JM, Rand TA, Kahmen A, Klein A-M, Buchmann N, Perner J, Tscharntke T. 2008. Resource heterogeneity moderates the biodiversity–function relationship in real world ecosystems. PLOS Biology 6: 947–956.

Van der Putten WH, et al. 2013. Plant–soil feedbacks: The past the present and future challenges. Journal of Ecology 101: 265–276.

Vicca S, et al. 2012. Urgent need for a common metric to make precipitation manipulation experiments comparable. New Phytologist 195: 518–522.

Wookey PA. 2008. Experimental approaches to predicting the future of tundra plant communities. Plant Ecology and Diversity 1: 299–307.

Yurkonis KA, Wilsey BJ, Moloney KA. 2012. Initial species pattern affects invasion resistance in experimental grassland plots. Journal of Vegetation Science 23: 4–12.

Hans J De Boeck (hans.deboeck@uantwerp.be), Sara Vicca, Ivan Nijs, and Matteo Campioli are affiliated with the Centre of Excellence PLECO (Plant and Vegetation Ecology) in the Department of Biology at the Universiteit Antwerpen (Campus Drie Eiken), in Wilrijk, Belgium. Jacques Roy and Alexandru Milcu are affiliated with Ecotron–Centre National de la Recherche Scientifique (CNRS), in Montferrier-sur-Lez, France; AM is also affiliated with CNRS–Centre d'Ecologie Fonctionnelle et Evolutive (CEFE), in Montpellier, France. Juergen Kreyling specializes in experimental plant ecology at the Institute of Botany and Landscape Ecology at Greifswald University, in Germany. Anke Jentsch specializes in disturbance ecology and Carl Beierkuhnlein specializes in biogeography at the Bayreuth Center of Ecology and Environmental Research (BayCEER) at the University of Bayreuth, in Germany. Abad Chabbi is affiliated with the Institut National de la Recherche Agronomique (INRA) Poitou-Charentes, in Lusignan, France, and with Laboratoire de Biogéochimie et Ecologie des Milieux Continentaux (BIOEMCO)–Centre INRA Versailles-Grignon, in Thiverval-Grignon, France. Terry Callaghan is affiliated with the Royal Swedish Academy of Sciences, in Stockholm, Sweden; with the Department of Animal and Plant Sciences at Sheffield University, United Kingdom; and with the Department of Botany at National Research Tomsk State University, in Russia. Claus Beier is affiliated with the Norwegian Institute for Water Research, in Oslo, Norway.