AGU Advances

COMMENTARY
10.1029/2022AV000676

Key Points:

- Need for increased investment of time and resources in foundational scientific activities in climate science
- Need for initiatives that reinforce the curiosity and scientific freedom of the early career climate researchers
- Urgent need for a coordinated action to provide resources to the early career researchers working in under-resourced environments

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

S. Jain,
shipra.npl@gmail.com

Citation:

Jain, S., Mindlin, J., Koren, G., Gulizia, C., Steadman, C., Langendijk, G. S., et al. (2022). Are we at risk of losing the current generation of climate researchers to data science? *AGU Advances*, *3*, e2022AV000676. https://doi.org/10.1029/2022AV000676

Received 9 FEB 2022
Accepted 13 JUN 2022

Peer Review The peer review history for this article is available as a PDF in the Supporting Information.

Author Contributions:

Conceptualization: Shipra Jain, Julia Mindlin, Gerbrand Koren, Carla Gulizia, Claudia Steadman, Gaby S. Langendijk, Marisol Osman, Muhammad A. Abid, Yuhan Rao
Project Administration: Shipra Jain, Valentina Rabanal
Resources: Valentina Rabanal
Writing – original draft: Shipra Jain, Julia Mindlin, Gerbrand Koren, Carla Gulizia

Are We at Risk of Losing the Current Generation of Climate Researchers to Data Science?

Shipra Jain[1,2], Julia Mindlin[3], Gerbrand Koren[4], Carla Gulizia[3], Claudia Steadman[1], Gaby S. Langendijk[5], Marisol Osman[3,6], Muhammad A. Abid[7], Yuhan Rao[8], and Valentina Rabanal[9]

[1]School of Geosciences, The University of Edinburgh, Edinburgh, UK, [2]Centre for Climate Research Singapore, Meteorological Services Singapore, Singapore, Singapore, [3]Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales, Departamento de Ciencias de la Atmósfera y los Océanos, CONICET – Universidad de Buenos Aires, Centro de Investigaciones del Mar y la Atmósfera (CIMA), CNRS – IRD – CONICET – UBA, Instituto Franco-Argentino para el Estudio del Clima y sus Impactos (IRL 3351 IFAECI), Buenos Aires, Argentina, [4]Copernicus Institute of Sustainable Development, Utrecht University, Utrecht, The Netherlands, [5]Climate Service Center Germany (GERICS), Helmholtz-Zentrum Hereon, Hamburg, Germany, [6]Now at Karlsruhe Institute of Technology, Karlsruhe, Germany, [7]Earth System Physics, The Abdus Salam International Centre for Theoretical Physics (ICTP), Trieste, Italy, [8]North Carolina Institute for Climate Studies, North Carolina State University, Asheville, NC, USA, [9]Servicio Meteorológico Nacional (SMN), Buenos Aires, Argentina

Abstract Climate model output has progressively increased in size over the past decades and is expected to continue to rise in the future. Consequently, the research time expended by Early Career Researchers (ECRs) on data-intensive activities is displacing the time spent in fostering novel scientific ideas and expanding the frontiers of climate sciences. Here, we highlight an urgent need for a better balance between data-intensive and foundational climate science activities, more open-ended research opportunities that reinforce the scientific freedom of the ECRs, and strong coordinated action to provide infrastructure and resources to the ECRs working in under-resourced environments.

Plain Language Summary Climate science research can be described by three key foundational activities: developing scientific theories and hypotheses, testing hypotheses using data and theory to generate scientific outcomes, and finally communication of scientific outcomes to inform climate actions, that is, adaptation and mitigation. The progress of our field is influenced by a balance between these activities. We, a group of early career researchers (ECRs), are concerned that the climate community is putting an excessive emphasis on data-intensive activities and the disproportionate investment of time and resources in these activities is leading to a displacement of more foundational scientific activities of our discipline. This not only impedes the scientific progress of our field but also hinders the development of the current generation of climate researchers as they struggle to strike a balance between data-intensive activities and foundational science. For the ECRs, this problem is further deepened by short-term employment contracts, constrained scientific freedom, and resource disparities. This makes our discipline less and less appealing and there is a risk of losing the present generation of climate researchers to data science.

1. Growing Climate Model Outputs

Over the last few decades, climate science has made large strides, partly through copious amounts of climate model output produced by the generations of the Coupled Model Intercomparison projects (CMIPs), an initiative under the World Climate Research Program (WCRP). There has been a progressive increase in the volume of climate model output, hereafter referred to as data, over the course of successive CMIP phases. Its latest phase, namely CMIP6, encompasses ~120 models from ~50 modeling groups, ~20 to 40 petabytes (PB) of output data, from 23 endorsed MIPs (Eyring et al., 2016). Several factors have contributed to the large increases in CMIP data such as increases in model resolution, the number of participating models and modeling centers, increase in the number of experiments from 11 in CMIP3 to 312 in CMIP6 (Petrie et al., 2021), and number of output variables (Balaji et al., 2018; Eggleton & Winfield, 2020; Juckes et al., 2020). Projected estimates show ~300 PB data to be produced within the next few years (Quobyte, 2019) and this is only one example of a global multi-institutional

Writing – review & editing: Shipra Jain, Julia Mindlin, Gerbrand Koren, Carla Gulizia, Claudia Steadman, Gaby S. Langendijk, Marisol Osman, Muhammad A. Abid, Yuhan Rao

data-centric climate research project. As the number of data-centric climate projects is constantly increasing, the data volume is also expected to continue to rise in the future.

While the growing data and data-centric projects offer one way to advance our understanding of the climate system, these projects are a huge responsibility in terms of data production, processing, and analysis, which is largely placed on the current generation of researchers. Early career researchers (ECRs), employed as research staff at the operational centers, doctoral candidates, and post-doc staff at the universities, form a major proportion of the workforce who spend tremendous amounts of their time producing and analyzing climate model datasets, often following prescribed research plans outlined by senior researchers or internationally coordinated endeavors. We are concerned that currently, our community is putting an excessive emphasis on data-intensive activities which consequently inhibits the progress of other important frontiers of climate research such as developing novel research ideas or developing scientific hypotheses.

Research is a creative process and creativity needs time and mental space. Thinking critically about scientific results, exchanging thoughts with other researchers in the community, or finding inspiration in routine scientific tasks including producing or analyzing model outputs—require time and mental space. An undue focus on data-intensive activities simply does not leave us with much time or mental space to indulge in foundational research activities. The current generation is increasingly being pushed toward data-intensive activities and the opportunities that allow us to focus on underpinning climate science questions are continually diminishing.

The success of the current data-centric research projects will depend on how effectively the data can be used by the climate community, and notably ECRs, over the coming years. This implies that over-emphasis on prescribed data-intensive activities will, in fact, refrain us from undertaking creative and novel endeavors, thereby limiting the value of these projects. The reviewers of this manuscript also pointed out that the excessive amounts of data are not only a burden for the ECRs but the entire climate community. If that is to be true, there is an even more pressing need to address this issue because we, as ECRs, are not keen to let this burden cascade to the generation of researchers succeeding us.

In this article, we highlight an urgent need for increased investment of time and resources in foundational scientific activities of our discipline, more open-ended research opportunities that reinforce the scientific freedom and curiosity of the ECRs, and coordinated action to provide equitable resources and infrastructure to the ECRs working in under-resourced environments to strengthen their prospects of pursuing a scientific career. Through this article, our hope is to encourage a dialogue on these issues which will ultimately help in advancing climate science over the next decades.

2. Our Challenges and Concerns

2.1. Making Space for Foundational Climate Science

The foremost challenge, which the increasing data and data-centric projects pose for us, is how to strike a balance between the time expended in data-intensive activities versus foundational climate science, and the fact that this decision is not always ours to make.

In a recent commentary, Emanuel (2020) has raised concerns regarding the risk of losing a whole generation of climate scientists to data. We share the author's concern on how investing a disproportionate amount of time in technical data-intensive activities may lead to a displacement of more foundational activities of our discipline, such as learning how to pose science questions or developing science hypotheses.

The scientific literature presents many examples where major breakthroughs in our field were driven by simple curiosity-driven questions which interrogated scientific data and theories. For example, Lorenz interrogated the amplification of round-off errors in model simulations to unveil the butterfly effect (Lorenz, 1963, 1969), also known as chaos theory. Following Lorenz's discovery, it seemed fundamentally impossible to make predictions with reasonable accuracy beyond two to 3 weeks and the future of long-range forecasting was thought to be doomed. But this predictability limit was further investigated and the prediction skill was demonstrated on longer time scales using slowly varying atmospheric and oceanic conditions such as changes in sea-surface temperatures (e.g., Charney & Shukla, 1981; Shukla, 1981). The unexpected discovery of the ozone hole (Farman et al., 1985) and the depletion of stratospheric ozone by anthropogenic emissions of chlorofluorocarbons (Solomon et al., 1986) led to profound advances in atmospheric chemistry and even changes in global

environmental policy (Solomon, 2019; Tripp, 1987; UN Montreal Protocol). The decades of collective effort led by the climate community to explain the Quasi-Biennial Oscillation (QBO) when neither observations nor theory behind tropical atmospheric waves that cause QBO existed at that time (Baldwin et al., 2001; Plumb, 1977). The identification of QBO disruption from a tiny signal in the observations much before it became a full-fledged disruption (Osprey et al., 2016). And more recently, the current efforts by the climate community to explain why some models under the CMIP6 initiative show higher climate sensitivity as compared to the previous CMIP5 generation (e.g., Forster et al., 2020; Sherwood et al., 2020; Zelinka et al., 2020). Scientists are investigating if such a high response of global surface temperatures to the external climate forcings arises due to the errors in climate models and the origin of these errors (e.g., Hausfather et al., 2022; Myers et al., 2021; Sherwood et al., 2020; Voosen, 2019, 2022). These examples show the importance of curiosity-driven research and the freedom to embark on unknown science territories to push frontiers in climate sciences.

Our field sometimes even requires skills that go beyond traditional scientific training such as a posteriori knowledge and intuitive guesswork which is often employed by forecasters to make weather predictions or develop seasonal and sub-seasonal outlooks of El Niño-Southern Oscillation, Indian Ocean Dipole, and Madden-Julian Oscillation when model results are not enough to reach a consensus. This implies that investment of time in the synthesis of the literature, critical and curiosity-driven thinking, theoretical comprehension, research writing, and communication, are needed to break new grounds. However, in contrast, our most common daily struggles are related to data science, for instance, downloading and transferring data, debugging codes, and solving compatibility issues between software libraries. Working with large amounts of data also poses the challenge of applying emerging new methodologies, such as artificial intelligence algorithms, effectively. It is not just about creating a few more libraries or learning a few more analysis tools to make research more efficient. As software and hardware are under continuous development, the data analysis challenges will continue to occur and demand time throughout our careers.

2.2. The Culture of Prescribed Science and the Problem of Continuity

With a tremendous increase in computational power and resources, at least in well-resourced nations with a strong tradition in climate research, it is easier than ever to run a model and produce output. Together with the strengthened international coordination for model simulations globally (e.g., CMIP), this has led to the immense growth of climate model output data, resulting in an increased need for analysis. On some occasions, the sole existence of the vast amounts of data, not scientific hypotheses, imposes the need to do more data analysis. For instance, in our experience, seniors would stress that "*these datasets are lying around and shall be analyzed*," spinning out entire research projects based on this imposed need. The data-centric projects focusing predominantly on data analysis also have the potential to reap fast results, which can quickly turn into publications, and therefore these are very fitting for the short-term projects. Due to the shorter project timelines, such data-centric studies demand the science objectives to be increasingly predefined. This, in turn, nurtures the culture of prescribed research.

Whilst an intricately prescribed research plan and pre-agreed objectives could be more productive in terms of published output and deliverables as compared to the open-ended research, it poses challenges for the ECRs to bring their ideas to the projects or design their own research paths. The pressure of producing deliverables and publications under tight timelines simply does not leave any time for the ECRs to deviate from the predefined research plan, or to confront the questions that have been handed to them by their seniors.

Traditionally, the purpose of postdoc jobs was to serve as an interim phase—a route to independence—for the ECRs. The phase which allows them the scientific freedom to confront current theories and research questions, to develop essential skills such as leadership and management, before they transition into more forefront permanent positions. However, in the current scenario, postdocs in our discipline are sometimes viewed as cheap laborers with constrained scientific freedom. Their role is limited to producing or analyzing model outputs while the science and decision making is steered by the senior researchers. This also generates a sense of unbelongingness in some ECRs as they struggle to take charge of the projects they are employed on and find their positioning in this field. This all makes our field less and less appealing.

In addition, the ECRs in short-term projects struggle to maintain continuity in their research which further confounds their ability to make meaningful contributions (Kreeger, 2004; Waaijer et al., 2017; Woolston, 2020). In summary, while it is possible to do incremental science and apply what we already know in short-term

data-centric projects, producing research that can break new grounds in climate science needs time—which this increasingly fast-paced and prescribed culture does not allow.

2.3. Data Procurement in Under-Resourced Regions and Institutions

In addition to the challenges outlined above, a large amount of data is a barrier itself, as it requires reliable computational facilities to store and process the data. Though multi-institutional data-driven projects, for example, CMIP, provide open access to data to develop regionally focused research, challenges remain for ECRs in under-resourced environments, sometimes even in high-income countries, to access and analyze data for climate research. For example, frequent power outages and internet disconnections in developing nations still make the big climate datasets challenging to access. The most recent pandemic only exacerbated the issue of research infrastructure inequity (Carr et al., 2021). Limited storage capacities in resource-lacking organizations force climate researchers to repeatedly download, post-process, and delete the raw data to make space for the datasets. While central facilities for data storage and analysis, such as JASMIN for UK researchers, are present for the ECRs in well-resourced countries, their counterpart ECRs in under-resourced nations struggle to find support, which puts even more pressure on these researchers and further limits their opportunities to contribute to science or to pursue a scientific career in their own nations (e.g., Dwivedi et al., 2022). There are growing movements to democratize research data and infrastructure via cloud computing technology to support the development of climate research in under-resourced regions, for example, Group on Earth Observation—Planetary Computer Program, Amazon Sustainability Data Initiative, and analysis-ready CMIP6 data on the cloud with Pangeo. However, these initiatives require new skills that are not yet included in the training of ECRs, thus still struggling to achieve their purpose.

3. Open Questions for the Climate Science Community and Potential Ways Forward

We are a generation that will make significant advances in our field using the vast amounts of climate data produced by internationally coordinated endeavors (e.g., CMIP, CORDEX). However, for a discipline like ours, solely analyzing data or applying statistics without deeper causal thinking has little to offer to our community (e.g., Kretschmer et al., 2021; Runge et al., 2019; Shepherd, 2021). To ensure that theory and our understanding of complex mechanisms develop at the same pace as data analysis methodologies, ECRs need to spend time in both handling data and gaining experience in more foundational activities of our discipline (e.g., Jakob, 2014).

The Intergovernmental Panel of Climate Change (IPCC) in its latest Working Group II report (IPCC AR6 WGII report, 2022) has highlighted that we are rapidly losing the window to secure a livable future, implying that the upcoming years are extremely crucial for climate change mitigation and adaptation. Mitigation can be favored by data-centric and top-down approaches because action has to aggregate on a global scale to reduce emissions. Whereas adaptation is a local problem and for climate research to inform local climate actions, we need visionary leaders who are capable of engaging in solution-driven science and drawing insight from multiple disciplines and lines of evidence—beyond solely using climate model data (Jakob, 2014; Rodrigues & Shepherd, 2022; Stevens et al., 2016). To ensure that the current ECRs are trained for this challenge, the research opportunities should be geared toward solving *pressing* and exciting research problems, with climate model data, and other approaches.

Here, we suggest potential ways forward to address the issues that we raised. We realize that this requires a shift in scientific culture and practice which can only be achieved when this is widely discussed and supported in the climate science community. Our suggestions should therefore not be considered as set in stone, as they are intended to be as starting points for a wider discussion.

3.1. Prioritizing Using Existing Model Outputs

There is already an abundance of model data available, much of it would perhaps take years to thoroughly analyze. Therefore, we think that the current efforts in our discipline should prioritize using the existing model data to its full potential and developing approaches to use it effectively to answer pressing research questions and to inform climate actions. If we are to produce more data on top of what is currently available, then we should first revisit which scientific and societal problems are most urgent and what type of data we need to address those questions.

The motivation for evermore data production must be clearer and should not be predominantly driven simply by the existence of international coordination structures or the availability of human or computational resources.

3.2. Stronger Push for Creative Over Prescribed Research

We recognize the need for benchmarking in climate modeling and comparison of different model generations. However, in addition to prescribed tasks of model evaluation and assessment, it is also important for the data-centric projects (e.g., CMIP) to be reasonably scaled which provides time and space for the ECRs to pursue exploratory research and undertake creative endeavors. If this time and space are not consciously assured, model inter-comparison will easily become a quality-control activity in place of a creative endeavor. In that light, we ask the climate community: are the additional scientific insights coming out of CMIP by producing more data and adding more models proportional to the time and resources going in? And perhaps most importantly, is it worth pushing the current and upcoming generations of climate researchers to keep producing data, or is this better invested in fostering novel scientific ideas? There is a need to further embark on these questions through an open inter-generational dialogue both at international and local levels. This dialogue could be through the existing participation of ECRs in the relevant WCRP activities, for example, Lighthouse Activities such as My Climate Risk or Explaining and Predicting Earth System Change (EPESC), Regional Information for Society (RifS), or engagement of the broader ECR community via surveys, workshops, and online meetings.

3.3. Sharing of Resources

The current remit of capacity building in under-resourced nations mainly involves the scientific and technical training of local researchers. The training of the researchers is necessary but not sufficient. The lack of accessible resources and reliable research infrastructure, particularly in under-resourced countries, pushes the capable and trained ECRs to migrate to other countries, which, in turn, disadvantages the research community in these regions (Jain, 2020). There exists a stark discrepancy in resource distribution within more well-resourced countries as well. For example, a third (33%) of all US higher education research and development expenditures in the geosciences, atmospheric sciences, and ocean sciences, is concentrated in only 10 institutions (fiscal year 2020). Half (50%) of these R&D expenditures occur in only 20 institutions (NCSES, 2021). Therefore, we emphasize an urgent need to create equitable globally accessible computational and storage solutions, similar to the Centre for Data Analysis (CEDA) or JASMIN in the UK, which allow researchers to perform data analysis at the site of data, without the need for processing or storage on local systems. Global access to computational infrastructure will benefit not only the ECRs but scientists of all career stages from under-resourced nations and organizations.

3.4. Dedicated Software Engineers and Data Managers

In addition to the global-centric resources, we urge scientific institutions to recognize the need for dedicated software engineers and data managers to reduce the burden of data management and analysis that is now predominantly on the ECRs. In addition to allowing ECRs to focus more on scientific interpretation, these specialists can increase efficiency, as they are better trained in coding, data management, and other practical computational tasks. This already exists in major climate modeling centers (e.g., NCAR, ECMWF, UK Met Office, and Barcelona Supercomputing Center) and some research organizations (e.g., eScience Center in the Netherlands) but we think that this should be a common practice across research institutions. This would require funding agencies to reserve a budget specifically for these tasks acknowledging that recruiting data engineers is now a normal scientific practice. Another potential benefit of this practice is that the data engineers can bring fresh ideas into research groups as they work on topics that may not be on the radar of the climate scientists themselves.

3.5. Streamlining Post-Production Activities

Joining forces to produce data is incredibly useful but leaving it to the next generation to navigate through the data challenges diminishes the purpose of such initiatives. For gigantic coordinated projects (e.g., CMIP), equally gigantic *coordinated* efforts focusing on post-production activities are needed. Streamlining data analysis, robustly identifying errors in current models, understanding their causes, highlighting those to the modeling groups to support model development, and developing interactive and robust communication methodologies

(beyond reports and briefings) that inform climate actions, need more attention and coordinated efforts from our community. To achieve some of these objectives, several data management and analysis tools have recently begun to emerge. Examples include the Pangeo Forge platform (Stern et al., 2022) and initiatives like Nicholls et al. (2021), where large amounts of data are transformed into an analysis-ready format, as well as diagnostic tools such as ESMValTool (Righi et al., 2020), which provides a set of scripts that can be applied to a wide range of data. This growing ecosystem of tooling can substantially reduce the time expended on routine tasks such as data download, data preprocessing, and basic model evaluation. However, these efforts are largely community-driven, and therefore professionalizing these initiatives and supporting them financially would be needed to ensure that they continue to develop and fulfill their purpose in the longer term.

3.6. Improving the Link Between Model Developers and Data Analysts

While most of the model development is undertaken by the operational modeling centers, a lot of data analysis and research that can inform model development initiatives is undertaken by the universities and research institutes. The weak link between the modeling centers and the research community elsewhere could be further strengthened to inform the developers about promising research that has the potential to improve their models and also increase the uptake of the data held by the former. Open forums on model development, both online and offline, targeted toward collecting feedback not only from the modeling groups but larger research communities (such as the post-CMIP6 survey coordinated by WCRP which gathers community input to shape the next phases of CMIP and optimize their usefulness) could be the way forward. Initiatives like NOAA testbeds (https://www.testbeds.noaa.gov/), which foster the transition of scientific advances from the research community to improve NOAA forecast products and services, have proven to facilitate the synergies between model producers and users and could be thought of for the CMIP initiative.

3.7. Opportunities to Maintain Continuity in Research

Though we recognize the urgency of our field, we also acknowledge that building science that can feed into the development of climate models and advance science is a complex, slow, and steady process. In order to achieve this, the ECRs need more stable, long-term employment opportunities that allow them to maintain continuity in their research. Increasing the share of funding toward independent ECR-led research positions, such as the Marie Skłodowska-Curie Actions as opposed to multi-million euro grants that eventually hire ECRs to perform prescribed research activities, can help achieve this objective. Another potential avenue for the ECRs is to tap into the longer-term focused research activities through community-driven organizations such as new WCRP Lighthouse activities. In return, more bottom-up efforts within these structured organizations that allow ECRs to spearhead research, with support and guidance from experts, can also provide a sense of continuity to the ECRs and shape the course of their scientific careers. However, embedding these activities in their daily routine without losing work-life balance would remain a challenge for the ECRs and the community.

3.8. Work Culture Improvements

The ECR community can benefit immensely by improving work-culture practices. However, implementing cultural shifts have always been difficult as they rely on individual efforts and structural changes. Nevertheless, we encourage supervisors to acknowledge that striking a balance between data analysis and underpinning science is an emerging challenge for the ECRs. Technical data analysis skills are obviously useful, and perhaps that is what makes us employable, but learning those skills is a never-ending process. Therefore, research positions must ensure that the balance between data-intensive and other foundational scientific activities is achieved from the beginning irrespective of the duration of the positions. Project leaders should learn to manage their own expectations, particularly from short-term projects, before imposing their expectations or interests on the ECRs. In summary, more opportunities that provide time and mental space to the ECRs to undertake creative endeavors can help build long-term and robust capability in our field.

3.9. Voice Your Opinion

We are aware that we are a small group of the ECRs and our concern may not be representative of the entire climate community. That is why we have created this online form (https://forms.gle/hbWgwKjbiytCdJ9X6) to collect feedback from other ECRs as well as experienced researchers within the climate community about their perspectives on growing data-intensive activities, and potential ways forward outlined in this piece. If you do not have access to this form, we encourage you to download Appendix A and send the filled form to us or discuss alternative options at datafatigue@yess-community.org. We plan to analyze and report the feedback received through this form in a follow-up perspective paper. We think that open-forum discussions and coordinated efforts at the community, organizational, and individual levels can help solve this problem and make our community grow. This will play a key role in shaping the future of climate science as well as the next generation of climate researchers.

Appendix A

A1. Your Perspective on Growing Data-Intensive Activities in Climate Sciences

Climate model data has been increasing progressively over the past decades. This is leading to a shift toward data-intensive activities in climate sciences which in turn is posing challenges for the current generation such as striking a balance between time expended in data-intensive and foundational climate science activities. Through this form, we want to know your perspective on these growing data-intensive activities in climate sciences and what we can do to help early career researchers (ECRs) achieve this balance.

Filling this form will only take a few minutes but will allow us to better understand the challenges you are facing and how we can best address them and make our community grow. You can also find a version of this form online at: https://forms.gle/hbWgwKjbiytCdJ9X6.

What can you share?

We are keen to know the challenges or opportunities the big climate data provides you. You are also welcome to share your concerns, perspective, experiences, or stories relevant to this topic.

What will we do with your feedback?

We plan to analyze the feedback received through this form and report the key findings in a follow-up perspective paper.

Please note that by filling out this form, you automatically provide consent to use the information for a follow-up perspective paper, and presentation in meetings and conferences. This form is anonymous by default and all feedback received through this form will be reported as anonymous. There is also an option to provide your name and email address at the end, in case you wish to be identified or contacted. Questions followed by an asterisk (*) are mandatory.

If you have any further questions on this form, please contact datafatigue@yess-community.org.

*Q1: Your country/region of origin**

*Q2: Your country/region of work**

*Q3: Your place of work**

- o Local/regional/national government (e.g., National Weather Services)
- o University
- o Non-university research institute
- o International agency (e.g., WCRP, WMO)
- o Private company
- o Other:

*Q4: Your career Stage**

- o Master student
- o PhD student

- o Postdoc
- o Research Scientist or Professor
- o Other:

*Q5: Do you identify yourself as an ECR (i.e., within 7 years of your highest degree)?**

- o Yes
- o No

*Q6: Does your day job or work include producing or analyzing climate model data?**

- o Yes, I produce or analyze climate model data
- o No, I do not produce or analyze climate model data
- o Other:

Q7: In your opinion, what key opportunities or challenges do big climate data and data-intensive activities pose for ECRs?

Q8: Are you personally affected by large data or data-intensive activities? If yes, please explain briefly how?

Q9: Would you agree that striking a balance between data-intensive and other foundational science activities is a challenge for the ECRs? In your opinion, what possible initiatives can help address this challenge?

Q10: Would you agree that ECRs in interim positions (e.g., post-doc jobs) lack scientific freedom and struggle to maintain continuity in their research? What possible initiatives would you suggest to address this challenge?

Q11: Would you agree that ECRs in under-resourced nations lack the basic infrastructure (e.g., computer software and hardware) to pursue a career in climate research? What possible initiatives would you suggest to address this challenge?

Q12: Any other relevant input or feedback you would like to provide?

Q13: In case you wish to be identified or contacted, please provide your name and email here.

Thank you for your feedback. In case you have any questions or thoughts you would like to share, please reach out to us by email at datafatigue@yess-community.org.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

No data was used in this manuscript.

Acknowledgments

The authors thank the Young Earth System Scientists (YESS) community for bringing our team together and providing us with a platform to discuss the thoughts presented here. The authors also thank Cecile B. Menard for her feedback on the initial version of this manuscript. The authors thank the editor Bjorn Stevens, reviewer Nadir Jeevanjee and two anonymous reviewers for their very constructive suggestions and criticism which have helped us immensely to shape some of the arguments. This piece has greatly benefited from the editor's advice on being less apologetic. The authors also thank Adam A. Scaife for sharing the story behind the discovery of QBO disruption and Nick Byrne for his comments on software packages for the Earth System Community. The authors did not receive any financial support for this manuscript.

References

Balaji, V., Taylor, K. E., Juckes, M., Lawrence, B. N., Durack, P. J., Lautenschlager, M., et al. (2018). Requirements for a global data infrastructure in support of CMIP6. *Geoscientific Model Development*, *11*(9), 3659–3680. https://doi.org/10.5194/gmd-11-3659-2018

Baldwin, M. P., Gray, L. J., Dunkerton, T. J., Hamilton, K., Haynes, P. H., Randel, W. J., et al. (2001). The quasi-biennial oscillation. *Reviews of Geophysics*, *39*(2), 179–229. https://doi.org/10.1029/1999rg000073

Carr, R. M., Lane-Fall, M. B., South, E., Brady, D., Momplaisir, F., Guerra, C. E., et al. (2021). Academic careers and the COVID-19 pandemic: Reversing the tide. *Science Translational Medicine*, *13*(584), eabe7189. https://doi.org/10.1126/scitranslmed.abe7189

Charney, J. G., & Shukla, J. (1981). Predictability of monsoons. *Monsoon dynamics*, *99*, 109–110. https://doi.org/10.1017/cbo9780511897580.009

Dwivedi, D., Santos, A. L. D., Barnard, M. A., Crimmins, T. M., Malhotra, A., Rod, K. A., et al. (2022). Biogeosciences perspectives on integrated, coordinated, open, networked (ICON) science. *Earth and Space Science*, *9*(3), e2021EA002119. https://doi.org/10.1029/2021EA002119

Eggleton, F., & Winfield, K. (2020). Open data challenges in climate science. *Data Science Journal*, *19*(1), 52. https://doi.org/10.5334/dsj-2020-052

Emanuel, K. (2020). The relevance of theory for contemporary research in atmospheres, oceans, and climate. *AGU Advances*, *1*(2), e2019AV000129. https://doi.org/10.1029/2019av000129

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Farman, J., Gardiner, B., & Shanklin, J. (1985). Large losses of total ozone in Antarctica reveal seasonal ClOx/NOx interaction. *Nature*, *315*(6016), 207–210. https://doi.org/10.1038/315207a0

Forster, P. M., Maycock, A. C., McKenna, C. M., & Smith, C. J. (2020). Latest climate models confirm need for urgent mitigation. *Nature Climate Change*, *10*(1), 7–10. https://doi.org/10.1038/s41558-019-0660-0

Hausfather, Z., Marvel, K., Schmidt, G. A., Nielsen-Gammon, J. W., & Zelinka, M. (2022). Climate simulations: Recognize the 'hot model' problem. *Nature*, *605*(7908), 26–29. https://doi.org/10.1038/d41586-022-01192-2

IPCC. (2022). *Climate change 2022: Impacts, adaptation, and vulnerability. Contribution of working group II to the sixth assessment report of the Intergovernmental Panel on Climate Change* [H.-O. Pörtner, D. C. Roberts, M. Tignor, E. S. Poloczanska, K. Mintenbeck, A. Alegría, et al., Eds.]. Cambridge University Press. In press. Retrieved from https://www.ipcc.ch/report/ar6/wg2/about/how-to-cite-this-report/

Jain, S. (2020). Hope, fear and climate scientists. Nature Index. Retrieved from https://www.natureindex.com/news-blog/hope-fear-and-climate-change-scientists-research-future

Jakob, C. (2014). Going back to basics. *Nature Climate Change*, *4*(12), 1042–1045. https://doi.org/10.1038/nclimate2445

Juckes, M., Taylor, K. E., Durack, P. J., Lawrence, B., Mizielinski, M. S., Pamment, A., et al. (2020). The CMIP6 data request (DREQ, version 01.00.31). *Geoscientific Model Development*, *13*(1), 201–224. https://doi.org/10.5194/gmd-13-201-2020

Kreeger, K. (2004). Short-term limbo. *Nature*, *427*(6976), 760–761. https://doi.org/10.1038/nj6976-760a

Kretschmer, M., Adams, S. V., Arribas, A., Prudden, R., Robinson, N., Saggioro, E., & Shepherd, T. G. (2021). Quantifying causal pathways of teleconnections. *Bulletin of the American Meteorological Society*, *102*(12), E2247–E2263. https://doi.org/10.1175/bams-d-20-0117.1

Lorenz, E. (1969). The predictability of a flow which posses many scales of motion. *Tellus*, *21*(3), 289–307. https://doi.org/10.1111/j.2153-3490.1969.tb00444.x

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, *20*(2), 130–141. https://doi.org/10.1175/1520-0469(1963)020<0130:dnf>2.0.co;2

Myers, T. A., Scott, R. C., Zelinka, M. D., Klein, S. A., Norris, J. R., & Caldwell, P. M. (2021). Observational constraints on low cloud feedback reduce uncertainty of climate sensitivity. *Nature Climate Change*, *11*(6), 501–507. https://doi.org/10.1038/s41558-021-01039-0

National Center for Science and Engineering Statistics (NCSES). (2021). *Higher education research and development: Fiscal year 2020*. NSF 22-311. National Science Foundation. Retrieved from https://ncses.nsf.gov/pubs/nsf22311/

Nicholls, Z., Lewis, J., Makin, M., Nattala, U., Zhang, G. Z., Mutch, S. J., et al. (2021). Regionally aggregated, stitched and de-drifted CMIP-climate data, processed with netCDF-SCM v2. 0.0. *Geoscience Data Journal*, *8*(2), 154–198. https://doi.org/10.1002/gdj3.113

Osprey, S. M., Butchart, N., Knight, J. R., Scaife, A. A., Hamilton, K., Anstey, J. A., et al. (2016). An unexpected disruption of the atmospheric quasi-biennial oscillation. *Science*, *353*(6306), 1424–1427. https://doi.org/10.1126/science.aah4156

Petrie, R., Denvil, S., Ames, S., Levavasseur, G., Fiore, S., Allen, C., et al. (2021). Coordinating an operational data distribution network for CMIP6 data. *Geoscientific Model Development*, *14*(1), 629–644. https://doi.org/10.5194/gmd-14-629-2021

Plumb, R. A. (1977). The interaction of two internal waves with the mean flow: Implications for the theory of the quasi-biennial oscillation. *Journal of Atmospheric Sciences*, *34*(12), 1847–1858. https://doi.org/10.1175/1520-0469(1977)034<1847:tiotiw>2.0.co;2

Quobyte. (2019). Digital infrastructure for a world-leading environmental data facility. Retrieved from https://www.quobyte.com/case-studies/stfc

Righi, M., Andela, B., Eyring, V., Lauer, A., Predoi, V., Schlund, M., et al. (2020). Earth System model evaluation tool (ESMValTool) v2. 0–technical overview. *Geoscientific Model Development*, *13*(3), 1179–1199. https://doi.org/10.5194/gmd-13-1179-2020

Rodrigues, R. R., & Shepherd, T. G. (2022). Small is beautiful: Climate-change science as if people mattered. *PNAS Nexus*, *1*(1), pgac009. https://doi.org/10.1093/pnasnexus/pgac009

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., et al. (2019). Inferring causation from time series in Earth system sciences. *Nature Communications*, *10*(1), 2553. https://doi.org/10.1038/s41467-019-10105-3

Shepherd, T. G. (2021). Bringing physical reasoning into statistical practice in climate-change science. *Climatic Change*, *169*(1–2), 2. https://doi.org/10.1007/s10584-021-03226-6

Sherwood, S. C., Webb, M. J., Annan, J. D., Armour, K. C., Forster, P. M., Hargreaves, J. C., et al. (2020). An assessment of Earth's climate sensitivity using multiple lines of evidence. *Reviews of Geophysics*, *58*(4), e2019RG000678. https://doi.org/10.1029/2019rg000678

Shukla, J. (1981). Dynamical predictability of monthly means. *Journal of the Atmospheric Sciences*, *38*(12), 2547–2572. https://doi.org/10.1175/1520-0469(1981)038<2547:dpomm>2.0.co;2

Solomon, S. (2019). The discovery of the Antarctic ozone hole. *Nature*, *575*(7781), 46–47. https://doi.org/10.1038/d41586-019-02837-5

Solomon, S., Garcia, R., Rowland, F., & Wuebbles, D. J. (1986). On the depletion of Antarctic ozone. *Nature*, *321*(6072), 755–758. https://doi.org/10.1038/321755a0

Stern, C., Abernathey, R., Hamman, J., Busecke, J., Wegener, R., & Sterzinger, L. (2022). Pangeo Forge: Crowdsourcing analysis-ready, cloud optimized data for ocean, weather, and climate science. In *102nd American Meteorological Society Annual Meeting*. AMS.

Stevens, B., Sherwood, S. C., Bony, S., & Webb, M. J. (2016). Prospects for narrowing bounds on Earth's equilibrium climate sensitivity. *Earth's Future*, *4*(11), 512–522. https://doi.org/10.1002/2016ef000376

Tripp, J. T. (1987). The UNEP Montreal protocol: Industrialized and developing countries sharing the responsibility for protecting the stratospheric ozone layer. *NYUJ Journal of International Law and Politics*, *20*, 733.

Voosen, P. (2019). New climate models forecast a warming surge. *Science*, *364*(6437), 222–223. https://doi.org/10.1126/science.364.6437.222

Voosen, P. (2022). Use of 'too hot' climate models exaggerates impacts of global warming. *Science*. https://doi.org/10.1126/science.abq8448

Waaijer, C. J. F., Belder, R., Sonneveld, H., van Bochove, C. A., & van der Weijden, I. C. M. (2017). Temporary contracts: Effect on job satisfaction and personal lives of recent PhD graduates. *Higher Education*, *74*(2), 321–339. https://doi.org/10.1007/s10734-016-0050-8

Woolston, C. (2020). Postdocs under pressure: 'Can I even do this any more?' *Nature*, *587*(7835), 689–692. https://doi.org/10.1038/d41586-020-03235-y

Zelinka, M. D., Myers, T. A., McCoy, D. T., Po-Chedley, S., Caldwell, P. M., Ceppi, P., et al. (2020). Causes of higher climate sensitivity in CMIP6 models. *Geophysical Research Letters*, *47*(1), e2019GL085782. https://doi.org/10.1029/2019GL085782