Organising a collaborative online hackathon for cutting-edge climate research

James Thomas[1], Emma J. Stone[2,3], Dann M. Mitchell[2,3], William Seviour[4], Clair Barnes[5], Hannah Bloomfield[3,6], Julia Crook[7], Hayley Jones[8] and Calum MacLeod[9]

[1]*Jean Golding Institute, University of Bristol, Bristol, UK*
[2]*Cabot Institute for Environmental Change, Bristol, UK*
[3]*Geographical Sciences, University of Bristol, Bristol, UK*
[4]*Global Systems Institute and Mathematics, University of Exeter, Exeter, UK*
[5]*Statistical Science, University College London, London, UK*
[6]*Meteorology, University of Reading, Reading, UK*
[7]*Earth and Environment, University of Leeds, Leeds, UK*
[8]*Met Office, Exeter, UK*
[9]*CDT Environmental Intelligence, University of Exeter, Exeter, UK*

Introduction

Hackathons are an exciting way in which a group of people who do not usually work together can collaborate on one or more self-contained challenges over a short but concentrated period of time. They may be from separate institutions and communities, so participation also fosters the creation of new networks and possible future collaborations. Hackathons are not a new idea – they originate from the field of software engineering (Briscoe and Mulligan, 2014; OpenBSD, 2021). When used for cutting-edge research in the climate science community, however, there are a unique set of barriers to overcome.

The size and complexity of datasets from climate modelling require significant domain knowledge and resources in order to analyse them (Schnase *et al*., 2016). The CMIP6 dataset, for example, is now a federation of 23 separate projects with an estimated up to 40 petabytes of data (Eyring *et al*., 2016; World Climate Research Programme, 2020). Servers often need enough storage to store datasets locally, so that they can be loaded in a timely manner. The processing power and memory of the servers also needs to be able to handle the load. Getting access to and becoming familiar with computation platforms can create a significant barrier to participation, and in a short hackathon event, there is no time to waste getting access to and configuring systems.

The COVID-19 pandemic has meant virtual events (including hackathons) have become more common. These can encourage wider participation, particularly from early career researchers (ECRs) or those with limited travel budgets. However, they have their own drawbacks, especially the need to prepare so much in advance because organising something on-the-fly is very difficult online. It is important to document what works well so that learning from previous events can help make future events even better. In the future, it is possible that hybrid or blended events (a combination of virtual and in-person) will become popular, bringing with them further challenges.

This article has been informed by the Climate Data Challenge hackathon series held in the spring and summer of 2021 by the Met Office's academic partner universities (Met Office, 2021b). In particular, the authors draw from their experiences of the CMIP6 Data Hackathon (Mitchell *et al*., 2022; Fung, 2021), a 3-day hackathon led by the University of Bristol's world-renowned Jean Golding Institute for Data Intensive Research and Cabot Institute for the Environment,[1] which supported over 100 participants across 10 project groups (Figure 1). The hackathon series demonstrated that in all cases, organising and running an online hackathon will take considerably longer than one might expect. We found that an

[1]The full series of hackathons was: Met Office: Marine & coastal, nature-based solutions and sustainable development (Met Office, 2021a); University of Leeds: Future weather in Africa (Marsham and The Leeds Met Office Hackathon Organising Committee, 2021; Crook *et al*., 2022); University College London: Extreme events (Brierley *et al*., 2021); University of Bristol: Analysis of CMIP6 data (Thomas and The CMIP6 Data Hackathon Organising Committee, 2021; Mitchell *et al*., 2022); University of Exeter: Sustainable finance (Harrison, 2021a); Universities of Oxford and Reading: Climate risks in future energy systems (Fallon *et al*., 2021; Harrison, 2021b).

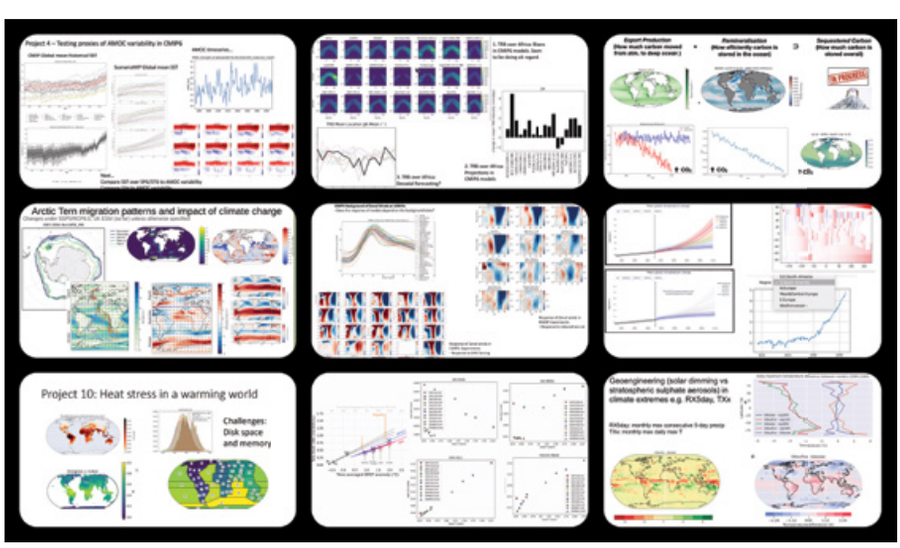

Figure 1. Group update slides from the end of day two (out of three) of the CMIP6 Data Hackathon (Mitchell et al., 2022). Each slide was presented by a group member to all the other participants of the hackathon, with around 2min allocated per group.

organising committee with representation from the science, data, and administrative domains was essential. In particular, giving participants access to a platform such as the JASMIN data analysis facility for environmental science (Lawrence *et al.*, 2013) negates many of the computational challenges they may face.

The rest of this article focuses on the following: why hackathons are beneficial to participants, structuring an event to remove barriers and promote inclusion, the methods that can be employed and next steps for organising a hackathon.

Advantages to participants

Many scientists will go into a hackathon hoping to develop brand new science or the next big scientific software library, going on to publish their work and present results at conferences. A subset of challenge groups will continue collaborating after the end of the event, and their ideas will turn into projects and eventually result in publications. But this will not be the case for everyone. Hackathons are great for getting communities engaged, working on something worthwhile and tackling 'low-hanging fruit' in newly-published datasets.

The benefits to participants are numerous. Often they will work with people outside of their normal sphere (those from other disciplines, institutions or industrial partners). This is especially useful for ECRs who often do not yet have large networks. A hackathon provides the opportunity to work on real-world problems and can lead to an engaging yet self-contained piece of work, much shorter-term than, for example, a PhD. And yet, the techniques that participants develop in working with large-scale datasets on cloud or high-performance computing infrastructure will be readily transferable to any part of earth system science – even the study of exoplanetary atmospheres. Participants will get exposure to best-practice software tools used across data science and research software engineering – even if they do not have time to get to grips with them, they are at least being nudged in the right direction.

Structuring an event to remove barriers and promote inclusion

Encouraging applications from ECRs is likely to be mutually beneficial, as they have the most to gain from a hackathon. Targeting where the event is advertised and emphasising the benefits of participating should increase the number of applications from prospective participants. Pre-requisite skills should be clearly but carefully stated: '*experienced user of Python*' may put some people off, when '*have done data analysis before, using Python or a similar programming language, or willing to learn about the tools we use*' may be more appropriate. A mix of participants – some from less data-focused disciplines who may have no coding experience at all – can bring new perspectives to a problem, so it is important not to inadvertently exclude them. However, this may require more curation of the make-up of teams, to ensure they have the mix of skills they need to carry out their project.

It is essential to take meaningful steps to assist groups that would not usually be able to attend an event. This includes promoting participation amongst typically underrepresented groups, such as (but not limited to) people of a specific age, disability, sexual orientation, religion, race, ethnicity or gender. Institutions should provide support and training for achieving this equality, diversity and inclusion. But even people outside of these groups may lack the computing resources to participate and conduct data analyses, or may have financial, time or other constraints. Video calls, although helpful, should not be essential – not everyone may have a fast enough internet connection. By using a shared computation platform with an interface such as JupyterLab (which is accessible with only a web browser, see Figure 2), the necessary computing power can be provided remotely. But some individuals will still struggle to get permission to attend for funding reasons – it will help them to justify their participation if the event includes (optional) keynote talks and training. Others may be unable to commit to full days or an entire week in front of a computer – they may be helped by an event split into a series of half-days, around school times (for those with caring responsibilities) or other work.

Finally, it is vital that the event is a safe environment for all participants, even in a virtual space. The event's code of conduct should be clearly signposted from the outset, with a transparent mechanism for reporting incidents, including a (monitored) reporting form. People will have different levels of comfort with participating openly in a group, so whilst they can be encouraged to switch their cameras on, they should not be forced to do so and sufficient warning should be given before group photos (or screenshots). Participation such as speaking and the giving of presentations can be encouraged from participants with less experience, but it should not be forced.

Methods

Organising an event

Of all the initial steps (Figure 3), the first question to answer is what type of hackathon to run. There is a spectrum of event types, from a completely pre-determined set of challenges (in the case of a data competition such as a Data Study Group, 2018) to a fully-flexible workshop with no defined goals (often with community events such as Connecting Bristol, 2019). Most hackathons will lie somewhere in between, with some curated challenges that could change slightly during the course of the event, as was the case with the Climate Data Challenge hackathon series. Hackathons can

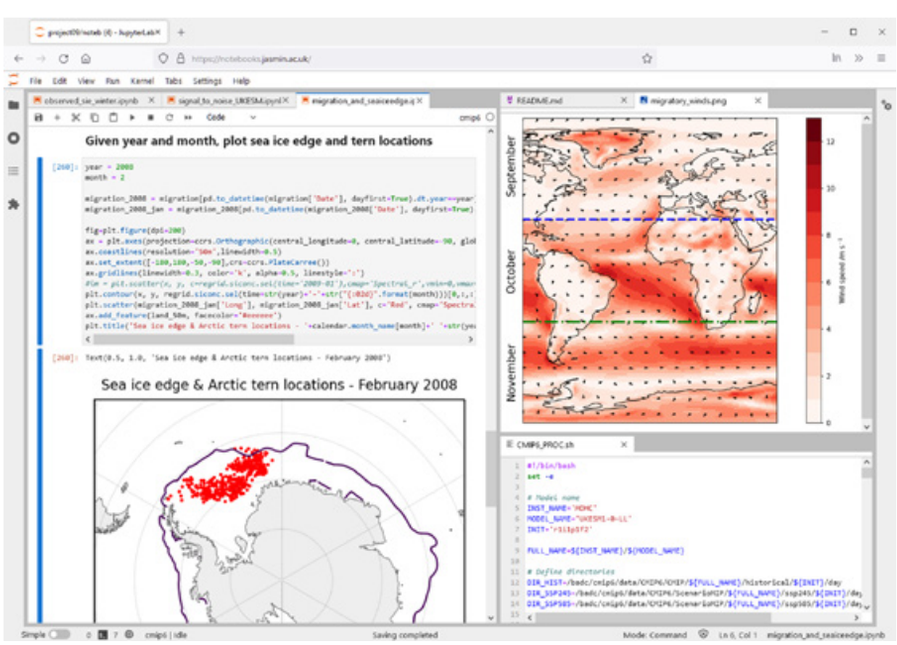

Figure 2. Screenshot of the JASMIN Notebook Service interface, showing the work of one cross-disciplinary group in the CMIP6 Data Hackathon (Mitchell et al., 2022) investigating the impacts of changing wind regimes and sea ice on the world's longest migrant (namely, the Arctic tern). The service uses JupyterHub, which runs Jupyter Notebooks on a shared computation platform, accessible to participants via a web browser. The text, code and figures were subsequently shared on the hackathon's GitHub repository.

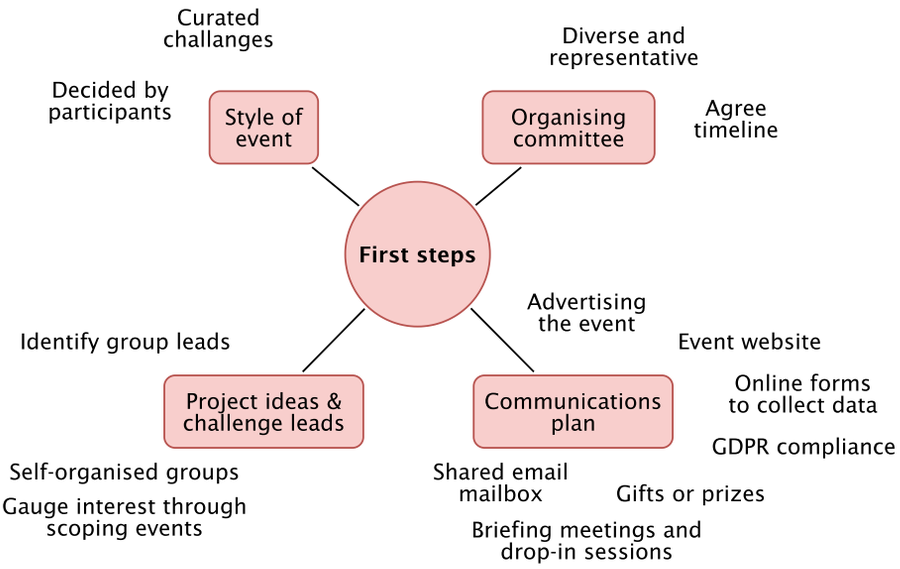

Figure 3. Common first steps taken by the organisers of the Climate Data Challenge hackathon series, held by the Met Office's academic partner universities in the spring and summer of 2021 (Met Office, 2021b).

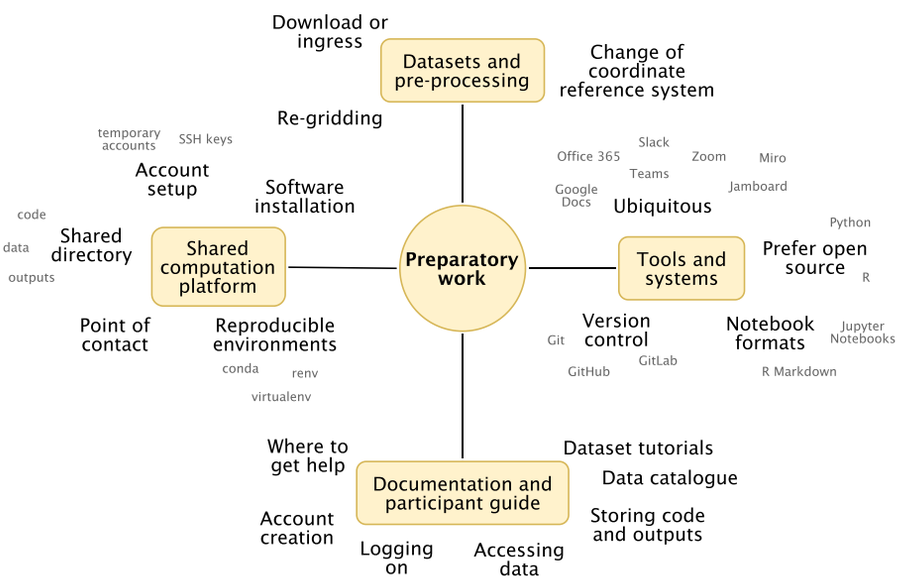

Figure 4. Preparatory work undertaken by organisers of the Climate Data Challenge hackathon series. The majority of the organisers made use of JASMIN, the UK's data analysis facility for environmental science (Lawrence et al., 2013), as a shared computation platform.

also be completely cooperative or contain a competitive element. These early decisions will influence the later organisation, including any pre-events that are run.

Gathering together an organising committee will likely be the next task. They should create a plan or timeline, with tasks assigned to the various members of the committee. If the event relates to a particular community or part of society, then it is strongly recommended that representatives from that community are included on the committee and that they play an active part of the event. The more diverse the group of organisers, the more likely it is that the event will be welcoming, enjoyable and worthwhile.

The committee will need enough time to prepare for the hackathon. The process (Figures 3–5) is likely to take considerably longer than expected, and for this purpose, some institutions provide a dedicated team of people that can help run events, or have connections with external events contractors (although either of these are likely to require funding). For instance, the Jean Golding Institute and Cabot Institute provided invaluable support throughout the CMIP6 Data Hackathon. If challenge leads are recruited as part of a curated event, then they should also understand the time commitment.

Communicating with participants

It will be necessary to exchange lots of information with participants in advance of the event (Figure 3), indeed, more for a virtual hackathon than a physical one. But too many communications can easily overwhelm people. A communications plan can help ensure that information is sent out in blocks. An event website (Thomas, 2021) can provide a central place for participants to find important information, linking to online forms and tools for the collection of registration data. One or more briefing meetings will likely be required with participants, and depending on the format of the event, these could be used to brainstorm projects. Gifts for participants or prizes for teams may encourage participation, but these will involve additional logistics and environmental impact.

Using familiar tools

There is a wide variety of tools and systems aimed at running an online event. Given enough time, it is possible to achieve a professional polished experience, however we suggest a pragmatic approach – using tools that are familiar to both organiser and participants, or ubiquitous in the field (Figure 4). The use of unusual Tool X or Software Y should be avoided unless they have a vital feature that cannot be easily replicated. Focusing on tools that allow participants to share information in different formats in real-time, with minimal setup, will mean they can spend more time getting to grips with the problem and less time becoming acquainted with the event system.

A shared computation platform

Using a shared platform (Figure 4), such as JASMIN, the UK's data analysis facility for environmental science (Lawrence *et al.*, 2013), ensures every participant has access to the same data, functioning set of computation tools and the resources to use those tools. On the other hand, these platforms can be complex and the procedures for setting up an account and connecting can be rather involved. To prevent this from becoming a barrier to participation, many platforms also have temporary accounts that participants can use for the duration of an event (on the JASMIN facility, these are called training accounts). They remove much of the account setup work and often enable participants to be non-academics or from an international audience, in situations where requiring a regular account on the platform could prohibit them from taking part.

Set-up time can be reduced by creating a reproducible environment with specified versions of software libraries pre-installed. Some minimal shell scripts can be used to help participants use the environment and access various data sources. The sharing of code, data and outputs is made easier if

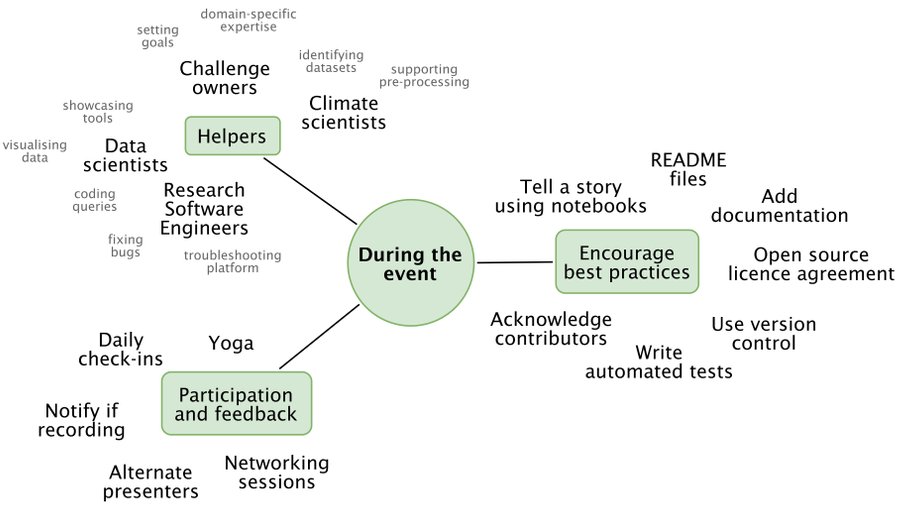

Figure 5. Ways in which the organisers of the Climate Data Challenge hackathon series provided support to the participants during their events. Participants were exposed to best practice tools and began to develop transferable skills, even if they did not have time to fully explore them during the event.

there is a shared directory that all participants can access (on JASMIN, this is a Group Workspace). This may also be a good place for storing the reproducible environment. It is a good idea to check with the administrators of a platform as to whether a shared directory is suitable for its intended use. Many systems have separate filesystems best-suited to either data or software, each with different access permissions and storage quotas.

Datasets in climate science (e.g. CMIP6 [Eyring *et al*., 2016]) can be vast. If a platform provides direct access to a dataset (such as via the CEDA Archive on JASMIN), then it can be helpful to link to a data catalogue and provide some simplified instructions for accessing datasets relevant to the event. It is helpful to have a mechanism for groups to request access to additional data, but in many areas of climate data science, these additional datasets will impact storage quotas.

Finally, not everything will run smoothly. A direct link to the support team that run the platform in case of emergency during the event will be hugely beneficial, as going through their regular ticketing system could cost valuable time. At a physical event, they might have a representative present, but for a virtual event, a private instant messaging channel could be used instead. This allows event helpers to triage problems and only use the support team if truly needed.

Providing enough support

We found that it is beneficial to provide two types of 'helper' at an event: climate scientists with domain-specific expertise in the datasets being used and data scientists or research software engineers with expertise in the platform and software engineering in general (Figure 5). Our experience has been that this works best if these helpers are not tied to a specific group, but can roam between them.

Depending on the field of the challenge owners, they may appreciate support from a climate scientist prior to the event, to identify the most relevant datasets and create an initial set of tasks for participants. They may also need support in pre-processing of data or re-gridding it to a different coordinate reference system. Having climate science support is not a substitute for the challenge owner attending the event themselves, however.

Independent data science support can be used to create and test documentation and resources before the event, solve programming issues during and promote best practices along the way. Learning about modern tools can be of benefit to both ECRs and well-established researchers alike.

Encouraging best practices

Organisers will naturally want the outputs of the hackathon to have longevity and to contribute to their community. However, poorly-organised, undocumented and confusing code is likely to be of little use to future research projects and could be a source of errors. There are many ways to document the outputs of a hackathon, such as encouraging participants to 'tell a story' using code notebooks with narrative text and figures in between code blocks. It can help to make the expected outputs of the hackathon clear by providing reminders of these in template README files for each group (Figure 5).

Encouraging participation and feedback

It is common in hackathons to ask for regular updates or 'check-ins' from groups, usually at the end of each day. These provide a convenient close to each day's efforts, allow participants to give feedback to each other and allow organisers to judge the progress being made by the groups (identifying who might need extra support). These should be short and need not use slides. Questions and suggestions can be handled through instant messaging, to save time. It is tempting to get group leads or challenge owners to present at these check-ins, but this could be a missed opportunity. Encouraging participants to rotate in this role gives them experience in presenting and can help make the event feel more inclusive (Figure 5).

At the end of the event, longer presentations using slides or live-demos can showcase the output of each group. Keeping each group to time is a challenge that is amplified at a virtual hackathon – several websites offer shared timers that can be helpful. It is important to make clear if presentations are recorded and to agree how and when these videos will be shared.

Although participants may raise comments and suggestions at any time during the event, it is customary to collect more formal feedback at the end. Online survey forms work well, however response rates are usually low. Return rates can be improved by scheduling a 5-min break during the closing presentations, where people can stop and fill the form in before the event has ended.

Plenary sessions

Talks and activities are a good way to break up a hackathon timetable and allow participants to take a break. Talks can range from keynotes from subject specialists to demonstrations or training for new tools, practices and systems. Some of the most valued talks take the form of short 'top tips' that participants can use both in the hackathon and their own research – data scientists will likely be willing to help with these.

Providing regular breaks is vital in a virtual event. It prevents people feeling tied to their computer and is part of an organiser's duty of care. In addition to scheduled breaks, physical activities such as desk-based yoga or meditation are generally popular. Ending a day with a networking session and group photo can also promote a collegiate atmosphere.

Next steps for organising an event

Each institution will have its own resources for running hackathons. They may have a team that can help with organisation, data scientists and research software engineers that can provide support beforehand and on the day, or even written resources from previous events. The websites from previous events can also be useful to get ideas of

what might work well and they often contain valuable resources that can be reused. We have also gathered a number of recommended resources and templates for running a climate science hackathon which are available on GitHub[2].

Acknowledgements

The 2021 Climate Data Challenge hackathon series, including the events hosted by Met Office Academic Partnership (MOAP) universities, was supported by the Met Office. We thank the JASMIN data analysis facility and team for providing support, computing services and access to the CMIP6 and other datasets via the CEDA Archive. We are also grateful for data science and administrative support provided by the Jean Golding and Cabot Institutes, University of Bristol.

The authors have no conflicts of interest to declare.

References

Brierley C, **Barnes C**, The UKCP18 Hackathon Organising Committee. 2021. *Hacking the climate*. http://github-pages.ucl.ac.uk/hacking-the-climate/ [accessed 23 November 2021].

Briscoe G, **Mulligan C.** 2014. *Digital innovation: The hackathon phenomenon*. Working Paper 6. CreativeWorks London, p. 13.

Connecting Bristol. 2019. Data and community hackathon – homelessness, housing and digital inclusion. 9 December. https://www.connectingbristol.org/data-and-community-homelessness-housing-and-digital-inclusion/ [accessed 2 February 2022].

Crook J, **Marsham JH**, **Rory F** ***et al***. 2022. The Leeds Africa Climate Hackathon – experiences of running a hackathon and highlights of results. *Weather*. (Submitted).

Data Study Group team. ***et al***. 2018. *CodeCheck: how do our food choices affect climate change?* Project report. The Alan Turing Institute: London, UK. https://doi.org/10.5281/zenodo.1415344.

Eyring V, **Bony S**, **Meehl GA** ***et al***. 2016. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.* **9**(5): 1937–1958.

Fallon JC, **Bloomfield HC**, **Brayshaw DJ** ***et al***. 2022. Understanding climate risk in future energy systems: an energy-climate data hackathon. *Bull. Am. Meteorol. Soc*. https://doi.org/10.1175/BAMS-D-21-0305.1.

Fung F. 2021. Cutting edge collaborative research – using climate data to advance understanding. Cabot Institute for the Environment blog, 14 June. https://cabot-institute.blogspot.com/2021/06/cutting-edge-collaborative-research.html/ [accessed 15 June 2021].

Harrison M. 2021a. Climate challenges for financial services – a data-driven viewpoint. 22 June. https://www.linkedin.com/pulse/climate-challenges-financial-services-data-driven-mark-harrison [accessed 23 November 2021].

Harrison M. 2021b. Climate risk in future energy systems. 2 July. https://www.linkedin.com/pulse/climate-risk-future-energy-systems-mark-harrison [accessed 23 November 2021].

Lawrence BN, **Bennett VL**, **Churchill J** ***et al***. 2013. Storing and manipulating environmental big data with JASMIN. *IEEE International Conference on Big Data 2013*. Silicon Valley, CA. pp 68–75.

Marsham J, The Leeds Met Office Hackathon Organising Committee. 2021. *Engaging narratives of future weather in Africa under climate change*. https://leedsmetofficehackathon.leeds.ac.uk/ [accessed 23 November 2021].

Met Office. 2021a. Climate data challenge – working in partnership for innovation. 2 April. https://blog.metoffice.gov.uk/2021/04/02/climate-data-challenge-working-in-partnership-for-innovation/ [accessed 23 November 2021].

Met Office. 2021b. Climate data challenge hackathon series. https://www.metoffice.gov.uk/weather/climate/cop/climate-data-challenge [accessed 23 November 2021].

Mitchell DM, **Stone EJ**, **Andrews OD** ***et al***. (2022). The Bristol CMIP6 Data Hackathon. *Weather*. https://doi.org/10.1002/wea.4161.

OpenBSD. 2021. *OpenBSD: Hackathons*. https://www.openbsd.org/hackathons.html [accessed 27 September 2021].

Schnase JL, **Lee TJ**, **Mattmann CA** ***et al***. 2016. Big data challenges in climate science: improving the next-generation cyberinfrastructure. *IEEE Geosci. Remote Sens. Mag.* **4**(3): 10–22.

Thomas J. 2021. *Hackathon website template*. Jean Golding Institute. https://github.com/Jean-Golding-Institute/hackathon-template [accessed 24 November 2021].

Thomas J, The CMIP6 Data Hackathon Organising Committee. 2021. *CMIP6 Data Hackathon*. https://cmip6moap.github.io/ [accessed 27 September 2021].

World Climate Research Programme. 2020. *Overview CMIP6 experimental design and organization*. https://www.wcrp-climate.org/wgcm-cmip/wgcm-cmip6 [accessed 2 February 2022].

[2]https://jean-golding-institute.github.io/hackathon-template/.

Correspondence to: J. Thomas

j.thomas@bristol.ac.uk

doi: 10.1002/wea.4199