



1300 Centre Street  
Boston, MA, 20131

February 9, 2018

Dear Dr. Chase, Dr. Hillebrand, and members of the editorial board:

Please consider our paper, entitled “How do climate change experiments actually change climate?” for publication as a Review & Synthesis in *Ecology Letters*. This manuscript is a substantially revised version of a previously submitted manuscript (ELE-01149-2017). We have incorporated the suggestions the referees, and this new manuscript includes large portions that have been rewritten, as well as new data and analyses. We include a point-by-point response to reviewer comments below.

The biological impacts of climate change have been widely observed around the world, from shifting species’ distributions to altered timing of important life events, and remain a major area of ecological research. With growing evidence and interest in these impacts, ecologists today are challenged to make quantitative, robust predictions of the ecological effects of climate change. One of the most important methods to achieve this goal is field-based climate change experiments that alter temperature and precipitation. The utility of these experiments, however, is directly dependent on the climate change they produce and how accurately researchers assess these changes.

Field climate change experiments have been used for decade and are still prevalent across diverse sub-disciplines (1,2) for cutting-edge climate change research. They critically offer the ability to create “no-analog” climate scenarios forecasted for the future, to isolate effects of temperature and precipitation from other environmental changes, and to examine non-linear responses to climatic changes. Yet, these experiments have recently been shown to estimate effects much smaller than those seen in long-term observational studies (3). Despite calls for improved methods (4,5), even sophisticated approaches appear to suffer from this discrepancy (6). Such results highlight the need to synthesize across studies and assess how realistically experiments alter climate conditions, as well as to develop novel approaches for applying experimental results to forecasting biological impacts of global climate change.

We address these major needs through a new database of daily climate data from 12 active warming experiments, containing an estimated 46 study years and 11629 study days of air and soil temperature and soil moisture data. Using this database we find that experimental climate change results may be interpreted in misleading ways, especially through the common practice of summarizing and analyzing only the mean changes across treatments. We show that such methods mask important variation in treatment effects over space and time. We also find that secondary, unintended treatment effects, which are rarely described or interpreted (e.g. soil drying with warming treatments), may lead to under- or over-estimation of climate change impacts. The implications of these complexities are likely to have important ecological consequences, across diverse systems. We describe a case study of spring plant phenology, in which a simple mean-focused analysis, ignoring secondary effects, leads to inaccurate quantification of species’ sensitivities to changes in temperature. We present our recommendations for future experimental design, analytical approaches, and data sharing that we believe will improve the ability of climate change experiments to accurately identify and forecast species’ responses.

Our author team brings together an international and interdisciplinary team of researchers, which bridges perspectives from ecology, climatology, and land surface modeling. It is comprised of many of the scientists who execute major warming experiments, as well as those who have raised concerns over the findings of such experiments. We expect our Review & Synthesis will lead to improved mechanistic understanding of climatic drivers of biological responses, and inspire innovative experimental design and analyses; we hope you will consider it for *Ecology Letters*.

Sincerely,

A handwritten signature in black ink that reads "Ailene K. Ettinger". The signature is fluid and cursive, with the first name "Ailene" and last name "Ettinger" clearly legible.

Ailene Ettinger Postdoctoral Fellow, Arnold Arboretum of Harvard University & Biology Department, Tufts University

*References mentioned in cover letter*

1. Reich, P. B., K. M. Sendall, K. Rice, R. L. Rich, A. Stefanski, S. E. Hobbie, and R. A. Montgomery. 2015. Geographic range predicts photosynthetic and growth response to warming in co-occurring tree species. *Nature Climate Change* 5:148-152.
2. Barton, B. T., and O. J. Schmitz. 2009. Experimental warming transforms multiple predator effects in a grassland food web. *Ecology Letters* 12:1317-1325.
3. Wolkovich, E. M., B. I. Cook, J. M. Allen, T. M. Crimmins, J. L. Betancourt, S. E. Travers, S. Pau, J. Regetz, T. J. Davies, N. J. B. Kraft, T. R. Ault, K. Bolmgren, S. J. Mazer, G. J. McCabe, B. J. McGill, C. Parmesan, N. Salamin, M. D. Schwartz, and E. E. Cleland. 2012. Warming experiments underpredict plant phenological responses to climate change. *Nature* 485:494-497.
4. Beier, C., C. Beierkuhnlein, T. Wohlgemuth, J. Penuelas, B. Emmett, C. Körner, H. Boeck, J. H. Christensen, S. Leuzinger, I. A. Janssens, et al. 2012. Precipitation manipulation experiments: challenges and recommendations for the future. *Ecology Letters* 15:899-911.
5. Kreyling, J., A. Jentsch, and C. Beier. 2014. Beyond realism in climate change experiments: gradient approaches identify thresholds and tipping points. *Ecology Letters* 17:125.
6. Menke, S. B., J. Harte, and R. R. Dunn. 2014. Changes in ant community composition caused by 20 years of experimental warming vs. 13 years of natural climate shift. *Ecosphere* 5:1-17.

*Reviewer Comments are in italics. Author responses are in plain text.*

REVIEWER 1

REVIEWER 2

REVIEWER 3

*This is a well written, often well-thought out submission. It has many strengths; all experimental manipulations are highly imperfect and we need to pay attention to what those imperfections mean for our interpretation or results. So, the authors are onto something important.*

Thank you!

*However, the paper also has several major weaknesses (which should be and can be remedied). Here are issues, in roughly descending importance:*

*(A) the authors lump and jointly analyze extremely different types of warming treatment approaches without carefully isolating which problems are associated with which. This is an extremely major problem, as it seems completely unhelpful to lump chamberless infrared warming, for example, with forced-hot air open-top chambers. The analyses also lump studies that warm both above- and below-ground with those that warm only aboveground or only belowground. Given these problems, the sample size of the study becomes problematic for a synthesis, although perhaps okay for a review. There are also other studies that could have been included. A major issue is the wide variation in experimental manipulation approaches among the 12 studies, as well as the different systems (i.e. ecosystem types, climate) and what that might mean for interpretation (which is largely ignored). To begin with, the paper simply needs to be much clearer about all this by providing more information in Table S1 (which I would much prefer to see in the main text) including information about the ecosystem type, and more information about each study. Additionally, although the authors (Clark et al 2014) call their apparatus an open-top chamber, the authors of the EL submission do not. It is not clear why, nor is it clear why only these OTCs and not others, have been included in this study. The paper would be more valuable if more examples of each warming approach were included. A separate issue is whether for purposes of this paper, continuations of the Harvard Forest and Duke Forest forced air experiments (assuming that is in fact what they are, I am guessing) should be considered as independent experiments for this analysis? If the same equipment and same approach were used, this seems like pseudo-replication to me. I think it would be extremely useful for readers if there was a Table that included passive warming, even though it has been studied frequently, as part of a qualitative comparison of passive OTC warming, active OTC warming, soil cable warming, infrared heater aboveground warming, and any combinations of aboveground and belowground warming that studies deployed.*

We greatly appreciate the reviewer's suggestion of separating our interpretation by warming type. In the our revised manuscript, we have taken the following steps to address these valuable points:

1. Color coded sites by warming type in Figures 2 and 3
2. *(B) the authors don't accurately convey projected future climate, and thus some of their criticisms, particularly of warming-induced soil moisture changes, are off-target; see comments below under ?specific points?.*  
*(C) the paper could help readers by clearly identifying (and separately interpreting) which studies directly and independently heated both above and belowground, just aboveground, or just below-ground; and which studies attempted to maintain a set elevation of temperature versus applied constant inputs of energy (e.g. constant wattage for infrared heaters).*  
*(D) the authors could do more to acknowledge diverse, alternative goals (does one want to isolate a single treatment variable or have a more ecologically realistic treatment that also simulates indirect effects? See below for more about this), rather than assuming that experiments that include indirect*

effects have major problems.

(E) the paper mentions cases where effects of experimental treatment resulted in different responses than natural variation, but ignored cases where effects of experimental treatment do result in similar responses as natural variation in the same driver,

(F) the paper could do more to frame concepts about what we should expect across different temporal and spatial scales when comparing experiment to observed patterns- i.e. we should not always expect them to be the same, as long-term indirect effects are part of the long-term observational record but not of the shorter-term experiments. Clearly if the short-term experimentalists claim that those results will hold true over long-term, that is a problem, but the actual outcomes shouldn't be expected to match. In other words, this may be more of an issue of how people interpret experiments versus observations, rather than any intrinsic problem with either.

Specific points 1. Abstract: Overall, reads well, except for the sentence about soil drying. ? Furthermore, warming treatments produce unintended secondary effects, such as soil drying.? As noted above and discussed below, sometimes these are intentional indirect effects, because they will likely occur in the future (see below). For example, in Rich et al 2015 GCB and related studies from that experiment, this (increased ET and thus reduced soil water) was an intentional secondary effect that matches natural processes that occur with climate warming (see below).

2. Line 52-53. Need to balance cases where experimental results do not match with cases where they do. The authors point out where responses are not the same as the temporal or spatial trends in nature, but ignore where it does. For example, the acclimation response of leaf respiration to temperature variation was similar across treatments as across time, in two independent experiments in two different system (Aspinwall et al 2016, New Phytol; Reich et al 2016, Nature). Another example is Carey et al 2016 (PNAS) which found that responses of soil respiration experimental warming often (and more often than not), well-match temporal responses of soil respiration to seasonal soil temperature change. Furthermore, sometimes one would expect an experimental treatment to match response to climate change, other types not. If we expect long-term responses (10, 20, 40, 80 years) to be solely driven by short-term mechanisms (as picked up in experimental of one or two or three years length), then perhaps they should match well. But if the long-term response involves feedbacks ? and there are many possible (compositional change due to competitive dynamics, trophic cascades, biogeochemical feedbacks, etc), then one might not expect the same response as in a short-term experiment. Check out Andressen et al 2016 (Adv Ecol Research) for how temporally variable response to experimental manipulations are. If authors of short-term experiments claim that their results should directly match long-term responses, then it is fair to point out that that is likely not to be true. But for me this is more an issue of how we interpret experiments, than any problem with experiments or observations. So we need to be extremely careful in what we ?expect? ? and the authors should provide a more nuanced presentation of this. The authors specifically state: "We use a case study of spring plant phenology to demonstrate how analyses that assume a constant and perfect treatment effect, ignoring secondary effects of warming treatments, lead to inaccurate quantification of plant sensitivity to temperature." But did the authors of the original paper really make such claims?

3. Lines 83-87 and soon after: In setting up their analysis the authors need to provide more information. How did these experiments warm plants and soils? What source of heat? Did they control temperature elevation via feedback control or deploy constant output heating (e.g. constant wattage lamps)? Did the study heat both aboveground and belowground independently? Some of this information is given in TableS1, but not enough and it should be put in main text. Did they analyse across all 12 studies despite the radical differences in experimental goals, designs, and methodology? It seems like they often did. Or did they stratify and isolate differences among these very different treatments, I guess as a reader we will find out, but it would be good to let reader know along the way. As there were 7 infrared studies (and I assume none of these warmed soils directly), two forced air, two forced air + soil warming, and one soil-warming, I worry about the extent to which this can be claimed to be a quantitative analysis across multiple

experiments other than for infrared studies (and  $n = 7$  is quite modest). This does not minimize the points they can and do make about the other approaches, but it is more of a ?review? than ?synthesis? in those cases. I also want to know as a reader how similar the IR approaches were for the 7 studies- did they vary in number of lamps, placement of lamps, constant wattage versus constant temperature elevation design, etc. We again thank the reviewer for pointing out where more specific information is needed in this manuscript. We have modified Table S1, so that it contains much for information. Move to main text? 4. Figure 1 Figure 1 is problematic to me as presented. Did all experiments intend to manipulate soil temperature or did some intend only to manipulate aboveground temperature? We know that in many contexts (types of ecosystems, time of year), aboveground infrared warming will not warm soils well. What depth are the soil T measurements shown at? The paper is guilty of mixing apples and oranges throughout, unintentionally painting different techniques with problems of other techniques by lumping different techniques in a single comprehensive analysis, which makes no sense. We thank the reviewer for pointing out the need for additional clarity and details about study methods, as well as the need to separate our interpretation by warming type. We now present above-ground temperature as well, to show variation in above-ground temperature, since, as the reviewer suggests, some warming methodologies warm the air, not the soil. We have also modified the legend of Figure 1, so that it now states the warming methodology of each site.

5. Figure 3 Problem: Figure 3 applies only to the forced-air treatment used in the study at Duke Forest and Harvard Forest. Lines 170-180 ? do these apply largely or entirely to systems like DK and HF where the treatment included a large structure?? It makes sense that there would be structural effects for the forced air OTC approach, just as there is for passive OTCs and for that matter, temperature controlled OTCs (e.g. in SPRUCE, the structures raise temperatures by 1-2C above a pure ambient). The authors should cite that as well. I believe this has been published, if not, is widely disseminated by spruce researchers. Moreover, why not include SPRUCE data here? In contrast to OTC example, it makes sense for studies with just a small amount of 3-d space having structure in it (i.e. posts and lamps aboveground, soil cables below) to have minimal structural effect. For example, Rich et al 2015 GCB found neither thermal nor biological differences between ambient (no structures) and sham treatments (structures).

6. Line 181 Most warming experiments calculate focal response variables relative to ambient controls (e.g., Price 182 Waser, 1998; Dunne et al. , 2003; Cleland et al. , 2006; Morin et al. , 2010; Marchin et al. , 2015), which our analyses suggest will not properly account for infrastructure effects. That is an unclear and perhaps unfair statement and highlights my concerns with the ?lumping? of apples and oranges in their analyses and presentation. First, regarding clarity- readers will be unsure if ?ambient controls? mean ambient environment without any structure or if it means the ?control? or non-heated treatment, which is ambient, but does have the structures in place. Second their point may be valid if structures matter, as when there are ?chambers? (as with OTCs of the forced-hot air approach) but if minimal environmental effects occur when there are IR lamps or soil cables, then this unfairly paints such approaches with a problem that only chambers have.

We thank the reviewer for pointing out the lack of clarity in this statement. In the revised manuscript, we define “ambient controls” as compared to “structural controls in Lines XX-XX. (...structural controls (i.e., ‘shams’ or ‘disturbance controls,’ which contained all the warming infrastructure, such as soil cables or infrared heating units but with no heat applied) and ambient controls with no infrastructure added.” We have worked in other parts of the manuscript to more clearly separate the different methods (forced hot air with OTCs, IR lamps, and soil cables) in our interpretation of the data. The data presented here, and in other studies (e.g. CITATIONS) suggest that there may be environmental effects from experimental infrastructure from all types of active warming. We therefore urge future researchers to collect, present, and interpret microclimate data from both ambient and structural controls; through additional analyses of these data, it will be possible for scientists to quantify the magnitude of “environmental effects” of experi-

mental infrastructure and from there decide if they are "minimal" or not. The studies that we have come across suggest that we do not yet know all the ways that experimental infrastructure affects biological responses so we shouldn't assume that effects are 0 for any warming type.

7. Line 192-202 *This section does a fair job of pointing out that most (all?) warming treatments have a difficult time warming to a target when soils are wet and/or plant are transpiring a lot. This was noted in Rich et al 2015 GCB. And moreover, this issue is particularly so if only aboveground warming is employed. Most (all?) grassland experiments that only warmed aboveground with IR heaters will have warmed the soil a moderate bit when vegetation was sparse but not so in treatments that create denser vegetation or times of year with denser vegetation or locations with denser vegetation. This is a major aspect of warming;; and variation within as well as among experiments is largely ignored in the current version of the paper. The authors do point out that this outcome – less warming when soils and/or plants are evaporating a lot– is accidental but does have a biophysical foundation and may match how warming from greenhouse forcing will influence future temperature elevations. So ? is this a problem or an accidentally brilliant technique? Again, the language could be more nuanced in discussing such issues. The EL submission should go further and discuss the temporal implications- studies with constant wattage will achieve markedly different amount of warming during times of year of low ET (e.g. spring in a grassland for instance where LAI is very low) versus high ET (summer in same system). and even feedback control studies will also have differences in such conditions, but much smaller ones. This is an important point (and an important nuance ? i.e. that constant wattage studies are particularly problematic in this regard). Here too the forced-air heating method should be contrasted- I do not know whether the problem is smaller or larger in that study, given that ambient air is treated before entering the chambers.*

18. Line 207 and lines 210-217 *"Thus, although active warming experiments may not be explicitly designed to manipulate soil moisture, soil moisture is unavoidably affected by changing temperatures." There are two issues here; one whether declining soil moisture with warming is ecologically realistic, and two, whether authors goals are to assess direct and indirect effects, or to isolate just direct effects. Thus I think the authors need to specify and be open to the goals of the experiments. If authors want to understand how temperature change influences plant function in isolation from other plausible indirect effects then it is fine to criticize them for ignoring the indirect effects or to applaud them for adding water to offset soil drying. But if authors intent was to examine joint effects of direct temperature effects on metabolism with indirect effects on soil moisture (e.g., Reich et al 2015, Rich et al 2015, see page 2338), they should be applauded for that. I am familiar with the cited studies and would think that the authors might have been interested in characterizing the way in which indirect changes (as in plant composition) would influence the trajectory of their studied system under climate change. Line 217 reads "The widespread presence of unintended secondary effects of climate change manipulations..." This wording is incorrect and thus an issue, because it basically says that everything in the lines above is a problem with those cited studies rather than their ecologically realistic intent! A key question of course is whether the indirect changes (as well as the direct T responses) are what we would expect with climate warming. See below for discussion of that.*

We thank the reviewer for calling attention to the need for clarification in our discussion of direct and indirect (including secondary and feedback) effects of warming in experimental climate change studies. As the reviewer notes, many authors conduct field experiments precisely because of the many "indirect" effects of warming that can affect species responses. This is one of the strengths of field experiments, which we acknowledge in the introduction (Lines 45-47: "Compared with indoor growth-chamber experiments, field-based experiments offer the possibility of preserving important but unknown or unquantified feedbacks among biotic and abiotic components of the studied systems."). In this new version of the manuscript, we have reworded lines 210-217 to make clear that the problem is not the presence of indirect effects of warming treatments, but of unrealistic indirect effects, which may be considered experimental artifacts that are unlikely to

occur outside of warming experiments. In order to know whether or not these non-temperature effects of warming treatments are realistic, we need to quantify them. We have rewritten That is the main point we wish to make.

*19. Figures 5 and 6. These go in a good direction but need tweaking and/or clarification. Figure 5 is presented as a problem of indirect effects. If the indirect effects are ecologically realistic ones, then one might argue that this is THE VALUE of realistic long-term field experiments, to be able to assess in an ecologically realistic way both the direct effects and the indirect effects that cascade through ecosystems over time.*

We agree that part of the value of long-term field experiments is that they may include indirect effects that are ecologically realistic, and might not be apparent from, for example, controlled laboratory studies manipulating temperature in isolation. The crux of the problem is in identifying when indirect effects are ecologically realistic and when they are experimental artifacts. The point we wish to make here is that we believe that critical steps toward distinguishing realistic indirect effects from unrealistic ones are quantifying, interpreting, and reporting these non-temperature effects in experimental studies. We have changed the text in the legends of Figures 5 and 6, as well as throughout the document, to clarify this. We now use the term “non-temperature” effects to refer to indirect effects of experimental warming that alter conditions other than temperature. *Regarding Fig 6, it is not clear from main text which studies went into this statistical model. From Table S15 it appears that data from 5 sites were used. Depending on which ones, and which warming treatments were used, we might interpret these results in a variety of fashions. For example, is soil drying is a realistic outcome of the treatments (see point 10 below) rather than a problematic artefact, then the impacts on phenology are ecologically realistic but must be interpreted through the lens of joint direct temperature and indirect moisture effects. This is enormously different perspective than the one the authors provide. If for some reasons the soil drying exceeds that expected in a warmer world, then the interpretation must be tempered because that indirect effect will be larger than one might expect in a warmer world.*

We thank the reviewer for pointing out the need for additional details about our analysis! We now list the studies included in the budburst analysis in the legends of Figure 6 and Tables S15-16

*10. Lines 273- 290 and the specific sentences below in italics “However, the northeastern United States has been trending wetter over time and is expected to be wetter in the future (Seager et al. , 2014; Shuman Burrell, 2017). The soil moisture changes in warmin experiments, and the biological changes they cause, may therefore represent an experimental artifact that is unlikely to occur with future warming.” This section raises important points, and does a good job of doing this with respect to effects of high ET in summer on realized warming (see point 7 above), but does not do a good job of building a foundation on the published literature regarding whether the future with be wetter or drier, which I focus on here. They cite two papers to claim that the northeast will have a wetter future. However, one, Shuman and Burrell, is about historic climate history not future projections. The second, Seager et al 2014, is relevant as it is a summary of model ensemble predictions of future P and E (and thus P-E). These results show that both P and E are projected to increase in both winter and summer for most of the US and Canada in 2021-2040 (compared with 1979-2005). However, while P-E (a rough surrogate of mean soil moisture trend) increases in winter in the east and decreases in the west, in summer the trend is reversed, with most of US and Canada, including the east, decreasing in terms of P-E. On top of that, increased rainfall intensity (more rain in fewer events), well documented in the northeast, means less recharge of soil moisture per unit P. Thus, Seager et al is consistent with other papers that suggest that in the warming future, soils will be functionally drier, because P (and recharge per P) will not keep up with ET, even in places where P might increase in the near future. So the idea pushed by the EL authors, that the northeast will be moister in the future is not consistent with the reference the authors use nor with a range of ecohydrologically oriented publications. This casts doubt on the utility of their claim that soil drying from experiments should be considered an artefact and suggests they need a thorough rewriting of all of the points of their paper that address this issue.*

We thank the reviewer for these comments. and have re-written the section to clarify the main point we wish to make- that warming and drying do not necessarily go hand in hand. Soil moisture trends are expected to vary by region, season, and even soil depth. We have added a citation of a recent paper to further emphasize this variation: Berg et al. 2017. Divergent surface and total soil moisture projections under global warming. *Geophysical Research Letters*, 44(1):236-44. It is possible that the soil drying we observe in warming experiments is consistent with changes to soil moisture that may occur with future warming in some regions; it is also possible that soils will get wetter in the future, in some regions, during some seasons, and at some depths.

*Summary*

*In summary, the paper has many important points but does so using less of the available data than would be preferable, with analyses that jumble together different warming techniques in a way that is not clarifying to a reader, and with some underlying premises that are either too narrow or actually wrong. Organizationally, I think that the paper would be far more useful if each section spoke about one of the approaches (passive aboveground warming, active aboveground warming with OTCs, active aboveground warming with IR lamps, active belowground warming with cables, combinations of these techniques) and presented data about the way that particular treatment achieved or failed to achieve its goals. Clearly in my opinion this paper is far far away from being in a publishable state at present. But, the topic is important and a thorough revision (perhaps with more data added) could be useful for both practitioners in the field and general readers who may not otherwise pay attention to important details of different kinds of experimental warming studies.*