Fossil Atmospheres: a case study of citizen science in question-driven palaeontological research

Article *in* Philosophical Transactions of The Royal Society B Biological Sciences · January 2019

DOI: 10.1098/rstb.2017.0388

4 authors, including:

Richard S. Barclay
Smithsonian Institution
39 PUBLICATIONS **282** CITATIONS

SEE PROFILE

Scott L Wing
Smithsonian Institution
193 PUBLICATIONS **8,467** CITATIONS

SEE PROFILE

PHILOSOPHICAL TRANSACTIONS B

rstb.royalsocietypublishing.org

Research

Cite this article: Soul LC, Barclay RS, Bolton A, Wing SL. 2018 Fossil Atmospheres: a case study of citizen science in question-driven palaeontological research. *Phil. Trans. R. Soc. B* **374**: 20170388.
http://dx.doi.org/10.1098/rstb.2017.0388

Accepted: 14 October 2018

One contribution of 16 to a theme issue 'Biological collections for understanding biodiversity in the Anthropocene'.

Subject Areas:
palaeontology, plant science

Keywords:
Ginkgo biloba, carbon dioxide, Zooniverse, palaeoclimate, citizen science

Authors for correspondence:
Laura C. Soul
e-mail: soull@si.edu
Richard S. Barclay
e-mail: barclayrs@si.edu

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.4274645.

THE ROYAL SOCIETY PUBLISHING

Fossil Atmospheres: a case study of citizen science in question-driven palaeontological research

Laura C. Soul[1,2], Richard S. Barclay[1], Amy Bolton[2] and Scott L. Wing[1]

[1]Department of Paleobiology, and [2]Department of Education and Outreach, Smithsonian Institution, National Museum of Natural History, 10th & Constitution Avenue NW, Washington, DC 20560, USA

LCS, 0000-0001-5352-886X; RSB, 0000-0003-4979-6970; AB, 0000-0003-2527-1985; SLW, 0000-0002-2954-8905

Palaeontologists increasingly use large datasets of observations collected from museum specimens to address broad-scale questions about evolution and ecology on geological timescales. One such question is whether information from fossil organisms can be used as a robust proxy for atmospheric carbon dioxide through time. Here, we present the citizen science branch of 'Fossil Atmospheres', a project designed to refine stomatal index of *Ginkgo* leaves as a palaeo-CO_2 proxy by involving citizen scientists in data collection through the Zooniverse website. Citizen science helped to overcome a barrier presented by the time taken to count cells in *Ginkgo* samples; however, a new set of challenges arose as a result. A beta-testing phase with Zooniverse volunteers provided an opportunity to improve instructions to ensure high fidelity data. Exploration of citizen scientists' estimates shows that volunteer counts of stomata are accurate with respect to counts made by the project's lead scientist. However, counts of epidermal cells have a wide range, and mean values tend to underestimate expert counts. We demonstrate a variety of approaches to reducing the inaccuracy in the calculated stomatal index that this variation causes. Zooniverse serves as an ideal tool for collection of palaeontological data where the distribution of fossils would be impossible, but where specimens can be easily imaged. Such an approach facilitates the collection of a large palaeontological dataset, as well as providing an opportunity for citizens to engage with climate research.

This article is part of the theme issue 'Biological collections for understanding biodiversity in the Anthropocene'.

1. Introduction

Citizen science is an established method for expanding the scale of scientific studies, while engaging the public in the scientific process [1]. In recent years, online citizen science interfaces have been shown to be a reliable way to collect large, accurate and precise datasets that can be used to address a wide variety of scientific problems [2]. The Zooniverse website, in particular, now serves as a hub for online citizen science participation and hosts numerous successful projects from many disciplines, including the physical sciences, medicine, literature and the social sciences [3]. Palaeontological research projects are almost completely absent from online citizen science platforms (although see Fossil Finder; www.zooniverse.org/projects/adrianevans/fossil-finder), though there are many transcription projects to 'digitize' museum specimen labels (e.g. *Notes from nature*: https://www.notesfromnature.org/). This absence is perhaps surprising in light of the rich history of collaboration between those professionally engaged in palaeontology and those traditionally referred to as 'amateur' palaeontologists. As a discipline based largely on prospecting and discovery, within palaeontology there have been several ways in which non-professionals have historically participated. Volunteers are frequently active participants and organizers of fieldwork,

where they find and collect specimens, and prepare them for subsequent research. These citizen contributors have been widely recognized within the professional community as vital to the progress of palaeontology, and each professional society has annual awards specifically for this purpose (e.g. Palaeontological Association Mary Anning Award, Paleontological Society Strimple Award).

Several decades ago, palaeontology shifted as a science from a largely exploratory discipline centred around specimen collection and description, to one that uses these specimens as the foundation for research questions driven by meta-analyses that aim to understand the evolution of life on geological timescales [4,5]. Non-professional collectors and preparators have continued in their involvement, but there has not been an accompanying shift to include such potential citizen scientists in collection of large morphological datasets, analyses, other aspects of question-driven inquiry, or subsequent use of the research outcomes. This is in contrast to other fields within the biological sciences where the involvement of citizen scientists in such activities is highly successful and has become commonplace (e.g. eMammal, Beluga Bits and Cochrane Crowd). The disparity can be attributed in large part to the 'one of a kind' nature of palaeontological data. Each measurement or observation must come from a fossil, and fossils are often rare and mainly housed in museums where access is mostly limited to professionals. There is scepticism on the part of the palaeontological research community, as there has been in the scientific research community more broadly, about having untrained volunteers handle specimens, and whether such volunteers could produce research quality data [6]. The lack of uptake may also relate to the less direct links between the science of palaeobiology and people's everyday lives [7,8]. The ways in which, for example, environmental or ecological monitoring projects can involve citizen scientists in research and policy [9] do not easily extend to palaeobiology, even though palaeontologists regularly employ similarly large datasets (e.g. www.PaleoBioDB.org and associated publications). Although palaeontologists have not worked extensively with citizen scientists, other than for fossil collection and preparation as mentioned above, the field is recognized as an effective gateway to further informal science learning [10] and can act as the beginning of a conversation about complex or controversial topics like evolutionary biology, geological time and modern global climate change.

Here, we present a case study of a palaeobotanical project, Fossil Atmospheres, for which the primary scientific goal is to refine the stomatal index proxy that is used for estimating atmospheric CO_2 concentrations in the geological past. An important step in the research is counting the number of stomata and epidermal cells on areas of leaf surface of standardized size in order to obtain stomatal index [11]. This step is a time consuming part of the research which results in a backlog of images to be processed. The lead project scientist takes an average of 14.5 min to classify each image, which equates to 73 eight-hour work days to classify all of the images (2424) we have currently made available on the Zooniverse, and is approximately one-fifth of the number of images that will eventually need to be classified in order to address the scientific research question. The research goal for the citizen science branch of the project presented here was to investigate the efficacy of citizen scientists for the collection of the data that is required from these images to achieve the research objectives. This was with a view to continuing to involve citizen scientists in the project, should the method prove to generate research quality data and be time effective.

Citizen science organizers should strive to balance the dual purposes of achieving research outcomes and providing a valuable experience for volunteers [1,12]. The project presented an opportunity to bring palaeontological museum specimens out from 'behind the scenes' so that they can be used to include members of the public in palaeontological research outside of traditional specimen collection. This potentially facilitates participation in palaeontology for a far broader audience. For example, it can reach cities, areas that do not have nearby fossil sites, or members of the public who are unable to participate in outdoor physical activity. Additionally, the research focus on climate change from a geological perspective makes Fossil Atmospheres a useful opportunity for climate change education.

In this paper, we outline the scientific background of the research project, the process we used for testing and implementing the citizen science user interface on the Zooniverse website, how we are validating the quality of the data collected by Zooniverse users and the performance and engagement of participants.

2. Scientific background

The Earth's climate during the Mesozoic and Early Cenozoic (100–50 million years ago) was much warmer than today [13] and was often punctuated by geologically short hyperthermal events [14–18]. The background warmth, and particularly the hyperthermals, are often attributed to increased atmospheric carbon dioxide (pCO_2; [19,20]) and are frequently cited as the best analogue situation for modern-day climate change. Despite the many lines of evidence that suggest CO_2 played a primary role in highly elevated temperatures, palaeo-pCO_2 proxy estimates for these time periods are sparse and sometimes inconsistent [21]. This is an active area of research within the palaeo-pCO_2 community, but disagreement among the proxies makes it clear that they are not all correct for all time intervals. Over the past decade, some of the marine and terrestrial proxies for pCO_2 have been revised by evaluating the underlying assumptions for each proxy [22–25], and there has been some convergence in estimates of palaeo-pCO_2 [25].

The primary scientific goal of the Fossil Atmospheres project is to evaluate the assumptions that govern the *Ginkgo* palaeo-pCO_2 proxy and then to apply the revised methodology to fossil material that comes from geological periods of global warmth, as well as the hyperthermal events that punctuate those warm periods. The *Ginkgo* palaeo-pCO_2 proxy depends upon calculating the stomatal index of both modern and fossil material [26,27]. *Ginkgo biloba* (figure 1) is the last living species in an order of gymnosperm trees that originated in the Permian around 270 Ma. The leaves of this order have remained morphologically similar through time and preserve relatively easily in the fossil record owing to their thick waxy cuticle. Leaves from living *G. biloba* individuals can easily be collected, and museums globally house many fossil and historical herbarium specimens.

Stomatal index is an area-independent measure of the proportion of epidermal cells on leaves that are stomatal

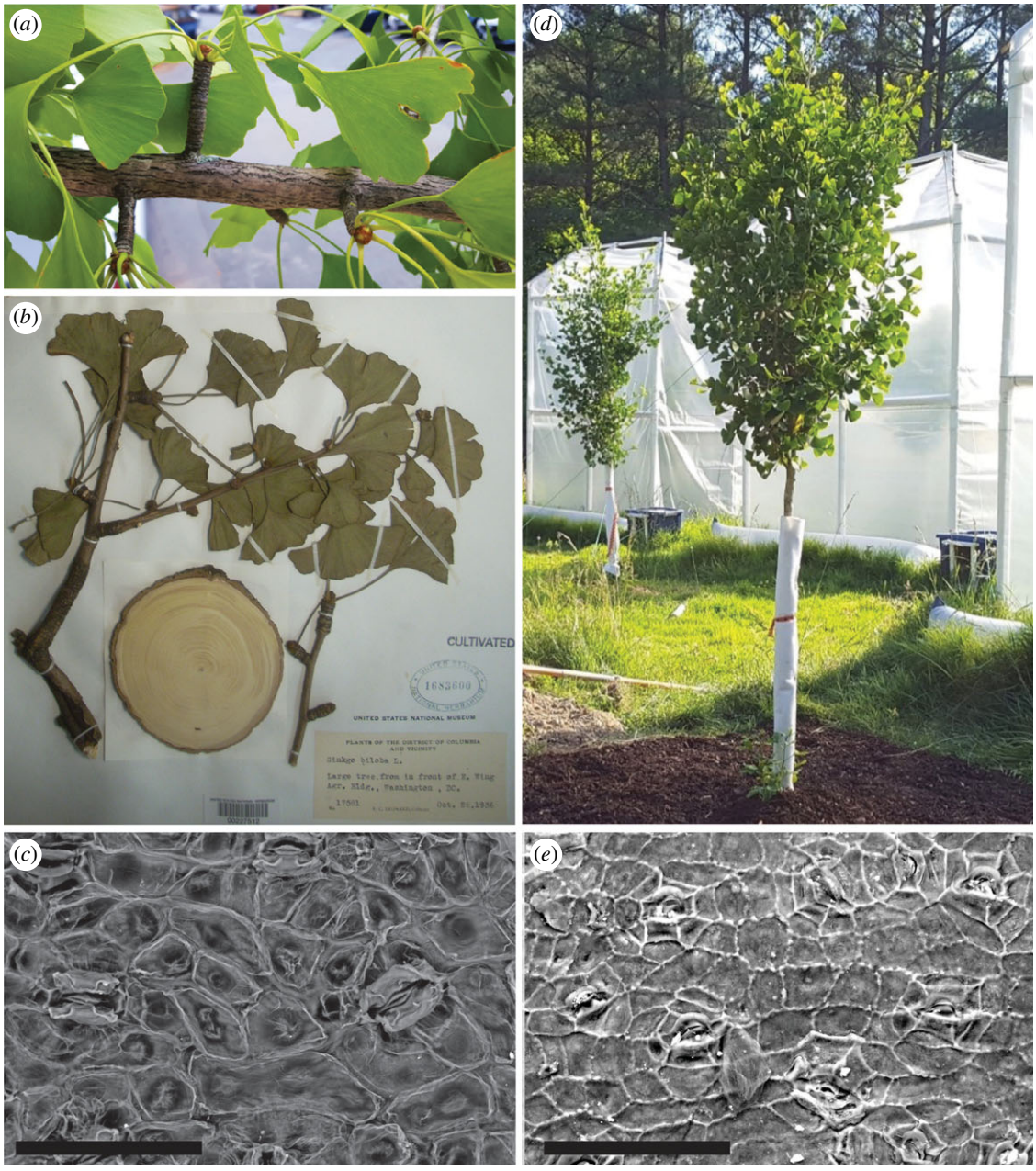

Figure 1. Images of modern *Ginkgo biloba* (*a*–*d*) and fossil *Ginkgo wyomingensis* (*e*). (*a*) Branch of *Ginkgo biloba* showing typical short shoot morphology where many leaves grow from a single node on the stem. (*b*) Herbarium sheet of *Ginkgo biloba* collected on 26 October 1936 near the Smithsonian in Washington, DC (USNM Herbarium Catalog no. 1683600). (*c*) Scanning electron microscopy (SEM) image of inside of lower surface of modern *Ginkgo biloba*. (*d*) Outdoor control tree from the Fossil Atmospheres experiment at the Smithsonian Environmental Research Center in Maryland. (*e*) SEM image of inside of lower surface of fossil *Ginkgo wyomingensis* collected from the Early Eocene of the Bighorn Basin of Wyoming (locality SLW0907). Scale bars, 100 µm.

pores (figure 1; [11]). Historical collections of *G. biloba* demonstrate that the stomatal index proxy for palaeo-pCO_2 is strongly correlated with pCO_2 over the range of 290–430 parts per million [22,28]. However, despite wide application of the *Ginkgo* palaeo-pCO_2 barometer in the past two decades [21,28,29], our understanding of pCO_2 in the fossil record is hindered because the morphological and physiological changes in *G. biloba* stomata under pCO_2 above 400 ppm are poorly constrained [22].

To investigate the relationship of *Ginkgo* to elevated pCO_2 conditions, we began an elevated pCO_2 experiment at the Smithsonian Environmental Research Center in Maryland. This experiment is designed to quantify the response of *Ginkgo* to elevated pCO_2 by growing 15 mature *G. biloba* trees in open-topped chambers in natural field conditions, with outdoor controls, and atmospheres in the chambers of 400, 600, 800 and 1000 ppm of CO_2. Each tree is regularly monitored for changes in stomatal index, and rates of photosynthesis and transpiration, to constrain parameters used in gas exchange models of palaeo-pCO_2. Local volunteers conduct the necessary daily maintenance of the experiment, as well as the data collection from the plants that can only be done at the experimental site. Samples collected from the trees are taken to the Smithsonian National Museum of Natural History in Washington, DC, where they are stored permanently. Selected specimens are worked on by local volunteers and processed to produce the imagery that is available to citizen scientists accessing the Fossil Atmospheres project on the Zooniverse website. Our experimental results will be used to infer palaeo-pCO_2 from stomatal features of Late Cretaceous–Palaeogene fossils of *Ginkgo wyomingensis* (nearly identical to extant *G. biloba*), allowing palaeo-pCO_2 estimates from these terrestrial fossils to be compared with records from other palaeo-pCO_2 proxies.

3. Material and methods

(a) Fossil Atmospheres on the Zooniverse

The Zooniverse (www.zooniverse.org) is a citizen science platform managed by personnel at the University of Oxford (UK) and the Adler Planetarium (Chicago, USA). It currently hosts close to 100 projects, with a citizen scientist user base of hundreds of thousands. Each project provides digital subjects (images, video or audio) that users classify by selecting tools to record observations of different kinds, in the same way that project scientists would. The Zooniverse has a back-end project builder that research teams use to build and beta test a user interface for their project, without the need for web programming [30]. As a possible approach to circumventing the concern of the palaeontological research community about having untrained volunteers handle delicate and sometimes rare specimens, we explored the Zooniverse as an interface through which non-professionals could collect data from museum specimens without receiving training in specimen handling.

Zooniverse users were asked to view scanning electron microscope images of the inside lower surface of leaves from living and historical *G. biloba* trees, as well as fossils, and to identify and mark all the stomata (gas exchange pore structures) and epidermal cells (all other cells) that fall within a box of a standardized size (300 × 300 µm), resulting in counts of their numbers (figure 2). Four tools were made available to do these tasks: point markers for (i) stomata, (ii) epidermal cells, (iii) not sure and (iv) a tool to draw an ellipse around an 'unclear patch'. These data are required to calculate stomatal index as defined by Salisbury [11]:

$$\text{stomatal index} = 100 \times \frac{\text{no. stomata}}{\text{no. stomata} + \text{no. epidermal cells}}. \quad (3.1)$$

The task of counting stomata and epidermal cells is an unusual one for people without botany training, so upon landing on the classification page, volunteers were given a short tutorial explaining the task. We designed the training material to reflect as closely as possible the way in which the project scientists would identify cells, thus contextualizing the training in science practice [31,32]. This tutorial can be accessed later at any time, as can other help features available on the website. These include help tabs for each task and a 'field guide', which we populated with example cell features, suggested counting procedures, and examples of images counted by an expert (R.S.B.).

Counting an image for stomatal index takes longer per image than most other Zooniverse tasks (see Results; [3]). We understood from the outset that this may cause challenges, but this was also identified during the beta-testing phase, a process that is mediated by the Zooniverse team, and provides valuable feedback for a successful launch of the citizen science project. As a result of the beta test, we added an additional question to the workflow: 'Have you marked every cell that you can see in this image'. If participants answer 'no', then this question is followed by another dialogue box with two response options: (i) 'There were too many cells to mark them all' and (ii) 'It was difficult to identify the cell boundaries'. This question was added to allow *post hoc* identification of classifications where the participant stopped part way through without completing, but submitted the classification anyway.

Workflow design can have important downstream effects on both volunteer retention and data quality [33]; we therefore used the beta test to inform modifications to our workflow structure. Following beta-test feedback, we instituted three separate workflows: (i) practice, (ii) easier count and (iii) stomatal count. The workflows are based on the zone of proximal development concept [34], which describes skill building as a range from undeveloped to developed capabilities. Through scaffolding, participants build skills through increasingly complex tasks supported by expert guidance [35,36]. The practice workflow contains five images, all of which are provided in the field guide with the lead project scientist's (R.S.B.) expert classification, for direct comparison on screen. Participants are automatically directed to the practice image set when arriving at the project classification page and are advised to complete the practice set before moving on to 'real' classifications. Feedback from participants during beta-testing suggested that many images were significantly more difficult than the practice set and therefore served as a barrier to continuing participation, causing people to give up. For this reason, the lead project scientist selected a set of 100 images from the full image set that show distinct and more easily identifiable cell boundaries. These 100 images constituted the 'Easier Count' workflow, as well as serving as the validation set for the current study to test citizen scientist accuracy relative to expert stomatal index estimates. Images were circulated until each had been classified by 50 unique project participants. The ideal number of repeat classifications for each image should minimize volunteer time spent, while maximizing accuracy and confidence. We began with 50 repeats so that these data could be analysed to determine an optimal number of repeat classifications beyond which the mean stomatal index estimate for an image did not change substantially.

A concern about 'incorrectly' identifying the cells and generating 'wrong' results for the researchers was frequently raised in the beta-testing feedback. Addressing motivation and self-efficacy is important for ensuring participants continue with the task as they build skills and confidence [37,38]. For this reason, the 'not sure' and 'unclear patch' tools were developed so that uncertainty about small areas of the image would not become a barrier to classifying an image and moving on to the next one. To encourage people to continue in spite of some uncertainty, we also included language about the need for multiple classifications, a consensus from many people, and the idea that scientists are also subject to biases and inaccurate identifications. There are also message boards through which participants can ask questions to clarify how to mark the images, or find out more detail about the scientific background of the project.

Although there is a true number of cells in each image, interpretation of where cell walls are visible, as well as which cells project into the standardized box (300 × 300 µm), is somewhat dependent on the person completing the classification. Because of this small amount of subjectivity in the process, there is the possibility of researcher bias in counting when the researcher has expectations for the range of the stomatal index based on knowledge of the source location for the sample. Zooniverse has an important advantage in this respect in that all images are presented without any metadata, which means that the participants are 'blinded', thus limiting inherent biases.

(b) Data analysis

Zooniverse allows data downloads that include the image identification number, username (if the user is logged in), the output from each individual classification of a subject and other associated metadata. The classification output is provided in a JSON string, which we processed using an R script (electronic supplementary material) to obtain stomatal count, epidermal cell count and point coordinates for each marker used in each classification for each image. The 100 images in the 'Easier Count' workflow were classified by the lead project scientist using the Zooniverse project interface. We used this information for exploratory analyses to better understand citizen scientist accuracy relative to expert classifications, and how to maximize data quality. Each of the 100 images was classified by 50 unique users (5000 classifications), and these data are the basis for all our analyses. Where the data were filtered, the sample size for analyses is slightly lower.

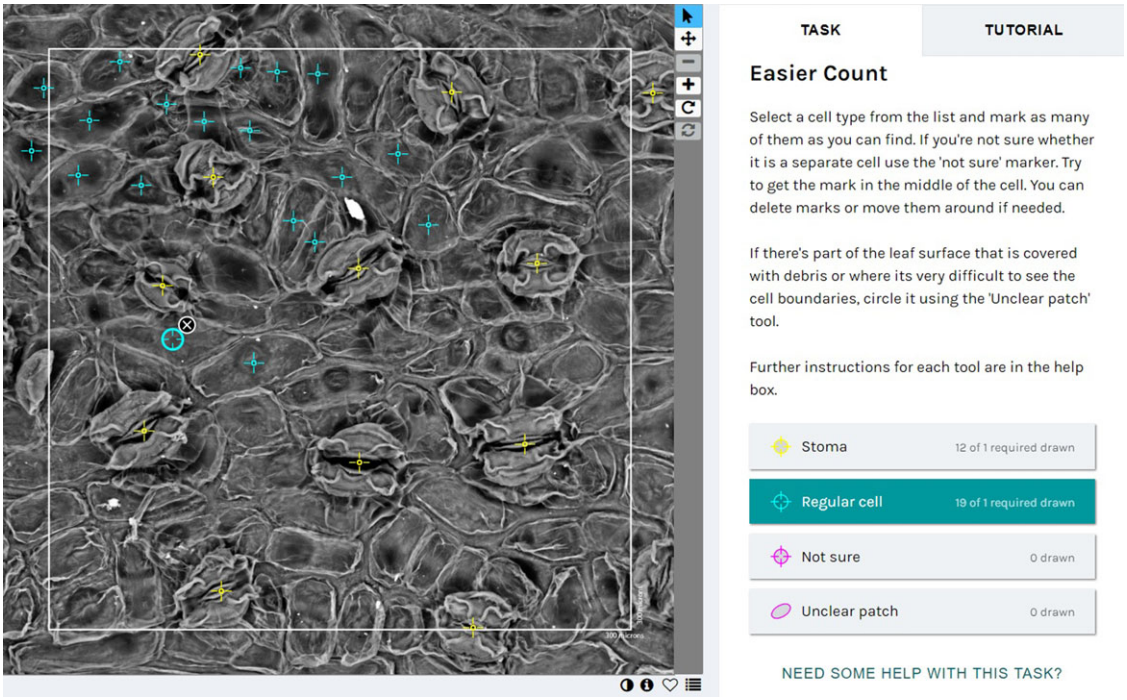

Figure 2. User interface for counting stomata and epidermal cells on the Zooniverse citizen science website (www.zooniverse.org/projects/laurasoul/fossil-atmospheres). These two cell types are all that is required to calculate the stomatal index for this image. The white 'counting box' over the SEM image is standardized for all images (300 × 300 µm). Participants are also provided with tools that allow for marking places on the images where the features are not obvious to them. Beta-testing demonstrated that without these 'unsure' tools, participants that were not able to identify the features with confidence often stopped analysing images.

To identify what the optimal number of repeat classifications is to balance loss of information and accuracy against use of participant's time, for each image we plotted the difference between the running mean and the final mean, as classifications were added. We then calculated the average of these differences across all images to find a reasonable cut-off. A gradual approach to the mean is expected; therefore, there is no non-arbitrary way to choose a cut-off, but we identified the point at which difference from the final mean was within one stomatal index unit.

To find the individual and average error in citizen scientist estimates of stomatal count, epidermal cell count and stomatal index, we compared citizen scientist estimates with expert estimates and calculated the differences for individual classifications, average differences for individual citizen scientists and average differences per image. We then filtered these data in several ways and compared original with filtered data using a *t*-test, to find whether filtering could reduce the error relative to expert classifications. We removed classifications from users who were not logged in, removed classifications that only had one stoma or one epidermal cell identified, and finally added 'not sure' counts to epidermal cell counts.

To investigate whether participants improved with experience, we compared the error in individual classifications (relative to expert classifications) with the experience a participant had (number of classifications made) when completing each classification, using linear regression, with individual participant as a fixed effect. Additionally, to find whether, and how, each individual participant improved in accuracy on classifying more images, we performed a breakpoint analysis using the R package *segmented* [39], which estimates the number and location of different regression relationships in time-series data. This analysis was applied to the same data as the regression analysis, separated by participant. This analysis can be used to identify whether there is an initial improvement followed by consistent accuracy, by modelling the data with two different regression slopes and measuring fit. If the point at which this switch in slope occurred was consistent across participants, it could be used as a 'burn-in' to discard inaccurate early estimates. We applied the breakpoint analysis for the 55 individual participants who had classified at least 20 of the validation image set of 100 images.

Finally, to investigate a possible route through which participants could be filtered prior to making full classifications, we tested whether an individual participant's error in stomatal count estimate could be used as an indicator of the likely error in estimating stomatal index, using a regression of mean stomatal count error against mean epidermal cell count error per participant.

4. Results

As of this submission Fossil Atmospheres has 2500+ participants who have collectively made 24 000+ individual classifications of 2424 images. The 'Easier Count' workflow was opened on 20 December 2017, and all images within that workflow had been classified by 50 unique users by 18 March 2018, in around two and a half months. The average time elapsed to complete an image on Fossil Atmospheres was 21.1 min, longer than the lead project scientist's average time of 14.5 min. The mean number of classifications submitted across all participants was 6, the most common number of classifications submitted was 1, and the maximum was 2449, which included repeat classifications.

The average of running means for the 100 images in the validation set approached the final mean gradually as expected (figure 3) and reached a difference of within one stomatal index unit of the final mean after around 30 classifications. In the future, the retirement limit will be set to 30 to reflect this.

Across all classifications by individual participants for the 'Easier Count' workflow, before filtering ($n = 5000$), the expert estimate of number of stomata (number counted by the lead project scientist) was identified in 36% of cases and was within one unit of the expert estimated count in 64%

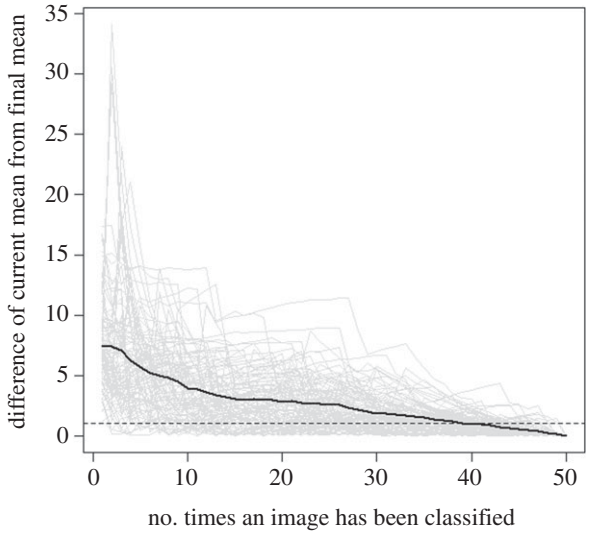

Figure 3. Running mean of classifications per participant (grey), and average across all users (black), plotted as difference between the current mean and the final mean after all classifications by the participant have been included.

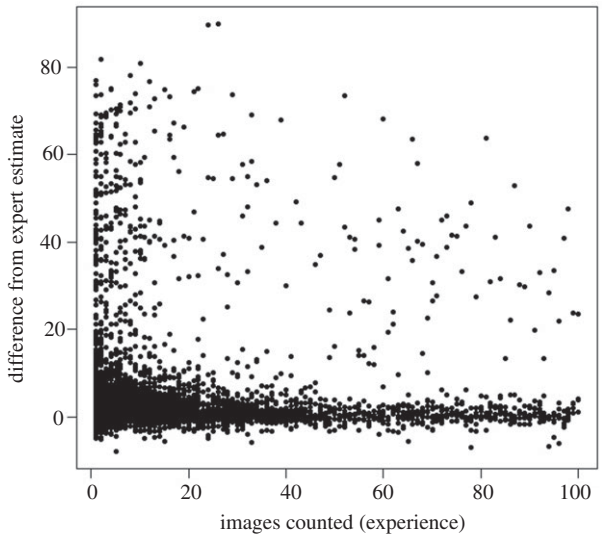

Figure 5. Individual classification accuracy in stomatal index estimate when compared with stomatal index estimates made by the lead project scientist, plotted in the order the classification was made by an individual citizen scientist.

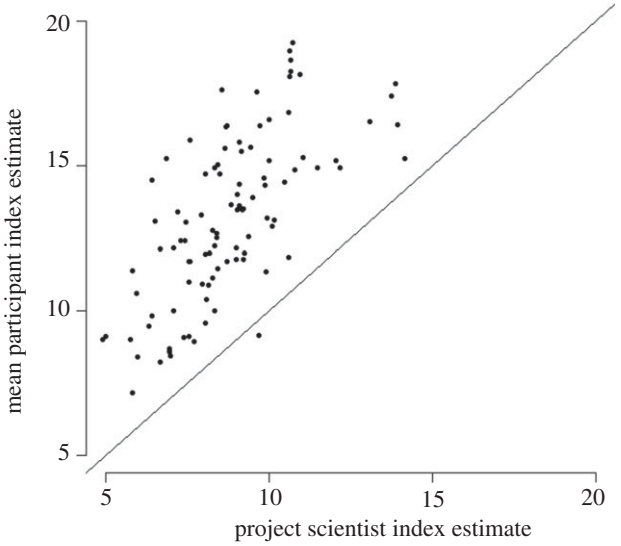

Figure 4. Comparison of mean citizen scientist stomatal index estimates versus lead project scientist's stomatal index estimates. Citizen scientists routinely undercount epidermal cells, leading to higher stomatal index values because the epidermal cell count is in the denominator of the equation (see equation (3.1)).

of cases. Citizen scientist estimated epidermal cell count only matched expert estimates in 0.8% of cases and was too few in 92% of cases. Averages of estimated stomatal index across all participants who classified an image were almost always higher than expert estimated indices (mean difference between participant estimates and expert estimates per image = 7.81; figure 4). Removing classifications made by users who were not logged in slightly improved average estimates (mean difference = 7.02), but a *t*-test did not show a significant difference ($p = 0.098$). Adding participants' 'not sure' counts to the epidermal cell counts slightly (but not significantly; *t*-test $p = 0.2423$) improved average estimates (mean difference = 7.26), but the participants' mean estimates were still consistently too high. Removing classifications where a participant had only marked one stoma or one epidermal cell (i.e. not completed the task) significantly improved average estimates (mean difference = 4.46, *t*-test $p < 0.0001$), and this filter was applied for the subsequent analyses.

Regression of the number of images a participant had counted (as a proxy for experience with the task) and index estimate difference (the difference between participant estimated stomatal index and project scientist estimated stomatal index, for each classification) showed a significant relationship, but with a very low positive slope of 0.02, suggesting that overall there was no improvement with experience (data shown in figure 5). Breakpoint analysis of participants' classification error in a time series demonstrated that a large majority of participants made either consistently accurate or consistently inaccurate estimates (figure 6), and, for the rare cases where individuals did show improvement followed by a plateau, there was no consistent breakpoint value across those individuals (electronic supplementary material).

The mean stomatal count error per user across all user classifications for the 'Easier Count' workflow is a significant predictor of their mean error in epidermal cell count (figure 7; $p < 0.0001$, $r^2 = 0.2796$).

5. Discussion

Fossil Atmospheres is a project designed to address research questions relating to atmospheric carbon dioxide and global climate change (in the modern and in the geological past), in which citizen scientists have been directly involved in the collection of data from museum specimens via an online user interface. This is a case study in presenting a relatively complex task to volunteers who do not receive training outside of the informational materials provided online.

There are three key lessons from this case study that could be applied to the development of other similar projects. (i) Acknowledge and address participants' uncertainty about their ability to complete the tasks satisfactorily. Feeling uncertain about one's skill set can be a barrier to prolonged engagement, and by addressing it, the participant feels motivated to continue. (ii) Scaffold the data collection process to

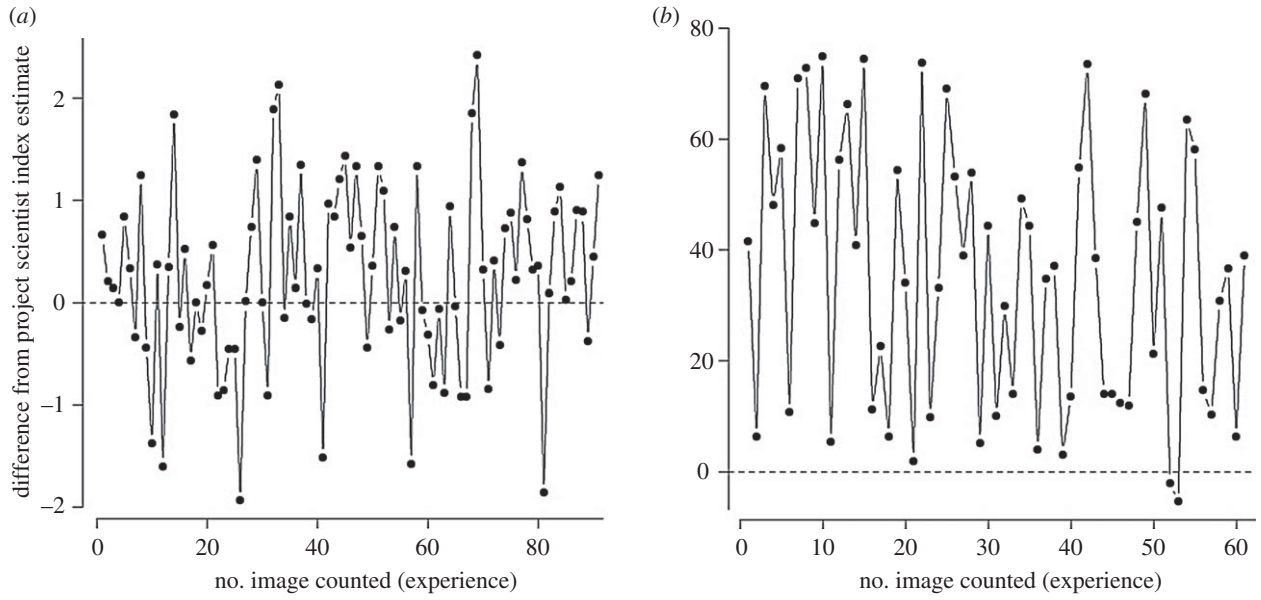

Figure 6. Accuracy of two example citizen scientist stomatal index estimates relative to expert estimates, plotted in the order the participant classified the images, as a proxy for experience in counting. (*a*) An example of a participant that makes consistently good estimates. (*b*) An example of a participant who remained inaccurate. Note the different *y*-axis scales.

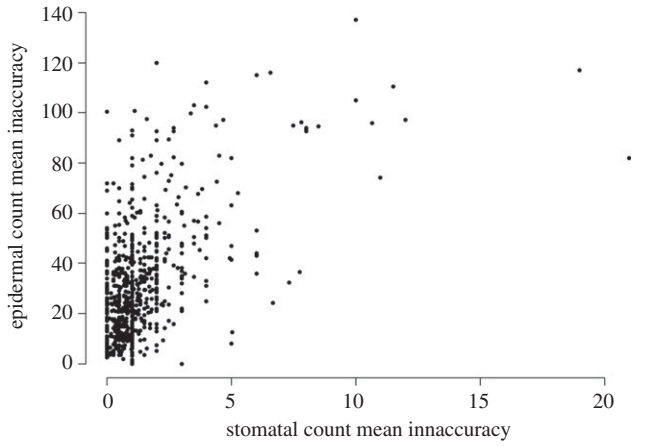

Figure 7. Stomatal count inaccuracy relative to expert counts by the project lead scientist, for each participant, plotted against epidermal cell count inaccuracy for each participant.

accommodate various levels of expertise so that participants work at the appropriate level while offering a higher achievement to work toward. Finally, (iii) provide data collection examples and annotated graphics to illustrate attributes of an expert classification. Participants use these materials to recognize and analyse discrepancies between their work and expert examples, and then apply what they have learned to their next task. Additionally, there are approaches that we took to refining the data that could be used for other online projects, particularly those hosted on Zooniverse, that present complex tasks for which data quality may be low. These are outlined in section (a) below.

Building a project on Zooniverse was a successful method for incorporating citizen science into the Fossil Atmospheres project. The interface provided all the tools necessary to complete the stomatal index estimation task, and many citizen scientists have participated, leading to many thousands of individual classifications made in a small fraction of the time it would take for the project scientists to complete the same number of classifications. This approach therefore has potential for other palaeontological projects where repeated measurements of specimens that can be easily photographed, but not easily handled, need to be made. However, data quality was low for a large proportion of individual citizen scientist classifications. We discuss the cause of this issue and possible solutions below.

(a) Accuracy and efficiency

Raw data from all classifications (by all participants) vary widely in accuracy and rarely approach expert stomatal index estimates (figure 4; electronic supplementary material). Therefore, some refining of the data is always necessary. Straightforward ways to somewhat improve average data quality were to remove classifications for which participants answered that they did not count all the cells in an image, and to use data only from participants who were logged in to the Zooniverse rather than submitting classifications as a guest.

Citizen scientists consistently do not count all of the epidermal cells within an image, which leads to an overestimate of the stomatal index relative to the expert value (equation (3.1); figure 4). For this reason, including 'not sure' markers as epidermal cells slightly improves participant accuracy. Even with these approaches, citizen scientist estimates are frequently still inaccurate to a degree that would introduce too large an error into an estimate of the relationship between stomatal index and pCO_2. One possible avenue of success is through measuring participant accuracy on stomatal count, which correlates with participant accuracy on epidermal cell count (figure 7), and would therefore provide an efficient first pass to identify individuals more likely to correctly identify epidermal cells. Alternatively, the 'easier' image set that we have used here as the validation test set could be used to identify citizen scientists who are likely to provide accurate data for other classifications.

The accuracy relative to expert counts of a large majority of participants is either consistently within 1–2% of the expert index estimate and therefore acceptable for use in further analyses (figure 6*a*), or consistently inaccurate and does not show improvement (figure 6*b*). A few participants improve with the number of classifications they have

completed, during approximately the first 10 classifications they complete, but this is rare (electronic supplementary material). Using these data to identify citizen scientists who are successfully making accurate estimates could allow us to use the data from these individuals exclusively, or to select which participants should continue to make classifications. While this approach has the potential to achieve the desired research outcome, it is an inefficient use of volunteer time if results from some volunteers are discarded, and is therefore not best practice for citizen science [9]. A possible solution to this issue would be to alter the set-up of the web interface to restrict who can continue to classify images.

(b) Machine learning approaches

The kind of image analysis we are enlisting citizen scientists for in this study is potentially suitable for machine learning. Indeed, an online interface called 'Stomata Counter' has recently been made available (through the efforts of Karl Fetter and the Fossil Atmospheres project scientists; www.stomata.science) that can count the number of stomata within images of leaf surfaces. This machine learning approach is supported by a convolutional neural network that is trained on about 5000+ micrographs from 700+ species of plants, including images from the Fossil Atmospheres experiment. As more images are piped through the system, along with a greater diversity of image types (e.g. scanning electron microscopy (SEM), transmitted light and epifluorescence) and from a greater diversity of species, the computer algorithm will improve over time as it gets 'trained'. Currently, citizen scientists are similarly accurate in their ability to correctly count stomata to the machine learning approach, but may soon be less accurate. More importantly, at least for the goals of the Fossil Atmospheres project, the computer cannot yet successfully count epidermal cells, so a human must make the counts. Our citizen science approach has the potential to be an efficient way to collect these data, and the current phase of our project is to assess the quality of the data and to determine how best to use the Zooniverse as a resource to collect more data than would be possible by the research team alone. Other citizen science projects (e.g. Snapshot Serengeti) have successfully used citizen science classifications as training data for machine learning algorithms, something that would not have been possible without the large number of images classified by Zooniverse users for that project. As our classification database expands, it may be possible to use it as training data for a machine learning approach that can estimate stomatal index, rather than solely stomatal number.

(c) Future directions

There are three avenues of further investigation we are pursuing in the long term, with an aim of improving each of the major goals of the project. They relate to data processing, citizen science performance and science education.

First, it is possible that better use could be made of existing data by aggregating all counts for each image (figure 8). As demonstrated above, most citizen scientists overestimate stomatal index because they do not find all of the epidermal cells (figure 4). However, when the coordinates for the markers for all participant classifications of an image are aggregated, they successfully identify all of the epidermal cells because different people fail to count different cells. Therefore, a clustering approach that aggregates all classifications would be advantageous. Current clustering algorithms have most frequently been designed for lower numbers of hard-to-discriminate clusters (e.g. hierarchical, *K*-means, distribution-based and density-based), whereas our data require an algorithm designed for a high number of well-defined clusters. An effective algorithm is not yet available. We are therefore working on developing a new method suitable for our purpose.

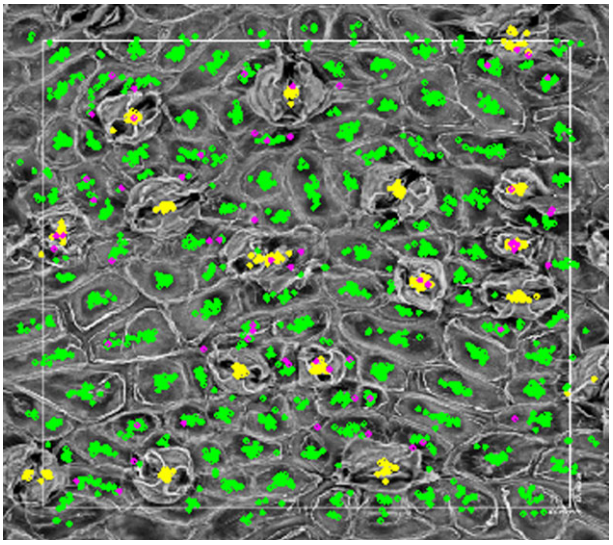

Figure 8. A typical example of the aggregate of all citizen scientist classifications for an image ($n = 50$ unique classifications). Different colours indicate the different classifications a participant has made for each marker: green for epidermal cells, yellow for stomata and pink for 'not sure'. White box is a standardized 300 × 300 µm. No single user identifies all the cells, but when all user classifications are taken together all the epidermal cells are marked.

Citizen science performance can also be improved. Interactions on the project talk boards showed that the information available on how to perform the task was not always read or understood. We have therefore held in-person training sessions and in a future study will compare the performance of volunteers who have been trained face to face with Zooniverse users who have only seen the online material, over the same period of time. If face-to-face training leads to greater success in counting, it suggests that we could increase efficiency by training a smaller set of citizen scientists to make very accurate estimates.

Finally, we are working on ways to increase the educational value of the project. We have provided information on the scientific background of the research, as well as background materials on plant science, palaeobotany and global climate change. One of the aims of Fossil Atmospheres is to create an opportunity to educate the public about climate change, and we are therefore designing evaluation materials to test for changes associated with participation in the project, in understanding of the subject matter and interest in science [40].

6. Conclusion

We have shown that online citizen science is a viable option for collection of data from natural history collections. Presenting images for classification on the Zooniverse leads to high

citizen scientist engagement and processing of more images than would be possible without such a resource. This approach to scientific data collection provides a unique opportunity to educate the public about biology and climate change, allows new audiences access to natural history specimens and has the potential to save researchers time that is better suited to implementing more complicated aspects of the scientific project, or conducting science that can only be done on site. This online data collection approach can therefore be well suited for palaeontological projects that require some kind of data to be collected from a large number of specimens, and where it is straightforward to photograph the specimens. However, this is with the caveat that even with careful thought and design of the instruction process, complex tasks such as the one required by Fossil Atmospheres can result in many inaccurate citizen scientist classifications and therefore low volunteer time efficiency. The 'wisdom of the masses' may prevail if methods of analysis can be developed that appropriately combine the efforts of multiple unbiased but imperfect observers.

Data accessibility. The datasets supporting this article have been uploaded as part of the electronic supplementary material.

Authors' contributions. L.C.S. and R.S.B. contributed equally to the development of the citizen science website on the Zooniverse. L.C.S. built the project worksite, conducted the data analysis, and wrote the initial draft of the manuscript, with revisions by R.S.B. All authors contributed equally to the intellectual design of the experiment conducted on the Zooniverse, contributed to the revision of the manuscript and signed off the submitted copy.

Competing interests. The authors declare no competing interests.

Funding. L.C.S. was supported through a 'Deep Time' fellowship from the Peter Buck Foundation at the Smithsonian National Museum of Natural History. R.S.B. was supported by grants from the Smithsonian's Scholarly Studies Program, and a grant from the Paleo Perspectives on Climate Change section of the National Science Foundation (no. 1804974).

Acknowledgements. This publication uses data generated via the Zooniverse.org platform, development of which is funded by generous support, including a Global Impact Award from Google, and a grant from the Alfred P. Sloan Foundation. We thank three anonymous reviewers whose comments improved the quality of the manuscript. We also thank Lucy Chang for assistance with data parsing. We also thank the many local volunteers on the Fossil Atmospheres project, in particular the efforts of Sal Bosco who prepared all of the specimens and Pamela Hamilton who took the scanning electron microscope images in the Imaging Lab run by Scott Whittaker at the Smithsonian National Museum of Natural History.

References

1. Bonney R, Cooper CB, Dickinson J, Kelling S, Phillips T, Rosenberg KV, Shirk J. 2009 Citizen science: a developing tool for expanding science knowledge and scientific literacy. *BioScience* **59**, 977–984. (doi:10.1525/bio.2009.59.11.9)
2. Cox J, Oh EY, Simmons B, Lintott C, Masters K, Greenhill A, Holmes K. 2015 Defining and measuring success in online citizen science: a case study of Zooniverse projects. *Comput. Sci. Eng.* **17**, 28–41. (doi:10.1109/MCSE.2015.65)
3. Simpson R, Page KR, De Roure D. 2014 *Zooniverse: observing the world's largest citizen science platform*. In *Proc. 23rd Int. Conf. on World Wide Web. Seoul, Korea, 7–11 April, 2014* (eds C-W Chung, A Broder, K Shim, T Suel), pp. 1049–1054. New York, NY: Association for Computing Machinery.
4. Sepkoski D. 2012 *Rereading the fossil record: the growth of paleobiology as an evolutionary discipline*. Chicago, IL: University of Chicago Press.
5. Smith JM. 1988 Palaeontology at the high table. *Did Darwin get it right?* Boston, MA: Springer.
6. Kosmala M, Wiggins A, Swanson A, Simmons B. 2016 Assessing data quality in citizen science. *Front. Ecol. Environ.* **14**, 551–560. (doi:10.1002/fee.1436)
7. Johnson MF, Hannah C, Acton L, Popovici R, Karanth KK, Weinthal E. 2014 Network environmentalism: citizen scientists as agents for environmental advocacy. *Global Environ. Change* **29**, 235–245. (doi:10.1016/j.gloenvcha.2014.10.006)
8. Kragh G. 2016 The motivations of volunteers in citizen science. *Environ. Scientist* **25**, 32–35.
9. Dickenson JL, Bonney R. 2012 *Citizen science: public participation in environmental research*. Ithaca, NY: Cornell University Press.
10. Moran S, McLaughlin C, MacFadden B, Jacobbe E, Poole M. 2015 Fossil explorers. *Sci. Children* **53**, 62–67. (doi:10.2505/4/sc15_053_04_62)
11. Salisbury EJ. 1927 On the causes and ecological significance of stomatal frequency with special reference to woodland flora. *Phil. Trans. R. Soc. B* **216**, 1–65. (doi:10.1098/rstb.1928.0001)
12. Sickler J, Cherry TM, Allee L, Smyth RR, Losey J. 2014 Scientific value and educational goals: balancing priorities and increasing adult engagement in a citizen science project. *Appl. Environ. Educ. Commun.* **13**, 109–119. (doi:10.1080/1533015X.2014.947051)
13. Zachos JC, Dickens GR, Zeebe RE. 2008 An early Cenozoic perspective on greenhouse warming and carbon-cycle dynamics. *Nature* **451**, 279–283. (doi:10.1038/nature06588)
14. Bowen GJ, Beerling DJ, Koch PL, Zachos JC, Quattlebaum T. 2004 A humid climate state during the Palaeocene/Eocene thermal maximum. *Nature* **432**, 495–499. (doi:10.1038/nature03115)
15. Koch PL, Zachos JC, Gingerich PD. 1992 Correlation between isotope records in marine and continental carbon reservoirs near the Paleocene/Eocene boundary. *Nature* **358**, 319–322. (doi:10.1038/358319a0)
16. McInerney FA, Wing SL. 2011 The Paleocene-Eocene thermal maximum: a perturbation of carbon cycle, climate, and biosphere with implications for the future. *Annu. Rev. Earth Planet. Sci.* **39**, 489–516. (doi:10.1146/annurev-earth-040610-133431)
17. Wing SL, Harrington GJ, Smith FA, Bloch JI, Boyer DM, Freeman KH. 2005 Transient floral change and rapid global warming at the Paleocene-Eocene boundary. *Science* **310**, 993–996. (doi:10.1126/science.1116913)
18. Zachos JC *et al.* 2005 Rapid acidification of the ocean during the Paleocene-Eocene thermal maximum. *Science* **308**, 1611–1615. (doi:10.1126/science.1109004)
19. Barclay RS, Wing SL. 2014 Increasing atmospheric CO_2 prior to the Paleocene-Eocene thermal maximum inferred from stomata of *Ginkgo adiantoides*, Bighorn Basin, Wyoming, USA. *Rend. Online Soc. Geol. Ital.* **31**, 23–24. (doi:10.3301/ROL.2014.26)
20. Huber M, Caballero R. 2011 The early Eocene equable climate problem revisited. *Climate Past* **7**, 603. (doi:10.5194/cp-7-603-2011)
21. Beerling DJ, Royer DL. 2011 Convergent Cenozoic CO_2 history. *Nat. Geosci.* **4**, 418. (doi:10.1038/ngeo1186)
22. Barclay RS, Wing SL. 2016 Improving the *Ginkgo* CO_2 barometer: implications for the early Cenozoic atmosphere. *Earth Planet. Sci. Lett.* **439**, 158–171. (doi:10.1016/j.epsl.2016.01.012)
23. Bolton CT *et al.* 2016 Decrease in coccolithophore calcification and CO_2 since the middle Miocene. *Nat. Commun.* **7**, 10284. (doi:10.1038/ncomms10284)
24. Breecker DO, Sharp ZD, McFadden LD. 2009 Seasonal bias in the formation and stable isotopic composition of pedogenic carbonate in modem soils from central New Mexico, USA. *Geol. Soc. Am. Bull.* **121**, 630–640. (doi:10.1130/B26413.1)
25. Breecker DO, Sharp ZD, McFadden LD. 2010 Atmospheric CO_2 concentrations during ancient greenhouse climates were similar to those predicted for AD 2100. *Proc. Natl Acad. Sci. USA* **107**, 576–580. (doi:10.1073/pnas.0902323106)

26. Woodward FI. 1987 Stomatal numbers are sensitive to increases in CO_2 from preindustrial levels. *Nature* **327**, 617–618. (doi:10.1038/327617a0)
27. Woodward FI, Bazzaz FA. 1988 The response of stomatal density to CO_2 partial pressure. *J. Exp. Bot.* **39**, 1771–1781. (doi:10.1093/jxb/39.12.1771)
28. Royer DL, Wing SL, Beerling DJ, Jolley DW, Koch PL, Hickey LJ, Berner RA. 2001 Paleobotanical evidence for near present-day levels of atmospheric CO_2 during part of the Tertiary. *Science* **292**, 2310–2313. (doi:10.1126/science.292.5525.2310)
29. Royer DL. 2003 Estimating latest Cretaceous and Tertiary atmospheric CO_2 from stomatal indices. In *Causes and consequences of globally warm climates in the Early Paleogene* (eds SL Wing, PD Gingerich, B Schmitz, E Thomas). *Geol. Soc. Am. Spec. Pap.* **369**, 79–93.
30. Trouille L, Lintott C, Miller GSH. 2017 DIY Zooniverse Citizen Science project: engaging the public with your museum's collections and data. In *Museums and the Web, 19–22 April 2017, Cleveland, Ohio*. See https://mw17.mwconf.org/paper/diy-your-own-zooniverse-citizen-science-project-engaging-the-public-with-your-museums-collections-and-data/.
31. Brown JS, Collins A, Duguid P. 1989 Situated cognition and the culture of learning. *Educ. Res.* **18**, 32–42. (doi:10.3102/0013189X018001032)
32. Lave J, Wenger E. 1991 *Situated learning: legitimate peripheral participation*. Cambridge, UK: Cambridge University Press.
33. Sprinks J, Wardlaw J, Houghton R, Bamford S, Morley J. 2017 Task Workflow Design and its impact on performance and volunteers' subjective preference in Virtual Citizen Science. *Int. J. Hum. Comput. Stud.* **104**, 50–63. (doi:10.1016/j.ijhcs.2017.03.003)
34. Vygotsky LS. 1978 *Mind in society*. Cambridge, MA: Harvard University Press.
35. Greenfield PM. 1984 A theory of the teacher in the learning activities of everyday life. In *Everyday cognition* (eds B Rogoff, J Lave). Cambridge, MA: Harvard University Press.
36. Wood D, Bruner J, Ross G. 1976 The role of tutoring in problem solving. *J. Child Psychol. Psychiatry* **17**, 89–100. (doi:10.1111/j.1469-7610.1976.tb00381.x)
37. Bandura A. 1997 *Self-efficacy: the exercise of control*. New York, NY: W.H. Freeman.
38. Dweck CS, Leggett EL. 1988 A social-cognitive approach to motivation and personality. *Psychol. Rev.* **95**, 256. (doi:10.1037/0033-295X.95.2.256)
39. Muggeo VM. 2008 Modeling temperature effects on mortality: multiple segmented relationships with common break points. *Biostatistics* **9**, 613–620. (doi:10.1093/biostatistics/kxm057)
40. Phillips T, Ferguson M, Minarchek M, Porticella N, Bonney R. 2014 *User's guide for evaluating learning outcomes in citizen science*. Ithaca, NY: Cornell Lab of Ornithology.