We would like to thank the editor and the four referees for their comments. We have been working toward a new version of the manuscript taking their respective comments into account. We include the comments from the referees in black, our responses in blue, and the modifications to the manuscript in red in this response file.
As we are talking about the review process, and to avoid confusion, we will refer to the participants to the IPCC review group project as "reviewer", and to the 4 referees of this manuscript as "referee", following the copernicus terminology.

Referee 1: Gary McDowell

Overall this is a very interesting study and an excellent piece of work. In particular, I noted that Figure 1 is excellent.
I have some specific comments/questions below, but my major comment is about "comments"
- I'm unfamiliar with the process that reviewers were participating in, and would greatly appreciate clarification (even a schematic, if that were thought helpful): Could you elaborate on the size and structure of "comments"? Why do people submit multiple comments? How long is a comment? What does it relate to e.g. is there a new comment for each issue raised? Is this roughly equivalent to commenting in a word processing document? It's not clear what the reviewers are actually submitting, and in what format.

The peer-review process in which we participated for the First Order Draft of the SROCC IPCC report is indeed quite specific. As for traditional peer-review, reviewers provide several comments that can either be specific to one or several lines, or relevant to the entire manuscript. Nonetheless, considering that there are several authors working on a chapter, each specific comment on a portion of the text is included as a single comment in a formatted-spreadsheet. During the review of the Casado et al. manuscript, we were told by Copernicus that the supplementary materials were not uploaded properly and were thus not accessible to the referee. This section of the manuscript contained explanations as to the process of comment submission during the group review (Page 9).

In addition, we have added a section to the manuscript on the IPCC review process which clearly sets out the review process and describes comments.

We will make sure with Copernicus staff that the supplementary materials are included.

It sounds like all reviews/comment are compiled into one long review to be submitted to the IPCC on behalf of APECS - is this the case? (Line 204) Could you also explain what you mean by a "group review" - this terminology may not be standard and it would be helpful to clarify the concept, especially if this is a phenomenon unique to the IPCC.

It's true that different concepts could be understood by the term "group review". We included an explanation (see line 135):

> "The review of the FOD of the SROCC was done as a group review, meaning that each participant of the group was only in charge of reviewing a small section of the report, and that all the comments were then combined and submitted together under the flag of APECS."

In general, we have provided further clarity on the review process for IPCC and the approach of APECS.

Other questions:
Line 68 - "Recognising that many of the reviewers had neither published a paper nor participated in a peer-review before" - is there evidence for this claim? Were they surveyed or are you merely surmising? Especially as this seems to be contradicted in line 86 - "Among the applicants, 72% had already reviewed a scientific document (such as a paper, a proposal or a scientific report). "

We asked participants when we were selecting them if they had participated to a peer-review or published a paper beforehand. It is true that "many" is not appropriate here. We modified to "some".

In Table 1, 5 of the 6 countries with largest representation are primarily English speaking. Is this distribution as expected for the population of researchers? Is it solely a reflection of the over-representation of English-speaking countries in the network advertising the call for reviewers as indicated in the text following in the table? Is it a reflection of a document being written in English? These are very minor questions, mostly my interest was piqued.

The distribution of the countries represented in our review is indeed closely related to APECS membership (numbers can be found in this presentation of APECS https://www.dropbox.com/s/n0pwe55um1l23uh/This_is_APECS_July%202019.pptx?dl=0). We agree with the referee that this is interesting. As this was not something we were necessarily satisfied with, in later developments of the project (FOD-SOD AR6 WGI, and FOD-AR6-WGII), we collaborated with other Early Career associations (i.e. Mountain Research Initiative (MRI), PAst Global ChangES Early-Career Network (PAGES-ECN), Permafrost Young Research Network (PYRN), and Young Earth System Scientists (YESS) community), and increased our worldwide distribution of participants.

This is an interesting question – and we have included a recommendation in the conclusion to encourage organisations in under-represented regions to participate as group review to enhance global engagement.

Line 139 - Could you clarify - was a student paired with a postdoc/early academic in each case? What was the format for exchanges to occur - were they connected in space and time or was all pairing remote/over email/internet exchanges?

Our review only took place online, but we wanted to provide a framework for more junior participants to have someone they could exchange with if they had questions. We added a specification on the nature of the exchanges:

> "The initial purpose was to promote interactions amongst participants during the review process, through online exchanges,"

Line 210 - Was there any data gathered on the length of training of reviewers? It is possible, for example, that a PhD student in the US could have as much time in training as a senior postdoc in the UK.

Unfortunately, we did not collect any such data. Estimating the "length" of the training of reviewers would be very difficult considering significant differences between countries educational systems, and personal life stories. Nonetheless, this would be a very interesting inquiry to make.

Referee 2 : Hans Visser

The study by Casado et al. represents an important piece of work. It is well written and well documented.

However, I have one serious comment: the paper is in many ways equivalent to the paper published by Lianne van der Veer, myself, Arthur Petersen and Peter Janssen in Climatic Change, 2014. Indeed, this paper is named in line 55, but more as a side remark.

I name a number of equivalent issues:

- We selected 90 PhD students, coming from 31 countries with global spread.
- We reviewed the IPCC AR5 WG II report containing 30 chapters and the corresponding SPM.

- We had a full training of the students with videos, provided network opportunities, interesting sessions with IPCC top speakers, such as Leo Meyer who was the main author of the AR5 Synthesis report, etc. We had a bonus of 250 euro for the best review and a bonus of 50 euro for students finding 'bloopers'.
- We designed a hand out with a systematic procedure for finding errors/bloopers/inconsistencies etc.
- The review resulted in a total of 3155 comments from which, after selection by an expert panel, a number of 1407 comments were included in the Government review.
- We made a comparison between the quality of PhD comments and experienced scientists.

Because of these large similarities it is my opinion that the style of paper should be changed considerately. As such, it is no problem that research team X repeats a study of team Y, published years before. However, one should start with writing this, including in the Abstract, and should highlight were the new study deviates from the study of team Y.

While the work the referee mentions (van der Veer et al, 2014) is indeed a precursor in many aspects, as the first published article describing a group review of an IPCC report, we are surprised by the description of our project as a mere repetition. Among the differences, we would like to highlight 5 that are significant:

1. First, in the paper published by the team of the review, the manuscript is about a government review of the IPCC report. Indeed, the referee and his team are all Dutch government employees. The difference between expert reviews and government reviews are not necessarily straightforward and can be seen in Figure R1.

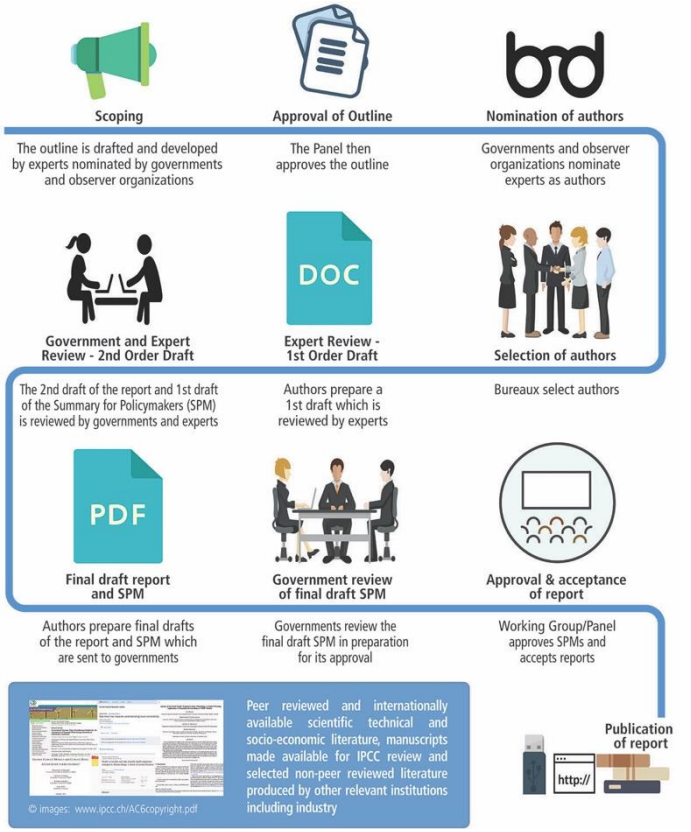


Figure R1: *different phases of the IPCC report peer reviews.*

One of the main aspects here, is that expert reviewers need to apply, and their expertise is checked by the Technical Support Unit (TSU) of the IPCC before they proceed to their

review. In the case of government review, the IPCC has to send the report to the different countries that are part of the United Nation Environmental Program.

2. Second, the group review in van der Veer et al. (2014) was organised by government employees, and using the network of science universities in the Netherlands. This results in significant differences in the structure of both projects. While the van der Veer et al. (2014) project had a vertical structure with supervision and organisation by senior government and university employees, the Casado et al. (this ms) project was solely handled by Early Career Scientists with a horizontal structure. What we call "chairs" were acting as focal points, making sure everyone was carrying out their assigned tasks on time and that information was given to everyone. However these chairs did not have more training or hierarchal importance than any other participants and everyone participated equally in the group review and proposed comments for submission. This has been highlighted in the manuscript (lines 113 – 115):

   > "The horizontal structure of this group review (all participants are ECS self-organising themselves to realise this review) strongly differs from the previous group review attempts of IPCC review project for which a government was organising the review and providing incentives to participant to participate in such review (van der Veer et al., 2014)."

3. Third, as a large number of the van der Veer et al. (2014) participants were part of a university network in the Netherlands, they had incentives to take part in the project as university credits. In general, actual financial incentives were proposed by the Dutch government to find errors. While it's amazing that the Dutch government organised this, this is unlike the actual review process in science, where no reward of any kind is given to reviewers. As such, our project is closer to an actual peer-review where the reviewer is doing this without expecting any financial compensation.
4. Fourth, the van der Veer et al. (2014) project only included PhD students, while our project was open to a large range of early career levels (from MSc. to early Professors).
5. Fifth, the van der Veer et al. (2014) project had a "healthy competition" among teams that were attributed the different chapters, which is significantly different from our structure.

In general, the group-review organised and described in van der Veer et al. (2014) had a lot of differences to one described in Casado et al. (this ms). We are for instance surprised that in their project, van der Veer et al. (2014) also recruited an expert panel of 10 scientists who discarded more than 50% of the comments made by the PhD students. The justification is that in most of the cases, the reviewers "misunderstood, had not read the underlying literature…". Considering that these reports are made for non-experts, we strongly believe that a comment which denotes that the reviewer missed the point of the report, or is not aware of a piece of an extensive body of literature, is valuable. Such comments may reveal that the report cannot stand-alone or needs to be more accessible, and it would be of use for the authors of the IPCC reports to consider such comments in their revisions.

We never intended to give the impression that we cited the manuscript of the authors as a side remark, having them very early on in the introduction. We strengthened the inclusion of their paper in our manuscript (Line 56):

> "While no individual possesses the required expertise to review an entire IPCC report, as a group, ECS have also proven to be efficient and motivated reviewers, providing added value to this type of manuscript as shown by van der Veer et al. (2014), who pioneered the organisation of groups of ECS to participate in the IPCC report review process."

We believe that a complete comparison of both group-review projects is outside the scope of this manuscript, considering the extensive differences in both methodology and results. Indeed, our results do not include an independent, external evaluation of the comments made, and thus we

cannot produce a quality control of the comments as done in van der Veer et al. (2014), and thus cannot contradict nor support the outcome of their study "It can be concluded that more comments does not guarantee better quality.". On the other hand, as they only included PhD students in their study, we cannot compare their results to ours which focus on the number of comments of different types and the time spent by reviewers at different career levels. We included a discussion about this point in the new version of the manuscript:

> "van der Veer et al. (2014) proposed that a large number of comments does not necessarily guarantee better quality, but as only a self-evaluation of the quality of the comments was done, we are not able to validate this point with our results."

Finally, while we agree that comparing both initiatives would be a very interesting idea, it would be difficult considering the extensive aforementioned differences. While preparing this manuscript, we carried out 3 more group reviews, and collected similar data from these different projects. A future study idea on our side could be to carry out an in-depth comparison of the different group reviews, for which collaboration with the referee would be interesting.

For example, we found it important the follow the conclusions presented in the SPM down to the lowest levels of the main text of the report. See Figure 1 below. As far as I can see, this is not followed in the present paper. Please give arguments why not.

Unfortunately, the Summary for Policy Maker (SPM) was not available for the review of the first order draft (FOD), which we participated to (see Figure R1). In general, expert reviewers are asked to mainly focus on the main text, while government reviews usually have more impact on the SPM.

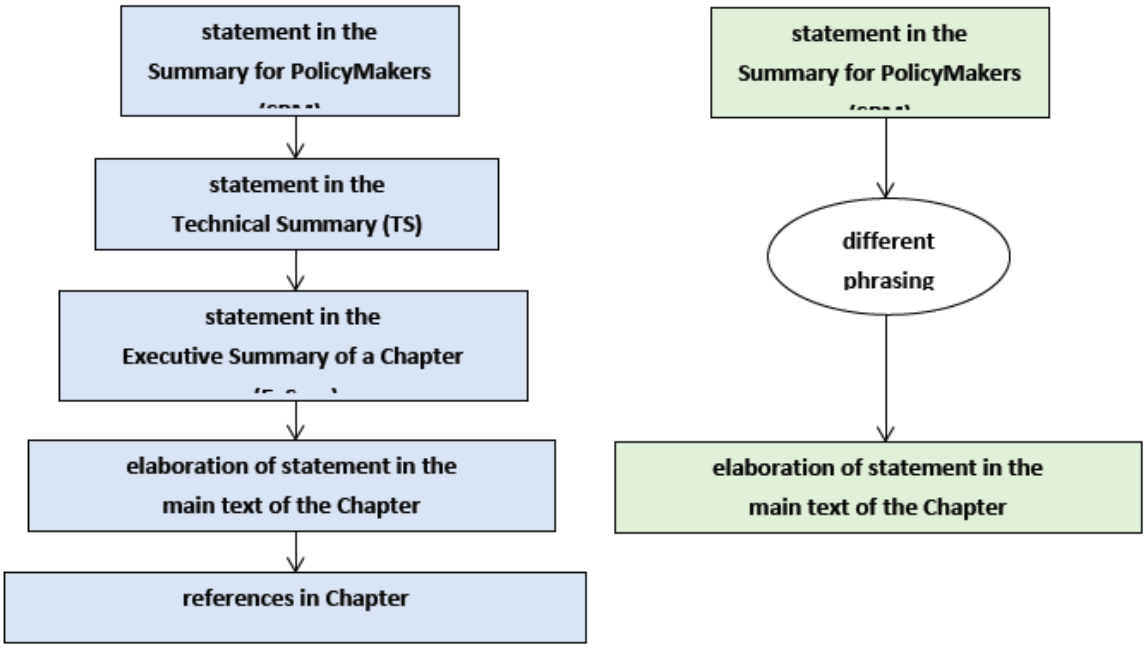


Figure 1. *Five levels of information can be distinguished in the IPCC WG II report. The left pyramid shows the ideal situation: a statement on the highest level (SPM) is founded in all the lower levels. The right situation is not ideal: a statement in the SPM is founded in the main text of the Chapter only, without any references to the peer-reviewed literature; moreover, the statement in the SPM has been rephrased (rephrasing is no error but care should be taken).*

I have two other comments on the paper. As a reader I am interested in the guidelines that were given to the students. An example of the guidelines we had in our 2014 study is reprinted in Table 1. Is it possible to give a similar table within the paper, with some explanation?

The guidelines given to the students are included in the guide, that apparently was not available during the peer-review, but that is planned as supplementary material. While we initially recommended to use codes similar to the ones provided by the referee in table 1 of the van der Veer et al. (2014) paper, (as can be seen in our guide page 8), it appears that the IPCC stopped using these codes. We were therefore requested to only categorise the comments as either "substantive" or "editorial". This change was mentioned in emails to the participants of the group review of the FOD, as well as during the training sessions. We have added a section on the IPCC review process, including for the SROCC.

Table 1. *Checkpoints that serve to signal potential weak points in the report under review. The typology is taken from PBL (2010). The list contains two types of errors (E1 and E2), and six types of comments (C1 through C6).*

Type	Description	Explanation
E1	Inaccurate statement	
E1a	Errors that can be corrected by an	For example, typographical errors, incorrect phrasing of part of a
E1b	Errors that require a redoing of the assessment of the issue at hand	Such as establishing a new range of numbers by revised calculations from the reference sources available during the assessment period, and/or rephrasing of the expert judgment including its uncertainty labelling.
E2	Inaccurate referencing	A reference to a wrong source, or source not correctly cited. In all cases,
C1	Insufficiently substantiated attribution	The *climate change component* of impacts/risks should be carefully characterized. If applicable, the role of *other factors* than climate change (e.g. population growth, industrialization, migration, and changes in land use and land cover) should be discussed appropriately? With regard to extreme events, it is particularly important to be careful with attributing events to anthropogenic climate change
C2	Insufficiently founded generalization	A proper argumentation is lacking or the evidence in the references does not justify a generalization or extrapolation of
C3	Insufficiently transparent exper t judgment	The reasoning behind an expert judgment, including the reasoning behind its level of likelihood and/or confidence, is not accessible to a non-expert reviewer. However, the reasoning should be transparent in all cases. Note: a lack of transparency does *not* imply the judgment to be wrong, since the authors may have had their reasons, and may have considered additional information or knowledge that was not

C4	Inconsistency o f messages	A message's content and/or confidence level change when going from the main text to a summary (SPM, TS or ExSum). The IPCC procedures require that all summary texts are consistent with the main text or lower level summaries.
C5	Untraceable reference	A reference in a statement cannot be found at all.
C6	Unnecessary reliance on grey referencing	A reference to a grey publication, although strong peer-reviewed journal references were available at the time of writing the concept report. Notice that grey literature is an indispensable part of many assessments since not all reviewed scientific journals. Please check if references to grey literature are weak: speeches, one-sided NGO reports, newspapers, opinion magazines

A second comment is on the comments found by the students in the SROCC report. No examples are given as far as I can see. The authors characterize comments/errors as 'substantive', but what is that? And were the level of comments comparable to those of experts? Please give some examples of what was found.

When the manuscript Casado et al. was first submitted, the report was still under review, as were the comments of other reviewers of the SROCC (not involved in our group review). Furthermore, information regarding the career stage of individual expert reviewers of IPCC reports is not available, only their affiliation, thus rendering such a comparison difficult. Now, the list of comments from all expert reviewers are available in the download page related to the report (https://www.ipcc.ch/srocc/download/) and such post analysis can be made following the same criteria we used to classify substantive or editorials comments among our participants comments submission.

As for the difference between substantive and editorial, we only used the same criteria as those provided by IPCC itself to categorise comments.

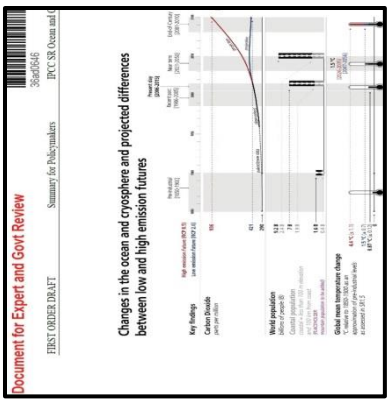
In Figure 2 an example is given from the SPM FOD of SROCC. There is a blooper in this graph, but which? I am curious if one of the students did find it?

As mentioned previously, the SPM was not included in our group-review, and was only made available at the next phase (the FOD of the SPM was provided during the Second order draft (SOD) of the full report).

In conclusion: to my opinion the set-up of the paper should change since it is largely repeating our article from 2014. At the other hand this is important work, well written, and certainly needs to be published! But not in the present form.

We hope that the changes made to our manuscript that highlight more the importance of the van der Veer et al. (2014) paper will convince the referee that we took his comments seriously. We would nonetheless argue that we did not repeat their study, both in the initiative itself, as well as in the results produced, as described above.

Figure2 *Graph from the FOD SPM of SROCC*



Referee 3: Anonymous Referee #3

General comment. This paper describes a very interesting experience in which Early Career Scientists (ECS), as a group, reviewed one of the Special Reports of the IPCC AR6. Moreover, it also briefly points to some issues for ECS' reviewers, and suggests potential solutions. The paper clearly describes all the procedure and presents some statistics about the participation and the outcome.

Specific comment. Although the Methods, and the Results and discussion sections are strongly focusing on the review of an IPCC report, the authors tried to apply what they learned from this experience to the review of a scientific paper. On my point of view this is a completely different exercise. I agree that the experience gained in the group review of an IPCC can be very useful when reviewing a scientific paper. However, the experiment presented in this paper can in no way be a basis to 'offer recommendations to editors of journal'. I don't mean that the authors opinion on that point is wrong or useless. I just say that there is no relation between the topic of the paper and this specific piece of conclusion. On the other hand, the paper is published in a journal with open review, as several other EGU journals. I was wondering how much ECS are taking this opportunity to submit (unsolicited) reviews.

We agree with referee #3 that our recommendations to journal editors included in the conclusion are 1. based on opinion, and 2. represent strong statements. While our manuscript focuses on the review of a scientific report, we've now included literature ("Ilgen et al, 2016") that provide similar results for scientific papers, and thus suggest that we could extend our results to peer-review in general. We would be happy to include more nuances in this opinion in the conclusion if the editor and the referees agree that we keep this section in the manuscript.
While IPCC reports are an assessment, they are grounded in the scientific literature and produced and reviewed by the scientific community, much like a review paper. As such, this group review experience is still very similar to reviewing a scientific paper.

As for the open-review of the EGU journals, while we think this is a very interesting initiative, it is not the case for all of the journals. We agree that it would be interesting to evaluate how much ECS are taking the opportunity to submit unsolicited reviews, but this is beyond the scope of the present manuscript.

I also have a few specific comments to the authors:
P2 l53: 'attracted comments . . . countries.' A reference would be welcome here
Added
P2 l64-67: 'APECS . . . Engineering.' This could more interestingly be moved to the recruitment section.

We believe that most of these figures are important in the introduction, and would prefer to keep them here together, rather than separate them between the introduction and the recruitment section.

P2 l77-79: This should be deleted as the same information is repeated later (p4 l155)

Modified accordingly.

P2 l 79-86: It would be interesting to know how many chapters each council member was chairing.

The number is quite variable, depending on time availability and interest of each members. Some of us only covered 1 chapter, some of us up to 4.

P3 l94-99: Why was it decided to select only a subset of the applicants for the review? Which were the criteria for this selection? More specifically, how was the motivation measured? How was the experience assessed? Were the applicants without experi- ence rejected (although giving experience was an objective as well)? What was the criteria on the country? Did you discriminate participants according to their country of residence?

Only a subset of the applicants was chosen as we did not have enough manpower to handle a larger group. In later reviews (SOD, AR6), we included more participants by having more chairs from different Early Career associations. We include more details on the application process that we used to evaluate motivation and experience (lines 88 and 89):

> "The call for participants, published on the APECS website, received 153 applications requesting information about applicants' credentials and motivations for participating in the group review"

and (lines 97 and 98):

> "We used the applications to evaluate (1) motivation, (2) experience and relevance of the application,…"

P5: section 2.4 could probably better fit in the Results section than in the Method one. Figure 1a and table 1 are providing exactly the same information. Delete one of them. Figure 1b: what is the y-axis?

Section 2.4. has been included as section 3.1. now.

Figure 1 provides an overview of all the important information from the project (including the country of origin of the different participants), while Table 1. aims at providing readers with pertinent data stemming from our manuscript. We believe both can be useful for readers. We would consider removing Table 1 if the editor agrees.

y-axis has been added to Fig 1.b.

P7 l208-209: 'rather than an entire chapter'. Please explain. I don't think that expert reviewers had to review one full chapter. Rather they were probably reviewing the sections connected to their expertise, possibly in several chapters.

Expert reviewers can indeed choose which sections of IPCC reports they would like to review. While some expert reviewers focus on sections of one or multiple chapters, those that review full chapters often provide a greater amount of commentary at the start of a chapter compared to the end, as evidenced from the greater number of comments received by the IPCC from the beginnings of chapters compared to the ends, which remains an issue for IPCC authors. Additionally, as the reports are not destined for specialists of each given section, but aim at providing an overview of the field, there is interest from IPCC authors to have reviewers going through sections not necessarily linked to their exact expertise.
We modified the text:

"The 388 other experts produced on average 26 comments per person, slightly less than the ECS participants (31.8, the difference being above 1 standard deviation) who had a specific set of 10 to 20 pages each to review, rather than an entire chapter."

P7 l211: How do you measure the 'quality and relevance' of the ECS comments?

We do not measure quality and relevance of the comments. We actually specifically do not measure this, as opposed to the van der Veer et al. (2014) study. We self-evaluated whether comments were substantive or editorial, using the guidelines provided by the IPCC. In general, throughout the manuscript, we made more consistent that we could not assess quality of the produced reviews.

P7 l254: How do you measure that the attribution process does not influence the quality of the comments?

We do not refer to quality of the comments throughout the manuscript. The comments were sorted between substantive and editorial because it was requested by the IPCC authors and that they were more interested by substantive comments. As no evaluation of the quality of comments was done, we cannot evaluate the impact of the attribution process.
We nonetheless received personal communications from several IPCC authors that our comments were very valuable to them, in particular due to the fact that all the sections of all the chapters received the same amount of attention.

P8 l 271-276: You listed several benefits for ECS, although number one is more a benefit for the report (and as written there, it lets the reader assume that the more senior scientists are not very rigorous). You also mention 'recognition' but I couldn't identify information about 'recognition' in the paper. What do you mean and how is it measured?

We discuss in the results of the post-review survey that a large number of participants believe that their participation to the project helped them improve their CV, which means they estimate that there is recognition in the scientific community for this type of project. In addition, after the publication of the report, participants are able to retrace their comments as well as to read the authors' answers to them, as a way of seeing how valuable their comments were (https://www.ipcc.ch/srocc/download/). Furthermore, participants in the group review have their names published online as proof of their participation in the review of the IPCC report.

We certainly did not want to imply that more senior scientists are less rigorous than ECS, but rather, that they may be more solicited for reviews and thus have less time. It is true that we have no proof of this, but on the other hand, having more potential reviewers in the community does lead to a reduced workload for everyone.

P9 l285-300: You wrote interesting suggestions to involve more ECS in the review process although as mentioned above, I don't think that it is a conclusion of your expe- rience but a 'personal' opinion. Moreover, all these suggestions would require a study to measure how much they are already taken into account.

See first comment.

Referee 4: Lonni Besancon

Disclaimer: I have not been involved before with the IPCC report reviewing (although I have read some chapters for curiosity or facts checking) and I don't work in climate science at all. I have,

however, been involved before on open science and open review efforts. I am also myself an ECR (postdoctoral fellow, PhD + 2 years).

In this submission, the authors present the results of a study on the impact of including ECSs in the reviewing of IPCC reports. The paper is well written and very interesting. It presents interesting data and effectively argues for a change in the recruitment of reviewers.

I nonetheless have some issues with the current submission that I will list below

First of all, I find the name ECS to be confusing. In many scientific communities I have been seeing ECRs used instead and I would argue that the authors should perhaps switch to using ECRs instead. Both are fine eventually, I just want the authors to know that in my field of expertise and the other fields I have contacts in, this is not a commonly used term.

We agree with the referee that the term ECR is also used in the literature. Our organisation, APECS, has the term 'ECS' in its name. While APECS also uses the term ECR in their communications, we chose ECS, both for consistency with the name of our organization (AP**ECS**), as well as because our group consisted of only scientists, but not all of them are researchers (for instance a few engineers were part of the group). We agree that to refer to an interdisciplinary group of researchers (from science and non-science fields), the term ECR is more appropriate. However, given the specific, scientific nature of our group of reviewers, we prefer to remain with the term ECS in the present manuscript. We define the term ECS in the first and second sentences of the abstract and the main text the manuscript, respectively. We include this modification (line 41):

> "Early Career Scientists (ECS, also referred to as Early Career Researchers (ECR))"

"Each chapter was distributed by the project leaders to the participants. Depending on the number of participants for each chapter, 10 to 20 pages were assigned to each participant. We attempted to assign whole sections as much as possible. We also attempted to balance the workload and in some instances, reviewers were assigned non-contiguous sections to even out the number pages they were responsible for." I am particularly skeptical about the consequences of that decision. I am not learned in how IPCC reviewing usually works (although I have read some chapters from IPCC reports before), I would argue that a thorough reviewing process probably cannot be completed if a whole chapter is not assigned to reviewers. In particular, I would personally feel extremely uneasy about having to review only parts of a scientific text. While it is true that, at times, I find myself reviewing only the parts of a scientific communication that I feel I have the expertise to review, I nonetheless read through the entire communication in order to understand the context/application/goal better. I would therefore like the authors to clarify what the impact of this decision to review parts of a chapter could be on the results they obtained. In particular, my concern is even more important if we consider the text two paragraph later, stating "Though participants themselves chose the chapter which they would have to review, a significant number of concerns were raised from participants that felt that the section they were assigned did not correspond to their particular expertise." I think this should be really clarified in the submission and potentially highlighted as a clear limitation of the work produced by the authors.

All the chapters of the report were available to the participants if they wanted to read it in its entirety, outside of their specific assigned section. Moreover, we requested that participants provide reviews at least for the designated pages, but did not restrict comments to these specific pages only. We modified the text to reflect this point (Lines 138 – 140):

> "Each chapter was distributed by the project leaders to the participants. Depending on the number of participants for each chapter, 10 to 20 pages were assigned to each participant to review, and in addition, the entire chapter was made available to participants, such that they could provide additional comments outside of the assigned pages, and be able to put into context their assigned section."

"The total workload of the participants was less than the project leaders, who spent an estimated 40 hours to prepare the project, participate in the webinars, read the different chapters in which they were involved, and sort all the comments." How is the workload of participants estimated? I would expect the workload of participants to be vastly heterogeneous, and in particular students to be, potentially, extremely rigorous in their reviewing assignment and therefore take quite some time to complete their review.

As part of a follow-up participation survey, we asked each participant the amount of time that they spent on the review. We included a sentence about this.

> "The total workload of the participants (obtained through a survey after the review) was less than the project leaders, who spent an estimated 40 hours to prepare the project, participate in the webinars, read the different chapters in which they were involved, and sort all the comments."

Line 175 "These three figures are not significantly different." Are the authors talking about statistical significance? If yes how was that determined? Dichotomous reports of statistical results have been shown to be particularly harmful by many different scien- tists (see references [A–F] for instance, although many others are available and usually cited in the the ones I provided hre) and I would like to advise against such reports (in particular if no exact numerical results are reported).

In the section the referee mentions, we are generally talking about statistical significance, and we agree with the referee on the interpretation of the p-value. However, in the case of line 175, given the small sample per category, we limited ourselves to study whether the averages plus/minus 1 standard deviation suggested differences and no statistical test was applied. Given that in all three cases these figures overlap, we inferred that they are not statistically different. In the other cases, since we have a larger database to work with (2155 comments), we applied tests, as described in lines 180 and 185. Yet, we were very careful to not actually say that the numbers were the same, nor different, but just that the differences were not significant, which, as the referee mentions, simply means that we cannot rule out the null-hypothesis, and not that there is no effect. To make this clear, we will modify the text in the following way:

> "The average number of comments per person was 31.8 ± 4.6 (errors on the average, sample size n = 61). The average and standard deviation of the numbers of comments for PhD, Post-docs and Early Career Academics are 39.9 ± 6.9, 28.9 ± 7.7 and 34.9 ± 9.5 respectively (sample sizes n = 26, 21 and 14 respectively). These averages are relatively similar to the range calculated from the overlap of the standard deviations, suggesting that these three figures are not significantly different."

Line 185: It is unclear what significance threshold the authors are using. Alpha = 0.05 is often used, yet the p-value obtained by the authors is greater than this. It is absolutely fine not to use a specific cut-off for significance but in this case I would argue that the authors should not use the word "significant" to describe their results. I also agree with the authors' results interpretation that this result seems interesting and definitely noteworthy, but the authors could simply interpret it in terms of strength of evidence obtained instead of using "significant".

We agree with the referee that significance thresholds are arbitrary and that no single value is correct, as described in the multiple manuscripts given by the referee. For example, a proof requires 5 standard deviations in physics while 3 in geosciences. We provided both the χ^2 and the p-value in order to put our results into context for the readers. To improve clarity, we included a more cautious description:

> "In contrast, the average time spent by reviewers from each academic levels was, at least slightly, significantly different ($\chi 2$= 7.16, p-value = 0.067)"

I would argue that figure 1a is not really appropriate. While having a map is always informative, the high level of clutter in Europe makes it very hard to read. A simple bar chart would have been more helpful in this case, and would have made comparisons much easier. Similarly, the donuts charts in figure 1c are very hard to compare between different categories of researchers. Also it would seem that

the time to get training on reviewing for the study presented was not included in the workload of participants, which can be problematic.

We respectfully disagree. For Fig. 1a, the advantage of a map is to visually present the bias toward Europe and North America (and to some extent Australia/New Zealand) that our group of participants displayed. Significant amount of work has been devoted to making the European countries readable, and we would additionally say that having the big clutter over Europe would almost be intentional as it reflects the geographic bias of our review. For Fig. 1c, we believe that these donut charts are actually quite easy to compare as we are only showing a limited number of values. As for the training time, while we agree that it is interesting, the amount of time is the same for all participants, as the amount of training was the same for everyone. What we instead wanted to provide was a visualization of the time reviewers spent on the review itself, in order to provide an indication of how much time an ECS would need to spend on reviewing part of an IPCC report, in the hopes that this information would motivate more ECS to engage in the IPCC report review process when they realize that the time required is actually not that substantial. Nevertheless, we have included in the caption of the figure the time spent on the training, in order to have a reference to compare to the numbers we show on the figure:

> "c) comments and time spent per participant by academic stage (the average training time was 3.5 hours per participants, regardless of the career level)."

Line 210: "The PhD students provided as many substantive comments as the more experienced participants of the group review (i.e. Post-docs and Early Career Aca- demics), thus the length of the academic career was ruled out as a factor in the ability to effectively produce reviews." seems like an overstatement. The number of substantive comments itselfs is not enough to rule out the length of academic career as a factor in the ability to produce good reviews. While we can trust the categorization made by project leaders, this number alone is not enough to make the conclusion that the authors are making. They are still many things that need to be considered: how critical are the comments made by the reviewers, do the comments identify critical bottlenecks/flaws. . . This specific conclusion should therefore be rephrased and take into account the limitation of this single metric as a measure of ability to review.

It is true that we do not have any metric to evaluate the quality of the comments (see also the discussion with Referee 2) and really tried to focus in this manuscript rather on the effectiveness of comments. At this point, we consider substantive comments not as "good" comments, but more as comments that will affect the substance of the report, which were specifically requested by the IPCC authors (in contrast to editorial comments, as copy editing is supposed to then take care of these, and the IPCC authors would rather not waste time dealing with these). It would be amazing to have an external evaluation of the quality of each individual comment to evaluate if there was an effect of career length, something similar to what is described in the manuscript of van der Veer et al. (2014), but unfortunately, we did not have the possibility to hire an external jury of experts to review the ~ 2000 comments that we produced, and as we only internally sorted comments between substantive and editorial, we decided not to discuss quality, and just limit our discussion to effectiveness, with the limitations highlighted by the referee here. As we do not mention any qualitative aspect of the produced review, we believe our conclusions stand. We mention caveats in the paper to highlight that we only discuss the effectiveness:

> "The PhD students provided as many substantive comments as the more experienced participants of the group review (i.e. Post-docs and Early Career Academics), thus the length of the academic career was ruled out as a factor in the ability to effectively produce reviews. A more comprehensive analysis would benefit from the use of several metrics to determine the quality of the review, and the link to the length of academic careers."

Line 218: "This latter aspect is particularly relevant for the climate change commu- nity considering the need for transparency in the peer-review process (Edwards and Schneider, 2001)." this comment is not particularly clear. The clear need for trans- parency in the PR process for climate science should first be made clear and then tied back to the claims of the authors.

The paper from Edwards and Schneider (2001) is one of the references explaining the need for transparency. We modified the sentence accordingly:

> "Considering the need for transparency in the peer-review process of IPCC reports following the claims of corruption in the peer-review process of the main Assessment Report 2 (AR2) (Edwards and Schneider, 2001), this latter aspect is particularly relevant for the climate change community."

Line 226: "The relatively reduced time commitment might make participating in future IPCC reviews more appealing to ECS." My comment here might be a bit outside of the scope of that submission and review. I will make it nonetheless. Reviewing can be appealing to ECRs if it is somehow also relevant for their careers. While open and non-anonymous reviews seem to be on the rise in many scientific communities, ECRs get a lot of pressure from the publish or perish system (leading to multiple calls for a change of system, see for instance [G]), from their supervisors and from peers which are incompatible with an involvement or a rigorous participation in the peer review sys- tem. While the authors then highlight the benefits that ECRs self-report they would get from participating in the reviewing process, this has to be contrasted with the limitations I have just highlighted. The figures provided by the authors in the next paragraphs are, still, very important and very nice indication that ECRs are willing to participate in the reviewing system and see value in it.

This is a good point. In general, it seems to us that more and more journals provide lists of their reviewers, even the anonymous ones, which is particularly important for ECS (for instance, for Geoscience communication, this can be found here: https://www.geosci-commun.net/2/referees.html). In the case of IPCC reports, the list of the referees is also published alongside the report. As a result, ECS can use this as a proof of commitment in the community, which is also well regarded, in order to advance their careers.

Line 254: "There is no evidence that this attribution process influenced the quality of their comments." This statement is quite bold and unsupported and ties back to one of my first comments at the beginning of this review. The manuscript should clearly state how the authors came up with this conclusion?

See response above (Referee #3).
We modify this sentence to be more neutral:

> "The influence of the attribution process could not be assessed during this project."

Figure 2a, I find it really interesting that the number differ from group peer review to individual review. Indeed, it would seem that the participants, according to the authors, did not communicate much within groups. I wonder if the authors can think of an explanation for that result somehow. Do they have some data on this?

We agree that it is interesting, and unfortunately do not have any explanation. We want to emphasize though, that there were a lot of exchanges within the group, just not necessarily as much as we expected for the pairs of PhD and Postdocs. We could imagine that doing an individual review is slightly more challenging, as you are by yourself.

Line 271: I would particularly like to highlight two things in the argument made by the authors here. First, point number 1 is not a direct benefit for an ECR, but a benefit for scientific communities, for the report at hand, and/or journals/publishers instead, and I would therefore argue that it should not be listed there. Second, I would argue that point 3 is not always true. Many fields and journals still have double/single blind reviewing processes and this argument simply does not apply in this case. This should be highly contrasted by the authors, should it stay in the final version of this manuscript. Instead of these benefits, I would mention that conduction reviews offers other advantages: 1/ better understanding how one's writing can influence how a paper is perceived and therefore improving on one's own writing skills, 2/ keeping up to date with recent lit- erature, 3/ participating and

understanding the reviewing process better and therefore getting an understanding of how one's paper is reviewed by others. Other benefits can be found, (in particular in open review processes [H,I,J,K]) and I wonder whether the authors should focus on these points instead.

We agree that open review processes are the way forward and would hope that more journals would opt for these, but we do not believe that our results provide any additional information regarding this. We agree that the structure of the paragraph could be improved. We modified to:

> "Participating as reviewers has many benefits for ECS, i.e. developing skills such as time management, responsible authorship, review and publication practices and getting recognition for critical review skills within the scientific community. For the scientific community, it also increases the reviewer pool, which could alleviate the workload of senior scientists, and hopefully enhances the scientific rigour of journal articles and reports that support policy making processes."

Line 295: this echoes previous findings from the literature (e.g., [J]).

Thank you very much, we included the citation in the manuscript.

Is the data going to be made available in some way by the authors? There are a few things that do not appear clearly in the manuscript. For instance how did project leader categorize comments? Was some form of agreement score computed? Or did all comments only get reviewed by one project leader? This is not very clear in the current manuscript. Making the data available would help to clear that but this should also be clearer in the manuscript too.

Some of the data will be made available. The number of comments of each category for each participant will be included, but only anonymously. The category of comments were following IPCC guidelines, and while we internally did the evaluation and could modify the categorization (substantive or editorial) for individual comments from reviewers, we also refrained from any qualitative discussions on these comments. We included an additional sentence to highlight this in the method section (lines 169 to 173):

> "We acknowledge that the comment sorting includes some level of subjectivity, as a single chair was evaluating each comment's categorisation (in addition to self-evaluation from the participants), as a result, we only report in the manuscript the total number of comments for each category and do not evaluate the quality of the review."

Update after having read the other reviews: I agree with RC2 that, should this work be close replication of previous work, it should be clearly mentioned in the submission. I don't think this hinders the work conducted by the authors in any way, but it should simply be made clearer. Should it not be a close replication of previous work, then the authors should clearly highlight what the differences are. I disagree with RC1 that figure 1 is excellent. The data is interesting, the visualization is not well chosen though. I agree with RC2 that the guidelines given to ECRs should be made available. This will, in particular help future replications of this work.

Overall, I am very positive about the work done by the authors and I simply think that some rewriting is needed to make some point clearer or more contrasted. I would personally argue that the revisions needed to accept this submission are minor and that the authors should be able to address all concerns raised so far (by myself and the other reviews I have read) in a short revision time.

REFs: [A] https://www.nature.com/articles/d41586-019-00857-9
[B] https://peerj.com/preprints/2921v2/
[C] https://journals.sagepub.com/doi/10.1177/0956797613504966
[D] https://link.springer.com/chapter/10.1007%2F978-3-319-26633-6_13
[E] https://hal.inria.fr/hal-01980268/file/alt.chi_2019_CPDI-authors.pdf
[F] https://discourse.datamethods.org/t/language-for-communicating-frequentist- results-about-treatment-effects/934

[G] https://opensciencemooc.eu/evaluation/2019/10/15/solve-research-evaluation/
[H] https://academic.oup.com/femsle/article/365/19/fny204/5078345
[I] https://doi.org/10.18243/eon/2017.10.8.1
[J] http://dx.doi.org/10.20944/preprints201905.0098.v2