Climate Data
Issues, Systems, and Opportunities for MPE June 2018

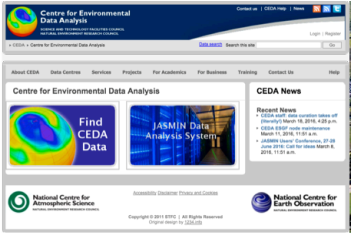
Et. Al. &
Bryan Lawrence

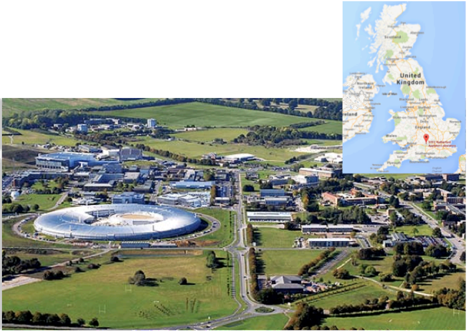
NERC SCIENCE OF THE ENVIRONMENT

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

CEDA

Centre for Environmental Data Analysis:





"to support environmental science, further environmental data archival practices, and develop and deploy new technologies to enhance access to data"

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.
- Embedded in existing school alongside mathematics and meteorology.
- Research groups include "Data Analytics", "Data Science and AI" and "*Advanced Computing for Environmental Sciences*".

  https://aces.cs.reading.ac.uk

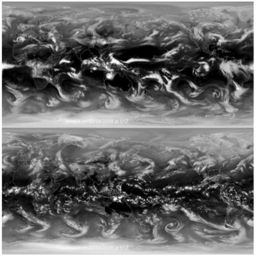
Outline

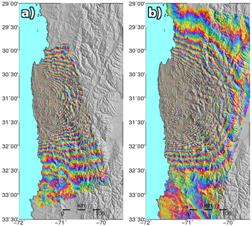
1. Characteristics of Environmental Science Data
2. Simulation, Models and Data
3. Smarter Computing (Software and Hardware)
4. Opportunties
5. Summary

What is Environmental Data? Diverse

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

What is Environmental Data?: Multiscale

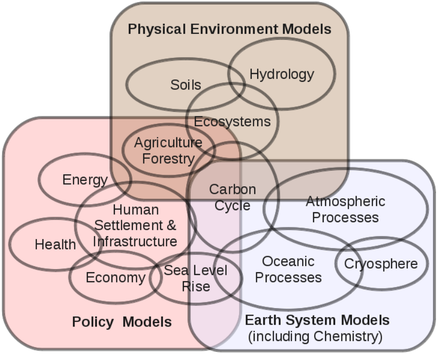




(Examples from JASMIN users:

- 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!

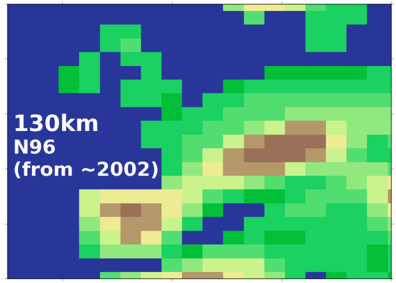


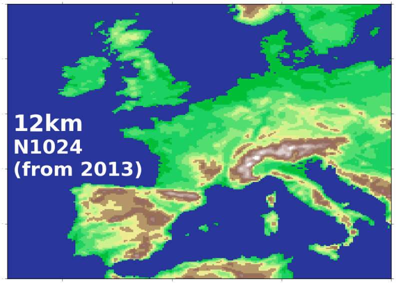
Many interacting communities, each with their own software, data (standards), compute environments etc.

Figure adapted from Moss et al, 2010

What is Environmental Data? Voluminous!

Europe within a global model …





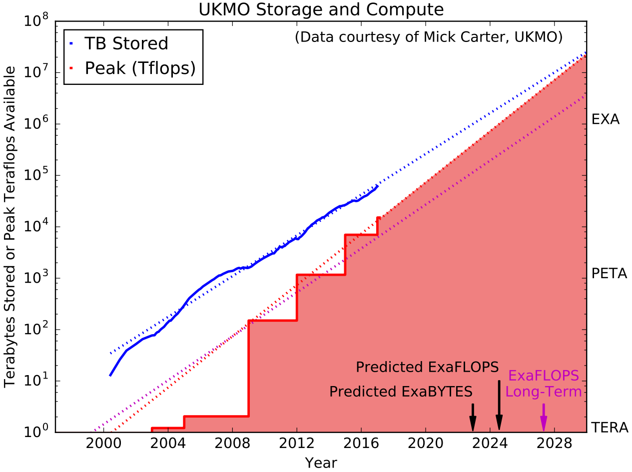
One "field-year" — 26 GB

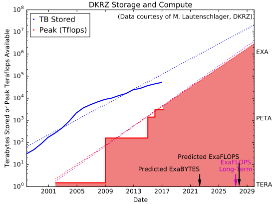
1 field, 1 year, 6 hourly, 80 levels
1 x 1440 x 80 x 148 x 192

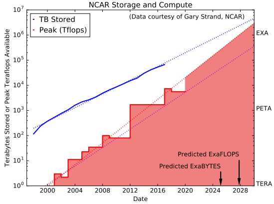
One "field-year" — >6 TB

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

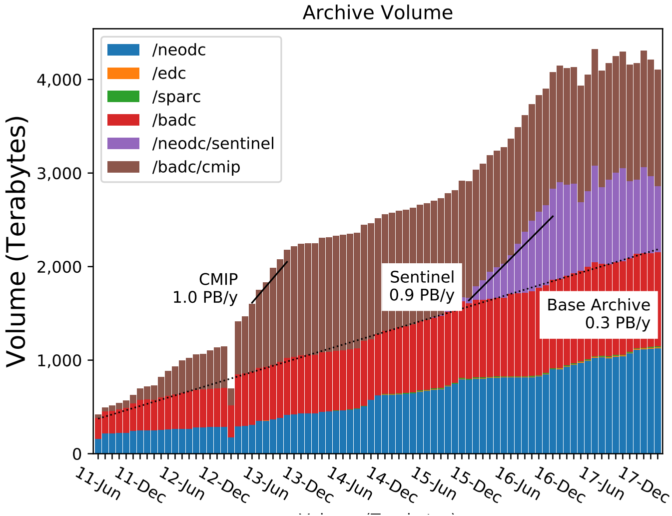
What is Environmental Data? Voluminous!



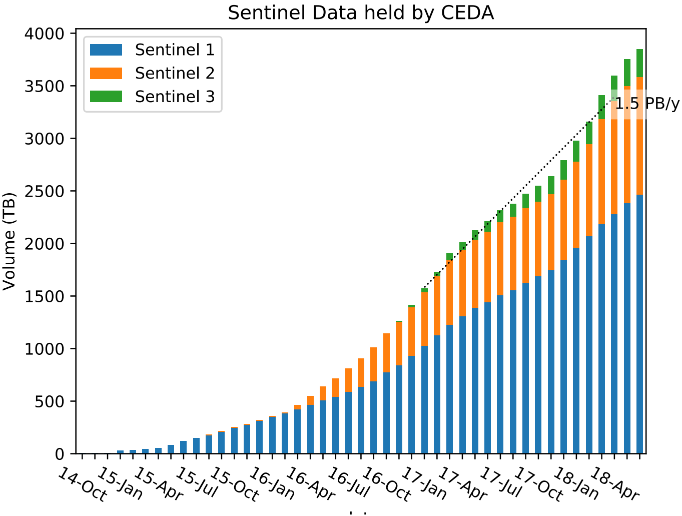




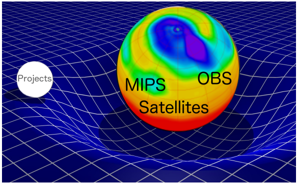
CEDA: Archive Growth



CEDA: Sentinel Growth

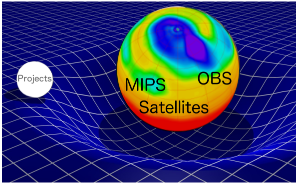


JASMIN — The Data Commons

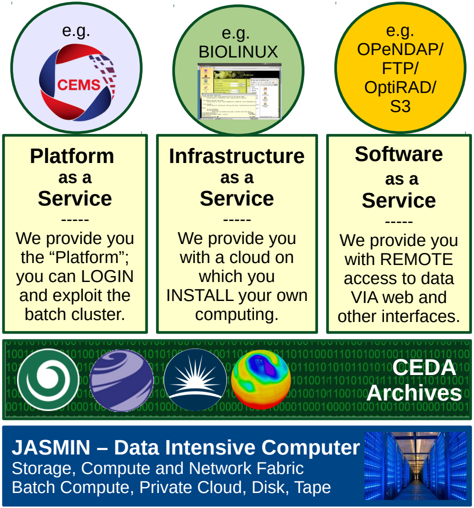


- Provide a state-of-the art storage and computational environment
- Provide and populate a managed data environment with key datasets (the "archive").
- Encourage and facilitate the bringing of data and/or computation alongside/to the archive!
- Provide FLEXIBLE methods of exploiting the computational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

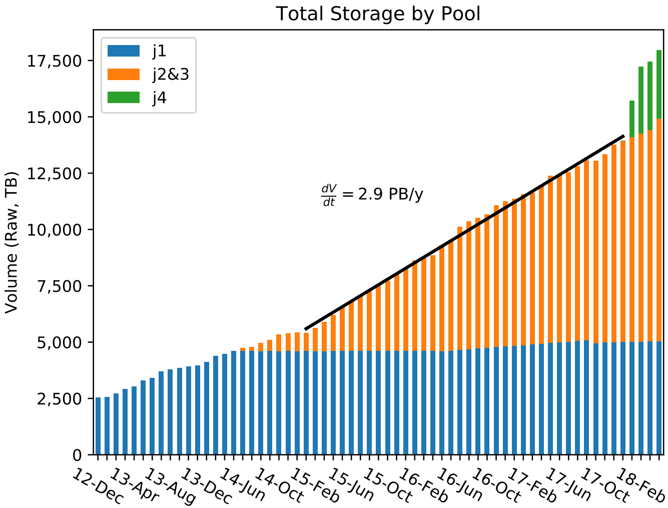
JASMIN — The Data Commons



- 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: Total Storage Growth



Climate Data in the context of the "Zettabyte Era"

- Global internet traffic per annum, 1.2 ZB in 2016, forecast to reach 3.3 ZB per annum by 2021 (9 EB/day).*
- Estimated power consumption of data centres globally: 1.5% of all global power (2015) - comparable to aviation! **

Climate Data in the context of the “Zettabyte Era”

- Global internet traffic per annum, 1.2 ZB in 2016, forecast to reach 3.3 ZB per annum by 2021 (9 EB/day).*
- Estimated power consumption of data centres globally: 1.5% of all global power (2015) - comparable to aviation! **
- A high resolution climate model needs to run somewhere between 0.5 and 5 Simulated Years Per real Day (SYPD)
- A real 1km model may have > 200 levels and o(10-100) prognostic variables, and a minimum useful ensemble size of 10.
- Assuming it is possible to integrate such as model at the required rate (which may be impossible),

Climate Data in the context of the "Zettabyte Era"

- Global internet traffic per annum, 1.2 ZB in 2016, forecast to reach 3.3 ZB per annum by 2021 (9 EB/day).*
- Estimated power consumption of data centres globally: 1.5% of all global power (2015) - comparable to aviation! **
- A high resolution climate model needs to run somewhere between 0.5 and 5 Simulated Years Per real Day (SYPD)
- A real 1km model may have > 200 levels and o(10-100) prognostic variables, and a minimum useful ensemble size of 10.
- Assuming it is possible to integrate such as model at the required rate (which may be impossible), then we get (0.5-5) x 200 x (10-100) x 1 PB =

  $10^3 \rightarrow 10^5$PB (100 EB) *per real day*.
- A lot wrong with this calculation, but …

* https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html ** Wikipedia, June 2018

SI Units, 2018

SI-prefix	Name	Scale
k kilo	thousand	10^3
M mega	million	10^6
G giga	billion	10^9
T Tera	trillion	10^{12}
P Peta	quadrillion	10^{15}
	(multi-PB)	10^{16-17}
E exa	quintillion	10^{18}
Z zetta	sextillion	10^{21}
Y yotta	septillion	10^{24}

SI Units, 2018

SI-prefix	Name	Scale	Status (2011)
k kilo	thousand	10^3	Count on fingers
M mega	million	10^6	Trivial
G giga	billion	10^9	Small
T Tera	trillion	10^{12}	Real
P Peta	quadrillion	10^{15}	Challenging
	(multi-PB)	10^{16-17}	Possible
E exa	quintillion	10^{18}	Aspirational
Z zetta	sextillion	10^{21}	Wacko
Y yotta	septillion	10^{24}	Science Fiction

From an orginal table by Stuart Feldman, Google

Challenging = Just about feasible for Google …
Far too easy to say "peta" and "exa" ...

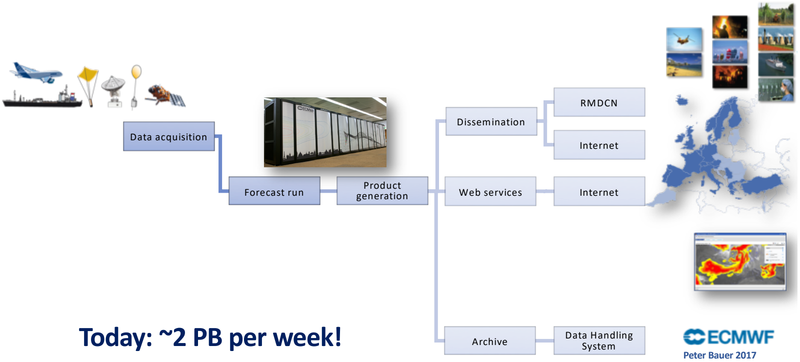
SI Units, 2018

SI-prefix	Name	Scale	Status (2011)	Status (2018)
k kilo	thousand	10^3	Count on fingers	Free
M mega	million	10^6	Trivial	Free
G giga	billion	10^9	Small	Free
T Tera	trillion	10^{12}	Real	Small
P Peta	quadrillion	10^{15}	Challenging	Real
	(multi-PB)	10^{16-17}	Possible	Challenging
E exa	quintillion	10^{18}	Aspirational	Possible
Z zetta	sextillion	10^{21}	Wacko	Aspirational
Y yotta	septillion	10^{24}	Science Fiction	Wacko

From an orginal table by Stuart Feldman, Google

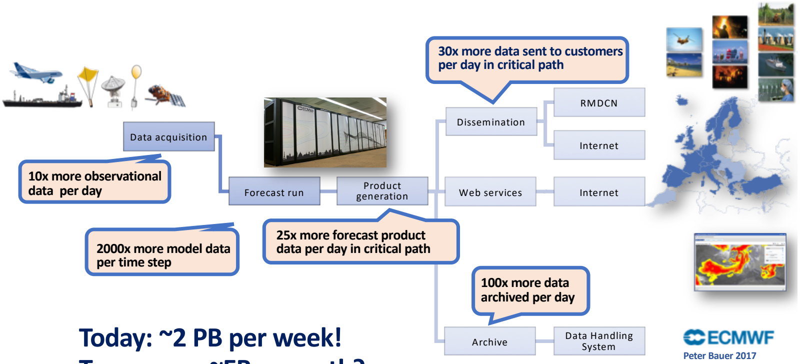
Challenging = Just about feasible for Google …
Far too easy to say “peta” and “exa” ...

What is Environmental Data? Part of Voluminous workflows!



Today: ~2 PB per week!

What is Environmental Data? Part of Voluminous workflows!

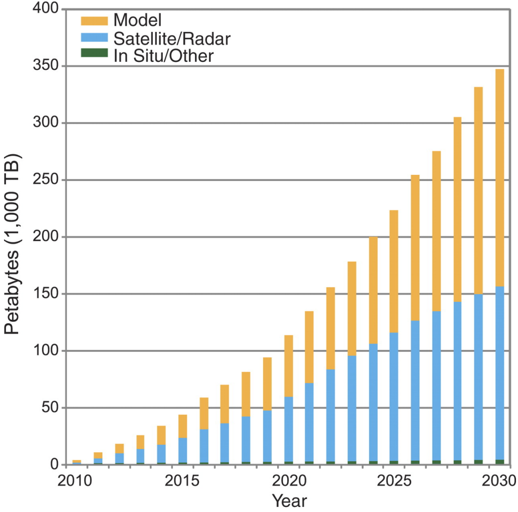


Today: ~2 PB per week!
Tomorrow: ~EB a month?

What is Environmental Data? Voluminous Global Sharing?

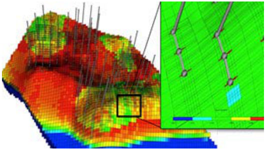
Fig. 2 The volume of worldwide climate data is expanding rapidly, creating challenges for both physical archiving and sharing, as well as for ease of access and finding what's needed, particularly if you're not a climate scientist.

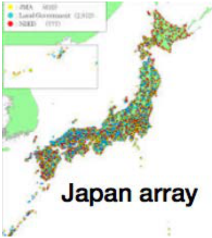
(BNL: Even if you are?)



J T Overpeck et al. Science 2011;331:700-702

Not only Weather and Climate have a volume problem

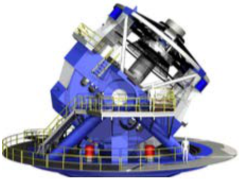

Reservoir Modelling ≈350 TB/run

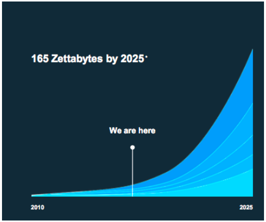


Seismic Network ≈150 TB/y


Copernicus/SWOT ≈4 PB/d


LOFAR/SKA ≈4EB/yr

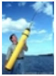
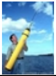

LSST/EUCLID ≈20 PB/night



Internet and IOT

(courtsey of Stéphane Requena – GENCI/PRACE)

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*)	 

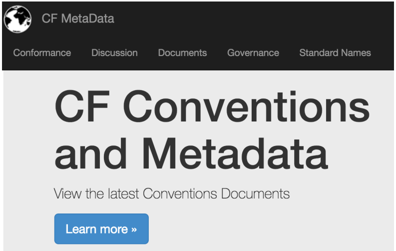
Classify by geometry, but that doesn't tell you how it stored, or what it is.

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).

Data Conventions - The Climate Forecast Conventions



`http://cfconventions.org`

A format is just a bucket:

- The CF conventions describe how to make data files self-describing.
- The conventions are a bit daunting, but there are some good software libraries that can make creation and usage of the cfconventions easy:
  - e.g. cf-python: `https://cfpython.bitbucket.io/`
- See also `https://doi.org/10.5194/gmd-10-4619-2017` for a description of the CF data model.

Exploiting a data model

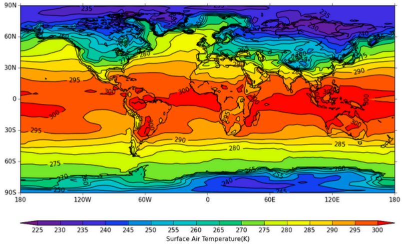
```
>>> f = cf.read('file.nc')[0]
>>> type(f)
<class 'cf.field.Field'>
>>> f
<CF Field: air_temperature(latitude(4), longitude(5)) K>
>>> print f
air_temperature field summary
-----------------------------
Data            : air_temperature(latitude(4), longitude(5)) K
Cell methods    : time: mean
Dimensions      : latitude(4) = [-2.5, ..., 5.0] degrees_north
                : longitude(5) = [0.0, ..., 15.0] degrees_east
                : time(1) = [2000-01-16 00:00:00] 360_day calendar
                : height(1) = [2.0] m
```

Exploiting a data model

```
>>> f = cf.read('file.nc')[0]
>>> type(f)
<class 'cf.field.Field'>
>>> f
<CF Field: air_temperature(latitude(4), longitude(5)) K>
>>> print f
air_temperature field summary
-----------------------------
Data
Cell methods
Dimensions
```

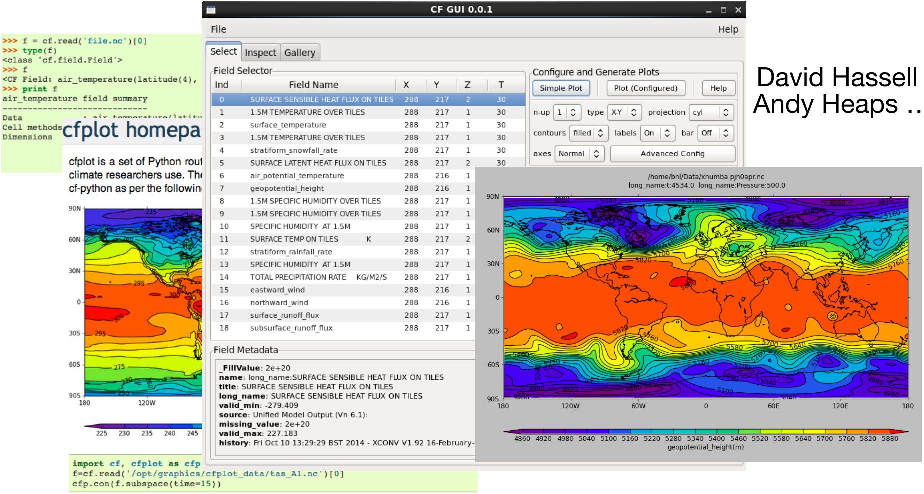
cfplot homepage

cfplot is a set of Python routines for making the common contour and vector plots that climate researchers use. The data to make a contour plot can be passed to cfplot using cf-python as per the following example.




```
import cf, cfplot as cfp
f=cf.read('/opt/graphics/cfplot_data/tas_A1.nc')[0]
cfp.con(f.subspace(time=15))
```

Exploiting a data model



David Hassell
Andy Heaps …

cf-python, cf-plot, and cf-gui — all built on the cf data model!

Direct Numerical Simulation

Primarily mathematical representation of a complex system of processes

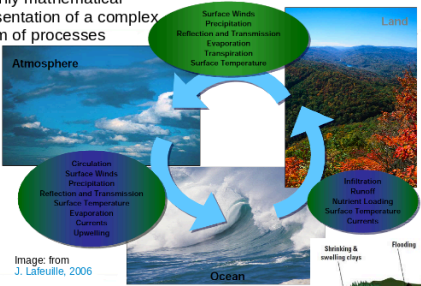
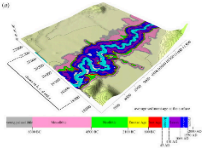
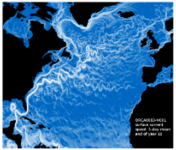
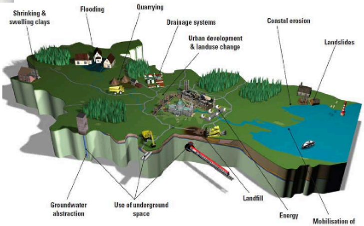


Image: from J. Lafeuille, 2006



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

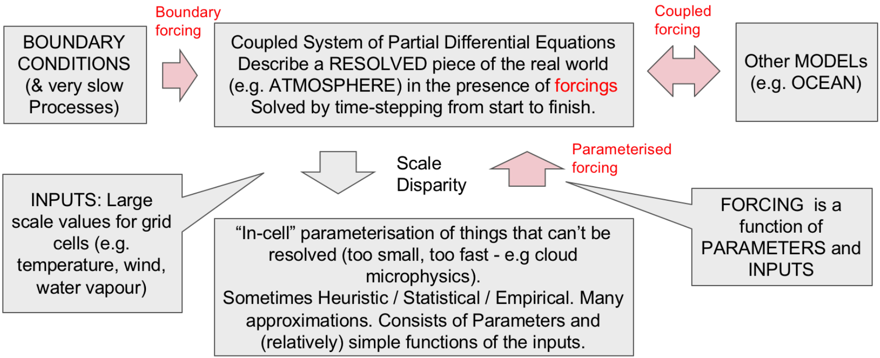




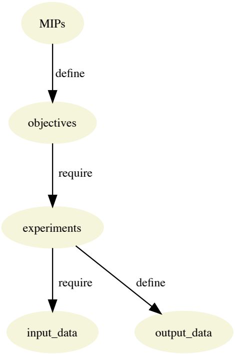
http://www.bgs.ac.uk/research/environmentalModelling/home.html

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

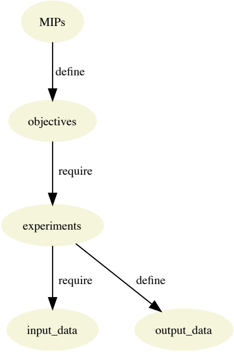
One slide introduction to numerical modelling

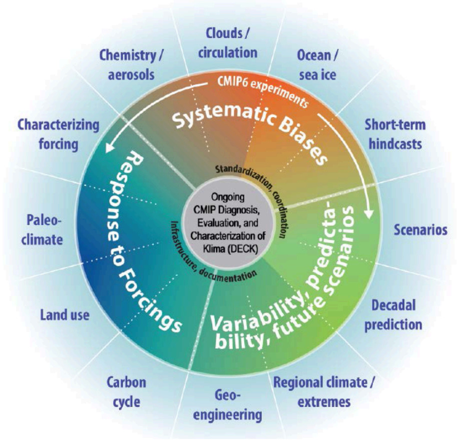


Model Intercomparison Projects - CMIP6

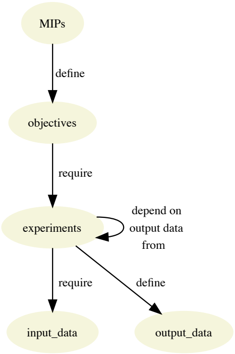


Model Intercomparison Projects - CMIP6

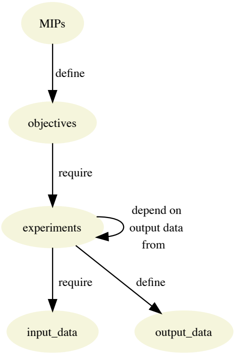


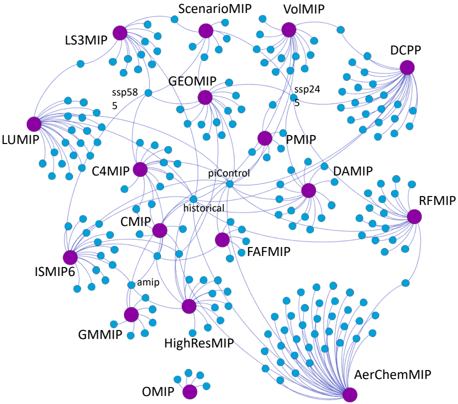


Model Intercomparison Projects - CMIP6



Model Intercomparison Projects - CMIP6

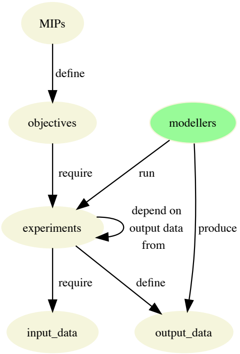




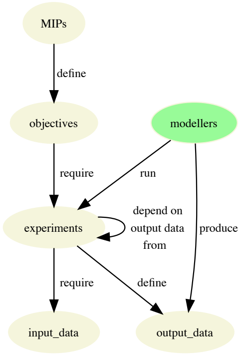
Complicated Experimental Interdependency!

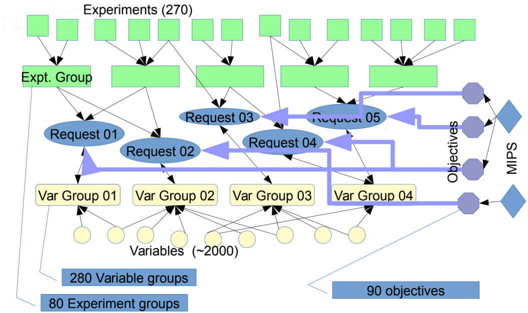
(Courtesy of Charlotte Pascoe and the ES-DOC project.)

Model Intercomparison Projects - CMIP6



Model Intercomparison Projects - CMIP6

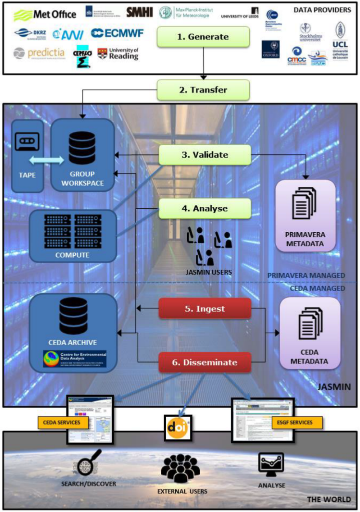




Complicated Data Requirements for Modelling Groups!

(Courtesy of Martin Juckes and his Data Request activity in support of CMIP6.)

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



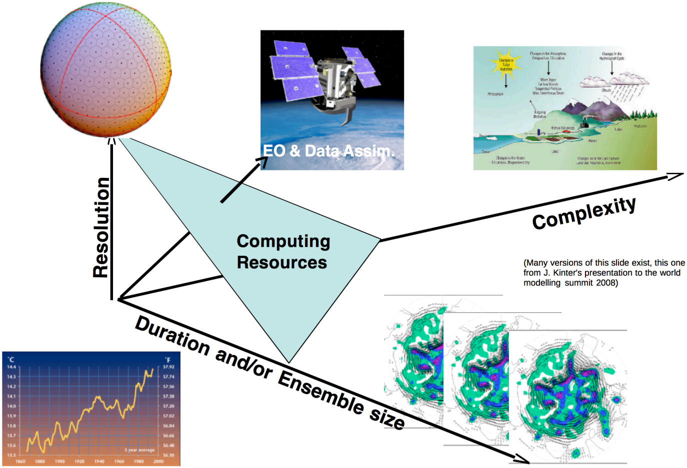
jon.seddon@metoffice.gov.uk

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!

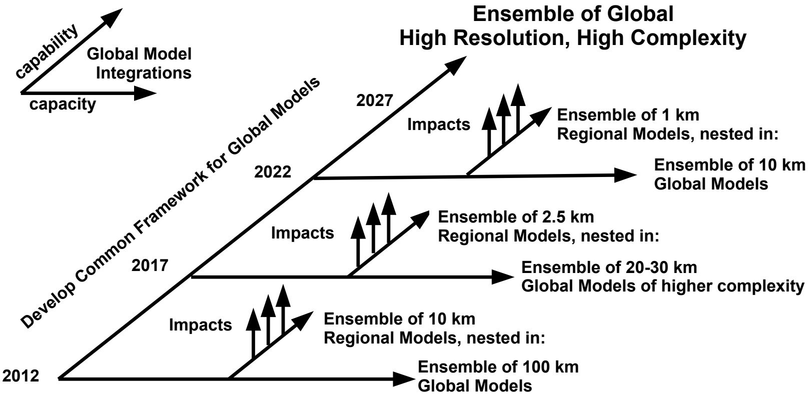
Give me more computing? Global Climate Modelling



(Many versions of this slide exist, this one from J. Kinter's presentation to the world modelling summit 2008)

Where is this going

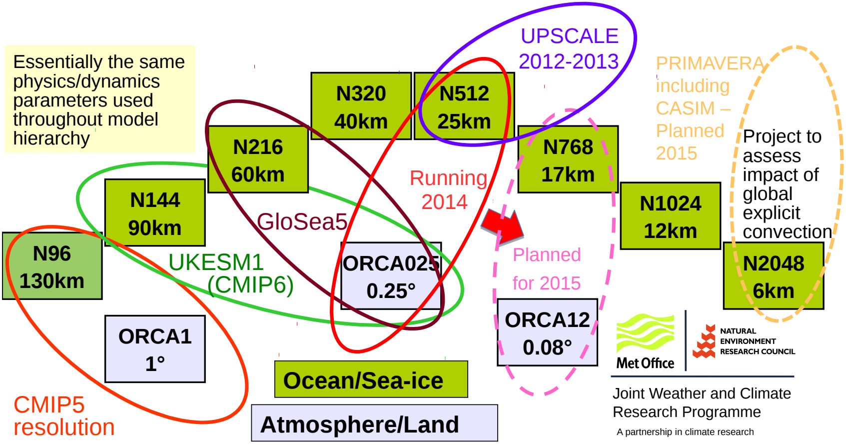
One of many views:



JWCRP Climate Modelling

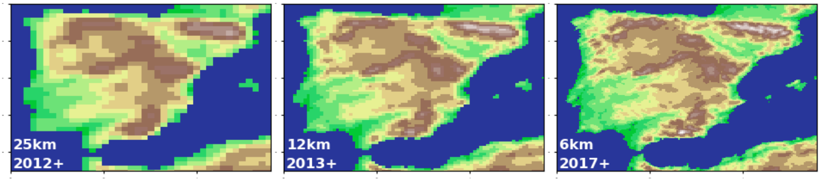
Earth System Modelling
PI C. Jones (NCAS at the Met Office)

High Resolution Climate Modelling
Joint PIs: P-L. Vidale (NCAS), M. Roberts (Met Office)



Joint Weather and Climate Research Programme
A partnership in climate research

Voluminous …and getting worse!



What about 1km? That's the current European Network for Earth System Modelling (ENES) goal! Consider N13256 (1.01km):

- 1 field, 1 year, 6 hourly, 180 levels
- 1 x 1440 x 180 x 26512 x 19884 = 1.09 PB
- Would take 760 seconds to read one 760 GB grid at 1 GB/s
- **Can no longer consider serial diagnostics!**

Stop writing data AND be much smarter!

Techniques for data reduction

1. Reduce temporal frequency of output
2. Compress Data
   - Lossless,
   - Lossey (how many bits do we really need?)
3. Reduce spatial freqency of output (real resolution is much lower than numerical resolution),
4. “In-Flight” Diagnostics
5. Ensemble Compression

First two are in use now, the next three are really important too ...

Stop writing data AND be much smarter!

Techniques for data reduction

1. Reduce temporal frequency of output
2. Compress Data
   - Lossless,
   - Lossey (how many bits do we really need?)
3. Reduce spatial freqency of output (real resolution is much lower than numerical resolution),
4. “In-Flight” Diagnostics
5. Ensemble Compression

First two are in use now, the next three are really important too ...

Smarter Data Use

Large volumes of data take a long time to *read* even if you can store them!

- Huge scope for better algorithms both for data reduction and when the data hasn’t been reduced, to exploit the data.

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

Uncommon (and inappropriate?) software solutions

Multiple tools

Contrast between two very types of workflow:

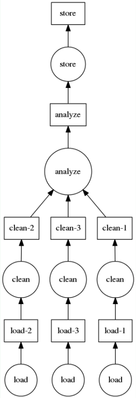
- Build Once: Many analysis tasks are build once, use once, throwaway. No room for optimisation (or MPI). *Need efficient libraries.*
- Repeatable: "build", "run", "move", "reduce/reformat", "analyse". *Much room for automation.*.

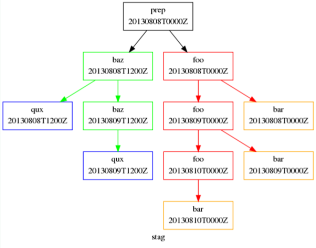
What to use? Plethora of architectures and tools out there

Spark DASK Pandas

Exploiting Concurrency

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

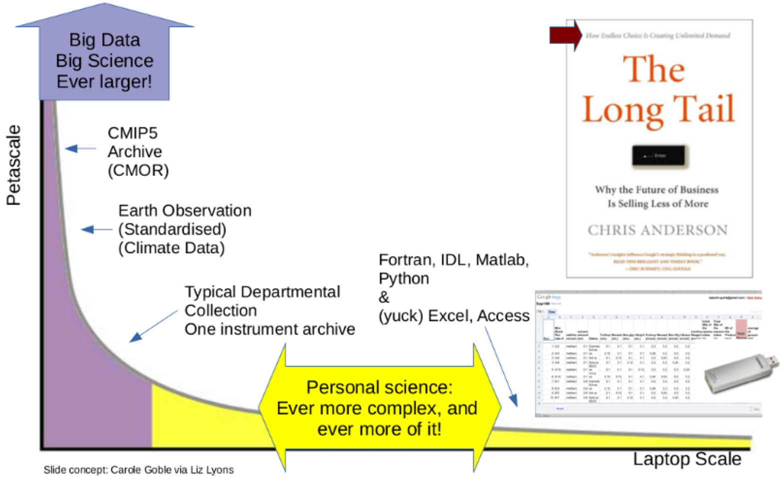




Much to do to harness tools to accelerate workflows!

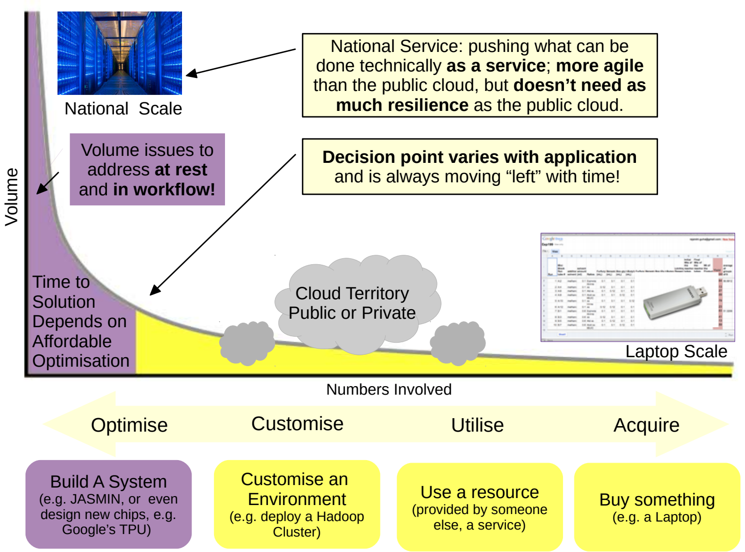
(These two examples: dask, and cylc, representing bespoke analysis and scheduling, reduction and proliferation.)

Wide Scope



Slide concept: Carole Goble via Liz Lyons

Wide Scope

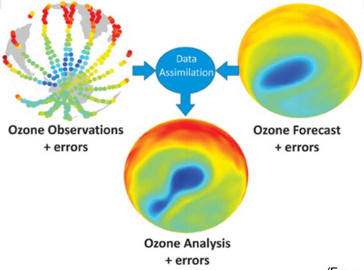

National Scale
National Service: pushing what can be done technically **as a service**; **more agile** than the public cloud, but **doesn't need as much resilience** as the public cloud.
Volume issues to address **at rest** and **in workflow!**
Decision point varies with application and is always moving "left" with time!
Volume
Time to Solution Depends on Affordable Optimisation
Cloud Territory Public or Private
Laptop Scale
Numbers Involved
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)

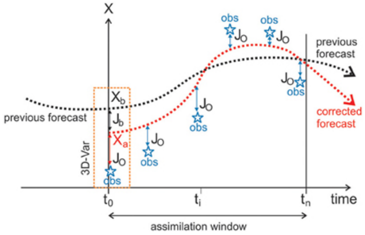
Application Opportunities

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.

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

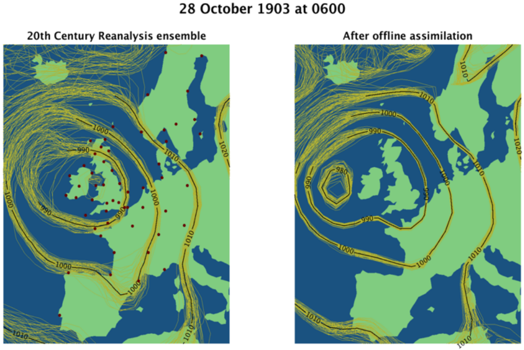
- Massive computing initiative.
- Heroic data iniative: 1.7 Billion Observations. 1 TB a year of output data.

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

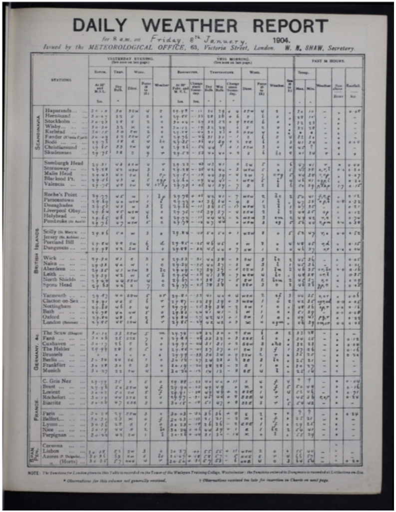
If we want to know about change, we need to know the baseline.



(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.

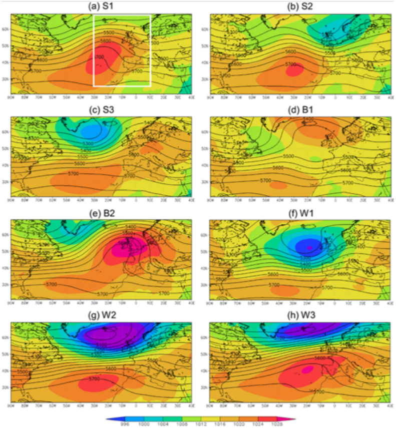
Opportunity: Large body of training data, and robust validation methodology.

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).



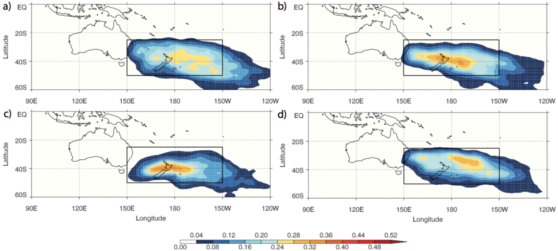
(Santos et al 2016, doi:10.1002/2015JD024399

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.

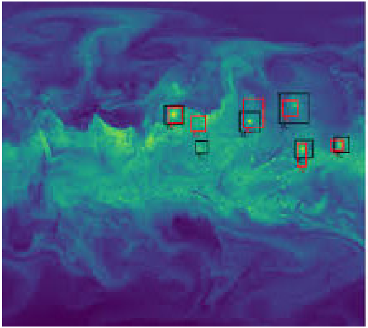
From J. Catto, 2018, doi:10.1175/JCLI-D-17-0746.1

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

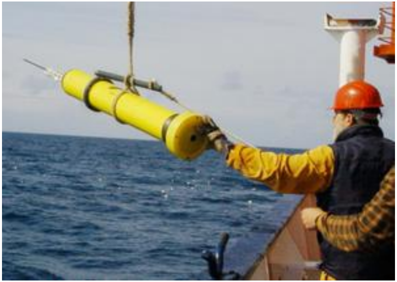
- Current Deep Learning implementations can take days to converge on O(10) GB datasets.
- Using a 15 TB climate dataset (768x768, 16 channels, 0.4M images)
- 9622 KNL nodes and sustained ≈ 12 PFLOP/s during classification
- Two HPC perspectives to consider for deep learning:
  1. How efficient is deep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

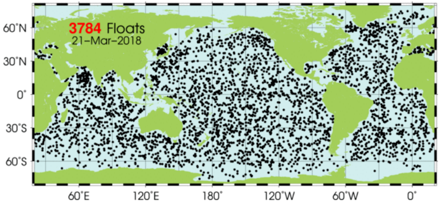


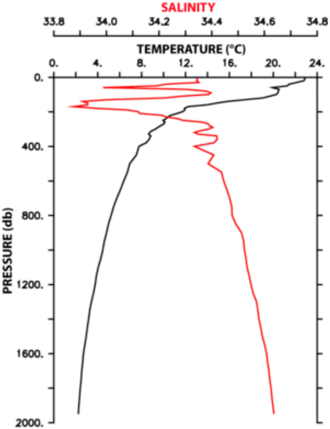
Tropical cyclones in water vapor: 95% confidence predictions in red, ground truth in black.

http://arxiv.org/abs/1708.05256

Understanding Southern Ocean Regimes - 1: ARGO

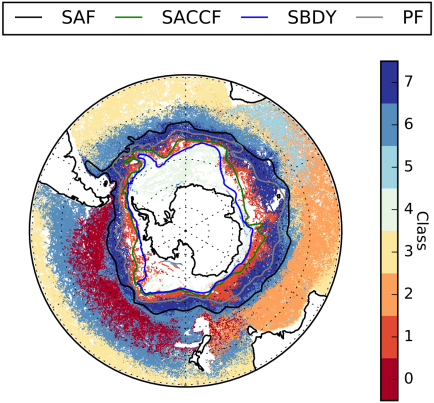






http://www.argo.ucsd.edu/About_Argo.html

Understanding Southern Ocean Regimes - 2: Unsupervised Learning



(Dan Jones, British Antarctic Survey)

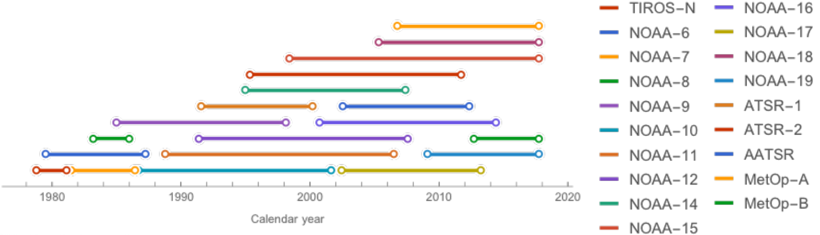
- Applying Gaussian Mixture Modelling to cluster Southern Ocean Argo profiles.
- The number of classes was determined using two statistical tests.
- Also shown are several classically-defined fronts of the Antarctic Circumpolar Current.
- Note that the cluster edges (roughly) line up with the fronts.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

Harmonisation of time-series (1)

Problem: Nominal radiance data 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 either of two questions:

Homogenisation: What are the calibration coefficients a_i, a_j that minimise the inter-sensor differences $L_i - L_i$?

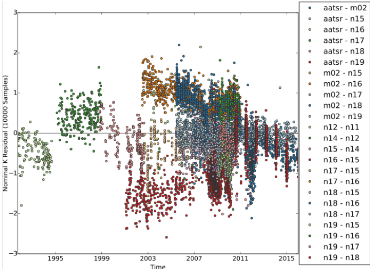
Harmonisation: What are the calibration coefficients a_i, a_j that minimise the differences between actual and expected inter-sensor differences $L_i - L_i - K_{i,j}$?



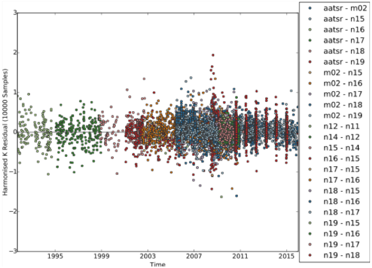
Harmonisation of time-series (2)

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

Nominal residuals:

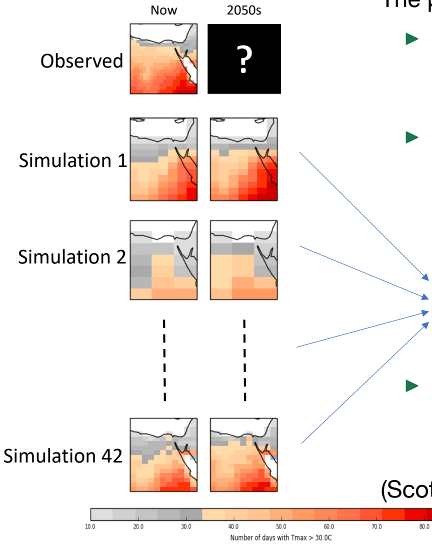


Harmonised residuals:



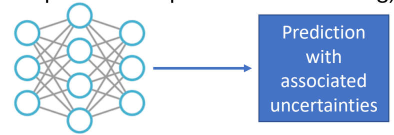
Early results using machine learning techniques (see `http://www.fiduceo.eu/content/propagating-uncertainty-climate-data-record`): successfully merging these data and removing the jumps that can create spurious trends in the climate data record.

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.

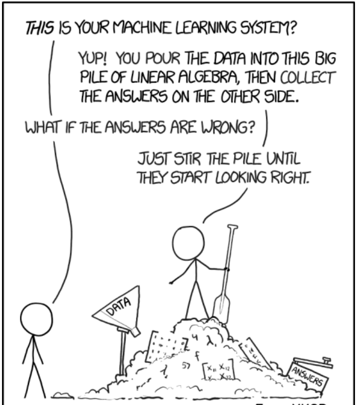
Take Care - Interdisciplinary Language is imprecise

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.

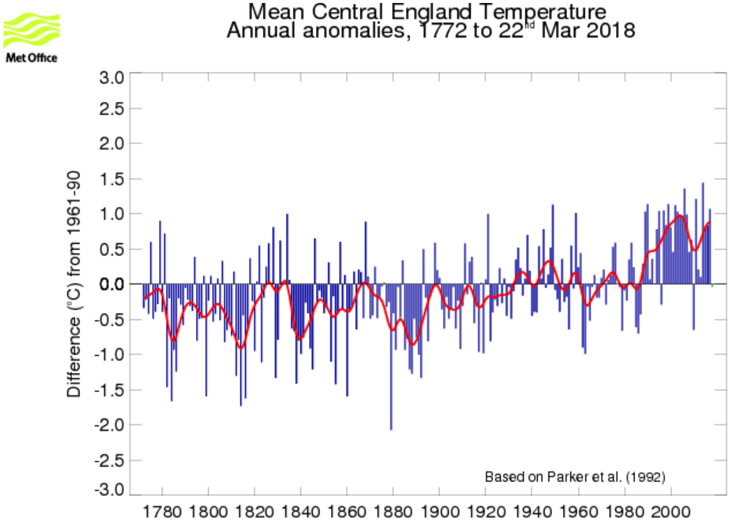


From XKCD

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

Summary

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



Summary

- Environmental data is messy, heterogenous, and volumnous.
  - The original description of “big data” talked about volume, velocity, and variety.
  - We then added value, veracity (provenance), voting (standards) ...
- Handling future volume will require changes to the way we think, from algorithms 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!

What the Data Deluge In Life Sciences Means For Exascale And Clouds

"Today, without a well executed software and data strategy, essentially the entire modern scientific method just simply falls apart."

"The next ten years will be critical because data will not only continue to be collected at an ever-faster rate, but we will also need to compute against all of it. At the same time."

(Anthony Philippakis, Broad Institute)

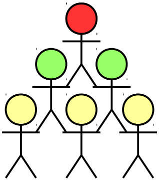
Data Infrastructure	Modernized Data Ecosystem	Data Management, Analytics, and Tools	Workforce Development	Stewardship and Sustainability
• Optimize data storage and security • Connect NIH data systems	• Modernize data repository ecosystem • Support storage and sharing of individual datasets • Better integrate clinical and observational data into biomedical data science	• Support useful, generalizable, and accessible tools and workflows • Broaden utility of and access to specialized tools • Improve discovery and cataloging resources	• Enhance the NIH data-science workforce • Expand the national research workforce • Engage a broader community	• Develop policies for a FAIR data ecosystem • Enhance stewardship

(NIH Data Plan)

Source: https://www.nextplatform.com/2018/06/14/what-data-deluge-means-life-sciences-exascale-clouds/

Modern Science: How do we work?

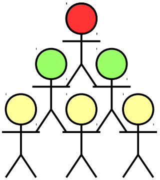
How we worked



PI stands on the shoulders of her postdocs and students (and as Newton would have said, the giants.)

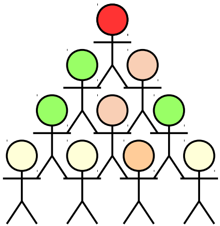
Modern Science: How do we work?

How we worked



PI stands on the shoulders of her postdocs and students (and as Newton would have said, the giants.)

How we work



PI stands on the shoulders of her postdocs, students, software engineers and data scientists.
(Are the giants down with the turtles?).