Received: 17 March 2022 | Accepted: 27 April 2022

DOI: 10.1002/env.2732

OPINION

WILEY

Framing data science, analytics and statistics around the digital earth concept

E. Marian Scott

School of Mathematics and Statistics, University of Glasgow, Glasgow, UK

Correspondence
E. Marian Scott, School of Mathematics and Statistics, University of Glasgow, Glasgow, UK.
Email: marian.scott@glasgow.ac.uk

Funding information
UKRI-NERC

Abstract

Environmental data science can be shaped around the concepts of data streams (or the data deluge), and both data driven and process models. Together they lead to the concept of a digital earth. In this short opinion piece, I reflect on some of the challenges in truly realizing the digital earth concept.

1 | INTRODUCTION

How we understand the environment, its connections and changes are based on data. How we generate those data is changing and continuing to evolve, which means also that the complexity of environmental systems can be studied in greater depth, and hidden connections can be explored. It also means that statistical methods need to evolve to deal with new data streams. It is in this landscape, that we often see, on the one hand, terms such as "data deluge," and data lake, and on the other, digital environment, digital twin, and digital earth describing the system we are studying. Together, these could be loosely described as comprising an environmental data science landscape. But are they just buzzwords, or do they reflect a change in how we think about the environment? I have used some recent pieces of work that I, with my colleagues and PhD students, have been involved in to reflect on "what is environmental data sciences." Most of the examples concern our observation of the freshwater environment, and are determined by the spatio-temporal nature of the data, whether from sensors or satellites.

1.1 | Why is the digital earth concept important?

The digital earth concept leads to the framing of our observation and modeling of the environment within a systems thinking approach, which is more holistic than our traditional silo-ed environmental sciences. We know that environmental and ecological systems are interconnected, but building the conceptual system has been and remains challenging and then turning this into a quantitative system remains a significant barrier to progress but progress is being made.

"Earth system science (ESS) is a rapidly emerging transdisciplinary endeavor aimed at understanding the structure and functioning of the Earth as a complex, adaptive system. ESS has produced new concepts and frameworks central to the global-change discourse, including the Anthropocene, tipping elements and planetary boundaries. Moving forward, the grand challenge for ESS is to achieve a deep integration of biophysical processes and human dynamics to build a truly unified understanding of the Earth System" (Steffen et al., 2020).

With the twin challenges of climate change mitigation and adaptation and reversing biodiversity loss, more recently framed in the separate but linked objectives of net zero for CO_2 emissions and nature positive there are substantial gains to be made in driving forward statistical and data analytic advances. This article offers some personal reflections and opinions on where we are and where we need to go, and adopts a digital earth framing of the general context, and uses a few use cases to suggest some of the areas where the biggest analytics gaps still exist.

Starting with net zero and nature positive, these are both recognized as aspects of the twin climate and biodiversity challenges. What is net zero? This can be loosely described as a global ambition (implemented at national level) to achieve a situation where greenhouse gas (GHG) emissions are reduced to zero. So where and how does environmental data science contribute? Given the complexity of the global carbon cycle (both natural and with its anthropogenic perturbations), to demonstrate progress toward reaching this ambitious and essential target (regardless of timescale), we need a measurement and verification system. As written in a recent Royal Society report (RS, 2020).

"Digital technology is embedded in our daily lives and in every sector, so it has a critical and growing part to play in delivering a net zero future. Nearly a third of the 50% carbon emissions reductions the UK needs to make by 2030 could be achieved through existing digital technology—from sensors to largescale modeling. For example, digital twins—virtual representations of physical assets—have already helped increase the yield of some wind farms by up to 20%, as well as improved their life span" (Royal Society, 2020). What is not said in this quote is that data are going to be vital to understanding the impact of our actions, and our trajectory toward these ambitious but essential targets.

What is nature positive? Nature positive is a simple phrase to capture a similarly ambitious target to halt biodiversity decline and to "turn the curve" (Leclère et al., 2020). "Nature positive arose from debate and refers to the ambition to not simply halt biodiversity loss but to reverse the decline" (JNCC, 2021). Historically our biodiversity records are at species level, so that we frequently see indicators of decline or recovery for individual species, but our ambition goes beyond single species, since in the natural world, connections are key.

1.2 | Definitions of data deluge, data flood, and data lake

An environmental data deluge is written about, usually in terms of new sensors, and new satellite missions. This clearly is related to *big data* discussions, which tend to focus on volume and velocity of data acquisition. In the big data debate, there has already been some discussion about the role of statistics, and analytics (Scott, 2018). But whether deluge or flood, there is perhaps a question about whether we have too much data (sometimes of less good quality) and indeed whether we need such volumes, or perhaps phrased in another way, can we make best use of them?

From an analytical and data science point of view, it is not uncommon to aggregate data- who has not aggregated data recorded every minute to hourly or hourly to day? Similarly, when considering time series of earth observation of images, we naively imagine that we have large quantities of data, we do, but actually when we begin our analysis, we often discover that the data are sparse, with large quantities of structured missingness due to cloud cover and other operational features. One response is to aggregate. In Elayouty et al. (2016, 2022) we aggregated data across different timescales, to allow us to explore potential nonstationarity while in Maberly et al. (2020), we considered single lake (mean) curves created from individual pixel curves rather than many individual time series. Whether we aggregate or not, depends of course on the questions of interest and the timescales at which these questions are most relevant, but also on the computational efficiency of whatever tools we are using, so in one sense the scientific question determines whether we have too much data.

In 2011, Porter et al. wrote "Developments in sensor design, electronics, computer technology and networking have converged to provide new ways of collecting environmental data at rates hitherto impossible to achieve. To translate this "data deluge" into scientific knowledge requires comparable advances in our ability to integrate, process and analyze massive data sets" (Porter et al., 2011).

So environmental data science is not simply about more data, but also about how we support new scientific insights using appropriate analytic tools. Data integration remains a key challenge in this space, as we tackle issues around spatial misalignment and change of support. In Wilkie et al. (2019) for instance we tackled the challenge of integrating (or fusing) satellite and in situ data, and increasingly we are facing challenges of satellite imagery at 1 km resolution, being integrated with imagery at 10 m resolution and with point data from local sensors.

Extending then the data deluge, we are also opening our eyes to further forms of data. As statisticians and modelers, we are perhaps most familiar with structured data (many forms of environmental data have geographical and temporal stamps), but semistructured and unstructured data speak to different forms of data, being collected in different ways. In

the systems approach, we are recognizing that there are many modes of observation, many sectors are involved, and so we are engaging more and more with citizens, resulting in a growth in qualitative data, from surveys, twitter feeds, videos and so forth as forms of citizen science data. Integration of qualitative, semiquantitative and quantitative date presents an area of significant analytics challenge. There remain significant governance, privacy and ethical concerns also around citizen science (Webber et al., 2019).

1.3 | Digital environment, a digital twin, a digital earth

Within the UK environmental sciences communities, the UKRI-NERC has operated a digital environment research program (https://digitalenvironment.org/) with the stated goal:

"to develop the digitally enabled environment which benefits scientists, policymakers, businesses, communities and individuals. · · ·to help support the creation of integrated networks of sensors (in situ and remote sensing based), and the methodologies and tools for assessing, analyzing, monitoring and forecasting the state of the natural environment. This will be done at higher spatial resolutions and at higher frequency than previously possible. Key pillars are acquisition of data, storage and processing of big data, data science and AI, visualization and decision support."

The language of a digital twin (https://www.digitaltwinconsortium.org/initiatives/the-definition-of-a-digital-twin.html) goes further.

"A digital twin is a virtual representation of real-world entities and processes, synchronized at a specified frequency and fidelity.

- Digital twin systems transform by accelerating holistic (system) understanding, optimal decision-making, and effective action.
- Digital twins use real-time and historical data to represent the past and present and simulate predicted futures.

Blair (2021) asked "can digital twins fill this void (*analytical tools to handle messy and complex environmental data*) and offer up the tools that environmental science needs in response to the pressures to scale up the science and to properly enable a new, more data-driven style of investigation." and proceeded to develop his argument that twins have the potential to be transformative, which I wholeheartedly endorse.

Within the EU (Nativi et al., 2021), there are equally significant developments, bringing the idea of a digital earth to realization. The "Digital Earth is a concept describing an interactive digital replica of the entire planet that can facilitate a shared understanding of the multiple relationships between the physical and natural environments and society. To do so, it needs to be able to display information in ways that are easily understood by multiple audiences (the public, decision-makers, scientists); and be constantly updated with data coming from sensors (space-based, airborne, in situ), citizens, and both public and private sectors." (https://joint-research-centre.ec.europa.eu/scientific-activities/digital-earth_en#:~:text=Digital%20Earth%20is%20a%20framework,the%20consequences%20of%20human%20activity).

The digital earth definition goes on "It must be able to focus on change (from the past, to present and future) and thus include not just data but also the outcomes of models and simulations to enable a wider understanding of the consequences of human action on the environment, and of environmental change on society."

The bold phrases and descriptors are one which resonate with statisticians, but the digital earth and digital twin concepts are perhaps something that we have not been heavily engaged in traditionally. I suggest therefore there are some missing key aspects, for example, what about uncertainty and inference?

Other authors have gone on to state "The basic components of a digital twin (essentially a model and some data) are generally comparatively mature and well-understood. Many of the aspects of using data in models are similarly well-understood. However, many interesting open questions exist, some connected with the volume and speed of data, some connected with reliability and uncertainty, and some to do with dynamic model updating (Wright & Davidson, 2020).

From these definitions, it is clear that there are at least three important parts in the digital twin of an object:

- a model or nested models of the object. We might ask what type of model, how are they being coupled, how is uncertainty represented and quantified?
- an evolving set of data relating to the object. There are many questions we might imagine, including what is the sampling frame for the data collection or how do we integrate real time and historic data and.
- a means of dynamically updating or adjusting the model in accordance with the data.

2 | DIGITAL TWIN(S), STATISTICAL MODELS, SYSTEMS

Given this basic list of component parts for a digital twin, from my statistical point of view, I might add some additional components and also reorder/expand the list to give the following:

A: conceptual model(s) of the systems under study.
B: a mathematical realization of that conceptual model, indeed in complex systems we could easily imagine nested models or a federated series of twins (Blair, 2021).
C: data (being streamed), the data deluge, real time and historic.
D: technology to update the model and the data (data assimilation).
E: ways to quantify and communicate uncertainty.
F: inference, an essential tool and the uncertainty needed to make the inference.
G: visualization and communication.

I have expanded these components individually, and used our experiences to reflect on some of the statistical and data science aspects.

2.1 | A and B a modeling framework

Like any modeling process, from the simplest structure of one response, one explanatory variable (or feature) to the most complex with hundreds of features, we have a strong statistical tradition in building models, including choosing which variables we need to measure and how they may be related. These are perhaps most commonly defined as data driven models, but it is likely that in environmental systems we may also have access to process models, perhaps based on sets of differential equations, and which represent best scientific understanding about the flows and interconnections. One challenge we have here concerns the use of multiple models where some at least are data driven, and how we couple the different models, including defining the boundaries of the system that we may be interested in. This is particularly important from a systems perspective. Similarly important are the scale or resolution reconciliations over both space and in time that are necessary as we try to couple models. We need to think as well about uncertainty propagation and quantification. We are all very familiar with (and agree with) "All models are wrong, but some are useful." attributed to George Box. But not with- "companies like Google, which have grown up in an era of massively abundant data, don't have to settle for wrong models. Indeed, they don't have to settle for models at all" (https://www.wired.com/2008/06/pb-theory/).

2.2 | C the data deluge

One defining characteristic of the digital earth are the multiple data streams we now deal with routinely, and this is a very strong characteristic of environmental data science. We might imagine this as a data hierarchy, from distributed sensors to satellite earth observation, with differing time and space stamps, but with increasingly with a focus on real time. We also need to recognize that historical data are likely to be much less temporally resolved (e.g., traditional water quality sampling campaigns for compliance purposes were often monthly) but that they are still inherently valuable.

Each data level will have its own uncertainty structure, that may or may not be quantified but certainly needs to be considered. As an example, from our own work, Gong et al. (2022) describe the uncertainty quantification in a time series of satellite observations and how those uncertainties are then incorporated into functional data approaches and smoothing.

As part of the deluge, we also need to pay attention to data quality assurance. Anomalies might be considered as nuisance but anomalies may also be events of interest. There has been a growth in development of machine learning methods including deep learning methods for anomaly identification and this is likely to remain an area for further growth. Missingness, and sparseness present further challenges (Gong et al., 2021) presented a computationally efficient, functional data approach to dealing with missing data in a series of satellite images.

As already mentioned, data integration or fusion plays a very big part in the digital earth. We have multiple challenges in developing methods to handle the diversity of data. Data fusion is very much a statistical issue, dealing with change of support issues. Our preferred approach has been to take a functional data approach as described in Wilkie et al. (2019).

This offers important computational efficiency but does make that fundamental assumption of smoothness which may or may not be realistic.

We all know, "There has been considerable change in the nature of data. In 1977 large and complex data sets were fairly rare, and little need was seen to attempt to analyze those few that did exist. Twenty years ago most data was still collected manually. The cost of collecting it was proportional to the amount collected. The goal was to carefully design experiments so that maximal information could be obtained with the fewest possible measurements." Jerome H. Friedman 2001 (in The Role of Statistics in the Data Revolution [2001]). Now another 20 years on, this statement still holds, certainly about the change, but still we miss the important role of experimental design and sampling. Environmental data science needs these considerations.

2.3 | D:Technology for coupling the models and data

One key feature of the digital twin concept is the idea that there should be a feedback between data and model, and that as new data arrive in real time, the model updates. The solution is data assimilation, which has been widely used in weather forecasting, but is also more generally in hydrological and oceanographic modeling. "Data assimilation provides an objective methodology to combine observational and model information to provide an estimate of the most likely state and its uncertainty for the whole Earth System. This approach adds value to the observations—by filling in the spatio-temporal gaps in observations; and to the model—by constraining it with the observations" (Lahoz & Schneider, 2014). There are a number of different data assimilation approaches, including Kalman filter and variational methods and as statisticians we will be familiar with some of the approaches. There is still more to be done in this space, with challenging aspects including handling the deluge of data and in the uncertainties (or errors) that might exist.

2.4 | E and F inference and uncertainty

Both inference and uncertainty are a statistician's bread and butter. We understand the importance of considering uncertainty and in quantifying it. We appreciate that uncertainty comes in many layers of the digital earth, from both data and models. We use posterior distributions to deliver both quantification and visualization of the uncertainty. We are familiar with making assumptions and then trying to demonstrate their validity. The Oxford dictionary defines statistical inference as "the theory, methods, and practice of forming judgments about the parameters of a population and the reliability of statistical relationships, typically on the basis of random sampling." I draw your attention to the bold phrase, which of course brings us back to the core of how do we observe? In recent years, with the discussion about big data, and ubiquitous data collection, there has been much debate about the population, and sampling frame and whether indeed we need to consider this. The quote below reflects a rather divergent view to one that I hold.

"if organizations aren't already thinking about phasing out sampling and other "artifacts" of past best practices, they are behind the curve. Data science is inherently diminished if you continue to make the compromise of sampling when you could actually process all of the data, Sampling is an artifact of past best practices; it's time has passed" (http://www.computerweekly.com/feature/Big-data-analytics-and-the-end-of-sampling-as-we-know-it).

2.5 | G: Communication and visualization

Our last topic, is paraphrased as "one picture is worth 1000 words." Whether we write about visualization or data storytelling, I argue we would all agree this is an important aspect of all that we do, and that there are always new developments, new platforms to help us achieve our goals of making our work accessible, understandable and actionable.

"Visualization is the means by which humans understand complex analytics and is often the most crucial and overlooked step in the analytics process. As you increase the complexity of your data, the complexity of your final model increases as well, making effective communication and visualization of data even more difficult and critical to end users. Data visualization is the key to actionable insights" (http://dataconomy.com/2017/05/big-data-data-visualization/).

I do not think there is more to add.

3 | CONCLUSION- WHAT ARE THE CHALLENGES?

Environmental data science is in my view (and colored by my experiences) characterized by:

- dealing with time, space and space–time data and so living in a world of data misalignment and change of support,
- by data driven learning,
- by handling data of widely varying quality (so the need for feature engineering), and
- modeling complexity, with a common recurring theme of nonstationarity.

Earth observation provides a key data resource which is evolving in quantities of data being routinely collected: with resolutions now at m scale, compared to the historical km or 100 s of meters. New instruments are being developed (e.g., lab on a chip), but one challenge we have is how we maintain the consistency of data to allow long time series to develop and how do we fuse/integrate all these data streams together.

Experience has shown that sensor networks may be spatially sparse since it is expensive to manage and maintain a detailed network, leading, in conclusion, to big and sparse data.

Data quality assurance suddenly becomes more important than ever - there may be a number of anomalies, issues with power and so forth but we need automatic ways of flagging and perhaps even rectifying them.

So in summary, my characterization of environmental data science is: big (or at least large) data, complex and varied in structure and relationships, complex trends over space and time; sparse but big data, the need to integrate multiple data sources with different resolutions. Big models, or simpler coupled models with connections, nonlinearities and feedbacks describing a system. In the applied setting, we need to pay due respect to the applied science. And finally, uncertainty evaluation and visualization to aid communication are key.

My plea—There are many common threads running from statistics, data analytics and data science (these are the ties that bind us all to a common goal- making sense from data). While the main evolution in data and the data landscape is natural and exciting, let us not forget some of the fundamentals and basics about design and sampling, and let us not overlook the challenges in data quality assurance.

ACKNOWLEDGMENTS

I would like to acknowledge my colleagues, collaborators and students. Working with them has helped shape my views expressed in this short opinion piece. I also acknowledge research funding received from UKRI-NERC Feasibility Study–NE/T005564/1, NERC Constructing a Digital Environment Grant.

DATA AVAILABILITY STATEMENT

No Data associated

ORCID

E. Marian Scott https://orcid.org/0000-0002-3709-0623

REFERENCES

Blair, G. (2021). Digital twins of the natural environment. *Patterns*, *2*, 1–3.

Elayouty, A., Scott, E. M., & Claire Miller, C. (2022). Time-varying functional principal components for non-stationary $EpCO_2$ in freshwater systems. *Journal of Agricultural, Biological and Environmental Statistics*, *19*, 1–7.

Elayouty, A., Scott, M., Miller, C., Waldron, S., & Franco-Villoria, M. (2016). Challenges in modeling detailed and complex environmental data sets: A case study modeling the excess partial pressure of fluvial CO_2. *Environmental and Ecological Statistics*, *23*(1), 65–87.

Friedman, J. H., (2001). The role of Statistics in the Data Revolution. *ISI review*, *69*(1), 5–10.

Lahoz, W., & Schneider, P. (2014). Data assimilation: Making sense of earth observation. *Frontiers in Environmental Science*, *2*(16), 1–28.

Leclère, D., Obersteiner, M., Barrett, M., Butchart, S. H., Chaudhary, A., De Palma, A., DeClerck, F. A., Di Marco, M., Doelman, J. C., Dürauer, M., & Freeman, R. (2020). Bending the curve of terrestrial biodiversity needs an integrated strategy. *Nature*, *585*, 551–556.

Porter, J. H., Hanson, P. C., & Lin, C. (2011). Staying afloat in the sensor data deluge. *Trends in Ecology and Evolution*, *27*(2), 121–129.

JNCC. (2021). https://jncc.gov.uk/our-role/the-uk/nature-positive-2030/

Gong, M., O'Donnell, R., Miller, C., Scott, M., Simis, S., Groom, S., Tyler, A., Hunter, P., Spyrakos, E., Merchant, C., Maberly, S., & Carvalho, L. (2022). Adaptive smoothing to identify spatial structure in global lake ecological processes using satellite remote sensing data. *Spatial Statistics*, *42*, 100615. https://doi.org/10.1016/j.spasta.2022.100615) (In press).

Gong, M., Miller, C., O'Donnell, R., & Scott, M. (2021). State space functional principal component analysis to identify spatiotemporal patterns in remote sensing lake water quality. *Stochastic Environmental Research and Risk Assessment*, *35*(12), 2521–2536.

Maberly, S. C., O'Donnell, R. A., Woolway, R. I., Cutler, M. E. J., Gong, M., Jones, I. D., Merchant, C. J., Miller, C. A., Politi, E., Scott, E. M., Thackeray, S. J., & Tyler, A. N. (2020). Global lake thermal regions shift under climate change. *Nature Communications*, *11*, 1232.

Nativi, S., Mazzetti, P., & Craglia, M. (2021). Digital ecosystems for developing digital twins of the earth: The destination earth case. *Remote Sensing*, *13*, 2119.

Royal Society. (2020December). Digital technology and the planet: Harnessing computing to achieve net zero.

Scott, E. M. (2018). The role of statistics in the era of big data: Crucial, critical and under-valued. *Statistics & Probability Letters*, *136*, 20–24.

Steffen, W., Richardson, K., Rockstrom, J., Schellbhuber, H. J., Dube, O. P., Dutreuil, S., Lenton, T. M., & Lubchenco, J. (2020). The emergence and evolution of earth system science. *Nature Reviews Earth & Environment*, *1*, 54–63.

Wilkie, C. J., Miller, C. A., Scott, E. M., O'Donnell, R. A., Hunter, P. D., Spyrakos, E., & Tyler, A. N. (2019). Nonparametric statistical downscaling for the fusion of data of different spatiotemporal support. *Environmetrics*, *30*(3), e2549.

Wright, L., & Davidson, S. (2020). How to tell the difference between a model and a digital twin. *Advanced Modeling and Simulation in Engineering Sciences*, *7*(13), 7–13.

Webber, K., Pallas, F., & Ulbricht, M.-R. (2019). Challenges of citizen science: Commons, incentives, organizations, and regulations. *The American Journal of Bioethics*, *19*(8), 52–54.

How to cite this article: Scott, E. M. (2023). Framing data science, analytics and statistics around the digital earth concept. *Environmetrics*, *34*(2), e2732. https://doi.org/10.1002/env.2732