Opportunities and Challenges for Data Science in (Big) Environmental Science

Et. Al. &
Bryan Lawrence

NERC SCIENCE OF THE ENVIRONMENT

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

Outline

1. (Why am I here?)
2. Characteristics of Environmental Science
3. Challenges - 1
4. Computing Environments
5. Opportunties and Examples
6. Challenges - 2
7. Summary

NCAS and Computer Science

NCAS

NCAS delivers national capability science and infrastructure

- Climate science, including climate change
- Atmospheric composition, including air pollution
- High Impact Weather, including processes.
- Facilities: Aircraft, Instruments, *Models, Data Centres (CEDA), HPC* etc

NCAS and Computer Science

NCAS

NCAS delivers national capability science and infrastructure

- Climate science, including climate change
- Atmospheric composition, including air pollution
- High Impact Weather, including processes.
- Facilities: Aircraft, Instruments, *Models, Data Centres (CEDA), HPC* etc

UoR: Computer Science

- A new department (2 years old) born from the ashes of a restructuring.
- Growing (hiring)!
- Embedded in existing school alongside mathematics and meteorology.
- Research groups include “Data Analytics”, “Data Science and AI” and “*Advanced Computing for Environmental Sciences*”.

Definitions

Data Science (Wikipedia)

Also known as data-driven science, is an interdisciplinary field of scientific *methods, processes, algorithms and systems* to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining.

Definitions

Data Science (Wikipedia)

Also known as data-driven science, is an interdisciplinary field of scientific *methods, processes, algorithms and systems* to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining.

What is big data and data science? (Google)

Dealing with *unstructured and structured* data, Data Science is a field that encompasses anything related to *data cleansing, preparation, and analysis*. Put simply, Data Science is an umbrella term for techniques used when trying to extract insights and information from data.

Definitions

Data Science (Wikipedia)

Also known as data-driven science, is an interdisciplinary field of scientific *methods, processes, algorithms and systems* to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining.

What is big data and data science? (Google)

Dealing with *unstructured and structured* data, Data Science is a field that encompasses anything related to *data cleansing, preparation, and analysis*. Put simply, Data Science is an umbrella term for techniques used when trying to extract insights and information from data.

What is big data and analytics? (Google)

Big data analytics is the process of examining large and varied data sets – i.e., big data – to uncover hidden patterns, unknown correlations, *market trends, customer preferences* and other useful information that can help *organizations* make more-informed *business* decisions.

Definitions

Data Science (Wikipedia)

Also known as data-driven science, is an interdisciplinary field of scientific *methods, processes, algorithms and systems* to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining.

What is big data and data science? (Google)

Dealing with *unstructured and structured* data, Data Science is a field that encompasses anything related to *data cleansing, preparation, and analysis*. Put simply, Data Science is an umbrella term for techniques used when trying to extract insights and information from data.

What is big data and analytics? (Google)

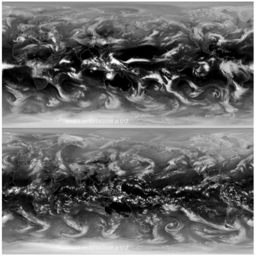
Big data analytics is the process of examining large and varied data sets – i.e., big data – to uncover hidden patterns, unknown correlations, *market trends, customer preferences* and other useful information that can help *organizations* make more-informed *business* decisions.

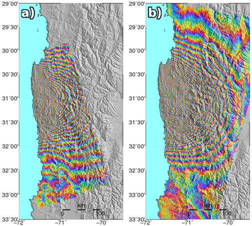
Environmental science has been doing *Big Data Science* since before I was a student.

What is Environmental Data? Diverse

NERC Data Catalogue, 21st of March, 2018: 5445 datasets:

What is Environmental Data?: Multiscale

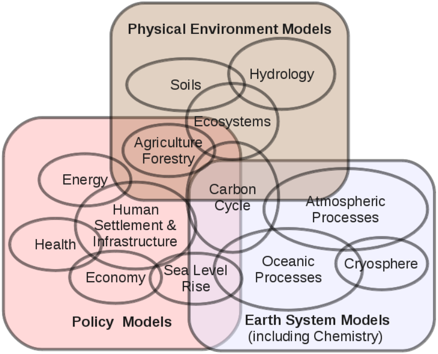




(Examples from JASMIN users:

- UPSCALE (courtesy of P.L. Vidale)
- COMET-LICS (http://comet.nerc.ac.uk/developing-licsar-automated-processing-sentinel-1-data/)
- CEH Wildlife Survey (Courtesy of Tom August).)

What is Environmental Science? Multidisciplinary!

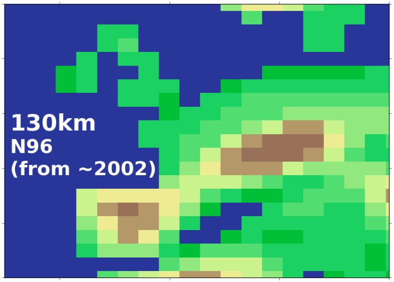


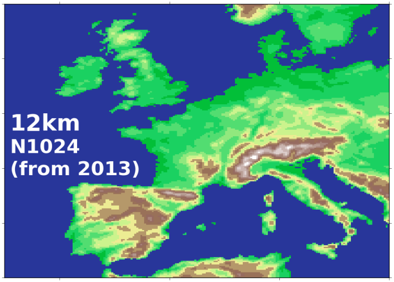
Many interacting communities, each with their own software, data (standards), compute environments etc.

Figure adapted from Moss et al, 2010

What is Environmental Data? Voluminous!

Europe within a global model …





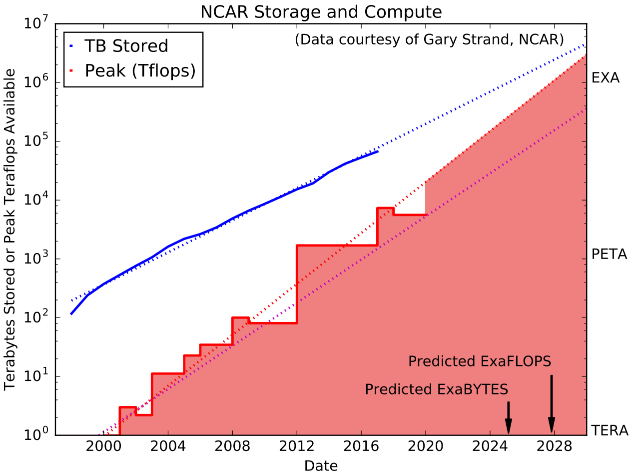
One "field-year" — 26 GB

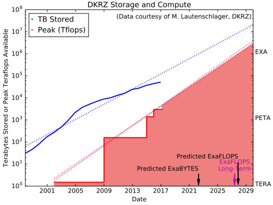
1 field, 1 year, 6 hourly, 80 levels
1 x 1440 x 80 x 148 x 192

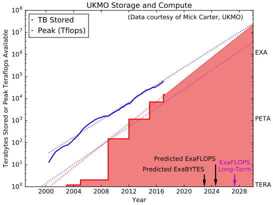
One "field-year" — >6 TB

1 field, 1 year, 6 hourly, 180 levels
1 x 1440 x 180 x 1536 x 2048

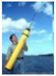
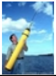
What is Environmental Data? Voluminous!







What is Environmental Data?: Sometimes clean, mostly messy!

PointSeriesFeature (*timeseries at a point*)	 
ProfileFeature (*vertical profile at a point*)	  
GridSeriesFeature (*series of multidimensional grids*)	 
SwathFeature (*single satellite sweep*)	
SectionFeature (*vertical section*)	 

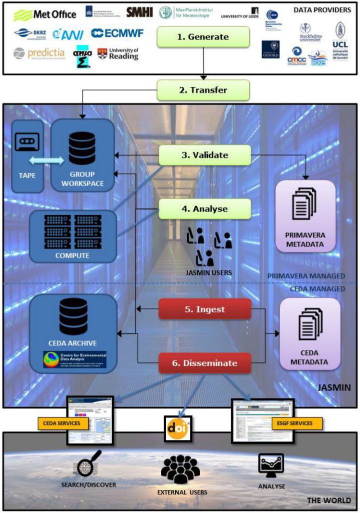
Classify by geometry, but that doesn't tell you how it stored, or what it is.

What is Environmental Data?: Sometimes clean, mostly messy!

Formats and Content Standards

- Disparate communities, disparate formats.
- Converging towards NetCDF (at least outside of the Met Agencies).
- (If your tool doesn't understand NetCDF, you wont be in business with much of environmental data.)
- But a format is just a bucket - can still label parameters in multiple ways, and there may be no text to get context ... if you can't understand the label, the data is useless.
- Massive importance of content standards (Climate Forecast Conventions, CMIP standards etc).

What is Environmental Data?: Sometimes clean, mostly messy!



jon.seddon@metoffice.gov.uk

PRIMAVERA and CMIP

Model intercomparison projects develop sophisticated standards and workflows:

- Simulations are designed to produce output in a common format with common metadata standards.
- ...but it still necessary to validate the output against those standards before publication into an archival and dissemination system.
- This is the *minimum* necessary to provide data into sophisticated data analysis pipelines!

Environmental Science Challenges

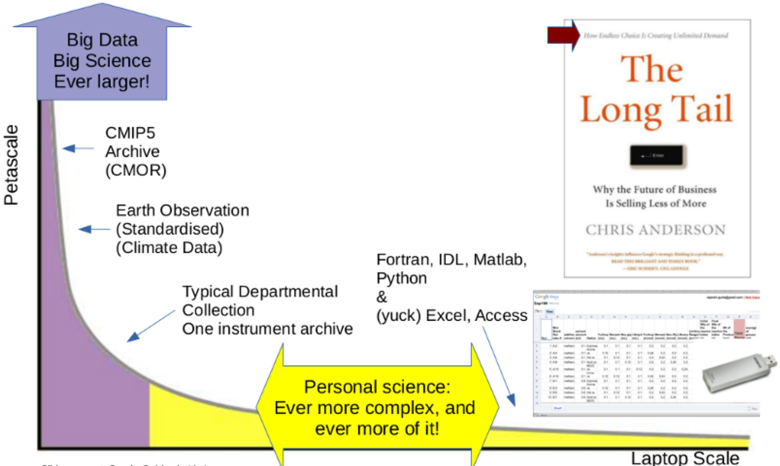
1. Diverse
2. Multiscale
3. Multidisciplinary
4. Voluminous
5. Often Messy,
6. Often NOT TEXT and/or
7. NOT simple columns of numbers

Environmental Science Challenges

1. Diverse
2. Multiscale
3. Multidisciplinary
4. Voluminous
5. Often Messy,
6. Often NOT TEXT and/or
7. NOT simple columns of numbers

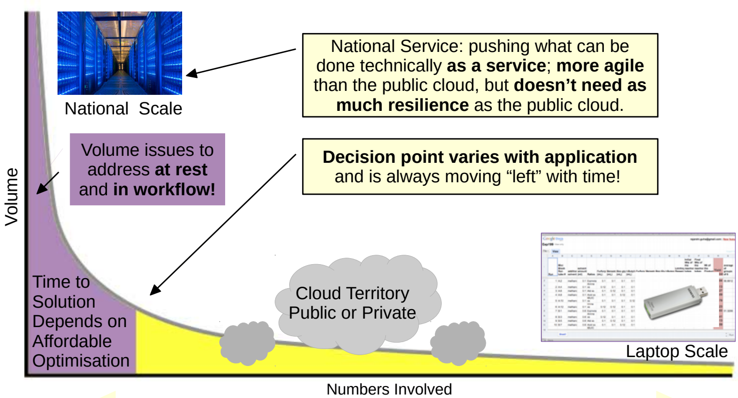
…but of course all of these are opportunities as well.

Wide Scope



Slide concept: Carole Goble via Liz Lyons

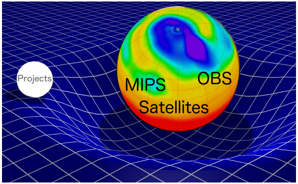
Wide Scope



Optimise Customise Utilise Acquire

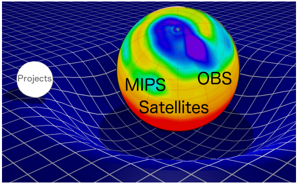
Optimise	Customise	Utilise	Acquire
Build A System (e.g. JASMIN, or even design new chips, e.g. Google's TPU)	Customise an Environment (e.g. deploy a Hadoop Cluster)	Use a resource (provided by someone else, a service)	Buy something (e.g. a Laptop)

JASMIN — The Data Commons

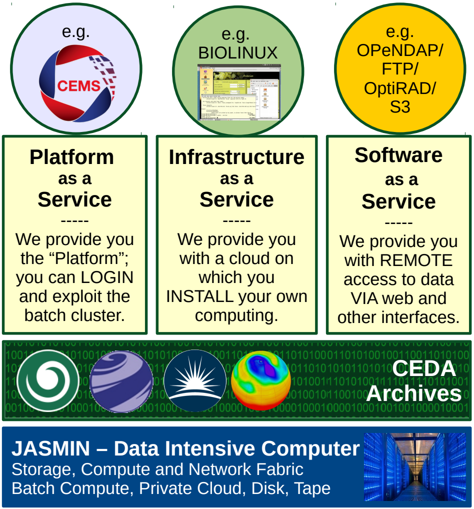


- Provide a state-of-the art storage and computational environment
- Provide and populate a managed data environment with key datasets (the "archive").
- Encourage and facilitate the bringing of data and/or computation alongside/to the archive!
- Provide FLEXIBLE methods of exploiting the computational environment.

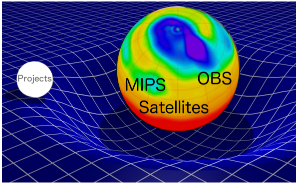
JASMIN — The Data Commons



- Provide a state-of-the art storage and computational environment
- Provide and populate a managed data environment with key datasets (the "archive").
- Encourage and facilitate the bringing of data and/or computation alongside/to the archive!
- Provide FLEXIBLE methods of exploiting the computational environment.



JASMIN — The Data Commons



- Provide a state-of-the art storage and computational environment
- Provide and populate a managed data environment with key datasets (the "archive").
- Encourage and facilitate the bringing of data and/or computation alongside/to the archive!
- Provide FLEXIBLE methods of exploiting the computational environment.

e.g.

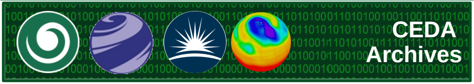
e.g. BIOLINUX

e.g. OPeNDAP/

Coming in 2018

JASMIN Phase 4/Phase 5

- Cluster-as-a-Service: spin up pre-canned SLURM, SPARK, and DASK clusters.
- Looking for guinea pigs.



JASMIN – Data Intensive Computer
Storage, Compute and Network Fabric
Batch Compute, Private Cloud, Disk, Tape

Common Software/Algorithm Patterns

Supporting a wide variety of algorithms and workflows: (but much to do to exploit parallelism)

"Big Data Ogres" by analogy with the Berkely Dwarves for computational patterns.

Different Problem Architectures, e.g:

1. Pleasingly Parallel (e.g. retrievals over images)
2. Filtered pleasingly parallel (e.g. cyclone tracking)
3. Fusion (e.g. data assimilation)
4. (Space-)Time Series Analysis (FFT/MEM etc)
5. Machine Learning (clustering, EOFs etc)

Important Data Sources, e.g:

1. Table driven (eg. RDBMS + SQL)
2. Document driven (e.g XMLDB + XQUERY)
3. Image driven (e.g. GeoTIFF + your code)
4. (Binary) File driven (e.g. NetCDF + your code)

Sub-Ogres: Kernels & Applications, e.g:

1. Simple Stencils (Averaging, Finite Differencing etc)
2. 4D-Variational Assimilation/ Kalman Filters
3. Data Mining Algorithms (classification/clustering) etc
4. Neural Networks

Modified from Jha et al 2014 arXiv:1403.1528[cs]

Uncommon (and inappropriate?) software solutions

Multiple tools

Contrast between two very types of workflow:

- Build Once: Many analysis tasks are build once, use once, throwaway. No room for optimisation (or MPI). *Need efficient libraries.*
- Repeatable: "build", "run", "move", "reduce/reformat", "analyse". *Much room for automation.*.

What to use? Plethora of architectures and tools out there

APACHE Spark DASK Pandas

Uncommon (and inappropriate?) software solutions

Multiple tools

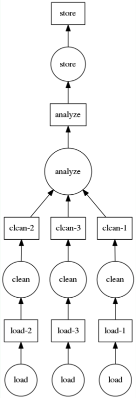
Contrast between two very types of workflow:

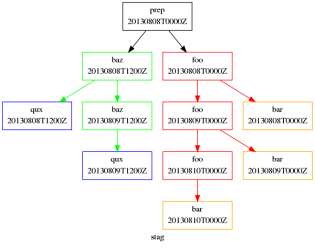
- Build Once: Many analysis tasks are build once, use once, throwaway. No room for optimisation (or MPI). *Need efficient libraries.*
- Repeatable: “build”, “run”, “move”, “reduce/reformat”, “analyse”. *Much room for automation.*.

What to use? Plethora of architectures and tools out there

Exploiting Concurrency

Whatever tools, need to get used to generating, understanding, and exploiting concurrency in more complicated ways:





Much to do to harness tools to accelerate workflows!

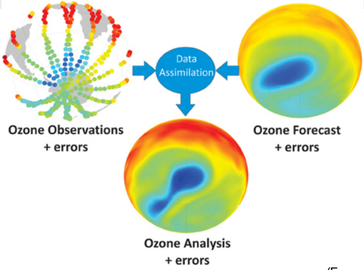
(These two examples: dask, and cylc, representing bespoke analysis and scheduling, reduction and proliferation.)

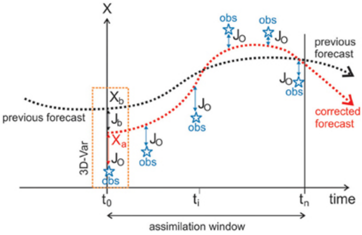
Application Opportunities

An eclectic set of applications:

1. Data Assimilation and Data Archaeology
2. Classification: from established practice to deep learning at scale.
3. Cleaning up earth observation data with machine learning.
4. Cleaning up the periphery of our models - whither machine learning in parameterisation?

Data Assimilation





(From Lahoz and Schneider2014)

Data Assimilation

DA is the process of using a model to interpolate (in space and time) between observations or to adjust a model trajectory towards observations. Always uses, and produces, error estimates. Typically used to

- Develop an *analysis* (or *re-analysis*) product, and/or
- To provide initial conditions for a model simulation.

Twentieth Century Reanalysis

Data Assimilation

Compo et al 2011. The Twentieth Century Reanalysis Project. DOI:10.1002/qj.776

- Delivers analyses of global tropospheric variability *and* of the quality of those analyses from 1871 to the present at 6-hourly temporal and 2 degress lat/long spatial resolution.
- Uses an Ensemble Kalman Filter (weighting 56 ensemble members and whatever observations were available (but not satellites).

Twentieth Century Reanalysis

Data Assimilation

Compo et al 2011. The Twentieth Century Reanalysis Project. DOI:10.1002/qj.776

- Delivers analyses of global tropospheric variability *and* of the quality of those analyses from 1871 to the present at 6-hourly temporal and 2 degress lat/long spatial resolution.
- Uses an Ensemble Kalman Filter (weighting 56 ensemble members and whatever observations were available (but not satellites).

Big and Expensive

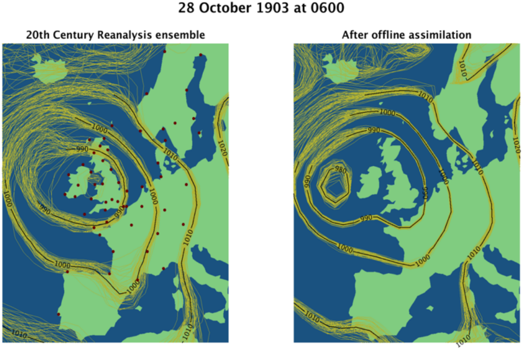
- Massive computing initiative.
- Heroic data iniative: 1.7 Billion Observations. 1 TB a year of output data.

Diverse Applications

- Early 20th Century Arctic warming
- Historical El Nino/Southern Oscillation events
- Decadal Atlantic hurricane variability
- Ocean ecology
- US Dust Bowl

Historical Observations: The benefits

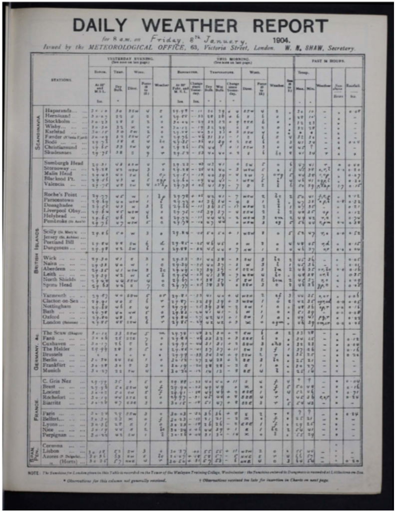
If we want to know about change, we need to know the baseline.



(Courtesy of Ed Hawkins, NCAS and UoR Meteorology.)

- An example of the potential benefit of combining old observations with retrospective data assimilation ("re-analysis").
- We get a much better understanding of historical weather!
- More understanding of extremes and tracks.

Depends on Data Archaeology


DAILY WEATHER REPORT
for 8 a.m. on Friday, 8th January, 1904.
Issued by the METEOROLOGICAL OFFICE, 63, Victoria Street, London. W. N. SHAW, Secretary.

Goal: Extract historical weather observations from paper records and exploit them in developing new re-analyses of past climate.

- Many thousands of historic records have been transcribed using volunteers (currently each record is transcribed by FIVE humans and compared).
- Low rate of progress; will take a decade just to do this particular dataset.

Opportunity: Large body of training data, and robust validation methodology.

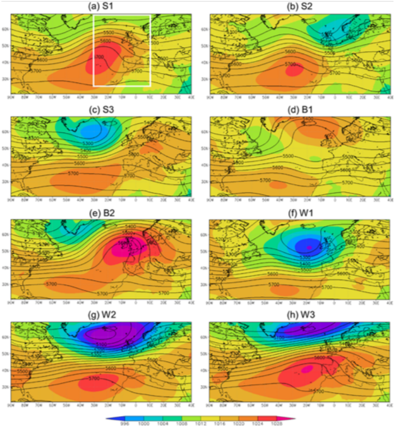
Classification: Lots of Prior Art

Cost733cat – A database of weather and circulation type classifications. Philipp et. al. (2010)

doi:10.1016/j.pce.2009.12.010

Catalogue of Types

- 23 methods, including 5 subjective and 18 automated methods with variants, totalling 72 classification schemes.
- Two main strategies: *Pre-defined types* (including subjective and threshold methods) and *Derived types* (including PCA, EOF, k-means etc, and combinations thereof).



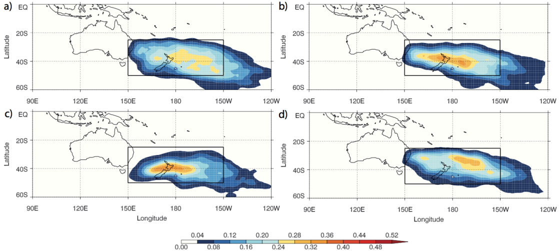
(Santos et al 2016, doi:10.1002/2015JD024399

Classification: Cyclones

Process Validation in Models. We want to understand how models do, or don't, simulate aspects of the different types of cylcones which occur - leads to confidence in predictions and projections.

K-Means Clustering

- Clustering of cyclone tracks - not images.
- Unsupervised, but need to select number of classes (can try variants).
- Validated by comparison with manual classification.



Track density for the four clusters identified, each has different impacts in terms of their precipitation (cluster 1 has the highest average precip), different seasonal cycles and genesis locations.

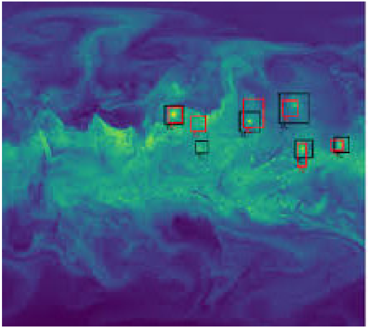
From J. Catto, 2018, doi:10.1175/JCLI-D-17-0746.1

Deep Learning at Scale

Deep Learning at 15PF: Supervised and Semi-Supervised Classification for Scientific Data

Kurth, Zhang, Satish, Mitliagkas, Racah, Patwary, Malas, Sundaram, Bhimji, Smorkalov, Deslippe, Shiryaev, Sridharank, *Prabhat*, Dubey

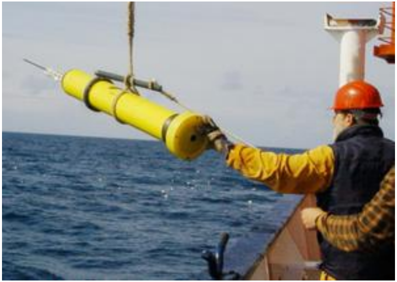
- Current Deep Learning implementations can take days to converge on O(10) GB datasets.
- Using a 15 TB climate dataset (768x768, 16 channels, 0.4M images)
- 9622 KNL nodes and sustained ≈ 12 PFLOP/s during classification
- Two HPC perspectives to consider for deep learning:
    1. How efficient is deep learning on a single node?
    2. How does it scale across a cluster of nodes?

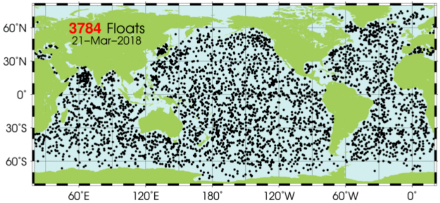


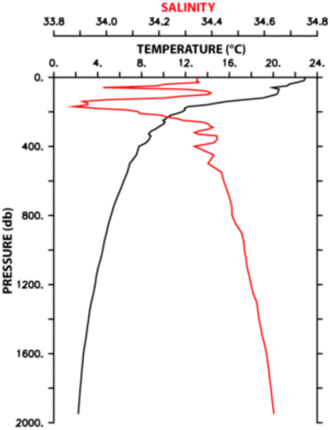
Tropical cyclones in water vapor: 95% confidence predictions in red, ground truth in black.

http://arxiv.org/abs/1708.05256

Understanding Southern Ocean Regimes - 1: ARGO

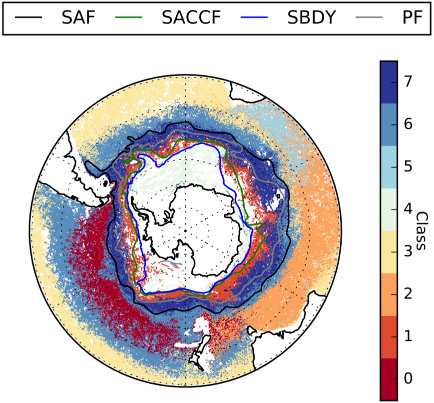






http://www.argo.ucsd.edu/About_Argo.html

Understanding Southern Ocean Regimes - 2: Unsupervised Learning



(Dan Jones, British Antarctic Survey)

- Applying Gaussian Mixture Modelling to cluster Southern Ocean Argo profiles.
- The number of classes was determined using two statistical tests.
- Also shown are several classically-defined fronts of the Antarctic Circumpolar Current.
- Note that the cluster edges (roughly) line up with the fronts. It suggests that GMM might be useful for front identification.

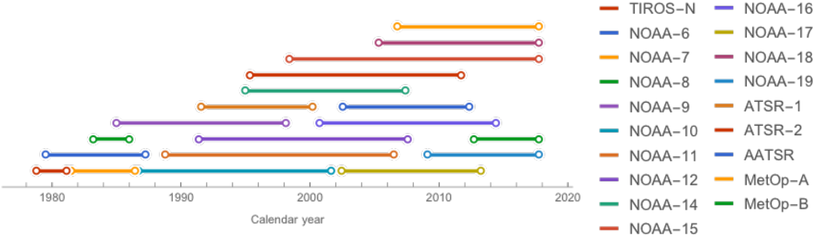
Harmonisation of time-series (1)

Problem: Nominal radiance data $\mathcal{L}_i$ obtained from different sensors $i, \ldots$ on board different satellites result in unexpected breaks in mean radiance and temporal trends when combined into multi-decadal fundamental climate data records. ML achieves this by answering 2 questions:

FIduceo

Homogenisation: What are the calibration coefficients a_i, a_j that minimise the inter-sensor differences $\mathcal{L}_i - \mathcal{L}_i$?

Harmonisation: What are the calibration coefficients a_i, a_j that minimise the differences between actual and expected inter-sensor differences $\mathcal{L}_i - \mathcal{L}_i - \mathcal{K}_{i,j}$?

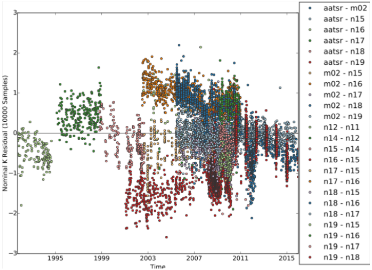


Harmonisation of time-series (2)

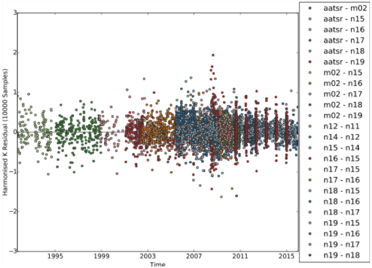
Ralf Quast, Ralf Giering (FastOpt, GmbH, Germany), Sam Hunt, Peter Harris (NPL, UK), Jonathan Mittaz, Michael Taylor (University of Reading, UK) (H2020 grant 638822)

Fiduceo

Nominal residuals:



Harmonised residuals:



Early results using machine learning techniques (see `http://www.fiduceo.eu/content/propagating-uncertainty-climate-data-record`): successfully merging these data and removing the jumps that can create spurious trends in the climate data record.

Direct Numerical Simulation

Primarily mathematical representation of a complex system of processes

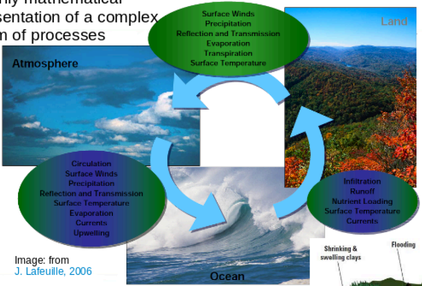
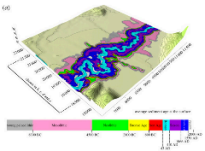
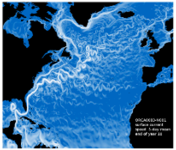
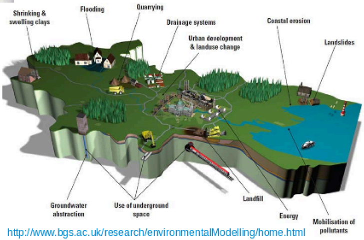


Image: from J. Lafeuille, 2006



Coulthard and Van De Wiel Doi: 10.1098/rsta.2011.0597

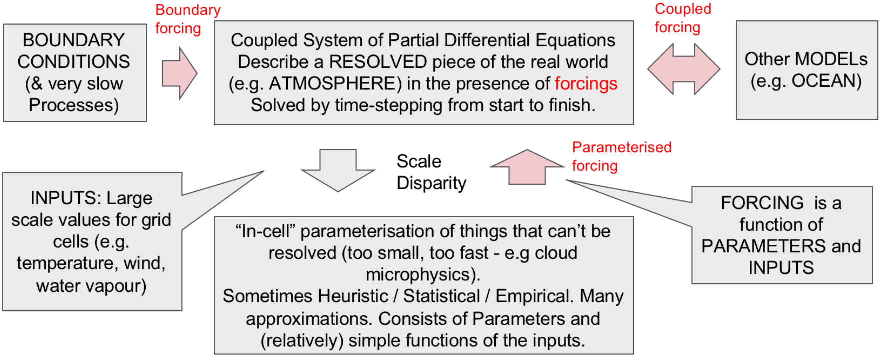




http://www.bgs.ac.uk/research/environmentalModelling/home.html

We want to observe and simulate the world at ever higher resolution! More complexity!

One slide introduction to numerical modelling



Machine Learning and Parameterisation

Optimising Parameterisations

Goal: Try to learn parameters of an *existing parameterisation* from observations and simulations.

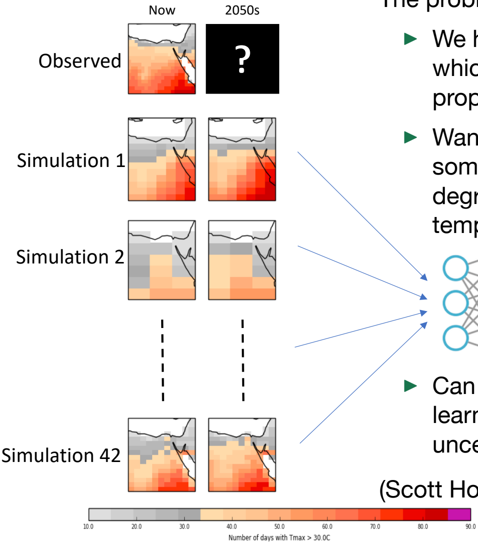
- Current methods for establishing parameters are based on "best estimates" within a range of physical possibilities, but these are generally done "one-by-one" - it is rare to try and establish the "correct" parameters across the complete set because of the large dimensionality.
- Tough computational problem! Currently addressed by brute force, if at all.

Replacing Parameterisations

Goal: Generate a new parameterisation (or replace an existing one) by *learning* rather than *modelling*.

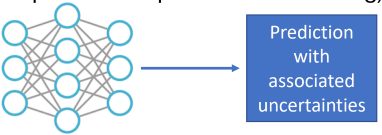
- We might want to replace a parameterisation (or even whole sub-model) to make it faster (e.g. emulating the behaviour of a complex model at lower resolution), because the scale disparities are too large, or the relationships are not known.
- Need to be careful about technique and applications and assumptions and constraints of stationarity of inputs.

Using Ensemble Output to develop new parameterisations



The problem:

- We have an ensemble of simulations which project/predict physical properties of the environment.
- Want to predict a climate indice at some specific location (e.g. growing degree days, or days where temperature requires airconditioning).



- Can apply a variety of machine learning approaches, but the need for uncertainties adds complexity.

(Scott Hosking, British Antarctic Survey)

Interesting Questions

How will climate change affect the global distribution of malaria?

July 2007 Tewkesbury flood: 3B€ loss!

Can we predict risk into the future?

What would be the impact of leakage from an oil and gas well in UK waters on the national economy, coastal and marine biodiversity and the well-being of the population affected?

How will climate change affect the incidence of road and rail closures due to landslides?

Take Care - Interdisciplinary Language is imprecise

Models

Are usually based on "Direct Numerical Simulation" even if some components are of necessity modelled with bulk statistical properties. Need to take care when talking with people for whom the word "model" can mean "statistical model".

Prediction

In climate science, model based prediction depends on confidence that the model is based on physical insight, and can predict emergent *physically sound* properties of change.

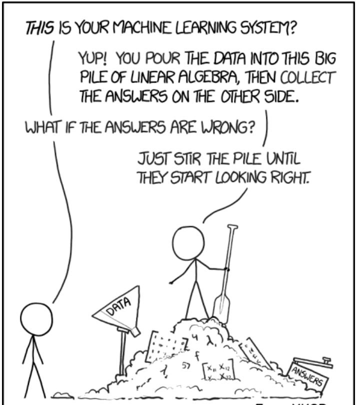
Take Care - Interdisciplinary Language is imprecise

Models

Are usually based on "Direct Numerical Simulation" even if some components are of necessity modelled with bulk statistical properties. Need to take care when talking with people for whom the word "model" can mean "statistical model".

Prediction

In climate science, model based prediction depends on confidence that the model is based on physical insight, and can predict emergent *physically sound* properties of change.

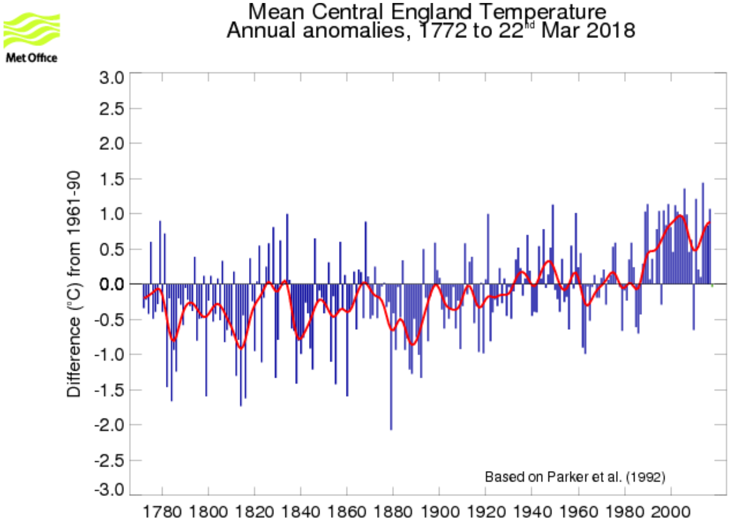


From XKCD

This is often fine, but when **prediction** is required, check assumptions and feedbacks!

Summary

Environmental science has been a *data science* since forever …



Summary

- Data science challenges abound: from those associated with data volume, velocity, variety, value, veracity (provenance), voting (standards), to the hardware and software platforms required.
- There are many pioneering interdisciplinary activities exploiting “modern” data science (aka machine learning, AI, and friends), and much scope for more!
- Like all interdisciplinary work, there is scope for misunderstanding and misapplication, but those risks sit right next to **opportunity**.