# Data Analytics for Sustainability and Environmental Risk (*DASER*)

*Coordinated by the Centre for the Study of Existential Risk (CSER, http://cser.org/)*
*University of Cambridge, UK*

# DASER-2016 Workshop Report

*12th October 2016 – David Attenborough Building, Cambridge, UK*
*Workshop organised by Scott Hosking (British Antarctic Survey) – Sponsored by CSER*

## 1. Attendees

Andrew Meijers - *British Antarctic Survey*
Anita Faul - *Centre for Scientific Computing, University of Cambridge*
Charlie Kennel - *Centre for Science and Policy, University of Cambridge*
David Simmons - *Willis Tower Watson*
Emily Shuckburgh - *British Antarctic Survey*
Herbert Lau - *Schlumberger*
Huw Price - *Centre for the Study of Existential Risk, University of Cambridge*
Ian Leslie - *Computer Laboratory, University of Cambridge*
Joel Gustafsson - *Max Fordham*
Kristen MacAskill - *Department of Engineering, University of Cambridge*
Paul Griffiths - *Department of Chemistry, University of Cambridge*
Paul Linden - *Department of Applied Mathematics and Theoretical Physics, U. Cambridge*
Peter Haynes - *Department of Applied Mathematics and Theoretical Physics, U. Cambridge*
Pierre-Philippe Mathieu - *European Space Agency*
Rob Doubleday - *Centre for Science and Policy, University of Cambridge*
Richard Turner - *Department of Engineering, University of Cambridge*
Scott Hosking - *British Antarctic Survey*
Seán Ó hÉigeartaigh - *Centre for the Study of Existential Risk, University of Cambridge*
Shahar Avin - *Centre for the Study of Existential Risk University of Cambridge*
Stephen Briggs - *European Space Agency*
Tatsuya Amano - *Centre for the Study of Existential Risk, University of Cambridge*
Tom Walters - *Google DeepMind*
Tony Phillips - *British Antarctic Survey*

## 2. Agenda

1. Emily Shuckburgh "*Introductions and objectives*"
2. Charles Kennel "*Making Climate Science More Useful*"
3. Andrew Meijers "*Reporting back from the 2016 6th International Climate Informatics workshop*"
4. Scott Hosking "*Overview of environmental datasets*"
5. *Discussion*

# 3. Background

**Challenge**: There exists a significant information gap between the sources of climate data and the kinds of local, actionable knowledge required for mitigation and adaptation policy and planning decisions. The causes of this gap are discontinuous and heterogeneous data sources, a plethora of loosely connected models of varying complexity, and limitations of expertise transfer across domains of climate science, environmental science, modelling and big data, policy, legal and financial decision making, and planning and civil engineering. The ultimate consequence is that it is difficult to make meaningful decisions and plans on the local and regional level in response to climate change.

**Solution**: Explore the use of big data and machine learning to harmonise the data sources and connect models and data at different scales. An example of what could be achieved is a system that continuously integrates diverse streams of live climate and environmental data and produces a set of key planetary vital signs. Another example is an interactive knowledge base for decision-makers that takes into account local and regional exposure to various climate-related risks. The systems would ideally have a fully transparent pathway from data to knowledge that can be investigated, challenged and improved upon by domain experts, showcasing the data sources, the methods underlying them and their reliability levels, the global/regional/local climate and environmental models and datasets and associated confidence/uncertainty, and the regional/local risk exposure assessments using decision-relevant metrics.

**Opportunity**: The challenge outlined cannot be tackled by climate scientists or environmental scientists alone. New capabilities in big data analytics, sensor networks and Earth observation have resulted in an opportunity to develop a much more sophisticated set of indicators of climate change and resulting impact. Cambridge seems well-placed to draw on expertise from these various disciplines.

If successful, having traceable connections from the processes affecting planetary vital signs through to the impacts relevant to society (e.g. food and water security, flooding, species loss), business (e.g. correlated extreme events, threshold exceedance) and decision-makers at all levels (e.g. risk of abrupt or irreversible change) would be valuable evidence-base for developing local, regional and global policy and planning.

# 4. Current research and collaborations at Cambridge

The University of Cambridge is home to a large range of relevant expertise across departments and interdisciplinary centres. The university also has active links to multiple stakeholders in government and in the private sector. These features appear to be essential for building a strong team to assess the impact of future climate change on sustainability and environmental risk, and building resilience using a range of methodologies such as Data Analytics and Machine Learning. Below is a partial list of relevant departments and centres in Cambridge, and external parties who have an active collaboration with researchers at the university.

| Department/Company | Area of expertise | Contacts |
|---|---|---|
| British Antarctic Survey | Climate science and modelling | Emily Shuckburgh, Scott Hosking, Andrew Meijers |
| Department of Chemistry | Air Quality, atmospheric air composition modelling | Alex Archibald, Paul Griffiths |
| Department of Applied Mathematics and Theoretical Physics | Climate science | Peter Haynes, Paul Linden |
| Centre for the Study of Existential Risk (CSER) | Interdisciplinary research on extreme risks | Seán Ó hÉigeartaigh, Shahar Avin, Tatsuya Amano |
| Computer Laboratory | Large scale computer systems, smart buildings | Ian Leslie |
| Machine Learning Group, Dept. of Engineering | Machine learning | Richard Turner |
| Sustainable Development Group, Dept. of Engineering | Sustainable development | Peter Guthrie |
| Centre for Science and Policy (CSaP) | Effective exchange of information between academia and government | Rob Doubleday |
| Cambridge Centre for Climate Science (CCfCS) | Climate science | Peter Haynes |
| Cambridge Institute for Sustainability Leadership (CISL) | Sustainability in the private sector | Jake Reynolds, Eliot Whittington |
| Centre for Risk Studies (CRS) | Risk modelling | Daniel Ralph |
| Cambridge Conservation Initiative (CCI) | Conservation policy, ecological modelling | Bhaskar Vira, Bill Sutherland |
| Cambridge Global Food Security | Food security modelling | Chris Gilligan |
| Leverhulme Centre for the Future of Intelligence (CFI) | Artificial intelligence, machine learning | Adrian Weller, Stephen Cave |
| Google DeepMind (*invited*) | Machine learning | Tom Walters |
| Willis Towers Watson (*invited*) | Insurance | David Simmons |
| Willis Group (*invited*) | Re-insurance | Rowan Douglas |
| Global Climate Observing System (GCOS) (*invited*) | Climate data | Stephen Briggs |

|  |  |  |
|---|---|---|
| Scripps Institute, UCSD (*invited*) | Climate science | Charlie Kennel |
| European Space Agency (ESA) (*invited*) | Climate data | Pierre-Philippe Mathieu |
| Max Fordham (*invited*) | Sustainability engineering | Joel Gustafsson |

# 5. Environmental Datasets

**Climate modelling projects**

A range of 'Big Data' climate model simulation datasets are available that lend themselves to environmental risk assessment. To assess extreme/rare environmental events we require large multi-model and multi-scenario ensembles of the order of 10,000s of years to produce robust statistics. These help answer questions such as: *To what extent are recent extreme events (e.g., heat waves, floods and droughts) attributable to natural variability or human activity? How will the risk of such high impact climate and weather events change over the next few decades?*
A few examples of modern climate modelling datasets include:

- CMIP6**:** https://www.wcrp-climate.org/wgcm-cmip/wgcm-cmip6
  *Available from ~2018, will become the primary dataset used as the basis for the IPCC Sixth Assessment Report (AR6) on Climate Change*
- HAPPI**:** http://www.happimip.org/
  *Designed to assess climate between a '1.5°C world' and a '2°C world' as the basis for IPCC Report following Paris COP (release 2018)*
- PRIMAVERA**:** https://www.primavera-h2020.eu
  *Organised around research themes including: Innovations in Modelling, Drivers of European Climate, and Climate Risk Assessment*

**Weather station:** https://www.ncdc.noaa.gov/oa/climate/ghcn-daily/

Weather station records from around the world are collated by the Global Historical Climatology Network, an integrated database of daily climate summaries from numerous sources (e.g., land surface stations) that have been integrated and subjected to a common suite of quality assurance reviews. GHCN-Daily contains records from over 75000 stations in 180 countries and territories with some extending to more than 175 years. Numerous daily variables are provided, including maximum and minimum temperature, total daily precipitation, snowfall, and snow depth; however, about two thirds of the stations report precipitation only.

**Global Atmospheric Reanalysis Products**

Global atmospheric reanalyses are meteorological data assimilation projects which aims to assimilate historical observational data (including balloon/sonde measurements and satellite products) usually spanning the period 1979-present (*the satellite era*), using a single consistent assimilation scheme throughout. This results in gridded datasets which are spatiotemporally representative of the real-word. Three examples of modern reanalysis products include:

- ERA-Interim: http://www.ecmwf.int/en/research/climate-reanalysis/era-interim
- MERRA-2: https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/
- JRA-55: http://jra.kishou.go.jp/JRA-55/index_en.html

# 6. Machine Learning

The DASER-2016 workshop included a presentation focused on currently used machine learning tools which may lend themselves to DASER-type research activities.

**“Reporting back from the 2016 6th International Climate Informatics workshop”**
*Andrew Meijers (Oceanographer, British Antarctic Survey) attended a workshop and hackathon hosted by NCAR ([https://www2.cisl.ucar.edu/events/workshops/ci2016](https://www2.cisl.ucar.edu/events/workshops/ci2016)). He provided an overview of the range of ongoing activities in use by the climate groups (mainly in the US) who attended. These included:*

- Regional climate downscaling and structured regression, including the use of Learn-alpha to identify the predictors of regional temperatures.
- Spatial infilling and statistical downscaling using LatticeKrig
- Extreme event prediction using quantile regression to calibrate and downscale global climate model output learning from distributions found within regional observations.
- Multi-model means & projections - Weighting of ‘expert’ model simulations which changes with time using a Generalised Hidden Markov Model. Bayesian updating, previous information incorporated into time evolving predictions.
- Paleoclimate reconstruction - identify relationships between spatial map of short instrument record and very sparse but long climate-proxy (e.g., ice core records) network. The major problem here is data fusion of many small datasets.
- Climate prediction challenges:
    - Often not ‘big data’ in the sense machine learning engineers are used to working with. Modelling datasets are very high dimensional and complex. Large modelling simulation projects usually create between 10s-100s of Tb of data. Data manipulation and algorithm scalability is a challenge.
    - Data is generally unlabeled, thus supervised learning is hard. Limited training and validation data (only ~50-100 years, sparsely observed).
    - Signals of change are mostly non-stationary, i.e., changes observed in the past may differ to those in the future.
    - Spatial inhomogeneity, with frequently non-Gaussian distributions (e.g. precipitation).
    - For risk management it is crucial to accurately predict outliers and extremes.
    - ‘Black box’ machine learning doesn’t always support physical understanding.

# 7. Use Cases

During the afternoon, we discussed use cases/exemplars where progress could be made with the skills already available within the Departments across Cambridge. One of these was the prediction of Arctic sea ice over seasonal timescales, by incorporating climate fields, e.g, the jet stream and global sea surface temperature patterns, as features within machine learning regression algorithms. Another use case discussed was reducing uncertainties in regional droughts in Africa on decadal timescales using calibration and downscaling climate model projections.