



WADITI - Developing a roadmap for a water quality digital twin

A scoping study under UK-SCAPE

Virginie Keller, Jan Dick, Anne Dobel, Alex Elliott, Matt Fry, Sam Harrison, Doran Khamis, Ezra Kitson, Yueming Qu, Gordon Blair & Steve Thackeray



UK Centre for
Ecology & Hydrology

04.07.2024

Contents

1.	Background	4
2.	Framing the requirements of a digital twin	7
2.1	Approach to identifying stakeholder needs	7
2.2	Analysis of stakeholder needs.....	8
	Feasible spatial scale.....	8
	Feasible temporal scale	8
	Most important determinands.....	8
	Output variables	9
	Output format	9
	Major obstacles	9
3.	Review of existing data and models.....	10
3.1	Review of existing data	10
	Manual sampling	11
	Sonde measurements	15
3.2	Review of existing models.....	18
3.3	Integrating data and models through data science methodologies	21
4.	Roadmap.....	25
4.1	Cyber-physical infrastructure.....	25
	Enabling FAIRness through standards, interfaces, conventions, and semantics ...	25
	Modularisation of assets into interoperable components	27
	Cloud based cyber infrastructure	28
4.2	Graphical interfaces, portals, and visualisations	29
4.3	Sustainability and long-term maintenance	30
5.	Case studies	31
5.1	Case study 1: FASE.....	32
5.2	Case study 2: nutrient and phytoplankton dynamics in rivers and lakes	36
	Benefits of digital twinning for PROTECH, QUESTOR and LAM	38
6.	Conclusions & Recommendations	41
	Acknowledgements	43



References 43

1. Background

Fresh waters supply several essential services to society, such as water supply for various sectors (including drinking water, agriculture, manufacturing), the leisure industry (e.g. bathing water), and habitats for fauna and flora (Millenium Ecosystem Assessment, 2005). Indeed, they make a disproportionately high contribution to global biodiversity; despite accounting for only 2.3% of the Earth's land surface area, fresh waters are estimated to host almost 10% of described animal species (Reid et al., 2019 and references therein).

Despite their overwhelming importance, fresh waters around the world are under increasing pressure due to the interacting effects of climate change, pollution, overexploitation, and socio-economic change (Dudgeon et al., 2006; Tickner et al., 2020). As a result, in Europe, only 40% of surface water bodies are achieving good ecological status, as required under the Water Framework Directive (European Environment Agency, 2018). In the UK, this figure drops down to only 14% (House of Commons, 2022). Compounding these trends, the research community continues to recognise further emerging threats to fresh waters, for which we have only limited insight (Reid et al., 2019; Stephenson et al., 2024). The impacts of this diverse array of pressures are challenging to predict, given that they may interact in complex ways (Jackson et al., 2016; Spears et al., 2021). It is therefore vitally important that we increase our understanding of freshwater ecosystem dynamics under current conditions and apply this enhanced understanding to make projections of likely future change and scenarios. Armed with such understanding, we would greatly facilitate adaptive management of freshwater resources and biodiversity.

Over recent decades, scientists across the UK, and internationally, have collected a wealth of environmental data using an ever-increasing array of approaches and technologies (Blair et al., 2019a; Thackeray & Hampton, 2020; Blair & Henrys, 2023). Long-term and high-frequency monitoring networks have been established and maintained (e.g. the Global Lake Ecological Observatory Network), gathering invaluable evidence on the changing state of fresh waters over time. In addition, observations are being made via remote sensing, professional survey, citizen science, and through the application of novel analytical techniques (e.g. molecular approaches, acoustics) (McCracken et al., 2024). Notwithstanding important challenges around the volumes, veracity, and heterogeneity of these collective measurements and observations (Blair et al., 2019a; Blair & Henrys, 2023), such data can be integrated through empirical modelling approaches and data science techniques to yield new insights into processes and states in fresh waters (Jarvis et al., 2023).

Environmental data are also essential to inform, test, and drive process-based models. Process-based models are valuable tools in ecological research and



management, providing digital representations of real-world processes that allow the investigation of impact scenarios in “virtual experiments” (e.g. Elliott et al., 2006; Hutchins 2012; Salk et al., 2022). However, such models also have recognised limitations. Typically, models show great variety (Blair et al., 2019b). Each model only captures specific aspects of the freshwater environment (e.g. rivers or lakes, droughts or floods, hydrology or water quality), or focuses only on particular aspects of environmental stress (e.g. point vs. diffuse sources of pollutants, macronutrients, metals). Furthermore, the structure and parametrisation of process-based models is typically static in time, limiting their ability to capture the constantly changing nature of the environment.

The use of data for the single purpose of process-based modelling (driving data, calibration, validation) is not optimal given that the wealth of data available to scientists contains invaluable information, outside of the process-modelling arena. New data science technologies can also unlock new knowledge from this information and facilitate efforts towards increasing the resilience of freshwater ecosystems. As such, data and model integration provide opportunities to both advance our understanding of fresh waters, and to build predictive capabilities that can inform conservation and management.

Over recent years, scientists have explored a new technological paradigm, the digital twin concept, to provide improved modelling capabilities that would enable more accurate forecasting, for potential use in decision making (Blair, 2021). The commonly accepted definition of a digital twin is a virtual representation of a system that is constantly updated to accurately represent the current state and behaviour of the system. In the environmental realm, the digital twin concept can be interpreted as a system that allows the integration of the plethora of information and technology available to scientists including monitoring observations, remote sensing data, process-based and data-driven models (Blair & Henrys, 2023). Crucially, digital twins include feedbacks on the way that we interact with the real environment (Siddorn et al., 2022), and can transform modelling into a learning process through the integration of incoming data (Blair & Henrys, 2023). Although digital twinning has a longer history in engineering (Rasheed et al., 2020; Blair, 2021) and urban drainage systems/utilities (Karmous-Edwards et al., 2019), recent developments are seeking to apply this approach in ways relevant to freshwater ecosystems. Examples include digital tools to address issues such as harmful algal blooms (Qiu et al., 2023), contaminant transport (Kim & Bartos, 2024), and water quality (Chen et al., 2023, Qiu et al., 2022, FLARE <https://flare-forecast.org/>, WaterWebTools <https://www.waterwebtools.com/>).

In this study, we develop a roadmap for a surface water quality digital twin, to provide now-casting and short-term forecasting of a range of environmental pollutants and ecosystem states in both rivers and lakes. Although shorter-term dynamics are our focus, we note that digital twins can be applied to longer-term system behaviour as well. Our aim was to integrate numerous aspects of water



WADITI - Developing a roadmap for a water quality digital twin |

quality and UKCEH expertise in monitoring networks, data science, hydrological and water quality, and river and lake ecosystem modelling. Thus, we reviewed current data streams, available models, and data science approaches. We built upon the existing body of work that considers how best to conceptualise environmental digital twins. This included knowledge acquired through:

- development of digital twins or similar systems under Land Insight (Fry et al., 2022) and UNIFHY (Hallouin et al., 2022, <https://unifyh.org.github.io/unifyh/index.html>),
- published scientific recommendations regarding the development of environmental digital twins (Siddorn et al., 2022; Blair, 2021), and the need to consult widely when delivering decision-grade knowledge (Chambers et al., 2021)
- external networks currently focussing on environmental digital twins, such as the NERC Constructing a Digital Environment Expert Network.

When designing and developing such tools, users' needs are an essential consideration. To maximise the benefits, and uptake, of the digital twin application by the wider community, we identified the needs of stakeholders and analysed their responses.

The WADITI project was funded as part of the National Capability programme UK-SCAPE (<https://uk-scape.ceh.ac.uk/>), Year 6. Aligned with the UKCEH Digital Strategy 2024-2030 (<https://www.ceh.ac.uk/about-us/digital-strategy>), the project is comprised of three inter-linked work packages (WP) (Figure 1), bringing together stakeholder community needs with in-depth understanding of digital twin architectures, data, and models.



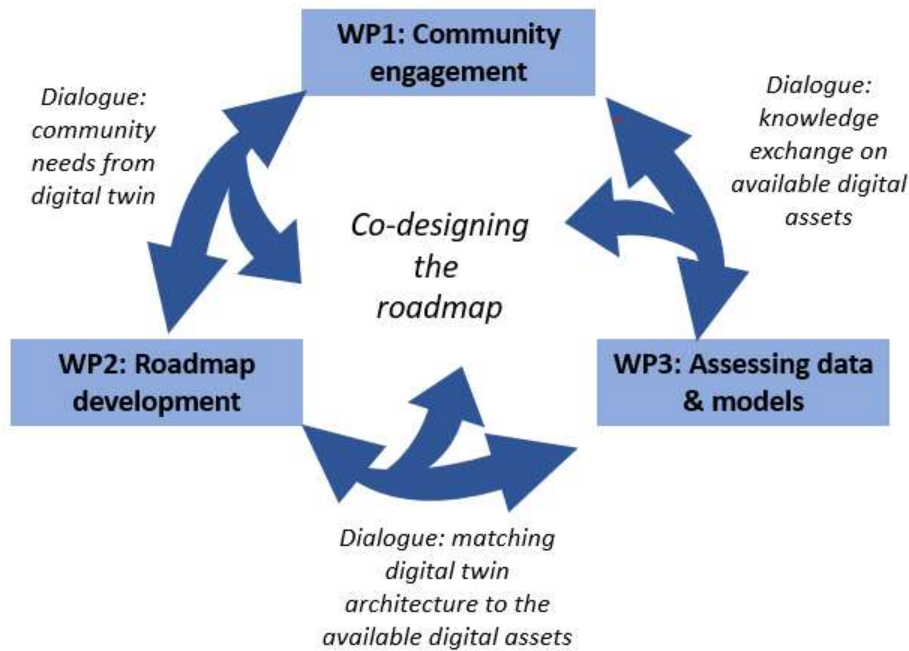


Figure 1. The WADITI Work Package (WP) structure.

This report presents the outcomes of the scoping study on the development of a water quality digital twin. The stakeholder requirements identified via WP1, “Community engagement”, are summarised in Section 2. A review of existing data and models is presented in Section 3. The proposed water quality digital twin roadmap is developed in Section 4, including case study examples. Finally, conclusions and recommendations are laid out in Section 6.

2. Framing the requirements of a digital twin

2.1 Approach to identifying stakeholder needs

Stakeholders’ views on the creation of a water quality digital twin were sought through an anonymous survey composed of closed and open questions. Full descriptions of the questionnaire and of stakeholders’ views are reported in Dick et al. (2023).

The questionnaire was designed by all members of the WADITI team to capture opinions on fundamental considerations when creating a roadmap for a water quality digital twin, including:

- Spatial and temporal scale,
- Determinands to consider,



- Outputs encompassing both content and format,
- Foreseen risks and limitations.

The questionnaire was distributed UK wide, using a snowball approach whereby the team encouraged respondents to share it with peers that might be interested in the topic, i.e., invitees invite others. The responses were delivered June/July 2023. On average the participants took ~20 min to complete the survey (see Dick et al. (2023) Annex 1 for full questionnaire).

2.2 Analysis of stakeholder needs

Fifty-nine participants responded and provided a wealth of viewpoints, from the perspectives of academia (~40%), industry (~25%), regulators (~15%), NGO's (~15%) and policy makers (~5%). The responses were fully analysed to identify stakeholder needs and are summarised here.

Feasible spatial scale

The majority (75%) of respondents who expressed a single scale preference considered that 'catchment scale' was the most feasible, useful, realistic, and deliverable scale for a water quality digital twin. The foremost rationales for preferring the catchment scale reflected that it was "*the usable spatial scale*" and "*more achievable to deliver*", reflecting the "*specific nature of each catchment*" and "*provides the highest resolution of data/information upon which to act*". In addition, respondents considered that "*Catchment-scale mirrors the currently active river basin management planning process*" and it was recognised that "*catchments often transcend national or even regional boundaries*", noting "*management that doesn't recognise this is insufficient*". Respondents who selected national as the single most relevant scale tended to focus on policy objectives while the regional scale was considered by some as a "*middle ground*".

Feasible temporal scale

There was no consensus on the most suitable temporal scale to deliver output from a water quality digital twin. Respondents considered the desired temporal scale to be dependent on the use envisaged for the digital twin. However, overall, 76% of respondents marked 'sub-daily plus daily' alone or in combination with other temporal scales as the most suitable temporal scale to deliver output from a water quality digital twin. This lack of agreement on temporal scale appears to result from respondents' concerns on (i) the use of the digital twin output, (ii) computational feasibility of running detailed digital twins, and (iii) availability of data at a suitable temporal scale.

Most important determinands

Nutrient concentrations were considered the most important determinands of running and standing water quality to include in a water quality digital twin, with



100% of respondents scoring these as a priority. The linkage between nutrients and cyanobacteria was noted by several respondents e.g. *“Algal blooms (and therefore nutrient concentrations) are of greatest concern to the general public so it would be most useful to be able to forecast these to inform management on the ground”*. Several other abiotic and biotic determinands were also suggested, partly depending on the required use of the resultant digital twin. Cyanobacteria, pesticides/herbicides and heavy metals were all selected by over 80% of respondents as being important-to-critically important to be included in a water quality digital twin.

Output variables

A combination of physico-chemical variables (selected by 52% of respondents) and biological variables (selected by 24% of respondents) were considered the most important outputs that a water quality digital twin should predict. However, respondent comments revealed the interdependencies between physico-chemical and biological variables e.g. *“I would have selected physico-chemical and biological variables ... as I believe the two go hand-in-hand. However, I would prioritise physico-chemical variables as prediction of nutrient concentrations, temperature, and dissolved oxygen concentration will allow better management of algal blooms and water for other biodiversity”*. Human wellbeing and financial impact as output variables were considered very important for decision making, they were however, recognised as more difficult to model.

Output format

Most respondents selected more than one output format (76%). An interactive web portal with visualisations of predictions was selected by 92% of respondents (often in conjunction with other output formats). It is clear from the output format selected and the accompanying written rationale that some respondents wished to have the data interpreted for them while others wanted access to the data so they could conduct further analysis themselves. For example, one respondent wrote *“Relevant knowledge that has been derived from the digital twin - don't make people have to work it out themselves...translated insights in an accessible way that doesn't involve more analysis”*, compared to another who wrote *“Really just worth having an API, everything else can build on top of that”*.

Major obstacles

The need for (near) real time data on both water quality and quantity was considered by most respondents to be a major obstacle to creating a fully functional digital twin of running and standing water quality. However, almost twice as many respondents considered that this situation was worse in terms of water quality compared to water quantity e.g. *“Water quantity is better quantified but needs to be linked to WQ [water quality], including monitoring at same locations”*. Information on water quality was acknowledged by several respondents as limiting, e.g. *“I consider the lack of near-real-time data on current water quality at sufficient representative*



locations to parameterise and update a digital twin to be the biggest obstacle because of the resource required to achieve this”.

3. Review of existing data and models

As an organisation focussing on environmental science and encompassing a range of expertise from monitoring to modelling across a range of systems, UKCEH is in an optimal position to access/curate data, develop models, and integrate these. As part of this study, reviews of available data and models have been undertaken, focussing on those that are most relevant to the requirements identified by the stakeholders (Section 2). Though our reviews were not intended to be exhaustive, they were conducted in a way that would represent a range of different freshwater ecosystem types and scales (running and standing waters, catchments), environmental states and processes (physical, chemical, biological), and modelling approaches (process-based, data driven).

The reviews were structured to address the various needs and challenges highlighted by our stakeholders, thus characteristics such as spatial and temporal resolution and the selection of determinands measured were the central aspects considered.

3.1 Review of existing data

There are two common data collection methods that provide most of our observational quantification of water quality in the UK, and these differ in their spatial, temporal, and ecological resolution. These are manual sampling, occurring at fortnightly-monthly intervals or less frequently, and automated high-frequency monitoring through sondes.

Manual sampling and lab-based analysis often gives information on a large suite of nutrients, contaminants, and biological measures. However, it is labour intensive and does not capture temporal dynamics occurring at frequencies higher than the seasonal scale. Automated sonde measurements are restricted to a smaller subset of water quality indicators and proxy variables but can reveal insights into the high-frequency dynamics of these variables (e.g. in response to extreme events, Woolway et al., 2018). Furthermore, these data have been collected over different spatial extents, both as part of national-scale Agency monitoring programmes, and also through long-term research programmes at sentinel sites (e.g. UKCEH monitoring in the Cumbrian Lakes, Loch Leven, and the Thames catchment). Increasingly, satellite remote sensing is also being used to assess the status of inland waters over broad spatial scales, delivering a subset of measures that are



observable based upon reflectance data (e.g. <https://3deo-portal.com/#/dashboard/scientificMap/UniversityOfStirling>).

A digital twin must be able to exploit these different data sources, capitalising on their individual strengths while robustly managing their weaknesses. Here, we give a summary of the existing data in each of the two main data classes identified above. It is, however, important to emphasize that although a significant amount of data is available and freely accessible, the intricacies of the data are seldom documented. It is therefore extremely important when using the data to establish close connections with data providers, to capture data particularities. Such intricacies can include changes in detection limit through time, operational deployment and redeployment of sondes.

Manual sampling

There is a large network of sampling sites covering most of the catchments in the UK (to a lesser or greater extent) where freshwater samples and measurements are collected. These manual sampling efforts are predominantly managed by the Environment Agency (EA) in England, the Scottish Environment Protection Agency (SEPA), the Natural Resources Wales (NRW) and the Northern Ireland Environment Agency (NIEA). The subset of determinands measured in these samples varies spatially and temporally, but there is a historical record dating back to the late 20th century. The measurement and analysis techniques deployed in these schemes has varied through time and space, including factors such as changing detection limits. The frequency of measurement, both temporal and spatial, will also vary depending on the determinand considered. Some determinands (e.g. biochemical oxygen demand (BOD), concentrations of nutrients, dissolved oxygen (DO), chloride) are routinely measured at higher temporal frequency than others such as metals and pesticides.

To illustrate this higher temporal frequency for the manual sampling network, we consider a subset of three routinely sampled determinands within the EA Water Quality Data Archive (<https://environment.data.gov.uk/water-quality/view/landing>): pH, DO, and ammonia concentration.

The evolution through time of the number of monitoring sites at which samples of these determinands were measured is captured in Figure 2. These are measured at most sites and in most samples: almost 10,000 sites within England had at least one sample taken where these determinands were measured in the year 2000. This number has been decreasing over time and stood at around 4,000 sites with a single sample in 2023. The number of sites where regular monthly measurements of these three common determinands were made is lower: around 3,000 in the year 2000 and less than 2,000 in 2023 (again within England).

For less frequently measured determinands, the manual sampling dataset can be sparse, both temporally and spatially (Table 1). Glyphosate, the most recorded herbicide in the data set, was not measured regularly (*regularly* meaning at least 11 samples per year) at any sites in England before 2022. In 2022, heightened interest



in the chemical translated into increased monitoring, with 63 sites measured for glyphosate regularly throughout the year. Coliforms were measured regularly at only 18 sites throughout 2022. Measurement of heavy metal concentrations in the river network has decreased since 2000, with regular measurement of zinc and nickel occurring at over 2000 and around 500 sites, respectively, in the year 2000 but only around 100 sites for both metals in 2022.

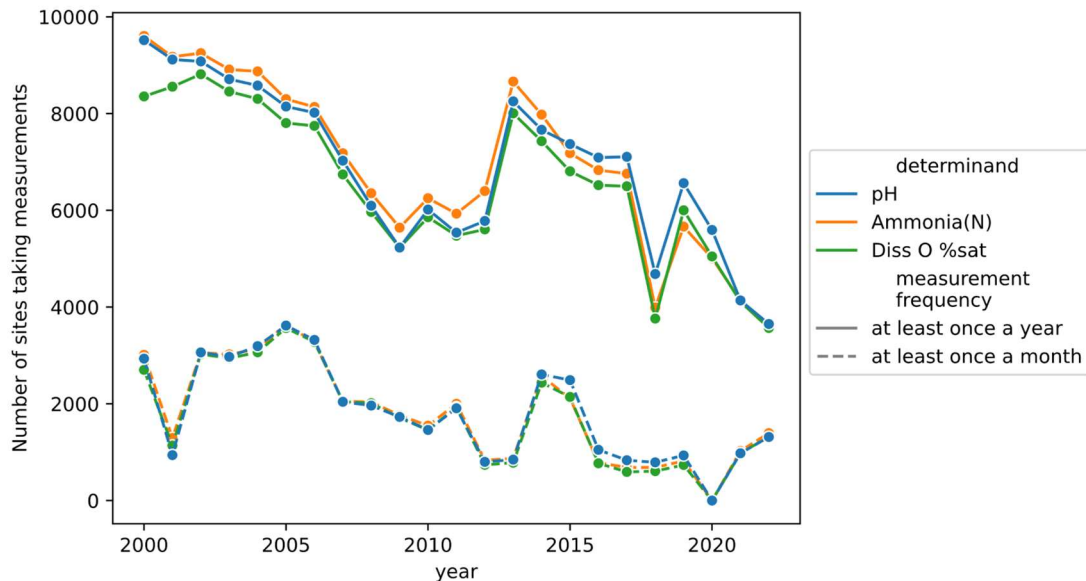


Figure 2. A line plot showing the number of sites with measurements of three key determinands in the EA Water Quality Archive.

Water quality data from long-term manual sampling of sentinel lakes and lochs are available from UKCEH, via the Environmental Information Data Centre (EIDC). This monitoring activity began in 1945 for the Cumbrian Lakes (initiated by the Freshwater Biological Association), and in 1968 for Loch Leven. Sampling has been conducted at weekly-fortnightly intervals throughout this time, and includes water temperature, and concentrations of oxygen, fractions of phosphorus and nitrogen, silica, and chlorophyll-a, as well as pH and alkalinity. Similarly, long-term data on a suite of physical, chemical, and biological ecosystem states have been collected from the River Thames and many of its tributaries, since 1997.

Any model attempting to predict a suite of determinands will achieve differing levels of calibration and validation based on the availability of data for each selected determinand. While some strong proxy relationships may be captured between commonly observed variables and infrequently sampled contaminants, this is unlikely to be the case for all target determinands. By incorporating contextual information on land cover, land use and effluent inputs into a data-driven model, it may be possible to overcome the data sparsity and generate more robust estimates for a larger list of determinands.



Another challenge posed by working with data from long-running sampling regimes is the possibility of methodological changes in data gathering or processing procedures. If metadata describing these changes are available, they can be represented as latent variables when using the data in statistical or machine learning models. If the metadata are missing, further analysis of the time evolution of the observations may be required to search for change points. This will be challenging due to the many confounding factors that are present in a complex environmental system.



WADITI - Developing a roadmap for a water quality digital twin

Table 1. Number of sites with monthly measurements for a selection of years in the EA Water Quality Archive (Monthly columns). The total number of sites with measurements is also indicated (Total columns). Only a selection of determinands are included.

ID	Determinand	Grouping	2022		2018		2014		2008		2000	
			Monthly	Total	Monthly	Total	Monthly	Total	Monthly	Total	Monthly	Total
948	Algae : Total cell count	Algae	0	0	0	0	0	0	0	0	0	0
111	Ammoniacal Nitrogen as N	Ammonia	1395	3588	685	3991	2574	7979	2036	6354	3013	9605
119	Ammonia un-ionised as N	Ammonium	1143	3123	419	3224	2190	6612	1814	5098	2532	8379
5889	Triclosan	Antimicrobial	60	131	0	25	22	445	0	6	0	0
85	BOD : 5 Day ATU	BOD	208	941	183	861	381	2478	623	3874	2880	9371
88	BOD : 5 Day	BOD	0	0	0	0	0	1	1	44	0	2
301	Carbon, Organic, Dissolved as C :- {DOC}	Carbon	481	1120	366	1811	632	1980	100	401	50	254
99	Carbon, Organic, Total as C :- {TOC}	Carbon	0	34	0	0	0	0	0	0	42	57
947	Chlorophyll a + b	Chlorophyll	0	0	0	0	0	0	0	0	119	508
92	Chemical Oxygen Demand :- {COD}	COD	11	232	18	312	43	1343	238	1404	206	1222
3461	Coliforms, Faecal : Presumptive : MF	Coliforms	0	0	0	0	3	38	8	540	4	271
3458	Coliforms, Faecal : Confirmed	Coliforms	0	0	0	0	3	38	5	182	4	86
2345	Escherichia coli : Presumptive : MF	Coliforms	0	0	0	0	0	0	0	0	0	0
6956	Escherichia coli	Coliforms	0	0	0	0	0	0	0	0	0	0
938	Escherichia coli : Presumptive : MPN	Coliforms	0	0	0	0	0	0	0	0	0	0
2348	Escherichia coli : Confirmed : MF	Coliforms	18	366	20	266	62	475	0	50	0	4
77	Conductivity at 25 C	Conductivity	1285	3498	647	3860	2550	7394	1634	5086	975	3386
62	Conductivity at 20 C	Conductivity	23	45	2	15	4	23	1	574	496	1640
9901	Oxygen, Dissolved, % Saturation	DO	1315	3568	608	3762	2437	7430	2013	5966	2702	8355
9924	Oxygen, Dissolved as O2	DO	1252	3148	547	3299	2375	6879	1913	5384	2076	6927
3002	Atrazine	Herbicide	9	53	19	35	52	108	49	162	67	316
3549	Mecoprop	Herbicide	62	160	8	30	32	122	33	142	55	337
3545	2,4-D :- {2,4-Dichlorophenoxyacetic acid}	Herbicide	66	156	5	15	24	102	29	183	54	334
3118	Linuron	Herbicide	62	134	2	10	14	64	20	81	39	280
3548	MCPA :- {4-Chloro-2-methylphenoxyacetic acid}	Herbicide	60	154	8	19	29	66	31	109	45	280
4065	Bentazone	Herbicide	60	149	6	13	14	91	2	41	23	122
3790	MCPB :- {4-Chloro-2-methylphenoxybutyric acid}	Herbicide	60	149	5	12	16	44	5	37	2	99
3009	Terbutryn	Herbicide	15	59	21	64	20	51	11	54	5	60
9477	Dichlobenil	Herbicide	0	4	1	3	3	12	8	39	22	112
3007	Glyphosate	Herbicide	63	130	0	0	0	3	0	6	0	23
3000	Diquat	Herbicide	0	0	0	1	4	6	0	4	0	5
73	Cypermethrin	Insecticide	26	186	6	190	21	103	20	92	10	132
9823	Permethrin (cis- and trans-)	Insecticide	0	0	4	18	18	20	11	30	7	124
9341	cis-Permethrin	Insecticide	39	142	6	194	26	57	12	84	10	66
5121	Metaldehyde	Insecticide	24	32	0	0	3	34	0	0	0	0
6051	Iron	Iron	20	150	70	357	107	692	135	792	325	1623
6460	Iron, Dissolved	Iron	124	568	234	1511	245	1327	152	1080	235	1172
6050	Manganese	Manganese	4	128	52	289	61	541	96	435	116	764
6458	Manganese, Dissolved	Manganese	75	381	178	782	161	793	85	317	34	337
6462	Nickel	Nickel	66	304	85	333	156	764	148	1008	436	1858
3410	Nickel, Dissolved	Nickel	118	591	256	991	329	1539	128	916	322	1258
117	Nitrate as N	Nitrate	1307	3187	525	3200	2676	7006	1924	5284	1864	6017
118	Nitrite as N	Nitrite	1356	3388	560	3593	2718	7421	1958	5450	1991	6248
116	Nitrogen, Total Oxidised as N	Nitrogen	1343	3463	642	3857	2819	7920	1991	6052	2943	9114
9889	Oestrone :- {E1}	Oestrogen	22	48	0	23	0	39	0	6	0	0
6982	PAH : Total :- {Polynuclear Aromatic Hydrocarbons}	PAH	0	3	0	0	0	12	1	6	0	0
555	DDT -pp	Pesticide	0	17	11	258	41	413	83	252	51	413
551	DDE -pp	Pesticide	0	17	11	257	41	409	83	244	59	384
539	DDT -op	Pesticide	0	17	11	257	41	365	29	196	57	365
581	DDE -op	Pesticide	0	4	4	6	6	14	3	36	23	165
575	DDT : Sum of components	Pesticide	0	13	3	246	20	330	13	123	0	50
7013	Pesticides : Total (Gamma-HCH, Dieldrin, Parathion) : SWAD	Pesticide	0	0	0	0	0	0	0	0	0	14
7213	DDT : Total Isomers (DDT op pp, DDE pp, TDE pp)	Pesticide	0	0	4	4	6	8	10	21	8	195
180	Orthophosphate, reactive as P	Phosphate	1348	3517	681	3895	2556	7816	1952	5933	2694	8663
348	Phosphorus, Total as P	Phosphate	450	1108	369	696	419	711	188	775	122	358
192	Phosphate :- {TIP}	Phosphate	238	299	119	157	92	198	35	270	113	303
1198	Salinity	Salinity	0	0	0	0	0	2	0	0	0	4
3863	Diclofenac	Steroid	0	0	0	0	0	39	0	0	0	0
135	Solids, Suspended at 105 C	Suspended solids	651	1706	314	1428	509	3469	1395	4382	1615	6381
76	Temperature of Water	Temperature	1494	3720	716	3937	2543	7587	2052	6129	2890	9065
6455	Zinc	Zinc	83	410	113	712	278	1401	1405	3934	2066	6242
3408	Zinc, Dissolved	Zinc	119	604	214	1064	248	1520	151	953	162	986
6423	Streptococci : Faecal : Presumptive : MF	nan	0	0	0	0	0	0	7	509	4	321
2551	Streptococci : Faecal : Confirmed : MF	nan	0	0	0	0	0	0	5	161	4	14
61	pH	pH	1315	3650	790	4684	2608	7665	1964	6095	2935	9519
3169	pH : In Situ	pH	4	61	31	187	29	240	0	22	14	917



Sonde measurements

The network of sondes taking real-time measurements is small when compared to that of the manual sampling sites. The use of sondes by the EA dates back to around 2008 but their use became more widespread in 2014. Data are made available via the EA Hydrology Data Explorer (<https://environment.data.gov.uk/hydrology/explore>). It is important to note that many sondes are currently deployed for a period of 4 to 5 consecutive months. Furthermore, it is noteworthy that one sonde may measure several determinands at once.

As illustrated in Figure 3, there are currently around 80 sondes monitoring DO, turbidity, water temperature, conductivity, ammonium, and pH. At a subset of these sites, concentrations of nitrate, chlorophyll, blue green algae and/or fluorescent dissolved organic matter are also monitored, and these are shown in Figure 4. Once again, high frequency sonde data are collected at UKCEH sentinel lake and river sites.

Utilising high-frequency sonde data in a digital twin will be pivotal to embedding sub-seasonal dynamics into feature relationship models (e.g. machine learning proxy variable-based predictors) and in the assimilation of near-real time observations to allow responsiveness to rapidly changing conditions (e.g. in storm overflow events). The sub-hourly resolution of sonde measurements (generally at a 15-minute time step) gives a very resolved view of how water quality and ecological state variables change in waterways. Within the proposed digital twin structure, operating at a daily (or coarser) time step, these high-resolution observational data could be used to estimate sub-temporal grid uncertainty and, to test the utility of the coarser time step in capturing rapidly evolving events.

A challenge for the digital twin will be to integrate manually collected and automatic sonde data which may or may not be co-located in space. If true co-location does not occur, or is rare, it may be possible to utilise data science approaches that account for spatial dependencies and covariance between manual and automatic monitoring sites that are near neighbours. Such an approach is discussed in Section 3.3.



WADITI - Developing a roadmap for a water quality digital twin |

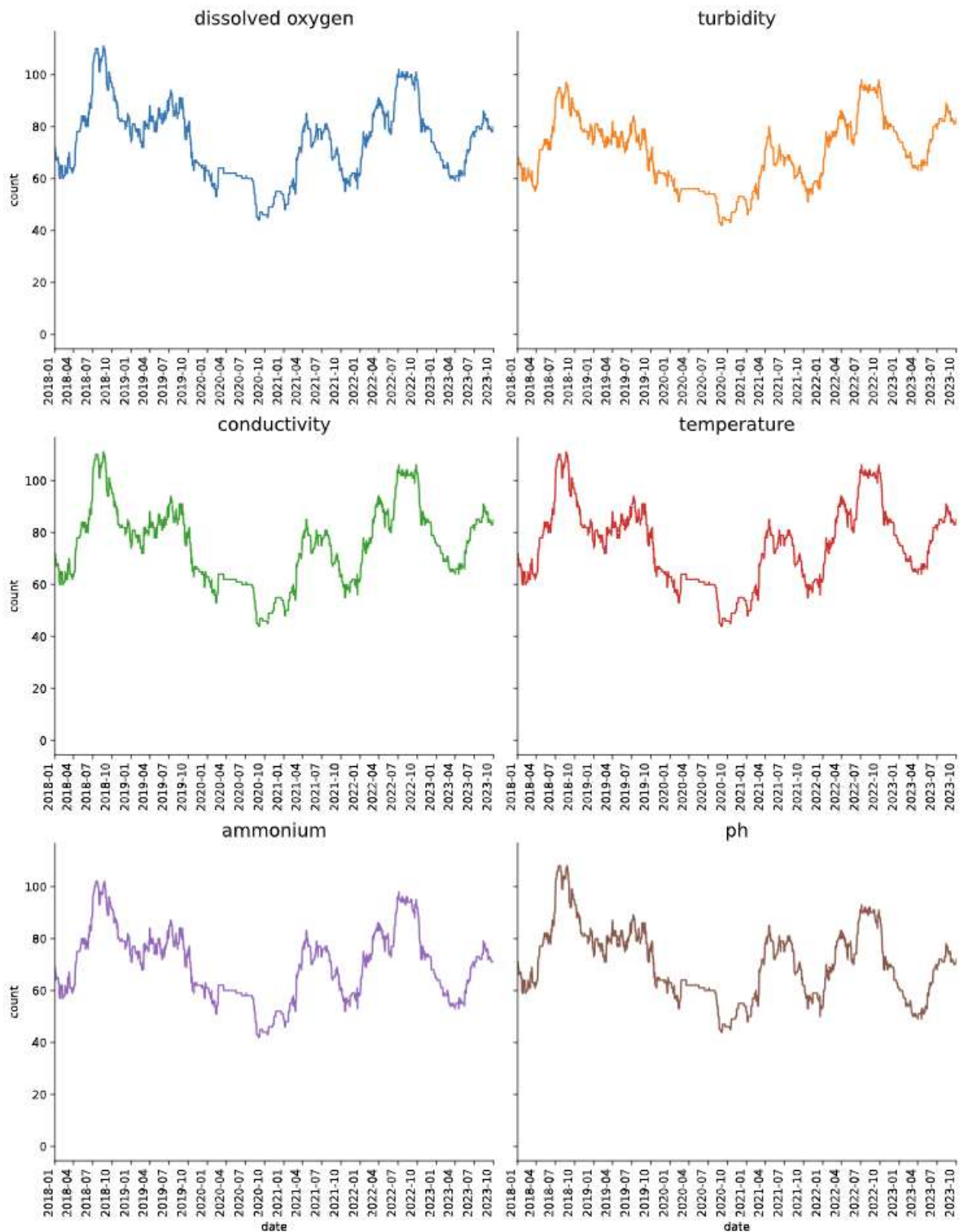


Figure 3. Line plots showing the number of active EA sondes measuring dissolved oxygen, turbidity, conductivity, temperature, ammonium, and pH between 2018 and 2023.



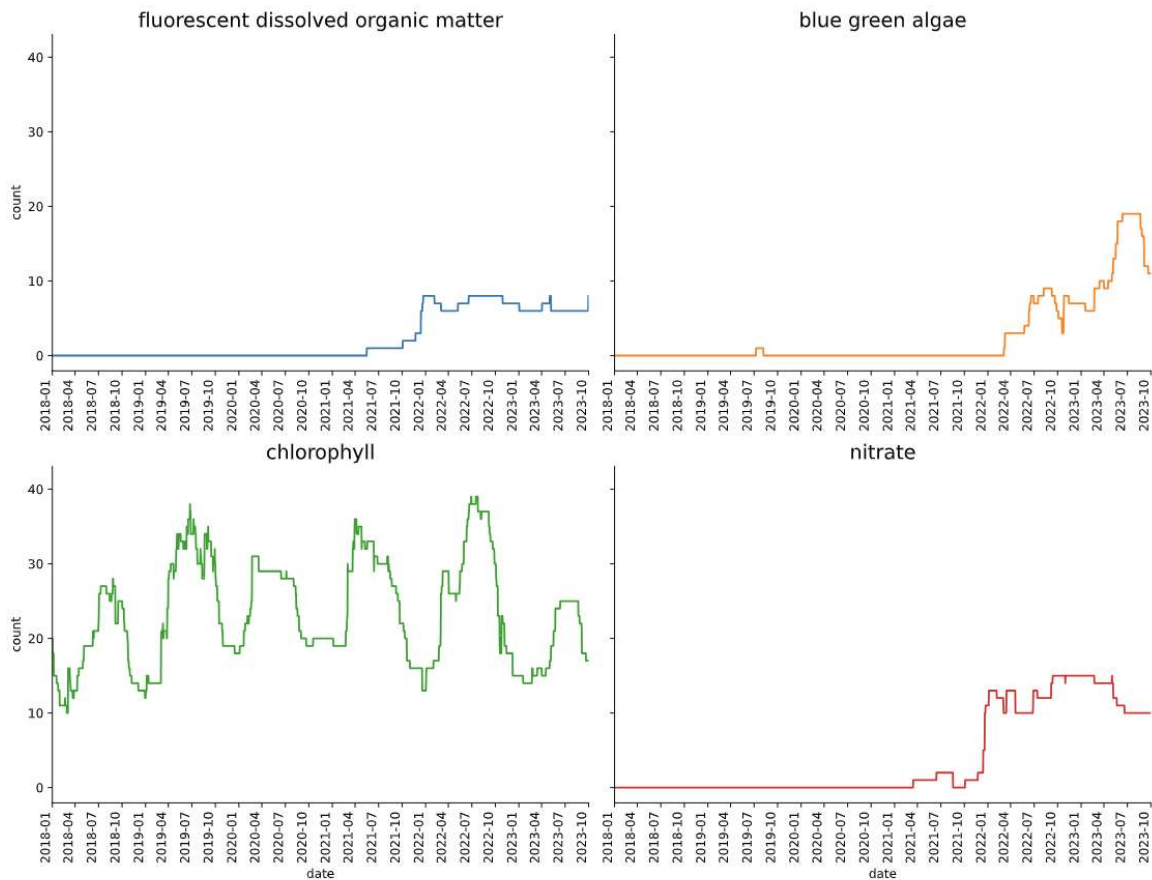


Figure 4. Line plots showing the number of active EA sondes measuring fluorescent dissolved organic matter, blue green algae, chlorophyll, and nitrate between 2018 and 2023.

The above considerations clearly demonstrate several aspects of the “Data Challenge” recognised by Blair et al. (2019a) and Jagadish et al. (2014). It is clear that a water quality digital twin will have to contend with the “four Vs” of environmental data: *volume*, *velocity*, *variety*, and *veracity*. Large *volumes* of water quality data are being collected by agencies, research institutes, the water industry, and charitable organisations, but these data show great *variety* (heterogeneity) in their spatial and temporal extent and resolution, determinands measured, and methods used. The *velocity* of data streams varies from near-real-time, for sondes, to time lagged for data generated by laboratory analysis of manually collected samples. Given that UK water quality data are collected by various organisations, and by a mix of professional surveyors and citizen scientists using different protocols, we also should expect that the *veracity* (reliability) of all of these data streams will vary. As a specific example, although we found evidence of FAIR principles being enacted through the EA Water Quality Archive (the API was well documented, and data easy to access), there existed many closely related determinands (with different identifiers), which represent the same chemical



measure, but analysed according to a different protocol. These protocols likely vary in their accuracy and limits of detection, but this is currently hard to resolve with the available metadata. All of this underscores the need for comprehensive, machine-readable metadata that can be streamed into the data pipelines underpinning a water quality digital twin, so that this information can be used in data processing.

3.2 Review of existing models

Within UKCEH, a variety of models have been, and are being developed for water quality purposes for both rivers and lakes. Within the remit of this scoping study, a comprehensive review, although not exhaustive, was undertaken of UKCEH models. To capture the breadth of models available in the water quality arena, other models commonly used in academia and the water industry were also added to the review.

Whilst undertaking the review, the following information was gathered for each model:

- Model name
- Description
- Developer organisation
- Type of model (e.g. process-based, empirical, data driven)
- Focal determinands
- Temporal scale
- Spatial scale
- Potential contribution to a near real time digital twin
- Useful links providing more detail on the model

A set of 29 models was reviewed, including 6 UKCEH owned models. This set included 8 models specifically designed for standing water.

Most of the models reviewed operate at a daily timestep, and several can operate at sub-daily timesteps as well. Some models such as LAM (Load Apportionment Model, (Bowes et al., 2008)), SAGIS (Source Apportionment GIS, (Comber et al., 2018)) and NIRAMS (The Nitrogen Risk Assessment Model for Scotland (Sample and Dunn, 2014)), adopt coarser time scales, either monthly or annual. It is important to note that the temporal resolution of a model is often dictated by the resolution of the input data. Therefore, many models can operate at a variety of time resolutions, with many models reviewed providing the possibility of using daily or sub-daily time scales. Some models, such as LF2000-WQX (LowFlows2000 Water Quality eXtension, (Williams et al., 2009)) adopt a stochastic approach, and thus describe the system in terms of long-term trends (e.g. flows and concentration). Such approaches allow the user to maximise the use of monitored data that may not have sufficient monitoring frequency (e.g. monthly water quality data).



Regarding spatial scale, most of the river models considered operate at a catchment and sub-catchment scale (e.g. SWAT, INCA, QUESTOR). There are however a few examples, such as SAGIS and LTSM (Long-term simulations of macronutrients (Bell et al., 2021)), that are national scale models, or that have a flexible scale (NanoFASE (Harrison et al., 2021)). With respect to the lake/reservoir focussed models, most of the models selected in this review consider a single lake/reservoir per simulation. Although such models may not have the innate ability to simulate a series of inter-linked reservoirs, this specific challenge could be addressed within a digital twin framework. To the authors' knowledge, the AQUATOX model (Park et al., 2008) is the only model reviewed here that is designed to simulate a river network including streams and multiple lakes and/or reservoirs.

Stakeholders identified nutrients as the priority determinands to include in a digital twin. These determinands are commonly modelled throughout the UK and across the world. Most of the models in this review can simulate nutrient loads or concentrations, apart from LF2000-WQX, WHAM (Windermere Humic Aqueous Model (Tipping et al., 2011)), some lake hydrodynamics models, and NanoFASE. The LF2000-WQX model was designed to represent the fate of “down-the-drain” chemicals in rivers, and thus focusses on those chemicals that enter the surface waters mainly via wastewater treatment works. WHAM is an equilibrium chemical speciation model that is best suited to metals in soil and water systems. NanoFASE was originally developed for nanomaterials and is best suited to particulate matter, such as microplastics, but is currently being updated to include functionality for pharmaceuticals. Several lake models, identified in this review, can simulate nutrients, including PROTECH (Phytoplankton RespOnses To Environmental Change (Elliott et al., 2010)), Delft3D (Deltares, 2024)), AED (Hipsey, 2022)) and PCLake+ (Janssen et al., 2019).

Despite some models being capable of modelling nutrients at a sub-daily resolution, these models will differ in terms of their level of complexity and representation of the environment. For the purpose of this review, the models suitable for modelling nutrients at a catchment scale and daily to sub-daily resolution are captured in Figure 5.











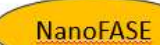






	Process Model	Data driven / statistical
Lake / Reservoir	    	
Running water – Catchment	     	  
Running water - National		

Figure 5. Water Quality models suitable for modelling nutrients at daily/sub-daily resolution. Note: NanoFASE, although not designed to model nutrients, has the flexibility to do so.

To summarise, there are several models available within UKCEH that provide the means to address stakeholder needs. These models encompass a range of processes, chemicals, scales (both temporal and spatial), environments (running/standing freshwater) and complexities. It is important to note that often models are designed to address specific questions, and subsets of ecological processes, and are static in time. UKCEH models are often operated on their own and may use inputs from various data sources including those described in Section 3.1.

This model review demonstrates the considerable expertise held within UKCEH in terms of water quality modelling. These models could form the basis of the modelling component within a digital twin but, to achieve this, the models would need to be translated and adapted into a modular form within a wider digital twin framework. Modularity is key as it would facilitate integration and thus, for instance, promote the interchangeability required for process representation, model linkages that allow simulation of connected freshwater environments, and a capacity to resolve processes across temporal and spatial scales.



Finally, the integration of models in a modular form within a digital twin would also allow the integration of models and available data, relaxing the static nature of model parameterisation, and opening up the possibility of model learning, as described in the following section.

3.3 Integrating data and models through data science methodologies

A digital twin represents an opportunity to design a system where mechanistic modelling and data science approaches work in tandem to harness both process understanding and insights from observational data. Some machine learning approaches may be challenging to apply in a near-real time framework where data are spatially and temporally sparse. It is certainly the case that the directed network structure of a river system creates challenges for the application of spatial statistics, though examples exist. Indeed, it may be the case that there is no plug-and-play solution currently in the literature for the specific needs of a water quality digital twin. We will first survey existing uses of data science in water quality research before examining how aspects of these methodologies might be adapted and combined for use in a digital twin.

There are many existing applications of machine learning techniques to various aspects of water quality, including source identification using deep belief networks (Liang, 2021), anomaly detection on sensor networks (Leigh, 2019; Liu et al., 2020), and modelling spatiotemporal relationships among sensors on an explicit network river structure using graph neural networks (Buchorn et al., 2023; Li et al., 2022; Ni et al., 2023). Spatial extrapolation of water quality indicators on river networks has also been studied using stochastic process kriging (Cressie et al., 2006; Garreta et al., 2009). More classical tree-based methods (random forests, gradient boosting trees) have been used frequently in recent years to predict determinands of interest from proxy variables, and to quickly identify feature importance between inputs and outputs (e.g. Schäfer et al., 2022; Wang et al., 2021). More broadly, Blair & Henrys (2023) explore possible connections between process models and data models pertinent to digital twinning. They highlight that data science methods can be used at each step of a process modelling pipeline, from data cleaning and quality control to data assimilation and model validation. In addition, data science techniques can be used directly for predictive modelling, where their strength in pattern-finding makes them good candidates for situations where no mechanistic model is known – at the expense of being “black boxes” with little interpretability. With all these approaches, data availability in terms of spatial and/or temporal coverage will govern how well the model performs and generalizes. This constraint must guide the choice of data science approaches that are used.

To set out how we could use the above approaches in a digital twin, we should take guidance from the modelling framework that we propose in the roadmap (Section



4). First, we aim to run a spatially explicit riverine model defined on river reaches connected in a network. A predictive data model should be able to “live” in the same space as the process models, therefore it should be aware of that network structure and should be able to receive input from and, send output to such a structure. Graph neural network approaches allow connectivity to be defined in this way, making them a good choice as an overall machine learning modelling framework.

Second, the set of determinands that the chosen process models predict is likely to be a small subset of the list of all determinands that we would like to know about and for which we have observational data. The data model should be able to learn functional relationships between all members of this full list of determinands by combining process model outputs with data where available. Because observations are not always available, the data model should be robust to missing variables and missing timepoints and be able to make predictions to fill those data gaps. Ideally, the model should quantify the uncertainty of its predictions. Geostatistical methods that are capable of learning from contextual information (like classification of river reaches through neighboring land cover and land use) may be key in ameliorating data sparsity issues. Network kriging using stochastic processes could act as a baseline prediction for data gap-filling with uncertainty quantification.

Third, there are likely to be variables calculated within the chosen mechanistic models that are not present in our list of important determinands. These extra model outputs could form a set of proxy variables that act as secondary inputs to the data model.

Fourth, outside the river waterways there are data describing the catchment’s geography in terms of land use, topography and soil composition, and meteorological data that inform on processes driving water, solute, and sediment movement through the catchment and into the waterways. It should be possible to use such “external” (to the river network) data as contextual fields that the data model can learn from.

Taking the constraints/guidance discussed in the previous paragraphs into consideration, an outline for implementing a data science framework alongside process modelling within the water quality digital twin could be the following: (i) “pre-processing” methods to clean, validate and gap-fill observational data; (ii) “processing” methods to predict variables not included in process model output; (iii) “post-processing” methods for anomaly and change-point detection in process and data model outputs, and for investigating model performance during normal and extreme scenarios.

The pre-processing methods should make use of correlative structures in the contextual data sources (described above) and water quality observations to find outlier periods or regions in spot-sampling data. Time series data from sondes and



autosamplers should undergo cleaning before being used as inputs into either data or process models. Where possible, infilling of time series could be achieved using Gaussian processes. Spatial extrapolation from spot-sampling sites may be possible through network kriging, again making use of land cover and other contextual fields to constrain the problem.

The predictive (“processing”) data model should have a graph structure with the nodes representing the river reaches in a catchment and the edges defining the adjacency matrix and directionality of the graph. Each node will have two input data streams and a single output data stream. The first input data stream will be the contextual data describing the geography and land use of the parcels of land adjoining each river reach. If possible, the subset of the upstream catchment from which water feeds directly into a river reach should be quantified and linked to the node as contextual data in the same way. The second input data stream will be dynamic in-river variables: model outputs of target determinands and other proxy variables at the current node as well as observational data on the full list of determinands of interest available for the current timestep. The output data stream is a vector of predictions and associated uncertainties for each of the water quality determinands at each node at the current time step. The architecture of the forward pass from inputs to outputs should be based on a combination of dense layers combining input data streams within a single node and causal graph convolutions to capitalise on the spatial structure inherent in the system and give the model the opportunity to learn how to share information across the river network. Between each process model timestep (each timestep of QUESTOR and PROTECH, for instance) the machine learning model inputs are collated, and the forward pass computed to generate the predictions for the full list of target determinands.

The second input data stream is likely to have many gaps because of the lack of model outputs or observational data for most of the full list of water quality determinands. We suggest three possible approaches for dealing with these data gaps:

(i) Utilising spatial statistical approaches based on kriging on the river network to extrapolate away from monitoring sites. This approach has the benefit of uncertainty quantification but may struggle with rarely observed variables and variables displaying strong temporal dynamics.

(ii) Using the dynamic information from the process-based flow model to approximate movement through the network structure by constructing a flow-driven adjacency matrix A and solving the equation:

$$\frac{dc}{dt} = A(t) c(t)$$

for each determinand, where c is a vector with length equal to the number of river reaches in the catchment and with values associated with the target determinand at



each of the river reaches. Between each model timestep $[t, t + \Delta t]$ we assume the flow matrix $A(t)$ can be approximated by A_t , a constant, and solve the resulting equation:

$$\frac{dc}{dt} = A_t c(t) \Rightarrow c(t + \Delta t) = e^{A_t \Delta t} c(t).$$

The value $c(t + \Delta t)$ represents the current outputs propagated through the river network under the action of the flow and can be used to form part of the inputs to the machine learning model at the next time step. This gives us a way to perform dynamics-informed gap filling for the second input data stream. We can also think of this approach as providing a framework for integrating the spatially sparse high frequency sonde data with the temporally sparse but spatially more dense manual sampling network data in the situation where sondes and manual samples are not co-located.

(iii) The data model could learn to make auxiliary predictions based solely on the contextual input data stream which we should always have access to. These predictions have value aside from the potential to fill gaps because they can act as a baseline prediction against which observations or other predictions can be compared to determine “hot spots” or “hot moments” of water quality events in the river network.

Training such a model will require running the process models across the period where historical observations exist. The process models need only be used once to generate their outputs across the data period; then, the machine learning model can be run on segments of the timeseries, minimizing the error between the predictions and observations. These observations can be from high resolution sonde data, structured spot sampling, ad-hoc manual sampling or even citizen science water quality data. To reduce overfitting, entire catchments could be left out of the training set to be used as validation, if the training set included an adequate distribution of land cover, land use and soil types.

The post-processing elements of the data science framework could be embedded in the digital twin model or be part of a dashboard for user interactivity and interrogation. Automated checks on model outputs – looking for model drift with respect to observational data; flagging spatially explicit change points in water quality to aid source identification; monitoring model performance in high-pollutant events – will increase the value of the digital twin system for stakeholders. Feedback into process or data models through data assimilation also falls under this heading but is covered elsewhere in this report.

It is worth noting that we have only discussed data science approaches for in-water data and processes, under the presumption that loads of potential contaminants from the terrestrial environment can be inferred from contextual layers like soil type and land use. The chosen water quality model may be part of a broader set of



models that includes terrestrial processes such as soil erosion, and therefore these data science considerations could also be extended to the terrestrial environment. The notable barrier to adopt this approach is a lack of real-time terrestrial monitoring data to inform such approaches.

4. Roadmap

Considering the stakeholder priorities, data sets and models identified above, there is a case that we should prioritise the development of a digital twin of catchment-scale nutrient and contaminant loads and concentrations, and associated effects on algal growth and oxygen concentrations in connected running and standing waters. Such a digital twin could be built upon agency and sentinel site datasets and existing water quality models. Below, we outline the digital infrastructure considerations that would apply to such a digital twin, as well as two case study applications focused upon specific models and water quality processes.

4.1 Cyber-physical infrastructure

Arguably what defines a digital twin over and above a model, or set of integrated models, is that a digital twin should be made accessible via a federated, interoperable infrastructure that enables the transfer of data to and from the twin for real-time or near real-time parameterisation. This is often referred to as Cyber-Physical Infrastructure (CPI), and the system by which data (information) is transferred through the infrastructure is one part of the so-called Information Management Framework (IMF). Considerations around the CPI and IMF are important for any digital twin, and significant work has already been done in scoping the requirements for an environmental Information Management Framework (IMFe) in a report led by the National Oceanographic Centre (NOC) and funded by NERC and the Met Office (Siddorn et al., 2022). Here, we further expand some of the concepts discussed in that report, making them more relevant to water quality digital twins.

Enabling FAIRness through standards, interfaces, conventions, and semantics

At the heart of any digital twin is the passing of data between various components – physical and virtual. Such operations require careful consideration of standards, conventions, and semantics. The overarching theme in these recommendations is that data and software should strictly follow the FAIR principles – ensuring they are Findable, Accessible, Interoperable and Reproducible. Steps that can be taken to meet these principles include, but are certainly not limited to:

- Standardised interfaces for any models used in the digital twin, including consideration as to whether dynamic coupling (passing of data on individual model timesteps) is required and whether the given interface is capable of



this. A compelling interface that is flexible enough to be used with the types of spatial and dynamic models used in water quality modelling is the Basic Model Interface (BMI): <https://bmi.readthedocs.io/en/stable/> (Hutton et al., 2020). Another interface worth considering is OpenMI (<http://www.openmi.org/>), though notably this has a steeper learning curve and fewer example interfaces (only C# and Java) compared to the BMI. It also has not seen any significant updates since 2010, whilst the BMI is actively maintained. Simpler models may not require such a prescriptive interface as these, but still should follow standards to ease usage and interoperability, such as RESTful APIs (Representational State Transfer). REST is broadly a standard that enables the transfer of data between two computer systems (e.g. models), usually via the internet, in a “stateless” manner – meaning that requests via the interface are completely independent of each other and not dependent on what requests have already been made.

- The use of standards and conventions around data management will facilitate easy passing of data between components, and minimise the risk of mistakes being made by, for instance, misinterpreted metadata such as grid projections. The standards and conventions include file formats, names of variables and metadata. In the context of water quality modelling, we recommend using NetCDF for spatial data (ideally Climate and Forecast (CF) compliant (Hassel et al., 2017)), or well-described text formats (e.g. CSVs) for lower-dimensional data.
- The variable names that the models and data use should at the very least follow a coherent and documented naming convention, but ideally should follow standardised naming rules such as CF or CSDMS Standard Names (https://csdms.colorado.edu/wiki/CSDMS_Standard_Names). The latter is arguably the most robust and compatible with our recommendation to use the BMI. This reduces the risk of variables being used incorrectly and makes it easier for researchers to couple components. It also raises the prospect of automated component coupling, as discussed further in the next section on modularity.
- To enable the virtual components of the digital twin to be ported to different cyber infrastructure, established conventions around defining computational environments should be used. At a bare minimum, instructions for setting up a computational environment for each component should be included, but better would be the use of robust package management (e.g. Conda), containerisation (e.g. Docker) and deployment (e.g. Kubernetes) tools to enable automation.
- Deviations from these standards and conventions are likely. For example, labelling variables with CF Standard Names is often not possible (the list of Standard Names is finite and skewed towards climate sciences). Where deviations occur, metadata are vital. We recommend that every component



within the digital twin is thoroughly described with metadata, giving the relevant context, background information, instructions for use, units, assumptions, etc.

Many of the above recommendations will require updates to existing models and data. Whilst the effort required for these updates might be significant, it is likely they will pay off in terms of future-proofing models and data for future use in integrated digital systems. By showcasing the use of standardised interfaces in our digital twin, we would promote the use of such standards across other developing digital twin systems and raise the prospect of coupling digital twin systems together through these interfaces. Ongoing projects such as DestinE (earth system digital twin), Land InSight (digital twin for UK soils; Fry et al., 2022) and BioDT (biodiversity digital twin) are potentially of relevance for integration with our water quality digital twin.

Modularisation of assets into interoperable components

The IMFe report (Siddorn et al., 2022) states that:

“In the vision of the CPI there is no one monolithic infrastructure, but myriads of component parts sharing and communicating to allow a powerful interconnectedness of information. The components of a cyber-physical infrastructure are modular, reusable and networked.”

This applies to all components of a digital twin; however, it is particularly pertinent with regards to models as legacy models often have monolithic codebases. Modularisation of code is not new, and arguably the development of object-oriented programming arose from this need for modularity, but the modularisation of models required for a digital twin goes beyond this.

The basic motivation for using a modular, component-based infrastructure is that it allows for much more flexibility. For example, components (which might represent individual physical processes) can be easily interchanged to provide alternative conceptualisations based on the scenario modelled. Indeed, it enables the integration of process-based descriptions with machine learning methods. It also brings added benefits for future model developments, enabling new components to be easily developed independently without disturbing the existing model, and making it possible for developers to work on separate components in parallel with each other.

If these components are based on flexible cyber infrastructures, this provides further capacity through the potential scalability of computing resources to meet differing computational demands. A commonly employed software architecture paradigm through which to achieve this scalable modularisation is by using *microservices*. Microservices are fine-grained, loosely coupled components acting through standardised interfaces, which can be flexibly coupled together to provide a variety of broader systems. They are widely used in the tech industry and underpin



many of the services that are ubiquitous in modern life, from searching the internet to streaming music and videos, but their use in scientific computing is currently very limited. It goes without saying that splitting an existing model up into microservices is a significant undertaking, and careful consideration would need to be given to the optimal level of granularity. For example, examining whether a microservice should cover all in-stream processes in one, or whether individual processes (e.g. sediment deposition, chemical degradation) should be served by separate microservices. Finer granularity brings more flexibility, but at the expense of increased overhead from microservices communicating with each other. It is also worth considering that the interfaces between microservices and the interfaces between complete modelling systems might be different and need to be carefully thought out. The model interfaces discussed above (BMI, OpenMI) relate to full model systems that have, for example, timesteps and spatial grids. This might not be relevant for microservices that comprise the model (e.g. they might be dimensionless), and API standards such as REST might be more relevant here. Consideration should be given to if and how different interfaces are compatible with each other. For example, can a BMI model be implemented as a web service in a RESTful manner?

Modularisation brings another important opportunity: self-assembled models. This is based on the idea that, instead of a human choosing and linking the individual components to make the broader system, machine learning is employed instead. For self-assembled models, a machine learning algorithm is presented with a set of potential components to use, and it optimises the arrangement of these to minimise some pre-defined cost, such as ability of the system to predict observational data. This could bring us much closer to realising “*models of everywhere*” (Beven, 2007; Blair et al., 2019b) – models that are adapted specifically to account for the heterogeneity of different places and time periods, thereby used as a learning process about those places. By enabling the digital twin to be automatically re-assembled to best fit a given scenario, we can learn which microservices – that is, which physical processes – best represent reality. Arguably, a digital twin is nothing but a system by which we can realise the models of everywhere concept.

It is worth reflecting that place-based digital twins are being proposed. For example, the [Forth-ERA project](#) aims to develop a “*digital observatory of the Firth of Forth’s entire water catchment*”. If this was implemented in a flexible, modular manner, then adapting this digital twin to another catchment would not just be possible, but it would also be an invaluable learning opportunity about the differences between catchments.

Cloud based cyber infrastructure

Though we have not explicitly said so until now, it is important to reflect that the resulting digital twin is expected to be “cloud native” – runnable in and accessible from the internet, if for no other reason than to allow easy data transfer between measurement instrumentation, the digital twin and end users. This obviously



requires cloud servers, or use of serverless technologies. Adopting the recommendations above around containerisation and deployment (namely, using Kubernetes or similar) and microservices (with the potential to use serverless computing platforms such as Amazon Web Services Lambda) has the added advantage that there is no need to purchase or maintain the physical server(s) hosting the digital twin. If the digital twin is deemed confidential enough to preclude hosting on external services, then using conventional containerisation and deployment tools would make setting up a computing system on internal systems, such as the UKCEH private cloud, easily achievable.

Physical infrastructure

The physical infrastructure element of the CPI relates largely, in the case of a water quality digital twin, to the monitoring network that provides data to the cyber infrastructure. A key aspect of digital twinning is the ability to learn from model predictions, and part of this learning could be used to refine the monitoring network. For example, model predictions may indicate a particular geographical location that is more dynamic than others, such as having more rapidly changing or a broader range of determinand concentrations. If this location is not well monitored by the existing network, this raises the possibility of moving monitoring sensors to better cover this area, potentially resulting in better model predictions. This ability to optimise not just model configuration, but also monitoring network configuration, is fundamental to what separates a digital twin from a digital clone or shadow that have largely one-way data flows.

4.2 Graphical interfaces, portals, and visualisations

A key consideration that arose from the stakeholder consultation (Section 2.2) is that stakeholders perceive a graphical user interface as important (92% of respondents saw the need for a such an interface). This is understandable, given that this would dramatically increase the accessibility of the digital twin, enabling users to visualise and analyse data without them having to implement their own analysis routines. Given that the output data from the digital twin is likely to be spatial and therefore use binary file formats like NetCDF, analysing such data without a visual interface would likely require programming skills.

The obvious contender for a visualisation platform for a catchment-based water quality digital twin is a map-based portal, most likely accessible via a web browser and ideally compatible with mobile/tablet viewing. Ideally, it would be possible to visualise time series at distinct catchment locations as well (e.g. observational data streams, model simulations, forecasts with associated uncertainty). The effort required to create such a system should not be underestimated, and we strongly recommend thorough investment in user-experience (UX) and user-interface (UI) experts throughout the creation of the twin. Given that this portal is likely to be what



most users access the twin through, it is imperative that it is as functional, intuitive, and as useful as possible, with well thought out visualisations that present the underlying data in intelligible ways. Websites such as [Our World In Data](https://ourworldindata.org/) (<https://ourworldindata.org/>) could be taken as inspiration due to their success in making complex data accessible to the public. Stakeholder co-design and robust testing (such as A/B testing) will be essential to the development of such visualisations. There is a significant amount of expertise within UKCEH in producing successful visualisation platforms, such as the UK Water Resources Portal (<https://eip.ceh.ac.uk/hydrology/water-resources/>).

Not all users will want access solely through a web portal, and as such it is also important to provide access to the underlying data by other means, ideally as a well-documented API following a standard interface such as REST. The ability to select and download data through the web portal would also be welcome.

4.3 Sustainability and long-term maintenance

“How will this be funded in the long-term?” might well be the most asked and also most difficult question to answer when it comes to discussing software systems. Typically, especially in scientific funding streams, funding is provided to create the system, but not to maintain the system beyond the lifetime of the project. In reality, long-term funding is essential to update the system: ideally, with the latest knowledge, but at minimum to fix security vulnerabilities and to keep it working. The same considerations apply to digital twins, but with the added complication of the network of instruments providing observational data. Another important aspect requiring funding consideration is long-term user support, such services are important to secure a positive organisational reputation.

Most importantly, there needs to be an acceptance, by funders but also by institutions developing digital twins, that a digital twin is *infrastructure* and needs to be treated as such. One wouldn't expect to fund the creation of a laboratory without available capital for the associated overheads (servicing machines, purchase of consumables, securing energy needs, etc). In the same way, one cannot expect to fund the creation of a digital twin in the expectation that it will remain functional in perpetuity without long-term funding.

Although beyond the scope of this study, careful consideration regarding long-term funding streams will be required. Ultimately, potential funding streams may be decided based upon who the beneficiaries of such a digital twin are. These beneficiaries are likely to come from various sectors including, for example, water companies, to support decision making, and public organisations or agencies to foster public awareness and support policy makers and regulators. It is important that digital twin developers are clear to their funders from the outset that the twin will only be functional for as long as funding remains available.



5. Case studies

Considering the stakeholder requirements (Section 2) and the models and data available (Section 3), it is clear that there is scope to develop a water quality digital twin with capabilities including forecasting of nutrient and algal dynamics in connected freshwater systems (rivers and lakes). Here we consider two case studies, each exploring a possible digital twin building on, and capitalising on, UKCEH current modelling expertise coupled with data currently available within the UK. These case studies showcase the use of some of the UKCEH models reviewed in Section 3.2 and data available, as presented in Section 3.1. The use of UKCEH developed models ensures the flexibility required for implementation of a digital twin, as highlighted in Section 4.

- Case study 1 was chosen as it focusses on utilising a water quality model that structurally complies to many requirements of digital twin. It lends itself to modularisation into a component-based system that will allow us to fully realise the vision of digital twins as a two-way learning process, where not only are models informed by monitoring data, but model output is used to reconfigure the model to best suit the given environment (offering a learning opportunity about that environment), and to optimise the monitoring network to provide the most pertinent data to the model and stakeholders. The model could relatively easily be adapted to model nutrients in rivers and lakes.
- In contrast, case study 2 was selected as it represents an opportunity to strengthen, within a digital twin framework, a methodology of model coupling already in place within UKCEH. This work focuses on the simulation of nutrient and phytoplankton dynamics in rivers and lakes. Crucially, this model coupling is currently underdeveloped, relying on manual approaches to transferring data between river and lake models. Here, digital twinning would create an opportunity to strengthen considerably UKCEH capabilities in modelling water quality and ecological states in connected freshwater ecosystems; automating model coupling and data integration and boosting our capacity to understand cross-system linkages that are ecologically and societally important. These models operate at waterbody scales but could be incorporated into a structure similar to that proposed in case study 1, adding considerable potential to simulate catchment scale processes and capture terrestrial-aquatic processes.

Thinking more broadly, there is also potential to implement both case studies within the same digital twin. There are multiple ways in which this case study integration could be implemented. If the determinand(s) of interest is common to both case studies, then either machine learning could be used to configure which (part of each) case study model(s) is used for predictions based on which best fits monitoring data, or alternatively an ensemble approach could be used where



(weighted) average results from both case studies could be used. If different determinands are of interest and they are not common to both case studies, then the case studies could be used separately to predict different determinands. Taken together, these case studies represent an opportunity to greatly enhance our capabilities in understanding and predicting water quality change and deterioration; an issue that has considerable social and political resonance.

5.1 Case study 1: FASE

The FASE (Fate And Speciation in the Environment) model is a catchment-or-broader scale spatiotemporal multimedia contaminant fate and exposure model. It was originally developed for nanomaterial exposure models and released as the NanoFASE model (Harrison et al., 2021) (Figure 6), but it is currently undergoing adaptation to pharmaceuticals and microplastics. In its current formulation, it is particularly suited to predicting terrestrial and surface water concentrations of particulates.

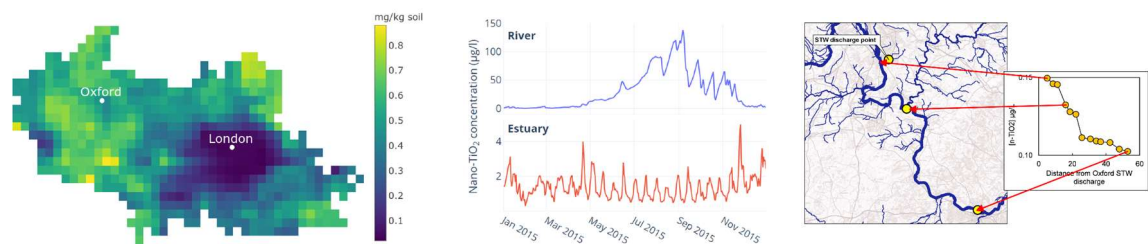


Figure 6. Example NanoFASE model outputs showing nano-TiO₂ concentrations in soils (left) and surface waters (middle, right) in the Thames catchment.

It is a gridded model, with a default (but flexible) spatial resolution of 5 km. Therefore, by default, it offers coarser spatial resolution than many traditional water quality models, such as QUESTOR. Again, by default, it uses a daily timestep, though this is split internally into sub-daily timesteps for surface waters when the model determines that daily timesteps would cause numeric instabilities. In terms of surface waters, it models rivers and estuaries and an extension to lakes and reservoirs is currently being developed, though this is not likely to be as sophisticated a model as dedicated standing water models such as PROTECH. Like other water quality models, inputs via atmospheric deposition must be provided by input data, and there is no atmospheric resuspension removal/transport process (which might be relevant for contaminants with hydrophobic properties such as PFAS, which can undergo significant atmospheric resuspension and transport (Sha et al., 2022)).

FASE requires a variety of meteorological, hydrological, and terrestrial input data (Figure 7). Usually, a separate hydrological model is used to provide surface and

subsurface runoff, which the model then routes in its surface water network. In the context of a digital twin, with the potential for real-time river flow monitoring data, this routing could be bypassed and flow data from monitoring could be used, likely supplemented by the data science approaches described above.

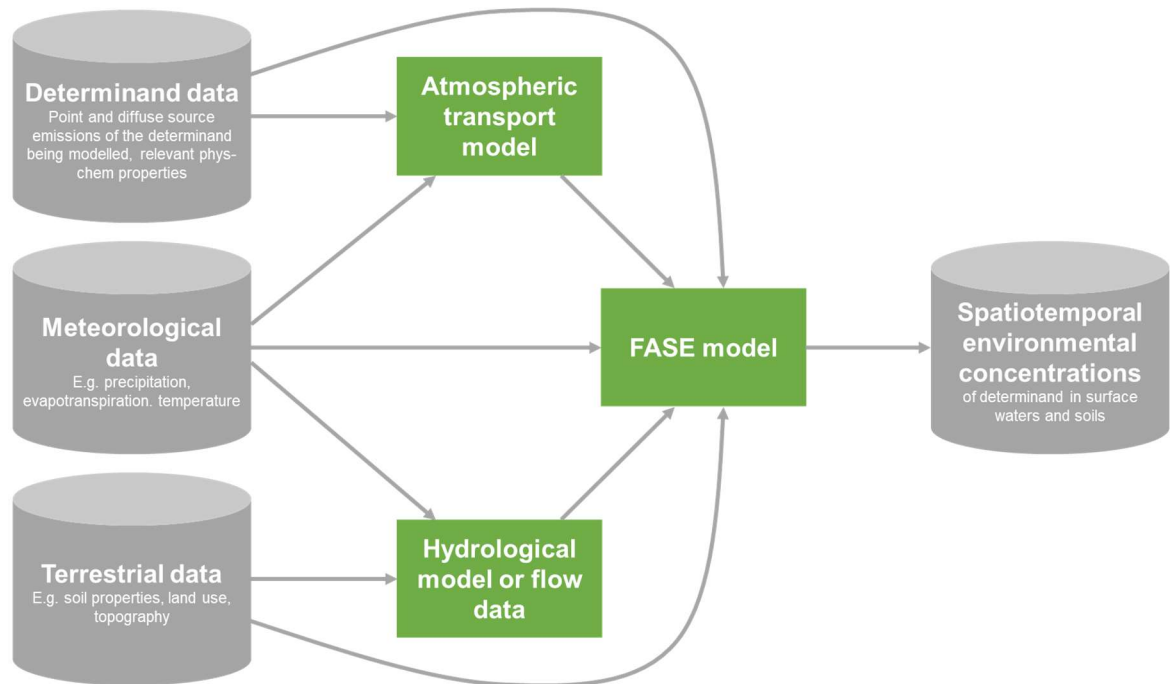


Figure 7. Typical data requirements and other models required to use the FASE model. Whether atmospheric model is required is dependent on the determinand being modelled.

Benefits of digital twinning for FASE

Over and above the next case study, FASE offers the potential advantage of including a specific model of the terrestrial environment, and it can be adapted to broader-than-catchment scales (though there was limited stakeholder demand for this). If the focus of the digital twin is anthropogenic contaminants such as plastics or down-the-drain chemicals, FASE already offers a framework for modelling such contaminants, and may be more easily adapted for this purpose than nutrient-focussed models. However, it is worth noting that the model does not currently simulate nutrient dynamics, which was seen as a key stakeholder need. Extension of the model could be made to facilitate this, though if the other advantages of using the model are not relevant for the particular use case, then it would be more logical to use a nutrient model like QUESTOR.

The model is written in modern Fortran, has been used extensively in cloud environments and is well tested within open-source software stacks. Therefore, its technical implementation in the cyber infrastructure hosting the digital twin should

be straightforward. Dynamic coupling can be achieved through a prototype implementation of the Basic Model Interface (Hutton et al., 2020), which so far has been used to test coupling the model to a hydrological model framework (eWaterCycle, Hut et al., 2022). Another advantage of FASE is that flexibility and modularity were a core requirement for its design, thus the model is written using an object-oriented approach. It is thus worth highlighting in the context of digital twins that, UKCEH has an Envision PhD project beginning in October 2024 that will be working to split the FASE model up into a component-based model using microservice architecture (see Section 4.1).

FASE within a digital twin

In its simplest form, the entire FASE model could be used in a digital twin system to predict near real-time water quality using sensor data, such as water flows from the National River Flow Archive. However, this would not embrace the aforementioned modularity and would be more akin to a digital “shadow” – there would be no two-way learning to optimise model results, it would simply be the monitoring network providing data to the model.

A more sophisticated approach would be to employ the FASE model as a set of components, e.g. microservices, and use the monitoring data to optimise exactly which components are used in a given scenario – a so-called self-assembled model system. This idea is outlined in Figure 8. More specifically:

- Each component could represent a particular physical process, such as soil erosion. For some processes, multiple algorithms will be available, and which algorithm produces the most realistic results is likely to depend on many factors, such as geographical location, catchment characteristics, or time of year.
- Driving data for the model, such as water flows, temperature, and pH, could be provided as input from the monitoring network. This may require the development of a data science approach to process the point data into the spatially gridded format required by the model. A “best guess” model configuration (selection of components) would be used to produce initial water quality predictions. For catchment scale predictions, this might also mean the implementation of a new or modified monitoring network, likely based on automated sonde measurements to provide (near) real-time observations at high frequency.
- If the monitoring network is capable of measuring any of the prediction variables, e.g. suspended sediment concentrations, then the predictions would be compared against these data and an error calculated. Again, a data science approach may be needed to process monitoring data. There is less need for these monitoring data to be real-time, and so here we could rely on manual sampling data, for example from the EA Water Quality Archive.



- This process would be repeated iteratively within either a numerical solver or another appropriate data science method, whereby the choice of components is varied until the optimum model configuration – i.e. the one with the smallest error – is chosen. The same principle could apply to calibration parameters required by individual components. This configuration is used to produce final water quality predictions, which are fed to a data visualisation platform for users to view.
- There is also the potential for model predictions to help refine the monitoring network itself, for example by moving sondes or sampling locations to regions that the model indicates are most relevant to study, such as where there are dynamic fluctuations in drivers or determinands, and high levels of prediction uncertainty.

This optimisation could be implemented on a cloud server (or serverless technology) as an automated process that happens continually, thereby always providing the best model configuration. If the digital twin is applied across a broad domain (e.g. at the national scale), the model system could be configured separately for different parts of this domain to produce the best fit. For example, different catchments might benefit from different model configurations, such as the use of different soil erosion algorithms for upland and lowland areas. This enables us to enact the *models of everywhere* paradigm within the same model system and without the requirement to configure and provision separate models for separate regions. It also lets us use the model as a learning process – to teach us about the physical processes that dominate in different domains.

The model is somewhat computationally intensive, and this might make iterative model runs in a numerical solver a lengthy process. If this became prohibitive to this vision of a digital twin, another option would be to create a model emulator, i.e. a machine learning algorithm that is trained to produce outputs as close to the process-based model as possible. The emulator could be periodically tested against the process-based model, and regularly against observational data, raising the prospect of using the emulator to identify improvements to be made to the process model.



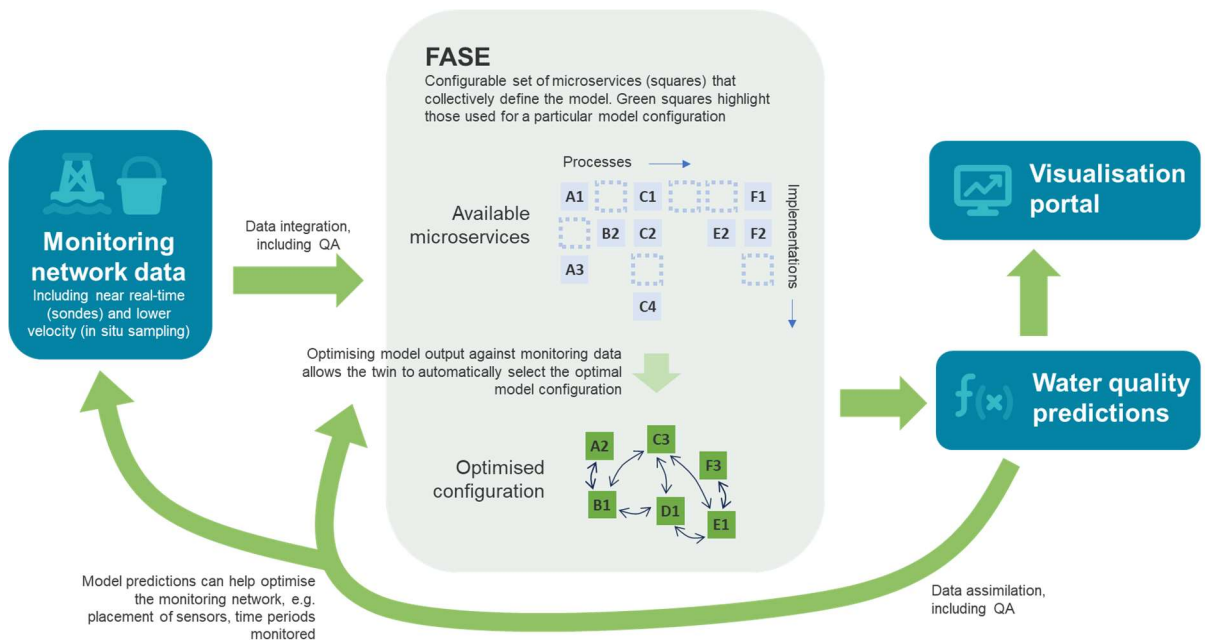


Figure 8. Schematic example of using FASE as a microservice-based configurable framework that is automatically configured based on monitoring network data.

5.2 Case study 2: nutrient and phytoplankton dynamics in rivers and lakes

The construction of a digital twin of nutrient and phytoplankton dynamics is much needed and highly relevant to current public water quality concerns. Such a tool would directly address key determinands identified by our stakeholders and provide functionality to advance understanding and management of pollution from wastewater treatment infrastructure and agricultural land (priorities identified by Stephenson et al., 2024). Furthermore, current UKCEH modelling activities in this arena have brought us to a state of readiness to develop such a twin. As a foundation we have, and are using, suitable models, namely:

- LAM (Load Apportionment Model)
- QUESTOR (Quality Evaluation and Simulation Tool for River Systems)
- PROTECH (Phytoplankton Responses to Environmental Change)

Inputs for these models are available and pre-processed outside the remit of the modelling suit. Similarly outputs from these models are currently analysed using a range of visualisation tools outside the modelling suit. What is currently missing is the integration of these disparate activities through advanced digital infrastructures. Here, we provide a brief description of the set of models currently in use.

LAM quantifies point and diffuse nutrient inputs by modelling their contribution to river nutrient concentrations as a power-law function of flow (Bowes et al., 2008). It is a statistical model that uses observed data (empirical component) for water quantity and water quality. It apportions sources of nutrients and other pollutants and predicts impacts on nitrogen and phosphorus loads of changing sewage treatment regimes, climate, and water flows. It is a simple model that does not require other catchment and land use information. The full version of the model is available as an Excel spreadsheet using macros. A newer version for phosphorus is available in R and will in time be deposited onto EIDC. This new version can access input data via CSV files or directly from the data source (e.g. NRFA data and Water Quality Archive data) via APIs where possible.

QUESTOR is an hourly 1-D model of river networks used for simulating eutrophication including dynamic solute transport based on a mass-balance approach (Pathak et al., 2021 & 2022). The river model produces time series (daily or hourly) of flow, temperature, nutrient and sediment concentrations, chlorophyll (phytoplankton biomass) and dissolved oxygen. The model has been used to assess water quality in a variety of UK river catchments ranging from small catchments of circa 50 km² to large river basins of approximately 10,000 km² (e.g. Hutchins et al., 2020; Hutchins et al., 2021). The river network is user-defined as a set of interconnected reaches bounded by different types of influences, namely weirs, abstractions, effluents, and tributary rivers. The model is programmed in Fortran and the input data required are stored within ASCII files. QUESTOR provides a comprehensive representation of biogeochemical processes driven by rates calibrated using water quality observations or estimated from other applications or literature.

PROTECH is a daily 1-D process-based phytoplankton community model that simulates the development of multiple algal populations or algal types in a lake or reservoir. Its developmental history and applications over the last two decades are well documented (Elliott et al., 2010; Elliott 2021). The basis of PROTECH is the interaction between the species-specific growth equations and the algae's environment. The bulk of the model is the environmental simulation which first feeds the biological response equations (e.g. for growth, grazing, movement) and then calculates the environmental impact for the next iteration. The logic relies on a series of loops, invoking a series of interdependent subroutines.



Benefits of digital twinning for PROTECH, QUESTOR and LAM

Currently, the linking of these models follows a data-coupling approach, achieved through the manual transfer of output files from one model to be used as input files by the other, as shown in Figure 9. This implementation, and the use of static input files, prevent near-real time configuration. Furthermore, the river and lake results are currently decoupled, and the outputs of the lake model PROTECH do not feed back into the catchment scale river model QUESTOR.

Integrating these models within a digital twin framework would permit near-real time water quality predictions of higher accuracy. Digital twinning would allow us to bridge important omissions currently identified in the manual coupling approach:

- Complete integration over catchment scale: currently the fluxes of algae biomass and nutrients are only considered from the river to a lake. In a digital twin, the entire catchment can be represented, including transfer of algal biomass from lake to river. This would enable better understanding of spatiotemporal changes in the water quality status of the whole system. Such a tool would be especially useful in a lentic-lotic continuum system (Figure 10), for the representation of algal flushes during high-flow periods.
- Access to near-real time data: As demonstrated in the data review (Section 3.1), water quality data can be sparse, in particular at a daily and/or sub-daily resolution. The NRFA data, and both the Water Quality Archive data and the sonde measurement data collected by the EA can be accessed automatically from the data provider via APIs. Such data access would enable near-real time water quality estimations. For example, data-driven models like LAM and other machine learning methods (Section 3.3) enhance continuous data integration. As such, this integration of approaches could improve process-based model performance and predictability. Including the use of remote sensing (e.g. UKCEH Land Cover map, NASA Surface Water and Ocean Topography data) and earth observation approaches would add considerable value to the digital twin.

To address these issues and achieve the inclusion of these modelling building blocks (LAM, QUESTOR and PROTECH) within a digital twin, the models will need to be re-written using a modular, component-based infrastructure such as the one described in Case Study 1 (Section 5.1).

Ecological and water quality forecasting is an important application of such a digital twin. The integration of the above modelling building blocks within a digital twin, along with data streams from multiple sensors, machine learning models, and process-based models, creates opportunities for more accurate and timely forecasting of impending high-risk events in the river network. This, in turn, enhances our ability to give early warnings, and determine mitigation options before events become critical. For example, the digital twin would enable the simulation of reservoir ecological status and algal bloom risk using information from connected rivers and could be used to test the impacts of different droughts scenarios (Elliott &



Bowes, 2022). The use of short-term river flow forecasts, such as Hydrological Outlook data, as input into a LAM approach could produce short-term forecasts of nutrient concentrations in the catchment system. Such an approach has been applied to estimate changes in nutrient concentrations under climate change (Charlton et al., 2018), and would build upon prior experience of applying PROTECH to short-term forecasting (Page et al, 2018). Then, QUESTOR could be used to assess changes at the river network scale, in a time series map of the watershed.

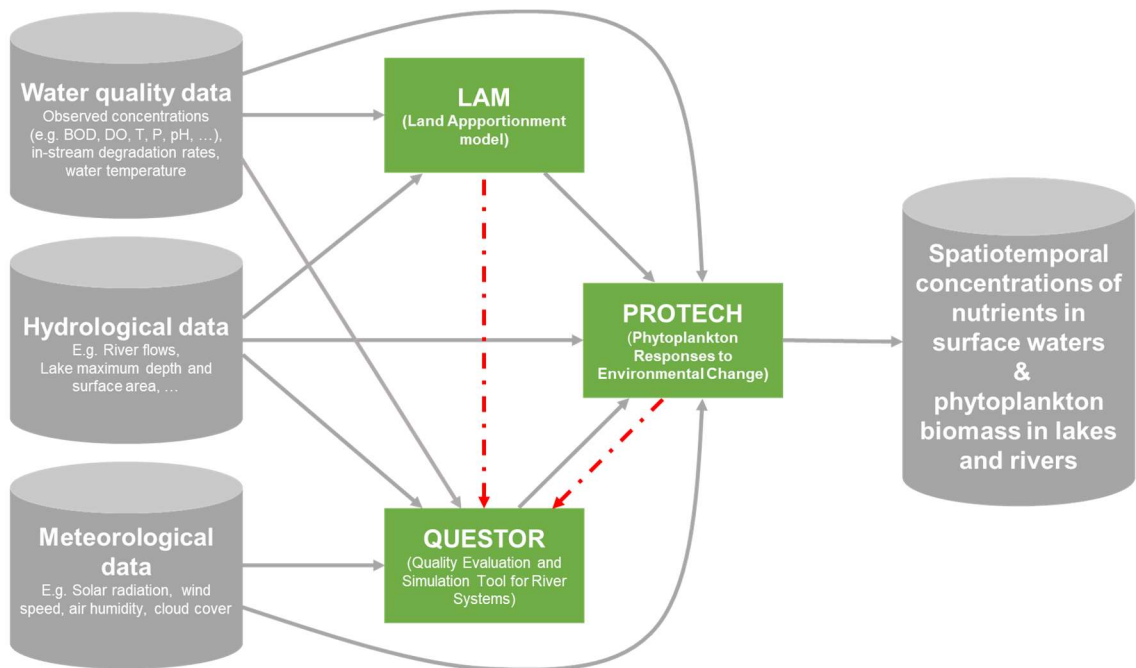


Figure 9. Typical data requirements and model linkages currently in place to simulate nutrient and phytoplankton dynamics in rivers and lakes. The red connection is anticipated to occur in future via twinning technology.

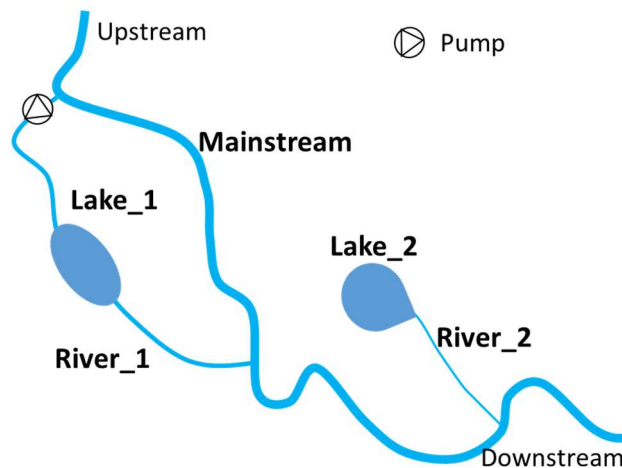


Figure 10. A conceptual diagram example of lentic-lotic continuum system.

We recognise that there are opportunities to integrate satellite data with *in situ* data streams, and models, within an operational digital twin forecasting system. Such integration would include important pre-processing steps:

1. Derivation of real-time temperature, rainfall, and river discharge estimates from relevant satellite products. Real-time temperature data can be generated from ERA-5 LAND (Munoz Sabater, 2019), precipitation from the NASA GPM IMERG v06 satellite constellation dataset (Huffman et al., 2019), and water flow from NASA's MODIS and upcoming SWOT mission. These data could be accessed and processed by Google Earth Engine, and then collated at a daily resolution for integration with *in situ* monitoring data.

2. Bias correction of satellite data through comparison with *in situ* monitoring data. Suitable *in situ* meteorological data are available from the UK Met Office, and flow data from the UK NRFA. Data-driven statistical or machine-learning models could be applied to eliminate biases in data distributions through matching of historical *in situ* and satellite data. Once pre-processed, these bias-corrected data could be used as part of the water quality monitoring framework described above. Process modelling using the bias corrected data would then yield estimations of several important system states (e.g. nutrient concentrations, dissolved oxygen, BOD, and chlorophyll-a concentration) at a catchment scale.

Key to such a forecasting system, though, is the co-design and agreement of suitable alerts and data visualisations that will be most useful to stakeholders in providing notice of possible water quality extremes (e.g., pollution discharge, combined sewer overflows discharge, high phosphate or nitrate water moving down the river system).

6. Conclusions & Recommendations

Based upon our survey of stakeholder needs, and upon our reviews of model and data availability, there is great potential to develop a digital twin focused on the cycling of nutrients through catchments of interconnected rivers and lakes, and of knock-on effects on algae (phytoplankton) growth and oxygen concentrations. Such a digital twin would also align well with recent priorities for fresh water, identified through the British Ecological Society (Stephenson et al., 2024). UKCEH, through decades of freshwater water quality and ecosystem modelling expertise, has developed a range of models that are suitable to form the core of this new technology. This digital twin could be developed through the integration of existing process modelling elements (QUESTOR, PROTECH), empirical models (LAM, machine learning), and existing water quality observational data from both manual sampling and automatic sondes. There is potential also to build upon this approach, adding the dynamics of other contaminants and adopting a component-based modelling system using FASE.

There is likely to be significant work required in re-engineering these models to make them best suited to a digital twin system. There will be a need to realise advances in automation and near-real time processing of data and running of models. Linking with other projects where such tasks are already seeing focussed efforts – such as the Net Zero UK-SCAPE integration project (where a code base has been created to automate the running of the JULES land surface model, along with the near-real time downloading and downscaling of meteorological driving data) and the Land Insight Digital Twin exemplar (where interactive dashboards for querying Digital Twin outputs were created) – will allow unified approaches to shared problems across the organisation. This will require software engineering expertise, which we are confident the organisation has or will have due to a recent focus on employing research software engineers (amongst other relevant roles such as software architects and user design experts).

Our findings on the heterogeneity of observational data availability (and likely quality), as well as stakeholder concerns regarding a lack of sufficient water quality data, suggest that initial work should identify areas with dense data coverage, for multiple determinands. It would be especially beneficial to focus on locations with high resolution sensor data that can be incorporated into data assimilation algorithms. These locations would be optimal for the development of place-based digital twins. The intricacies of heterogeneous data must also be considered, such as changes in detection limits, operational deployment and redeployment of sondes. These metadata are not accessible when accessing the data using modern technologies such as APIs. It is therefore vital to involve data providers during the



design and throughout the lifetime of the digital twin to ensure that the data are used in the most appropriate manner.

Based upon this report, we have identified several important steps that would need to be undertaken as part of the overall roadmap towards integrating existing models and data within a digital twin:

- 1) Identify any other digital twins in development that could federate with that proposed here, and opportunities to share learning and realise linkages between twin architectures.
- 2) Rewrite and implement the process models in a component-based, interoperable architecture, such as using microservices. Engineer data interfaces for initialisation and model driving that can utilise observed or estimated data and can be generalised for use in state or parameter update data assimilation methodologies.
- 3) Define data requirements: this includes selecting variables to predict (based upon community priorities) and assessing input data availability, resolution, and quality. An assessment of the accessibility and velocity of monitoring data (e.g., automated, manual) would also be conducted at this stage.
- 4) Write data processing pipelines, to enable the collection, quality assurance, modification, and delivery of data. This step is often referred to as data integration. It will contain a set code to retrieve and manipulate data required as an input, and a set of code to manipulate and disseminate output data. The latter will most likely occur via an API to provide data accessibility, including use of digital twin output in a portal. This step will also need to consider long-term data storage.
- 5) Define data science methodologies to enclose process models: preparatory methods to clean, validate and gap-fill observational data for input into data models and process models; predictive data models to sit alongside process models to target wider list of water quality indicators; post-processing methods to automate insight generation and monitor model performance.
- 6) Write a time-based job scheduler script to schedule the run of the digital twin at specific times, regularly draw data from APIs, run pre- and post-processing pipelines, automate model performance monitoring tasks and schedule regular backups.

One delivery mechanism for this development is the single-centre National Capability programme (“ACCESS-UK”), which will bring together sustained data collection at intensively monitored sites (Cumbrian Lakes, Loch Leven, the River Thames) with digital tools and expertise, known local-regional stakeholder communities, and mechanisms to co-design digital twin development with the wider UK academic, regulatory, and policy communities.

Successful delivery of a functional digital twin, or perhaps several that can federate when required, will require a strong collaborative ethos across and beyond UKCEH.



The data, models, ecosystem and process understanding, and data science and software engineering expertise required for delivery, span sites and Science Areas. As such, we must invest time, energy and resources into bringing together and sustaining a transdisciplinary team that can deliver this new capability at the whole-UKCEH scale and, in doing so, demonstrate community leadership.

Acknowledgements

This work was conducted under the *Develop a roadmap for a UKCEH water quality digital twin (WADITI)* project, funded under the NERC funded UK-SCAPE project (grant number NE/R016429/1).

References

Bell, V.A., Naden, P.S., Tipping, E., Davies, H.N., Carnell, E., Davies, J.A.C., Dore, A.J., Dragosits, U., Lapworth, D.J., Muhammed, S.E., Quinton, J.N., Stuart, M., Tomlinson, S., Wang, L., Whitmore, A.P. & Wu, L. (2021). Long term simulations of macronutrients (C, N and P) in UK freshwaters. *Sci. Total Environ.*, 776, 145813, <https://doi.org/10.1016/j.scitotenv.2021.145813>.

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

Blair, G. S., Henrys, P., Leeson, A., Watkins, J., Eastoe, E., Jarvis, S. & Young, P. J. (2019a). Data Science of the Natural Environment: A Research Roadmap. *Front. Environ. Sci.*, 7, <https://doi.org/10.3389/fenvs.2019.00121>

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.(2019b). Models of everywhere revisited: A technological perspective. *Environ. Modell. Softw.*, 122, 104521, <https://doi.org/10.1016/j.envsoft.2019.104521>.

Blair, G. S. (2021). Digital twins of the natural environment. *Patterns*, 2, 100359, <https://doi.org/10.1016/j.patter.2021.100359>

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

Bowes, M. J., Smith, J. T., Jarvie, H. P. & Neal, C. (2008). Modelling of phosphorus inputs to rivers from diffuse and point sources. *Sci. Total Environ.*, 395, 125-138, doi:10.1016/j.scitotenv.2008.01.054



Buchhorn, K., Santos-Fernandez, E., Mengersen, K. & Salomone, R. (2023). Graph Neural Network-Based Anomaly Detection for River Network Systems, arXiv 2304.09367, <https://doi.org/10.48550/arXiv.2304.09367>

Chambers, J. M., Wyborn, C., Ryan, M. E., Reid, R. S., Riechers, M., Serban, A., Bennett, N.J., Cvitanovic, C., Fernández-Giménez, M.E., Galvin, K.A., Goldstein, B.E., Klenk, N. L., Tengö, M., Brennan, R., Cockburn, J. J., Hill, R., Munera, C., Nel, J. L., Österblom, H., Bednarek, A. T., Bennett, E. M., Brandeis, A., Charli-Joseph, L., Chatterton, P., Curran, K., Dumrongrojwathana, P., Durán A. P., Fada, S. J., Gerber, J. D., Green, J. M. H., Guerrero, A. M., Haller, T., Leimona, A. I. H. M. B., Montana, J., Rondeau, R., Spierenburg, M., Steyaert, P., Zaehringer, J. G., Gruby, R., Hutton, J. & Pickering, T. (2021). Six modes of co-production for sustainability. *Nat. Sustain.*, 4(11), 983-96, <https://doi.org/10.1038/s41893-021-00755-x>

Charlton, M. B., Bowes, M. J., Hutchins, M. G., Orr, H. G., Soley, R., & Davison, P. (2018). Mapping eutrophication risk from climate change: Future phosphorus concentrations in English rivers. *Sci. Total Environ.*, 613, 1510-1526.

Chen, H., Fang, C. & Xiao, X. (2023). Visualisation of environmental sensing data in the lake-oriented digital twin world: Poyang Lake as an example. *Remote Sens.*, 15, 1193, <https://doi.org/10.3390/rs15051193>

Comber, S.D.W., Smith, R., Daldorph, P., Gardner, M.J., Constantino, C. & Ellor, B. (2018). Development of a chemical source apportionment decision support framework for lake catchment management. *Sci. Total Environ.*, 622–623, 96-105, <https://doi.org/10.1016/j.scitotenv.2017.11.313>.

Cressie, N. A., Frey, J., Harch, B. & Smith, M. (2006). Spatial prediction on a river network. *J. Agric. Biol. Envir. S.*, 11 (2), 127-150.

Deltares (2024) Delft3D-FLOW User Manual: Simulation of multi-dimensional hydrodynamic flows and transport phenomena, including sediment. Available at: https://content.oss.deltares.nl/delft3d4/Delft3D-FLOW_User_Manual.pdf (last access: 27 March 2024).

Dick, J., Dobel, A., Fry, M., Harrison, S., Qu, Y., Khamis, D., Keller, V. & Thackeray, S.J. (2023). Water quality digital twin survey. Edinburgh, UK Centre for Ecology & Hydrology, 95pp. <https://nora.nerc.ac.uk/id/eprint/536363/>

Dudgeon, D., Arthington, A. H., Gessner, M. O., Kawabata, Z. I., Knowler, D. J., Lévêque, C., Naiman, R. J., Prieur-Richard, A. H., Soto, D., Stiassny, M. L. J. & Sullivan, C. A. (2006). Freshwater biodiversity: importance, threats, status and conservation challenges. *Biol. Rev.*, 163-182, <https://doi.org/10.1017/S1464793105006950>



Elliott, J. A., Jones, I. D. & Thackeray, S. J. (2006). Testing the Sensitivity of Phytoplankton Communities to Changes in Water Temperature and Nutrient Load, in a Temperate Lake. *Hydrobiologia*, 559, 401-411.

Elliott, J.A., Irish, A.E. & Reynolds, C.S. (2010). Modelling phytoplankton dynamics in fresh waters: affirmation of the PROTECH approach to simulation. *Freshwater Reviews*, 3, 75-96

Elliott, A. J., (2021). Modelling lake phytoplankton communities: recent applications of the PROTECH model [in special issue: New, old and evergreen frontiers in freshwater phytoplankton ecology: the legacy of Colin S. Reynolds] *Hydrobiologia*, 848 (1), 209-217, <https://doi.org/10.1007/s10750-020-04248-4>

European Environment Agency (2018). European Waters - Assessment of status and pressures 2018. EEA Report No 7/2018.

Elliott, J.A. & Bowes, M.J. (2022). PROTECH simulations of QE2 and Wraysbury reservoirs using scenarios based on 1921-23, 1929, 1933-35 and 1943-44 droughts. Wallingford, UK Centre for Ecology & Hydrology, 35pp, <https://nora.nerc.ac.uk/id/eprint/533929>

Fry, M., Blair, G., Blyth, E., Khamis, D., Wiggins, M. & Tso, M. (2022). "Land InSight: a Digital Twin of UK soil carbon and water". Presented at CDE22 – Construction a digital environment Conference, Birmingham 11-12 July 2022.

Garreta, V, Monestiez, P & Ver Hoef, J. M. (2009). Spatial modelling and prediction on river networks: up model, down model or hybrid? *Environmetrics* 21, 439–456. <https://doi.org/10.1002/env.995>

Hallouin, T., Ellis, R. J., Clark, D. B., Dadson, S. J., Hughes, A. G., Lawrence, B. N., Lister, G. M. S. & Polcher, J. (2022). Unify V0.1: A Community Framework for the Terrestrial Water Cycle in Python. *Geosci. Model Dev.*, 15, 9177–9196, <https://dx.doi.org/10.5194/gmd-2021-419>

Harrison, S., Keller, V. D., Williams, R. J., Hutchins, M., & Lofts, S. (2021). NanoFASE model (Version 0.0.4) [Computer software]. <https://github.com/nerc-ceh/nanofase.git>

Hassell, D., Gregory, J., Blower, J., Lawrence, B. N., & Taylor, K. E. (2017). A data model of the Climate and Forecast metadata conventions (CF-1.6) with a software implementation (cf-python v2.1). *Geosci. Model Dev.*, 10, 4619–4646, <https://doi.org/10.5194/gmd-10-4619-2017>

Hipsey, M.R., ed. (2022). Modelling Aquatic Eco-Dynamics: Overview of the AED modular simulation platform. Zenodo repository. DOI: 10.5281/zenodo.6516222.



House of Common (2022). Water quality in rivers. Fourth Report of Session 2021–22. HC 74

Huffman, G.J., Stocker, E. F., Bolvin, D. T., Nelkin, E. J. & Jackson Tan (2019). GPM IMERG Final Precipitation L3 Half Hourly 0.1 degree x 0.1 degree V06, Greenbelt, MD, Goddard Earth Sciences Data and Information Services Center (GES DISC), Accessed: 12/07/2024, 10.5067/GPM/IMERG/3B-HH/06

Hut, R., Drost, N., van de Giesen, N., van Werkhoven, B., Abdollahi, B., Aerts, J., Albers, T., Alidoost, F., Andela, B., Camphuijsen, J., Dzigan, Y., van Haren, R., Hutton, E., Kalverla, P., van Meersbergen, M., van den Oord, G., Pelupessy, I., Smeets, S., Verhoeven, S., de Vos, M. & Weel, B. (2022). The eWaterCycle platform for open and FAIR hydrological collaboration, *Geosci. Model Dev.*, 15, 5371–5390, <https://doi.org/10.5194/gmd-15-5371-2022>

Hutchins, M. G. (2012). What impact might mitigation of diffuse nitrate pollution have on river water quality in a rural catchment? *J. of Environ. Manage.*, 109, 19–26, <https://doi.org/10.1016/j.jenvman.2012.04.045>

Hutchins, M. G., Harding, G., Jarvie, H. P., Marsh, T. J., Bowes, M. J. & Loewenthal, M. (2020). Intense summer floods may induce prolonged increases in benthic respiration rates of more than one year leading to low river dissolved oxygen. *J. Hydrol. X*, 8, 100056, doi: <https://doi.org/10.1016/j.hydroa.2020.100056>

Hutchins, M. G., Qu, Y. & Charlton, M. B. (2021). Successful modelling of river dissolved oxygen dynamics requires knowledge of stream channel environments. *J. Hydrol.*, 603, 126991.

Hutton, E. W. H., Piper, M. D. & Tucker, G. E. (2020). The Basic Model Interface 2.0: A standard interface for coupling numerical models in the geosciences. *J. Open Source Softw.*, 5(51), 2317, <https://doi.org/10.21105/joss.02317>

Jackson, M. C., Loewen, C. J. G., Vinebrooke, R. D. & Chimimba, C. T. (2016). Net effects of multiple stressors in freshwater ecosystems: a meta-analysis. *Global Change Biol.*, 22, 180–189, <https://doi.org/10.1111/gcb.13028>

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

Janssen, A. B. G., Teurlincx, S., Beusen, A. H. W., Huijbregts, M. A. J., Rost, J., Schipper, A. M., Seelen, L. M. S., Mooij, W. M., & Janse, J. H. (2019). PCLake+: A process-based ecological model to assess the trophic state of stratified and non-stratified freshwater lakes worldwide. *Ecol. Model.*, 396, 23–32, <https://doi.org/10.1016/j.ecolmodel.2019.01.006>.



Jarvis, S. G., Mackay, E. B., Risser, H. A., Feuchtmayr, H., Fry, M., Isaac, N. J. B., Thackeray, S. J. & Henrys, P. A. (2023). Integrating freshwater biodiversity data sources: Key challenges and opportunities. *Freshwater Biol.*, 68, 1479-1488, <https://doi.org/10.1111/fwb.14143>

Karmous-Edwards, G., Conejos, P., Mahinthakumar, K., Braman, S., Vicat-Blanc, P. & Barba, J. (2019). Foundations for building a digital twin for water utilities. *Smart Water Report*, <https://swan-forum.com/wp-content/uploads/2022/07/Foundations-for-Building-a-Digital-Twin-for-Water-Utilities.pdf>

Kim, M. G. & Bartos, M. (2024). A digital twin model for contaminant fate and transport in urban and natural drainage networks with online state estimation. *Environ. Modell. Softw.*, 171, 105868, <https://doi.org/10.1016/j.envsoft.2023.105868>

Leigh, C., Alsibai, O., Hyndman, R. J., Kandanaarachchi, S., King, O. C., McGree, J. M., Neelamraju, C., & Strauss, J. (2019). A framework for automated anomaly detection in high frequency water-quality data from in situ sensors, *Sci. Total Environ.*, 664, 885-898, <https://doi.org/10.1016/j.scitotenv.2019.02.085>

Li, L., Lv, NN & Li, W. (2022). Research on Application of Graph Neural Network in Water Quality Prediction, *Int. J. Artif. Intell. Tools*, 31:01 2250018, <https://doi.org/10.1142/S021821302250018X>

Liang, L. (2021). Water Pollution Prediction Based on Deep Belief Network in Big Data of Water Environment Monitoring, *Sci. Program.*, 2021:8271950, <https://doi.org/10.1155/2021/8271950>

Liu, J., Wang, P., Jiang, D., Nan, J. & Zhu, W. (2020). An integrated data-driven framework for surface water quality anomaly detection and early warning, *J. Clean. Prod.*, 251, 119145, <https://doi.org/10.1016/j.jclepro.2019.119145>

McCracken, M., Cavers, S., Banin, L., Bowler, D., Bush, A., Coskeran, H., Gerard, F., Groom, Q., Henrys, P., Hill, J., Jarquin, M., MacKechnie, C., Pocock, M., Read, D., Thompson, J., Thackeray, S., Widdicks, K., Williams, E. & Roy, H. (2024). Mapping and Monitoring Biodiversity, Part I: Scoping Exercise. Final Report to Defra, 93pp.

Millennium Ecosystem Assessment, (2005). Ecosystems and Human Well-being: Wetlands and Water Synthesis. World Resources Institute, Washington, DC.

Muñoz Sabater, J. (2019). ERA5-Land monthly averaged data from 1950 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). DOI: 10.24381/cds.68d2bb30 (Accessed on 12-07-2024)

Ni, Q., Cao, X., Tan, C. Peng, W., & Kang, X. (2023). An improved graph convolutional network with feature and temporal attention for multivariate water quality prediction. *Environ Sci Pollut Res* 30, 11516–11529, <https://doi.org/10.1007/s11356-022-22719-0>



Page, T., Smith, P. J., Beven, K. J., Jones, I. D., Elliott, J. A., Maberly, S. C., Mackay, E. B., De Ville, M. & Feuchtmayr, H. (2018). Adaptive forecasting of phytoplankton communities. *Water Research*, 134, 74-85, <https://doi.org/10.1016/j.watres.2018.01.046>

Park, R. A., Clough, J. S. & Coombs Wellman, M. (2008). AQUATOX: Modeling environmental fate and ecological effects in aquatic ecosystems. *Ecol. Model.*, 213 (1), 1-15, <https://doi.org/10.1016/j.ecolmodel.2008.01.015>.

Pathak, D., Hutchins, M., Brown, L.E., Loewenthal, M., Scarlett, P.M., Armstrong, L.K., Nicholls, D.J., Bowes, M.J., & Edwards, F.K. (2021). Hourly Prediction of Phytoplankton Biomass and Its Environmental Controls in Lowland Rivers. *Water Resources Research*, 57.

Pathak, D., Hutchins, M., Brown, L.E., Loewenthal, M., Scarlett, P., Armstrong, L., Nicholls, D., Bowes, M., Edwards, F. & Old, G. (2022). High-resolution water-quality and ecosystem-metabolism modelling in lowland rivers. *Limnology and Oceanography*, 67(6), 1313-1327.

Qiu, Y., Duan, H., Xie, H., Ding, X. & Jiao, Y. (2022). Design and development of a web-based interactive twin platform for watershed management. *Trans. GIS*, 26, 1299-1317, <https://doi.org/10.1111/tgis.12904>

Qiu, Y., Liu, H., Liu, J., Li, D., Liu, C., Liu, W., Wang, J. & Jiao, Y. (2023). A Digital Twin Lake Framework for Monitoring and Management of Harmful Algal Blooms. *Toxins*, 15, 665, <https://doi.org/10.3390/toxins15110665>

Rasheed, A., San, O. & Kvamsdal, T. (2020). Digital Twin: Values, Challenges and Enablers from a Modeling Perspective. *IEEE Access*, 8, 21980-22012, doi: 10.1109/ACCESS.2020.2970143

Reid, A. J., Carlson, A. K., Creed, I. F., Eliason, E. J., Gell, P. A., Johnson, P. T. J., Kidd, K. A., MacCormack, T. J., Olden, J. D., Ormerod, S. J., Smol, J. P., Taylor, W. W., Tockner, K., Vermaire, J. C., Dudgeon, D. & Cooke, S. J. (2019). Emerging threats and persistent conservation challenges for freshwater biodiversity. *Biol. Rev.*, 94, 849-873.

Salk, K. R., Venkiteswaran, J. J., Couture, R. M., Higgins, S. N., Paterson, M. J. & Schiff, S. L. (2022). Warming combined with experimental eutrophication intensifies lake phytoplankton blooms. *Limnol. Oceanogr.*, 67, 147-158, <https://doi.org/10.1002/lno.11982>

Sample, J. & Dunn, S. M. (2014). Spatially distributed modelling in support of the 2013 review of the Nitrates Directive. CD2014_02. Available online at: crew.ac.uk/publications

Schäfer, B., Beck, C., Rhys, H., Soteriou, H., Jennings, P., Beechey, A. & Heppell, C. M. (2022). Machine learning approach towards explaining water quality



dynamics in an urbanised river. *Sci Rep* 12, 12346, <https://doi.org/10.1038/s41598-022-16342-9>

Sha, B., Johansson, J. H., Tunved, P., Bohlin-Nizzetto, P., Cousins, I. T. & Salter, M. E. (2022). Sea Spray Aerosol (SSA) as a Source of Perfluoroalkyl Acids (PFAAs) to the Atmosphere: Field Evidence from Long-Term Air Monitoring. *Environ. Sci. Technol.*, 56, 228-238, <https://doi.org/10.1021/acs.est.1c04277>

Siddorn, J., Blair, G., Boot, D., Buck, J., Kingdon, A., Kloker, A., Kokkinaki, A., Moncoiffe, G., Blyth, E., Fry, M., Heaven, R., Lewis, E., Marchant, B., Napier, B., Pascoe, C., Passmore, J., Pepler, S., Townsend, P. & Watkins, J. (2022). An Information Management Framework for Environmental Digital Twins (IMFe). National Oceanography Centre, 33 pp, <https://noc.ac.uk/files/documents/about/NOC%20IMFe%20Summary%20Report2.pdf>

Stephenson, I. Thackeray, S. J. & Ransome, E. (2024). Delivering biodiversity: priority actions for fresh water, British Ecological Society, London, UK. https://www.britishecologicalsociety.org/wp-content/uploads/2024/03/BES_Delivering-biodiversity_priority-actions-for-fresh-water.pdf

Spears, B. M., Chapman, D. S., Carvalho, L., Feld, C. K., Gessner, M. O., Piggott, J. J., Banin, L. F., Gutiérrez-Cánovas, C., Solheim, A. L., Richardson, J. A., Schinegger, R., Segurado, P., Thackeray, S. J. & Birk, S. (2021). Making waves. Bridging theory and practice towards multiple stressor management in freshwater ecosystems. *Water Res.*, 196, 116981, <https://doi.org/10.1016/j.watres.2021.116981>

Thackeray, S. J. & Hampton, S. E. (2020). The case for research integration, from genomics to remote sensing, to understand biodiversity change and functional dynamics in the world's lakes. *Global Change Biol.*, 26, 3230-3240, <https://doi.org/10.1111/gcb.15045>

Tickner, D., Opperman, J. J., Abell, R., Acreman, M., Arthington, A. H., Bunn, S. E., Cooke, S. J., Dalton, J., Darwall, W., Edwards, G., Harrison, I., Hughes, K., Jones, T., Leclère, D., Lynch, A. J., Leonard, P., McClain, M. E., Muruven, D., Olden, J. D., Ormerod, S. J., Robinson, J., Tharme, R. E., Thieme, M., Tockner, K., Wright, M., & Young, L. (2020). Bending the Curve of Global Freshwater Biodiversity Loss: An Emergency Recovery Plan, *BioScience*, 70, 330–342, <https://doi.org/10.1093/biosci/biaa002>

Tipping E., Loftis S., & Sonke J. E. (2011). Humic Ion-Binding Model VII: a revised parameterisation of cation-binding by humic substances. *Environ. Chem.* 8, 225-235.



Williams, R. J., Keller, V. D., Johnson, A. C., Young, A. R., Holmes, M. G., Wells, C., Gross-Sorokin, M., & Benstead, R. (2009). A national risk assessment for intersex in fish arising from steroid estrogens. *Environ. Toxicol. Chem.*, 28(1), 220-230, doi: 10.1897/08-047.1

Wang, F., Wang, Y., Zhang, K., Hu, M., Weng, Q. & Zhang, H. (2021). Spatial heterogeneity modeling of water quality based on random forest regression and model interpretation, *Environ. Res. J.*, 202:111660, <https://doi.org/10.1016/j.envres.2021.111660>.

Woolway, R. I., Simpson, J. H., Spiby, D., Feuchtmayr, H., Powell, B. & Maberly, S. C. (2018). Physical and chemical impacts of a major storm on a temperate lake: a taste of things to come? *Clim. Change*, 151, 333-347.



Contact

enquiries@ceh.ac.uk

[@UK_CEH](#)

ceh.ac.uk

Bangor

UK Centre for Ecology & Hydrology
Environment Centre Wales
Deiniol Road
Bangor
Gwynedd
LL57 2UW

+44 (0)1248 374500

Edinburgh

UK Centre for Ecology & Hydrology
Bush Estate
Penicuik
Midlothian
EH26 0QB

+44 (0)131 4454343

Lancaster

UK Centre for Ecology & Hydrology
Lancaster Environment Centre
Library Avenue
Bailrigg
Lancaster
LA1 4AP

+44 (0)1524 595800

Wallingford (Headquarters)

UK Centre for Ecology & Hydrology
Maclean Building
Benson Lane
Crowmarsh Gifford
Wallingford
Oxfordshire
OX10 8BB

+44 (0)1491 838800



Disclaimer goes here

