frontiers in Environmental Science

ORIGINAL RESEARCH
published: 14 August 2019
doi: 10.3389/fenvs.2019.00121

Data Science of the Natural Environment: A Research Roadmap

Gordon S. Blair[1,2]*, *Peter Henrys*[2], *Amber Leeson*[1], *John Watkins*[2], *Emma Eastoe*[1], *Susan Jarvis*[2] *and Paul J. Young*[1,3]

[1] Data Science Institute, Lancaster University, Lancaster, United Kingdom, [2] Centre for Ecology and Hydrology, Lancaster Environment Centre, Lancaster, United Kingdom, [3] Pentland Centre for Sustainability in Business, Lancaster University, Lancaster, United Kingdom

Data science is the science of extracting meaning from potentially complex data. This is a fast moving field, drawing principles and techniques from a number of different disciplinary areas including computer science, statistics and complexity science. Data science is having a profound impact on a number of areas including commerce, health, and smart cities. This paper argues that data science can have an equal if not greater impact in the area of earth and environmental sciences, offering a rich tapestry of new techniques to support both a deeper understanding of the natural environment in all its complexities, as well as the development of well-founded mitigation and adaptation strategies in the face of climate change. The paper argues that data science for the natural environment brings about new challenges for data science, particularly around complexity, spatial and temporal reasoning, and managing uncertainty. The paper also describes a case study in environmental data science which offers up insights into the promise of the area. The paper concludes with a research roadmap highlighting 10 top challenges of environmental data science and also an invitation to become part of an international community working collaboratively on these problems.

Keywords: data science, earth and environmental sciences, complex systems, uncertainty, spatial and temporal reasoning

OPEN ACCESS

Edited by:
Xuan Zhu,
Monash University, Australia

Reviewed by:
Quanxi Shao,
Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia
Yuichi S. Hayakawa,
Hokkaido University, Japan

***Correspondence:**
Gordon S. Blair
g.blair@lancaster.ac.uk

Specialty section:
This article was submitted to Environmental Informatics, a section of the journal Frontiers in Environmental Science

Received: *19 April 2018*
Accepted: *24 July 2019*
Published: *14 August 2019*

Citation:
Blair GS, Henrys P, Leeson A, Watkins J, Eastoe E, Jarvis S and Young PJ (2019) Data Science of the Natural Environment: A Research Roadmap. Front. Environ. Sci. 7:121. doi: 10.3389/fenvs.2019.00121

INTRODUCTION

Data science is emerging as a major new area of study, having significant impacts on areas as diverse as eCommerce and marketing, smart cities, logistics and transport, and health and well-being (Dhar, 2013; Provost and Fawcett, 2013). To date, there has been little work on data science applied to the understanding and management of the natural environment. This is surprising for two reasons. Firstly, studies of the natural environment are increasingly data rich with a pressing need for new techniques to make sense of the accelerating amount of data being captured about environmental facets and processes. Secondly, climate change is such a major challenge and one would anticipate that data science researchers would be drawn toward this area and the many rich data challenges. However, this is not yet happening.

This paper examines the potential of data science for the natural environment. More specifically, the paper has the following main objectives:

1. To define and map out the emerging field of environmental data science;
2. To systematically discuss the major data challenges in environmental science;
3. To draw up a research roadmap, highlighting the most significant areas requiring further research and collaboration.

As an additional objective, the paper aims to draw others to this field to create a worldwide, cross-disciplinary community to progress the field of environmental data science.

The paper draws on experience from a strategic partnership between the Data Science Institute (DSI) at Lancaster University and the Centre of Ecology and Hydrology (CEH) to create a world-leading Centre of Excellence in Environmental Data Science (CEEDS). This partnership builds on excellence in (cross-disciplinary) environmental science in both CEH and Lancaster, the national capability offered by CEH in terms of data sets and modeling capabilities, and the wide range of cross-disciplinary methodological and computational skills that are present in the Data Science Institute.

Note that our scope is deliberately broad in considering all areas of environmental science, including but not limited the geosphere, hydrosphere, biosphere, and atmosphere; our experiences indicate that there are many more commonalities than differences when considering the data science challenges in the various aspects of the natural environment.

The paper is structured as follows. Section Data Science of the Natural Environment looks more closely at the motivation for a data science of the natural environment. Section Challenges then examines the core challenges associated with environmental science arguing that the challenges are both unique and significant. Following this, section Data Science of the Natural Environment: Revisited provides a more refined statement of the nature and scope of environmental data science. Section Case Study: Modeling Extreme Melt Events on the Greenland Ice Sheet presents a case study of environmental data science in practice, highlighting the potential in this area. The paper then concludes with a series of overall observations culminating in a research roadmap—highlighting the top 10 research challenges of environmental data science.

DATA SCIENCE OF THE NATURAL ENVIRONMENT

What Is Data Science?

Data science is the *science of extracting meaning from complex data*, hence supporting *decision-making in an increasingly complex world* (Baesens, 2014). Many commentators use the term "big data" (Mayer-Schonberger and Cukier, 2013; Jagadish et al., 2014; Reed and Dongarra, 2015) as a synonym for data science. We avoid this term as it emphasizes the "big" whereas the real challenges lie in the *complexity* and *heterogeneity* of the underlying data sources—discussed further below.

Researchers agree that data science is an interdisciplinary challenge and a series of data science research institutes have been created, drawing on statistics, computer science, artificial intelligence (AI), social sciences, psychology, economics, health, and so on. These include the Alan Turing Institute in London and Data Science Institutes in Berkeley and Columbia. For some, the emphasis is on algorithmics and computation. We argue that data science research should be *problem-driven* to ensure that algorithmic and computational breakthroughs are targeted toward real-world problems; and that the most significant and transformational breakthroughs will emerge from research where the *disciplinary boundaries become permeable* and a range of researchers work together on problems situated in the real world. Furthermore, researchers working on the problem domain should not just be end users but should be first *class citizens* in the resultant collaborations. This *situated* philosophy is at the heart of data science research at Lancaster.

We argue that the potential for environmental data science is enormous and indeed understanding, and managing the impact of, environmental change is a grand challenge for the emerging subject of data science. Before developing this argument further, we look more closely at the nature of the environmental sciences.

A Focus on the Natural Environment

Understanding of the natural environment is increasingly important as society struggles to respond to the implications of a changing climate and anthropogenic pressures on finite natural resources, and their impacts on water, energy and food security, infrastructure, human health, natural hazards, and biodiversity. This is also a major cross-disciplinary challenge involving, for example, ecologists, hydrologists, soil scientists, biologists, chemists, physicists, and statisticians. With the need to influence policy and derive well-founded adaptation and mitigation strategies, there is also an increasing emphasis on social science and communication of science.

More generally, it is possible to observe a significant shift in this area toward a *"big" science*, which is a science that is more integrative and collaborative. This represents a *cultural shift* away from individual scientists working within their own (siloed) discipline, with the emphasis now on understanding the full complexities of the natural environment in all its facets. The prime example of this is the move toward natural capital and ecosystem services (Helm, 2015; Potschin et al., 2016). Natural capital is concerned with the world's stocks of natural assets, including its soil, water, air, energy sources, and all living entities on the planet. The study of ecosystem services then investigates the sustainable and integrated management of complex ecosystems in the support of the services we need to live, hence reifying the complexity of this management in all its facets including environmental, social, health, and economic considerations (Muller et al., 2010). Future Earth is a further example of a major initiative seeking cross-disciplinary insights (in their case around global sustainability)[1].

The environmental and earth sciences, as with other areas of science, are also increasingly data-driven (Hey et al., 2009). In parallel, there is a move toward more open data, leading to an *open science*, and a science that is more transparent and potentially repeatable and/or reproducible[2].

A Data Science of the Natural Environment?

It is clear that, given the challenges outlined above, earth and environmental sciences should be fully embracing data science

[1] www.futureearth.org

[2] http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf

and should be at the forefront of this initiative. There are pockets of excellent data science work being carried out in the environmental community, and we reference several examples in throughout the paper, but it will become clear that this work needs to be significantly extended. Similarly, data scientists should be reaching out to this community to offer support. In reality, again, this is not happening to the extent it should. Other areas of science are much further on in embracing data science, most notably physics (Philip Chen and Zhang, 2014) and life sciences (Marx, 2013) (including a wealth of work in bioinformatics; Greene et al., 2014; Wang et al., 2015). This leaves a significant semantic gap in: (i) the integration of highly complex data sets, (ii) transforming this underlying data into new knowledge, for example around ecosystem services, and (iii) informing policy around, for example, appropriate mitigation and adaptation strategies in the face of climate change.

In summary, there should be a strong symbiotic relationship between data science and the earth and environmental sciences. Earth and environmental sciences need data science, and data science should be responding to the intellectual challenges associated with complex and heterogeneous data. More profoundly, data science should be woven into the very fabric of earth and environmental sciences as we seek a new kind of science and subsequently intellectual breakthroughs that can transform society. Finally, we note that while data science can have a significant impact on the earth and environmental sciences, we qualify this by stating that it is clearly not a "silver bullet" in terms of understanding and responding to environmental change; it must sit alongside other initiatives in the spheres of politics, economics, and so on.

CHALLENGES

The Data Challenge

Data is central to earth and environmental sciences with significant investments in techniques for managing a wide range of environmental data. The data challenge is quite distinct from many fields of science with the most striking factor being the heterogeneity of the underlying data sources and types of data, hence the inappropriateness of the term "big data" in this field (as discussed in section What Is Data Science? above). More specifically, data science is often annotated using the four "V"s of data: volume, velocity, variety, and veracity (Jagadish et al., 2014). While in many areas of data science, consideration of volume and velocity dominate, in the environment variety and veracity (accuracy/precision) are the most important characteristics. This is not to diminish the first two properties as there are areas where there are very large data sets and where the processing of such data sets can be challenging, e.g., in climate science (Schnase, 2017), but this only helps to exacerbate the issue of variety when considered alongside other data sources. We look at the issues of variety and veracity in more detail below.

Environmental data comes from a wide variety of sources and this is increasingly rapidly with new innovations in data capture:

1. Large volumes of data are collected via *remote sensing* where environmental phenomena are observed without contact with the phenomena, typically from satellite sensing or aircraft-borne sensing devices, including an increasing use of drones. This includes passive sensing, such as photography or infra-red imagery, and active sensing, e.g., RADAR/LIDAR. The increasing availability of open satellite data, in particular, is a major trend in earth and environmental sciences. For example, the EU Copernicus programme and the associated Sentinel missions, or NASA's LandSat archive are regularly mined for data for a variety of applications (e.g., Langley et al., 2016).
2. Other data are collected via *earth monitoring systems*, which consist of a range of sensor technologies more typically in close proximity with the observed phenomena. Such sensors will monitor a range of parameters around the atmosphere, lithosphere, biosphere, hydrosphere, and cryosphere. Examples include weather stations and monitoring systems for water quality. Historically, such sensing technologies would be placed in the field and visited to periodically download data. It is more common now to have telemetered data providing real-time access to such data streams. Developments around the *Internet of Things* (IoT) also have the potential to dramatically increase the level of monitoring in the natural environment through real-time access to dense deployments of a wide variety of sensors (Atzori et al., 2010; Nundloll et al., 2019).
3. Significant quantities of data are collected through *field campaigns* involving manual observation and measurement of a range of environmental phenomena and these are increasingly supplemented by *citizen science* data collected by enthusiasts with strong exemplars in the areas of soils data (e.g., through the use of a mobile application called MySoil; Shelley et al., 2013) and biodiversity (e.g., RSPB's Big Garden Bird Watch; Godard et al., 2010).
4. There are large quantities of *historical records* that are crucial to the field. Many of these are digitized but, equally, significant quantities of potentially important information are not, particularly at a local level. Important examples of historical records in the UK context, for example, include: geological survey data and samples, managed by the British Geological Survey (BGS), and meteorological and ice observation records going back to the 1800s as managed by the British Antarctic Survey (BAS).
5. *Model output* is also a significant generator of environmental data with results from previous model runs often stored for subsequent analysis (see section The Spatial/Temporal Challenge for a more in-depth consideration of modeling).
6. Significantly, there is growing interest (as in many fields) of exploiting *data mining*, discovering data, and data patterns from the web and social media platforms, such as seeking images showing localized water levels during periods of flood (Cervone et al., 2016) or seeking evidence of air quality problems and impacts on human health (Mei et al., 2014). This area is in its infancy but is likely to grow massively over the next few years.

Together, this adds up to the potential for having environmental data at an unprecedented scale, hence providing major opportunities for science but also key challenges. In particular,

it should now be very apparent that *variety* is a crucial and central issue in environmental data. Data is captured from a wide variety of sources about a wide variety of natural phenomena. The underlying data are highly heterogeneous in terms of how the data are stored and exchanged. Some of the data will be structured, others unstructured (structured data is highly organized with the structure captured by a data model or scheme, whereas unstructured is not). The majority of the data will be quantitative but important qualitative data will also be present. Some of the data will be lodged in an environmental data center and decorated with appropriate meta-data to enhance discovery and interpretation. Other data will be held on individual scientists' PCs. Researchers at the University of Chicago's Computation Institute refer to this latter phenomenon as the "long tail of science," whereby vast amounts of data are not available for sharing or community analyses due to lack of resources and tools to make them accessible[3]. The data will also cover different geographical regions and be at different spatial scales, which can impact on integration (and ditto with the temporal dimension)—see also section The Uncertainty Challenge below. Some advances have been made around managing variety, particularly around *interoperability standards* for environmental data, e.g., the Inspire Directive 2007/2/EC, which covers a wide range of sources of spatial data, and also more domain specific proposals such as Water ML from the OGC. In addition, many researchers see *linked data* as a promising technology to capture the complex interrelationships between different data sets in the natural environment (Bizer et al., 2010; Hitzler and Janowicz, 2013). Similarly, a number of initiatives are looking at the *semantic web*, and the development of ontologies as a means of describing and subsequently supporting the integration of disparate data sets (Raskin and Pan, 2005; Compton et al., 2012). Linked data is a "set of best practices for publishing, sharing, and interlinking structured data on the Web [and] its main objective is to liberate data from silos"[4]. Berners-Lee et al. (2001) define the semantic web as "a web of data that can be processed directly and indirectly by computers," and ontologies then have the important role of capturing the meaning of data, including complex relationships across data.

Veracity is also increasingly important, particularly given new developments alluded to above. For example, how reliable is data emanating from citizen science collection methods? (The answer will also vary significantly depending on the level of expertise of the citizen.) Satellite observations may be of lower fidelity when compared to *in situ* observations. Similarly, with the growth of the Internet of Things it is likely that expensive and hence almost certainly more accurate instruments may co-exist with dense deployments of cheaper, less reliable, sensors and hence the provenance of data sources must be both stored and factored into data analyses.

Arguably the major trend is to creatively bring together different data sources in terms of understanding relationships across phenomena and also to constrain uncertainty by linking different observed data readings. To be effective though, the data science issues around the four "V"s need to be addressed, especially around variety, and veracity.

Summary of data challenges

Managing the variety and heterogeneity in underlying sources of data, including achieving interoperability across data sets;

Reducing the long tail of science and making all data open and accessible through environmental data centers;

Ensuring all data are enhanced with appropriate semantic meta-data capturing rich semantic information about the data and inter-relationships;

Ensuring mechanisms are in place to both record and reason about the veracity of data;

Finding appropriate mechanisms and techniques to support integration of different data sets to enhance scientific discovery and constrain uncertainty.

The Modeling Challenge

Modeling is the principal tool for understanding the environment and forecasting or projecting environmental change. Modeling enables us to make sense of the data that emanates from the various observation and monitoring techniques described above and, from this, to make predictions about the future and analyze "what-if" scenarios. Models can usefully be classified as either *process models*, which attempt to capture and/or abstract over the underlying physical processes being considered, or *data-driven models*, which are based on empirical statistical fits to observations or data derived from more complex models. Many environmental process models are often heavily parameterized, where complex phenomena, or phenomena acting at small scales, are captured by semi-empirical methods and approaches.

Model simulations are often combined to form *ensembles* in order to explore uncertainty and sensitivities. Ensembles may consist of a number of single model runs or of a number of different models. Single model ensembles may be used to explore the sensitivity to *initial starting conditions* (e.g., Kay et al., 2015) or to investigate the uncertainties associated with the model structure (e.g., *perturbed physics ensembles*, where parameters are varied within their uncertainty; e.g., Beven and Binley, 1992; Carslaw et al., 2013). Multi-model ensembles often include individual models run with the same input data in order to compare predicted outcomes, such as the global climate model intercomparisons conducted in support of the IPCC process (CMIP5; Taylor et al., 2012). These are referred to as "*ensembles of opportunity*" to emphasize that they are not full explorations of uncertainty.

There is also significant interest in integrated modeling, where multiple environmental and impact models may be combined to address complex real-world problems, particularly problems requiring higher-order systems thinking and holistic solutions (Laniak et al., 2013). Global climate (or Earth system) models can be viewed as integrated modeling systems, where different aspects of the environment (atmosphere, land, ice, oceans etc.) are simulated by individual model components or sub-models. These models often have sophisticated software architectures to

[3]https://voices.uchicago.edu/compinst/blog/unwinding-long-tail-science/

[4]https://www.coar-repositories.org/community/events/archive/repository-observatory-third-edition/coar-talks-ir-cris-interoperability/second-edition-linked-open-data/7-things-you-should-know-about-open-data/

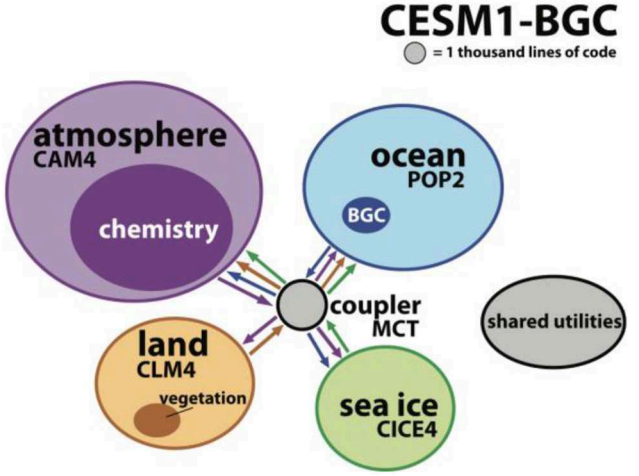

FIGURE 1 | The software architecture of the Community Earth System Model (CESM), taken from Alexander and Easterbrook (2015) and reproduced with the permission of the authors.

manage the associated couplings between the model components (Alexander and Easterbrook, 2015), as illustrated, for example, by the Community Earth System Model (CESM) developed at NCAR, shown in **Figure 1**. As can be seen, CESM is made up of different model components, namely CAM4, POP2, CLM4, and CICE4 (representing the atmosphere, oceans, land, and sea ice resp.), interconnected by a coupler, MCT. Each of these model components is a complex model in its own right, perhaps integrating several other smaller model components (e.g., clouds, chemistry, and radiative transfer in the atmospheric model).

There is also interest in integrated modeling in other areas of earth and environmental sciences ranging from combining a small number of models to investigate inter-relationships between phenomena (Thackeray, 2016), through to complex integrated modeling for an understanding of nature's contribution to people (hence requiring models of economic and social factors, for example; Alvarez et al., 2015; Harrison et al., 2015). This is a very demanding area featuring a number of data science challenges, such as quantifying and propagating uncertainty, and dealing with complex phenomena, interdependent variables, and feedback loops. These integrated modeling systems will often include complex process models coupled with data-driven models, perhaps relying on archived output of process models rather than coupling components online.

One key challenges for integrated environmental modeling is *variety*, with models being developed by different groups for a wide range of environmental phenomena, operating at a wide range of scales and temporal/spatial resolutions, and perhaps with different representations and data types for the same phenomena. At the software level, models are often written in different languages, with Fortran featuring heavily as a legacy language in many process models, and specialized environments such as R, Matlab, Python, or Julia being used for data-driven models. There is some support for integrating models together, including the Earth System Modeling Framework (ESMF) or the OGC standard OpenMI (Open Modeling Interface). However, major interoperability challenges remain, amplified by the level of heterogeneity in the data being exchanged (as discussed in section The Data Challenge above).

One method to promote model integration and shared access is by taking advantage of the cloud and cloud standards, such as Web Services. Such an approach would also enable modelers to take advantage to the elastic storage and computational resources available in the cloud, permitting management of large ensemble simulations and their analysis with cloud machine learning tool kits for instance. However, existing cloud services are not yet suited to supporting the execution of environmental models and ensemble or integrated model runs. For example, popular Platform as a Service (PaaS) offerings such as MapReduce and Apache Spark (Dean and Ghemawat, 2004; Zaharia et al., 2016) have a simple computational model whereby the same computation is carried out on different partitions of large data sets. This is quite different from the requirements of complex models, where different parts of the data might be tightly coupled (e.g., atmospheric and ocean circulation in climate models).

There are also significant issues around the *veracity* of models and, given unknowns in the accuracy and precision of different models, understanding the impacts when they are combined (e.g., Wilby and Dessai, 2010). Models are mostly trained, calibrated, or evaluated against historical data, while recognizing that the past is not necessarily a good indicator of the future, especially given that climate change may take the environment to states outside observations. There has been a long-standing interest in combining models with observations to partially address this problem, a field known as data assimilation (e.g., Lahoz et al., 2010). Historically, most work in this area has been carried out in the context of numerical weather prediction, but data assimilation has also been applied to ecology (Niu et al., 2014), the carbon cycle (Williams et al., 2005), and flood forecasting (Yucel et al., 2015; see also Park and Xu, 2017). More generally, there is huge potential in combining process models with data-driven models to achieve deeper understanding on environmental change and the uncertainties associated with such change.

Summary of modeling challenges

Moving models to the cloud to support open and shared access to a range of environmental models;

Providing interoperability between the full range models, including process and data-driven models;

Supporting the construction of a range of possible ensemble models;

Supporting integrated modeling including potentially highly complex and multi-faceted models for natural capital assessment;

Reasoning about and managing uncertainty in model runs, including in ensembles and across integrated modeling frameworks.

The Complexity Challenge

The earth is a complex system, even more so when considerations of the earth are folded together with economic and social concerns. Dealing with this inherent complexity is a major

challenge for data science. Complexity is itself a major area of study, reflected in the emergence of complexity science as a subject in its own right. Kastens et al. (2009) usefully define a complex system as one that "exhibits the following characteristics:

- feedback loops where change in a variable results in either an amplification (positive feedback) or a dampening (negative feedback) of that change;
- many strongly interconnected variables, with multiple inputs contributing to observed outputs;
- chaotic behavior, i.e., extreme sensitivity to initial conditions, fractal geometry, and self-organized criticality;
- multiple (meta)stable states, where a small change in conditions may precipitate a major change in the system;
- a non-Gaussian distribution of outputs, often where outcomes that are far away from the average are more likely than you might think."

Of the many definitions of complex systems, this really resonates with studies of the earth and the environment. For example, an analysis of feedback loops has been shown to be core to understanding rebound effects in climate change where technological innovations have failed to slow the rate of emission of greenhouse gasses (Greening et al., 2000; Jarvis et al., 2012). Similarly, it is well-understood that in the natural environment everything is interconnected and hence the second bullet strongly aligns with research in this area. As a final example, extreme value theory has emerged to explain phenomena that are far removed from the average and associated with rare events that may otherwise be regarded as outliers. Unsurprisingly, extreme value theory has been applied successfully to environmental science, including in flood prediction (Tawn, 1988).

Dealing with this complexity represents a major challenge for earth and environmental sciences and folding in new methods to deal more explicitly with feedback loops and interconnected variables across spatial scales, would represent a significant breakthrough in many areas of environmental science. Complexity also represents a major challenge for data science with data science also offering interesting perspectives on how to handle complexity, for example the role of machine learning in dealing with and responding to emergent phenomena and in dealing with surprises in complex systems.

Summary of complexity challenges
Managing the complexity of the underlying phenomena, particularly in terms of understanding feedback loops and inter-dependent behaviors, chaotic behaviors, and also extremes;
Developing new data science techniques to deal with and respond to emergent behavior and other complex phenomena.

The Spatial/Temporal Challenge

Studies of the environment are often related to reasoning about natural phenomena across space and time. Estimating spatial or temporal patterns in data and deciphering the effects of covariates across the domain of interest is key to this. Many environmental processes exhibit spatial and/or temporal structure and describing and quantifying such structure is important for enabling robust inference, in terms of patterns and covariate effects, to be drawn. Evaluating such structure within a modeling framework means having to account for second order properties, that is to say the dependence between observations, rather than simple mean effects, which poses many practical modeling challenges. In many instances the dependence structure is captured through a covariance matrix, but in other cases alternative measures may be more appropriate, see for instance Coles et al. (1999) and Davison et al. (2012) for an introduction to tail dependence measures used in extreme value analysis. Inclusion of such second order properties involves additional stochastic processes within the model and standard likelihood based approaches are unsuitable. There has been a large body of work over many years in both time series analyses and spatial statistics where the goal has been to estimate inherent spatial/temporal structure within the observed data and to exploit this for predictive purposes. Traditional examples include ARIMA models, Kriging, Gaussian processes, and spatial point processes. Such approaches have been continuously developed, modified, and improved over many years to overcome limiting assumptions or allow for greater flexibility. Excellent summaries of these approaches are provided in Brockwell et al. (2002) for time series methods and Cressie (1993) and Gelfand et al. (2010) for spatial methods.

Technical developments in several fields have created the opportunity to observe, simulate, and forecast our environment at unprecedented scales of space, time, and complexity. This has led to deluge of spatially and temporally referenced data. Traditional methodological approaches are, however, not well-suited for the era of big data and most scale poorly to handling large spatio-temporal data sets. The computational demands become limiting as the need to handle increasingly large covariance matrices becomes infeasible. There is therefore an increasing challenge to develop approaches that can handle large spatio-temporal data sets and to estimate the fine scale structure within. Heaton et al. (2018) provide a nice summary of the state of the art and comparison of approaches to handle large spatial data, but admit that further research is required for the spatio-temporal setting and for optimizing computational run times. There is therefore a real opportunity here for data science to significantly enhance the state-of-the-art and provide intellectual breakthroughs in capabilities for spatial/temporal reason across large-scale environmental data sets.

A further key challenge is to increase the *spatial and temporal resolution* of predictions. For example, global climate models typically operate at a resolution of 2 degree grids (equivalent to 200 km at the equator) and there is a desire in the community to significantly increase the resolution, for example to 5 km grids or even 1 km grids in the longer term. Similarly, researchers wish to develop air quality predictions at the level of street "canyons" (Reis et al., 2015) offering more localized warnings of health risk and richer mitigation and/or adaptation strategies. Such analyses typically exploit the availability of multiple data sources via statistical downscaling or data fusion approach. However, challenges remain since there are often mismatches in terms

of either, or both of, spatial or temporal scale, and spatial and temporal overlap or representativeness. It is also sometimes the case that different data sources do not measure the same processes. For all of these reasons, it can be very hard to achieve integration across different data sets and models. These are largely unresolved issues in earth and environmental sciences and a key challenge for environmental data science.

Finally, we see that there is great potential in incorporating spatial dependence structures within analytical frameworks that are traditionally focused at single sites and hence evaluating marginal effects. Examples here include extreme value analysis where return levels are typically estimated for each site independently and changepoint analyses where changes are identified on a site-by-site basis. Incorporation of spatial structure into models of this type not only results in more physically realistic models, but also enables a better understanding of the underlying processes and allows sharing of information between sites; the latter is particularly helpful for sites for which there may be little, or no, data. This is a highly active area for research and an area where novel data science insights can provide a significant step forward. Such insights might contribute both to the development of completely novel modeling approaches or, as in the case of extreme value analysis, to the development of a more accessible implementations of existing multivariate and/or spatial methodology.

Summary of spatial/temporal challenges

Providing a range of data science techniques to support sophisticated reasoning across space and time, including areas such as clustering, propagation, and extrapolation, particularly for big data;

Develop data science techniques to achieve the required level of spatial and temporal resolution in scientific studies;

Support the integration of data and models that operate at different spatial and temporal scales;

Support the extension of typically marginal analyses to incorporate spatial and/or temporal structure.

The Uncertainty Challenge

This challenge is arguably the defining one for environmental data science. Uncertainty can emanate from a wide variety of sources, including:

- Uncertainty (or veracity) of the underlying data sources/observations, with this becoming even more significant given the newer sources of data discussed in section The Data Challenge;
- Uncertainty related to the choice of model(s) used in experiments;
- Uncertainty related to model structure, including consideration of inter-dependent variables, parameterizations of unresolved processes (e.g., clouds in global climate models), and altogether missing processes and feedback loops;
- Uncertainty related to the initial conditions and assumptions for model runs and the potential sensitivities to small changes in these parameters;
- Uncertainty in the scenarios used for projection (e.g., we do not know what will the greenhouse gas emissions will be in 25 years);
- Uncertainties related to the accuracy of the data used to calibrate or train the models.

Crucially, these all need to be made explicit and folded into a *reasoning framework* to assess uncertainty. This is an area that is not well-developed in the earth and environmental sciences. Notable counter-examples include the work on UncertWeb (Bastin et al., 2013), which offers a set of mechanisms and tools to represent and support reasoning about uncertainty in modeling scenarios (including the use of UncertML to capture meta-data related to uncertainty), EQUIP[5] and QUMP[6]. In hydrology, GLUE (generalized likelihood uncertainty estimation) has been developed to reason about uncertainties in hydrological modeling (Beven and Binley, 2014). Other approaches to managing uncertainty include Differential Adaptive DREAM (Vrugt et al., 2009) and Bayesian Total Error Analysis (Kavetski et al., 2005). Beven and Lamb (2014) also discuss the important aspect of reasoning about cascading uncertainties in integrated modeling. These are examples of good environmental data science being carried out, but not necessarily rolled out across the sub-disciplines of environmental science (another example being data assimilation in numerical weather prediction as mentioned above).

Uncertainty can also usefully be divided into *aleatory and epistemic* sources of uncertainty (Beven and Young, 2013). The word aleatory is derived from the Latin word for die or a game of dice and hence represents random variability that derives from "irreducible natural variability" (Beven, 2015). In contrast, epistemic uncertainty arises from lack of knowledge and hence the uncertainties can be reduced by the availability of new knowledge. In other words, aleatory uncertainty can be captured stochastically through "odds" whereas for epistemic uncertainty additional information is always required to assess the level of uncertainty. If this information becomes available, then epistemic uncertainties can become aleatory in nature but the danger is that such uncertainties may be irreducible. Such uncertainties are very hard to deal with through stochastic means. They may then appear "rather arbitrary in their occurrence" and equate to "surprises," which must then be dealt with in associated model structures (Beven, 2015).

To deal with such uncertainties, Beven (2007) has long argued for new approaches to modeling. His hypothesis is that such epistemic uncertainties will never be accurately captured by probabilistic models and he proposes an approach to models which they deem *models of everywhere*. In this approach, models operate at very fine spatial resolution, associated with particular places, and many such models co-exist. With this approach, it is then possible to collect local data including data from local historic records and derived from local knowledge (for example from farmers), and this knowledge can help resolve both styles of uncertainty and, in particular, deal with surprises. While derived

[5] www.equip.leeds.ac.uk

[6] https://www.metoffice.gov.uk/research/applied/international/precis/qump

for hydrology, such an approach has promise for other fields of environmental modeling in managing different sources and styles of uncertainty. Other approaches may also be applicable in this context. For example, machine learning may have a role in dealing with emergent properties and surprises emanating from models and their interactions.

Another interesting possibility is to consider *adaptive strategies* in response to estimates of uncertainty and indeed in response to more general contextual information around model execution and environmental observation. There is a strong literature in Computer Science around adaptive/self-adaptive/autonomic computing (Kephart and Chess, 2003; McKinley et al., 2004; Cheng et al., 2009) and there is potential to apply this work in the area of environmental understanding. One example would be to link environmental models and Internet of Things technology whereby the uncertainty in models is used to drive the volume and velocity of the sampled data from the Internet of Things infrastructure. There has been some work in adaptive environmental modeling, for example the use of adaptive mesh refinement to adapt the resolution of model execution in response to the sensitivity or turbulence of the area being modeled (Cornford et al., 2013). There is the potential to go much further though in terms of *self-organizing environmental modeling frameworks* that adjust their modeling strategies and approaches to reduce uncertainty and more generally achieve the overall goals of the modeling experiment. Once again, there is strong potential to employ machine learning and other related data science methods in this context.

Summary of uncertainty challenges
Reifying uncertainty as a first class entity in all aspects of environmental science related to data and models;
Providing a framework to support reasoning about uncertainty;
Developing data science techniques to deal with epistemic uncertainties including emergent events and surprises emanating from the underlying complexity of the systems being observed or modeled;
Based on uncertainty and other contextual information, seek adaptive strategies for sampling and model execution.

The Cross-Disciplinary Challenge

The cross-disciplinary space associated with environmental data science is shown pictorially in **Figure 2** below.

This is not just a matter of bringing researchers together from different disciplines. We argue that this requires new means of organization, new methods and indeed a fundamentally new culture of working; and that this is at the heart of the promise of data science as an emerging area of study. This contrasts significantly with current modes of organization and working in universities, research labs and funding councils where research is often categorized and, by implication, siloed.

It is also important, as discussed above, that data science is situated in the problem domain and that we have a data science *of* the natural environment and not data science *for* the natural environment. This also poses significant challenges for the modes or organization discussed above.

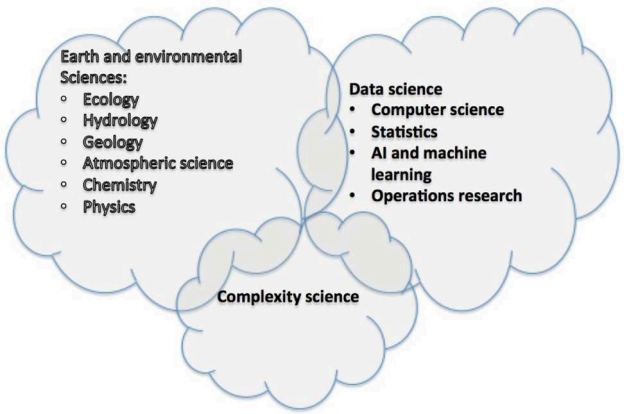

FIGURE 2 | The cross-disciplinary nature of environmental data science.

Summary of cross-disciplinary challenges
Bringing together of a wide variety of disciplines in new data science initiatives targeting the natural environment;
The identification and discovery of new means of organization and fundamentally new modes of working to expedite and maximize innovation at the interface between the many disciplines involved;
To ensure the resultant data science is embedded in the problem domain.

DATA SCIENCE OF THE NATURAL ENVIRONMENT: REVISITED

Building on the discussion above, environmental data science is concerned with *the development of data science principles and techniques for sense making and decision support related to the natural environment*. From the discussion in section Challenges, it is apparent that environmental data science is *distinctive with a set of challenges that are unique to this area* (particularly when considered collectively). For example, very few other application domains have the same rich legacy of process models, which then must be combined with a data models to develop a more complete understanding; the spatial and temporal dimensions are highly distinctive; the level of heterogeneity across data and process is high; when coupled with issue around uncertainty and complexity, this is a uniquely challenging but exciting field.

One of the over-arching themes that comes across from considering the challenges is that of *integration*: of a rich variety of data sources, of models, of data with models, and most profoundly of disciplines to work together in interpreting the associated data and models to achieve new scientific insights through a new integrative science. Some of the building blocks of this are more obvious and straightforward such as the role of cloud computing in providing a common platform for the technological aspects of integration, complemented by emerging standards to ensure interoperability. Such platforms also enable a more open and collaborative approach to science, providing a catalyst and common focus for the necessary cross-disciplinary collaboration. Other aspects are more challenging, particularly

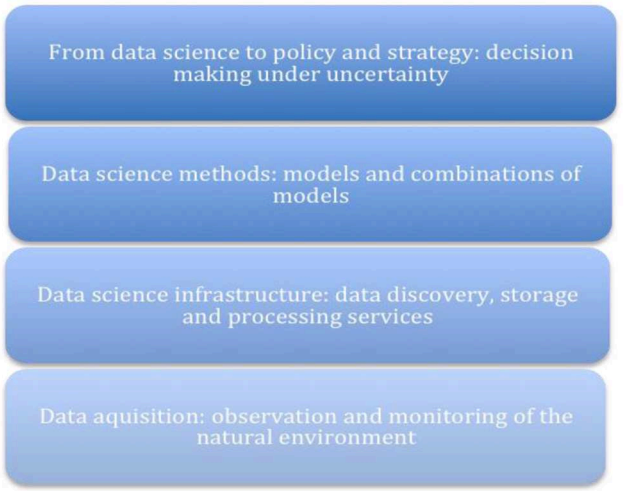

FIGURE 3 | Key data science themes.

the challenges of meaningful cross-disciplinary collaboration, and this requires profound shifts in culture, method, and organization to be effective.

The second key over-arching theme is that of maintaining *broad vision* and not getting too narrow in the definition of data science. This means embracing a rich set of potential data sources, new methods of modeling and data interpretation and of course the breadth and richness that comes from multiple disciplinary perspectives on key scientific problems. Environmental data science is not about better process modeling for earth and environmental science. Nor is it about deep learning on unstructured environmental data. It is about the possibilities of transformation and intellectual breakthrough when we embrace the full breadth and diversity that should be apparent from the discussions above and seek innovations at the interfaces between disciplinary perspectives as well as learning from best practices in other areas, for example the deep understanding of integrated modeling in the climate science community or the insights into data assimilation in weather prediction.

Data science is a new and emerging area of study and environmental data science is in its infancy. The topic can usefully be broken down into a series of over-lapping and mutually supporting themes as shown in **Figure 3**.

The first area is *data acquisition*, embracing the breadth of existing and emerging techniques to provide a significant step change in the observation and monitoring of the natural environment (see section The Data Challenge).

Building on this, there is a need to provide appropriate data science infrastructure supporting *data storage, discovery, and processing capabilities*. This builds on innovation in the area of cloud computing, but many research challenges remain before this infrastructure is fit for purpose for the challenges of environmental data science (Elkhatib et al., 2013).

Continuing upwards in this diagram, there is a need to provide appropriate *data science methods* to help make sense of the plethora of environmental data. This is arguably the biggest area of potential innovation. The core of environmental data science is providing novel methods and combinations of methods to solve particular scientific challenges and problems. This includes combinations of process models and data-driven models, with the latter drawing on areas such as spatial and temporal statistics, machine learning, deep learning, extreme value theory, changepoint analysis, and optimization, offering a rich "playground" for innovation and offering new tools to scientists in responding to the challenges alluded to above (An example of such a combination is provided in section Case Study: Modeling Extreme Melt Events on the Greenland Ice Sheet below).

The final area is that of supporting *decision making in an uncertain and complex world*, and this involves the development of new methods of decision support aligned with ways of communication of data-driven scientific output and its translation into new understanding and policy development. This can also draw on new developments in visualization to aid the interpretation and understanding of the underlying complex and inevitably messy data. This is arguably the most important area but also the most complex and under-developed.

CASE STUDY: MODELING EXTREME MELT EVENTS ON THE GREENLAND ICE SHEET

Purpose of the Study

Modeling Extreme Melt Events on the Greenland ice sheet (MEMOG) was originally a feasibility study supported through the EPSRC funded SECURE network. The aim of the project was to assess the potential for integrating process and stochastic models to improve forecasts of future Greenland ice sheet melting, working at the boundary between data science and environmental science. We found that process-based models that are currently used to simulate future Greenland ice sheet melting (Regional Climate Models—RCMs), and the associated contribution to global sea level rise, underestimate present-day melting because they do not capture extremely high temperatures (**Figure 4**). Preliminary investigations (unpublished) also suggested that statistical models associated with Extreme Value Analysis can potentially be used to downscale RCM predictions of temperature to give better agreement with observed behavior. This has led to two main research priorities: (1) improving the representation of processes in RCMs such that they simulate extreme temperatures with greater fidelity and (2) developing *data-driven* models of extreme melting on Greenland to support work conducted with the process-based RCM. The latter is a key focus of current research in the EPSRC funded Data Science for the Natural Environment (DSNE) project in which we aim to use such models to (a) quantify the spatial/temporal structure of extreme melt events (b) make predictions on the risk of future extremes and (c) downscale RCM predictions to improve their fidelity with respect to observed behavior.

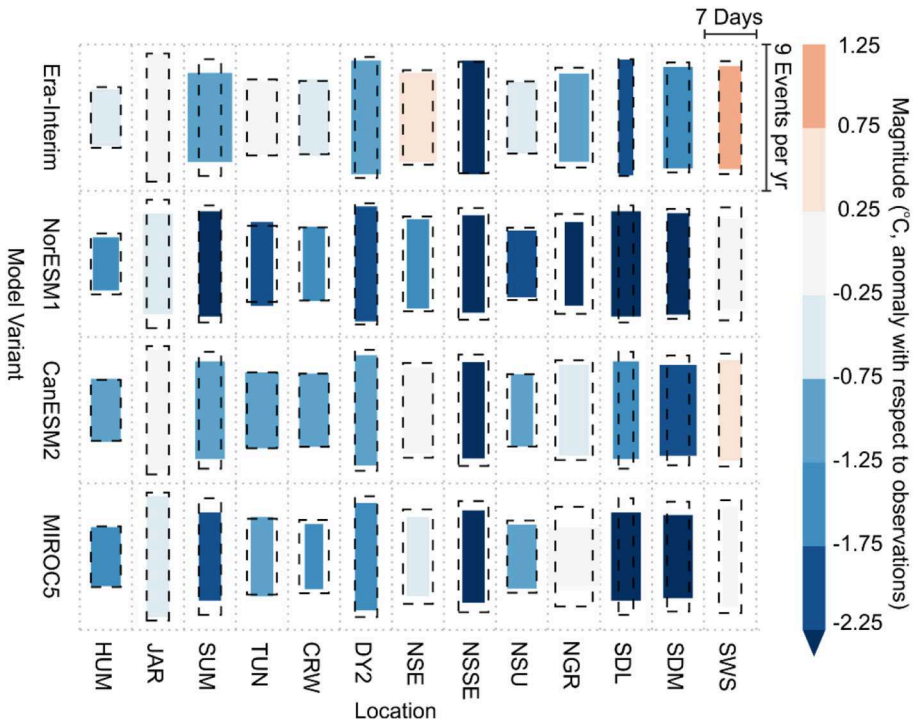

FIGURE 4 | Median frequency, duration, and magnitude of extreme temperature events simulated by four model variants at 13 locations. Frequency is denoted by the height of each box, duration is indicated by the width of each box, and observed values are given by the dashed black boxes. Box colors indicate the departure of the modeled magnitude from the observed value, blue colors indicate an underestimate, and red colors indicate an over estimate (Leeson et al., 2017).

Data Science Perspective

Extreme value analysis (EVA) provides a tool-kit of asymptotically motivated statistical methods that can be used to statistically model the extreme values of a data set and predict the size and frequency of unusually large (or small) events in the future. In an environmental context this could include modeling extreme wind speeds, temperatures, droughts, precipitation, wave heights etc. It has previously been used with great success in many environmental applications, e.g., flood forecasting, however it has never been considered in estimates of cryospheric change. Historically, EVA models made the assumption of independent extreme events and had limited capacity to account for temporal and spatial trends. For many climate-driven processes, such limitations are a major weakness as they limit the ability to account for climate change, which has a strong signal in Greenland (Hanna et al., 2012). In recent and ongoing research, several methods have been developed to deal with these issues by the use of covariates, random effects (also known as latent processes) or multivariate methods (Eastoe and Tawn, 2009, 2012). Using a subset of these state-of-the-art techniques, in the MEMOG project we modeled the frequency, distribution and magnitude of statistically extreme temperature events in *in-situ* observations and contemporaneous RCM predictions at 13 sites. We then used these data to (1) develop a climatology of extreme temperature events in the observational record, (2) analyse the performance of the RCM in terms of reproducing these extremes, and (3) make preliminary (unpublished) investigations into developing a method by which EVA can be used to downscale RCM output to reproduces extreme events with greater fidelity.

Analysis

This work is proving to be an excellent platform to explore the potential for an integrated modeling approach to ice melt prediction. We have achieved success with our marginal (site-wise) approach and the next step is to incorporate spatial elements. In order to do this however, there are a number of issues that our current efforts aim to overcome. These include:

1. **Sparsity of *in-situ* observations.** In order to independently model extreme events, i.e., without using the spurious RCM output, one would ideally want direct observations. The Greenland ice sheet is 1.71 million km^2 and yet there are only ∼20 weather stations on its surface from which it is possible to acquire temperature and melting data. As such, it is not yet possible to model the spatial dependence of extreme melt events using these data.
2. **Data heterogeneity.** While gridded satellite-derived observations of temperature and melting covering the entire ice sheet do exist, these data are a derived product that suffer from heterogeneity. For example, neighboring pixels in the dataset may have been acquired at different times of day and thus are not directly comparable. In addition, there are no observations during periods of cloud cover, and since

these periods tend to be associated with higher temperatures than usual it is not possible to assume these data are "missing at random" for statistical modeling purposes.
3. **High volumes of data.** While observational data are sparse, RCM output is abundant (order of Tb) and continuous in both time and space. This provides an opportunity in that it enables us to explore the spatio-temporal dependence of extreme events (albeit in the model space only) however it also presents additional challenges in terms of both necessary computational power, and devising meaningful data-reduction techniques (e.g., clustering) in order to enable useful inference from the data.

Lessons Learned for Environmental Science

While this work is focused on Greenland ice sheet melting, lessons learned during this process are eminently transferable to other areas of Environmental Science.

1. It is insufficient to test process model fidelity against aggregated data such as annual, or even seasonal, means; "outliers" are important when it comes to overall model performance. Here, we were able to perform a more robust assessment using EVA to compare modeled vs. observed extreme events and found that the RCM misses 16–41% of melt energy at selected locations, largely due to poor representation of temperature extremes.
2. The heterogeneity inherent in environmental data requires a high degree of innovation in applying data science methods. For example, in this study we found that the strength of extremal dependence between observations and climate model output varied between sites. This necessitated the use of a sufficiently flexible bivariate EVA model that could then be applied across a number of heterogeneous locations regardless of the type of extremal dependence. Studying the spatial dependencies in extremal behavior revealed by this study is now a key part of ongoing work.
3. Understanding the physical drivers of why RCMs may not represent extreme events is difficult as they are extremely large and complex models comprising many interconnected processes. However, by using EVA we may be able to correct for this at least. This is important because assessments of the ice sheet contribution to sea level rise (i.e., total melting) are used for policy and decision-making.
4. Integration of process and statistical models presents in itself a novel research challenge and further effort is needed in order to determine principled ways of using the EVA model to drive the RCM output into states that do not naturally arise from integrations of model physics.

This case study, on combining process and stochastic models, demonstrates how models of different kinds can usefully be combined to better represent the reality, in this case, of extreme events. We see many other innovative combinations of process models and data-driven or stochastic models, for example the use of changepoint analysis or machine learning alongside process models (a couple of studies have also recently used machine learning in this way to attempt to derive patterns that are indicative of El Niño occurrences, a complex phenomenon that has so far eluded traditional process-based analyses; Lima et al., 2015; Chalupka et al., 2016). We also see data science methods being usefully combined in different ways to create hybrid approaches, for example the use of changepoint analysis with machine learning to discover patterns of higher-level events resulting from fundamental change in the environment. This is core to our vision of a future environmental data science—that is, by enabling innovation at the interfaces between disciplines and approaches, through bringing the different groups of researchers together in multi-disciplinary teams.

A RESEARCH ROADMAP

Building on our analyses and experiences documented above, we present a research roadmap for data science for the natural environment in terms of a *top 10 set of research challenges*[7]. This is not necessarily intended to be complete but rather to highlight from our perspective some of the key challenges that must be addressed to achieve a form of maturity in this area.

Challenge 1: To encourage and enable a *cultural shift toward open science*, that is toward a science that is more collaborative and integrative through open approaches to data, models and knowledge formation, and also toward a science that is more transparent, repeatable and reproducible.

Challenge 2: To build on the benefits of cloud computing, but offer *levels of abstraction* (and associated services) that are much better suited to the domain of science, including high-level support for running complex, integrated modeling in the cloud.

Challenge 3: To address *complexity* more fundamentally and explicitly, in particular, seeking data science techniques that recognize and resolve key issues around feedback loops, inter-dependent variables, extremes and reasoning about emergent behavior.

Challenge 4: To provide techniques and frameworks to both reify *uncertainty* in scientific studies and also reason about the cascading uncertainties across complex experiments, e.g., in integrated modeling frameworks.

Challenge 5: To seek *adaptive* techniques driven by considerations of uncertainty and also the goals of a scientific study, including adaptive approaches to sampling or gathering of data and adaptive modeling.

Challenge 6: To seek approaches that deal with *epistemic uncertainty* in environmental modeling, noting the important links with dealing with emergent behavior in complex and irreducible phenomena.

Challenge 7: To seek *novel data science techniques* and, in particular, innovative *combinations of data science techniques* that can make sense of the increasing complexity, variety and veracity of underlying environmental data, exploiting also multiple data sets including real-time streaming data.

[7]It is important to stress that there is excellent work in many of these areas in the environmental sciences but this work is rather fragmented and it is clear that a more integrated approach is required.

Challenge 8: To seek innovations in modeling by *combining process models with data-driven or stochastic modeling techniques* and also seeking ways of assimilating a range of data sources more generally into steering model executions.
Challenge 9: To incorporate sophisticated *spatial and temporal reasoning*, including reasoning across scales, as an integral aspect of environmental data science and not something that is just provided through separate tools such as GIS tools.
Challenge 10: To discover *new modes of working, methods and means of organization* that enable the required level of cross-disciplinary collaboration as required to address the grand challenges of earth and environmental sciences and, more specifically, environmental data science in its contribution to these grand challenges.

These research challenges cross-cut the themes of data acquisition, infrastructure, methods and policy making as illustrated in **Figure 3**. The overarching challenge is then to overcome the 10 challenges above in an *end-to-end environmental data science* (from acquisition right through to policy and strategy) and to apply such techniques in responding to the many problems around the management of the natural environment.

CONCLUDING REMARKS

This paper has discussed the emergent area of environmental data science arguing that there is an important symbiotic relationship between the fields of data science and earth/environmental sciences: data science has a lot to offer in terms of a deeper understanding the natural environment and in informing mitigation and adaptation strategies in the face of climate change; this domain of application has much to offer in terms of data science with its unique combination of challenges, challenges that require significant breakthrough and innovation in data science methods.

The contributions of the paper are: (i) a definition of the field of environmental data science; (ii) a systematic analysis of the range of challenges in environmental data science; (iii) a research roadmap in the form of 10 key research challenges that, if addressed, would lead to significant progress in environmental data science.

The paper sets out with the additional objective of reaching out to researchers working in this space to create an international community to address the very significant challenges in this area. In retrospect, the creation of such an international community would dwarf the other contributions in terms of long-term significance. We invite you to this international effort.

AUTHOR CONTRIBUTIONS

GB was the principal author for the text, and responsible for editing the whole manuscript together. Other authors contributed to the underlying research and to its analysis, and also contributed specific sections of text. AL and EE were the main authors for section Case Study: Modeling Extreme Melt Events on the Greenland Ice Sheet, while PY and AL led the writing of The Modeling Challenge and PH and EE led on The Spatial/Temporal Challenge.

FUNDING

This work was partially supported by the following grants: DT/LWEC Senior Fellowship (awarded to GB) on the Role of Digital Technology in Understanding, Mitigating, and Adapting to Environmental Change, EPSRC: EP/P002285/1; Models in the Cloud: Generative Software Frameworks to Support the Execution of Environmental Models in the Cloud, EPSRC: EP/N027736/1; Data Science of the Natural Environment, EPSRC: EP/R01860X/1; Modeling extreme melt events on the Greenland ice sheet, SECURE Network, EPSRC: EP/M008347/1 (FP2016008AL); NE/R016429/1 as part of the UK-SCAPE programme delivering National Capability.

ACKNOWLEDGMENTS

We would like to acknowledge our colleagues in the Centre of Excellence in Environmental Data Science, the Data Science Institute at Lancaster, and the Ensemble research group for providing such a stimulating cross-disciplinary environment and context for this research.

REFERENCES

Alexander, K., and Easterbrook, S. M. (2015). The software architecture of climate models: a graphical comparison of CMIP5 and EMICAR5 configurations. *Geosci. Model Dev.* 8, 1221–1232. doi: 10.5194/gmd-8-1221-2015

Alvarez, J. L., Yumashev, D., Whiteman, G., Wilkinson, J., Hope, C. K., and Wadhams, P. (2015). "Is the Arctic an economic time bomb?: Integrated assessment models can help answer this question," in *Proceedings of the 11th International Conference of the European Society for Ecological Economics* (Leeds).

Atzori, L., Iera, A., and Morabito, G. (2010). The Internet of Things: a survey. *Comput. Netw.* 54, 2787–2805. doi: 10.1016/j.comnet.2010.05.010

Baesens, B. (2014). *Analytics in a Big Data World: The Essential Guide to Data Science and its Applications, 1st Edn.* Wiley Publishing.

Bastin, L., Cornford, D., Jones, R., Heuvelink, G. B. M., Pebesma, E., Stasch, C., et al. (2013). Managing uncertainty in integrated environmental modelling: the UncertWeb framework. *Environ. Model. Softw.* 39, 116–134. doi: 10.1016/j.envsoft.2012.02.008

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web. *Sci. Am.* 2841, 34–43. doi: 10.1038/scientificamerican0501-34

Beven, K. (2007). Towards integrated environmental models of everywhere: uncertainty, data and modelling as a learning process. *Hydrol. Earth Syst. Sci.* 11, 460–467. doi: 10.5194/hess-11-460-2007

Beven, K. (2015). Facets of uncertainty: epistemic uncertainty, non-stationarity, likelihood, hypothesis testing, and communication. *Hydrol. Sci. J.* 61, 1652–1665. doi: 10.1080/02626667.2015.1031761

Beven, K., and Binley, A. (1992). The future of distributed models: Model calibration and uncertainty prediction. *Hydrol. Process.* 6, 279–298. doi: 10.1002/hyp.3360060305

Beven, K., and Binley, A. (2014). GLUE: twenty years on. *Hydrol. Process.* 28, 5897–5918. doi: 10.1002/hyp.10082

Beven, K., and Lamb, R. (2014). "The uncertainty cascade in model fusion," in *Integrated Environmental Modelling to Solve Real World Problems*, eds A. T. Riddick, H. Kessler, and J. R. A. Giles (London: Geological Society of London), 255–266. doi: 10.1144/SP 408.3

Beven, K., and Young, P. (2013). A guide to good practice in modeling semantics for authors and referees. *Water Resour. Res.* 49, 5092–5098. doi: 10.1002/wrcr.20393

Bizer, C., Heath, T., and Berners-Lee, T. (2010). Linked data – the story so far. *Int. J. Semant. Web Inf. Syst*. 5, 1–22. doi: 10.4018/jswis.2009081901

Brockwell, P. J., Davis, R. A., and Calder, M. V. (2002). *Introduction to Time Series and Forecasting, Vol. 2*. New York, NY: Springer. doi: 10.1007/b97391

Carslaw, K. S., Lee, L. A., Reddington, C. L., Pringle, K. J., Rap, A., Forster, P. M., et al. (2013). Large contribution of natural aerosols to uncertainty in indirect forcing. *Nature* 503, 67–71. doi: 10.1038/nature12674

Cervone, G., Sava, E., Huang, Q., Schnebele, E., Harrison, L., and Waters, N. (2016). Using Twitter for tasking remote sensing data collection and damage assessment: 2013 Boulder Flood Case Study. *Int. J. Remote Sens*. 37, 100–124. doi: 10.1080/01431161.2015.1117684

Chalupka, K., Bischoff, T., Perona, P., and Eberhardt, F. (2016). "Unsupervised discovery of El Nino using causal feature learning on microlevel climate data," in *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence (UAI'16)* (Arlington, VA: AUAI Press), 72–81.

Cheng, B. H., Lemos, R., Giese, H., Inverardi, P., Magee, J., Andersson, J., et al. (2009). "Software engineering for self-adaptive systems: a research roadmap," in *Software Engineering for Self-Adaptive Systems*, eds B. H. Cheng, R. Lemos, H. Giese, P. Inverardi, and J. Magee, Lecture Notes in Computer Science, Vol. 5525 (Berlin; Heidelberg: Springer-Verlag), 1–26. doi: 10.1007/978-3-642-02161-9_1

Coles, S., Heffernan, J., and Tawn, J. (1999). Dependence measures for extreme value analyses. *Extremes* 2, 339–365. doi: 10.1023/A:1009963131610

Compton, M., Barnaghi, P., Bermudez, L., García-Castro, R., Corcho, O., Cox, S., et al. (2012). The SSN ontology of the W3C semantic sensor network incubator group. *Web Semant.* 17, 25–32. doi: 10.1016/j.websem.2012.05.003

Cornford, S. L., Martin, D. F., Graves, D. T., Ranken, D. F., Le Brocq, A. M., Gladstone, R. M., et al. Lipscomb, W.H. (2013). Adaptive mesh, finite volume modeling of marine ice sheets. *J. Comput. Phys.* 232, 529–549. doi: 10.1016/j.jcp.2012.08.037

Cressie, N. A. C. (1993). *Statistics for Spatial Data.* New York, NY: John Wiley & Sons. doi: 10.1002/9781119115151

Davison, A. C., Padoan, S. A., and Ribatet, M. (2012). Statistical modeling of spatial extremes. *Stat. Sci.* 27, 161–186. doi: 10.1214/11-STS376

Dean, J., and Ghemawat, S. (2004). "MapReduce: simplified data processing on large clusters," in *Proceedings of the 6th conference on Symposium on Operating Systems Design & Implementation (OSDI'04), Vol. 6* (Berkeley, CA: USENIX Association).

Dhar, V. (2013). Data science and prediction. *Commun. ACM*. 56, 64–73. doi: 10.1145/2500499

Eastoe, E. F., and Tawn, J. A. (2009). Modelling non-stationary extremes with application to surface level ozone. *J. R. Stat. Soc. C.* 58, 45–55. doi: 10.1111/j.1467-9876.2008.00638.x

Eastoe, E. F., and Tawn, J. A. (2012). Modelling the distribution of the cluster maxima of exceedances of subasymptotic thresholds. *Biometrika.* 99, 43–55. doi: 10.1093/biomet/asr078

Elkhatib, Y., Blair, G. S., and Surajbali, B. (2013). "Experiences of using a hybrid cloud to construct an environmental virtual observatory," in *Proceedings of the 3rd International Workshop on Cloud Data and Platforms (CloudDP '13) ACM* (New York, NY), 13-18. doi: 10.1145/2460756.2460759

Gelfand, A. E., Diggle, P., Guttorp, P., and Fuentes, M. (eds.). (2010). *Handbook of Spatial Statistics* (Boca Raton, FL: CRC Press). doi: 10.1201/9781420072884

Godard, M. A., Dougill, A. J., and Benton, T. G. (2010). Scaling up from gardens: biodiversity conservation in urban environments. *Trends Ecol. Evol*. 25, 90–98. doi: 10.1016/j.tree.2009.07.016

Greene, C. S., Jie Tan, J., Ung, M., Moore, J. H., and Cheng, C. (2014). Big data bioinformatics. *J. Cell. Physiol.* 229, 1896–1900. doi: 10.1002/jcp.24662

Greening, L. A., Greene, D. L., and Difiglio, C. (2000). Energy efficiency and consumption – the rebound effect – a survey. *Energy Policy* 28, 389–401. doi: 10.1016/S0301-4215(00)00021-5

Hanna, E., Mernild, S. H., Cappelen, J., and Steffen, K. (2012). Recent warming in Greenland in a long-term instrumental (1881-2012) climatic context: I. Evaluation of surface air temperature records. *Environ. Res. Lett.* 7. doi: 10.1088/1748-9326/7/4/045404

Harrison, P. A., Dunford, R., Savin, C., Rounsevell, M. D. A., Holman, I. P., Kebede, A. S., et al. (2015). Cross-sectoral impacts of climate change and socio-economic change for multiple, European land- and water-based sectors. *Clim. Change* 128, 279–292. doi: 10.1007/s10584-014-1239-4

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., et al. (2018). A case study competition among methods for analyzing large spatial data. *J. Agric. Biol. Environ. Stat.* 1–28. doi: 10.1007/s13253-018-00348-w

Helm, D. (2015). *Natural Capital - Valuing Our Planet*. Yale University Press.

Hey, T., Tansley, S., and Troll, K. (eds.). (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research.

Hitzler, P., and Janowicz, K. (2013). Linked data, big data, and the 4th paradigm. *Semant. Web* 4, 233–235. doi: 10.3233/SW-130117

Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., et al. (2014). Big data and its technical challenges. *Commun. ACM* 57, 86–94. doi: 10.1145/2611567

Jarvis, A. J., Leedal, D. T., and Hewitt, N. (2012). Climate-society feedbacks and the avoidance of dangerous climate change. *Nat. Clim. Change* 2, 668–671. doi: 10.1038/nclimate1586

Kastens, K. A., Manduca, C. A., Cervato, C., Frodeman, R., Goodwin, C., Liben, L. S., et al. (2009). How geoscientists think and learn. *Eos Trans*. 90, 265–266. doi: 10.1029/2009EO310001

Kavetski, D., Kuczera, G., and Franks, S. W. (2005). Bayesian analysis of input uncertainty in hydrological modeling: 2. Application. *Water Resour. Res.* 42:3. doi: 10.1029/2005WR004376

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., et al. (2015). The Community Earth System Model (CESM) large ensemble project: a community resource for studying climate change in the presence of internal climate variability. *Bull. Am. Meteor. Soc.* 96, 1333–1349. doi: 10.1175/BAMS-D-13-00255.1

Kephart, J. O., and Chess, D. M. (2003). The vision of autonomic computing. *Computer* 36, 41–50. doi: 10.1109/MC.2003.1160055

Lahoz, W., Khattatov, B., and Menard, R. (eds.). (2010). *Data Assimilation: Making Sense of Observations*. Springer Science & Business Media. doi: 10.1007/978-3-540-74703-1

Langley, E. S., Leeson, A. A., Stokes, C. R., and Jamieson, S. S. R. (2016). Seasonal evolution of supraglacial lakes on an East Antarctic outley glacier. *Geophys. Res. Lett*. 43, 8563–8571. doi: 10.1002/2016GL069511

Laniak, G. F., Olchin, G., Goodall, J., Voinov, A., Hill, M., Glynn, P., et al. (2013). Integrated environmental modeling: a vision and roadmap for the future. *Environ. Modell. Softw.* 39, 3–23. doi: 10.1016/j.envsoft.2012.09.006

Leeson, A. A., Eastoe, E., and Fettweis, X. (2017). Extreme temperature events on Greenland in observations and the MAR regional climate model. *Cryosphere* 12, 1091–1102. doi: 10.5194/tc-12-1091-2018

Lima, C. H. R., Lall, U., Jebara, T., and Barnston, A. G. (2015). "Machine learning methods for ENSO analysis and prediction," in *Machine Learning and Data Mining Approaches to Climate Science,* eds V. Lakshmanan, E. Gilleland, A. McGovern, and M. Tingley (Springer), 13–21. doi: 10.1007/978-3-319-17220-0_2

Marx, V. (2013). Biology: the big challenges of big data. *Nature* 498, 255–260. doi: 10.1038/498255a

Mayer-Schonberger, V., and Cukier, K. (eds.). (2013). *Big Data: A Revolution That Will Transform How We Live, Work and Think* (London: John Murray Publishers).

McKinley, P. K., Sadjadi, S. M., Kasten, E. P., and Cheng, B. H. C. (2004). Composing adaptive software. *Computer* 37, 56–64. doi: 10.1109/MC.2004.48

Mei, S., Li, H., Fan, J., Zhu, X., and Dyer, C. (2014). "Inferring air pollution by sniffing social media," in *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)* (Piscataway, NJ). doi: 10.1109/ASONAM.2014.6921638

Muller, F., de Groot, R., and Willemen, L. (2010). Ecosystem services at the landscape scale: the need for integrative approaches. *Landsc. Online* 23, 1–11. doi: 10.3097/LO.201023

Niu, S., Luo, Y., Dietze, M. C., Keenan, T. F., Shi, Z., Li, J., et al. (2014). The role of data assimilation in predictive ecology. *Ecosphere* 5:65. doi: 10.1890/ES13-00273.1

Nundloll, V., Porter, B., Blair, G. S., Emmett, B., Cosby, J., Jones, D., et al. (2019). The design and deployment of an end-to-end IoT infrastructure for the natural environment. *Future Intern.* 11:129. doi: 10.3390/fi11060129

Park, S. K., and Xu, L. (eds.). (2017). *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications, Vol. 3*. Springer Science & Business Media. doi: 10.1007/978-3-319-43415-5

Philip Chen, C. L., and Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: a survey on big data. *Inf. Sci.* 275, 314–347. doi: 10.1016/j.ins.2014.01.015

Potschin, M., Haines-Young, R., Fish, R., and Kerry Turner, R. (2016). *Routledge Handbook of Ecosystem Services*. New York, NY: Routledge.

Provost, F., and Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big Data*. 1:1. doi: 10.1089/big.2013.1508

Raskin, R. G., and Pan, M. J. (2005). Knowledge representation in the semantic web for Earth and environmental terminology (SWEET). *Comput. Geosci.* 31, 1119–1125. doi: 10.1016/j.cageo.2004.12.004

Reed, D. A., and Dongarra, J. (2015). Exascale computing and big data. *Commun. ACM* 58, 56–68. doi: 10.1145/2699414

Reis, S., Seto, E., Northcross, A., Quinn, N. W. T., Convertino, M., Jones, R. L., et al. (2015). Integrating modelling and smart sensors for environmental and human health. *Environ. Model. Softw.* 74, 238–246. doi: 10.1016/j.envsoft.2015.06.003

Schnase, J. L., et al. (2017). MERRA analytic services: meeting the big data challenges of climate science through cloud-enabled climate analytics as a service. *Comput. Environ. Urban Syst.* 61, 198–211. doi: 10.1016/j.compenvurbsys.2013.12.003

Shelley, W., Lawley, R., and Robinson, D. A. (2013). Technology: crowd-sources soil data for Europe. *Nature* 496:300. doi: 10.1038/496300d

Tawn, J. A. (1988). Bivariate extreme value theory: models and estimation. *Biometrika* 75, 397–415. doi: 10.1093/biomet/75.3.397

Taylor, K. E., Stouffer, R. J., and Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bull. Am. Meteor. Soc.* 93, 485–498. doi: 10.1175/BAMS-D-11-00094.1

Thackeray, S. J., et al. (2016). Phenological sensitivity to climate across taxa and trophic levels. *Nature* 535, 241–245. doi: 10.1038/nature18608

Vrugt, J. A., ter Braak, C. J. F., Diks, C. G. H., Robinson, B. A., Hyman, J. M., and Higdon, D. (2009). Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. *Int. J. Nonlin. Sci. Numer. Simul.* 10, 273–290. doi: 10.1515/IJNSNS.2009.10.3.273

Wang, B., Li, R., and Perrizo, W. (eds.). (2015). *Big Data Analytics in Bioinformatics and Healthcare* (Hershey, PA: IGI Global). doi: 10.4018/978-1-4666-6611-5

Wilby, R. L., and Dessai, S. (2010). Robust adaptation to climate change. *Weather* 65, 180–185. doi: 10.1002/wea.543

Williams, M., Schwarz, P. A., Law, B. E., Irvine, J., and Kurpius, M. R. (2005). An improved analysis of forest carbon dynamics using data assimilation. *Glob. Change Biol.* 11, 89–105. doi: 10.1111/j.1365-2486.2004.00891.x

Yucel, I., Onen, A., Yilmaz, K. K., and Gochis, D. J. (2015). Calibration and evaluation of a flood forecasting system: utility of numerical weather prediction model, data assimilation and satellite-based rainfall. *J. Hydrol.* 523, 49–66. doi: 10.1016/j.jhydrol.2015.01.042

Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., et al. (2016). Apache Spark: a unified engine for big data processing. *Commun. ACM* 59, 56–65. doi: 10.1145/2934664

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.