A Data Science of the Natural Environment: Applied to Flood Risk Management

21st November, 2017

Prof. Gordon Blair, EPSRC Senior Fellow in Digital Technology and Living with Environmental Change

Introducing me!

- Distinguished Professor of Distributed Systems, School of Computing and Communications, Lancaster University
- Theme Lead (Environment), Data Science Institute, Lancaster
- EPSRC Senior Fellow in Digital Technology and Living With Environmental Change (DT/LWEC)
- Senior Fellow, Centre for Ecology and Hydrology

… and part-time shepherd!

Overview of the talk

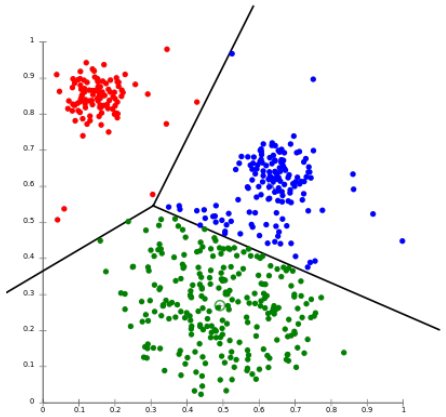
- Part I: Environmental data science
  - What is environmental data science
  - Challenges in the area
  - Scope of our work
- Part 2: Digital Technology and Living with Environmental Change
  - Aims of my fellowship
  - Approach
  - The first sprint on flooding

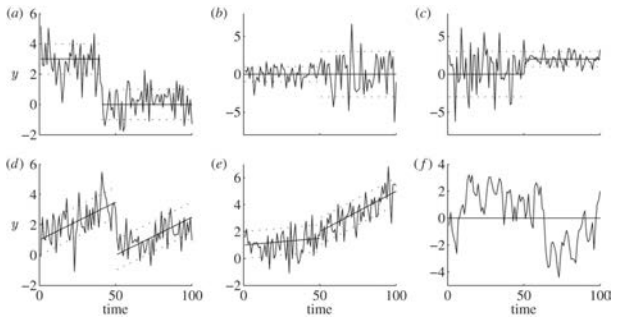
Part 1

Environmental data science

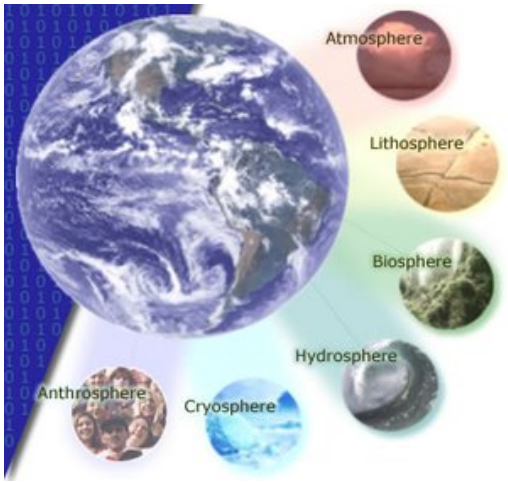
What is environmental data science?

- The development of data science principles and techniques for sense making and decision support related to the natural environment
- Focus on methodological innovation
- Importance of integration





Challenges of the area



- The complexity challenge
- The data challenge
- The modelling challenge
- The cross-disciplinary challenge
- The spatial/ temporal challenge
- The uncertainty challenge

Scope of environmental data science

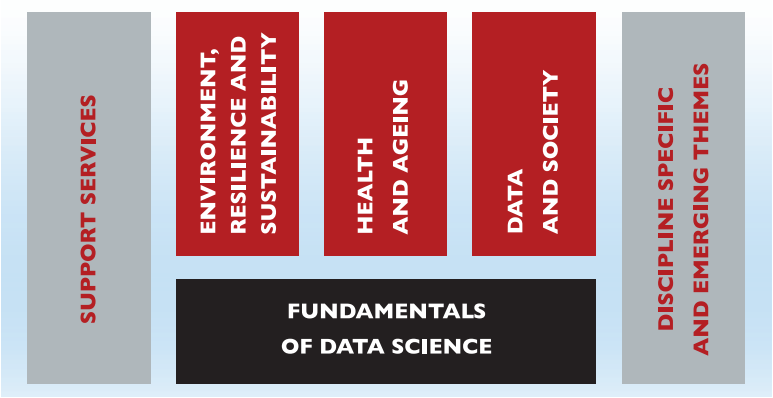
From data science to policy and strategy: decision making under uncertainty

Data science methods: models and combinations of models

Data science infrastructure: data discovery, storage and processing services

Data aquisition: observation and monitoring of the natural environment

Centre of Excellence in Environmental Data Science



+

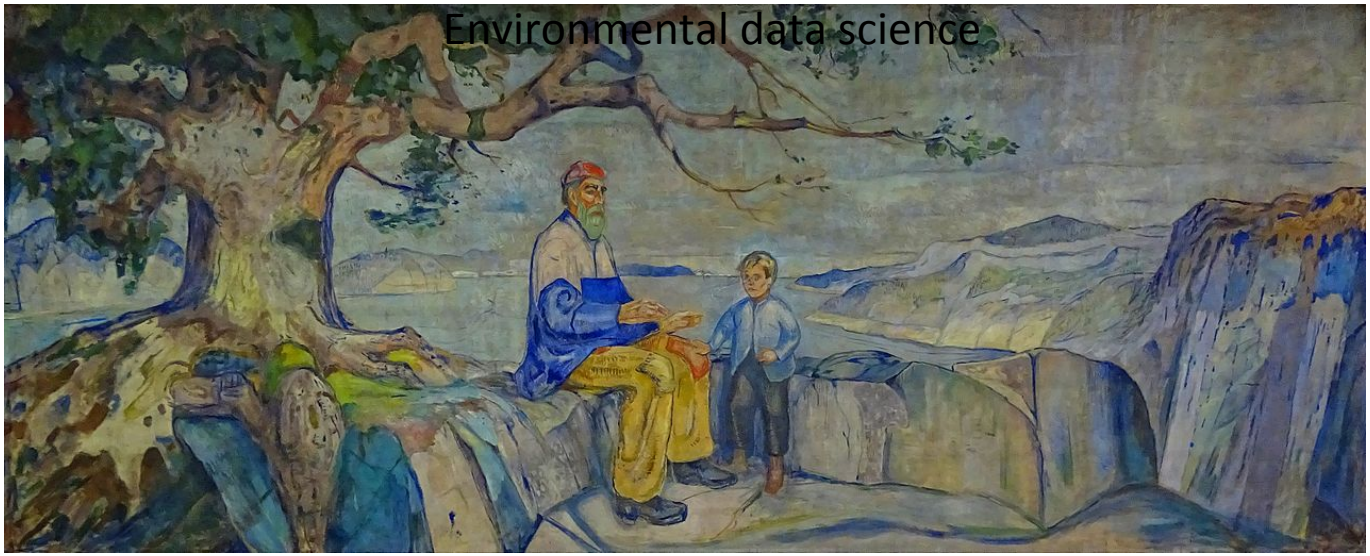
CEH Centre for Ecology & Hydrology
NATURAL ENVIRONMENT RESEARCH COUNCIL

A research roadmap

- **Challenge 1**: To encourage and enable the required level of ***cultural shift towards open science***, that is towards a science that is more collaborative and integrative through open approaches to data, models and knowledge formation, and also towards a science that is more transparent, repeatable and reproducible.
- **Challenge 2**: To build on the benefits of cloud computing, but offer ***levels of abstraction*** (and associated services) that are much better suited to the domain of science, including high-level support for running complex, integrated modelling in the cloud.
- **Challenge 3**: To address ***complexity*** more fundamentally and explicitly and, in particular, seek data science techniques that recognise and resolve key issues of complex systems including feedback loops, inter-dependent variables, extremes and also to detect and manage emergent behaviour.
- **Challenge 4**: To provide techniques and frameworks to both reify ***uncertainty*** in scientific studies and also reason about the cascading uncertainties across complex experiments, e.g. in integrated modelling frameworks and in ensemble approaches.
- **Challenge 5**: To seek ***adaptive*** techniques driven by considerations of uncertainty and also the goals of a scientific study, including adaptive approaches to sampling or gathering of data (including in real-time in an Internet of Things), and also in adaptive modelling approaches.
- **Challenge 6**: To seek approaches that deal with ***epistemic uncertainty*** in environmental modelling, noting the important links with dealing with emergent behavior in complex and irreducible phenomena.
- **Challenge 7**: To seek ***novel data science techniques*** and, in particular, innovative ***combinations of data science techniques*** that can make sense of the increasing complexity, variety and veracity of underlying environmental data, exploiting also multiple data sets including real-time streaming data.
- **Challenge 8**: To seek innovations in modelling by ***combining process models with data-driven or stochastic modelling techniques*** and also seeking ways of assimilating a range of data sources more generally into steering model executions.
- **Challenge 9**: To incorporate sophisticated ***spatial and temporal reasoning***, including reasoning across scales, as an integral aspect of environmental data science and not something that is just provided through separate tools such as GIS tools.
- **Challenge 10**: To discover new fundamentally ***new modes of working, methods and means of organisation*** that enable the required level of cross-disciplinary collaboration as required to address the key grand challenges of earth and environmental sciences and, more specifically environmental data science in its contribution to these grand challenges.

Part 2

Digital Technology and Living with Environmental Change

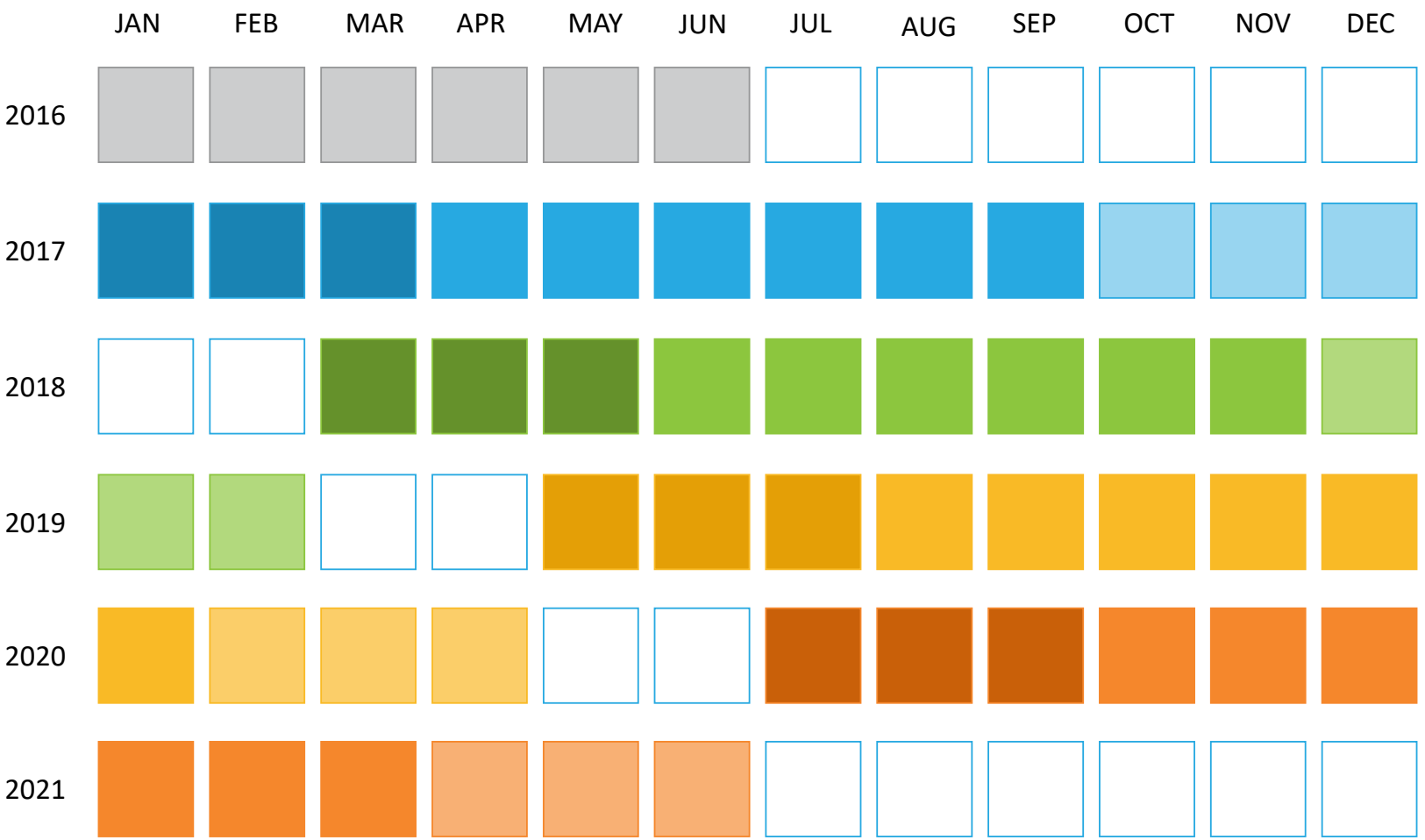


Overall aims of the fellowship

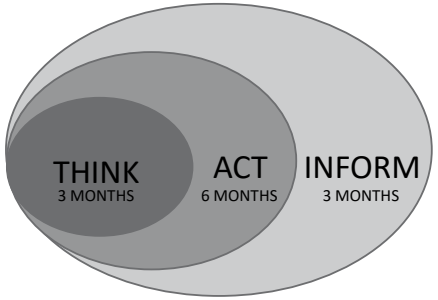
Working together for digitally inspired integrated environmental science

- Digital technology to explore grand challenges in environmental science
- Cross cutting themes:
  - Managing complexity & uncertainty
  - Raising abstraction
  - Developing a software architecture for deploying cloud technologies

ENSEMBLE FIVE YEAR PLAN



SPRINT CYCLE:



COLOUR CODING:

DISSEMINATION

FLOOD

BIO-DIVERSITY

SOIL

ECO-SYSTEM SERVICES

Principal Investigator and Co-Investigators


Prof. Gordon Blair


Dr Marie Ferrario


John Watkins


Dr Amber Leeson


Prof. Jon Tawn

Academics


Prof. Keith Beven


Prof. Jon Whittle


Dr Paul Young

Researchers


Richard Bassett


Graham Dean


Liz Edwards

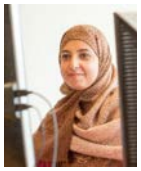

Pete Henrys


Dr Susan Jarvis


Louise Mullagh


Dr Vatsala Nundloll


Jordan Phillipson


Faiza Samreen


Dr Will Simm


Dr Ross Towe

Affiliates


Dr Bran Knowles


Dr Jess Davies


Dr Victoria Janes Bassett

Engagement


Mike Berners-Lee


Claire Dean


Harriet Fraser


Rob Fraser

Support


Carol Airey


Esther Carrington

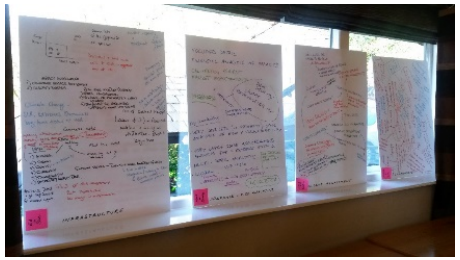

Pete Lloyd

ensemble

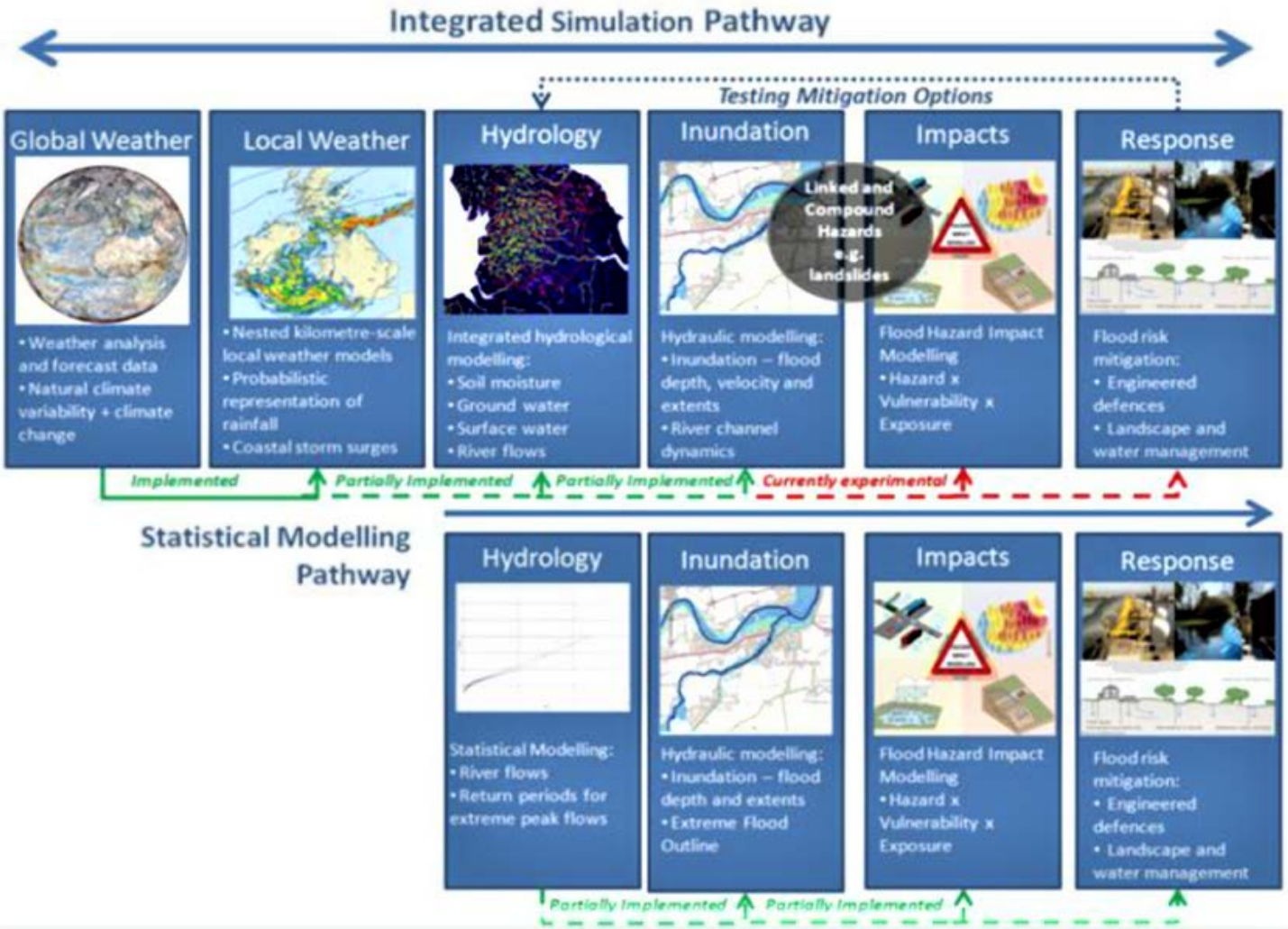
Flood sprint: initial workshop

Key outcomes:

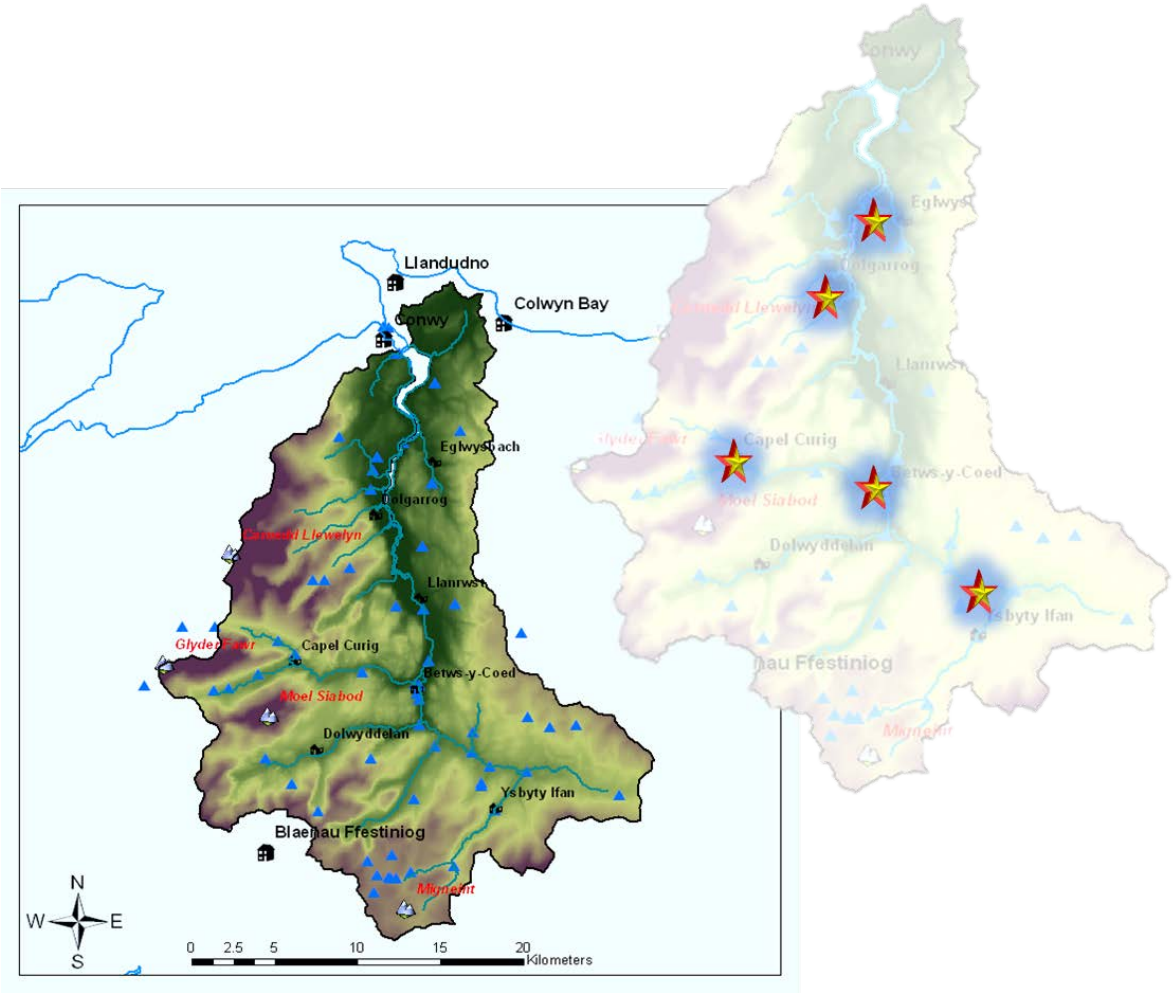
- Identification of four storyboards to drive research
  - The case for place: models of everywhere
  - A data-centric view
  - Towards more agile infrastructure
  - Let's work together
- "We are on the brink of something new and exciting building on: i) a move towards a more open approach to flood risk assessment, and ii) a shift in balance towards a more data-centric perspective to complement the existing process model-centric approach."



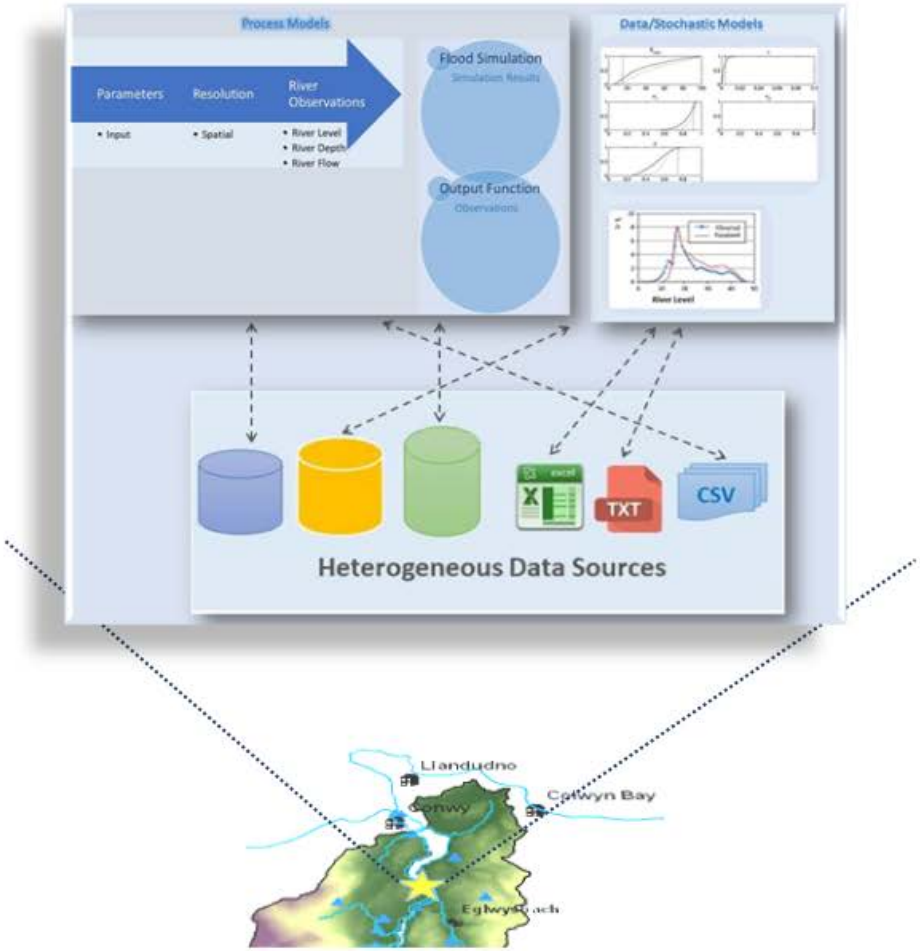
Context: National Flood Resilience Review



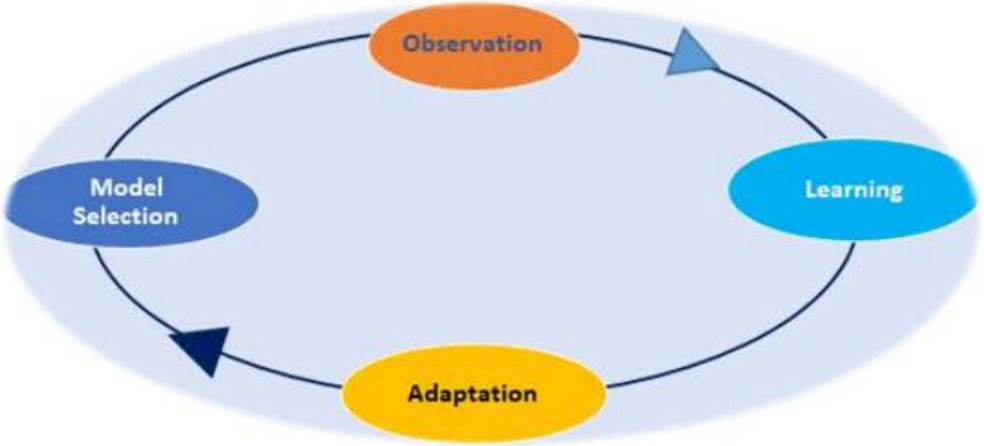
Models of everywhere revisited

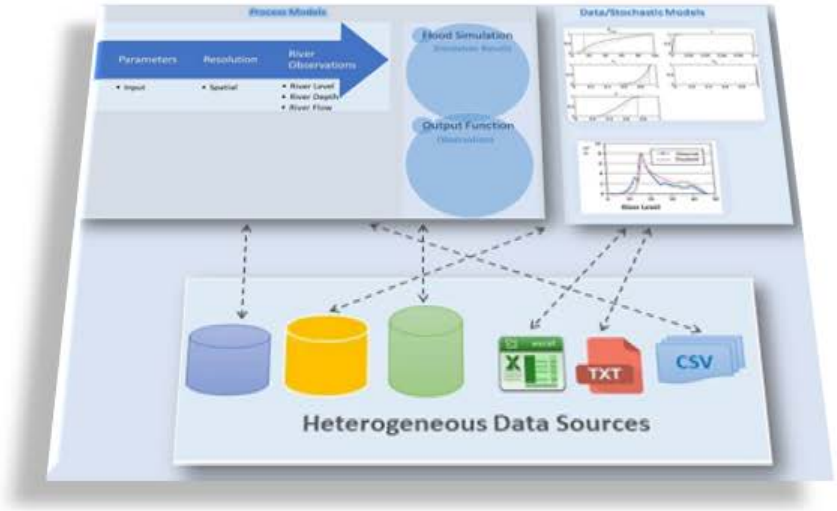


Models of everywhere revisited



Models of everywhere revisited

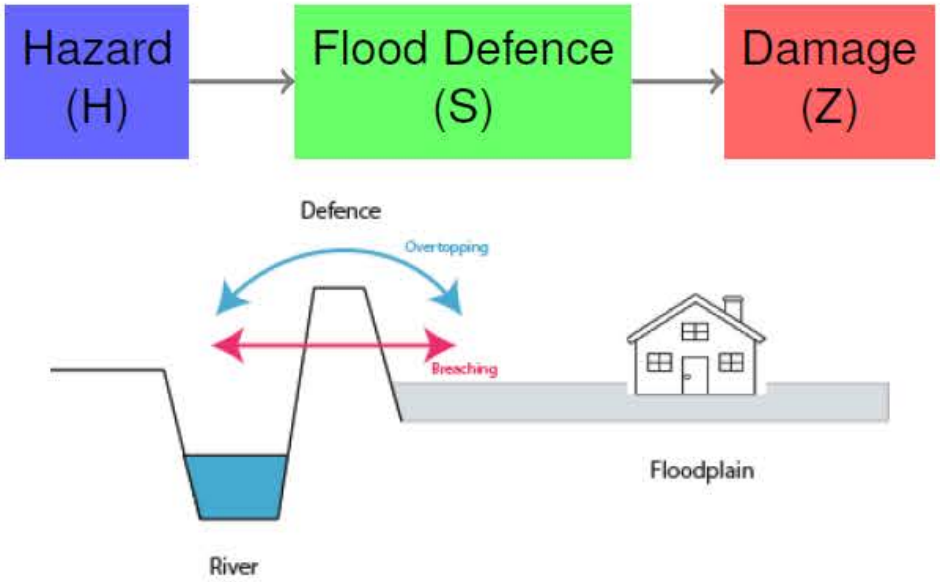




Focus: communities at risk project

- Data set provided by JBA Consulting
- Aim to improve flood risk information for a range of drivers
- Covers Derbyshire, Nottinghamshire and Leicestershire
- Access to the Property Impact Estimator Spreadsheet
- Links river flow gauge to each individual property at risk of flooding within Flood Zone 2

Communities at Risk Project (contd)

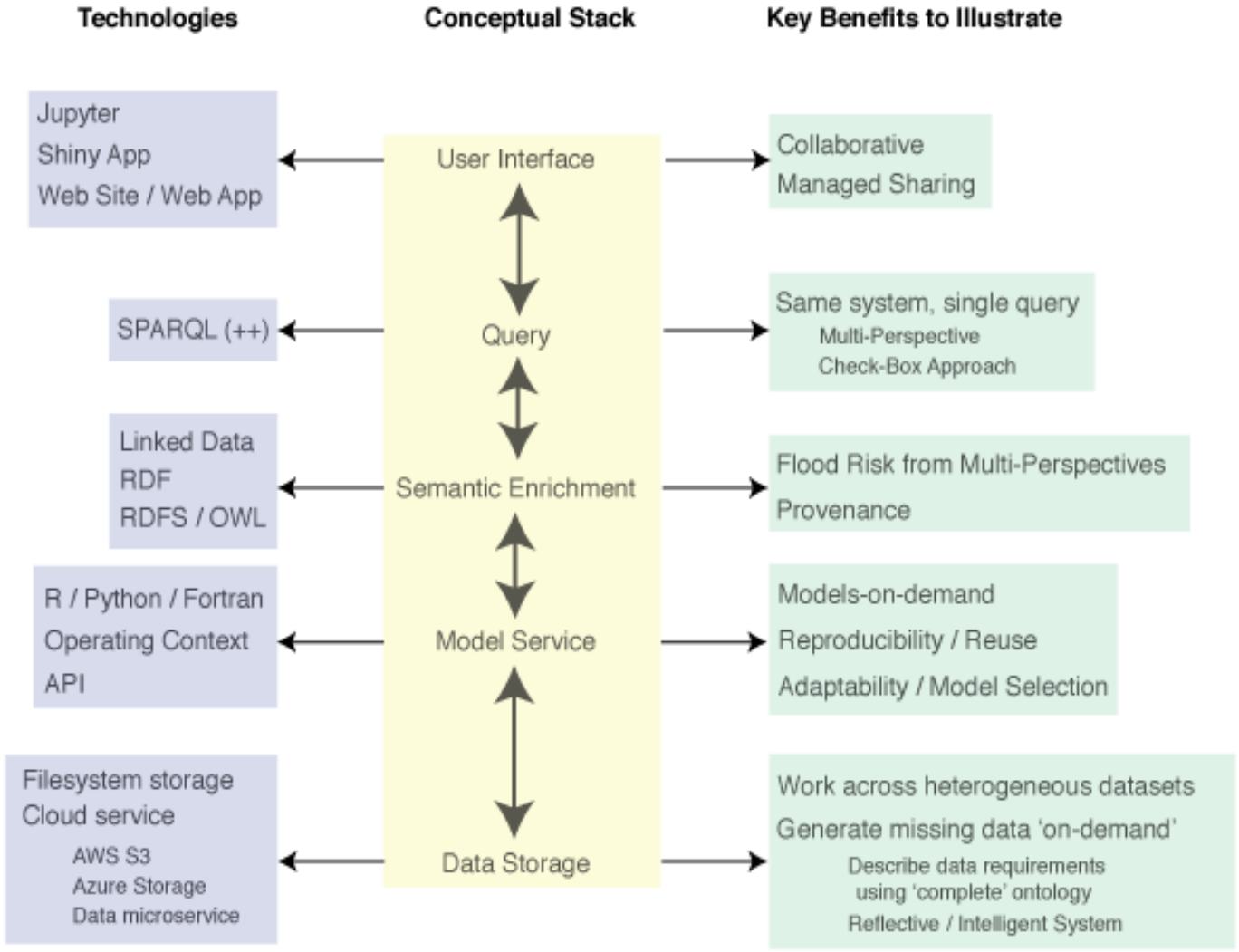


$$\mu_Z = \int_H \square f(H)dH$$
$$= \int_H \int_S \square P(S \mid H)dS \; f(H)dH$$
$$= \int_H \int_S Z(S,H)P(S \mid H)dS \; f(H)dH$$

Due to computational complexity

$$\mu_Z = \sum_H \sum_S Z(S,H)P(S \mid H)f(H)$$

Story-Led Demonstrator
(Used to illustrate and evaluate key research ideas)



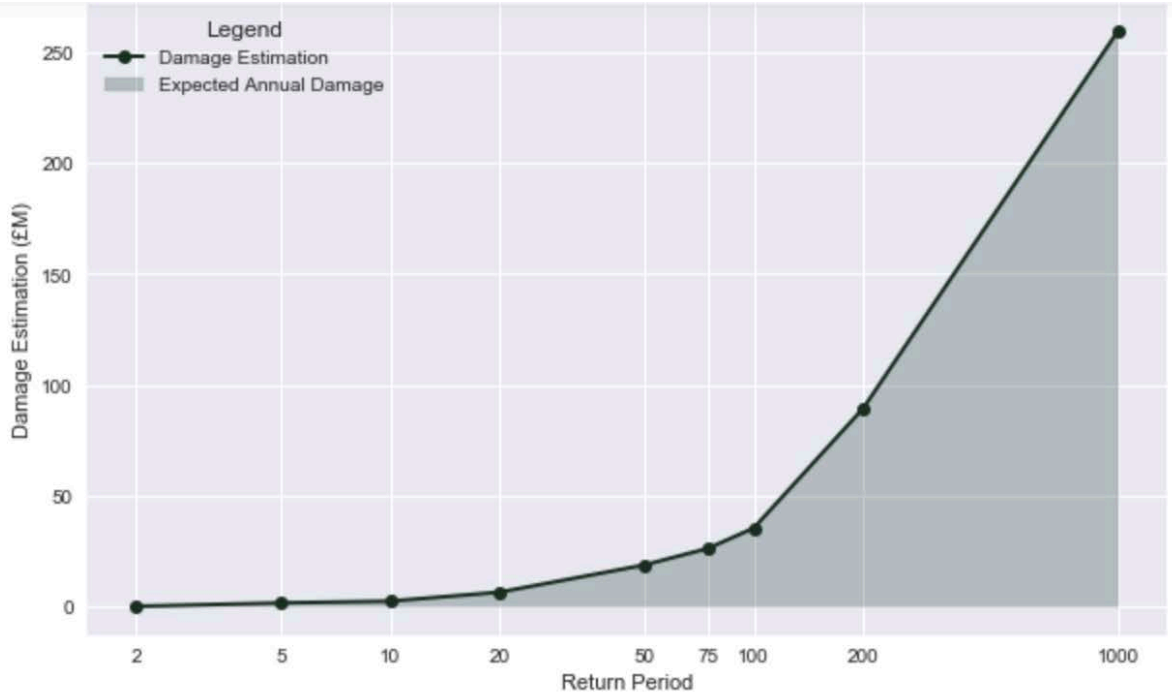
Evaluation: enhanced risk analysis

"As a flood risk manager, I want to understand the flood risk of Newark and the associated uncertainties. This flood risk assessment should combine information from all available data sources. Any inconsistencies in the modelling assumption should be made visible as well potential intervention strategies to reduce and mitigate flood risk. "

Evaluation: exploring the data space

"As a data scientist working with DSI Lancaster, I want to explore and understand all available data sources that I have to inform decisions about the flood risk of Newark. This includes model outputs, and an increasingly wide range of data sources including, for example, Section 19 reports and data extracted from social media. I also want to access meta-data associated with all this data, including provenance information. I am particularly interested in using this data to discover potentially new levels of uncertainty in the data."

Example output



EXPECTED ANNUAL DAMAGE

	Onset of Flooding Assumption		
	Optimistic	Midpoint	Pessimistic
Current Defence Scheme:	£2,990,942.12	£3,312,795.07	£3,751,284.98
50 Year Defence Scheme:	£1,892,017.33	£2,022,983.93	£2,203,772.42
100 Year Defence Scheme:	£1,605,100.95	£1,721,811.84	£1,887,153.00
200 Year Defence Scheme:	£1,332,623.57	£1,409,444.48	£1,497,270.56
500 Year Defence Scheme:	£737,728.62	£762,208.60	£767,279.36

School of Computing & Communications | Lancaster University

Thank you for listening!

… any questions?

- Hydraulic model has been run for a finite set of return periods, e.g. 2,5,10,20,50,… years (x-axis)
- For this property flooding occurs somewhere between 20 and 50 year return period. This can be seen because the estimate of damage change to being non-zero at the 20 year return period point
- What we don't know is the exact change in behaviour between the 20 and 50 year return period i.e. the exact return period that causes flooding.
- From this graph we know it lies somewhere between the 20 and 50 year return period. We can make some assumptions about where this point lies in order to calculate the expected annual damage estimate. These assumptions are framed in terms of the property owner (colour coded as in the figure):
  - Optimistic - flooding occurs as close to the 50 year return period as possible
  - Pessimistic - flooding occurs just after the 20 year return period as possible
  - Midpoint – halfway in between the 20 and 50 year return period; the 35 year return period
- This changing assumptions can have a huge impact on the expected annual damage therefore there is huge uncertainty around the risk analysis and the resulting decisions that are made.