# Rethinking data-driven decision support in flood risk management for a big data age

Ross Towe*[1,6], Graham Dean[1,7], Liz Edwards[1], Vatsala Nundloll[1], Gordon Blair[1], Rob Lamb[2,3], Barry Hankin[3,4], and Susan Manson[5]

[1]School of Computing and Communications, Lancaster University, United Kingdom
[2]JBA Trust, Skipton, United Kingdom
[3]Lancaster Environment Centre, Lancaster University, United Kingdom
[4]JBA Consulting, Warrington, United Kingdom
[5]Environment Agency, Beverley, United Kingdom
[6]Shell Research, London, United Kingdom
[7]UK Center for Ecology and Hydrology, Lancaster, United Kingdom

May 16, 2020

**Abstract**

Decision-making in flood risk management is increasingly dependent on access to data, with the availability of data increasing dramatically in recent years. We are therefore moving towards an era of big data, with the added challenges that, in this area, data sources are highly heterogeneous, at a variety of scales, and include a mix of structured and unstructured data. The key requirement is therefore one of integration and subsequent analyses of this complex web of data. This paper examines the potential of a data-driven approach to support decision-making in flood risk management, with the goal of investigating a suitable software architecture and associated set of techniques to support a more data-centric approach. The key contribution of the paper is a cloud-based *data hypercube* that achieves the desired level of integration of highly complex data. This hypercube builds on innovations in cloud services for data storage, semantic enrichment and querying, and also features the use of notebook technologies to support open and collaborative scenario analyses in support of decision making. The paper also highlights the success of our agile methodology in weaving together cross-disciplinary perspectives and in engaging a wide range of stakeholders in exploring possible technological futures for flood risk management.

*Keywords: big data, cloud computing, data hypercube, data science, flexible querying, semantic web, uncertainty.*

*r.towe@lancaster.ac.uk

# 1 Introduction

Flood risk management (FRM) is an increasingly important topic as societies around the world are faced with a significant rise in the number of extreme events (Beniston and Stephenson, 2004; Jongman et al., 2012). Data are a fundamental part of FRM, however recently there has been an explosion in the availability of heterogeneous data, with the volume, variety and quality of data increasing all the time. The end result is a complex web of highly heterogeneous structured and unstructured data, often at different scales and with different levels of veracity.

This paper examines the potential of a data-driven approach to support decision making in FRM. The overall goal of the paper is to investigate a suitable software architecture to support this new approach, starting from the premise that cloud computing has the appropriate elastic capacity and associated supporting services to facilitate such an approach. Specifically, the paper addresses the following key challenges:

1. Identifying appropriate data representation techniques to capture and integrate highly heterogeneous data (referred to as a hypercube of FRM data);

2. Identifying appropriate means to support discovering, navigating and querying complex structures;

3. Making uncertainty explicit and providing a conceptual framework to reason about uncertainty;

4. Supporting a more open and collaborative style of science and decision making building on a more data-driven approach.

We aim to address the need for integrated risk modelling frameworks that consider all aspects of flooding. The work is inspired by discussion of integrated risk modelling and data analysis in the UKs recent National Flood Resilience Review (NFRR) (Her Majesty's Government, 2016; Tawn et al., 2018) and international efforts to identify and understand disaster risk through better integration of models and data (World Bank, 2016). As seen in Figure 1, the NFRR showed that the current approach of integrated modelling for flood risk is disjointed with the dynamical and statistical pathways existing independently of one another. Being able to consider both pathways concurrently clearly leads to a better informed understanding of all flood risks. However, the development of an over-arching integrated modelling framework is a major challenge.

Implicit in this is the uniting of the two pathways, that is, process modelling and statistical modelling. This is a major task, and beyond the scope of this paper. For the purposes of the paper, we assume a data-centric view where we have access to the inputs, outputs and specifications of process models and statistical models alongside other sources of data.

This is also a major cross-disciplinary challenge requiring input from (amongst others) computer scientists, data scientists, and expertise associated with FRM. FRM in itself also incorporates many

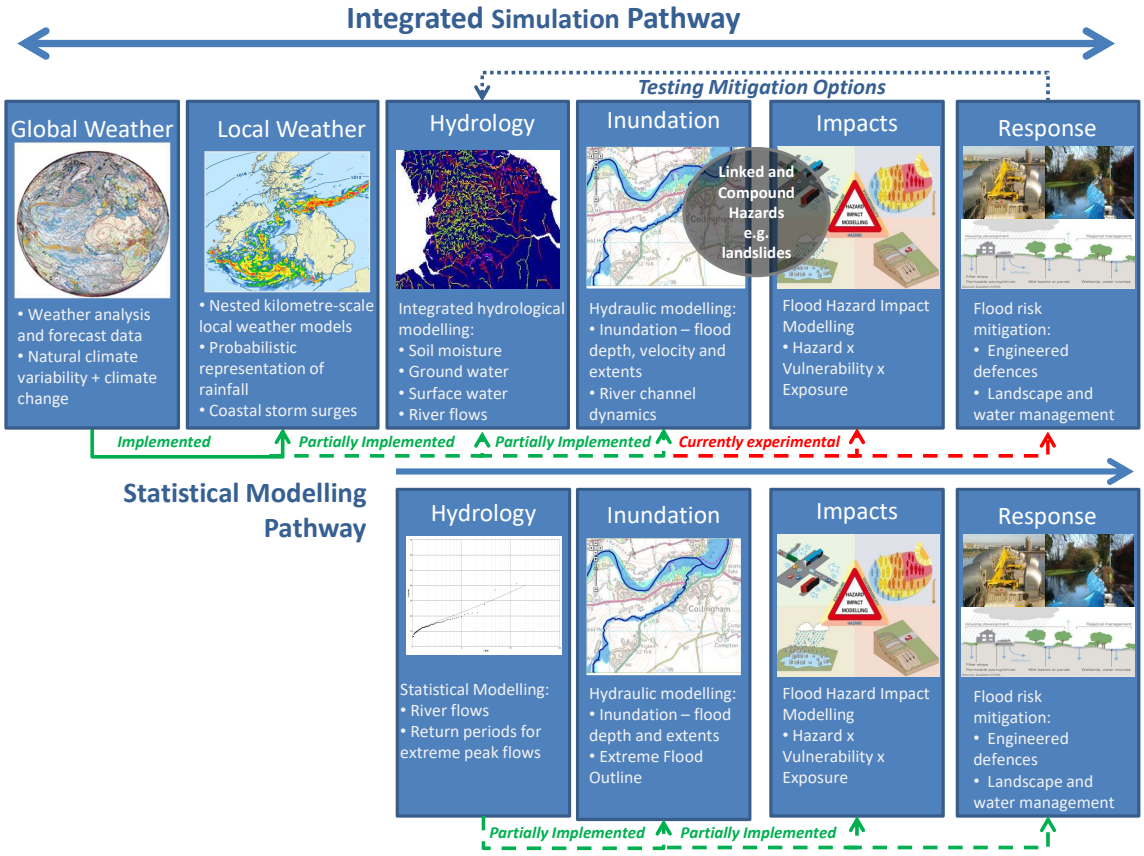



Figure 1: Schematic showing the key elements of an integrated flood risk approach. Image is taken from the National Flood Resilience Review.

different disciplines all of whom communicate in a variety of differing languages. These challenges need to be overcome if the ambitions of various national and international strategies are to be achieved (Environment Agency, 2019; Welsh Government, 2019; The World Bank, 2018).

The research method we have chosen is specifically designed to support the necessary cross-disciplinary dialogues. Traditionally, the commissioning of large and complex software projects would involve establishing the requirements at the start of the project before entering a (lengthy) design and implementation phase with the requirements reviewed on delivery (Boehm, 1988). Our approach is to work in a more agile way to allow for an ongoing dialogue between the various collaborators and disciplines as an intrinsic part of the process. To scope the study and provide a baseline for comparison, we took an existing project in FRM implemented using traditional structures and re-engineered the approach using our proposed new software architecture. We then evaluated the approach against the overall goal and associated challenges as described above, taking

input from partners representing different stakeholders and disciplinary backgrounds.

The paper is structured as follows, Section 2 reviews the current state of data and modelling with the FRM sector with Section 3 building on this to describe future visions for the sector and introducing our proposed hypercube approach. The approach is applied in Section 4 through a specific case study. Section 5 includes evaluation of the ideas, details on how to transfer these prototypes into data-driven decision-support for FRM and the barriers that need to be overcome in order for these methods to become operational.

# 2 Data and modelling in Flood Risk Management

## 2.1 Drivers for risk analysis

A risk-based approach recognises that uncertainty is a fundamental problem in planning for and responding to flooding (Beven and Hall, 2014). Statements about the risk of flooding can be interpreted as attempts to describe the probability of floods occurring, or reflect the distribution of potential consequences incurred as a result of flooding. The concept of FRM acknowledges that, given uncertainties about climate, weather, hydrological processes and human interventions, it may not be possible to eliminate the risk of flooding entirely (Fleming, 2002; Alfieri et al., 2016). Rather, risk management involves seeking an understanding of both the probability and consequences of flood events, and the residual risk that remains after accounting for mitigation measures (such as flood defences). Very often, the goal of the analysis may be characterised as enabling rational, proportionate decisions to be made about mitigating the risk. A conceptual framework for risk analysis distinguishes between the hazard (the physical sources of the risk), the vulnerability of physical, social and environmental systems that are exposed to the hazard, and the consequential impacts. Different flavours of this model often co-exist, such as the source-pathway-receptor concept (Holdgate, 1980; Samuels, 2009) or the hazard-vulnerability-loss concept commonly used by the insurance sector in modelling natural hazards (Mitchell-Wallace et al., 2017). No matter how it is conceptualised, the risk analysis may have to be implemented and applied at multiple scales, and within complex social, economic, fiscal and political contexts (UNISDR, 2015).

## 2.2 Scale and context

The multiple scales and drivers for flood risk analysis make it hard to achieve a unified, coherent view of risk. For example, in many territories, local detailed models are applied in support of decisions about flood defence design or flood warning systems at the scale of an individual community. On the other hand, regional, state, national or even supra-national decision-making may require flood risk analysis at a much larger scale of aggregation. Although it is appealing to build the larger scale analysis up from local studies, there will inevitably be a patchwork of different models, with gaps, reflecting different priorities and levels of detail, which may not be easy to aggregate to a

larger scale. This is a current topic of discussion within the community (Environment Agency, 2018c). Similar issues arise in the treatment of time. Models and data sets applied for short-term predictions need to meet different priorities (e.g. operational speed and robustness, integration with real-time data feeds) than those developed to assess very long-term risk (where flexibility to test what-if scenarios or generate a wide array of output risk metrics may be more prominent).

Layered onto this mosaic of different modelling and data priorities are the organisational structures associated with FRM in a given territory. A mixture of public agencies and private enterprises may be involved in creating, using and maintaining models and data sets. Differing priorities lead to the fragmentation in approaches to modelling and data; for example, in the UK, different modelling approaches are applied to predict national patterns of river flow over the next 7 days (Environment Agency, 2018e) and to simulate national patterns of runoff for potential extreme rainfall scenarios to support planning decisions (Environment Agency, 2018d). While our analysis is largely informed by a United Kingdom perspective, and more specifically FRM in England, the value chains associated with flood risk models and data share many similarities in other countries; see NFRR, Annexe 10 (Her Majesty's Government, 2016), for a comparison between England, France, the Netherlands, Australia, Japan and the USA.

### 2.3 Technical approaches

There is a strong tradition of process-based modelling in FRM, with conceptual rainfall-runoff models embedded within systems developed for short-term flood warning (Beven, 2007; Environment Agency, 2018e; Kauffeldt et al., 2016; NERC CEH, 2019). Prediction of flood depths and extents has centred on hydraulic models, that have evolved from spatially one-dimensional channel models based on either energy or momentum equations (Brunner, 1995; US Army Corps of Engineers, 2016) to the spatially two-dimensional, dynamic models based on the Shallow Water Equations that are now routinely applied (Hunter et al., 2007; Environment Agency, 2013). In some situations, 3D Computational Fluid Dynamics modelling is also being applied in FRM, although this is yet to be cost effective as a baseline approach (Teng et al., 2017). Alternatives also include more data based methods such as PDM (Probability Distributed Model) for flood forecasting (Moore, 2007).

Alongside the use of physically-based models, FRM has long taken advantage of an understanding of the statistics of extremes (Benson, 1968; Gumbel, 1958), which enables mathematically and scientifically justified extrapolation from observable data to consider more extreme scenarios, as required for a robust risk analysis. Equally there are good practices in utilising and managing data sets, with carefully maintained and curated long term hydrometric records (NERC CEH, 2018; US Geological Survey, 2016; Grand River Conservation Authority, 2016) being crucial resources, and a particular focus on establishing credible data around extremes.

However, the emergence of disruptive concepts and technologies such as cloud computing and data science are prompting questions about how best to exploit such developments to make sense

of complex, heterogeneous information about flooding (Demir and Krajewski, 2013). There is a growing recognition of the importance of new types of data in future flood risk assessments (Environment Agency, 2018c). There remains a significant gap in both the literature and practice as to how this can and should be achieved including how to combine process-driven and data-driven paradigms.

# 3 Vision for a Data-centric Approach for Flood Risk Management

## 3.1 Agile research methodology

As mentioned in Section 1, we adopt an agile approach as the core methodology underpinning the research. Agile is an umbrella term for a range of practices that embrace the values and principles expressed in the Manifesto for Agile Software Development[1]. This results in set of agile methods that iterate towards a solution through "early and continuous delivery" of software, and that are intrinsically reflective throughout. While generally used in software development, we are using *agile as a research methodology* in exploring different software architectures for future FRM (Ferrario et al., 2013, 2016). The agile process started with a workshop, which brought together the research team and stakeholders with additional thought leaders from the flood sector[2]. This workshop produced four key outputs:

1. a strong consensus on the need for a more data driven approach;
2. the identification of the key elements underpinning such a data-centric approach;
3. the construction of a community of researchers and stakeholders as required to progress with an agile approach;
4. a set of user stories (storyboards) as input into the subsequent software development process.

This then fed into an iterative, agile process involving the rapid development of software prototypes and various mechanisms to achieve community feedback including monthly show-and-tells and more intensive sessions with smaller selected groups. The results were then presented in a follow up workshop, with this validating the results and establishing pathways to take the work forward.

## 3.2 The role of storyboards

Storyboards are a commonly adopted practice within the software development community (Cohn, 2004) and provide a representative picture of the specific challenges and perspectives for stakeholders within a particular domain. The storyboard generated from the workshop is presented in Table 1. This storyboard played a special role in our agile methodology in providing both the initial input

[1] http://agilemanifesto.org/

[2] https://www.ensembleprojects.org/

into the iterative development process and the means of evaluation for the eventual prototypes (De Nicola and Navigli, 2009). To add richness to the storyboard, it was considered from the perspectives of two key *personas*: from the perspective of a data scientist, Alicia, looking to make sense of the underlying (inevitably complex and heterogeneous) data, and from the perspective of a flood risk manager, Bashir, seeking to make decisions based on the available data and its analyses. Note that in practice a number of stories emerged from the workshop but for ease of presentation these have been consolidated into this one storyboard.

| **Perspectives** | |
|---|---|
| **Data scientist** | **Flood Risk Manager** |
| Our data scientist, Alicia, is researching the impacts of flooding in Newark and wants to access and query multiple data sets regarding different facets of the environment to seek trends and causality. Her key requirement is to have a single repository that integrates structured and unstructured information and includes meta data such as provenance so she can make appropriate data selection, and use tools and methods that account for the contextual information.<br>In her work, she becomes frustrated by missing data and wants this to be explicit so that she can take remedial actions in her analyses. She also notes inconsistencies when looking at flooding impacts at different scales, in particular the local data and associated data on national trends in response to extreme events. As a data scientist, she wants to consolidate these views and reach well-informed conclusions about the impact of flood events. | Our flood risk manager, Bashir, has important decisions to make regarding investments in Newark following a flood event. Bashir has to account for his decisions to stakeholders including local residents, but he is mindful that local flood groups sometimes mistrust data, believing it is partial, omitting their experiential accounts of flood impact. Consequently he is keen to include detailed local assessments in his evaluation and decision making. Local Environment Agency reports are particularly helpful to resolve this uncertainty as they include actual flood level measurements related to the flood event.<br>Following lobbying from the local business community, Bashir has a particular concern for the way small, independent businesses were affected, and wants to explore this dimension of the data. Bashir wants to learn from experiences in other catchments that have dealt with comparable events with intense rainfall. He's interested in the effectiveness of different defence intervention strategies and alternative natural flood management approaches and to use this knowledge to assist future decision making.<br>Bashir needs to be able to communicate his findings to audiences with different interests and competencies. He wants to be able to present the outputs of his Environmental Risk Assessment as place based visualisations, which allow the data to be observed at different scales. This involves zooming in on particular properties and showing comparative visualisations of the impact of different management strategies on the wider community. It is important that he is able to do this in a transparent manner that makes visible missing data, uncertainties and inconsistencies. |

Table 1: Storyboard from the perspectives of a data scientist and a flood risk manager.

Looking more closely at our storyboard, both Alicia and Bashir want to be able to systemically query all available data on a single platform using a single tool. This would enable them both to have the potential to use all of the available data efficiently. In particular, Alicia wants to be able to compare and assess a number of different models in their ability to represent the Newark flood event. She is particularly interested in using the most complete data sets. This can conflict with the wishes of clients, who want a nationally consistent approach, but Alicia hopes to convince them of the value of a more detailed local approach. Bashir wants to be able to construct pictures of flood risk quickly and accessibly in order to communicate the impact of flood events and alternative management strategies.

Further analysis of the storyboard led to the following key insights: i) the key enabler is to achieve a strong level of integration across highly heterogeneous data; ii) it is necessary to support multiple perspectives over this data, as different actors and stakeholders require different things; iii) transparency is crucial in the subsequent decision-making process. These three insights provided the high-level requirements to seed the subsequent agile process, leading to the (iterative) development of our demonstrator.

### 3.3 A story-led demonstrator

The key output of the agile process is the development of a story-led demonstrator to allow exploration of the added value of a data-driven approach to support decision making in FRM. The architecture of this demonstrator is shown in Figure 2. The architecture takes the form of a *conceptual stack* consisting of multiple levels, with the diagram also showing the technologies that can be used to implement each level and the associated benefits for stakeholders and users of the system.

At the lowest level in the stack, we have the underlying cloud storage service. The major trend in cloud computing is to move towards offering a heterogeneous set of relatively primitive storage services. Most cloud providers offer variants of traditional networked file stores, block stores (that is underlying fixed size blocks of data), or more sophisticated object stores that contain name-value pairs, where the name is an arbitrary index to identify and locate an object and the value is of arbitrary length and content and could be of any type such as image, video, raw data and so on. These options can be used to capture unstructured data, that is data that does not have a pre-defined data model (cf. schema). Many cloud providers will also offer abstractions for more structured data, often in the form of tables, viewed as more efficient and scalable compared to full relational capability (cf. NoSQL). Some providers will also offer relational databases, at least partially for legacy reasons. The advantages of cloud-based solutions are that they bring all this data, structured and unstructured, together in one place, and that the cloud provider ensures reliability and scalability. There is as yet no integration across the different data sources.

The next level up is our model service, supporting the execution of both data/statistical and process models. In practice, we focused on data model support using a combination of R and Python

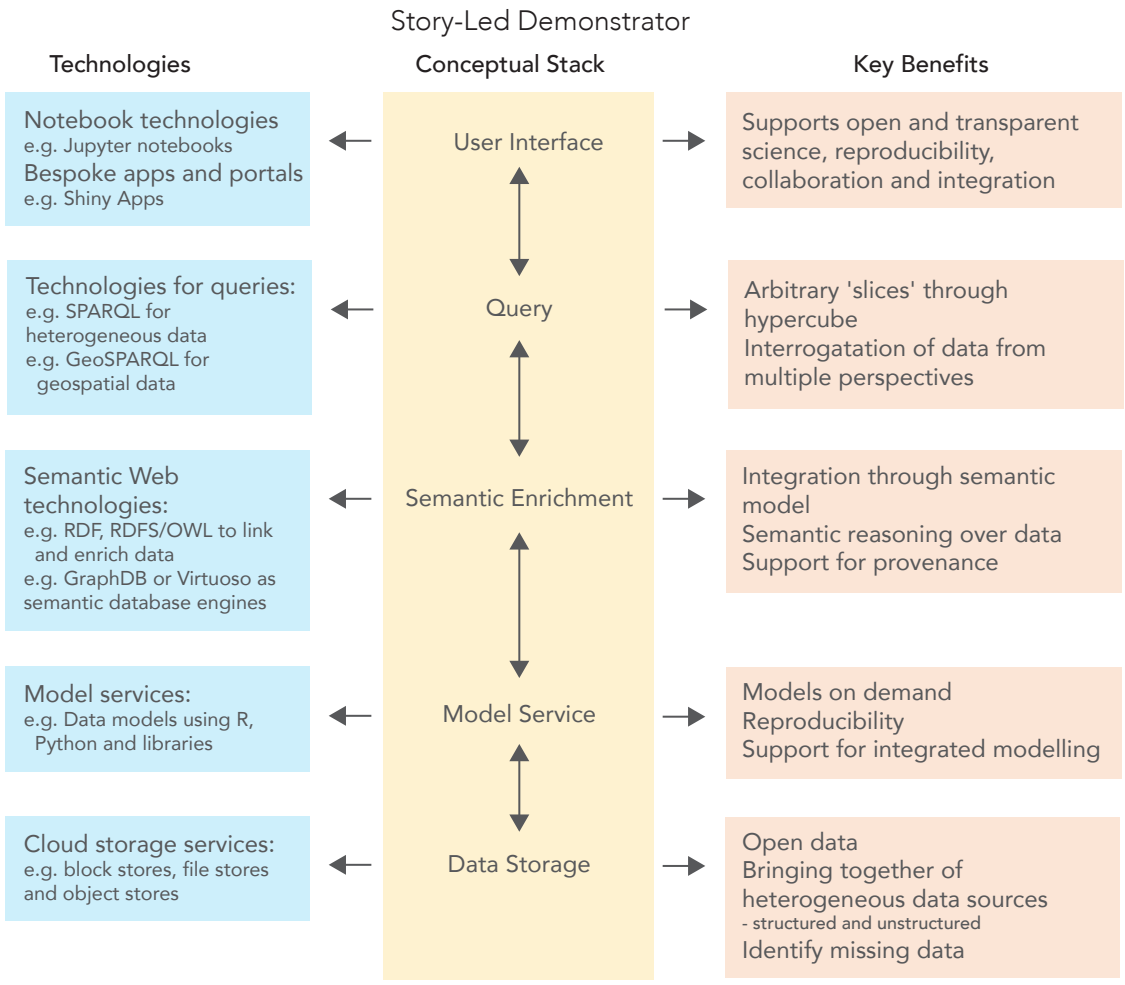



Figure 2: Story-led demonstrator mapping technologies to key benefits these technologies would illustrate for the flood risk management sector.

and associated libraries. For process models, we stored model outputs as data in the underlying data store. Bringing together models in the cloud facilitates open, transparent and reproducible science and paves the way for integrated modelling (Simm et al., 2018).

The next layer up is core to constructing the hypercube (see below), achieving the necessary integration and semantic enrichment. This is achieved through the use of Semantic Web technologies to decorate and semantically enrich the data and support rich, flexible queries over the resultant integrated, linked web of data. In practice, we used the technologies of RDF and RDFS/OWL to represent the semantically enriched data and provide a linked data solution. We also experimented with two contrasting database engines to implement the hypercube, namely GraphDB and Virtuoso (the former being a graph database and the latter being a hybrid, multi-model data store embracing both SQL and NoSQL approaches).

Layered on top of this, we offer querying facilities through the emerging standard SPARQL enabling rich queries from a variety of perspectives, for example finding all available data around a specific flood event and location, or discovering the financial impact on small businesses in a given region. We also experimented an extension to SPARQL, GeoSPARQL supporting the interrogation of geospatial data. This allows different stakeholders to make arbitrary queries on the underlying data, for example to discover how small businesses were impacted by a particular flood event, or how bringing together all the available data about this particular flood to assess more accurately the nature of the flood and its impacts. We can effectively examine the hypercube from a wide range of different perspectives.

Finally, the top level of the stack uses notebook technology, and in particular Jupyter notebooks, a user interface that intrinsically supports openness, sharing, collaboration and reproducibility. This enables the users to present different scenarios, reports and investigations to a variety of stakeholders. The architecture also enables the development of more specialist interfaces using technologies such as Shiny Apps or general web interfaces/ portals (Hunter et al., 2018). Note that Jupyter readily supports all the underlying technologies used in the layers below.

### 3.4 Defining flood events as a $n$-dimensional hypercube

Within FRM, data are rich and multidimensional. For example sources include flood source data (surface, sewer, fluvial, groundwater, coastal, reservoir, etc), spatial data (pathways and hazard), temporal data (past, real-time, forecast), probability domain (to fully describe risk), defence asset and fragility data (pathways), receptor data (properties, people, infrastructure), vulnerability data (people, property resilience), impact data and qualitative data. There has also been a proliferation of statistical and process data generated from a range of sources which have the potential to support the decision-making processes for FRM. Typically, these sources are held by different organisations, in different formats/structures, at different locations, and the data, which have been collected for multiple purposes, have varying resolutions, and temporal and geographical dimensions. Computing techniques such as *cloud computing* can be used to bring together disparate data sources in flexible ways to accommodate live and constantly updating data streams, as well as the addition of new data as and when they become available. Such an approach can also embrace model outputs as additional data sources, with the potential to run models on demand to fill in gaps. Cloud computing has the additional advantage of offering large-scale and on-demand (elastic) capacity to support this integration, in terms of storage and processing. The end result of the data-centric approach described above is a complex web of both structured (for example observational weather data) and unstructured data (for example data from post event flood reports) that can be referred to as an $n$-dimensional *hypercube*, capturing this vast range of factors that can be used to define a flood event (as illustrated in Figure 3).

A specific instance of the knowledge graph represented by the hypercube concept can be imagined

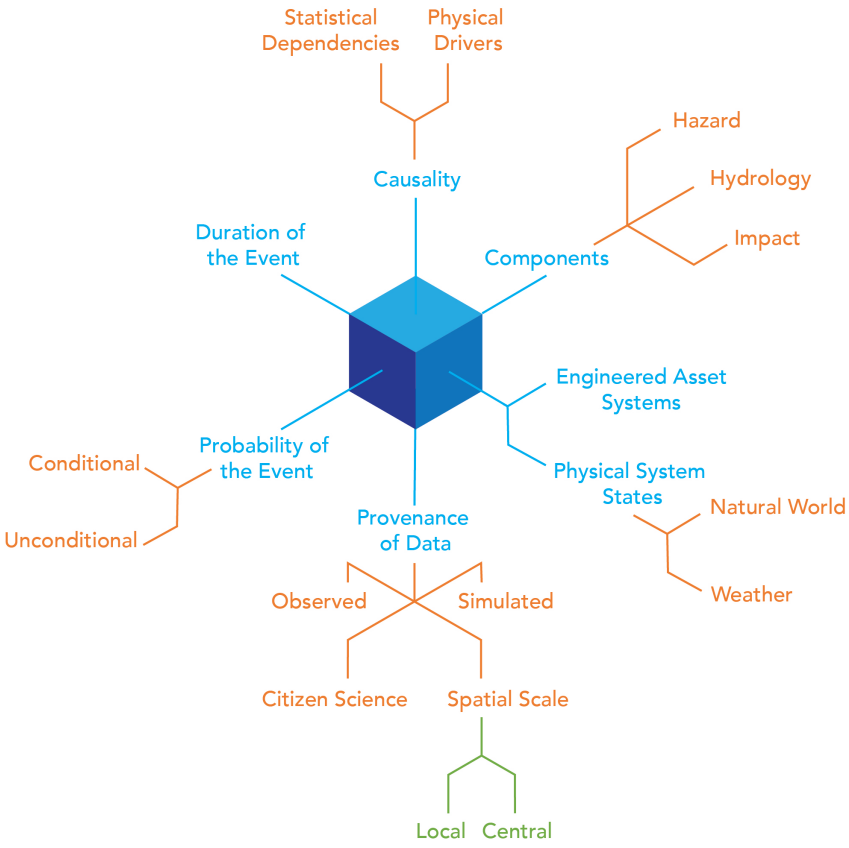



Figure 3: The concept of a data hypercube for a flood event.

for a real or simulated flood event. Here, we illustrate the idea for the floods of Autumn and Winter 2000 across the UK in Figure 4. Information shown in black text represents data that was available during or soon after the event. The text in red represents information that became available later, capturing research that added to our knowledge and understanding of the event. In particular, this includes research about the influence of climate change on the floods in 2000, which could not be quantified at the time but has since been assessed through event attribution studies (Kay et al., 2011). Also shown is an example of how the knowledge in the hypercube can continue to grow; here by the future addition of estimates of economic cost, based on a methodology that was developed after subsequent flood events (Environment Agency, 2018b). The knowledge embedded within the example is real, but is shown for illustration of the concept and represents only a subset of the data and metadata that could be included in a full implementation.

### 3.5 Integration of different data sources

Through thinking of flood events as a $n$-dimensional hypercube, it allows us to integrate together both structured and unstructured data, representing different scales and time series to provide the richest picture of a flood event. Once we can assimilate heterogeneous data within models, it becomes feasible to introduce data sets that would previously have been incompatible with the

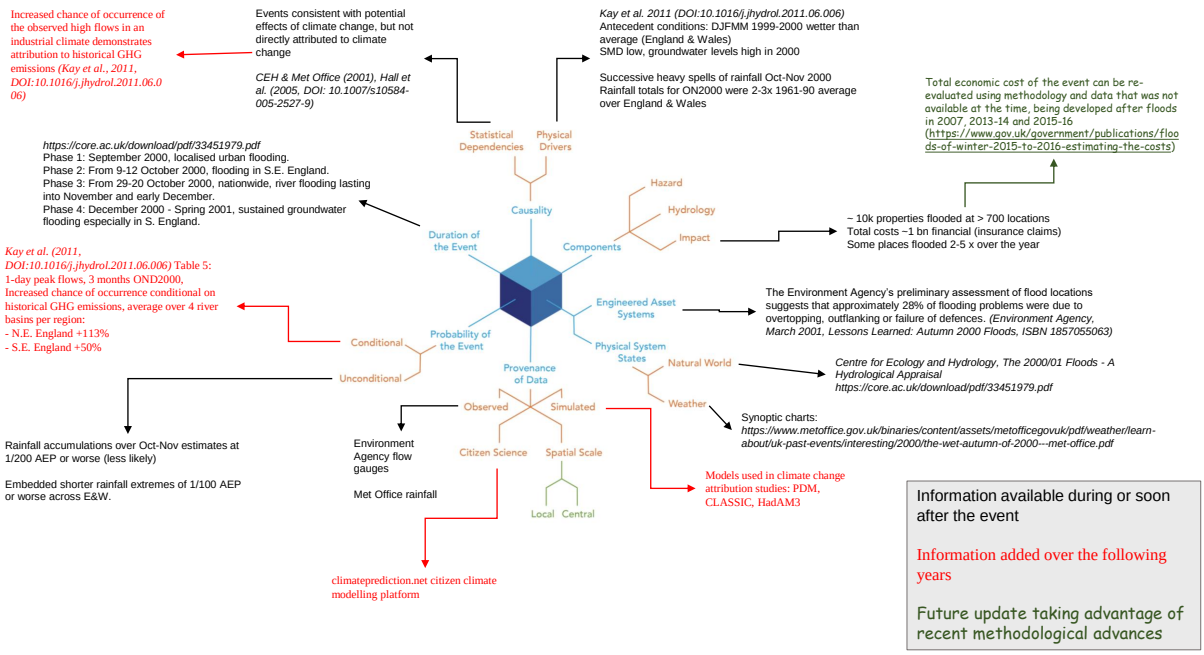



Figure 4: An instance of the data hypercube populated with information about a specific flood event, in this case the flooding that happened in the Autumn and Winter of 2000 in the UK.

model structure. This in turn opens up the potential for developing novel methods for gathering quantitative and qualitative data. Solving data consolidation challenges to manage a wide variety of data is one step towards reducing uncertainty and building a more detailed appreciation of flood risk, but, for this to have the greatest outcome, people must be engaged in contributing their data and experiential knowledge (Edwards et al., 2017). Data science methods such as data mining or machine learning can also be used to mine data from alternative sources such as documents or from social media.

Bringing together these data sources also exposes gaps and reveals opportunities for reducing the fuzziness (caused by spatial gaps, gaps in the hazard data in the probability domain as well as gaps in quality) using alternative data sources. However, it is essential that inconsistencies within the combined data are made visible.

### 3.6 Decorating data with meta-data to support semantic reasoning

When models and heterogeneous data are integrated, it is critical that people can track the provenance of individual data streams due to the variation in data collection methods, model assumptions, technologies and techniques, spatial resolution, age of the data, and general quality of the data.

This can be addressed by retaining and enhancing *metadata* (information about the data itself) to expose uncertainty within decision making. Semantic web technologies allow us to establish links

between data sets which are not only understood by humans but also by machines; this notion allows us to efficiently query across a large number of heterogeneous data sets, at different spatial and temporal scales. Ontologies are powerful data models that can represent an aspect of a domain, with metadata surrounding the domain used to enrich the data. The ontology not only represents a schema of the different domain concepts, but it also has a reasoning capability which can be used to infer new knowledge based on the existing knowledge created in the ontology. This approach also goes significantly beyond supporting provenance, allowing arbitrary semantic reasoning about the domain.

In our proposed approach, semantic triples are used to describe the ontological relationships in terms of a subject-object and predicate, an example is given in Figure 5 and Figure 6 shows how these semantic triples form a ontology that can be reasoned over. The ontology developed for our hypercube approach is based on the Environmental Impact Ontology (Garrido and Requena, 2011). Although the discussion in Sections 3.5 and 3.6 has primarily focussed on data, models can also be semantically enriched, for example including information on the computational hardware and date for which the model was run.

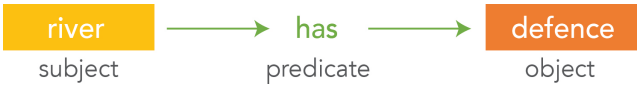



Figure 5: The subject-predicate-object structure of a semantic triple.

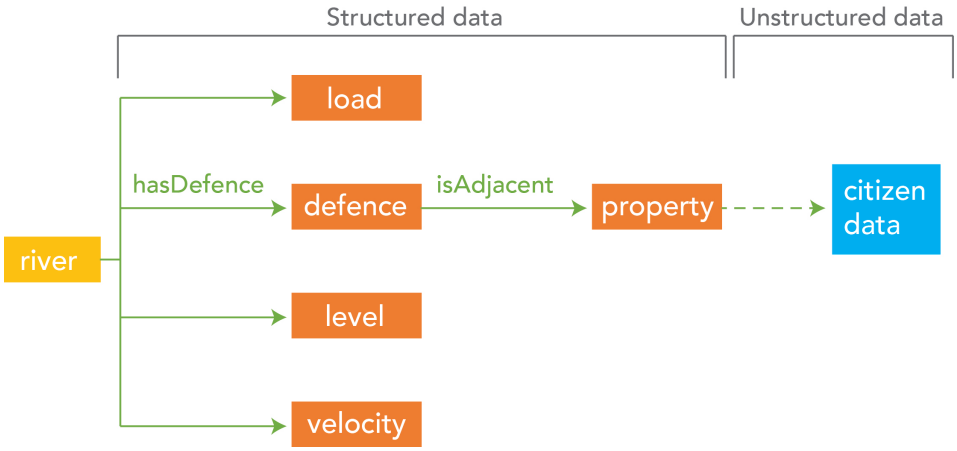



Figure 6: Diagram to show the relational structure of items within an ontology.

### 3.7 The use of data science methods to extract meaning from complex data

Through integrated data and making it easily queryable, we can then start to think about gaining a greater understanding of both the qualitative and quantitative data sets that are available. Data science allows us to extract meaning from complex data sets. This is a fast moving field, drawing

principles and techniques from a number of different disciplinary areas including computer science, statistics and complexity science (Dhar, 2013; Provost and Fawcett, 2013).

We argue that FRM in particular has a lot to gain from innovations in the data science literature, with techniques emerging to aid sense making and reasoning based on the complex web of data alluded to above. This offers the opportunity to complement the rich array of process models emanating from engineering and physical or social sciences with an equally rich set of data-driven models (Blair et al., 2019a). Methods are often not easily accessible and require expert knowledge and support from data scientists/statisticians. Therefore, there should be more of a focus on making all of the models open, easily integrated and reproducible in support of more transparent decision making. We firmly believe that such an approach offers rich scope for innovation in support of FRM.

### 3.8 Decision making from these different data sources

Once the data has been integrated and analysed, the next step is to support arbitrary investigations, for example through queries over this data to meet the needs of different stakeholders and, more generally, perspectives on the data. This represents the ability to take arbitrary slices through the hypercube. This is important for example to enable the exploration of specific places or events, using a multitude of models to test hypotheses from the perspective of multiple users with different concerns. This increases the value of the data because it can be utilised by far more decision makers to ask questions about flood risk that are specifically tailored to their needs and concerns, in ways that have not been previously possible without substantial investment for bespoke and time-limited structures.

These arbitrary slices can also be a collection of models and data, which can be combined together to form a scenario. Scenario analysis has a crucial role to play in supporting decision making around FRM (Harvey et al., 2012). In our context, we define a scenario as a data-driven investigation by a given stakeholder or stakeholders, typically in support of a given FRM decision, for example an investment in appropriate defences or natural flood management schemes. Scenarios effectively become first class entities in the underlying hypercube, and previous scenarios can be discovered as with any other data, and used to inform or inspire future scenario-based investigations. A (national) scenario library would effectively negate organisational data silos and instead move towards integrated analyses in the cloud. This in turn would support novel data combinations providing a much richer picture of flood risk and potentially reducing uncertainty (Uusitalo et al., 2015).

# 4 How the data-driven approach can support flood risk management decisions: An illustrative example based on real data and the storyboard approach

The proposed hypercube approach is illustrated through the creation of a notebook to better understand and query the flood risk to a particular area. An integrated flood risk analysis is undertaken for this area and the benefits of adopting a notebook based approach along with the integration of semantic web technologies. The hope is that the proposed hypercube approach empowers stakeholders to make better use of existing information in FRM decisions by integrating data, models and documents to enable queries on a single platform. In this section, we illustrate how these aspirations can be met through a specific case study from the East Midlands, UK, and an agile-inspired storyboard narrative.

## 4.1 East Midlands Communities at Risk data set

The research was carried out in the context of an Environment Agency initiative known as East Midlands Communities at Risk (Environment Agency, 2018a). The goal of the work was to bring together very large volumes of process-based model data sets (80 detailed hydraulic river models and 1000s kms of more generalised broad-scale models). These models are typically of the kind used to produce the national Flood Maps. This data was linked with flood levels predicted at 150,000 individual properties (some with topographic surveys of threshold elevations, i.e., the precise level at which water would start to flood the building) and 200 local river gauges (Figure 7). Combing these data sets helps to understand the risk profiles of the different properties under a number different scenarios and mitigation strategies.

One of the features of the 'East Midlands Communities at Risk' was its incorporation of different data types and qualities (detailed/generalised model; presence or absence of threshold survey; date and currency of topographic survey data).

The organisation of this data required a lot of experience and effort to be converted into an appropriate format. The work also produced a suite of interactive maps (Hankin et al., 2017) to visualise how properties may be impacted based on the reading at the community gauge. The process-based models and their outputs consist of a large amount of data, but the process of deriving level-to-level relationships form this information based on geospatial queries enables this information to be condensed into a much smaller data volume because only a part of each model domain or simulation will be important in determining each level-to-level curve. The process-based models embody physical data including topography and drainage network layout. It is recognised that in reality there could be multiple physical variables that determine the flood hazard at a specific location (e.g. upstream flow, downstream water level, position of flood gates). In general, the approach demonstrated here supports such many-to-one mappings, although here we have illustrated

only the simplest 1:1 (level-to-level) cases.

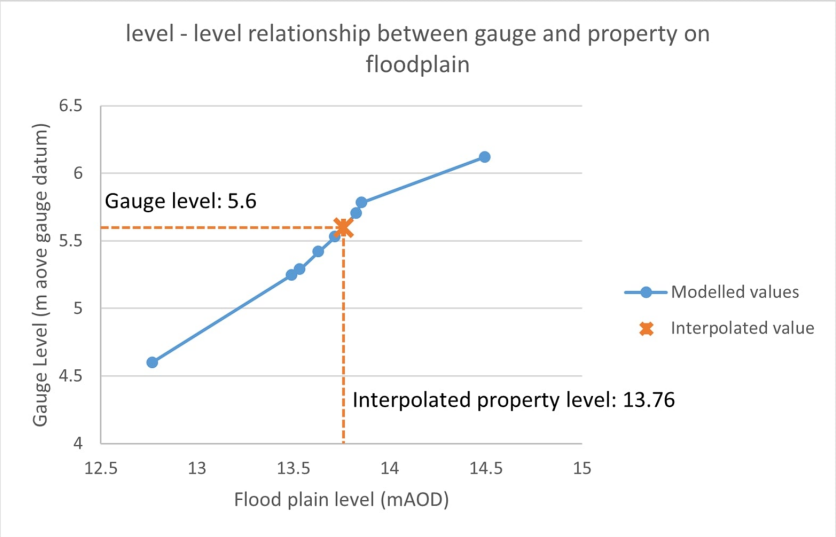



Figure 7: The figure shows the relationship between the severity of an event (recorded as return period) against the associated damage for a particular property. The blue circular points represent the return periods for which the flood inundation model was run and the blue line is interpolated. It is computationally intensive therefore prohibitively expensive to model values for every flood plain level, so models are generally run for specified values. The figure shows that property levels can be interpolated by amalgamating results of model runs.

The work adopts an integrated risk assessment approach involving a source-pathway-receptor model. The team and partners were keen to investigate how a datacentric approach might bring in additional sources of data, reveal new data science techniques and create more flexible ways of interrogating and working with the 'East Midlands Communities at Risk' data and associated analyses methods.

## 4.2 Expected annual damage

In our demonstration of the hypercube approach we adopt the definition of expected annual damage that is derived from the source-pathway-receptor approach for integrated risk assessment (Shaw et al., 2010).

Expected Annual Damage (EAD, the area under the impact versus probability curve) is a commonly-used estimate of the long term annual average damages from flooding. It can contribute to the understanding of flood risk, and is important in economic appraisal of the cost-efficiency of FRM proposals. Other social and economic factors may be considered in appraisal, such as the distribution of risk with respect to different social groups, and our approach could generalise to include them. The hypercube approach not only allows the EAD to be calculated for a number of

variables of interest to flood risk managers, but also enables the models used in this calculation and their associated uncertainties to be retrieved (Hall et al., 2003).

Although the estimate of EAD is typically presented as a single number it is the product of a number of different models with differing assumptions. These assumptions are not normally apparent but the hypercube approach is designed to make them visible (Metin et al., 2018).

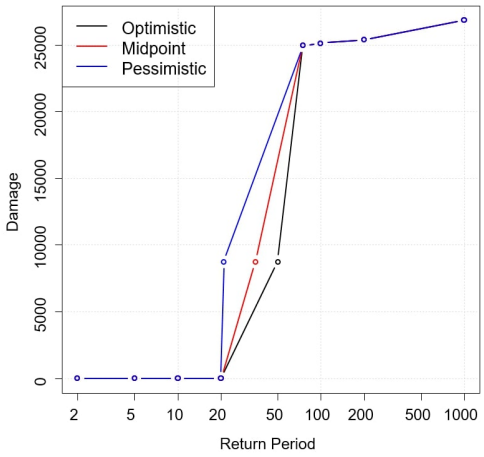
The EAD should ideally integrate over all possible values of the hazard as well as multiple states of the flood defence (see Figure 7). However, this is computationally expensive as a costly flood inundation model is run for each of the combinations of the hazard and state of the flood defence. Consequently, for most studies, the model is only run for a finite number of events. Further assumptions are made including the spatial resolution of the inundation models. These simplifications to the estimates of the damage make the problem tractable but induce further uncertainty.

A key result of the discretisation of the relationship between the severity of the event and the associated damage is that the onset of flooding is unknown. In an ideal world, the exact return period for which damages start to be incurred would be known, to understand this further we can think of the calculation of EAD in Figure 8 for a single property as being

$$\text{EAD} = \int_{Rp_{\text{min}}}^{Rp_{\text{max}}} D(Rp)dRp = 0 + \int_{Rp_*}^{Rp_{\text{max}}} D(Rp)dRp,$$

where $D$ is the damage function evaluated at a particular return period ($Rp$) with $Rp_{\text{min}}$ and $Rp_{\text{max}}$ are the minimum and maximum return periods for which the damage function was estimated. The damage function is zero up and until the onset of flooding, which has the associated return period $Rp_*$. The return period $Rp_*$ itself is unknown as the damage function is only evaluated at a finite number of return periods. A typical assumption would be to assume that flooding occurs somewhere between when the model runs estimate zero damage to non-zero damage (midpoint assumption) with associated return period $Rp_*^{mid}$. The return period $Rp_*$ can also be bound as follows, $Rp_*^{pes} < Rp_* < Rp_*^{opt}$, we refer to these bounds as being the pessimistic and optimistic cases for when the first observation of estimated damage occurs. The pessimistic case $Rp_*^{pes}$ assumes that the damage occurs after the final estimate of zero damage. Whereas, the optimistic case $Rp_*^{opt}$ assumes that damages occur just before the first estimate non-zero of non-zero damage.

The expected annual damage calculations in Figure 8 show that different assumptions about the onset of flooding cause a variation of 300 GBP in the estimate of expected annual damage. This estimate of expected annual damage is only for a single property but the impacts of the different assumptions for the onset of flooding can propagate through the calculations of risk and impact. Further simulations can narrow down this window to determine a more accurate estimate of the onset of flooding. These assumptions mean that the likelihood of flooding occurring could be underestimated or that the estimate of damage may be under or over estimated. It is therefore key to expose this uncertainty to make sure that decisions are as well informed as possible.




| **Assumption** | **EAD** (£) |
|---|---|
| Optimistic $Rp^{opt}$ | 651.26 |
| Midpoint $Rp^{mis}$ | 741.58 |
| Pessimistic $Rp^{pes}$ | 990.10 |

Figure 8: Left: The relationship between the severity of an event (recorded as return period) against the associated damage for a particular property (estimated in sterling). The circular points represent the return periods for which the flood inundation was run. The red line shows the midpoint assumption, the black (blue) shows the pessimistic (optimistic) assumption about the onset of flooding. Right: Estimates of expected annual damage for the return period damage curve calculated for each of the onset of flooding assumptions.

Other sources of data can also provide valuable additional information in calculating the EAD, for example real time flood levels can be used to produce real time estimates of damage. Combining these data sources would allow for real time warnings for properties or areas that are now in a higher risk category of flooding. Furthermore, other unstructured sources of data such as news reports should be retrieved to provide a richer picture of flood risk and determine whether models are now out of date. With the examples given here, we aim to illustrate how different sources of uncertainty can be exposed, building towards a shared understanding of the quality of evidence about flood risk. This is not a static view of uncertainty, but rather one in which new information can be incorporated, as illustrated in the flood event data object illustrated in Figures 3 and 4. Overall, the idea is to enhance confidence in a decision process, rather than to validate or assess uncertainty in specific flood risk data sets.

### 4.3 Estimates of expected annual damage for Newark

Focusing on Newark, we are able to retrieve estimates of expected annual damage with example output shown in Figure 9. The existing estimates of expected annual damage are retrieved from the Communities at Risk data set as well as the estimates with the different assumptions of the onset of flooding.

This is also presented in tabular form as if all of the properties in Newark were protected by a

flood defence up to a certain level. This allows the end user to gain a quick understanding of the potential impact of future flood schemes and the levels of protection that are economically viable.

Another key aspect of Figure 9 is how the expected annual damage estimates are broken down according to flood warning areas associated with different gauging stations within the Newark area. This clearly shows that the economic risk in terms of EAD differs between different locations, which may influence the prioritisation of future flood defences and this information can also be visualised spatially and can be displayed at an individual property level. Finally, information about existing flood defence schemes can also be viewed. In Figure 9, we have provided a small snapshot of information about a (hypothetical) proposed flood defence scheme.

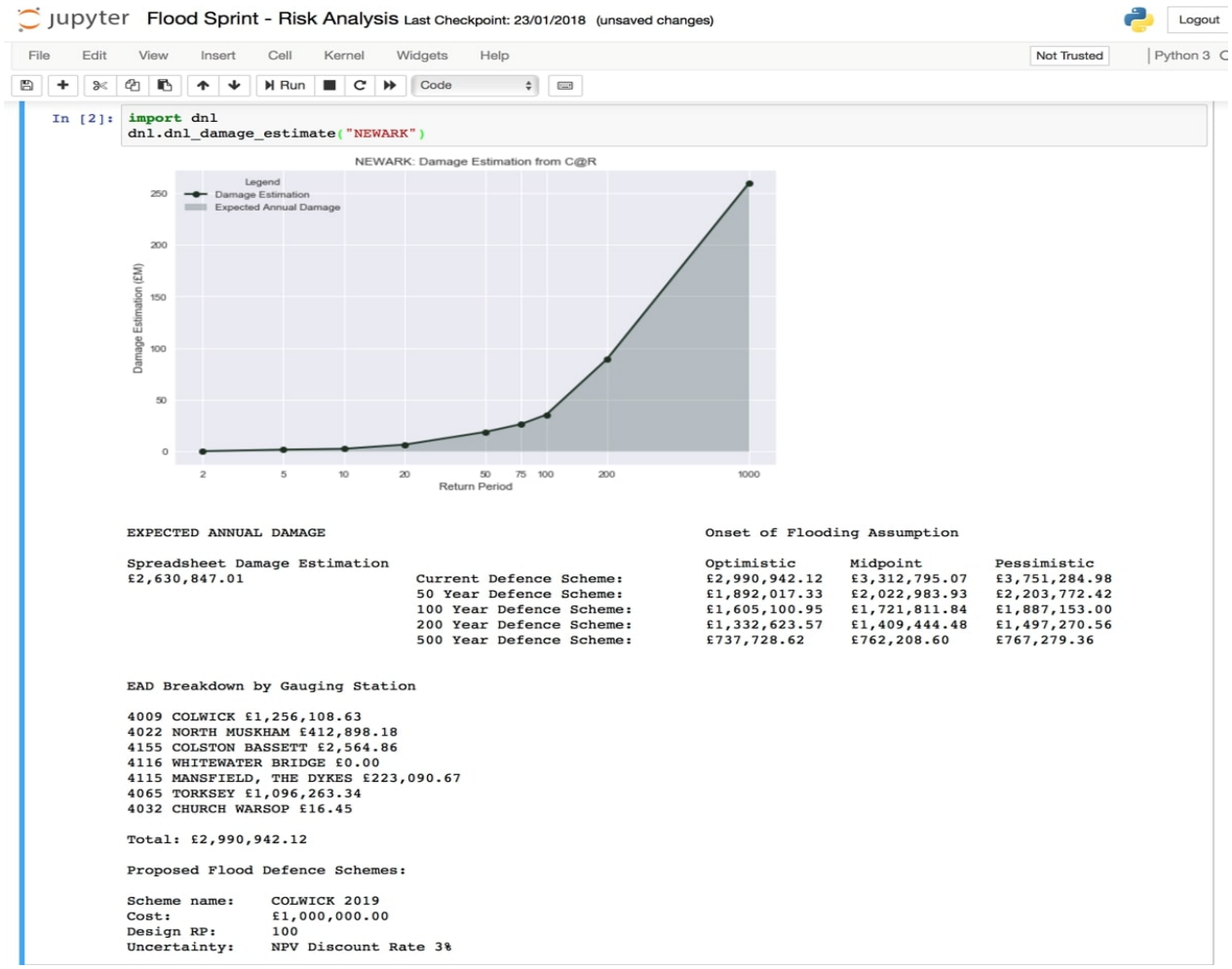



Figure 9: Estimated expected annual damages with associated metadata for Newark-on-Trent from the flood risk demonstrator. The expected annual damage has been broken down by gauging stations with information also provided on proposed flood defence schemes.

### 4.4 Incorporating unstructured data

Utilising the ideas given in Sections 3.5 allows us to incorporate additional sources of information from different unstructured sources. To illustrate this concept, we focus on Section 19 Reports, which are flood investigation reports produced by the Risk Management Authorities (including Local Authorities and the Environment Agency) in the UK in response to flood events. They are unstructured documents, often stored as PDF files, but there is no data model to support the extraction of key information. They do though contain rich qualitative and quantitative data collected from stakeholders post-flood, including information such as the severity of the event, depth of inundation, the locations that were flooded and the demand on the emergency services. Natural Language Processes (NLP) (Bird et al., 2009; Chowdhury, 2003) techniques are used to extract meaning from unstructured data such as these reports. These processed data are stored using a semantic structure that references domain concepts, enabling links to be made with existing structured data (as shown in Figure 6). This integration is done at the metadata level, where the data atoms are represented by domain concepts that are linked through the ontological schema. Note that the techniques deployed focus mainly on objectively extracting raw information, and there is no inference from this data given the levels of uncertainty this can introduce; any inference is in higher levels of the architecture where the assumptions and methods used can be explicit (recorded in notebooks). More details of this integration can be found in Nundloll et al. (2020).

The estimates of damage from these Section 19 reports could be included as an extra point on the depth-damage curve in Figure 7. In many cases, the estimates of damage and the severity of the event may also be uncertain. This uncertainty could also be captured in any calculation of expected annual damage. In some cases combining this new event information with the existing output from the hydraulic model may visualise potential inconsistencies. Such information should be readily available to any end user to help ensure that existing data and models used to support FRM decisions can be updated or reinterpreted appropriately to take account of subsequent events.

Figure 10 shows an example for Colwick through accessing information from a news report to determine the area the event has affected. This additional information has been combined with the Communities at Risk data set to update the estimate of expected annual damage, therefore reducing the estimated damage to this particular property.

### 4.5 Evaluation of the hypercube approach

The notebook based approach allows us to easily query and understand the value of the data available and ultimately expose those parts of the hypercube where data are scarce. This section also has shown that our approach enables flood risk managers to carry out risk assessments and obtain estimates of expected annual damage. In addition, the notebook based approach can incorporate additional sources of unstructured information. These extra sources of information provide contextual information and allow the ability to query in a more flexible and responsive manner than

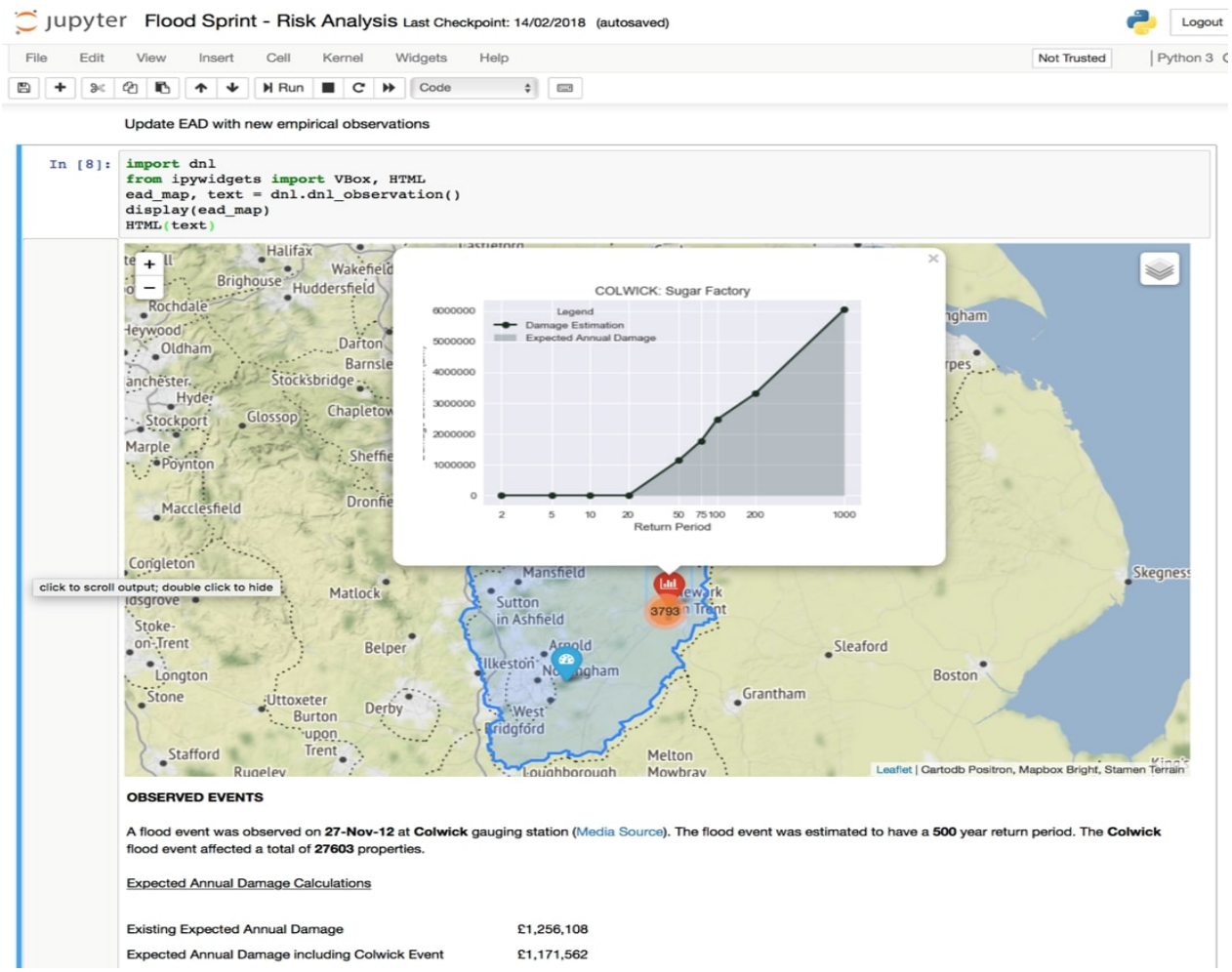



Figure 10: Incorporating unstructured data from a news report about Colwick into the estimates of expected annual damage. The plot shows the updated depth damage for a property that has been affected by this new event.

existing flood risk software tools.

When this is applied to the characters in our storyboard we find that Alicia can assess the relative strengths and weaknesses of different modelling assumptions in the estimation of flood risk. With this knowledge Alicia is now able to select the most appropriate models for the analysis of flood risk in Newark.

Bashir is able to quickly visualise and understand the impact of possible future decisions to mitigate against flood risk. As the hypercube approach has the capacity to integrate data sources that were previously excluded from this kind of risk assessment, Bashir not only has a richer picture of flooding but is able to be responsive to local residents, who may feel that their input has not traditionally been included in decision making processes.

In order to further assess the viability of the hypercube approach, we assess whether the visions

for the sector (as discussed in Section 3) are illustrated by the hypercube approach. This evaluation is shown in Table 2, for a number of visions the hypercube approach has provided clear benefits, however further research is required to address all of the visions and associated challenges to fully accommodate a data-centric approach for FRM.

| **Vision** | **Evidence** |
|---|---|
| Integrating heterogeneous data through cloud computing | The development of an ontology has provided a means of connecting structured and unstructured data sources enabling systematic queries across disparate data sets. The development of the ontology also allows for new data to be incorporated as and when they become available. An extension would be to enable reasoning using the existing knowledge in the ontology to infer new knowledge as well as rerunning models to account for these new observations. |
| Incorporating new sources of data | The flood risk analysis for Newark-on-Trent has utilised both qualitative and quantitative data, which typically is not available as part of a standard flood risk analysis. Further development is needed to process and support real time information from social media sources such as Twitter. |
| Tracking provenance through enhanced metadata | The use of semantic web technologies and the development of an ontology has meant that any notebook query is able to return the relevant data as well as the associated metadata. Users can choose the level of detail and amount of metadata that is returned. |
| Embracing the peculiarities of place | The Newark case study shows that the hypercube approach can be used to understand place at a variety of scales from fine grain property level returns, needed by the insurance industry to regional overviews, needed for long term infrastructure planning. |
| Multi-perspective data queries | Separate notebooks have been developed to cater to the differing needs and requirements of Alicia and Bashir. These notebooks can be extended to capture varying needs of the large number of stakeholders who have an interest in flood risk. More research is needed to adapt the design of the notebook interface for use amongst non-specialist audiences. |
| Embracing data science | Natural language processing techniques have been used to extract data from unstructured data, in particular Section 19 reports. The notebook provides a basis for developing and supporting other data science techniques, however this is not possible in the current version of the notebook. |
| Supporting scenario libraries | The notebook approach has shown that a number of different scenarios can be stored and queried from different perspectives. The benefit of the notebook approach is that these queries can be stored and rerun when more data are incorporated into the notebook. This idea provides a basis for a number of different scenarios to be considered and reasoned over. This could result in the ability to determine where the largest gaps in the data exist and the best way to fill them to reduce this uncertainty. |

Table 2: Relating the outputs of the hypercube approach to the visions for the flood risk management sector as discussed in Section 3.

# 5 Overall Evaluation

This section presents our overall evaluation including our main findings including limitations, challenges and future work.

The most significant finding from the research is the *success of the agile methodology* in exploring potential digital futures for FRM. Our experience is that agile is already familiar to software development teams working in the flood management industry. Here, we moved away from a focus purely on software development, and applied agile principles to enable participatory research on FRM and digital technologies. The approach was successful in drawing in stakeholders to the point they were an intrinsic part of the team (contrast this to a more traditional approach that would create an inevitable separation between the development team and the users). The methodology also encouraged cross-disciplinary dialogue where different perspectives were shared (contrast this with the traditional approach that inevitably emphasises the role of the software developer). Finally, the approach allowed us to investigate potential futures that could not be perceived by any one constituent group, allowing us to iterate towards something that is both exciting; meeting the needs of the community and its various participants; and also is technically feasible.

It should be noted that the agile methodology has limitations as it requires concerted commitment by stakeholders to ensure the demonstrator reflects the needs of the community. This requires frequent input that has to be accounted for financially. Different stakeholders may have different priorities that require negotiation during the development process.

A second significant finding is the potential importance of *notebook technology* in support of more open and transparent science. Notebooks effectively allow experiments and analyses to be documented, including documentation of assumptions. Importantly, code, documentation and outputs are tied together, and in a way that is immediately reproducible.

The third important finding is that *contemporary cloud computing technologies* have real potential to support the next generation of FRM systems (Simm et al., 2018). Cloud computing is fast becoming a mature area and the benefits in terms of open and collaborative science are already well known[3]. Within the FRM industry, cloud technology has already been applied for the flexibility it offers in providing access to computing resource for data- and processing-intensive tasks. Cloud computing is also now populated with a range of exciting new services and our exploration has shown that such services can usefully be combined to support a data-centric approach for FRM. In particular, we effectively combined cloud storage services and semantic web technologies to construct a pilot hypercube, and this was readily extended with natural language processing services to mine data from unstructured reports. Importantly, such services have been designed to scale to massive data-sets, and our experiments have also indicated that they can readily accommodate highly heterogeneous and inter-linked data-sets. However, the limits of this have not yet been tested for an operational FRM system (Blair et al., 2019b).

[3]http://www.evo-uk.org/

Both notebook and cloud computing technologies require particular skill sets that may not be widely available to organisations that would benefit from this approach. Without the availability of these skills, the demonstrator is limited to the small group of users, who understand and trust the architecture and have the capacity to incorporate new sources of data. The next phase in this work should include research into the dissemination of the necessary skills to enable widespread adoption of the hypercube approach.

The research also highlighted a number of areas requiring further attention. The biggest research challenge as we see it is to bring together effectively a data-centric approach to FRM with the large legacy of process models around hydrology and flooding more generally. For this to be made possible, data has to be made open and accessible, though this has security and data ownership implications. There is also a need to investigate the generality and transferability of the approach. The current approach is designed to be *extensible* within the area of FRM. The underlying technology should readily accommodate additional dimensions of the hypercube and associated additional data sources. The approach should also *scale* to much larger instantiations of a hypercube as this is a key property of the cloud technologies adopted. We also see real opportunities in extending the hypercube beyond flood risk data to other aspects of environmental ecosystems to allow studies (DEFRA, 2018), etc. Note that this has implications in terms of our methodology. Incremental changes and extensions can be made through the normal iterative cycles inherent in the agile approach, whereas larger changes such as embracing other environmental concerns would inevitably mean revising the whole methodological cycle from initial workshop through creating storyboards to iterative development.

# 6 Conclusion

This paper has explored how to take advantage of the increasing amounts of data available to support FRM of the future. In particular, the paper has explored technological support for such a data-centric approach, seeking insights into software architecture principles and techniques building on the promise of cloud computing. The paper has identified a series of existing technologies which, when combined, readily meet our requirements. These include techniques for underlying data storage, semantic enrichment of this data, and mechanisms to query this data, including spatial dimensions of this data. Such an approach also helps to identify the veracity of data and to highlight missing data and inconsistencies, crucial in supporting risk management. Notebook technologies also proved to be successful in terms of supporting openness, transparency and reproducibility. The agile methodology proved to be crucial in supporting effective collaboration.

Ongoing research is investigating how to marry a more data-centric approach to the important legacy of process models. Finally, we are developing more robust and complete open source demonstrators of the proposed technological approach with a view of influencing future generations of FRM architectures both in the UK and elsewhere.

**Acknowledgements**

This research was supported by EPSRC Grant EP/P002285/1 (Senior Fellowship on the Role of Digital Technology in Understanding, Mitigating and Adapting to Environmental Change) and JBA Trust project W16-5841. We wish to thank the stimulating discussions we have had from a wide range of researchers and practitioners in partner organisations, including: JBA Consulting, the Environment Agency, United Utilities, the European Centre For Medium Range Weather Forecasts, the Oasis Loss Modelling Framework, and the Environmental Change Institute (Oxford University). The flood event "hypercube" concept originated from a discussion with Neil Hunter and Steve Hutchings. This research was completed whilst RT and GD were employed by Lancaster University.

## References

Alfieri, L., Feyen, L., and Di Baldassarre, G. (2016). Increasing flood risk under climate change: a pan-european assessment of the benefits of four adaptation strategies. *Climatic Change*, 136(3-4):507–521.

Beniston, M. and Stephenson, D, B. (2004). Extreme climatic events and their evolution under changing climatic conditions. *Global and Planetary Change*, 44(1-4):1–9.

Benson, M. A. (1968). Uniform flood-frequency estimating methods for federal agencies. *Water Resources Research*, 4(5):891–908.

Beven, K. (2007). *Environmental modelling: An uncertain future?* CRC press.

Beven, K. J. and Hall, J. (2014). *Applied uncertainty analysis for flood risk management*. Imperial College Press London.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Blair, G., Henrys, P., Leeson, A., Watkins, J., Eastoe, E., Jarvis, S., and Young, P. (2019a). Data science of the natural environment: a research roadmap. *Frontiers in Environmental Science*, 7:121.

Blair, G. S., Beven, K., Lamb, R., Bassett, R., Cauwenberghs, K., Hankin, B., Dean, G., Hunter, N., Edwards, L., Nundloll, V., Samreen, F., Simm, W., and Towe, R. P. (2019b). Models of everywhere revisited: A technological perspective. *Environmental Modelling & Software*, 122:104521.

Boehm, B. W. (1988). A spiral model of software development and enhancement. *Computer*, 21(5):61–72.

Brunner, G. W. (1995). Hec-ras river analysis system. hydraulic user's manual. version 1.0. Technical report, HYDROLOGIC ENGINEERING CENTER DAVIS CA.

Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology*, 37(1):51–89.

Cohn, M. (2004). *User stories applied: For agile software development*. Addison-Wesley Professional.

De Nicola, A.and Missikoff, M. and Navigli, R. (2009). A software engineering approach to ontology building. *Information systems*, 34(2):258–275.

DEFRA (2018). A green future: Our 25 year plan to improve the environment. Technical report, UK Government.

Demir, I. and Krajewski, W. F. (2013). Towards an integrated flood information system: centralized data access, analysis, and visualization. *Environmental Modelling & Software*, 50:77–84.

Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, 56(12):64–73.

Edwards, E. R., Mullagh, L., Towe, R. P., Nundloll, V., Dean, C., Dean, G., Simm, W. A., Samreen, F., Bassett, R., and Blair, G. S. (2017). Data-driven decisions for flood risk management.

Environment Agency (2013). SC120002 - 2D benchmarking - evaluating the latest generation of the hydraulic models for FCRM purposes. Technical report, Environment Agency, Bristol.

Environment Agency (2018a). Communities at risk project. Technical report, Environment Agency.

Environment Agency (2018b). Estimating the economic costs of the 2015 to 2016 winter floods: LITlit 10736. Technical report, Environment Agency.

Environment Agency (2018c). National flood risk assessment 2 (nafra2). Technical report, Environment Agency, Bristol.

Environment Agency (2018d). Risk of flooding from surface water depth: 0.1 percent annual chance. Technical report, Environment Agency, Bristol.

Environment Agency (2018e). Sc110003 - evaluating and improving the grid-to-grid (g2g) model for flood forecasting in rapid response catchments. Technical report, Environment Agency, Bristol.

Environment Agency (2019). Draft national flood and coastal erosion risk management strategy for england. Technical report, Environment Agency, Bristol.

Ferrario, M. A., Simm, W., Forshaw, S., Gradinar, A., Smith, M. T., and Smith, I. (2016). Values-first se: research principles in practice. In *Proceedings of the 38th International Conference on Software Engineering Companion*, pages 553–562. ACM.

Ferrario, M. A., Simm, W., and Whittle, J. (2013). Speedplay: Managing the other edge of innovation. *DE2013: Open Digital. RCUK*.

Fleming, G. (2002). Learning to live with riversthe ice's report to government. In *Proceedings of the Institution of Civil Engineers-Civil Engineering*, volume 150, pages 15–21. Thomas Telford Ltd.

Garrido, J. and Requena, I. (2011). Proposal of ontology for environmental impact assessment: An application with knowledge mobilization. *Expert Systems with Applications*, 38(3):2462–2472.

Grand River Conservation Authority (2016). River and stream flows - grand river conservation authority.

Gumbel, E. J. (1958). *Statistics of extremes*. Courier Corporation.

Hall, J. W., Dawson, R. J., Sayers, P. B., Rosu, C., Chatterton, J. B., and Deakin, R. (2003). A methodology for national-scale flood risk assessment. In *Proceedings of the Institution of Civil Engineers-Water Maritime and Engineering*, volume 156, pages 235–248. London: Published for the Institution of Civil Engineers by Thomas Telford.

Hankin, B., Lamb, R., Craigen, I., Page, T., Chappell, N., and Metcalfe, P. (2017). A whole catchment approach to improve flood resilience in the eden. winning entry to the defra flood modelling competition 2016. Technical report, Winning entry to the Defra Flood Modelling Competition 2016.

Harvey, H., Hall, J., and Peppé, R. (2012). Computational decision analysis for flood risk management in an uncertain future. *Journal of Hydroinformatics*, 14(3):537–561.

Her Majesty's Government (08 September 2016). National Flood Resilience Review. Technical report.

Holdgate, M. W. (1980). *A perspective of environmental pollution*. CUP Archive.

Hunter, N., Lamb, R., Towe, R. P., Warren, S., and Wood, E. (2018). Spatial joint probability for FCRM and national risk assessment multivariate event modeller-user guide: SC140002/R3. Technical report, Environment Agency.

Hunter, N. M., Bates, P. D., Horritt, M. S., and Wilson, M. D. (2007). Simple spatially-distributed models for predicting flood inundation: a review. *Geomorphology*, 90(3-4):208–225.

Jongman, B., Ward, P. J., and Aerts, J. C. (2012). Global exposure to river and coastal flooding: Long term trends and changes. *Global Environmental Change*, 22(4):823–835.

Kauffeldt, A., Wetterhall, F., Pappenberger, F., Salamon, P., and Thielen, J. (2016). Technical review of large-scale hydrological models for implementation in operational flood forecasting schemes on continental level. *Environmental Modelling & Software*, 75:68–76.

Kay, A. L., Crooks, S. M., Pall, P., and Stone, D. A. (2011). Attribution of autumn/winter 2000 flood risk in england to anthropogenic climate change: A catchment-based study. *Journal of Hydrology*, 406(1-2):97–112.

Metin, A. D., Dung, N. V., Schröter, K., Guse, B., Apel, H., Kreibich, H., Vorogushyn, S., and Merz, B. (2018). How do changes along the risk chain affect flood risk? *Natural Hazards and Earth System Sciences*, 18(11):3089–3108.

Mitchell-Wallace, K., Jones, M., Hillier, J., and Foote, M. (2017). *Natural catastrophe risk management and modelling: A practitioner's guide*. John Wiley & Sons.

Moore, R. J. (2007). The pdm rainfall-runoff model. *Hydrology and Earth System Sciences Discussions*, 11(1):483–499.

NERC CEH (2018). National river flow archive (NRFA). Technical report, NERC CEH, Wallingford.

NERC CEH (2019). Hydro-jules. Technical report, NERC CEH.

Nundloll, V., Blair, G., Hankin, B., Dean, G., Edwards, L., Lamb, R., and Towe, R. (2020). A semantic approach to tackle data integration for a flood domain. *In preparation*.

Provost, F. and Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big data*, 1(1):51–59.

Samuels, P. (2009). Language of risk: project definitions. *T32-04-01*.

Shaw, E. M., Beven, K. J., Chappell, N. A., and Lamb, R. (2010). *Hydrology in Practice*. CRC Press.

Simm, W. A., Samreen, F., Bassett, R., Ferrario, M. A., Blair, G., Whittle, J., and Young, P. J. (2018). SE in ES: Opportunities for software engineering and cloud computing in environmental science. In *Proceedings of the 40th International Conference on Software Engineering: Software Engineering in Society*, pages 61–70. ACM.

Tawn, J., Shooter, R., Towe, R., and Lamb, R. (2018). Modelling spatial extreme events with environmental applications. *Spatial statistics*, 28:39–58.

Teng, J., Jakeman, A. J., Vaze, J., Croke, B., Dutta, D., and Kim, S. (2017). Flood inundation modelling: A review of methods, recent advances and uncertainty analysis. *Environmental Modelling & Software*, 90:201–216.

The World Bank (2018). Proceedings from the 2018 UR forum on understanding risk: Disrupt.communicate.influence. Technical report.

UNISDR, U. (2015). Sendai framework for disaster risk reduction 2015–2030. In *Proceedings of the 3rd United Nations World Conference on DRR, Sendai, Japan*, pages 14–18.

US Army Corps of Engineers (2016). Hec-ras river analysis system. hydraulic user's manual. version 5.0. Technical report, HYDROLOGIC ENGINEERING CENTER DAVIS CA.

US Geological Survey (2016). National water information system data available on the world wide web (usgs water data for the nation).

Uusitalo, L., Lehikoinen, A., Helle, I., and Myrberg, K. (2015). An overview of methods to evaluate uncertainty of deterministic models in decision support. *Environmental Modelling & Software*, 63:24–31.

Welsh Government (2019). The draft strategy for flood and coastal erosion risk management in wales. Technical report, Welsh Government, Bristol.

World Bank (2016). Solving the puzzle: Innovating to reduce risk. global facility for disaster reduction and recovery (GFDRR). *The World Bank*.