The ARNOLD ARBORETUM of HARVARD UNIVERSITY

1300 Centre Street
Boston, MA, 20131

March 29, 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 of the referees, and this new manuscript includes new data, analyses, and text, as well as many revisions to previous figures and tables. 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 we as researchers interpret these changes.

Field climate change experiments have been used for decades and are still prevalent across diverse subdisciplines (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 thoroughly assess how experiments alter microclimate conditions, and to develop novel approaches for analyzing the large amount of microclimate data available from climate change studies to apply experimental results to forecasting biological impacts of global climate change.

We address these major needs through a new database of daily climate data from 15 active warming experiments, containing an estimated 59 study years and 14,913 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 in temperature across treatments. We show that such methods mask variation in treatment effects over space and time. We also find that secondary treatment effects, which are rarely thoroughly described or interpreted with biological responses, may lead to under- or over-estimation of biological responses to climate change. These complexities are likely to to be relevant 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,

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. Betan-court, 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. Korner, 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 com-position 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

The central argument made in this manuscript is:

1. *Active climate-warming experiments yield unintended climate effects, in addition to achieving or not achieving the "target temperature."*
2. *These unintended effects can be either non-temperature responses or responses of any kind that vary over time and place.*
3. *Ecosystems can respond to non-temperature climate effects, and to local spatial and temporal variation in climate.*
4. *Analyses based on mean temperature response will be misleading.*

I am familiar with several of the climate experiments discussed here and with other active warming experiments that are not included. Not a single sentence about climate responses in this manuscript will be surprising to the experimenters I talk to, except in so far as their intentions have been mischaracterized.

Notions like "unintended" have to be used cautiously in the academic literature. In this paper the word is misused to a level that borders on insulting. When the experiments I am most familiar with were originally conceived, designed, and proposed to NSF, there was explicit and clearly stated understanding that active heating would dry soil, affect date of snowmelt in the spring and snow accumulation in autumn, alter frost incidence, affect soil temperature and other climate responses over heterogeneous aspects and slopes, and influence soil temperature response as a function of varying soil moisture via the Bowen ratio effect. The intentions were spelled out in both NSF proposals to launch the experiments and in the journal papers that ensued.

In this manuscript, we seek to call attention to important ways that active warming experiments alter microclimate that are not often explicitly interpreted in analyses focused on biological responses in active warming experiments. We thank the reviewer for pointing out that we were not clear about this in the earlier version, and for pointing out some poor choices of language in our original manuscript. We do not wish to insult researchers conducting climate experiments; many of the authors on this manuscript fall into this category, and we appreciate all the hard work put into and valuable insights gained from active warming experiments. We have rewritten much of the manuscript to be more clear and have worked to avoid any potential insulting language. For example, we have removed the word "unintended" throughout the manuscript. In addition, we now discuss more thoroughly insights made in previous analyses of the secondary and complex effects of active warming- as the reviewer points out, many scientists have conducted in-depth analyses of the complex abiotic effects, including soil drying and altered freeze-thaw cycles, in single experiments (e.g., Kimball 2005, Kimball et al. 2008, McDaniel et al. 2014, Pelini et al 2011). We feel the critical next step, to improve understanding and forecasting of climate change impacts outside of an experimental context, is to integrate this level of detail directly with analysis of biological responses. See below for more on this point.

Moreover, despite this manuscripts continual reference to a target temperature, in a number of cases there was no fixed target temperature because the experimental design set infrared flux or heating coil wattage to a constant value rather than thermostatted the temperature response.

We thank the author for pointing out that, although the majority of climate change studies do describe explicit targets in their own manuscripts, some active warming experiments do not have fixed target temperatures. Twelve of the 15 studies in our database describe targets in their own manuscripts (the three that do not are exp0, exp, and exp). In addition, all studies did report a mean level of warming. In the revised manuscript, we seek to make this distinction more clear by using the word "reported temperature" rather than target, when relevant. In addition, we have added a column to Table S1 for

"warming control": "fixed" denotes fixed wattage (i.e. there was not a target temperature maintained by the warming equipment); "feedback" denotes that it was themostatted temperature response (i.e., the equipment was set to try to maintain a target temperature). One key point we wish to make in our manuscript is not whether or not studies match their target warming, though this may be interesting to some readers, but is that there is often variation (in space and time) in the amount of warming experienced by organisms in these experiments (e.g. Figure 2). By using one mean value (whether target or reported) in analyses, this variation is not analyzed, interpreted, nor understood, despite the fact that it may have important implications.

The authors clearly have very selectively read (or at least cited) the literature, and appear to have selected papers that ignore the points made above. For example, there is a thorough analysis of factors influencing both temperature responses and soil moisture and melt date responses, and spatial variation in responses across plots, and temporal variation of responses, and effects of vegetation cover on climate responses, in the Harte et al. 1995 Ecological Applications paper, yet that is not even cited here.

We thank the reviewer for bringing to our attention this reference. We now cite it in several places in the manuscript (e.g. Lines XX, Lines XX, Lines XX-XX) . This reference offers excellent detailed analysis of soil climate data in one experiment at one site. Indeed, other active warming study authors have taken a similar approach of conducting detailed analyses of climate data separately from biological data (e.g., Kimball 2005, Kimball et al. 2008, McDaniel et al. 2014, Pelini et al 2011). One of the points we wish to make is that we believe our understanding of biological processes will be enhanced by integrating such detailed analyses of climate data with the analyses of biological data (e.g. phenology). In addition, although there are multiple studies that have independently analyzed climate data from climate change experiments (e.g. Harte et al. 1995, McDaniel et al. 2014, Pelini et al 2011), a meta-analysis that synthesizes across multiple studies and warming types is lacking to date.

Moreover, the complex structure of climate responses has in fact been utilized in many of the analyses that experimenters have carried out. ANCOVA tests, multiple regressions and many other types of analysis have been used to determine how ecological responses depend upon the multidimensionality of altered climate conditions, including spatial and temporal variation as well as non-temperature responses.

We found that 10 out of the 15 studies in our database treat warming as a constant, categorical factor in their analysis (i.e. ANOVA was used). In the supplemental Table X, we now include a column for analysis type for each paper associated with each study in our database.

There are many more technical and scholarly issues with this paper. Lines 198-202 ignore the bigger issue here which is the soil, not air, temperature response triggered by precipitation. We thank the reviewer for pointing out that soil effects are also important, and we have added information about soil temperatures precipitation treatment effects on soil temperatures as well (now lines XX to XX). *Lines 215-216 will mislead the reader into thinking that carbon feedback is a local effect, and it ignores the evidence that both negative and positive feedbacks can arise from the processes described.* We have reworded the sentence in an attempt to clarify that both negative and positive feedbacks can arise, leading to local and large-scale feedbacks (Lines XX-XX).

Lines 243-245 repeat the discredited notion that all warming experiments generate phenological responses that differ from observational responses at the same site.

We thank the reviewer for pointing out our need to be more clear: we did not wish to suggest that *all* warming experiments generate responses that differ from observational responses, and have reworded the sentence to highlight that this discrepancy can exist and has been observed in diverse species, warming types, and locations (Lines XX-XX).

A sentence in the paper that I do agree with is lines 265-267, which points out that in some locations

global climate change may cause soil to be wetter not drier. This has been pointed out many times in the past and, in full realization of that possibility, some researchers augment heating with water additions. I suppose it is helpful to continually remind ecologists of this basic point.

Please see our response to Reviewer 3 for more discussion on this.

In summary, this could conceivably have been a useful paper 30 years ago, when the first active heating experiments (e.g., The Harte subalpine meadow RMBL experiment and the Melillo forest soil warming experiment) were under design. But those early experimentalists fully grasped at the time all the points listed in a – d above. I cannot speak to the understanding possessed by all experimenters (anymore than one can know the intentions of all of them) but my reading of the literature suggests that this paper will not provide new insight, except barely possibly to those who have not been engaged in or thought about climate-ecosystem interactions.

REVIEWER 2 *Overall, this is a well written and long overdue paper examining micro-climatic variation, experimental artifacts and indirect consequences of global change experiments. It shines a spot light on the imperfections of these experiments (that they do not necessarily achieve their target effects and they have unintended indirect consequences); highlights the need to better quantify direct, indirect and unintended consequences; advocates for experimental artifact controls as well as ambient controls; and offers insights on improvements in interpretation of results (Box 1).*

I do believe that individual global change experimental papers typically discuss artifacts and actual vs target amounts of temperature, precipitation (or CO2 or N or S deposition) change in global change experiments, and a tremendous amount of energy, funds and ingenuity have gone into maximizing ability to simulate a prescribed change, and monitoring impacts. That said, experiments and funds are imperfect, and this paper provides some bounds on that imperfection for a small number of experiments.

Funding agencies should take note of this paper, as the inclusion of more sophisticated monitoring systems as well as experimental artifact controls add significantly to the cost of global change experiments. Researchers often need to trade off additional controls or monitoring vs better measuring direct effects, measuring a wider suite of direct effects, or adding true replicates. This paper is a reminder that the research community should not be lulled in to accepting these trade-offs.

That said, the paper would have been more compelling with more case studies and better separation of the case studies between above and belowground heating.

More case studies: I appreciate that this is a conceptual paper, using a handful of studies to illustrate points. Perhaps the authors can comment that these types of artifact issues should receive more attention in future, more robust syntheses or meta-analyses.

We agree that additional case studies are needed and have added a sentence about the need for additional syntheses or meta-analyses (Lines XX). We have also added three additional studies to our database, to broaden the scope of our paper a bit.

Above vs belowground heating: We know that aboveground heating techniques do not do a great job of warming soils, and that warming soils does not warm air. The authors need to do a better job of discussing this, and separating out expected impacts of above and belowground heating, and above and belowground experimental infrastructure artifacts. Perhaps the experiments can be named in such a way that the reader can separate this out in the figures. I found myself going back and forth between the figures and tables to try to understand what experiments were driving the trends.

We greatly appreciate the reviewer's suggestion of separating above vs. belowground heating (a similar point was made by Reviewer 3). In the our revised manuscript, we have taken the following steps to address this valuable point:

1. Using different shapes to represent different warming types in Figures 2 and 3.

2. Adding "warming type" as a predictor to the following analyses: (Table X), (Table X).)
3.

More detail on where measurements are made: We know that the soil warming experiments will best achieve their target temperatures at the depth to which they are controlled (typically 5-10 cm depth), and that the warming will decline with depth. I expect the same is true for air temperature. It is not clear from the figures at what depth soil (or air) temperatures were measured.

We have worked to include more detail on measurements and experimental design in Table S1, as well as throughout the main text of the manuscript, including in Figures 1-3.

The word ?climate?: Climate is defined as the statistics of weather, usually over a 30-year interval. So, the title is misleading, as is the name of the ?C3E? acronym. Climate change experiments don?t change climate. They change microclimate or climatic variables in an attempt to simulate future or alternate climates. For example, in line 22, ?To investigate how climate change experiments actually change climate?, climate change experiments don?t change climate, they alter micro climate, or key components of climate to simulate that change. And Climate from Climate Change Experiments, is really MicroClimate from Climate Change Experiments. I find the title ?cute? but not accurate.

Not sure what to do here...

Specifics on Figures and Tables. In general I found the figures difficult to interpret, perhaps because the authors lump above and belowground heating techniques which will have a very different impact on above and belowground warming, as well as indirect consequences. Perhaps the authors can differentiate between above and belowground heating experiments (a different symbol?).

Figures 1 and 2. Can authors use ?air? temperature instead of above ground temperature? Unless the aboveground temperature is tissue temperature. ?Air? temperature would be parallel to ?soil? temperature.

We agree that air temperature is a more clear parallel to soil temperature, and thank the reviewer for pointing out that this may be unclear to readers. However, infra-red heating (the warming type used in the majority many experiments in our database) does not directly warm the air and air temperature is frequently not monitored in these experiments. Instead, surface temperature or canopy temperature are often measured and reported. For this reason, we show above-ground temperature, which includes air, canopy, and surface temperature. We have tried to make this more clear to the reader by....

Figure 2. Again, can the author?s distinguish between air and soil warming studies? Soil warming studies are going to do a much better job at controlling soil temperature across time and blocks, and are not as susceptible to aboveground microclimate factors such as wind.

We now use different shapes to represent different warming types in this figure

Figure 4. Soil moisture is notoriously difficult to measure, and microsites right next to each other could have different soil moisture contents in time and space depending on soil organic matter distribution, root channels, etc. This inherent variability should be better noted.

We thank the reviewer for making this point. We now discuss the variability within plots, as well as across treatments (Lines XX-XX).

Table S1 and S2. Can these be combined? I find myself trying to concatenate them.

We have now combined these into one large table (Table S1).

Table S2. Is Target warming air or soil targets? At what depth in soil, or height above the soil?

This information is now in Table S1. We specify whether target warming was air or soil (if designated in

the cited reference) and the depth at which soil and aboveground (air, canopy, or surface) temperatures were measured.

Note on accessing data: It strikes me as rude to call out the authors who either declined to share data or didn?t respond. There are many reasons for declining or not responding (including change of address), and perhaps there is a better way to phrase this:"Note that we were unable to include the following studies because authors declined to share their data or did not respond: (Schwartzberg et al., 2014; Moser et al., 2011; Caron et al., 2015; Ellebjerg et al., 2008)." There are also dozens of other studies that could be included, or contacted. Why just call out these. This seems to put the negative onus on the authors of the other studies, rather than the authors of THIS study didn?t dig deeper to get more data!

The sentence about studies that did not share data was included in our original manuscript for two main reasons. 1) It is useful and important information for potential users of the database, who may need to know how comprehensive the data are and why some data are not included in our database. 2) An earlier reviewer of the manuscript asked why one or more particular studies were not included in our database. We followed standard protocols for meta-analyses to identify potential studies for the C3E data base (i.e. we used reproduceable search terms, and had specified rules for inclusion). We then contacted all authors of these potential studies; some authors did not respond, despite multiple requests, and some authors declined to share data. For this new version, we have added three additional studies, found by searching publicly available databases for data that have been added since our original literature review (using the same search terms). We have removed the sentence from this manuscript, and include it with the meta-data of the database, so that the information is available for those who need it.

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 thank the reviewer for drawing attention to lack of clarity about our rules for inclusion in the previous manuscript version. We have tried to emphasize more clearly in this new version that our database focuses on active warming studies: any climate change experiment that actively controls temperature by warming it. In lines XX-XX we say : We focused on *in situ* active-warming manipulations because recent analyses indicate that active-warming methods are the most controlled and consistent methods available for experimental warming...We do not include passive warming experiments because they have been analyzed extensively already and are known to have distinct issues, including extreme reduction in wind, overheating and great variation in the amount of warming depending on irradiance and snow depth...."

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

1. Using different shapes to represent different warming types in Figures 2 and 3.
2. Adding "warming type" as a predictor to the following analyses: (Table X), (Table X).)
3. Adding a table (SX) that compares warming in passive OTCs (from XX), to the active OTCs, soil cable warming, and infrared heating studies in our database.

(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'.

We thank the reviewer for this point and have re-written the section. Please see below under 'specific points' for details.

(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 belowground; and which studies attempted to maintain a set elevation of temperature versus applied constant inputs of energy (e.g. constant wattage for infrared heaters).

We thank the reviewer for this suggestion, and we have taken the steps described above (in part A) to identify and interpret different heating methods.

(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.

We agree with the reviewer that the previous version of our manuscript did not adequately reflect the diverse goals associated with climate change experiments, especially their inclusion of indirect effects. We have rewritten much of the introduction and discussion to address this. Here are a few specific areas where we have modified text:

(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,

The reviewer brings up an excellent point! We have now added discussion of two cases when responses to experimental warming treatments appear to match responses due to climate change observed in long-

term plots (altered daily temperature range and shifts from non-woody to woody vegetation, Lines XX-XX) .

(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.

NEED TO ADDRESS THIS

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).

We have removed the word "unintended" from this sentence in the abstract.

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 reviewer's point is critical. We agree that an enormous challenge of using data from short-term experiments to understand future responses is in how we interpret the experimental results. We have worked to provide a more nuanced presentation of the challenges associated with interpreting experiments in light of potential differences between short- and long-term responses by adding the following:

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?

We have rewritten this statement so to read:

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 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 (Figure SX). We have also modified the legend of Figure 1, so that it now states the warming methodology of each site, as well as the depth at which temperature was measured. Table S1 gives additional methodological details.

5. 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 mall 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).

The SPRUCE data were not included in the earlier version because no studies had been published at the time of our original literature review that met our search criteria. At the reviewer's suggestions, we searched the SPRUCE online database () and contacted SPRUCE authors about phenology and microclimate data. We were told that "Phenology data from the SPRUCE warming study will be posted soon for full public consumption and use, but are being held pending publication of a couple of project articles currently being considered by journals." We also attempted to get daily microclimate and phenology data from a study that predates SPRUCE (Gunderson et al 2012); however we were unable to acquire the daily (non-summary) microclimate and phenology data, as the original first author of the manuscript has retired.

6. Line 181 Most warming experiments calculate focal response variables relative to ambient controls (e.g., Price and 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 experimental infrastructure and from there decide if they are minimal. The studies that we include in our analysis, as well as a additional studies we have come across in reviewing literature and writing this manuscript, suggest that we do not yet know all the ways that experimental infrastructure affects biological responses so we should not assume that effects are zero 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.

8. 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 this section (Lines XX-XX) in an effort to more clearly make this point.

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.

To do: 1. Add type of aboveground temp measurement to Table S1 2. Add whether target is soil or air or no target 3. Add to Figure 1: warming methodology of each site, as well as the depth at which temperature was measured 4. Email Christy and ask her about progress...

References mentioned in response to reviewers

1. Kimball 2005. *alksje* 5:148-152.
2. Kimball et al. 2008.
3. McDaniel et al. 2014.
4. Pelini et al 2011.
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 com-position caused by 20 years of experimental warming vs. 13 years of natural climate shift. *Ecosphere* 5:1-17.