UKCEH

Our planet. Decoded.

Towards the development of a water quality digital shadow

Stakeholder engagement & prototyping

Steve Thackeray, Virginie Keller, Faiza Samreen, Oladimeji Awe, Mike Bowes, Jan Dick, Alex Elliott, Sam Harrison, Michael Hutchins, Ellie Mackay, Nathan Shaw, Yueming Qu & John Wallbank.

July 2026

Table of contents

1. Introduction

1.1 Background to the Windermere Digital Shadow Project

The water quality of freshwater ecosystems is of increasing concern to the public and policymakers, and emerges from complex interactions between physical, chemical, and biological drivers and processes. We need to improve our ability to understand and, ultimately, forecast these changes if we are to manage risks of exposure to water quality issues and implement effective management and restoration. Windermere, England's largest lake, exemplifies these public concerns around water quality, given it's known long-term responses to varying nutrient inputs, climate change, and the introduction of non-native species.

To address this need, enabled by our Digital Research Infrastructure (DRI), UKCEH are developing a Windermere Water Quality Digital Shadow by integrating process-based and empirical models with long-term and high-resolution data from a range of observation networks located in Windermere and its catchment. This initiative is supported by the National Capability for UK Challenges programme, funded by NERC, and follows a previous scoping study under the programme UK-SCAPE (Dick *et al*, 2023; Keller *et al*, 2024).

A digital shadow is a one-way digital representation of a physical object or system (such as a lake) that automatically collects and reflects/models/analyses data from it. These data flow from the physical system to the digital model, allowing you to monitor, analyse, and understand the system in near real time. The digital shadow is a precursor to a digital twin; a two-way digital representation that feeds back from the digital to the physical system by, for example, influencing the way humans interact with the physical system by influencing the design and execution of monitoring programmes.

This report arises from a 1-year data integration project commissioned by the NC-UK programme. In this project, we build upon reported UK-scale stakeholder needs with respect to water quality digital twinning (Dick *et al*, 2023) and a proposed digital twin case study on nutrient and phytoplankton dynamics in rivers and lakes (Keller *et al*, 2024). Our work is focused on the ultimate goal of a place-based digital twin for Windermere.

Given the limited timeframe of the current project, we initially focused our efforts on a digital shadow that automates the flow of *in situ* monitoring data into coupled river and lake water quality models, and their subsequent simulation of nutrient concentrations and phytoplankton dynamics. Specifically, our first step was the automated coupling of the river biogeochemical model QUESTOR (to simulate river inflows) and the lake phytoplankton model PROTECH (to simulate Windermere itself). The goal, at this stage, was to automate the modelling process end-to-end, from accessing model driving data from APIs, through the transfer of data between models, to the delivery of water quality simulations. Subsequent steps are to incorporate additional river models into the architecture, to allow for a blending of their respective inputs into PROTECH, and the creation of a digital shadow user interface.

1.2 Brief hydrological overview of Windermere

Windermere is, by area, the largest natural lake in England. The lake comprises a north and south basin, separated by a region of shallow water and islands, and the two basins differ in their morphometry and water quality (Mackay *et al.*, 2023). The north basin is deeper than the south (64 m and 42 m maximum depth, respectively) and has a longer mean water retention time (180 and 100 days, respectively). Based upon total phosphorus (TP) and chlorophyll-*a* concentrations, the lake is classified as at moderate ecological status according to the Water Framework Directive (WFD). Several other waterbodies are found in the Windermere catchment: Grasmere, Rydal Water, Loughrigg Tarn and Elterwater to the north, and Esthwaite Water and Blelham Tarn to the west. The major inflows to the lake are the rivers Rothay and Brathay, Troutbeck, Blelham Beck, Mill Beck and Cunsey Beck (Figure 1.1). Three of these inflows are monitored via an NRFA gauging station with upstream catchment draining areas varying from approximately 57 km^2 (River Brathay) to 19 km^2

(Cunsey Beck). The lake drains to Morecambe Bay via the River Leven. There is a river flow gauging station downstream of Windermere. Morphometric, water quality and catchment data for Windermere can be found via the UK Lakes Portal.

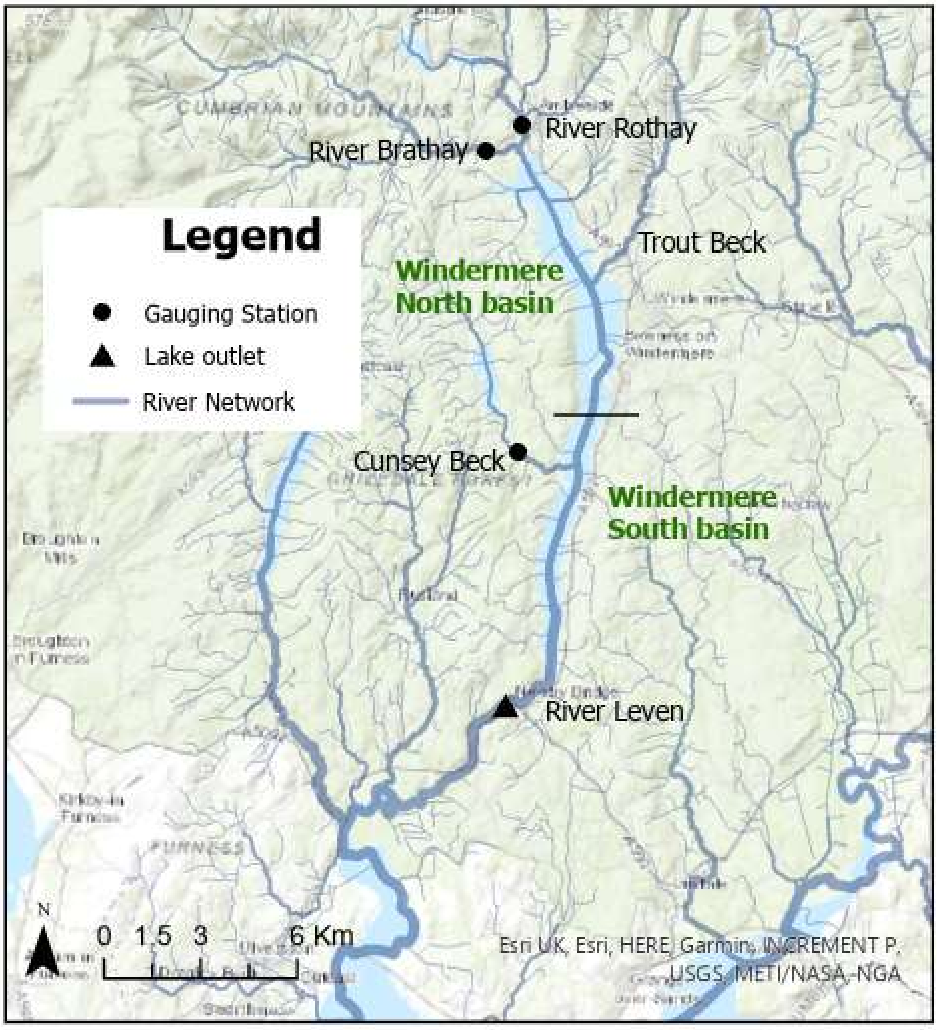


Figure 1.1 The Windermere catchment, UK.

2. Stakeholder engagement

2.1 Introduction

As part of the Windermere Digital Shadow project, it was identified that modelling tool co-design was essential to develop wider community interest in the work and ensure that the development and its outputs were fit-for-purpose in the context of Windermere. Engagement activities were focussed on a single in-person workshop, that was held on 6th November 2025, as part of existing stakeholder engagement in the area, using the Cumbrian Lakes Research Forum as a way to engage with potential users of the digital shadow. Stakeholders attending the meeting were invited via a distribution list that had developed over the last 10 years. Following an overview of the Windermere Digital Shadow concept, participants were first asked to consider what their needs were for a water quality digital shadow in the context of specific variables, locations, timescales and metrics. Information on these aspects were captured via an online survey. The second activity in the workshop was a broader discussion about what the digital shadow might look like, asking the participants to think about different components of the interface that they would like to see included. These discussions were captured on paper and later digitised and collated into themes.

The results from the workshop detailed in this report will be shared with the participants and considered when developing the digital shadow outputs. The final project outputs will be shared via a future engagement event, and a follow up survey carried out to gain further feedback that will be used to develop future project plans.

2.2 Methods

The stakeholder workshop was developed to enable the capture of two important types of information as part of the co-development of the water quality digital shadow. It builds on the broader survey carried out as part of a previous project (Dick *et al*, 2023) and therefore considered more specific issues relating to the practical implementation of the Windermere digital shadow project. We assessed 1) the specific information needs that users have around water quality and 2) what the interface that they would interact with would look like. The workshop forms an important step in the process of the project development (Figure 2.1).

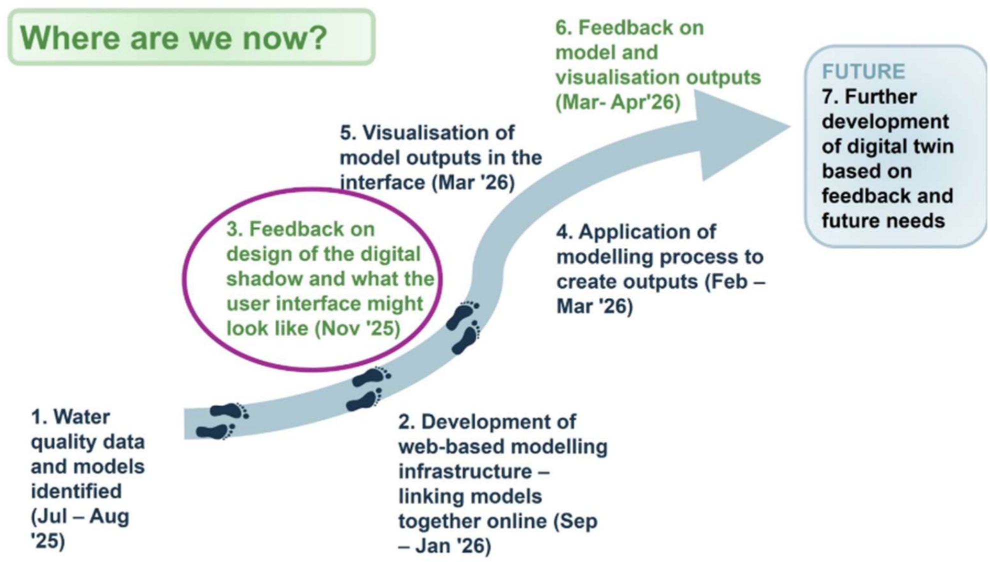


Figure 2.1 Windermere Digital Shadow development process and feedback stages (as shared at the Cumbrian Lakes Research Forum in November 2025).

2.2.1 Stakeholder workshop

39 attendees at the 2025 Cumbrian Lakes Research Forum participated in the discussion sessions. Favourable Ethical Opinion for this study was obtained from the UKCEH Human Research Ethics Committee (HREC0100). The nature of the event was explained in advance, in a 'Participant Information Sheet', along with the purposes for which workshop participant response data would be used (See Appendix 6.1).

Windermere Digital Shadow content survey
A survey process was utilised to determine the elements of water quality that workshop participants deemed most desirable for inclusion in a digital shadow of Windemere. The survey was designed as a mixture of categorical and open questions (Appendix 6.2).

Participants' feedback was captured in a questionnaire co-designed by the core UKCEH research team. The questionnaire was created using UKCEH approved software JISC, the online survey tool designed for academic research, education, and public sector organisations. Most of the questions were obligatory (i.e. respondents could not move to the next page without adding some character in the allotted box). They were advised of this fact and the reasons explained (fullness of participants views captured). The link to the questionnaire was distributed to participants via a QR code during the event (allowing 20 min for completion), and via a web link in a follow up email.

Responses to categorical questions were analysed using the UKCEH subscription-based version of Microsoft Office (updated monthly with the latest features and security patches), and the final open question was analysed with the assistance of Microsoft 365 Copilot.

Interface discussion
To enable discussion and data capture around the appearance and functionality of the digital shadow interface, an overview of the topic of interface design was provided and a question posed about how users might work through a three-step interaction with the results from the digital shadow. Workshop participants were split into groups with a facilitator who enabled the discussion within the group and recorded ideas onto flip chart paper and sticky notes, including text and diagrams. These discussions took around 30 minutes. The ideas generated were then fed back to the whole group in the final 10 minutes of the session. Following the workshop, the ideas were collated and grouped into themes.

2.3 Results

2.3.1 Survey results

Seventeen survey responses were obtained, mostly at the workshop on 6th November 2025. Only two respondents completed the survey afterwards (on the 27th November and 1st December 2025), following a reminder email after the workshop. It is not possible to tell exactly how long respondents took to complete the survey, as one participant opened the survey during the workshop (at 12:19) yet did not complete it until the evening after the workshop had finished (at 17:58), apparently taking 5 hrs and 39 min over this task. Most respondents (76%) appear to have completed the survey during the period allocated at the workshop (survey open for less than 20 min), indicating a relatively high level of engagement.

Participant characteristics
Almost equal numbers of participants self-reported their affiliation as either Government Agency, University, or Commercial Business (29 - 24%), with 18% affiliated to a Research Institute and one individual representing a Non-governmental Public Body (Fig 2.2). One participant wrote their affiliation as Water Company, which was reassigned to Commercial Business in this report.

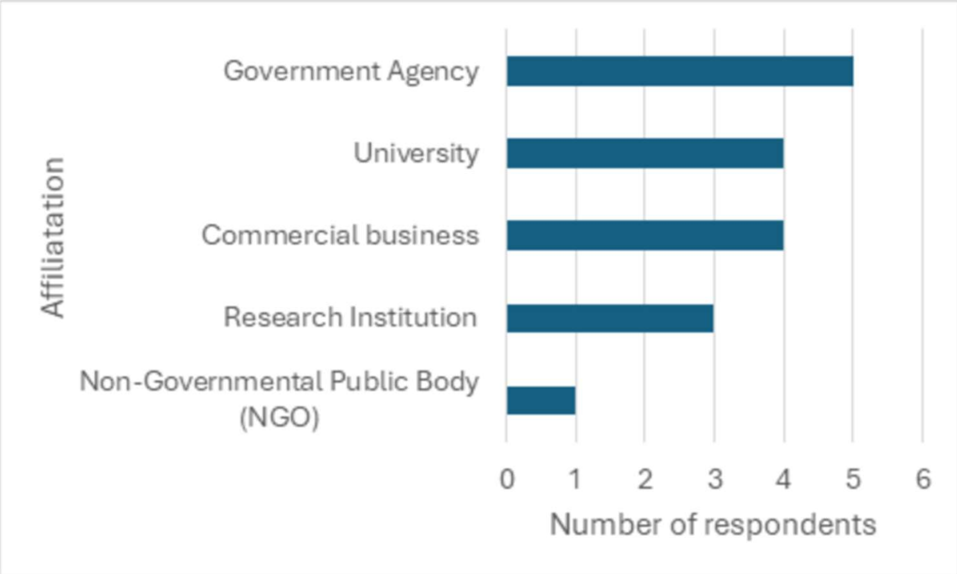


Figure 2.2 Participant responses to the question "*What type of organisation/institute are you affiliated with? (Please select all that apply)*".

Almost twice as many respondents were female (65%) compared with male (35%), and almost half (47%) self-reported their age as between 18 - 34 years. The others were relatively evenly split between age categories (i.e. 2 - 3 respondents in each category) with only one person reporting that they were over 65 years of age (Fig 2.3).

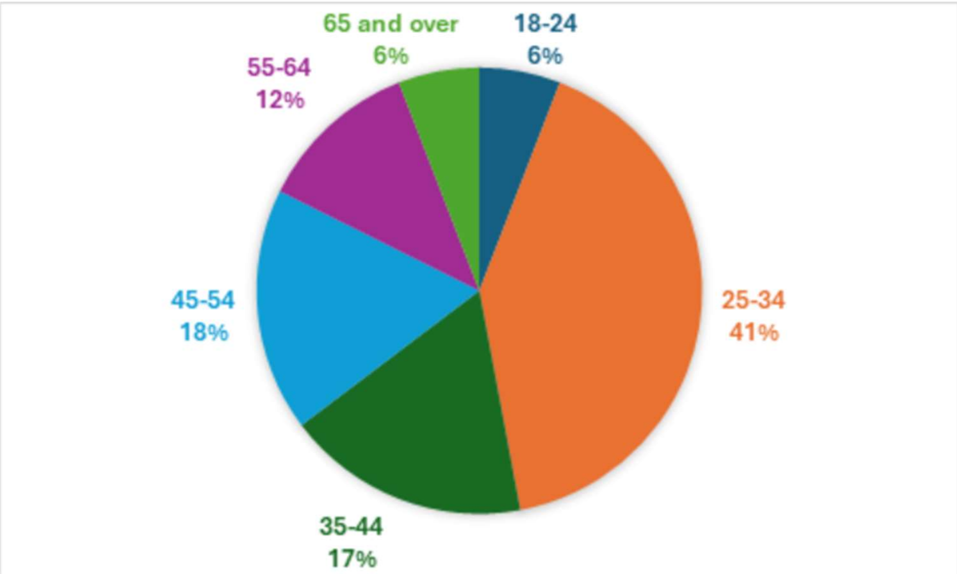


Figure 2.3 Participants responses to the question "*Which age group do you belong to?*"

Physico-chemical variables: which are desired and at which spatial and temporal scales?
Half of the respondents (Fig 2.4) selected either all of the variables suggested, or all except "Incident solar radiation (water column average for rivers)", indicating an interest in a broad range of physico-chemical variables (the variable list was presented in a randomised order in each survey). Nitrogen and phosphorus, in various forms, were the most frequently selected (94% and 88% of respondents selected these variables, respectively). Other variables selected by most respondents (82%) were dissolved oxygen, water temperature, and river water quantity (flow, velocity, depth), and dissolved organic carbon (DOC, 76% of respondents).

Most respondents (94%) wished for physico-chemical variables to be summarised at the scale of individual input rivers (Fig 2.5). Over half (59%) requested variables to be summarised, spatially, for each lake basin (North/South). In comparison, 35% and 29% of respondents desired aggregation of these variables as a total lake input (including run-off from small areas) and for the whole lake (both basins combined), respectively. Importantly, choices of spatial scale for river inputs and the lake itself were not mutually exclusive. Individual respondents requested data summaries for both individual and total lake inputs, and for both the whole lake and individual basins.

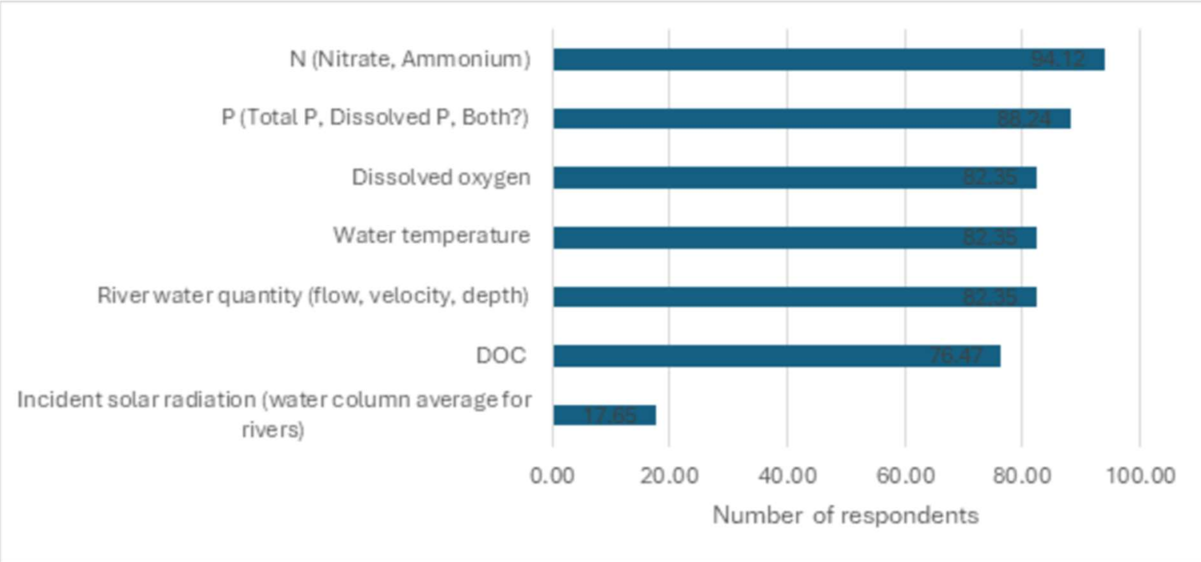


Figure 2.4 Participant responses to the question *"From your perspective, which physico-chemical variables would be the most important outputs from the digital shadow? (Please select all that apply)".*

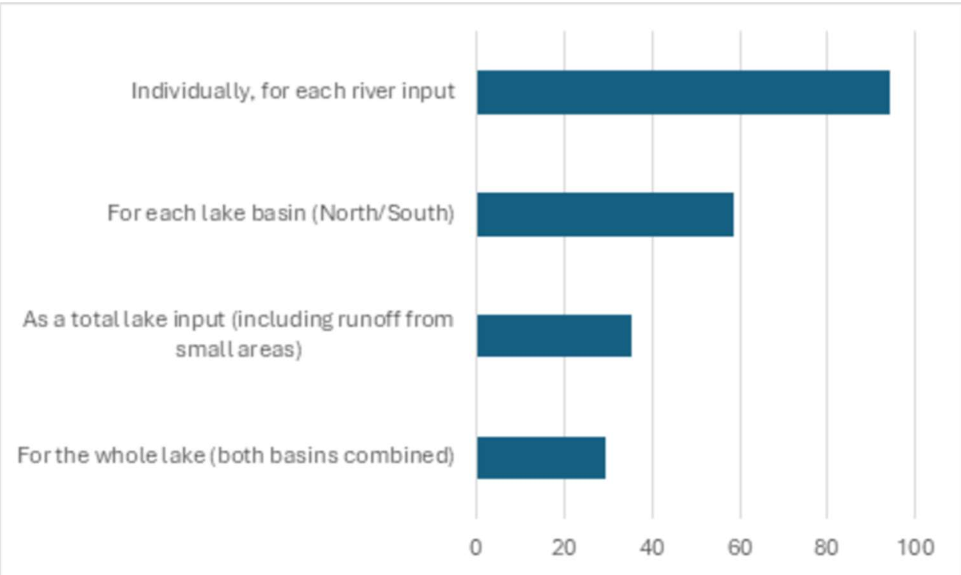


Figure 2.5 Proportion of participants responses to the question *"How do the variables you selected above need to be summarised, spatially, to be of most use to you? (Please select all that apply)".*

There was no clear overall preference regarding the time scale over which physico-chemical variables should be summarised (Fig 2.6), although the longer time scales appeared slightly more popular. Monthly and yearly scales were selected by 59%, and 53% of respondents, respectively, while weekly and daily scales were selected by 47% and 41% of respondents, respectively. Two respondents selected all temporal scales offered, and six respondents selected only a single time scale, with one specifying quarterly/seasonally (which was not one of the preassigned categories).

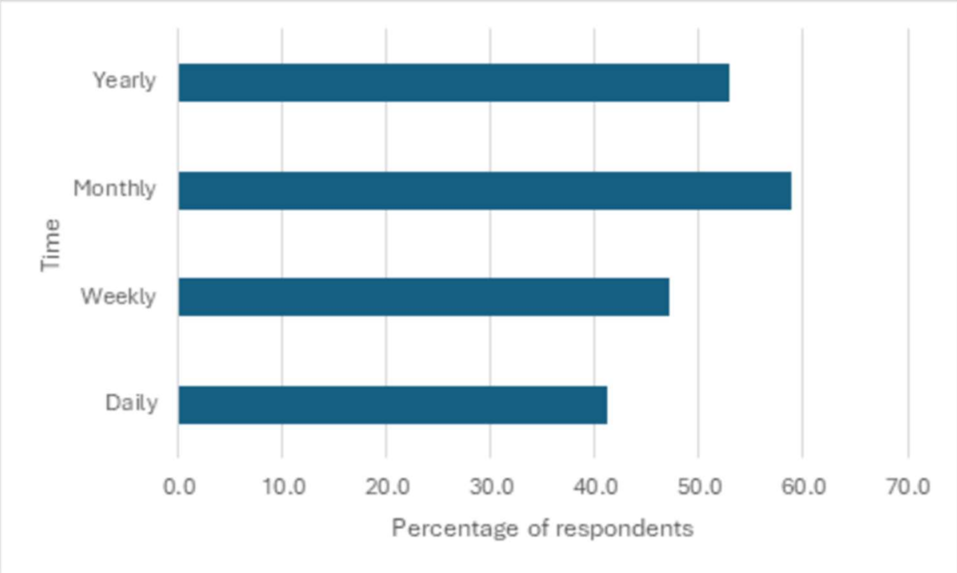


Figure 2.6 Proportion of participants responses to the question "*What timescales should these variables be output over, to be of most use to you? (Please select all that apply)*".

Algal variables: which are desired and at which spatial and temporal scales?
Many respondents (Fig 2.7) selected chlorophyll concentration as the algal variable most important to their work, while less than half (41%) selected the less precise option of information on the most abundant algal group on each day/timestep (e.g. cyanobacteria, diatom etc.). Only three respondents selected both options.

Regarding the spatial scale at which algal data should be summarised, more respondents selected for each river input and lake basin separately than options summarising the lake as a whole (Fig 2.8). However, approximately half of respondents selected more than one spatial summary option. Respondents were forced to select a single temporal resolution over which algal data should be summarised due to an error with the online survey that precluded multiple choices. With this constraint in place, monthly-resolution data were most frequently selected (Fig 2.9).

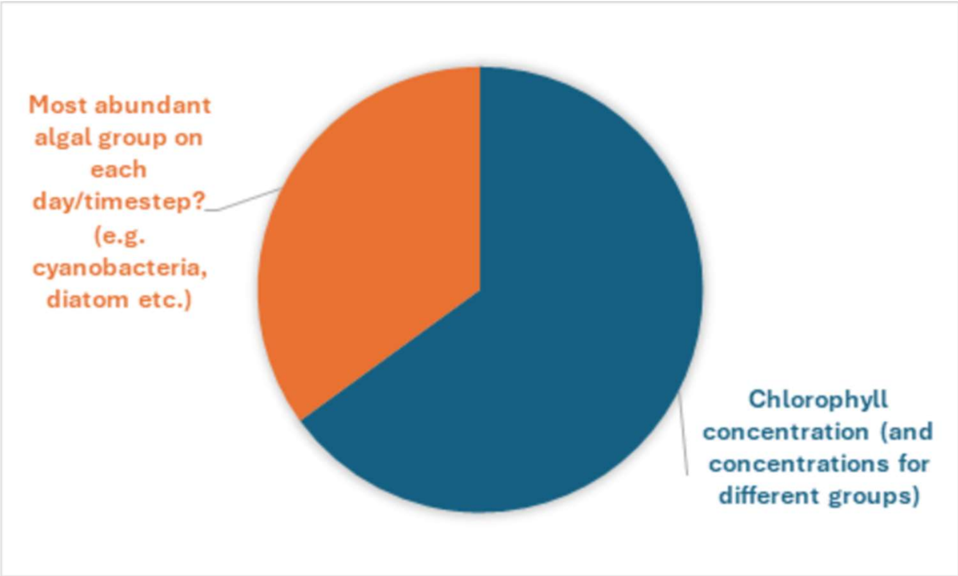


Figure 2.7 Proportion of participants responses to the question "*What algal variables are most important to you in your work? (Please select all that apply)*".

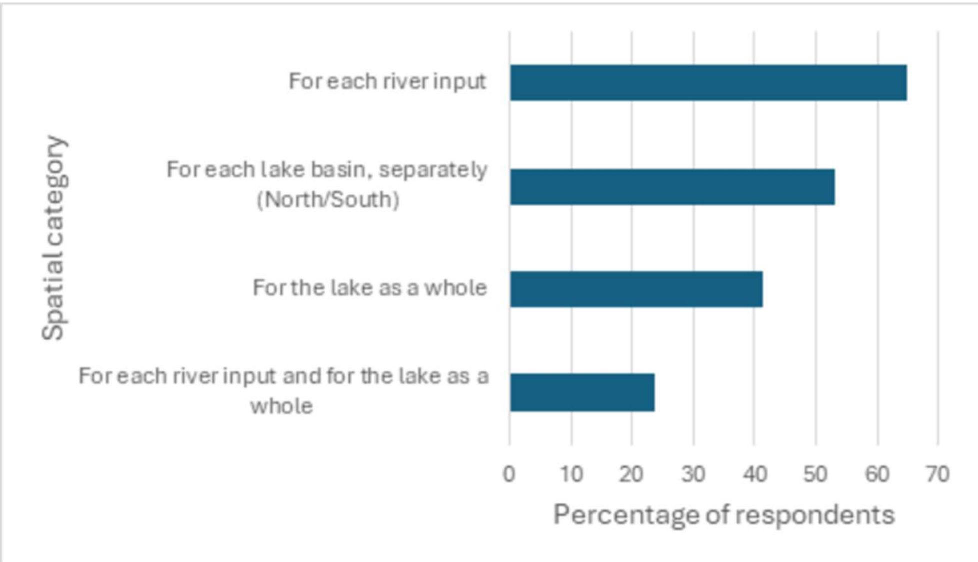


Figure 2.8 Participant responses to the question "*What spatial scales should these algal variables be output over, to be of most use to you? (Please select all that apply)*".

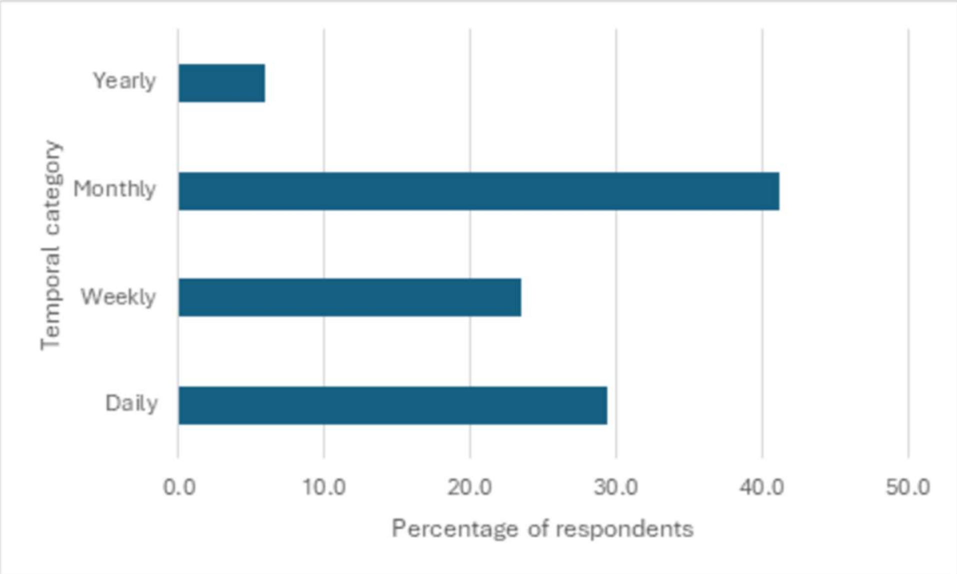


Figure 2.9 Participant responses to the question "*What timescales should these algal variables be output over, to be of most use to you? (Please select all that apply)*".

Metrics for digital shadow outputs
Our survey offered the opportunity to select different metrics to quantify the variables output by the digital shadow. The option of concentrations of nutrients, algal biomass and algal groups was selected by 82% of respondents (Fig 2.10) as a desirable metric to indicate water quality. Loads (nutrients) were selected by 71% of respondents, and just over a half (53%) considered that exceedance of thresholds would be suitable. Almost a quarter of respondents (23%) selected all three metrics offered, and no one offered an alternative metric to the options offered.

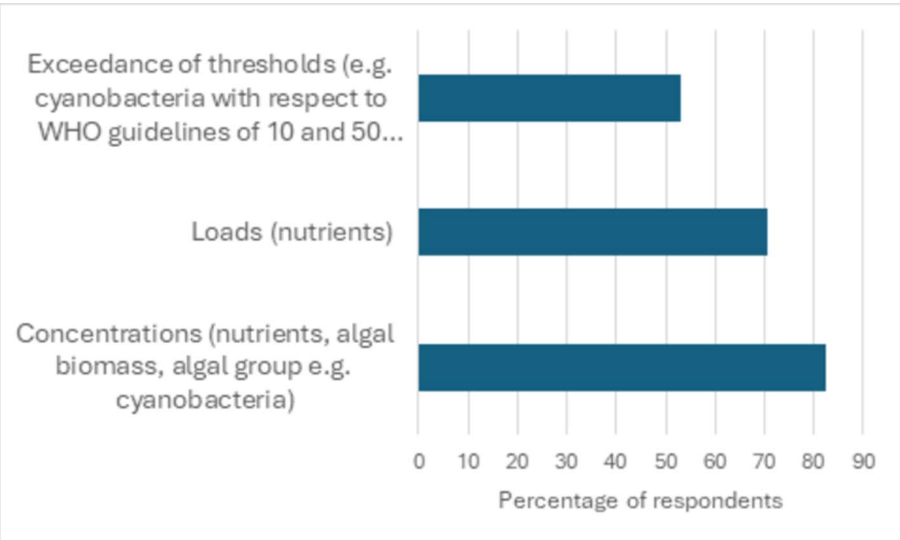


Figure 2.10 Participant responses to the question "*What metrics would you like to see from the Digital Shadow? (Please select all that apply)".*

Additional free-text responses
To conclude the survey, a final open question was asked: *We are keen to develop the digital shadow so would welcome any other thoughts you have e.g. are there different variables you would like reported with different metrics and over different timescales or is there anything else you would like to add to help us improve upon the digital shadow presented today?*

A wide variety of responses were received (Table 2.1). The two respondents who submitted responses after the workshop (and probably did not attend) stated that they required clarity on exactly what a 'digital shadow' would ultimately look like but were enthusiastic about the concept (respondent numbers #16 & #17). One respondent commented they had nothing to add (#3).

The responses obtained from those in attendance at the workshop revealed both strong interest and a relatively consistent set of priorities across stakeholders. While perspectives varied in emphasis, from operational needs to strategic planning and public communication, the overarching message was that the Digital Shadow was seen as a high value, forward-looking tool whose usefulness would depend on the scope, quality, and accessibility of data, and forecasting capabilities.

A prominent theme across the responses was the desire for more granular, site-specific information (e.g. #1) capable of showing where water quality pressures originate and how different activities contribute to local conditions. This aligns with a broader call for improved diagnostic and attribution capabilities (e.g. #6, #7, #8, #9, #10) emphasising that stakeholders value outputs that help them understand not only what is happening but also why.

Table 2.1. Participants response to the final open question "*We are keen to develop the digital shadow so would welcome any other thoughts you have e.g. are there different variables you would like reported with different metrics and over different timescales or is there anything else you would like to add to help us improve upon the digital shadow presented today?*".

Respondent number	Response to final open question calling for further thoughts
#1	Some very site specific data would be very useful in regard to showing what areas and activities are contributing to water quality issues
#2	Real time or near real time data
#3	Nothing else
#4	Need to take into account abstraction, daily/annual licence maximum but also recent actuals and potential for drought permit implementation
#5	Totals and dissolved totals of nutrients would be very helpful
#6	For incident response trigger thresholds for ammonium and DO in rivers for fish would be super useful so we can predict if fish kills are likely or not in order to prepare and respond
#7	Discussions around using flood forecasts as inspiration. Some wanted to use it as a means to look at large-scale catchment interventions whereas others like the algal bloom forecasts and nutrients and to use it as a public facing tool. This would then feed up to influencing policy by showing its value.
#8	Being able to understand conditions that result in blue green algal blooms [BGA], and in turn then being able to forecast them and communicate to the public. Some public perception that nutrients = BGA but not that simple and lots of contributing factors. Want to communicate that to people and improve scientists understanding of blooms and occurrences.

#9	The models should probably be developed with well established and extensively studied water quality parameters, but it would interesting to extend into emerging pollutants and contaminants of concern such as plastics and pharmaceuticals. A platform that people could use to predict water quality and plan sampling and leisure activities could be useful. Adding in pollutants that affect human health may lead to funding streams to help the digital shadow gain greater traction and wider reach than focusing on environmental health (though it shouldn't!)
#10	Would be really helpful to have a collection of data accessible - combining work down by different organisations. Long term forecasting would be good for planning future actions. Short term modelling good for public, and predicting WQ & ecological response to weather events etc. Sounds like a big job! :)
#11	Need to try and balance the value to strategic catchment scale planning and wider policy implications against the need for local delivery on the ground to start constructing the building blocks of change and justify interventions to funding organisations. Also balance between outputs for research and understanding, outputs for delivery and outputs for communications to the wider public and policymakers.
#12	Seasonality changes in relation to temperature, daylight hour and climate change related events such as flood, drought etc.
#13	Historical annual P load values from all STWs, and EA apportionment from all possible years back in time
#14	Would be useful to be able to view graphs of concentrations over time.
#15	Capturing more biodiversity abundance/composition based data (could be things like eDNA or zoo/phyto-plankton taxonomic counts/classification) this would be best at a monthly resolution (or more frequent if possible) to capture changing communities overtime
#16	I confess I'm a little unclear on what this 'digital shadow' will look like and how it will operate, but sounds neat and important. I look forward to seeing what emerges.
#17	Sorry I was not at the Forum but very interested (and quite excited) to see that this idea is being proposed

Several respondents highlighted or implied the need for real-time or near real-time data (e.g. #2, #4, #12), particularly to support operational decision making and rapid incident response. Closely linked to this was a strong emphasis on incident response thresholds, e.g. for ammonium and dissolved oxygen concentrations, to better predict fish kill risk (respondent #6) and enable faster, more informed interventions. These comments reflect a desire for the Digital Shadow to act as both an early warning system and a practical support tool for those managing the impacts of environmental events.

In terms of variables and metrics that could be included within the Digital Shadow, multiple respondents referenced the importance of nutrient data, both totals and dissolved fractions, as essential for understanding water quality dynamics and pressures (e.g. #5, #6, #7, #8). There was also interest in expanding beyond more established monitored variables to include emerging pollutants such as microplastics and pharmaceuticals (respondent #9), and an acknowledgement of evolving regulatory and public health concerns. Complementing these chemical measures was a call for enhanced biological data, including eDNA and plankton community composition (respondent #15), with the desire that these data should be collected regularly enough to reflect changing ecological conditions.

Several responses stressed the importance of forecasting (e.g. #6, #9 #10), with references to flood forecasting systems as an inspirational model for further progress (respondent #7). Stakeholders expressed interest in both short-term event-based forecasts (e.g., algal blooms, pollution events) and longer-term catchment-scale projections useful for planning and policy development. The need to understand and predict blue-green algae (BGA) blooms was a recurring focus, particularly the need to explain the complex drivers behind blooms, to counter simplistic public assumptions (respondent #8).

Survey respondents also raised themes of data integration and accessibility, expressing a clear wish for the Digital Shadow to act as a hub that draws together data from multiple organisations (e.g. respondent #10), enabling coherent and consistent access across sectors. There was also a call for intuitive outputs, including graphs of concentrations over time (respondent #14) and inferred a need for accessible dashboards (e.g. #7, #9, #10) that serve the differing needs of operational teams, researchers, policymakers, and the public.

Several respondents reflected on the need to balance these different audiences and functions. There was recognition that the Digital Shadow must support strategic catchment planning, inform policy, underpin local delivery, and also communicate meaningfully with the public (e.g. #7, #11). This balancing act was seen as important but challenging, particularly in terms of prioritising development efforts (e.g. #9, #10).

Seasonality, climate extremes, and long-term change were all highlighted as important elements to capture (e.g. #6, #10, #12, #13). Interest in historical datasets, including long-term phosphorus loads and Environment Agency apportionment data (respondent #13), underscores the need for strong temporal context to detect trends and evaluate management interventions.

Finally, the sentiment overall was positive, with excitement and anticipation of the final product expressed. Taken together, these responses painted a clear picture of stakeholder expectations: the Digital Shadow should be **diagnostic, predictive, integrative, and accessible**, operating at multiple timescales and serving a broad range of users. The final product is expected to combine robust scientific metrics with practical real-world applications and effective public communication.

2.3.2 Interface discussion theme

The afternoon discussion session focussed on ideas to consider when developing an interface for interacting with the digital shadow outputs. Participants discussed their ideas in smaller groups and recorded the results on flipchart paper either using text, diagrams or images to convey their suggestions. These paper outputs were then collated and the points raised grouped into seven themes. The following section provides a summary of the content of these themes.

Audience
An important theme raised by all the groups was that of who the target audience was for the digital shadow interface. This recognised that different users would expect different levels of engagement, have different needs and interests and require different levels of information regarding the tool and its outputs. There was a clear preference to be inclusive and provide a multi-function/ multi-purpose platform that would provide content for different user groups. These were categorised as the public and community groups, citizen scientists and scientists/ specialists. There was expectation that information would be provided at an increasing level of detail and complexity at each of those levels to support the needs of the user.

User interface
Workshop participants expressed that the user interface of the water quality digital shadow should be simple to navigate, allowing users to find and visualise parameters and locations of interest. As specific enablers of this,

they suggested the provision of video demonstrations of how to use the tool, and search/filtering functions that allow ready access to bespoke information. Some participants articulated a desire for a map-based interface, allowing sites to be explored across national, regional, and catchment scales, and that such an approach would be enhanced by the provision of location information using multiple coordinate systems (e.g. GPS coordinates and What3Words).

One workshop breakout group developed a detailed vision for the user interface, based around the concept of a landing page with three panes. The first pane would be an interactive spatial viewer that allows querying of data. The second was a learning zone targeted at a public audience, providing explainers of key topics, the context for the digital shadow, the constituent data, and simple forecast information. Finally, the third pane would focus on a more detailed treatment of data, including tables, summary statistics, and interactive model scenarios.

Data visualisation
Visualisation of data within the interface was another theme identified by workshop participants in all the groups. Enabling an interface to provide different plotting options, that could be selected from, was seen as important to allow users to visualise the data in ways that were meaningful to them. These audience specific visualisation could range from numerical summaries of data such as summary statistics to more visual representations of data via graphs or infographics. This included the option for plots of data at specific locations and the ability to view maps where model outputs for could be overlayed alongside other layers of information such as the location of pollution sources or the location of monitoring sites.

A second element to the discussions on data visualisation centred around the representation of modelled outputs. There was reference to the need to distinguish between monitored and modelled data, enabling a comparison to be made between model fit to observations and to provide users with a way to distinguish between different data types that they may be more familiar with or have greater trust in. Although not mentioned explicitly in this context, representation of uncertainty in modelled and observed data could be considered in this context. Participants also expressed an interest in being able to view plots of modelled outputs of scenarios to illustrate the impact of changing the model drivers and the consequences for downstream water quality.

Data availability
Amongst the workshop participants, there was great interest in the data that would underpin a water quality digital shadow, whether it be the monitoring data used to drive the constituent ecosystem models, or the outputs from those models. There was a clear need for data to be open access, and downloadable in a usable format. By providing the data behind the digital shadow outputs and visualisations, we would build a sense of trust in the system itself.

However, beyond the specific data sets feeding into the developing digital shadow (and twin), participants also articulate a need for a better way to discover and explore data collected by the multiple organisations that are active in the region (e.g. regulators, research institutes like UKCEH, the water industry, universities, citizen scientists and charities). It was felt that access to this “data landscape” for the region would be very enabling for the wider stakeholder community. More specifically, there was an identified need to be able to access and explore both summary statistics (e.g. means, minima and maxima) and visualisations of these diverse data streams.

Understanding and interpretation
Following on from the discussion about the audience of the interface, workshop participants also emphasised the need for the tool to aid understanding and interpretation of the data being presented. These ideas built on suggestions grouped under the data visualisation theme, specifically around enabling the interpretation of data in the context of other information. There were suggestions that the Windermere focal system should be contextualised within the ‘bigger picture’ of water quality, through comparison with information from a national scale. As part of the visualisation, it was suggested that different colour codes and plotting symbols be used to differentiate data sources or types and that these data were shown alongside standards and thresholds such as the site-specific Water Framework Directive (WFD) targets for water quality. Suggestions were also made about ensuring plots were easy to understand using simple ‘traffic light’ colour schemes or the plotting of ‘hot spots’ to pinpoint areas of concern or focus for intervention. It was felt that being able to show

the effect of “interventions” within the modelling was important, both the effect of positive and negative measures such as the implementation of restoration activities and the impact of pollution sources.

Several suggestions were also made in relation to the provision of background material to aid interpretation. These included the use of examples to show how to interpret results presented, providing an explanation of parameters and jargon or provision of a glossary and providing explainers for what the modelled scenarios represent and considering how they could be visualised. The most comprehensive approach to the provision of background information was the development of a handbook that would explain the system.

Another area highlighted by workshop participants was the need to communicate or visualise the likelihood of outcomes and the uncertainties associated with the modelling. This was felt to be important but also inherently difficult to communicate to broad audiences. Finally, there was concern around the misinterpretation of results and the need to include a disclaimer and encourage the use of “common sense” when using the tool and interpreting the outputs.

Infrastructure
Workshop discussions yielded other desirable features of a water quality digital shadow infrastructure. Participants noted that the interface for the digital shadow should be compatible with multiple devices (e.g. smartphones and tablets) and suggested that the tool could be made into an App, to increase ease of use. One participant suggested that access and use of the digital shadow would be greatly improved if there was no need for passwords; a feature that would require consideration from a cybersecurity perspective.

User and community engagement
The final theme discussed by the workshop participants centred around user and community engagement with the tool. Suggestions included simple additions such as the provision of explanatory text accompanying the interface to provide clarity in the purpose and the novelty of the work to more complex ideas around the ‘gamification’ of the tool. The idea here was to provide more educational information to build understanding of the knock-on effects of changes to water quality, such as the impact on the broader ecology of the lake, including effects on fish species population dynamics and links to the wider food web. Another aspect identified was the need to make connection with the real world. Connecting the tool to places within the catchment via QR codes linked to the tool via an App or web browser link would enable people to relate to locations being modelled by the digital shadow. There were also suggestions for building more participation for residents and citizen scientists with a ‘what can I do?’ element. This could link to activities led by organisations working in the catchment such as the Love Windermere partnership, and its partner organisations.

3. Water quality digital shadow prototype

3.1 Defining a "minimum viable product"

Outcomes from our stakeholder engagement activities (Section 2) provided clear direction for our digital shadow prototyping process. Based upon that feedback, we felt that early developments should focus on nutrient and phytoplankton dynamics, as focal aspects of water quality, and that we should develop a tool that will be able to represent dynamics for individual water bodies (each lake basin, and each river inflow/outflow) and produce information that can be summarised over multiple time scales, using varied visualisation approaches.

Given the limited time frame (1 year) for the current project, it was important to define an achievable, "minimum viable product" that could be developed within the constraints of the available resources, and that would represent the first essential steps for digital shadow, and subsequent digital twin, development. As a goal, we determined that this "product" should include the implementation of an integrated modelling chain (linking river and lake models) within a cyber-physical infrastructure, with automated access to the near-real-time model driving and validation data. We also decided that this structure should be modular, ensuring that it would be flexible enough to enable the inclusion of additional models and data streams in the future, as well as the virtual-to-physical system feedback loop that would transform a digital shadow into a digital twin.

To develop our proposed "minimum viable product", we focused on the automation of "external" data flows into and among models, rather than "internal" aspects of individual model structures. In this initial phase, the focus of the current work has been on the development of the architectures to enable near real-time coupled modelling of rivers and lakes, and not on the validation of the modelled outputs with respect to observed freshwater ecosystem behaviour. As a consequence, access to the developing tool has been restricted to the UKCEH modelling and development team. To develop a functional digital shadow, and twin, of water quality in the Windermere catchment, we would need to undertake such validation work to ensure that the tool can produce robust insights that can be used by decision makers. This is the ultimate goal; however, it will require additional resources and time to achieve.

3.2 Data and model flows

As a result of the scoping study under the programme UK-SCAPE (Keller *et al*, 2024), the following empirical and process-based models were identified as key components to be integrated within the modular digital architecture that underpins the water quality digital shadow:

- Phytoplankton RespOnses To Environmental Change (PROTECH)
- QUality Evaluation and Simulation TOol for River systems (QUESTOR)
- Cascade
- Land Apportionment Model (LAM)

These models, developed by UKCEH, existing observational data on water quality, river flows and climatology form the building blocks of the proposed digital shadow.

A detailed data-model flow diagram was created to enable the building of our "minimum viable product" (Figure 3.1).

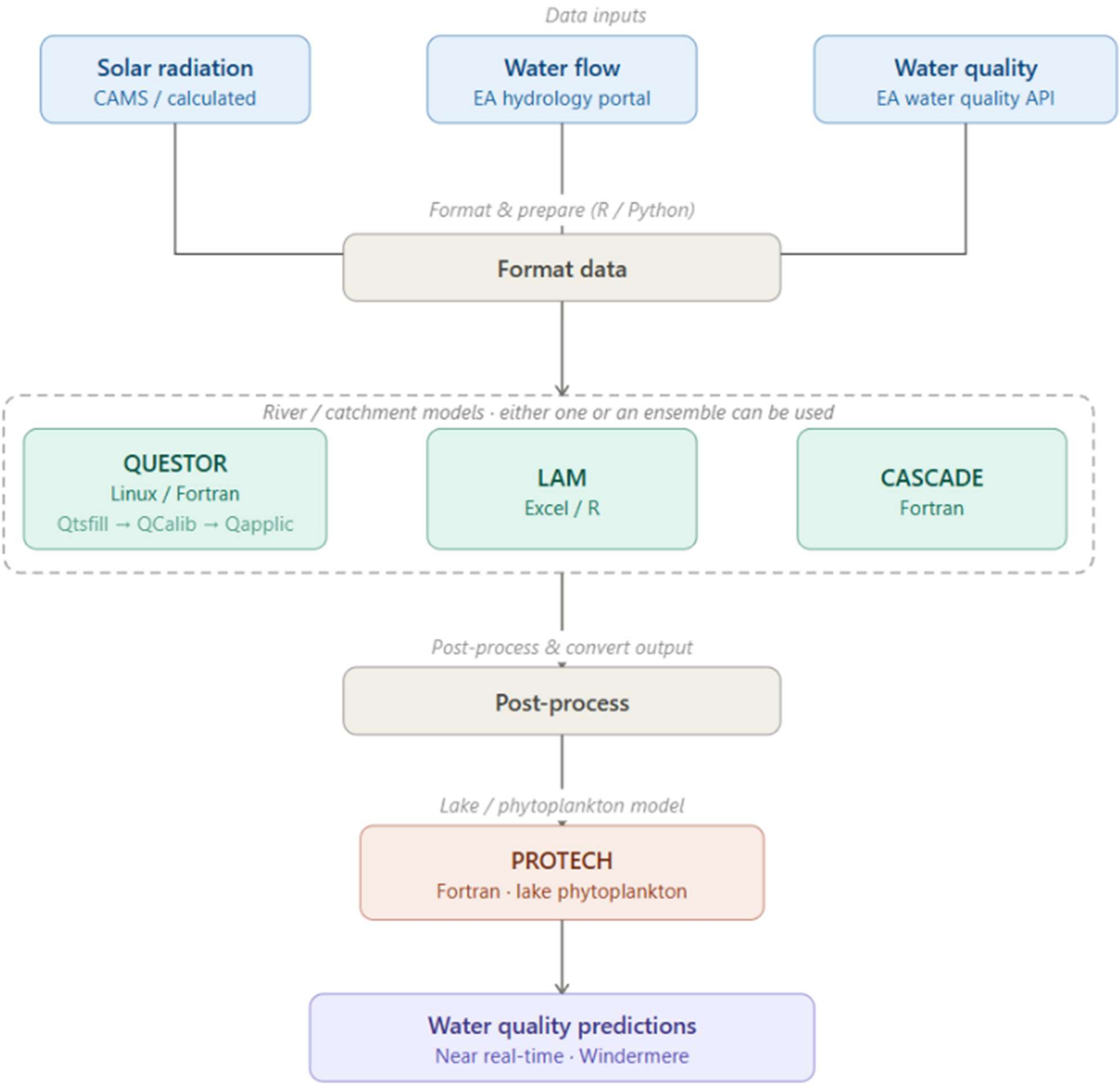


Figure 3.1. Data-model flow diagram for the prototype water quality digital shadow.

3.2.1 PROTECH

PROTECH is a process-based 1-D lake and reservoir ecosystem model that simulates phytoplankton dynamics, nutrient concentrations and physical water column processes at a daily time step. PROTECH has been used to model Windermere (Elliott, 2012; Environment Agency, 2026), splitting the lake into two connected basins (North and South). PROTECH requires, as drivers, daily river inflow discharges, associated nutrient concentrations, and meteorological data including wind speed, cloud cover, air temperature and air humidity (taken from a Met Office observation station or UKCEH's Windermere automatic water quality monitoring buoy). However, for this study we used meteorological data that had been created from climate models used in an earlier study with the Environmental Agency. River flow data and water quality can be sourced from the EA and is accessible via APIs. The three river models described below can simulate the required inflow driving data, allowing us to explore more realistic catchment scenarios. Through the current development of the water quality digital shadow, we have automated the coupling of QUESTOR (Section 3.2.2) to PROTECH. Although not currently implemented in the digital shadow, the Cascade (Section 3.2.3) model and LAM (Section 3.2.4) were set up and applied to model Windermere. These additional models are ready for future incorporation in the digital shadow. This advance is foundational to a more near-real-time model implementation.

3.2.2 QUESTOR

QUESTOR is a river water quality model that simulates phytoplankton dynamics, nutrient and sediment concentrations, water flows and temperature at hourly or daily time steps. QUESTOR is a process-based 1-D model of flow, water temperature, nutrient dynamics, eutrophication and ecosystem metabolism in river networks (Pathak et al., 2022). A network comprises a set of reaches bounded by influences such as headwater tributaries and treated effluents. Flow routing in reaches of constant-width and variable-depth is informed by riverbed condition and gradient and defined using non-linear equations to relate travel-time, water depth and discharge. By linking flow routing to biogeochemical processes (nutrient transformations, water column and riverbed heterotrophic respiration, reaeration, phytoplankton metabolism, particulate settling), the reach structure represents advection and dispersion. The model requires solar radiation, air temperature, river flow and water quality time series as inputs. A simplified version of the model was used in the tool development (Hutchins et al., 2026) which aggregates pollutant inputs and simplifies reach structure. However, the equations and time-series input data requirements are identical. This simplified model structure was developed for national scale applications but can accept data and attributes that are more locally relevant to the catchment to which the model is applied. Raw time series data are available from Met Office (air temperature), NRFA (river flows), and Copernicus (solar radiation). Catchment and river attributes define other model input values and parameters. The time series URLs and how the attributes are accessed and derived is described by Hutchins et al. (2026). Here, we used flow and nutrient data from the Windermere outflow, the River Leven, as the data input to QUESTOR, rather than collating the individual inflow data from the catchments upstream of the lake. This decision was made as a pragmatic approach to the available resources in the digital twin project, as QUESTOR was already being set up to model this river under parallel research projects, and the focus here was on developing a functional digital architecture rather than accurate simulations *per se*.

3.2.3 Cascade

Cascade is a catchment-scale model simulating water flow and water quality, typically applied at 50 m spatial and hourly/sub-hourly temporal resolutions. It is driven by gridded rainfall data, with a transition time between rainfall arriving at each grid square, and the response at selected locations: generally, a body of standing water or a river. Although Cascade was not implemented in the current digital shadow, it will form part of the digital twin and thus was set up for Windermere, ready for incorporation into the modular digital architecture (Figure 3.1). Note that Cascade can estimate river flows at any ungauged location and thus is a valuable tool for implementation within the digital twin.

Within the Windermere catchment, Cascade was set up to calculate stream flow and a series of determinand concentrations from diffuse sources (phosphate, nitrate, and DOC). All watercourses with a catchment area of at least 50 ha upon entering the lake were selected for incorporation into the model, amounting to 23 locations with a combined area of approximately 90% of the full Windermere catchment. For each watercourse, the IHDTM flow direction and elevation grids were used to estimate river slopes, widths (Davies et al. 2024), and catchment areas that were subsequently used to estimate surface water travel times based on the Manning equation. Phosphate and nitrate concentrations are estimated using export coefficients based on each grid cell's Land Cover Map class and soil-layer dependent runoff with further seasonal adjustments. For DOC concentrations an improved scheme was used, based on long-term DOC production trends in UK Upland Water Monitoring Network data, with a temperature and run-off dependent adjustment factor following Monteith et al. (2023). Table 3.1 lists the main inputs into the Cascade model, with selected gridded inputs mapped in Figure 3.2.

A preliminary comparison between Cascade outputs and observations is shown in Figure 3.3 for the Trout Beck catchment (Figure 3.2). The upper row shows the considerable agreement between the modelled and observed flow timeseries, with correlation, $r = 0.83$ and little bias (-5%). Nevertheless, there is some inaccuracy in the sub-daily modelled response, probably arising from a sub-optimal hydrological-parameter/travel-time pairing. This could be easily rectified and is eliminated if flows are aggregated to daily means ($r = 0.93$ in this case). The lower three rows of Figure 3.3 show a reasonable agreement between the typical modelled and observed nutrient concentrations. However, the agreement between modelled and observed temporal trends is poor for nitrate and phosphate ($r = 0.27$ and $r = -0.25$ respectively). In contrast, the correlation for DOC, featuring the improved production scheme, is much higher ($r = 0.79$), perhaps indicating scope for improvement for nitrate and phosphate, also.

Table 3.1: Inputs into the Cascade model. File names marked with an asterisk are shown in Figure 3.2.

File name	Resolution	Source	Description
temperature.csv	point/daily	Levens Hall (UKMO)	Temperature at known elevation
rg_out_dat/DateMMDD.csv	1km/hourly	EA Hydrology Explorer	Rainfall input (interpolation of EA RGs)
travelTime_to_monitoring.asc*	50m/static	Manning's eqn + estimated widths	Travel time to sub-catchment outlet
outflow.raster.asc	50m/static	UKCEH IHDTM	D8 flow direction grid
lcm.raster.asc*	50m/static	UKCEH LCM2007	Land cover map classes
peatysoils.asc*	50m/static	Peaty Soils (Natural England)	Class 1 ("Deep Peaty Soils"), used in DOC calculation
elevation.raster.asc	50m/static	UKCEH IHDTM	Elevation grid
accumarea.raster.asc	50m/static	UKCEH IHDTM	Accumulated catchment area
saar.raster.asc	50m/static	SAAR 1991-2020	SAAR long-term rainfall
evapsq.raster.asc	50m/static	-	Maps 50m cells to 40km MORECS square
evapBySq.asc	40km/static	MORECS PE	Long-term monthly average of MORECS PE
tomonitoringPoint.asc*	50m/static	-	Maps grid cells sub-catchment index
rainfall_index.asc	50m/static	-	Maps 50m cells to 1km rainfall input

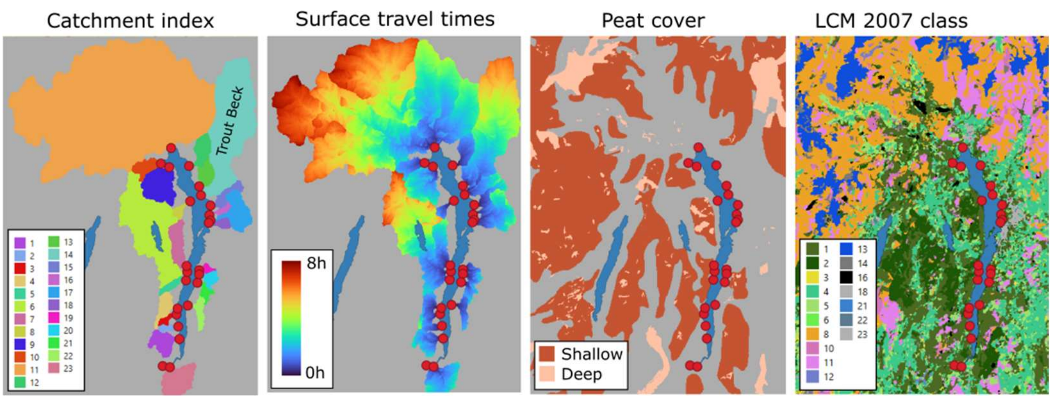


Figure 3.2: Selected static gridded inputs for the Windermere Cascade model. Large water bodies (Windermere, Esthwaite Water, Coniston Water) are shown as blue polygons. Outlets of the 23 sub-catchments are shown as red points. Outputs for the Trout Beck sub-catchment (see left panel) are explored in Figure 3.3.

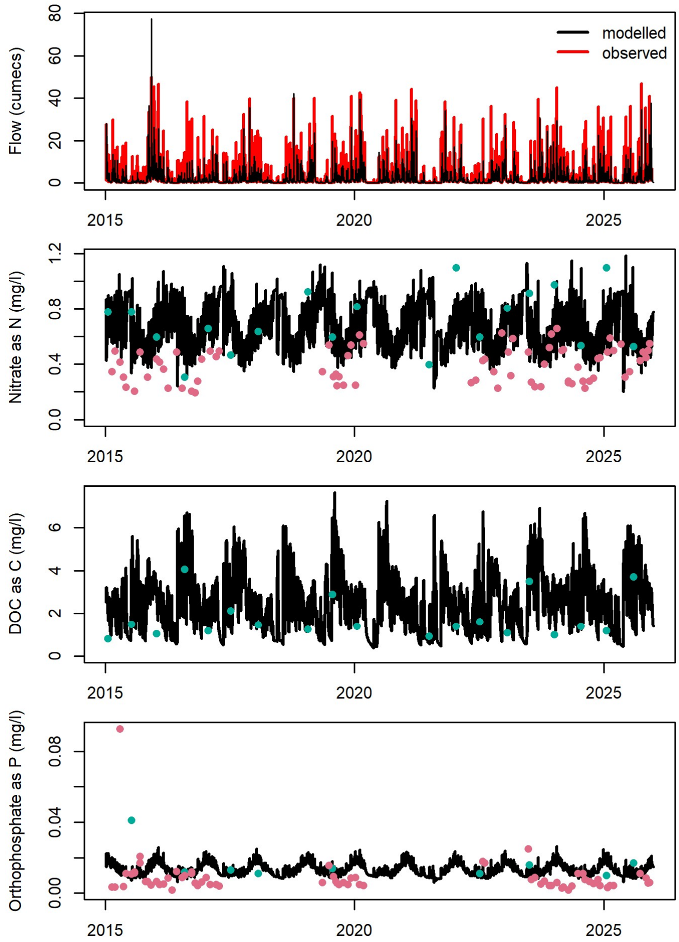


Figure 3.3: Hourly Cascade outputs compared to observations for the Trout Beck sub-catchment. Upper panel: modelled flows compared to Calgarth gauging station observations (Hydrology Data Explorer, 2026-03-26). Remaining panels: modelled nutrient concentrations (lines) compared to measurements from "Trout Beck at Windermere" (red points) and "Thickholme (NY40/1A)" (green points) obtained from the EA Water Quality Explorer (2026-03-25). For phosphate, measurements of both "Orthophosphate, filtered as P" and "Orthophosphate, reactive as P" are shown without distinction.

3.2.4 Load Apportionment Model

As for CASCADE, the Load Apportionment Model (LAM), was set up for subsequent inclusion in the digital architecture. Following inclusion, LAM would establish the relationship between phosphorus concentration, nitrate concentration and river flow, based on the infrequent (monthly) water quality data gathered by the EA. If these relationships are robust, and cover the full range of flow conditions, the LAM could estimate daily phosphorus and nitrate concentrations, based on real-time mean daily flow data. These modelled nutrient loads would then feed into PROTECH as an inflow, to simulate Windermere water quality. As new EA water quality and flow data become available, the LAM could automatically refine the nutrient concentration / flow relationship, providing more accurate load estimates. For smaller, ungauged rivers and streams that have water quality data, modelled flows from other models such as CASCADE could be used to produce LAM phosphorus and nitrate load estimates.

LAM is based entirely on understanding the relationship between river flow and nutrient concentration data. This simple approach makes the model effective in catchments with limited or potentially incorrect GIS data. LAM is based on a statistical approach that correlates water quantity and water quality. Thus, it requires for the nutrient and the flow observed data to be co-located and openly available. As this is a statistically based approach, it can only be applied if there is a clear and consistent relationship between flow and a particular nutrient concentration. Furthermore, it is reliant on the water quality data being accurate and covering the full range of flows at a particular site.

For the purposes of this project, due to budget and time constraints, LAM could not be fully integrated in the digital shadow. It was, however, set up and run for Windermere inflows, to showcase its capabilities. Due to the lack of observed flow and water quality data for most of the small stream inputs into Windermere, LAM could only be applied to the three main inputs: the Rivers Rothay and Brathay, and Cunsey Beck. Monthly observed in-stream concentrations for total phosphorus (TP), ortho-phosphorus and nitrate were available from the Environment Agency Water Quality Archive since the 1980s. Daily river flow data were available from the National River Flow Archive (NRFA). The relationships between total P and flow for the three rivers are shown in Figure 3.4, alongside the model fit.

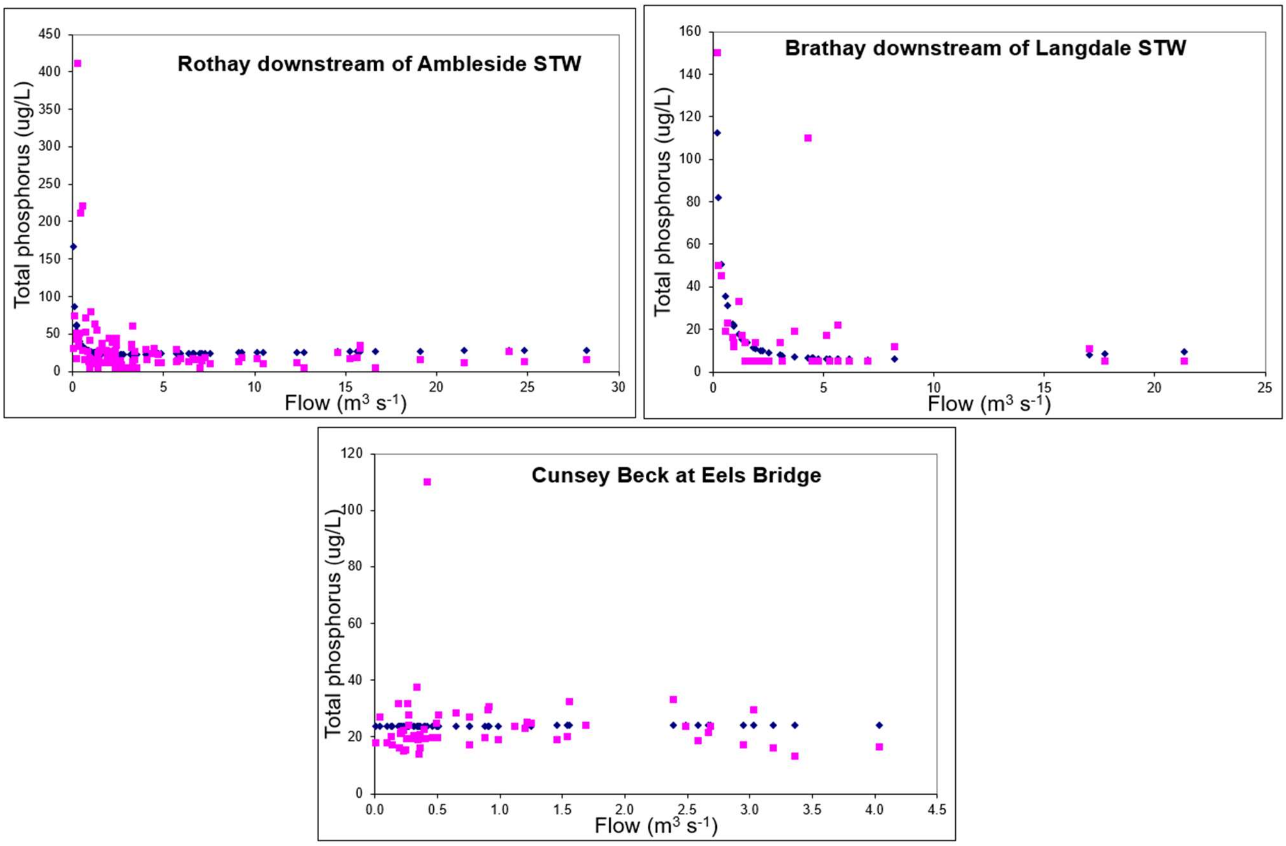
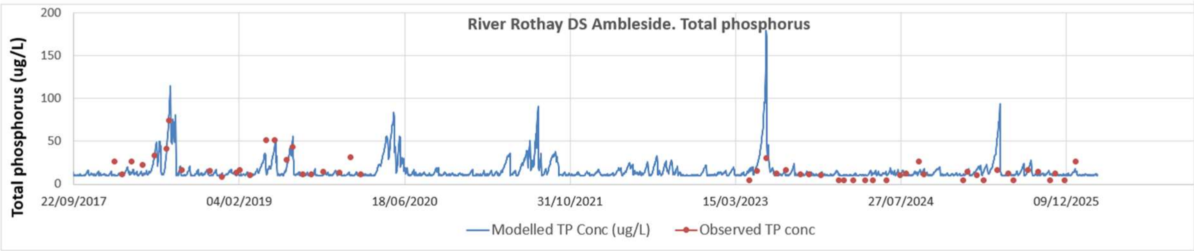


Figure 3.4 'Total phosphorus – flow relationships estimated using the Land Apportionment Model for three inputs into Windermere. Pink symbols = observed data. Blue symbols = LAM model fit.

The Rivers Rothay and Brathay show high concentrations at very low flows, indicating that there must be a constant phosphorus input to the river, probably from sewage treatment works. This constant input is rapidly diluted as the flow increases, and for most of the flow range, the TP concentrations remain relatively constant, indicating significant diffuse phosphorus sources. In contrast, Cunsey Beck had no high concentrations at low flow, indicating no point source inputs and a dominance of diffuse inputs. The ortho-phosphorus data produced similar relationships, but many of the P concentrations were below the limit of detection (LOD) (10 µg/L). For these data points, it was assumed that the concentration was half the LOD. Due to this source of error, the concentration / flow relationships were not as predictive, and the modelled outputs will be less accurate but still usable within the digital shadow platform.

The fitted models were then applied to the daily mean flow data, allowing daily TP concentrations to be estimated (Figure 3.5) for the river inflows. Within the digital shadow, this approach could generate “near real time” daily TP concentrations and loads, based on daily flow data, and be fed into the PROTECH model. In addition, LAM would be run using all the data available at the site thus the model fit may change as more data become available. Thus, the accuracy of the modelled nutrient estimates will improve as the digital shadow is in operation.



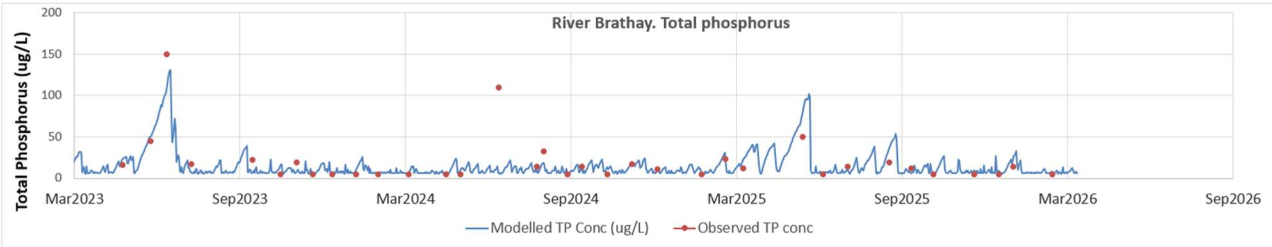


Figure 3.5. Modelled daily total phosphorus concentrations from LAM for River Rothay and Brathay.

3.2.5 Data – model linkages

By automating linkages between the identified models, their driving data, and their outputs, it would be possible to work towards near real-time simulation of water quality in connected river-lake systems; the core capability of a water quality digital shadow. For this stage of digital shadow development, our priorities were to:

1. Automate the linkages between QUESTOR and PROTECH, so that river water quality simulations produced by the former model are automatically fed into the latter. This would enable automated simulation of lake water quality responses to changes in inflowing river water quality.

2. Develop pipelines for the automated ingestion of water quality and weather data, as inputs into QUESTOR and PROTECH, to drive coupled modelling. This would enable near real-time simulation.

3. Prepare CASCADE and LAM for inclusion in the above framework, during subsequent digital shadow and twin development. This would enable the blending of inflow water quality simulations for inclusion into PROTECH, or the dynamic optimisation of river inflow model for each lake inflow, based upon measures of model performance.

3.3 Digital architecture and interfaces

To achieve our objectives, the digital architecture used for the project needed to satisfy three main criteria:

Criterion 1: Allowing the models to be run programmatically so that they could be linked.

Criterion 2: Having access to third party data for model inputs.

Criterion 3: Allowing users to access the architecture and run the models (currently, during this initial construction and testing stage, access has been restricted to UKCEH staff, but wider community access would be a key focus for future development).

3.3.1 Notebook technology, DataLabs and JASMIN cloud

We considered Jupyter Notebook technology as a potential solution because it provides an interactive environment where scientists, researchers, and developers can write, execute, document, and share code within a single platform. Jupyter offers strong platform independence, allowing notebooks to run across Windows, Linux, and macOS environments, as well as cloud-based platforms such as Google Colab, JupyterHub, and Azure Machine Learning. This flexibility makes it easier to develop reproducible scientific workflows and collaborate across different computing infrastructures.

We selected DataLabs, deployed on the JASMIN cluster, as a potential platform for delivering notebook-based access to computational models, data, and analysis workflows. JASMIN is a globally unique data analysis facility and provides storage and compute facilities that enable data-intensive environmental science. DataLabs is a cloud-based digital platform designed to support collaborative, open, transparent, and data-driven science (Hollaway *et al*, 2020; Green *et al*, 2024). It allows users to flexibly combine multiple elements within one integrated ‘environment’, and for this reason it is sometimes classified as a Virtual Research Environment (VRE). DataLabs provides a project-based environment in which users can create and run Jupyter notebooks while benefiting from integrated access to storage, computational resources, and containerised services. This made it a suitable candidate for supporting scientific users who require an interactive and reproducible environment for developing, executing, and sharing analyses.

The use of Jupyter notebooks within DataLabs offers platform independence, enabling workflows to be accessed through a web browser regardless of the user's local operating system. It also provides a familiar interface for scientists and researchers, combining code, documentation, visualisations, and results within a single environment. Through DataLabs, notebooks can interact directly with containerised models, code repositories, and backend processing pipelines hosted on the JASMIN cluster.

An additional advantage of using Jupyter was its ability to serve as an interface to backend model pipelines. While computational models were executed within containerised pipeline environments, users interacted with them through notebooks that contained the relevant code, documentation, parameters, and contextual information required for analysis. This approach abstracted much of the complexity associated with deploying, managing, and scaling models on containerised platforms, allowing end users to focus on scientific workflows and data analysis rather than infrastructure management.

Furthermore, the Jupyter ecosystem can be extended to create user-friendly graphical interfaces and web applications using tools such as Streamlit, Gradio, Voila, Dash, and Panel, enabling complex computational workflows to be made accessible to a wider range of users without requiring direct interaction with code.

The proof-of-concept architecture was designed as a modular MVP, providing flexibility for deployment across different public cloud providers or other infrastructure environments. This modular approach enables components to be reused, replaced, or extended as requirements evolve, while also supporting the development of custom user interfaces where needed. As a result, the MVP provided a strong foundation and architectural blueprint that can be deployed elsewhere and enhanced in future implementation phases.

We tailored and made our technology decisions based on the specific requirements and constraints of the use case, while ensuring alignment with the overall objectives and selecting a solution that is both practical and technically coherent, as discussed below.

Using AWS (Amazon Web Services), an on-demand cloud computing platform provided by Amazon which allows the deployment of modern software services, was also discussed and, although this would address Criterion 2, as the input data already exist, the setup required to enable Citeria 1 and 3 – and the associated cost – meant a large chunk of the project time would be spent before implementing the digital shadow.

The users mentioned in point (3) are diverse and would need to interact with the tool in different ways, as demonstrated in the stakeholder engagement outcomes (Section 2). Depending on their needs, requirements will differ with some users preferring a UI (User Interface) to run the digital shadow and others requiring access to the various elements of the pipeline and the possibility to analyse the outputs. Thus, based upon the above considerations and a need to improve accessibility and inclusion, we created two tools to run the digital shadow:

1) *A Streamlit application*: It provides a UI for users to run the shadow via the click of the button and view the outputs in the same window.

2) *A Jupyter notebook*: It exposes the code behind the Streamlit application and allows users to run certain steps, explore different parts of the shadow and make edits to components as part of an analysis.

3.3.2 Implementing the digital shadow

Underpinning the current digital shadow are the QUESTOR and PROTECH models, which needed to be deployed within DataLabs. This was achieved by using container technology, whereby a piece of software is deployed within a small container with everything it needs to run. This allows a separation of software (good for manging multiple pieces of software and for debugging) and means the software can be easily moved between different computing environments. For the digital shadow, each model was put into its own container.

This required porting the model code (written in Fortran) into a Linux environment (which DataLabs runs on); the first challenge encountered during the development process. The resolution of this challenge required some additional support from in-house IT. Once deployed within DataLabs, the models could be run from a terminal which allowed them, and their inputs and outputs, to be linked together. This type of challenge is quite common, whereby individual systems are developed in relative isolation and over the course of many years using a range of technologies. A push towards development taking place in a shared infrastructure using common architectural principles and software engineering best practices would start to alleviate these types of challenges.

Although Fortran is often used for modelling purposes, for example within QUESTOR and PROTECH, it is not suitable for use a digital twin wrapper as it lacks native cloud connectivity (e.g. APIs, IoT, …), lacks text and data parsing, and does not provide GUI or dashboard support. Therefore, we used Python to create the digital shadow infrastructure, in this case Python is referred to as the language of the wrapper. Python is commonly used as a digital twin wrapper as it allows to connect high-performance engineering simulations with modern cloud, AI, and data analytics tools commonly used. Python allows for a modularised approach allowing each ‘step’ in the digital shadow as displayed in Figure 3.6. Python is used for both Streamlit and Jupyter notebooks.

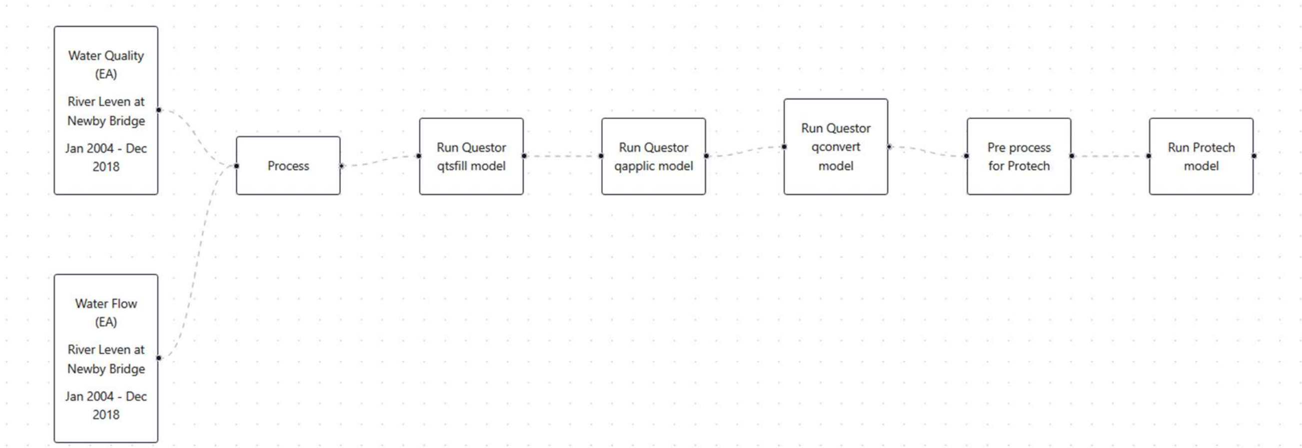


Figure 3.6 'Steps' in the digital shadow pipeline. Note: the current implementation uses Environment Agency API data from the Windermere outflow on the River Leven as a 'proof of concept', a real-world implementation would necessitate the use of inflow data from upstream catchments.

Some of the existing scripts that were used for both the 'Process' and 'Preprocess for PROTECH' steps in Figure 3.6 were originally written using R and so needed to be converted. Although, R code can be used within a Python environment, it is much cleaner and consistent to use the same language throughout. This also allows future developers a better chance of understanding the codebase. As the scripts were relatively small, it was a straightforward process to convert them to Python.

The QUESTOR pipeline required some external data from the Environment Agency (EA) API. Water quality and water flow data were extracted at runtime using scripts that connect to the API and request data for a particular site and period. The EA have no limitations on the number of requests that can be made to the API meaning the shadow can be run as many times as the user would like (for example, if running for a variety of sites and different periods).

Once the prototype pipeline has been run, the user can explore the inputs to, and outputs from, QUESTOR and PROTECH, in 2 ways (Figure 3.7):

1) A graphical user interface where different variables for each input/output can be plotted.

2) The data for each input/output can be downloaded as a CSV.

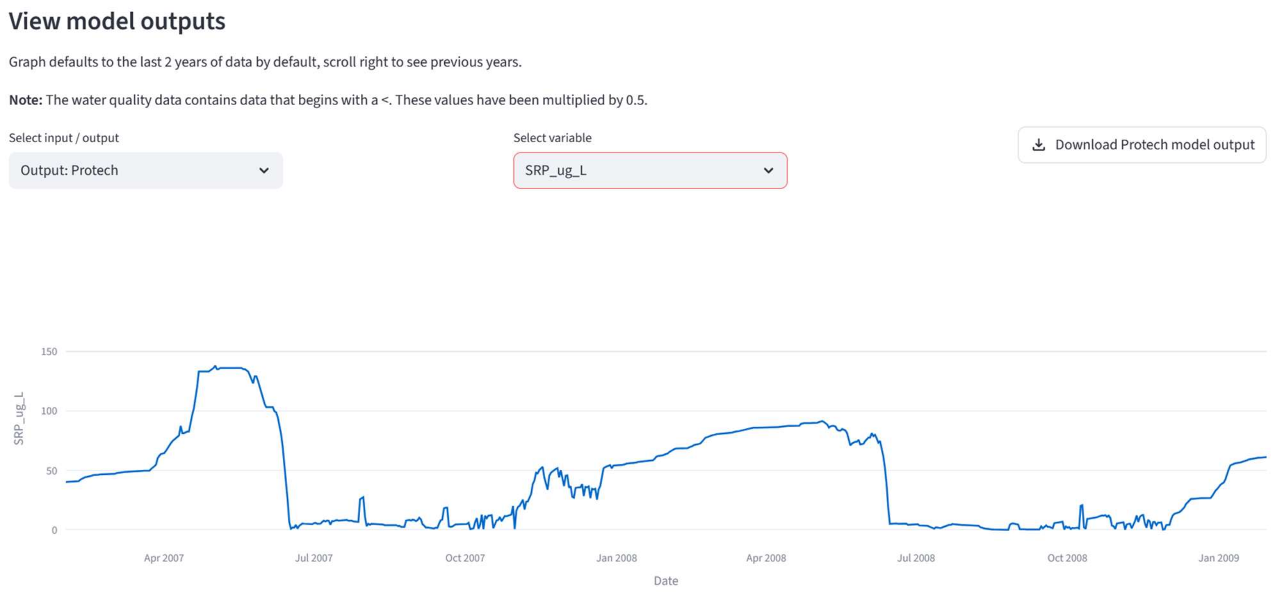


Figure 3.7 The graphical user interface and download options presented to the user. Note that SRP is Soluble Reactive Phosphorus.

3.4 Future development

3.4.1 Infrastructure

For the minimum viable product, the application and notebook are currently hosted on DataLabs. However, the tools have been developed to be platform-independent, so that future developments could be moved elsewhere, onto a cloud infrastructure such as Amazon Web Services (AWS), the provider used by UKCEH. This would bring several benefits:

1. A customised user interface could be built according to stakeholder and researcher requirements, rather than being restricted by what functionality Streamlit provides. This could improve the user experience and how the data are presented. The engagement conducted within this project has provided several desirable design features to guide future development and, once these are acted upon, there would be an opportunity for future engagement and co-design with end users, to further refine and optimise functionality.

2. Ancillary datasets could be accessed directly from UKCEHs internal storage on the cloud. For example, we are collecting Environment Agency data directly from their API, but these data already exist within AWS and so could be referenced directly without needing to re-fetch the data from the API. This reduces unnecessary duplication of data.

3. This would align with UKCEH's Digital Strategy by hosting the project infrastructure within an existing Digital Research Infrastructure (DRI), for example by leveraging established pipeline infrastructure to integrate additional processing workflows. This approach would enable a shared and common infrastructure for core services, improving reproducibility and scientific coherence across projects. It would also reduce duplication and fragmentation, meaning that changes to underlying components would have minimal downstream impact and fewer links or dependencies would need to be updated or maintained. From an IT perspective, consolidating onto a common stack also simplifies operational support, as teams would maintain a single, consistent infrastructure rather than multiple independent and siloed systems.

4. Moving to cloud infrastructure, such as AWS, would provide additional resilience and long-term continuity. The MVP has been designed with portability in mind and can be migrated to alternative infrastructure with minimal backend re-engineering, while maintaining core functionality and flexibility across environments. This is further supported by the use of open-source, community-supported software stacks such as notebook technology, which enhances interoperability and reduces platform dependency.

3.4.2 Technical developments

The minimum viable product developed here is a foundation upon which to build new functionalities. To further develop the digital shadow, there will be two main priority steps:

- Inclusion of the LAM and CASCADE models into the digital architecture (Figure 3.1): these models were applied and calibrated (where relevant) to the Windermere catchment as part of this project.

- Automated data access to the appropriate data repositories: within this project, this was achieved for the water quality data available from the EA, as a demonstrator. This aspect will need further development to access additional input data streams, such as flow and climate data.

These steps are essential to transitioning from a digital shadow to a digital twin. The inclusion of other water quality models would enable either 1) ensemble modelling of river inflows into lake ecosystems and associated estimation of model uncertainties, or 2) a more adaptive approach to connected river-lake simulation, whereby the selection of river model could be optimised for each individual river inflow/sub-catchment according to comparisons of river model performance. Automated access to observed data would enable real-time forecasting. These additional developments will complete a functional digital shadow and provide the digital infrastructure for a digital twin. A set of AI functionalities could then be added to enhance the capabilities of the digital twin, as referred to in the previous scoping study for this work, under the programme UK-SCAPE (Keller *et al*, 2024). Such functionalities could include using machine learning tools to select the most suitable input data for each daily time step, spatial extrapolation of water quality input data and/or indicators, and selection of the most suitable water quality model depending on the river inflow being modelled (Figure 1.1). A final consideration on the use of this infrastructure for the purposes of a digital twin is the current use of executable files for the river (QUESTOR, CASADE and LAM) and lake (PROTECH) models. Ultimately, we need to decide whether the fact that the models, and thus the process representation, are fixed and only input data can vary, is sufficient for the resulting infrastructure to produce a digital twin. If it is not deemed sufficient, then we may need to consider alternative methods to integrate these models in the digital infrastructure.

Following this work on the digital architecture, the performance of the overall modelling framework should be assessed in relation to observed ecosystem behaviour. We should ask whether the system is capable of reproducing real ecosystem dynamics and assess where the major modelling uncertainties still remain e.g. in which locations, at which times, at which scales, and for which variables do model simulations depart significantly from observed behaviour. By using system knowledge to diagnose the likely underlying causes of such discrepancies, we will be able to identify key measurement/data needs that should be incorporated into monitoring programmes. As such, the digital system will feed back to how we monitor the physical system, in-keeping with the digital twin paradigm.

There have been recent calls for “real-time pollution alerts” at Windermere, following concerns around public health. By building on the current minimum viable product, and developing towards a functioning digital twin, we would create a framework that could be used to enable near-term forecasting of water quality and meet this need. Key to this ambition, however, will be the coordination of monitoring efforts by different organisations who are active at the lake, and the development of capabilities for near-real-time data collection, quality checking, and integration. It is worth noting that quality checking could in time be undertaken as part of the digital twin.

As a more general point, the potential of digital twins to enhance our understanding and management of the water environment is widely recognised, and several examples of the development of such tools exist. In any future work, it will be important to share learnings among projects, if we are to fully realise this potential. As specific examples, digital twins are currently being developed for water distribution networks, the impacts of extreme heat on water industry assets, river ecosystems (including tools to enable restoration), and ecosystem service provisioning in lakes. There are also active initiatives to develop forecasting tools for water quality and algal blooms in lakes and reservoirs. Through collaboration with this wider community, it may ultimately be possible to create interconnected digital twins of different elements of the water environment (including water supply infrastructures); an ambition that currently forms the focus of the IDEATION project.

3.4.3 Sustainability

Sustainability has been embedded as a core principle throughout the design and development of the current prototype and will continue to be a key consideration for the future development. From an architectural perspective, we have adopted modular and scalable design practices to ensure that additional features can be integrated with minimal additional work, reducing waste in both development effort and computational resources. Future iterations of the prototype tool will adhere to this principle. Where appropriate, these future iterations will align the deliverables within the Floods and Droughts Research Infrastructure (FDRI) project and are likely to adopt a “lift and shift” approach to leverage its core infrastructure, enabling reuse rather than duplication. This strategy will support sustainability by minimising the number of parallel pipelines and lowering the demand for additional computing capacity. FDRI uses Amazon Web Services and elastically managed services to allow resources to scale dynamically with demand, helping to optimise energy usage and avoid over-provisioning. The current prototype also leverages the use of container technology to support sustainable software development practices, and this approach will continue to be adopted going forward, as it enables lightweight, portable, and efficient deployment of services while ensuring consistent environments and maximising resource utilisation.

3.4.4 Metadata

Another important area of future development will focus on systematically capturing metadata and provenance information associated with pipeline execution and processing steps. This will enable greater transparency and provide the necessary context to support the discoverability, accessibility, and reproducibility of data, methods, tools, and execution environments. Building on our existing work around FAIR (Findable, Accessible, Interoperable, Reusable) data principles and enhanced discovery of data, methods and tools, funded through FDRI and NC-UK, this capability will extend the depth and quality of information available to users. It will offer rich contextual metadata about datasets, models, and processing stages, allowing scientists and other users to better understand interactions between components, as well as the underlying scientific context and complexity involved in the workflows. It is worth noting that this will also contribute to ensuring that the data used or produced by the digital shadow or digital twin are AI-ready.

3.4.5 Interface co-design

Future work will also focus on the interface and interaction design of the tool, adopting a co-design approach with users to better understand their needs, expectations, and preferred ways of engaging with the system. Stakeholder engagement within the current project has provided us with invaluable first steps in this direction. This will involve iterative feedback and collaboration to shape how users interact with workflows, initiate executions, and explore associated metadata and provenance information. Such an approach will help ensure that the platform remains intuitive, user-centred, and aligned with real scientific practices. The future developments will ensure alignment and consistency with ongoing work under NC-UK and on our DRI, ensuring harmonised development across related initiatives.

4. Acknowledgements

We gratefully acknowledge all participants who contributed to the Digital Shadow survey and workshop. Your time, insights, and thoughtful engagement through discussions, written responses, or completion of the survey, were invaluable to the development of this work. We particularly thank those who attended the workshop on the 6th November 2025 for their active participation, collaborative spirit, and willingness to share expertise across sectors. Your contributions have played a central role in shaping the direction, priorities, and design of the Digital Shadow, and we sincerely appreciate your commitment to improving water quality understanding and management.

This work was supported by NERC, through the UKCEH National Capability for UK Challenges Programme NE/Y006208/1.

5. References

Davies, H. N., Rameshwaran, P., Bell, V. A., Dadson, S. (2024) Spatially consistent physical characteristics of UK rivers: 1-km data. *Geoscience Data Journal*, 11(3), 284–291. https://doi.org/10.1002/gdj3.209

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

Elliott, J. A. (2012) Predicting the impact of changing nutrient load and temperature on the phytoplankton of England's largest lake, Windermere. *Freshwater Biology*, 57, 400-413.

Environmental Agency (2026). Nutrient futures for Windermere. Environment Agency, Bristol. 78pp. https://assets.publishing.service.gov.uk/media/6a057b2a22977ebc82cb3f25/Nutrient_futures_for_Windermere_-_report.pdf

Green, D. P., Lord, C., Widdicks, K., Roebuck, J., Gouldsbrough, L., Hollaway, M. & Blair, G. (2024) DataLabs and discoverability: a summary report on the DataLabs enhancements project. UK Centre for Ecology & Hydrology, 17pp. https://nora.nerc.ac.uk/id/eprint/539039/

Hollaway, M. J., Dean G., Blair, G. S., Brown, M., Henrys, P. A. & Watkins, J. (2020). Tackling the Challenges of 21st-Century Open Science and Beyond: A Data Science Lab Approach. *Patterns*, 1, 100103, https://doi.org/10.1016/j.patter.2020.100103

Hutchins, M., Gardner, E., Gechelle, L., Johnson, A., Qu, Y., Sadykova, D., Bowes, M., Read, D. & Robertson, C. (2026). What controls current and future background vulnerability of rivers to eutrophication and pathogens? *Journal of Hydrology*, 665, 134713.

Keller, V., Dick, J., Dobel, A., Elliott, A., Fry, M., Harrison, S., Khamis, D., Kitson, E., Qu, Y., Blair, G. & Thackeray, S. J. (2024) WADITI - developing a roadmap for a water quality digital twin. A scoping study under UK-SCAPE. UK Centre for Ecology & Hydrology, 50pp., https://nora.nerc.ac.uk/id/eprint/538010/

Mackay, E. B., Rankin, G., McShane, G., Rhodes, G., Dodd, B., James, B. J., Clarke, M., Kelsall, M., Moran, S. Keenan, P. O., Pereira, M. G., Guyatt, H., Hunt, A., Salisbury, E., Chetiu, N., Grant, H., Bell, K., Clark, R & Thackeray, S. J. (2023). Lakes Tour 2021. UK Centre for Ecology & Hydrology report, 129pp, https://nora.nerc.ac.uk/id/eprint/535668/1/N535668CR.pdf

Monteith, D.T., Henrys, P. A., Hruška, J., *et al* (2023) Long-term rise in riverine dissolved organic carbon concentration is predicted by electrolyte solubility theory. *Science Advances,* 9, eade3491, https://doi.org/10.1126/sciadv.ade3491

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 modeling in lowland rivers. *Limnology and Oceanography*, 67(6), 1313–1327, https://doi.org/10.1002/lno.12079

6. Appendices

6.1 Appendix 1: Participant Information Sheet

Participant Information Sheet

Title of Project: *Water Quality Digital Twin*

Introduction
We would like to invite you to take part in a research study. Joining the study is voluntary and your decision. To help you decide you need to understand why the research is being done and what it would involve from you. Please ask questions if anything you read is not clear or you would like more information, either to a member of the project team or by contacting Ellie Mackay, ellcka@ceh.ac.uk, and 01524 595904.

What is the purpose of the study?
The UK Centre for Ecology & Hydrology (UKCEH) is conducting research on how to develop a digital twin, linking environmental observations with environmental models for water quality, using the Windermere catchment as a case study. The results of the initial study are here (Keller et al. 2025 *WADITI - developing a roadmap for a water quality digital twin. A scoping study under UK-SCAPE)* . This follow-on research will actualise the design of the digital twin including the key variables, metrics and timescales of water quality outputs that would be relevant to you and your work and how you would like to visual and use these outputs.

Why have I been asked/offered the opportunity to take part?
You have been invited because we want to get input to our work in the Lake District from the key stakeholders who have interests in the management of lakes and their catchments. We are really interested to hear from you because we think that tools like the digital twin may be of use to you and your work in the future.

Do I have to take part?
No. Participating in this research project is voluntary, and entirely your decision/choice. If you do not want to take part, that is ok.

We, the research team, will discuss the study with you at the Cumbrian Lakes Research Forum on 6th November 2025 and give you a copy of this information sheet. If you agree to take part, we will then ask you to sign a consent form. If you are unable to attend please consider offering your opinion by completing the survey.

What will happen to me if I take part; what will I have to do?

We would like you to get involved in the discussion sessions during the Cumbrian Lakes Research Forum where we will ask for your opinions and ideas for how we should design the digital twin during the project. We will take these ideas with us to inform the development of this project and provide us with ideas for what the appetite for future developments in the digital twin could be. We will provide you with the summary of the discussions that are carried out in the workshop and welcome any feedback that you may have. We would also like to take some photographs of participants talking together during the discussions to help illustrate our report and for use in Social Media reporting of the event. If you would prefer not to be included in these photos, please let us know.

What are the possible risks and disadvantages?

The risks to participating are low, as we will work with only generic personal data about you (affiliation, age group and gender) and will anonymise any responses you provide to us for the project. Your contact details, your name and email will not be associated with these data and will be stored securely in a separate location.

What are the possible benefits?
We cannot promise the study will help you but the information we get from the study will help our knowledge and understanding of the needs of potential users of digital twin tools to ensure that what we develop is of use.

What if something goes wrong?
If you have a concern about any aspect of this study, you should ask to speak to the researchers who will do their best to answer your questions. The main point of contact is Ellie Mackay, ellcka@ceh.ac.uk, and 01524 595904.

If you remain unhappy and wish to complain formally, you can do this by contacting the UK Research Ethics Committee ukcehresearchethics@ceh.ac.uk.

Can I change my mind about taking part?
Yes. You can withdraw from the research project at any time. You just need to tell the research team, or contact Ellie Mackay, ellcka@ceh.ac.uk and 01524 595 800.

If you withdraw from the study, we will destroy all your contact details, we will continue to use the data/information collected by you up to your withdrawal as the data is anonymised.

What will happen to information collected about me?
We will need to use information about you for this research project. All information collected about you will be kept securely. Only the research team and ethical committees who check that the study is being carried out properly will be allowed to look at information about you.

Information about you will be you're:
- name
- contact details (email)

We will keep all information about you safe and secure on our servers in UKCEH.

Please see below for explanation

Where data has been anonymised, the original information will be securely deleted to prevent any reversing of the 'anonymisation' process. In most cases, if this deletion does not take place, then the data is classified as 'pseudonymised' rather than 'anonymised', and, is still considered personal data. Full personal data is simply using the details gathered for the project with no omissions.
In all three cases personal data should be kept to a minimum.

Your personal details, meaning your name and contact details, will be kept in a different safe place to the other study information and will be stored securely on servers at our site in Lancaster.

At the end of the project, the study data will be archived at Lancaster. The data will be made available to other researchers worldwide via our data repository the EIDC. Your personal information will not be included and there is no way that you can be identified.

What are your choices about how your information is used?

You can stop being part of the study at any time, without giving a reason*, but we will keep information that you have provided that we already have collected up to your withdrawal.*

Where can you find out more about how your information is used?
You can find out more about how we use your information

- At: https://www.ceh.ac.uk/contributors-ceh-science-public-interest OR https://www.ceh.ac.uk/contributors-ceh-science-funded-through-competitively-won-income
- by asking one of the research team
- by sending an email to Ellie Mackay (ellcka@ceh.ac.uk)

What will happen to the results of this study?
The research results will be published in reports or journal articles. We will share these project outputs with all project participants. Your personal information will not be included in the reports and there is no way that you can be identified from these.

Who is organising and funding this study?
UKRI is the funder of this research project; UKCEH is the overall project lead and they have full responsibility for the project including the collection, storage and analysis of your data, and will act as the Data Controller for the study. This means that we are responsible for looking after your information and using it properly.

Who has reviewed this study?
All research involving human participants is looked at by an independent group of people, called a Research Ethics Committee, to protect your interests. This study has been reviewed and given favourable opinion by UK Centre for Ecology & Hydrology Human Research Ethics Committee (UKCEH-HREC add number when known).

Further information and contact details
Thank you for taking time to read this information sheet.

If you would like any further information, please contact Ellie Mackay who can answer any questions you may have about the study.

Contact details:
Ellie Mackay
ellcka@ceh.ac.uk
01524 595904

6.2 Appendix 2: Survey questions

Water Quality Survey - Final

You are invited to participate in this survey, funded by UKRI-NERC, which aims to create a water quality digital shadow of Windermere and its catchment. Following a co-developed roadmap from a previous project, we aim to deliver a digital shadow as a first step towards a water quality digital twin (WQDT).

The water quality of freshwater ecosystems is of increasing concern to the public and policymakers, and emerges from complex interactions between physical, chemical, and biological drivers and processes. There is a crucial need to improve our ability to understand and, ultimately, forecast these changes in water quality. Enabled by our Digital Research Infrastructure (DRI), we will develop a Windermere Water Quality Digital shadow by integrating process-based and empirical models with long-term and high-resolution data from a range of observation networks located in Windermere. This initiative is funded by the National Capability for UK Challenges (NC-UK) programme.

A digital shadow is a one-way digital representation of a physical object or system that automatically collects and reflects/models/analyses real-time data from it. The data flow from the physical system to the digital model, allowing you to monitor, analyse, and understand the system. It's a passive and observational system. It is a first step towards a digital twin which, by contrast, has a bidirectional relationship, and thus allows for controlling and optimising the physical system as well as monitoring it.

Source: https://www.flexcon.it/news-events/digital-twin-vs-digital-shadow-vs-digital-model/

No personal data which can identify you will be collected, however we do ask at the end for general demographic (age, gender) and role (institutional type, career stage) data in order to describe the cohort of individuals offering their opinions and ensure we can report to the funders Equity, Equality, and Diversity criteria.

The information you provide will be captured electronically via this survey tool or as a paper copy. The data will be stored to support analyses and any potential future publication documenting this co-production process. We intend to archive the anonymized data for future research use; however, there will be no way for these data to be linked to individual survey participants.

Please see full Participants information Sheet for further details or contact Jan Dick (jand@ceh.ac.uk) if you can not access the full Participants information Sheet.

If you consent to take part in this survey please tick this box ☐

1. From your perspective, which physico-chemical variables would be the most important outputs from the digital shadow?

(Please circle all that apply).
P (Total P, Dissolved P, Both?)
Dissolved oxygen
N (Nitrate, Ammonium)
DOC
Incident solar radiation (water column average for rivers)
River water quantity (flow, velocity, depth)
Water temperature

2. How do the variables you selected above need to be summarised, spatially, to be of most use to you?

(Please circle all that apply).
Individually, for each river input
For the whole lake (both basins combined)
As a total lake input (including runoff from small areas)
For each lake basin (North/South)

Other (please specify)

3. What timescales should these variables be output over, to be of most use to you?

(Please circle all that apply).
Daily
Weekly
Monthly
Yearly
Other (please specify)

4. What algal variables are most important to you in your work?

(Please circle all that apply).
Chlorophyll concentration (and concentrations for different groups)
Most abundant algal group on each day/timestep? (e.g. cyanobacteria, diatom etc.)
Other (please specify)

5. How do the algal variables you selected above need to be spatially summarised, to be most useful to you?

For each river input
For the lake as a whole
For each river input and for the lake as a whole
For each lake basin, separately (North/South)
Other (please specify)

6. What timescales should these algal variables be output over, to be of most use to you?

(Please circle all that apply).
Daily
Weekly
Monthly
Yearly
Other (please specify)

7. What metrics would you like to see from the Digital Shadow?

(Please circle all that apply).
Exceedance of thresholds (e.g. cyanobacteria with respect to WHO guidelines of 10 and 50 ug/L cyanobacteria chlorophyll a).
Loads (nutrients)
Concentrations (nutrients, algal biomass, algal group e.g. cyanobacteria)
Other (please specify)

8. We are keen to develop the digital shadow so would welcome any other thoughts you have e.g. are there different variables you would like reported with different metrics and over different timescales or is there anything else you would like to add to help us improve upon the digital shadow presented today?

UKCEH

Our planet. Decoded.

For more information please contact:

[EMAIL US]
Enquiries@ceh.ac.uk

[LEARN MORE]
ceh.ac.uk

[FOLLOW US]
@uk_ceh