Received: 7 July 2022 | Revised: 17 November 2022 | Accepted: 10 January 2023

DOI: 10.1002/env.2789

OPINION

WILEY

The role of data science in environmental digital twins: In praise of the arrows

Gordon S. Blair | **Peter A. Henrys**

UK Centre for Ecology & Hydrology (UKCEH), Lancaster Environment Centre, Lancaster, UK

Correspondence
Gordon S. Blair, UK Centre for Ecology & Hydrology (UKCEH), Lancaster Environment Centre, Lancaster, UK.
Email: gblair@ceh.ac.uk

Funding information
Engineering and Physical Sciences Research Council, Grant/Award Numbers: EP/P002285/1, EP/R01860X/1; Natural Environment Research Council, Grant/Award Number: NE/R016429/1

Abstract

Digital twins are increasingly important in many domains, including for understanding and managing the natural environment. Digital twins of the natural environment are fueled by the unprecedented amounts of environmental data now available from a variety of sources from remote sensing to potentially dense deployment of earth-based sensors. Because of this, data science techniques inevitably have a crucial role to play in making sense of this complex, highly heterogeneous data. This short article reflects on the role of data science in digital twins of the natural environment, with particular attention on how resultant data models can work alongside the rich legacy of process models that exist in this domain. We seek to unpick the complex two-way relationship between data and process understanding. By focusing on the interactions, we end up with a template for digital twins that incorporates a rich, highly dynamic learning process with the potential to handle the complexities and emergent behaviors of this important area.

KEYWORDS

adaptive modeling, adaptive sampling, complex systems, digital twins, environmental data science, environmental modelling

1 | INTRODUCTION

The concept of digital twins initially emerged in the engineering domain to mean a digital or virtual representation of a physical artifact and one that is constantly updated to represent the current structure and behavior of that artifact (Blair, 2021). The concept of digital twins can also be applied to the natural environment, providing us with important new tools to understand and manage the natural environment in all its facets and at a variety of scales (Blair, 2021). For example, we can envision a digital twin for a river catchment looking at issues such as floods, droughts and water quality; we could equally look at developing digital twins for healthy soils operating at field/farm level, regionally or nationally and answering different land use questions, for example, around Net Zero ambitions; we could develop a digital twin for the whole earth system and interactions between the atmosphere, land, oceans and biodiversity under different climate scenarios. For some, the twins may operate in real-time or near real-time; for others, we are looking at supporting decision-making over potentially very long timescales. Ultimately, the digital twin should be reflective of the questions and scenarios that it will be used to answer.

Environmental digital twins are fueled by the unprecedented amounts of data available concerning the natural environment from a wide variety of sources including remote sensing data from satellites, aircraft and drones through to the potentially dense deployment of earth-based sensors. Citizen science data and data mining from the World Wide Web or historical archives also contribute significantly to the available environmental data. Finally, model output represents a significant generator of environmental data with results from previous model runs often stored for future analyses. Each of these sources of data may differ in many different ways (e.g., in their uncertainties or coverage and resolution across space or time), and the challenge is to overcome these differences to enable a coherent picture to be established.

Data science has an important role to play in making sense of the resultant highly complex data. Given the nature of the data, there is a need though for tailored data science techniques designed to deal with such complexities (Blair, Henrys, et al., 2019):

1. Considering the well known 4 'V's of big data, in many areas of data science volume and velocity are dominant concerns but, in this area, variety and veracity are often bigger concerns with there being a need to integrate highly heterogeneous data that may vary in terms of uncertainties associated with the measurements (and modeling);
2. Given the harsh environment in which this data is often measured, there is a need to deal with messy data, perhaps with large measurement error or with significant numbers of missing values;
3. There is a need to support reasoning around spatio-temporal data, including reasoning across scales;
4. Given the nature of environmental events and climate change, there is a need to develop techniques that fully address extremes and non-stationary properties.

Significant progress has been made in developing data science techniques for the natural environment and indeed this special issue is testimony to the advances that are being made. The toolbox is now well populated and includes techniques to investigate extremes, detect change points, separate trends from seasonality and noise and so on, often operating on multivariate, spatio-temporal data. Machine learning techniques are also increasingly being deployed. This includes both supervised learning that can be trained on data sets and unsupervised learning to detect hidden patterns, clusters or other structures in the input data. Hybrid approaches and deep learning approaches, that exploit computational resources, are also being investigated particularly in classification problems, for example, related to satellite data.

Clearly, data science has a central role to play in constructing digital twins of the natural environment. Indeed, for some, digital twins are all about data science: taking complex streaming data and using, for example, artificial intelligence techniques to make sense of this data and present it to stakeholders and decision-makers. This is not our view when it comes to environmental digital twins. For us, environmental digital twins are all about blending understanding and insight from both data models (i.e., statistical or AI-based models derived from the data) and process models (i.e., complex simulations based on an understanding of scientific processes and their interactions). We would go as far as to say that this relationship between data and process models is the defining characteristic of environmental digital twins. It is this blended approach that allows the twin to be an accurate representation of the current state, and also to react appropriately to given scenarios or perturbations. However, the exact nature of this two-way relationship is poorly understood and under-explored.

The overall aim of this short paper is to reflect on the modeling needs of environmental digital twins and, in particular, to unpack the two-way relationship between process and data models, and associated understanding. Our hope is that this will result in a richer, more sophisticated approach to environmental digital twins, informing underlying software architectures and also providing a roadmap of research challenges where some elements of the inter-relationships are poorly understood.

2 | THE RELATIONSHIPS BETWEEN PROCESS AND DATA MODELS

In our time in working with environmental scientists, we have gained a rich admiration of the work on environmental models, and process models in particular. Process models capture the current underlying scientific understanding encoded in mathematical models of the various underlying components and interactions. In many ways, scientific understanding evolves along with the associated models; that is, process models represent the best of breed of current scientific understanding. They reflect the state of the art knowledge of how a particular system may respond to

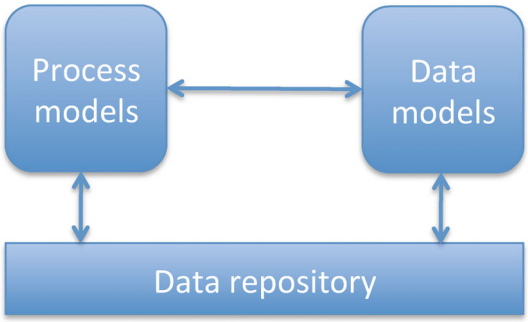


FIGURE 1 Outline interactions between process and data models

a given set of inputs. They are therefore fundamental to the science. This is why we are so adamant that you cannot replace process models with understanding derived from data alone: this would be detrimental to science, to scientific understanding and, indeed, to explainability of scientific phenomena. Having said that, process models are far from perfect:

1. Process models do not as yet fully take advantage of the emergence of big data as documented above;
2. They are often rather complex entities that take a long time to run, and this introduces limitations in terms of understanding sensitivities to different parameter options and to understanding uncertainties;
3. Historically, it has taken a long time for process models to be updated; for example, to update a component can take years of research and a period of agreement across the community.

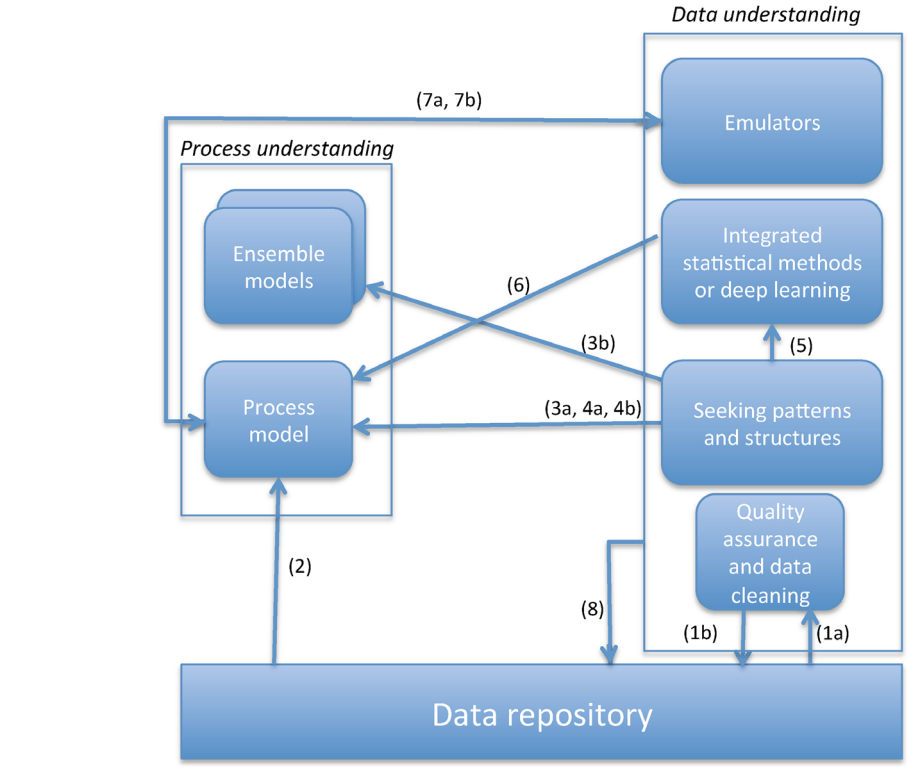
More philosophically, process model approaches represent a positivist approach to science where hypotheses are investigated through rigorously designed modeling experiments. In contrast, data driven models are often more lightweight (both computationally and structurally) and this means they can be run many times to determine sensitivities to different parameters and assumptions and to determine uncertainties. Data science can be used in a positivist manner but can also be used for more bottom-up, emergent discovery, for example for identifying potential patterns or structures that were previously unknown, whilst also explicitly representing stochastic elements within the system. It is therefore very natural to look at how we can take advantage of the potential synergies between both these modeling paradigms.

Figure 1 shows a very simple schematic representation of process and data models working together, also interacting with the underlying data repository. There are a number of arrows in this diagram – but what exactly do they mean?

We start off by recognizing that data models can be used for a variety of purposes. First of all, it is increasingly common to use statistical or machine learning techniques to carry out automated or semi-automated quality assurance on the underlying data and this is particularly important for digital twins where we may be dealing with real-time or near real-time streaming data. A first use of data models is therefore to *quality assure* and *clean* this data (1a) by identifying and potentially repairing anomalies/outliers and missing values, with the resultant cleaned data then going back into the repository (1b). This (potentially cleaned) data can then feed into process models through a process of *data assimilation* whereby the process model is nudged into a new state to be consistent with current observations (2). It may be that additional data models are required to feed the data into the process models at the appropriate scale to be assimilated. Data assimilation is effectively determining if the process model is consistent with current observations and, if not, nudging the model into a new state.

As mentioned above, data science techniques are very good at identifying patterns and structures in the data, for example potential clusters, correlations, extremes, and change points. We can now ask higher-order questions of the process model – is the process model able to represent these potentially significant patterns or events and if not, why not, providing an extra level of *model validation* (3a). This is particularly powerful when combined with an ensemble approach, allowing the weighting of different models to be altered depending on how well they represent emerging events, or removing a member from the ensemble, c.f. Beven's *model invalidation* (Beven & Lane, 2019) (3b).

Going further, the insights gained from observations and data models can lead to the concept of *adaptive modeling* more generally (4a). This may be limited in scope due to the often black box parameterization of the models but, given the complexity of the contemporary models and sheer range of parameters, this can in itself be hugely significant as model runs are optimized according to observation and insight. Looking forward, it is beneficial to imagine more open structures



Legend
1a: raw data;
1b: cleaned data;
2: data assimilation;
3a: model validation;
3b: model invalidation;
4a: adaptive modelling (black box);
4b: adaptive modelling (open architecture);
5: deep dive into data;
6: adaptive modelling (new behaviour);
7a: investigate sensitivities, uncertainties and interactions;
7b: further adaptive modelling;
8: adaptive sampling/monitoring.

FIGURE 2 Detailed interactions between process and data models, as described in Section 2

where models have an explicit software architecture consisting of components with their interactions and where this very *component structure* can be adapted to the same observations and insights (4b). This means that model implementations can be optimized to represent environmental behaviors as measured at a given place, and can change over time to represent what is currently being observed. From this, we are starting to see the relationship between process and data models resulting in modeling as a learning process, another of Beven's key tenets (Beven, 2007), and a key building block of models of everywhere (more completely referred to as models of everywhere and everything at all times (Blair, Beven, et al., 2019)). This may seem quite futuristic but such adaptive techniques have been employed successfully in other areas of complex software systems, for example in the field of autonomic/adaptive computing (Kephart & Chess, 2003; McKinley et al., 2004).

There are a several additional dimensions to this inter-relationship still to explore. The first is in response to the observations and insights where there is an indication of some scientific process or interaction that is not understood and, importantly, not captured by the process model. This may be investigated further (5) by utilizing contemporary data science methods that can exploit the strengths of different sources of data, whilst allowing and quantifying different error structures, to explore complex multivariate responses and patterns. Approaches such as integrated

spatio-temporal statistical modeling, or deep learning investigations (key themes of the DSNE Project mentioned below in the acknowledgements) are proving to have great potential in this area (Isaac et al., 2020; Wang et al., 2018; Wilkie et al., 2019). The intention is to dig deeper into the relationship between different variables and responses, potentially at fine spatial and temporal scales that may only be observed across multiple data sources, with a view to enhancing our understanding of the science and underlying processes. This, in turn, will ideally result in *enhancements* to process models (6), albeit at a different timescale to the adaptive techniques discussed above. Such steps though are important to enhance scientific understanding.

It has already been noted that process models are complex entities that can take a long time to run. Given this, it may be prohibitive to run such models many times and, yet, this may be important to fully understand their behaviors and sensitivities to different parameter settings, and also to appreciate the uncertainty space associated with different executions and assumptions. For the same reasons, it may be impractical to look at coupled modeling where the output from one model may provide inputs to another, perhaps in arbitrary patterns and configurations. For both these reasons, model emulators are often generated from perceived behaviors of process models, taking an abstract, data driven view of the process model by mapping inputs to outputs, often using machine learning or statistical approaches such as Gaussian processes. Emulators, that are far quicker to execute than their numerical counterparts, can then be used to *understand sensitivities and uncertainties*, and *interactions* in integrated modeling experiments (7a). This may in turn result in *changes in how process models are configured and parameterized* (7b).

Another advantage of data models is that they may have an intrinsic understanding of the uncertainties associated with the data, parameters, inter-variable relationships, and those associated with space and time. This in turn can drive *adaptive sampling* or monitoring (8) approaches whereby further data elements can be specifically targeted to efficiently reduce uncertainties (Phillipson et al., 2019). These adaptive approaches can take many forms, from dynamic control of sensor networks to provide more data at a given time and at different locations, through to optimal design of field surveys or citizen science data gathering activities, perhaps constrained by other factors such as cost. The overall aim, here, is to optimize the information obtained to minimize uncertainty.

This builds up to a much richer presentation of the interactions between process and data models as shown in Figure 2. It is interesting to note that with the final elements we complete the circle back to the data, representing a rich and sophisticated dynamic learning process – effectively our definition of what is a digital twin.

3 | IN PRAISE OF ARROWS

Writing this paper has been a fascinating exercise, as we have deliberately focused not on the entities of process and data models, but on the interrelationships between them. It is all too easy to draw diagrams with boxes and arrows but, often, the complexities of the arrows are over-looked in contrast with the deep investigation that is carried out inside the boxes. Hopefully, this exercise has demonstrated the importance of arrows in understanding the architecture of digital twins and the importance of interactions in realizing the full potential of this exciting technology. In particular, we have gone from a relatively static view of digital twins to a fully dynamic learning process that can capture the nuances of complex environmental behavior at different spatial locations and across time. Let us give much more attention to 'arrows' as they often hold the key to understanding complex systems, their interactions, and emergent behavior.

ACKNOWLEDGMENTS

This work was strongly influenced by discussion in the Centre of Excellence in Environmental Data Science (CEEDS), a joint initiative involving the UK Centre for Ecology & Hydrology (UKCEH) and Lancaster University. We thank the many members of CEEDS who have contributed to such vibrant and thoughtful discussions around digital twins of the natural environment. This work was partially supported by the following grants: (i) a grants from the Engineering and Physical Sciences Research Council in the UK (EPSRC) on Data Science of the Natural Environment (EP/R01860X/1); (ii) an EPSRC Senior Fellowship (awarded to Blair) in the Role of Digital Technology in Understanding, Mitigating and Adapting to Environmental Change (EP/P002285/1); (iii) a National Capability grant from the Natural Environment Research Council in the UK (NERC) funding the UK-SCAPE program (UK Status, Change and Projections of the Environment) (NE/R016429/1); and (iv) a further NERC grant investigating Information Management Framework for Environmental Digital Twins (IMFe).

ORCID

Gordon S. Blair https://orcid.org/0000-0001-6212-1906
Peter A. Henrys https://orcid.org/0000-0003-4758-1482

REFERENCES

Beven, K. (2007). Towards integrated environmental models of everywhere: Uncertainty, data and modelling as a learning process. *Hydrology and Earth System Sciences*, *11*, 460–467. https://doi.org/10.5194/hess-11-460-2007

Beven, K. J., & Lane, S. (2019). *Invalidation of models and fitness-for-purpose: A rejectionist approach, chapter 6*. In C. Beisbart & N. J. Saam (Eds.), *Computer simulation validation - fundamental concepts, methodological frameworks, and philosophical perspectives* (pp. 145–171). Springer.

Blair, G. S. (2021). Digital twins of the natural environment. *Patterns*, *2*, 10. https://doi.org/10.1016/j.patter.2021.100359

Blair, G. S., Beven, K., Lamb, R., Bassett, R., Cauwenberghs, K., Hankin, B., Dean, G., Hunter, N., Edwards, L., Nundloll, V., Samreen, F., Simm, W., & Towe, R. (2019). Models of everywhere revisited: A technological perspective. *Environmental Modelling & Software*, *122*, 104521. https://doi.org/10.1016/j.envsoft.2019.104521

Blair, G. S., Henrys, P. A., Leeson, A., Watkins, J., Eastoe, E., Jarvis, S., & Young, P. (2019). Data science of the natural environment: A research roadmap. *Frontiers in Environmental Science*, *7*. https://doi.org/10.3389/fenvs.2019.00121

Isaac, N. J., Jarzyna, M. A., Keil, P., Dambly, L. I., Boersch-Supan, P. H., Browning, E., Freeman, S. N., Golding, N., Guillera-Arroita, G., Henrys, P. A., & Jarvis, S. (2020). Data integration for large-scale models of species distributions. *Trends in Ecology & Evolution*, *35*(1), 56–67.

Kephart, J. O., & Chess, D. M. (2003). The vision of autonomic computing. *Computer*, *36*(1), 41–50. https://doi.org/10.1109/MC.2003.1160055

McKinley, P. K., Sadjadi, S. M., Kasten, E. P., & Cheng, B. H. C. (2004). Composing adaptive software. *Computer*, *37*(7), 56–64. https://doi.org/10.1109/MC.2004.48

Phillipson, J., Blair, G. S., & Henrys, P. (2019). *Uncertainty quantification in classification problems: A Bayesian approach for predicating the effects of further test sampling*. In S. Elsawah (Ed.), *MODSIM2019, 23rd international congress on modelling and simulation*. Society of Australia and New Zealand. https://doi.org/10.36334/modsim.2019.B1.phillipson

Wang, C., Puhan, M. A., Furrer, R., & SNC Study Group. (2018). Generalized spatial fusion model framework for joint analysis of point and areal data. *Spatial Statistics*, *23*, 72–90.

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.

How to cite this article: Blair, G. S., & Henrys, P. A. (2023). The role of data science in environmental digital twins: In praise of the arrows. *Environmetrics*, *34*(2), e2789. https://doi.org/10.1002/env.2789