A digital twin of Earth for the green transition

For its green transition, the EU plans to fund the development of digital twins of Earth. For these twins to be more than big data atlases, they must create a qualitatively new Earth system simulation and observation capability using a methodological framework responsible for exceptional advances in numerical weather prediction.

Peter Bauer, Bjorn Stevens and Wilco Hazeleger

The European Union (EU) intends to become climate neutral by 2050, and the set of policies designed to bring about this green transition — the European Green Deal — was announced in December 2019 (ref. [1]). Accompanied by €1 trillion of planned investment, Green Deal policies aim to help the world's second-largest economy sustainably produce energy, develop carbon-neutral fuels and advance circular products in energy-intensive industrial sectors with zero waste and zero pollution.

A key element of the Green Deal is its dependence on the 'digital transformation' — an openly accessible and interoperable European dataspace as a central hub for informed decision making. The EU identified two landmark actions to support the necessary information systems: GreenData4All[2] and Destination Earth[3]. Whereas GreenData4All will develop the European approach to discover, manage and exploit geospatial information, Destination Earth aims to construct highly accurate models, or 'digital twins', of the Earth to monitor and predict environmental change and human impact in support of sustainable development. Aligned with the new Digital Europe funding programme[4], Destination Earth is expected to start in 2021, and the first, high-priority digital twins serving extremes prediction and climate change adaptation will start production in 2023 (ref. [5]).

Credit: Map of Layerace / Freepik

The timing couldn't be better. Advances in high-performance computing have reached a point that now make it possible to model the Earth system with much greater physical and spatial fidelity[6]. By combining this new class of models with advances from past investments in Earth system prediction and observations[7], digital twins promise to close substantial and recalcitrant gaps in our ability to look into the future. These gaps presently hinder reliable projection of climate change signals at regional scales[8,9], undermining confidence in assessments of the vulnerability of human welfare and capital to present and future weather extremes and catastrophic climatic shifts. Digital twins can and must close these gaps, but to do so will require coordinated development across scientific disciplines.

Digital twin

A digital twin of Earth is an information system that exposes users to a digital replication of the state and temporal evolution of the Earth system constrained by available observations and the laws of physics.

While we are familiar with a plethora of observation-based monitoring tools that document our impact on the environment, only physics-based simulation models can help us to grasp the causes of change and explore options for future adaptation and mitigation actions. The ongoing step change in the physical content of Earth system models is making them amenable to approaches that harmonize the physical laws they encode with ever more extensive observations to provide the best possible estimate of the state of our planet. Hence, digital twins must focus exactly on how best to realize this convergence of the modelling and observation worlds.

A methodological framework for the twin's architecture already exists in the form of data assimilation, which numerical weather prediction has developed with success over decades[10]. Data assimilation combines data from different observational sources with physical Earth system model simulations to derive an estimate of the state

Box 1 | Digital twin architecture

The concept of a digital twin comes from industrial production and space technology engineering, and aims to optimize design and operations of complex processes through a highly interconnected workflow[21]. A digital twin of Earth would fully integrate observations with an Earth system model and human subsystems for, for example, water, food and energy resource management, to assess the impacts on, and influences from, these subsystems on Earth system trajectories. The twin would allow us to assess possible changes and their causes consistently across local and global spatial scales and over timescales stretching from days to decades.

The twins combine a physical model with observations using a data assimilation framework that is based on the foundations of information theory and high-dimensional optimization problems. Beyond its daily use in weather prediction, it offers an objective approach to produce a realistic replica of a highly nonlinear system with many degrees of freedom. The optimum combination of observations with model simulations allows one compensating for gaps in the other: not all variables are observed everywhere all the time, and models contain approximations for unknown and spatially unresolved processes; the models fill the observational gaps in space and time in a physically consistent way and observations help constrain model approximations. Full traceability of errors in all of these components is built into data assimilation, so better observations and models will receive a stronger weight and the state estimation will return higher-quality output.

The twin architecture is a step change in Earth system modelling because:

- It combines simulations and observations at much greater spatial (km-scale globally, hm-scale regionally) and thereby physical realism;
- It can both monitor and predict, even on multi-decadal time scales, to account for natural (for example, volcanos) and human (for example, land-use change, emissions) perturbations;
- It can operationalize the assessment of model process errors through weather prediction techniques and constrain missing physics through observations and machine learning;
- It can evaluate the information content of observations and help optimize entire observation networks.

The digital twin architecture sits much closer to the user because:

- It provides a framework for optimization, whether that be of disaster prevention or adaptation strategies; this can be done at scale and across sectors — from food and energy security to coastal protection;
- It gives non-experts full access to data and data analytics toolkits using a scalable cloud-based infrastructure;
- It represents high-impact events at scale and in terms of quantities that users are already familiar with.

Digital twins, envisioned in the figure as a federated infrastructure for both data production and user access, are accompanied by interactive tools to allow users to directly intervene and perturb the digital twin workflow. The twin can guide reinsurance companies in quantifying their exposure to risks from extreme storms in different climate scenarios, help foresters to assess different management strategies and allow farmers to weight the impact of different agriculture policies.

It could also allow an assessment of measures intended to adapt to future extreme weather events or an identification of the types of human perturbations more likely to lead to catastrophic changes in regional subsystems. Adaptation measures could be assessed using downscaling to urban infrastructure scales to quantify the impact of city planning on the local effects caused by heatwaves or inundations.

of a system. Starting with a model prediction of what the state could be, the Earth system model acts as a physically consistent integrator of observational information in four dimensions, namely space and time. This framework has many similarities with machine learning and therefore offers numerous opportunities for performance upgrades arising from this exciting new discipline. Applying this framework to climate change poses new challenges, for instance through control-theoretic approaches to scenario construction or in parameter optimization.

Building on data assimilation, a digital twin must also deliver an interface and development toolkit that makes it easy to configure and efficient to use, and that can execute all of its tasks in a highly efficient virtual environment. The origins and novel aspects of digital twins for the Earth system are summarized in Box 1.

Realizing this vision will require two major breakthroughs in today's information systems: one for information completeness and quality, and another for information access and intervention.

A leap in information quality

To produce a new quality of information, a digital twin of Earth must reshape the classical data assimilation framework by (i) creating systems capable of solving much larger problems; (ii) incorporating human systems, also as part of the prediction problem; and (iii) benefitting from the synergy of advances in information theory and digital information technology.

First and foremost, a digital twin would become a data assimilation instrument that continuously cycles real-time, highly detailed, high-resolution Earth system simulations and ingests observational information from all possible instruments — including novel observatories like miniaturized satellites[11], drones in the Arctic[12], undersea cables and buoys[13], smart sensor arrays in crop fields[14] and mobile phones within the expanding internet of things[15] — also to estimate uncertain model parameters and surrogate missing process details. The resulting Earth system state simulation will be more realistic and lend itself to more reliable predictions.

Second, a digital twin must fully integrate the human dimension of the Earth system and allow activities or policies to be optimized by the system, subject to new benefit norms like socioeconomic loss or sustainability. Human impacts such as greenhouse gas emissions, pollution, land-cover change or water management need to be added at the relevant scales. Including, for example, an economic model, models for solar and wind power, and models for agriculture management in the twin would allow for the addition of observed data from these spheres and produce the best available twin for Earth in the Anthropocene. This twin would produce more reliable predictions and allow fully multivariate testing of scenarios: what is the economic impact of adding wind power plants in an area once Europe is carbon neutral? What agricultural policy is the most sustainable given future weather regime pattern changes?

Third, the systems must be agile in ways that allow them to benefit from advances in information theory whereby unknown processes and missing constraints are filled in by data-driven learning[16]. Exploiting the full range of edge, cloud and high-performance computing, smart sensors and smart data communication networks[17] offers entirely new efficiencies but has not been used in synergy to its full potential in this context. Generic software infrastructures would serve all applications. Computing and data-handling tasks should be handled by a federated infrastructure covering the range from data extraction and analytics applied close to the user to extreme-scale data assimilation and simulation tasks run on supercomputers.

The end-user and application of such a digital twin would include national governments to understand future climate change impacts and scenarios on food security, water resources and geopolitics; energy providers to design deep mantle geothermal energy resourcing; and the reinsurance industry to test risk mitigation and disaster management.

A leap in information intervention

An exciting aspect of a digital twin is the potential to break the paradigm of classical Earth system prediction models with fixed and static flows of information managed by layers of experts. Here, the challenge will be to design a digital twin that allows users to intervene, extract information and influence the system trajectory across time and space, as done — albeit often unwittingly — in the real world.

Enhancing information quality requires a step change in computational complexity, in part from the need to incorporate a much broader range of digital technologies to support data production and information extraction. At the same time, it needs to hide this complexity from the user so that users can run and configure complex workflows and access the information in ways that don't require expert intervention. The solution will require a multi-layered software infrastructure that is based on abstractions with a data assimilation framework at its core, and where tasks like simulations, observational data ingestion, post-processing and so on are treated as objects that are executed on federated computing infrastructures, feed data into virtual data repositories with standardized metadata, and from which a heavily machine-learning-based toolkit extracts information that can be manipulated in any possible way.

This approach turns the classical, linear value chain concept — whereby data move from primary sources to users with selective value-adding applied on the way — inside out. Data assimilation methods offer this flexibility: they can estimate the response of the system based on the priorities and choices of the user, while the user can increase or decrease the twin's complexity, add new observations and try new tools to explore possible futures, test trial intervention scenarios and even assess uncertainty by probing well-observed past periods. The flexibility paradigm will prove indispensable for a system that aspires to allow users to interact with and experience the system's sensitivities and fallibilities, and thereby better inform decision making[18]. It also provides a basis for entirely new application communities to innovate around their information content in ways we can't yet even imagine.

Digital twins as substrates

Destination Earth, and its idea of digital twins, foresees the scale of effort required to deliver the knowledge base needed to manage the green transition, also offering a model for others to follow. For this to be successful, a substantial part of the investment needs to be made in the multi-layered software infrastructure. Only then will it achieve the necessary breakthrough in information quality and information intervention. Such an infrastructure must achieve very high computing throughput rates of its physical models, support concurrency in the execution of impact models and be capable of processing billions of observations per day. This needs a radical redesign to provide: (i) options for performance optimization and on-the-fly auto-tuning to achieve the best time and energy to solution on future supercomputers; (ii) a workflow capable of integrating machine-learning techniques to accelerate computing, compress data and extract sector-specific information; and (iii), application interfaces that allow a diversity of users to easily and intuitively apply their expertise[19]. The software will not avoid the need for large investments in extreme-scale hardware. Extreme-scale computing creates extreme-scale data, so both aspects need to be developed together, but with machine learning as a bridge between data and information, it should, however, be possible to reduce the hardware footprint.

Such a digital twin approach touches on all aspects of the present workflow for producing Earth system information. With contemporary limits to computational progress[20], this will only be achieved through international collaboration among Earth system scientists, users from impact sectors, economics, computer scientists and industry. Once achieved, the twin will be the catalyst for yet unprecedented scientific and technological innovation because it can do so much more while being much easier to use. While beyond the scope of this Comment, we fully acknowledge that including the human dimension means including non-technocratic approaches from social sciences to ensure that the end-user uptake becomes reality and will be effective.

We estimate the level of effort to be significant and to cut across the Earth system, application, social and computational sciences. Firstly, Earth system simulation and data assimilation systems need to be advanced to more physically represent the system's state and evolution at much finer spatial resolution, seamlessly from days (weather) to decades (climate), at a high rate of throughput. Secondly, Earth system science must be fused with the entire range of impact sectors, also to include new and disparate data sources, as well as to translate scientific data into decision making. Thirdly, achieving optimum computing and data handling performance while allowing user interaction with data and machine-learning tools made accessible through cloud infrastructures will require sustained coordination.

These efforts will, however, not be realized through a patchwork of funding carried into production by individual groups or even labs. The European Green Deal and funding allocated for Destination Earth are compelling because they represent a concerted effort at the necessary scale, and because the Earth science community is well organized and increasingly cognizant of how its efforts must be directed to provide an adequate scientific basis for decision making in the Anthropocene. ☐

Peter Bauer[1]✉, Bjorn Stevens[2] and Wilco Hazeleger[3]

[1]*Research Department, European Centre for Medium-Range Weather Forecasts, Reading, UK.* [2]*Max-Planck-Institute for Meteorology, Hamburg, Germany.* [3]*Faculty of Geosciences, Utrecht University, Utrecht, the Netherlands.*

✉*e-mail: peter.bauer@ecmwf.int*

Published online: 1 February 2021
https://doi.org/10.1038/s41558-021-00986-y

References

1. *The European Green Deal* (European Commission, 2019).
2. *A European Strategy for Data* (European Commission, 2020).
3. Destination Earth (DestinE). *European Commission* https://ec.europa.eu/digital-single-market/en/destination-earth-destine (2020).
4. Europe investing in digital: the Digital Europe Programme. *European Commission* https://ec.europa.eu/digital-single-market/en/europe-investing-digital-digital-europe-programme (2020).

5. Voosen, P. *Science* https://doi.org/10.1126/science.abf0687 (2020).
6. Satoh, M. et al. *Curr. Clim. Change Rep.* **5**, 172–184 (2019).
7. Bauer, P., Thorpe, A. & Brunet, G. *Nature* **525**, 47–55 (2015).
8. Shepherd, T. *Nat. Geosci.* **7**, 703–708 (2014).
9. Palmer, T. & Stevens, B. *Proc. Natl Acad. Sci. USA* **116**, 24390 (2019).
10. Blayo, E., Bocquet, M., Cosme, E. & Cugliandolo, L. F. (eds) *Advanced Data Assimilation for Geosciences* (Oxford Univ. Press, 2014).
11. Stephens, G. L. et al. *IEEE T. Geosci. Remote* **58**, 4–13 (2019).
12. Sheridan, I. *Roy. Soc. Open Sci.* **7**, 191494 (2020).
13. Levin, L. A. et al. *Front. Mar. Sci.* **6**, 241 (2019).
14. Saiz-Rubio, V. & Rovira-Más, F. *Agronomy* **10**, 207 (2020).
15. Marjani, M. et al. *IEEE Access* **5**, 5247–5261 (2017).
16. Reichstein, M. et al. *Nature* **566**, 195–204 (2019).
17. *ETP4HPC's SRA-4* (European Technology Platform for High-Performance Computing, 2020).
18. Hallegatte, S. *Glob. Environ. Change* **19**, 240–247 (2009).
19. Bauer, P. et al. *Nat. Comput. Sci.* (in the press).
20. Khan, H. N. et al. *Nat. Electron.* **1**, 14–21 (2018).
21. Tao, F. & Qi, Q. *Nature* **573**, 490–491 (2019).

Acknowledgements

We thank Y. Schrader for her graphical assistance.

Author contributions

P.B. conceived the Comment and P.B., B.S. and W.H. contributed to the writing and revision of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Peer review information *Nature Climate Change* thanks Antje Weisheimer and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.