We thank the editor and reviewers for their helpful reviews; the manuscript is stronger because of them. The comments are below, followed (in bold) our responses. All line numbers by us reference the marked-up copy of the revised manuscript.

EDITOR

Comments to the Author:
Thank you for addressing the reviewer comments. Both reviewers were positive about the paper and I also think it is an important contribution to the paleo-CO2 literature. The revisions proposed address most of their concerns. I note a few issues below that I would like to see addressed before final acceptance and ask that you submit a revised version of the paper that includes the revisions you proposed and addresses these further comments:

1) Line 98-101. Can you clarify this discussion of uncertainty? The issue is whether all studies using only stomatal density ignore the two sources of error you mention, or just some do. The wording here is not clear on that point.

We have clarified this point (lines 99-101). Very few studies propagate both sources of uncertainty (the Beerling 2009 study, mentioned at the end of the sentence, is one such study).

Line 102. Smaller than what? With fossil taxa? Please clarify.

We have clarified this point (smaller than with gas-exchange proxies; line 103).

Line 116-118. This sentence strikes me as awkward it refers to "elements" but then phrases them as questions. Can you reword?

Done (lines 116-117)

Line 121. Can you provide a more informative title for the table?

Done (line 122)

Line 170-171. How are the ambient CO2 values known?

References added to the Mauna Loa and Harvard Forests databases (lines 171-173).

Line 239. Regarding Milligan et al., please just provide the details of the method. Even if the paper is in press the reader would benefit.

Done (lines 242-243). Also, the Milligan paper is now published, with a doi.

Line 320-321. I think it would help here to explain that you do not have measurements of d13C-CO2.

We now note this (lines 324-325); we also mention this in the results (lines 508-510)

Line 342-346. I understand why you use the 2/3 range, but you should explain here why. Also, what does it mean to say that 28% overlap the target at 95% confidence if they are really noisy (that is, is this a

useful statement). And, how is the target value established? Assumed 400 ppm?

We now say that the 2/3 range is a close equivalent to +/- 1 standard deviation (lines 349-350). The target value is assumed to be 400 ppm (we now state this directly on line 171).

We agree that this is a noisy signal, and the reader can decide whether the signal:noise ratio is good enough for their purposes. We included this statement on the account of a direct request by one of the reviewers.

Line 526. Can you replace "tough" with a more specific term?

We replaced this word with "durable" (line 533) (toughness is a quantitative trait made by plant ecologists).

With best wishes, Ed Brook

REVIEWER #1

General comments: Fossil leaf gas-exchange based CO2 models are currently going through the "rigorous testing" phase and as the authors of this paper point out, this mechanistically, rather than empirically calibrated proxy, shows considerable promise. It is therefore of high relevance that studies, such as this one, are presented that provide quantification of potential confounding factors. In this case, the authors test three potential confounding factors (photorespiration, leaf temperature and canopy position) and provide quantifications on how these factors influence final CO2 estimates. They are capable of eliminating two of these factors as insignificantly affecting CO2 estimates (photorespiration and leaf temperature). The third factor, canopy position, is determined to strongly skew CO2 estimates, but the authors point out that it is possible to identify leaves that grew in lower canopy positions, based on leaf micromorphology and an uncharacteristically wide δ13C range. This paper is a relevant contribution towards quantification of the potential error in fossil leaf gas-exchange based CO2 models, and apart from minor suggested amendments, I have no problem with seeing this study being published.

Specific comments: In the materials and methods section, the authors lay out the specific ways that they are testing modern plants for potential bias in reconstructed CO2. In the appendix all the specific plants are listed with their input values and reconstructed CO2. However, from reading the methods section I get the impression that not each plant is being tested for the same potential confounding variable (photorespiration, leaf temperature and canopy position). It would be very helpful if there was a table that outlines specifically which plants were tested for what, or at least that this was made clear in the appendix, because in the main body of text it is hard to follow.

We now include this information in column E of the supplemental table.

In several places in the manuscript, including the abstract, it is mentioned that the random error propagation of the Franks et al. gas exchange model is better than uncertainty estimates of other leading paleo-CO2 proxies. It would be very helpful for the untrained reader to see some proof of this

statement in the form of a table that lists 1) the different CO2 proxies, 2) a method of error quantification, 3) the actual amount of uncertainty in those CO2 proxies and 4) the references to the case studies where this was tested. Such a table would lend credibility to the statement that gas-exchange models are quantifiably better than other CO2 proxies.

There are of course two elements of uncertainty: precision (spread of possible solutions) and accuracy (comparison to true answer; can only be quantified for times when CO2 has been measured). The abstract brings up the theme of accuracy (28% mean error rate). In the main text (section 3.1), the mean error rate is compared generally to that in other CO2 proxies by referencing the summary work of Franks et al. (2014).

The error propagation scheme noted by the reviewer is related to precision. We only mention precision in the Introduction by referencing what others have found (Franks et al., 2014). It is not a focal point of the current study.

The reviewer may (also) be referencing the paragraph in the Introduction where we argue that studies using other stomatal-based proxies probably overstate the accuracy and precision of their CO2 estimates (lines 98-107). Our arguments here are conceptual only—there are no data we can summarize in a table, unfortunately. The point we are trying to make is that the reported accuracies and precisions associated with these other methods—when applied to plants living today (not fossils)—are better than what we find with gas-exchange methods. But this is partly because these other methods are based on empirical calibrations with…present-day plants. So excellent accuracies and precisions are not particularly surprising. But when you apply these other methods to fossils that are millions of years old, the present-day empirical calibrations are likely less appropriate.

Final specific comment is on the title itself, for which I would like to suggest that the authors include what specifically is being tested. I.e. “Sensitivity of . . .. CO2 concentration to x, y & z”. There are other variables that the model is sensitive to and I believe the title would be more informative if the specifics were included.

The largest block of data (40 species) is “general” testing, that is, estimating CO_2 from field-grown trees without isolating any single confounding factor (summarized in Figure 2). Thus, it would not be fully representative to say that we were only testing the model for the influence of canopy position, temperature, and photorespiration.

Technical corrections: I could not find any spelling or styling errors in the manuscript. The paper is very well constructed and easy to follow.

REVIEWER #2

The authors present a sensitivity analysis of a mechanistic model (Franks model) to predict paleoatmospheric CO2. They explore several specific areas; the effect of gc(op)/gc(max), A0, temperature, photorespiration and leaf canopy position on the accuracy of CO2 estimates produced by the model. In doing so, the paper adds clarity, certainty or recommendations to the model for fossil application, all of which are important additions, especially as this model is being using in a growing number of research projects. Although the paper is an important contribution, it would benefit from clarity or expansion in certain areas:

1) Aims, methods and appendix: The aims and methods section is hard to follow. This may be due to the fact the aims and rationale are mixed in with the methods. It is unclear from the text or appendix data whether all or a subset of the data is being used for each of the analysis performed. A summarised table in the methods section containing the information on the analysis being performed, data source and parameters used or tested would be beneficial (i.e. a summary of the methods in tabular format). Similarly, in the appendix, additional information on the origin of the data, sample number per species, which data points/values are measured vs estimated/assumed and a direct comparison of measured vs model estimated CO2 would greatly improve clarity.

We now present a tabular summary of our study design (new Table 1).

In the Supplemental Table 1, we now give the sample size (column F), the target (i.e., correct) CO2 concentration (column G), and whether the input was measured or inferred (color coding of column headers). And column E gives what part of the study was addressed (general testing, temperature, or canopy position; reviewer #1 also asked for this information). We are not sure what is meant by "additional information on the origin of the data" beyond what is listed in column A and stated in the main-text Methods.

2) Statistical analysis: Accuracy was evaluated by the degree of error rate. These claims can be strengthened by using statistical analysis. How well the model predicts CO2 could be assessed by whether or not the estimates are statistically significant different (or hopefully not) from measured CO2 values.

We have added information about whether individual estimates depart from the target CO2 concentrations (lines 350-352 and 425-427).

3) gc(op)/gc(max) and A0 (section 3.1): This section gives details about when both gc(op)/gc(max) and A0 values are either known or values from Franks et al. 2014 are used, but it would be nice to see these two parameters evaluated separately i.e. how much does gc(op)/gc(max) alone improve estimates and the same for A0. Does one contribute more than the other for improving error rates?

We have added this information (lines 357-358).

Additional comments:
Line 86. Sensitivity saturates for some but not all taxa. See Haworth et al 2011.

We have added the qualifier "in many species".

Line 93. A Nearest living relative or equivalent approach also get around the issue of extinct taxa.

This is true for the stomatal ratio method, but these CO2 estimates are not meant to be quantitative in the same manner as estimates from the "full calibration" methods or the gas-exchange methods (as noted in the previous paragraph).

Line 156. Alternative approaches for fossils have been suggested such as estimating fossil A0 using scaling relationships between vein distance and assimilation rate however they are not discussed here (EG Montanez et al., 2016).

We have added a citation to the Montanez paper

Introduction – general comment. Critical published assessments of the Franks model are not cited (eg McElwain et al. 2016) yet they raise issues associated with parametrization of A0 and the insensitivity of CO2 estimates to variation in gamma star values which are both important discussion points in this manuscript in lines 454 -456 and 497-499.

As per a later comment, we have added a citation to McElwain et al. 2016 regarding gamma star on line 472.

Our study does not focus on the parameterization of A0, and so the associated literature does not seem relevant to the Introduction. Our study focuses on temperature, photorespiration, canopy position, as well as a general and broad test of the method.

Paragraph 201-217: A some information is missing here: chamber model/make, duration plants were grown in the chamber, light levels. What were measured vs set chamber conditions for temperature, light and CO2 (i.e. similar to how humidity is reported)

Chamber make/model (lines 216-217) and duration of experiment (line 233) are given. We have added information about light intensity as well as the standard deviations for temperature and CO2 concentrations in lines 217-221.

Lines 232: Stomatal density/stomatal measurements and leaf stable carbon isotopes were performed on the same leaves. Clarify how this was partitioned, e.g. was the leaf divided into 2 or was a whole punch used for carbon isotopes, etc.?

We now clarify our methodology in lines 241-242. We used either a hole punch or razor to remove two adjacent sections of leaf tissue near the leaf centers, avoiding major veins.

Lines 235: As Milligan et al is in review, I suggest adding more detail here on how δ13Ca of chamber CO2 was calculated. δ13Ca values of supplemented CO2 can be very negative and can vary between cylinders, unless the CO2 gas has a specific δ13Ca. What is the capacity of these cylinder, in L?

This paper is now published. In short, a mixing line was established based on direct d13C measurements of lab air, chamber air, and cylinder CO2 (= pure CO2). We were fortunate that the d13C of the cylinder was close to the well-mixed atmosphere (the d13C in most cylinders we have used in other experiments is much more depleted). We used only the single cylinder for the duration of the experiment. The target CO2 concentration (500 ppm) was not much higher than the CO2 concentration inside the lab (~440 ppm), so we did not use much CO2.

Figure 1: Does this need to be on a log scale? 1000 or 2000ppm are not very high values and the log scale visually skews data and error bars. A difference plot between measured and estimates plotted on a non-log scale would improve this figure.

We prefer a log scale because it is easier to differentiate estimates at the low-end of the CO2 scale, and because the uncertainties scale in a logarithmic fashion.

Line 351: Please provide supporting data for this statement in tabular form. What are the error rates of other proxies?

This information was summarized by Franks et al. (2014), so we prefer not to repeat it here.

Line 355: Might be helpful to report standard deviation of CO2 estimates, here and throughout the text.

We now report the range that encompasses two-thirds of all estimates (lines 350-352). (Because the individual estimates are not normally distributed (tail at the high end), reporting a standard deviation can be misleading.)

Line 411 to 413. Reporting of the difference between estimated and measured CO2 here is incomplete. Only means of all species investigated are provided rather than species-based diffeences or errors. For some species the error is substantial whereas other taxa show very small errors.

As per an earlier comment, we now report the species-level differences on lines 425-427; no individual species-level test was significant (line 414).

Line 454 to 456. This supports the findings of McElwain et al 2016 Paleo 3 but it is not cited. “This compensation point (Γ * in Eq. (2) is temperature, species and O2 dependent (Ethier and Livingston, 2004) but Franks et al. (2014) account only for the temperature dependency in the new paleo-CO2 proxy model. Allowing Γ * to vary in response to prevailing paleoatmospheric O2 concentration [O2] (Γ* = 1.78 × [O2]), which is known to have varied widely (10% to 30%) through the Phanerozoic (Bergman et al., 2004; Belcher and McElwain, 2008; Berner, 2009), would increase the precision of paleo-CO2 estimates but only fractionally.”

We have added a citation to McElwain et al. (2016 Palaeo3) (line 472).

Lines 500 to 506: A number of papers have suggested methods of estimating A0 to improve the accuracy of CO2 estimates using the Franks model but they are not discussed. This section would provide a good opportunity to discuss the proposed ideas and solutions.

This section deals with living leaves, where A could be measured directly. Measuring A wasn't part of our study design, unfortunately. In this section we are discussing possible reasons for noise in our mixing-model calculations. With regards to fossils, we are not recommending that our mixing model be used (line 527: "We note that our mixing-model strategy cannot be applied to fossils because…"), so the question of how to constrain A in fossils within the context of the mixing model is moot. Our take-home message for fossil applications is to avoid shade leaves (line 535), and we provide specific measurements that can be made on fossils to make this distinction, including vein density (lines 536-540).

Section 3.4: Have any values for δ13Ca been measured or are all calculated for this section? Is there any data set (from the literature or otherwise) this could be compared to? i.e. a dataset where known δ13Ca is compared to itself when calculated as per the manuscript? This would strengthen this section. If δ13Ca has only been calculated/inferred for this section without a comparison to measured δ13Ca I think claims on the effect of δ13Ca (or low canopy plants) on the model should be softened.

We made no direct measurements of understory d13Ca (multiple measurements over a growing season, and at different daytime hours, would be needed to calculate a representative mean value). As the reviewer correctly notes, we instead are assuming a well-behaved two end-member mixing model. We have added a note of caution related to this on lines 508-511.

Appendix: The authors used both known and general values for gc(op)/gc(max) and A0 to evaluate error rates but no measured values of either gc(op)/gc(max) or A0 are given in the appendix or text.

The Appendix summarizes all new data presented in the study (with the key graphics being Figures 2, 5, and 7). For these data, we *only* used "default" values of gop/gmax and Ao; that is, we did not measure these inputs on our leaves. As noted in the Introduction, this was a purposeful strategy because we wanted to test the CO2 model in a manner that would be similar to how most (but not all) folks will be applying the model to fossils. A "worst-case" test, if you will.

In the Introduction, we do summarize some of the already-published data (Figure 1). For these estimates, either gop/gmax or A0 were measured, and in most cases both were measured (lines 143-146). These data are not in the Appendix because they are already published and are not central to our study.

As the reviewer noted, we did additionally "degrade" these estimates by re-doing them assuming default values for gop/gmax and A0. We did this so we could compare them more directly to our estimates (lines 355-357).