relevant libraries should be introduced such as handling arrays, data frames, and plotting. A student is likely to address these issues as a postgraduate, but formally stressing these concepts from the outset should prevent inefficiency and promote reproducibility. Since this module includes coding it should come after 'programming'.

iii. Introduction to visualising environmental data

Visualisation of data has a number of functions, from exploratory plots to assist analysis, to charts and diagrams for papers or to influence or communicate with end-users. One plot does not suit all and this needs to be considered along with caution for contouring, colour schemes, etc. to ensure that data are not distorted by the plotting. Effective visualisation of large datasets requires coding and familiarity with the underlying data, so should come after 'programming' and 'environmental data'. Effective visualisations are important for communicating with other users of research, hence it is a core skill to be able to produce well laid out plots as a faithful representation of the data, and this even includes tables of data.

iv. Data management

The effective management of data generated by oneself and the use of data generated by others is a key component of research reproducibility and whose importance can be underestimated by domain scientists. This module can be separated into two sections, firstly the context and secondly the practicalities. The context of data management includes the description of the research data life-cycle, publisher or funder requirements, international initiatives on open data, the basics of metadata, ethics, and the increasing importance of persistent identifiers. The practicalities include producing data management plans, organising data, using metadata, applying persistent identifiers, citing, security, data protection, sharing, licencing, and more. The data management module would be useful at any stage during early research training. A great deal of data management training materials exist online, it is motivating researchers to engage with the training and put the ideas into practice that is the greater challenge.

v. Interdisciplinary data exchange

The ability to more easily share data is a major challenge for interdisciplinary data-intensive global environmental change research. There are at least two underlying issues to address, firstly, researchers need the confidence to openly share data and models (e.g., to know what limitations and uncertainties another domain's data has), and secondly, that specific words have different meanings to different people and different scientific communities. The interdisciplinary module is aimed at bringing different disciplines together in a workshop to examine relevant issues and, in a peer-to-peer way, discuss practical solutions and real examples. This training workshop could be held at any stage, but it may be more effective learning slightly later in research training when each attendee has a solid grounding in their own data and research, or for any early or mid-career.

OPTIONAL

The optional modules explore methods for improving efficiency and robustness of data-intensive research and/or expand the techniques of early and mid-career researchers.

vi. **Software development ideas for scientific coding**

The rigour of software development has concepts that are of benefit to researcher programmers, specifically for reproducibility and reliability. This module builds on the structure of the first core module and is agnostic to the programming language. Diagramming data flows, testing and error handling are all topics that a data-intensive researcher could employ to make sharing their code, and reusing their own code at a later date, more feasible. Whilst many data-intensive researchers may find their work too iterative to align with the rigour of software development, they are useful concepts to know and important to use if code is operationalised. It would be advantageous for PhD students to appreciate these fundamental skills, but they are generally more applicable to learners who know they are going to be immersed in coding.

vii. **Object-orientated programming**

Individual data-intensive researchers may need to employ object-orientated programming, but for those who do not, an introduction to the techniques would allow for professional development and would present alternative solutions with which to experiment.

viii. **Introductory data science topics**

A range of training suitable for mid-career researchers would be valuable professional development. These have been grouped under ‘data science’ but in reality are current topics that could apply to global environmental change research or topics for researchers with autonomy (e.g., to decide on their own compute, storage, and databases). Machine learning, data mining and neural networks are widely used in other scientific domains, but have had less take-up in environmental science while offering interesting solutions, conceptual challenges, and produce new avenues of investigation.

ix. **Data organisation**

Beyond data organisation for data management, tools exist to track workflow and these may be of use to researchers to instil through more automated processes data and work controls.

PRINCIPAL INVESTIGATOR

x. **Data management plans and data repositories**

Anecdotal evidence continues to suggest that project leaders may not see data management as a priority. This briefing would cover high level issues such as expectations of funders and others, key aspects of data management plans, and reinforce identifying and allocating team roles for effective data management. The briefing would encourage Principal Investigators to retain overall oversight of data management as much as they would the scientific rigour in a project.

xi. **Overview of skills for data intensive research**

Within research departments there is an increase in Digital or Hybrid Scientists who write libraries and plotting routines, maintain systems, and perform data transformations. These people currently mostly lie outside of the routine metrics of credit for academic researchers and, therefore, may have stifled career pathways and may struggle to move between institutions. This briefing would highlight the issue and encourage debate, including presenting ways to address imbalances.

4. Delivery and format

The curricula workshop and other discussions have produced a number of recommendations on effective ways to engage domain scientists and ensure maximum relevance. In delivery of the curricula, emphasis should be on practical sessions with real-life examples, reproducibility of research and interdisciplinary work highlighting data sharing. Further recommendations follow:

Format

- Use real-life examples and, if possible, set aside sessions to deal with an individual's real research data. Design training acknowledging what a researcher needs.
- Encourage attendees to bring their own devices, if possible, with the intention that software and examples installed will work after the training. Dissemination is possible in other ways and the fewer barriers to continued experimentation the better.
- Training needs to be used soon after delivery or it will be forgotten. Good online resources for post-course reference, using a follow-up webinar presentation, and/or making a trainer available on a specified day for email questions or calls or other troubleshooting methods are ways to make the training survive into the workplace.
- Emphasise the tangible benefits of the efficiencies (e.g., in programming, data management), collaboration between countries and interdisciplinary data sharing.
- Online delivery is a good way to complement face-to-face training, for example, bringing all attendees to a known common level before starting a course. Online has the asynchronous advantage, addressing time-zone issues and availability, with discussion boards then useful to promote peer-to-peer learning.
- A helpful format for snippets of online training is to host alongside data repository access. For example, UK Data Service provide short videos next to the data access.
- Teaching of a curricula need not be on consecutive days and breaks between learning is an advantageous way to include further learning and project work.
- Assessment is a positive way to ensure that a module has been understood (outcomes-based) and certificates and certification should only be provided if there is, at the very least, an assurance of complete attendance.

Audience

- Exploiting conference training slots is a good way to attract a diverse group of domain scientists for short taster training sessions (one day or less).
- The use of ingress, kick-off meetings, and mid-term meetings of funded projects is a way to gain a wider audience than by self-selection. Separate training activities could be aimed at potential future applicants for upskilling and connecting with the widest community.
- The team approach: training should not be compulsory for everyone on a funded project, but selection of individuals for particular upskilling could be encouraged, for example, via nomination by the Principal Investigators.
- Provision of training aligned with the curriculum serves as messaging from the funding agencies that these skills are necessary and valued.

- Training could be linked around particular research challenges, including using data from other domains, such as hackathons.
- Some funded projects have held summer schools (e.g., GOTHAM) and it would be beneficial to include elements of the curricula as part of the school.

The Software and Data Carpentry approach to training delivery is to take domain experts and provide them with 'train the trainer' sessions to ensure that content is conveyed in an interactive way. The above recommendations should be viewed similarly as guidelines for curricula delivery.

5. Mid-career skills gaps

A remit of the AT4 work was to include topics relevant for augmenting the skills of mid-career researchers. Whilst it is indicated that the core skills are for postgraduate students, mid-career researchers may not be proficient in all aspects of the five core modules of the curricula because they have never attained proficiency in the underlying skills or because they have not kept up-to-date with developments. Therefore, the core skills may be relevant to mid-career researchers and awareness of the curricula may motivate some to upskill, but this group may not recognise this need in themselves. As careers progress it is more difficult to assess necessary skills, but if mid-career researchers are practitioners of data intensive activities then they should be proficient in all core topics.

Proficiency in the optional topics, such as object- orientated programming, machine learning or e-infrastructure, may be vital to a mid-career researcher (or even PhD students) in certain situations. The optional topics are included as curricula as they would be useful introductions and awareness raising to widen the skills and knowledge at their disposal and provide new options for examining and processing data. The assertion that no one has to be expert in everything also applies and many projects will call on a range of skills from several individuals.

Encouraging mid-career researchers to explore and be trained in new methods and technologies is challenging. The obstructions include, the time required, costs, and inertia (both personal and institutional). The skills gap report included ways to engage and encourage mid-career researchers, the two most popular being '*making recognition of digital skills part of career progression*' and '*providing full financial support for training e.g., including travel*'.

The Principal Investigator briefings are targeted at mid-career researchers. The briefings provide the essentials of strategic topics and, given their importance, relative brevity, and potential to be disseminated online, could be made compulsory for Belmont Forum-funded projects.

6. Belmont Certification

An early consideration of AT4 in 2017 was whether to establish 'Belmont Certification'. The practicalities necessary for resourcing and administering such certification requires examinations of the options and resourcing decisions by the Belmont Forum or individual partner agencies.

These issues were raised in the skills gap report, but there are attractions of certification that warrants further exploration of the topic here.

At the curriculum workshop there was discussion on the difference between accreditation, certification, and certificates. Certificates can be awarded for each course completed, which is consistent with the notion of assessment either purely for attendance or with some level of achievement via testing. Certificates would be most appropriately awarded by the training provider. It is possible that a training provider could approach the Belmont Forum to use its logo and, therefore, tacit approval of a course.

Accreditation is a role only achievable by a recognised awarding body, such as a university or learned society, so the Belmont Forum or individual agencies could partner to provide accreditation. Following discussions with an accrediting body, it would seem most likely that accreditation would be provided for the curricula as a whole, all 'core skills', and the accreditor would potentially adapt curricula to ensure correspondence to their own values or requirements.

Certification, as used here, is assumed to be an expression of endorsement by the Belmont Forum, recognition that training activities could be shown to align with the curricula and training delivery can be compared against the format recommendations above. An extension of certification is to certify an individual who can prove that they are competent in all the core curricula. All of these activities require resourcing and for agencies to be comfortable with the actions of external groups.

Advantages of certification raised at the curriculum workshop and elsewhere include:

- Ability to point students to a Belmont Forum-stamped course to be able to identify preferred activities and courses.
- Recognition of relevant training by Belmont Forum would be useful if it helps build a course reputation.
- Some cultures and countries place great importance on certification from respected organisations, so provision of certification may be more valuable than at first glance.
- Endorsement [of courses] by the Belmont Forum could be based on a tick-box approach based on whether a course had appropriate content and delivery.
- Benefit of using the core skills as comparable to a driving license for data intensive environmental scientists.
- Belmont Forum researchers could undertake a self-assessment questionnaire to determine which skills they already have and which areas need improvement. The curricula could be 'badges', so a researcher could determine the badges they require to achieve Belmont Certification.

Disadvantages and issues with certification raised at the curriculum workshop and elsewhere include:

- Formal certification is more appropriate for repositories themselves or experts managing the data than for researchers.
- Keeping a course current is more important than certification.

- Once someone has got to the level of earning a PhD they are less interested in gaining professional qualifications; they direct their own research and similarly their learning.

The original motivation for 'Belmont Certification' implied the approval of individual courses that were aligned with the curricula, but the necessary scrutiny and auditing is an overhead which is difficult to address. A practical and passive solution would be to promote the curricula as an overarching mould without specifics on course recommendations or attainments. This promotion would begin to raise awareness amongst researchers. The concept of the driving license or badges is an appealing way to view and formalise the core skills, and some of the assessment could be automated once devised, but the scheme would require administration.

7. Existing training

As part of the skills gap analysis report, a list of existing courses and resources (section 8) was developed from suggestions of survey contributors and other direct e-I&DM contacts.

Of the five core curricula, some have extensive and well-organised existing resources and others are more limited. Programming and introduction to visualisation have some existing courses such as the Data Science Boot Camp and Software and Data Carpentry, as well as these topics being covered in a range of summer schools. Data management is covered by a number of open source materials from well-established groups including MANTRA and DataONE. Data Carpentry offers workshops on ecology, genomics, geospatial data, and biology, but does not have a tailored offering for data intensive environmental topics. The geospatial data workshop is taught in R, although teaching it in Python would be more valuable for global environmental change researchers. The interdisciplinary data exchange is partially covered by some of the data management materials, but the face-to-face mixed classes allowing for discussions and collaborative work is the strength of this, and is an area currently under-trained.

Many funding agencies will already have existing short training courses that could be openly shared between partners, allowing maximum exposure of the materials and upskilling. These training activities are lower profile than, for example, open online courses, so effort would be required to catalogue and investigate inter-country attendance.

The conclusions of comparing existing training with the core curricula are:

- The core curricula that would benefit most from prioritising development of new courses are 'Environmental data: expectations and limitations' and 'Interdisciplinary data exchange'.
- A great deal of open source material already exists on data management and, therefore, it is encouraging researchers to appreciate its importance, learn about it, and action best practice, which is more important than developing new courses.
- Where short courses exist that align with the curricula, these could be made openly available to partner funding agencies.

8. Conclusions and recommendations

This report provides recommendations for curricula to address the digital skills gaps for data intensive global change research (section 3). Alongside the curricula are a number of specific recommendations on delivery and format (section 4) emphasising practical sessions with real-life examples, reproducibility of research and interdisciplinary work highlighting data sharing.

The curricula themselves have sufficient detail to allow a training provider to determine if their course fits, without being overly prescriptive nor including too much specificity on formats or software that may make the curricula look dated quickly. The curricula are divided into 'core', 'optional' and 'Principal Investigator'. The core skills are a contemporary view of how to handle, understand, and effectively share large environmental datasets, with an emphasis on reproducibility. The optional skills would introduce more specialised activities, providing relevant topics for those wishing to expand their skills into newer data intensive domains. The optional skills are relevant to mid-career researchers wishing to grow and update. The Principal Investigator briefings are targeted at mid-career researchers. The briefings provide the essentials of strategic topics and, given their importance, relative brevity,!and potential to be disseminated online, could be made compulsory for funded projects.

Finding ways to upskill mid-career researchers remains a challenge and relies on self-motivation. The delivery and format recommendations (section 4) are likely to assist in maximising engagement with training by emphasising relevance and using real-life data. The acknowledgement that not all mid-career researchers would be proficient in the core skills is an opportunity to use the curricula as an informal benchmark for researchers, encouraging some to pursue training. The inclusion of the optional modules is to introduce alternative, yet well-established methods for dealing with data intensive issues. The skills gap report highlighted ways to engage mid-career researchers, the top two were '*making recognition of digital skills part of career progression*' and '*providing full financial support for training e.g., including travel*'.

Belmont Certification remains an initiative which has advantages, but requires resourcing. The least-cost option consistent with certification is to promote the curricula to raise awareness of the importance of these key topics amongst the global environmental change research community.

The work carried out on cataloguing existing training has shown that a great deal of open source material already exists on data management. Since these courses are available for researchers to appreciate its importance, learn about it, and action best practice, encouraging engagement and uptake is more important than developing new courses. Conversely, very little training appears to be available on the core modules of 'Environmental data: expectations and limitations' and 'Interdisciplinary data exchange', and these courses would benefit most from prioritising development of new courses. It seems likely that the environmental data materials would be contained in masters' modules and could be repurposed into short courses. The interdisciplinary module is less likely to exist with simultaneous strong data focus, and to develop this course would provide a worthwhile and novel contribution to global environmental change research.

The options for the role of partner agencies in upskilling researchers are:

i. Endorsement of the curricula. Clear and active communication by funding agencies that the curricula are vital skills and a valuable reference for providing training activities.
ii. Explore sharing existing short courses, especially that align with the curricula, between agencies.
iii. Consider funding priority training activities identified by the curricula. Priority core topics are 'Environmental data: expectations and limitations' and 'Interdisciplinary data exchange'. The Principal Investigator briefings are also a priority.
iv. Provide opportunities or directing funded groups to access training, ensuring these resources are widely and openly available.
v. Decide whether the advantages of 'Belmont Certification' are sufficient to resource a scheme.
vi. Consider funding training activities based around curricula as part of the process of implementing Collaborative Research Actions.
vii. Consider targeting a funding call on the re-use of data, especially incorporating data outside of one's domain.

Actions for 2018-2019

Of the list of recommendations, the first three are the priority. The corresponding goals and activities align with the recommendations and would be appropriate activities to further the curricula work presented in this report.

i. Endorsement: disseminate, communicate and influence. Engage with the community to promote upskilling aligned with the curricula by writing blogs/articles, webinars, meeting with organisations, highlighting in funding call text/meetings and presenting at relevant conferences.
ii. Convene a party of funding agencies to create a working group on sharing of existing short courses, making attendance open to partners, with a pilot scheme to be developed.
iii. Work with the Belmont Forum to develop priority training activities, which may be in partnership with other organisations. The training delivery can be within funding activities, such as meetings, or otherwise.
iv. Review curricula and track emerging issues to remain current e.g. the need for publishing research software.
v. Continue to liaise with CODATA-RDA summer school, DCC, ESIP, DataONE, UNSD and others.

9. Acknowledgements

This work has been funded by the Natural Environment Research Council of the UK and prepared for the e-Infrastructure and Data Management project of the Belmont Forum. I would like to thank all the participants of the curricula workshop held in Vienna April 2017 for their engagement and professionalism, and especially Rowena Davis of the e-IDM in the preparation of this report and her insight into prioritising inter- and trans-disciplinary data.

Vicky Lucas 14 September 2017
Human Dimensions Champion, e-I&DM and
Training Manager, Institute for Environmental Analytics, v.lucas@the-iea.org

APPENDIX A – Details of curricula

A = first year PhD; B = final year PhD; C = postdoc, D = mid-career, Principal Investigators.

CORE

i. **Programming for data intensive research (for those who already code)** (A) – 4 DAYS
   - Unix/Linux, shells, syntax and using the command line
   - Coding standards and style guides e.g. PEP 8
   - Version control using GIT, individual and team
   - Commenting and documenting code
   - Modularising code
   - Domain specific data formats and libraries
   - Debugging and basic error handling
   - Introduction to testing and design
   - Open source licensing and code
   - Importance of metadata
   - Optional mention of parallelising code
   - Programming language agnostic, but most likely taught in Python
   - Introduction to Notebooks, e.g. Jupyter, for sharing code and documents

ii. **Environmental data: expectations and limitations** (A) – 3 DAYS
   - Uncertainties in environmental measurements
   - Instrumentation examples, calibration, precision
   - Spatial data - issues for gridded data and projections
   - Cleaning data and dealing effectively with missing or corrupt data
   - Combining datasets from multiple sources e.g. points in time and space vs averages in time and space
   - Using representative data – assessing suitability of other sources of data e.g. from a different location or using modelled instead of measured data
   - Using numerical model outputs e.g. climate simulations
   - Getting data into a usable format; documenting data workflows
   - Using historical data e.g. transcribed hand written records and scanned documents
   - ISO 8000 on data quality and mention of other relevant standards
   - Metadata, provenance and documentation

iii. **Introduction to visualising environmental data** (A and B) – 2 DAYS
   - Data presentation and labelling - tables and plots
   - Plotting data for analysis
   - Data visualisation for papers and presentations
   - Scripts for reproducible figures

iv. **Data management** (A and B) – 2 DAYS
   - Research data, collected by you and collected by others, formats
   - Data management plans

- Journal and funder data policies
- Reproducibility and organising data: file names, README, structures and versions
- Metadata for describing, finding and making data reusable
- Citing and publishing data
- Data security including documentation, backing up, checksum and wider issues
- Best practice for data including ethics, transparency and data protection
- Persistent identifiers and introduction to research objects
- Data sharing, preservation, licensing and trusted repositories
- Data licensing using machine-readable standards (e.g. Creative Commons licenses)
- The European INSPIRE directive and similar initiatives

v. **Interdisciplinary data exchange** (B, C and D, mixed classes engineers, social and environmental scientists) – 2 DAYS
- Sharing and open data – data standards and metadata, interoperability standards
- Introduction to semantic vocabularies across domains and ontologies e.g. Envo
- Syntactic ways to encode data
- Discussion of uncertainties e.g. surveys, model output
- The use of expert judgment in environmental science e.g. emission pathways
- Practical collaborative project work e.g. hackathon or bring your own data
- Thinking of end-users in project design and throughout the project lifecycle
- Data and software citation and publication
- Reusability and reproducibility across domains

OPTIONAL

vi. **Software development ideas for scientific coding** (C and D) 3 DAYS
- Design methodologies, diagramming and data structures
- Unit and integration testing
- Requirements capture
- Error handling and debugging

vii. **Object orientated programming** (C and D) 3 DAYS
- Object orientated programming: analysis design and implementation
- Design patterns and advanced design methodologies
- Exception handling and exception classes
- Testing strategies, testing classes
- Debugging with classes

viii. **Introductory data science topics** (C and D) 1 DAY each
- Relational and non-relational databases
- Advanced visualisation
- Machine learning
- Data mining, including text mining

- Research computational infrastructure - cloud, HPC etc. (some country dependence)

ix. **Data organisation** (A to D) 1 DAY
- Best practice on research science workflow, intro to Taverna, Kepler or Vendor
- Intermediate to advanced use of Jupyter or other notebooks for sharing

PRINCIPAL INVESTIGATOR

x. Data management plans and data repositories (<0.5 DAYS)
- Expectations of funders, publishers, repositories and other stakeholders
- Data life cycle and the requirements of data management plans
- Legal considerations
- Data security, privacy and sensitive data
- Team organisation for effective data management – designating duties

xi. Overview of skills for data intensive research (1 hour briefing)
- Software and computing skills as a team resource
- The Hybrid or Digital Scientist - cases e.g. from Lesley Wyborn or Bryan Lawrence
- Giving credit to supporting roles e.g. CASRAI CRediT