Science & Technology Facilities Council

# Environmental and Climate Data Challenges

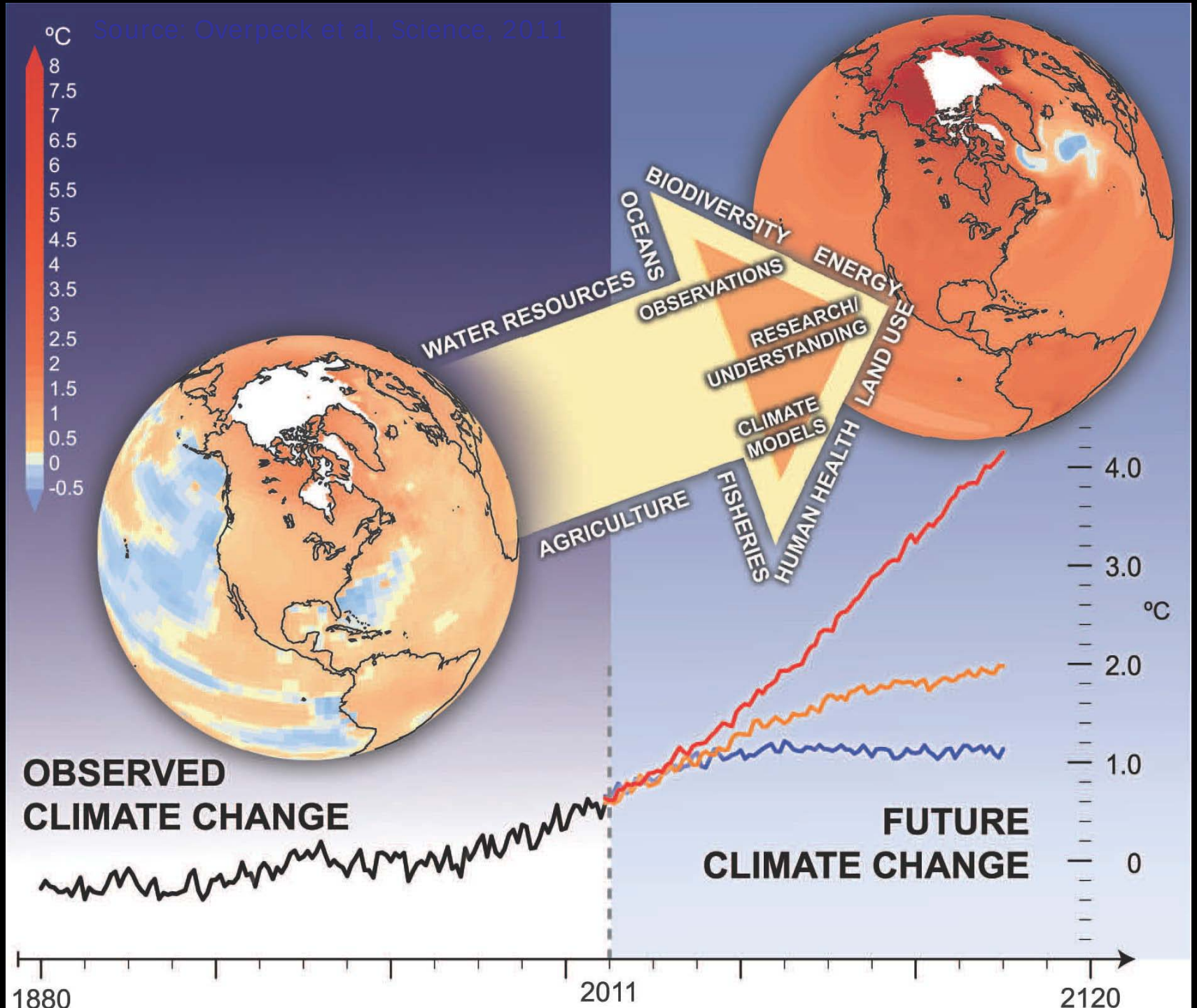

## Mike Ashworth

Associate Director

Computational Science & Engineering Department

STFC Daresbury Laboratory

mike.ashworth@stfc.ac.uk


Source: Overpeck et al, Science, 2011
°C
8
7.5
7
6.5
6
5.5
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
-0.5
BIODIVERSITY
ENERGY
LAND USE
HUMAN HEALTH
FISHERIES
AGRICULTURE
WATER RESOURCES
OCEANS
OBSERVATIONS
RESEARCH/
UNDERSTANDING
CLIMATE
MODELS
OBSERVED
CLIMATE CHANGE
FUTURE
CLIMATE CHANGE
4.0
3.0
°C
2.0
1.0
0
1880
2011
2120

# Outline

- The Climate Data Deluge – Simulation and Earth Observation
- On The Path to Exascale
- ICE-CSE and Data-Intensive Computing

# Outline

- The Climate Data Deluge – Simulation and Earth Observation
- On The Path to Exascale
- ICE-CSE and Data-Intensive Computing

Science & Technology Facilities Council

# The World in Global Climate Models

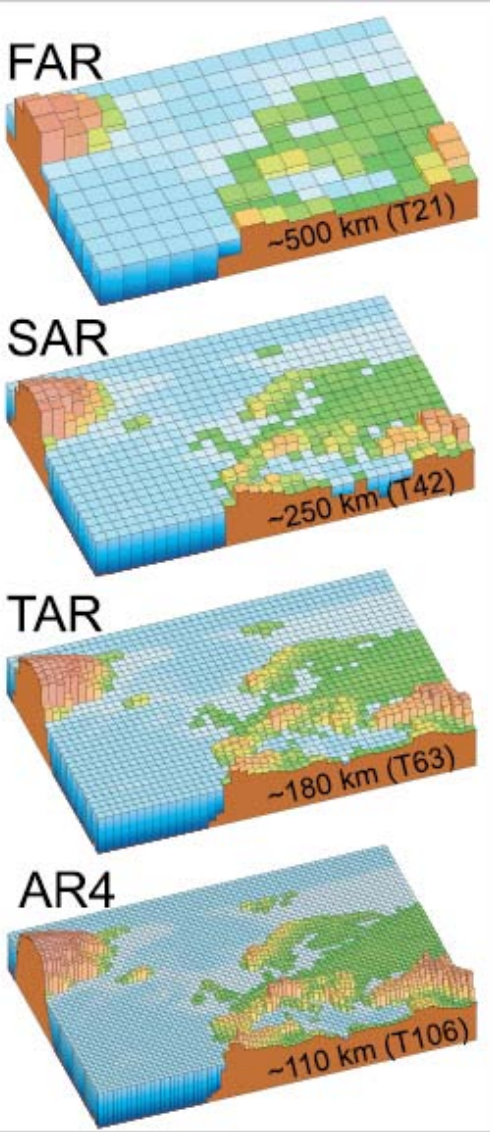
**FAR:1990**

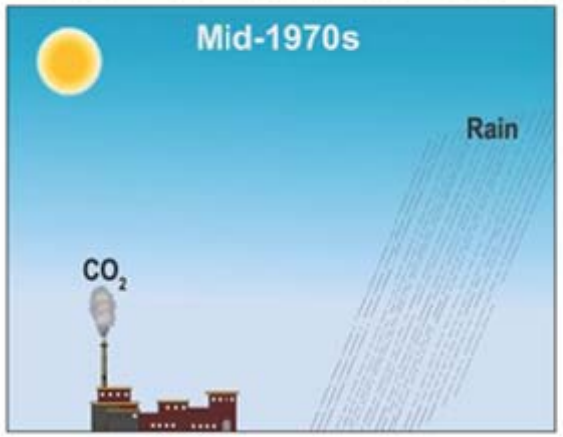
**SAR:1995**

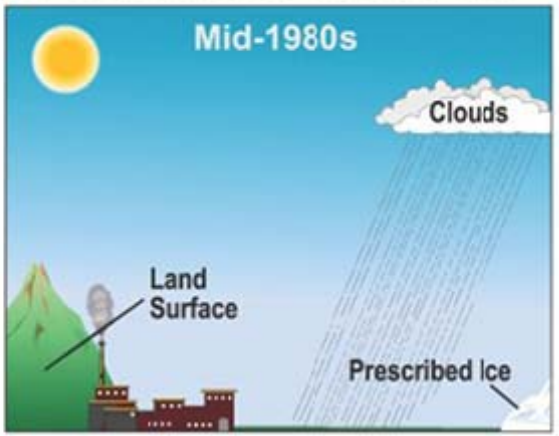
**TAR:2001**

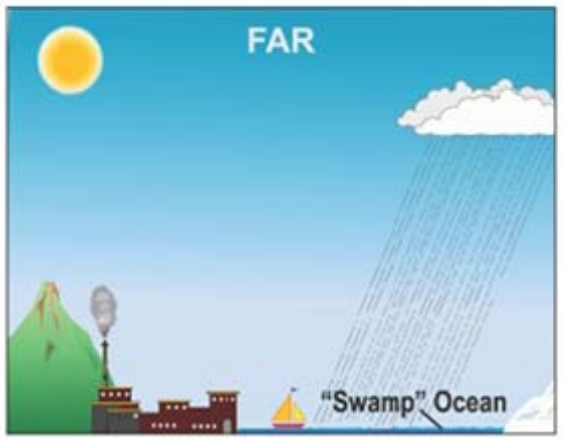
**AR4:2007**

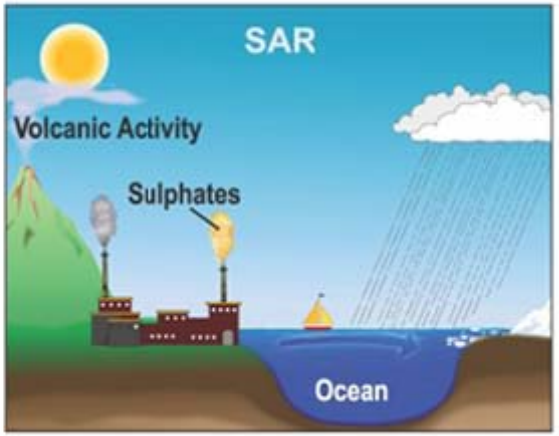
**AR5:2013**

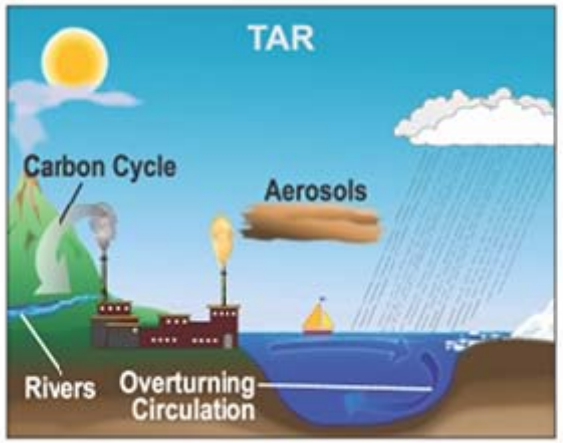


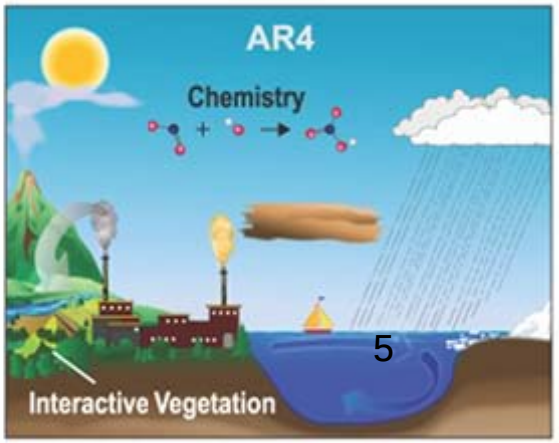



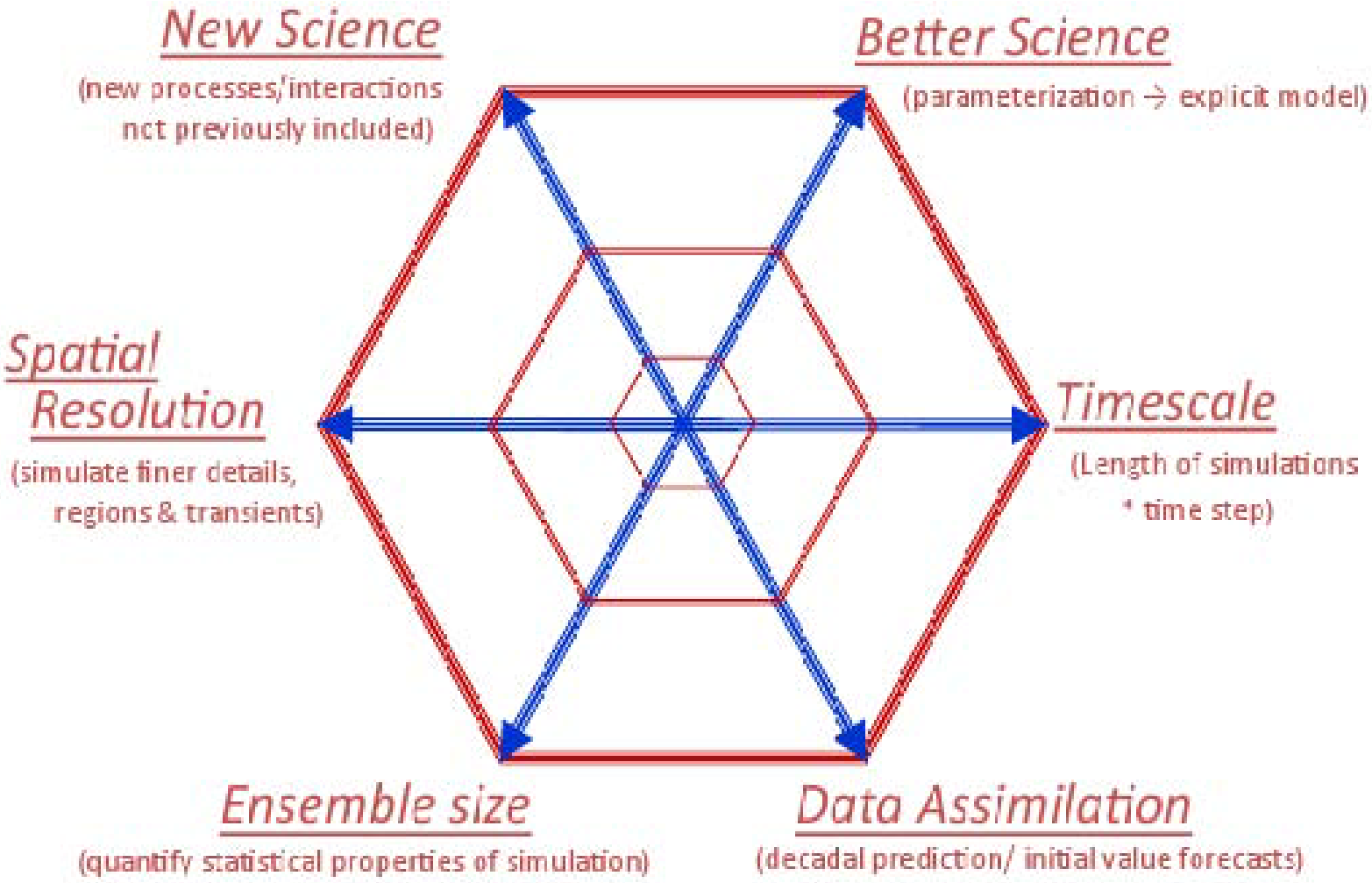



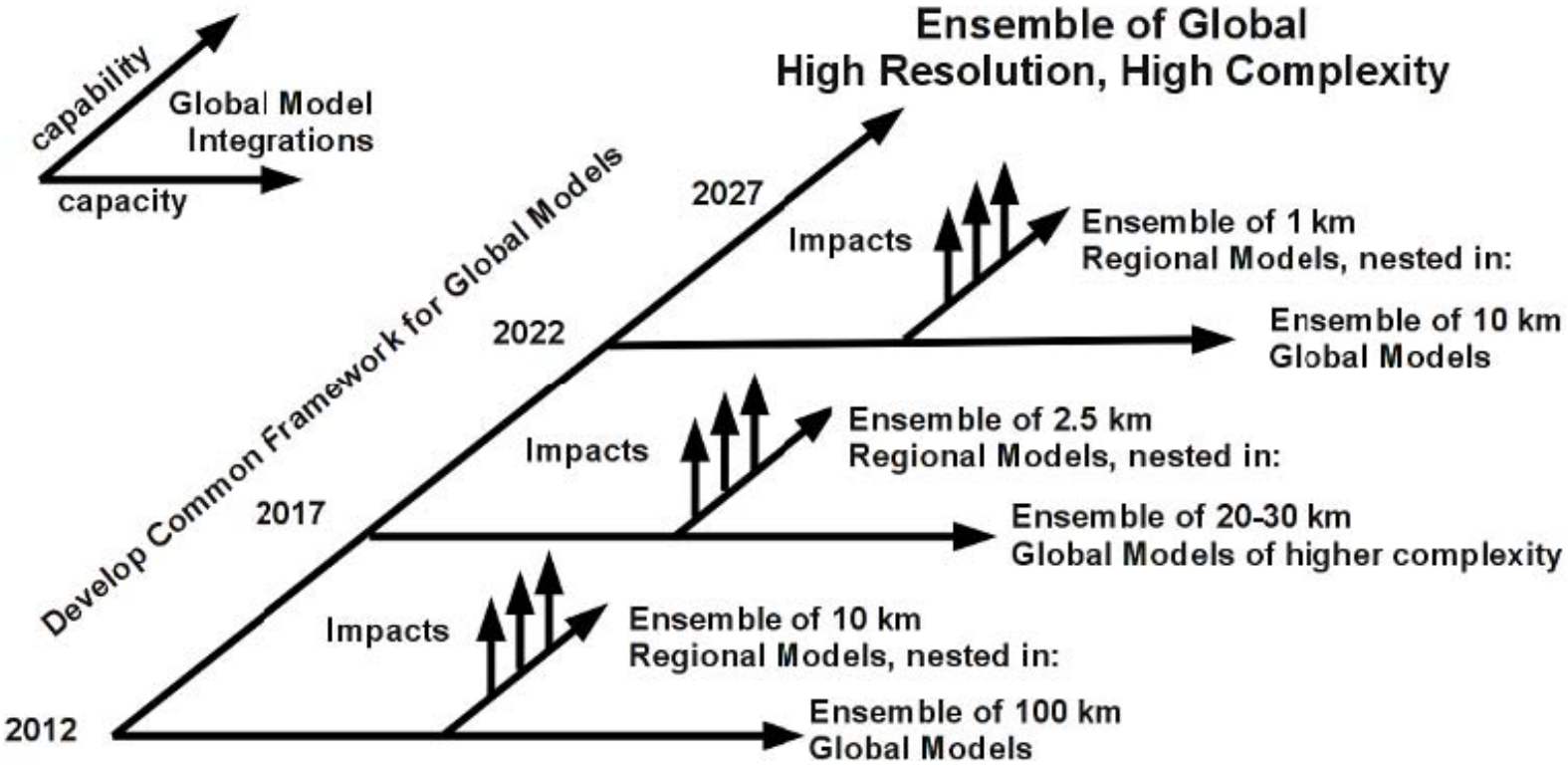



Science & Technology Facilities Council

# Outline



# Climate Simulation Roadmap



**British Atmospheric Data Centre**

NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE
NATURAL ENVIRONMENT RESEARCH COUNCIL

| | CMIP5 | CMIP6 | CMIP7 |
|---|---|---|---|
| Year | 2012 | 2017 | 2022 |
| Power factor | 1 | 30 | 1000 |
| Npp | 200 | 357 | 647 |
| Resolution [km] | 100 | 56 | 31 |
| Number of mesh points [millions] | 3.2 | 18.1 | 108.4 |
| Ensemble size | 200 | 357 | 647 |
| Number of variables | 800 | 1068 | 1439 |
| Interval of 3-dimensional output (hours) | 6 | 4 | 3 |
| Years simulated | 90000 | 120170 | 161898 |
| Storage density | 0.00002 | 0.00002 | 0.00002 |
| **Archive size (Pb) (atmosphere)** | **5.31** | **143.42** | **3766.99** |

$N_{pp}$ = Number of mesh points pole to pole

$N_g$ = Total number of spatial mesh points = $O(N_{pp}^3)$

$N_v$ = Number of variables ~ $\sqrt{N_{pp}}$

$N_e$ = Ensemble size ~ $N_{pp}$

$N_t$ = Time steps per simulated year ~ $N_{pp}$

$N_y$ = Years simulated per intercomparison ~ $\sqrt{N_{pp}}$

$$\text{Cost} \sim N_{pp}^6$$

O(40) decrease in storage density needed to bring this estimate in line with Overpeck et al.

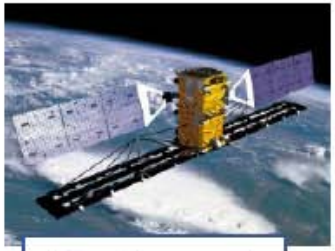
# Current and future EO missions (excerpt)

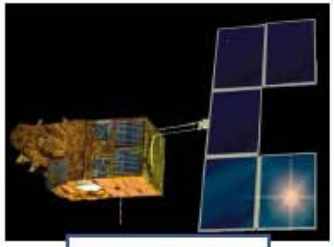

Envisat

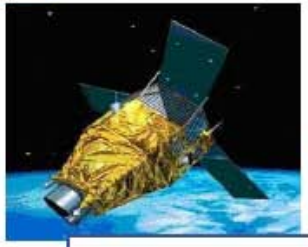
ERS-2


Explorers

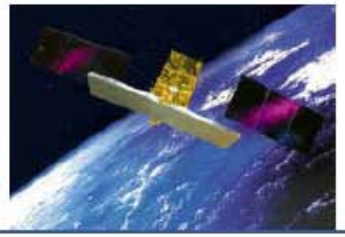

Sentinels


Radarsat

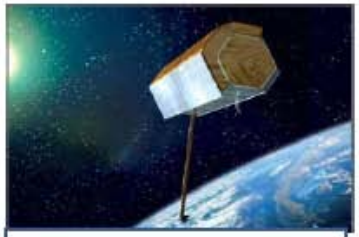

SPOT


Pleiades

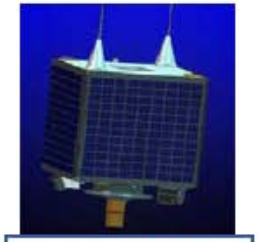

Jason-2


Cosmo-Skymed


ODIN


Terrasar-X

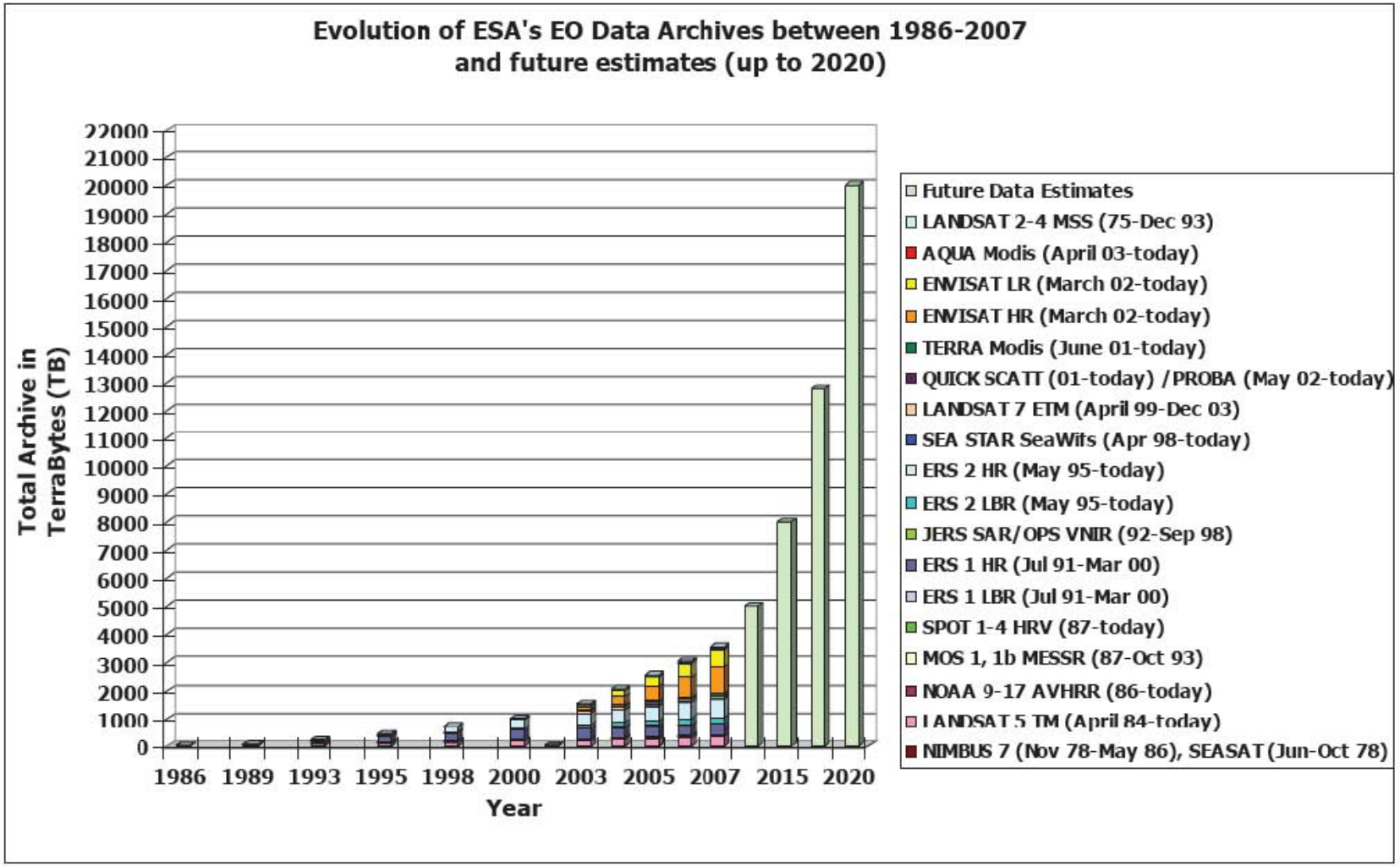

Rapideye


UK-DMC


METOP

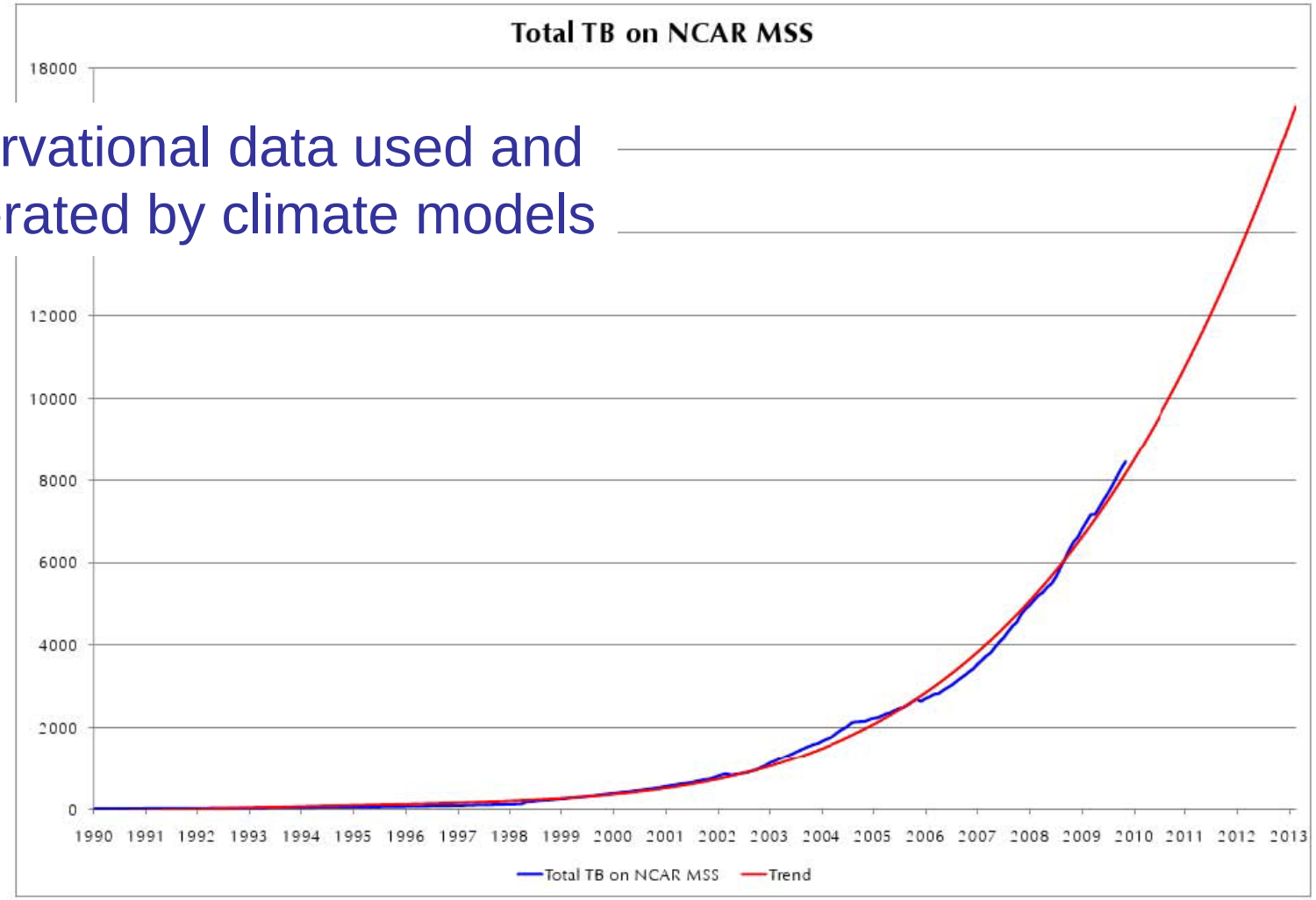
Science & Technology Facilities Council

# Earth Observation Data Archives



# NCAR's Mass Storage System

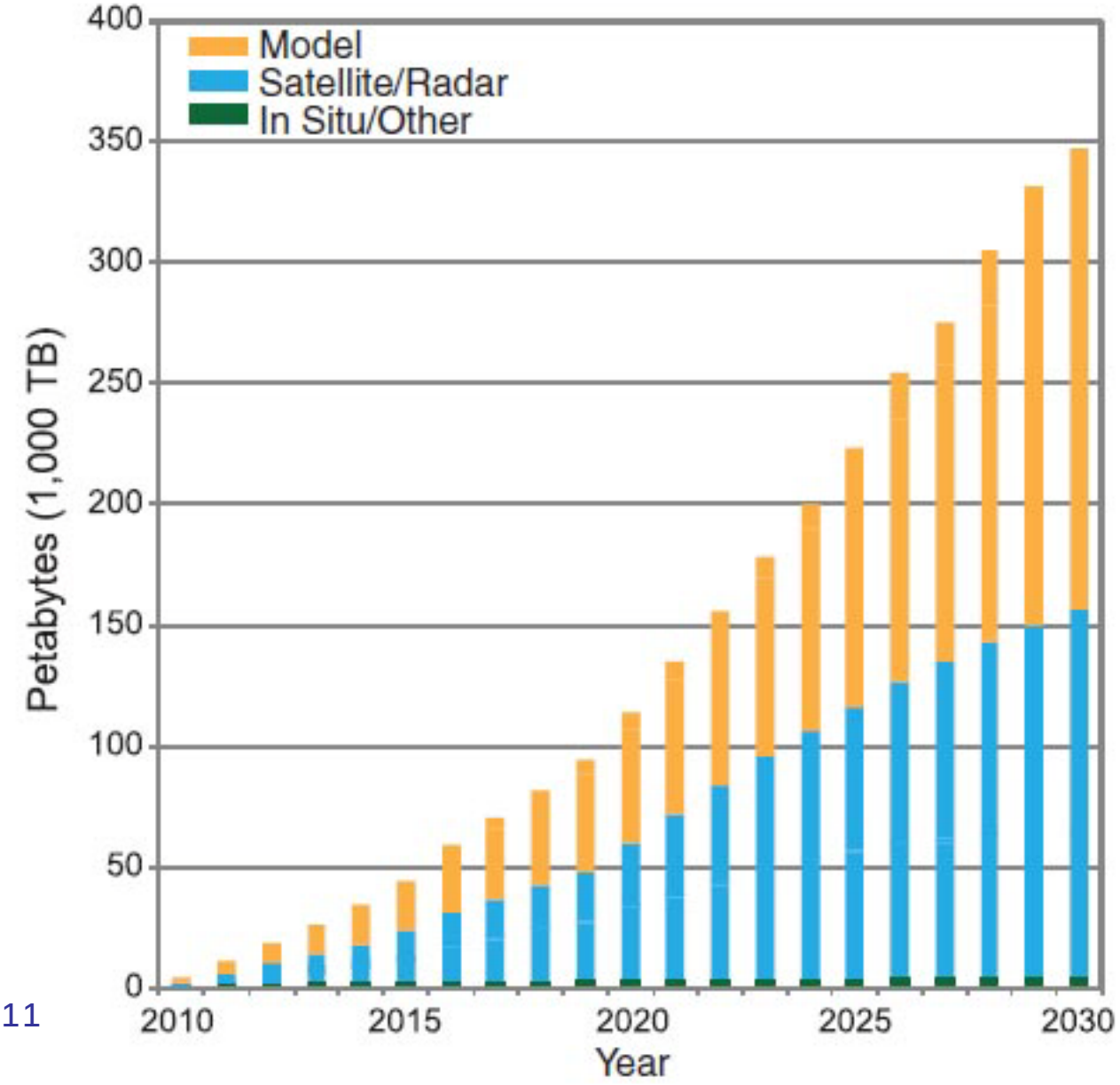
observational data used and generated by climate models



Science & Technology Facilities Council

# Simulation **AND** EO

projected increase in global climate data holdings for climate models, remotely sensed data, and in situ instrumental/proxy data




Source: Overpeck et al, Science, 2011

# Outline

- The Climate Data Deluge – Simulation and Earth Observation
- On The Path to Exascale
- ICE-CSE and Data-Intensive Computing

Science & Technology Facilities Council

# G8 project ExArch

## Climate Analytics on Distributed Exascale Data Archives

Martin Juckes (Lead PI, STFC, UK), V. Balaji, B.N. Lawrence, M. Lautenschlager, S , . Denvil, G. Aloisio, P. Kushner, D. Waliser, S. Pascoe, A. Stephens, P. Kershaw, F. Laliberte, J. Kim, S. Fiore [UK, US, France, Germany, Canada, Italy]

Research into exploitation of exascale computational resources

Focus on 10-year time horizon

Start: March 1st, 2011 Duration: 39 months

Budget: 1.44 million Euros Effort: 246 staff months

UCLA

Institut Pierre Simon Laplace

DKRZ

UNIVERSITY OF TORONTO

Princeton University

Centro Euro-Mediterraneo per i Cambiamenti Climatici

RAL Space

Harwell International Space Innovation Centre

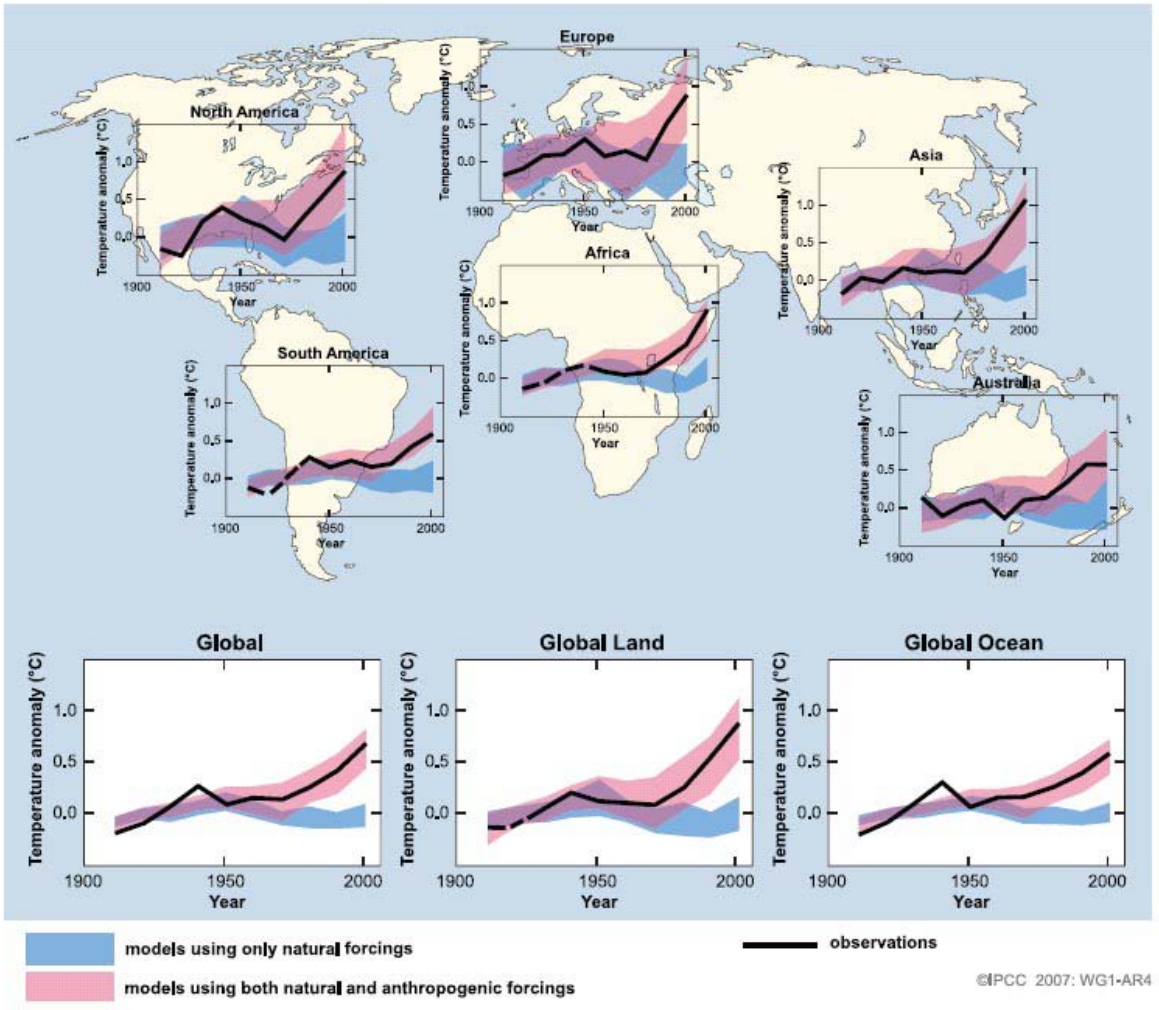
# G8 project ExArch

## ExArch Science Driver

Uncertainty at Regional Scale:
Need for regional scale policy information from global-scale research

## Elements of ExArch

- Query Syntax
- Web Processing Service
- Common Information Mode
- Processing Operators & Quality Control
- Scientific Diagnostics
- Earth Observation Data for Model Evaluation
- Grid Computing




Reference: Figure from 2007 IPCC SPM

# System Evolution to Exascale

| Systems | 2011 K computer | 2019 | Difference Today & 2019 |
|---|---|---|---|
| **System peak** | **10.5 Pflop/s** | **1 Eflop/s** | **O(100)** |
| **Power** | **12.7 MW** | **~20 MW** | |
| System memory | 1.6 PB | 32 - 64 PB | O(10) |
| Node performance | 128 GF | 1,2 or 15TF | O(10) – O(100) |
| Node memory BW | 64 GB/s | 2 - 4TB/s | O(100) |
| Node concurrency | 8 | O(1k) or 10k | O(100) – O(1000) |
| Total Node Interconnect BW | 20 GB/s | 200-400GB/s | O(10) |
| System size (nodes) | 88,124 | O(100,000) or O(1M) | O(10) – O(100) |
| Total concurrency | 705,024 | O(billion) | O(1,000) |
| MTTI | days | O(1 day) | - O(10) |

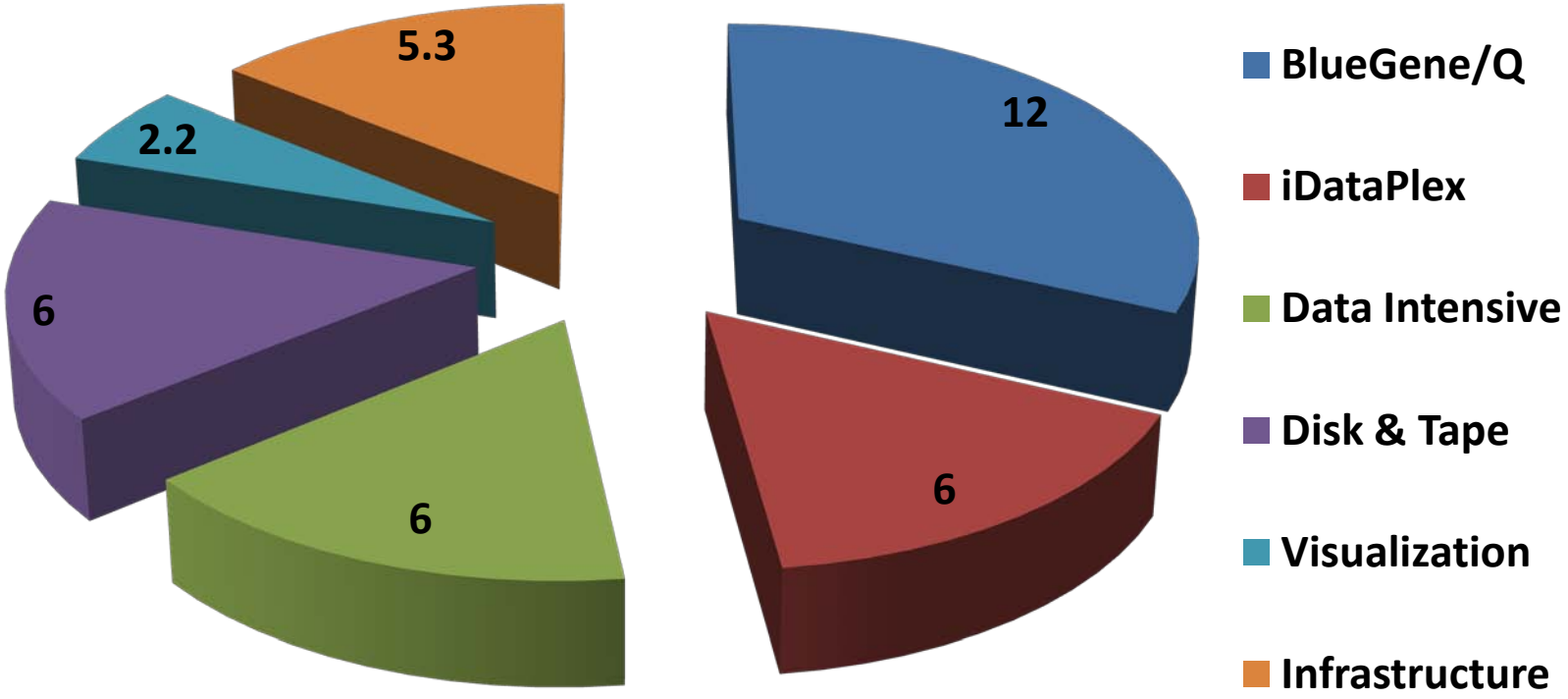
# Outline

- The Climate Data Deluge – Simulation and Earth Observation
- On The Path to Exascale
- ICE-CSE and Data-Intensive Computing

# ICE-CSE
# International Centre of Excellence in Computational Science and Engineering

- David Cameron confirmed £10M investment into STFC's Daresbury Laboratory. £7.5M of this will be used to upgrade the Campus computing infrastructure
- Chancellor announced £145M for e-infrastructure at the Conservative Party Conference
- David Willetts visited the next day and indicated £30M further investment in CSED

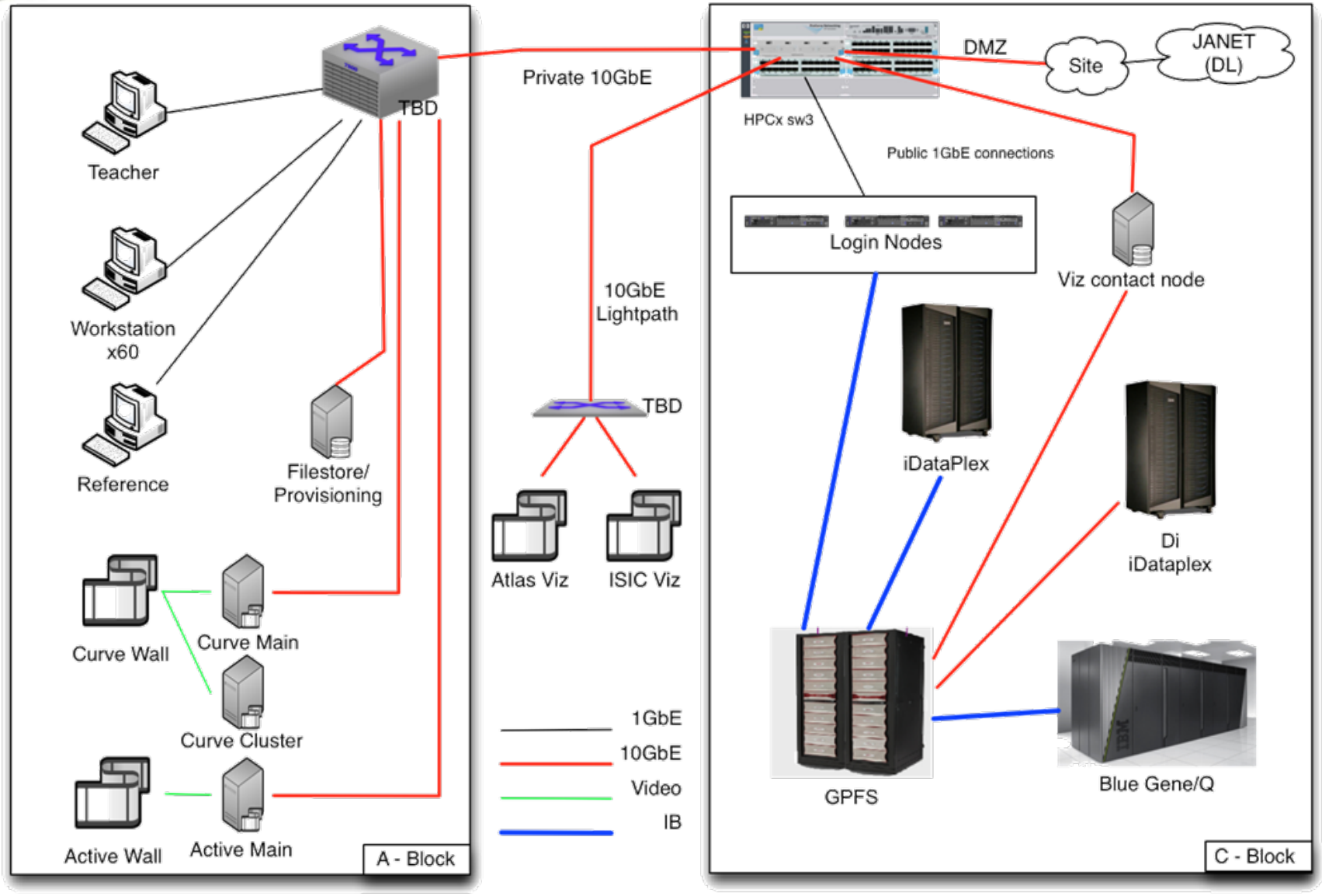
# ICE-CSE capital spend 2011/12




approximate capital spend £M

Total £37.5M

Science & Technology Facilities Council



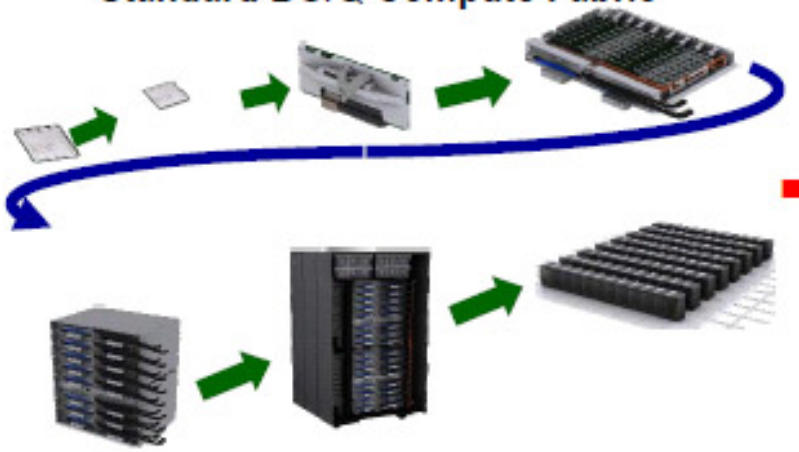
# ICE-CSE Data-intensive Projects

- Climate simulations using high-resolution ensembles for uncertainty estimation
- SKA - development and demonstration of prototype software for the SDP
- Bio – large-scale mining of genomics and other *omics data sets
- CFD – analyses of data bases of turbulent flow data generated from large-scale Direct Numerical Simulation

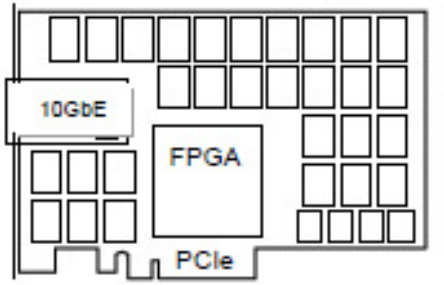
# Blue Gene Active Storage Concept: BG/Q + Solid State Storage

IBM

### Standard BG/Q Compute Fabric



### Solid State Storage Device – PCIe Flash




| Example PCIe Flash | 2012 Targets |
|---|---|
| Flash Capacity | 2 TB |
| I/O Bandwidth | 2 GB/s |
| IOPS | 200 K |

**Recipe:**

- **Remove 512 BG/Q compute nodes**
- **Add 512 PCIe SSD Cards**

### Target Applications

- Parallel File and Object Storage Systems
- Graph-based algorithms
- Join
- Sort
  - "order by" queries
- "group by" queries
- Map-Reduce (heavy reduce phase)
- Aggregation operations
  - count(), sum(), min(), max(), avg(), …
- Data analysis/OLAP
  - Aggregation with "group by"…
  - Real-time analytics

Source: Robert Germain, IBM

**Key architectural balance point: All-to-all throughput roughly equivalent to Flash bandwidth**

### BG/Q Active Storage Rack

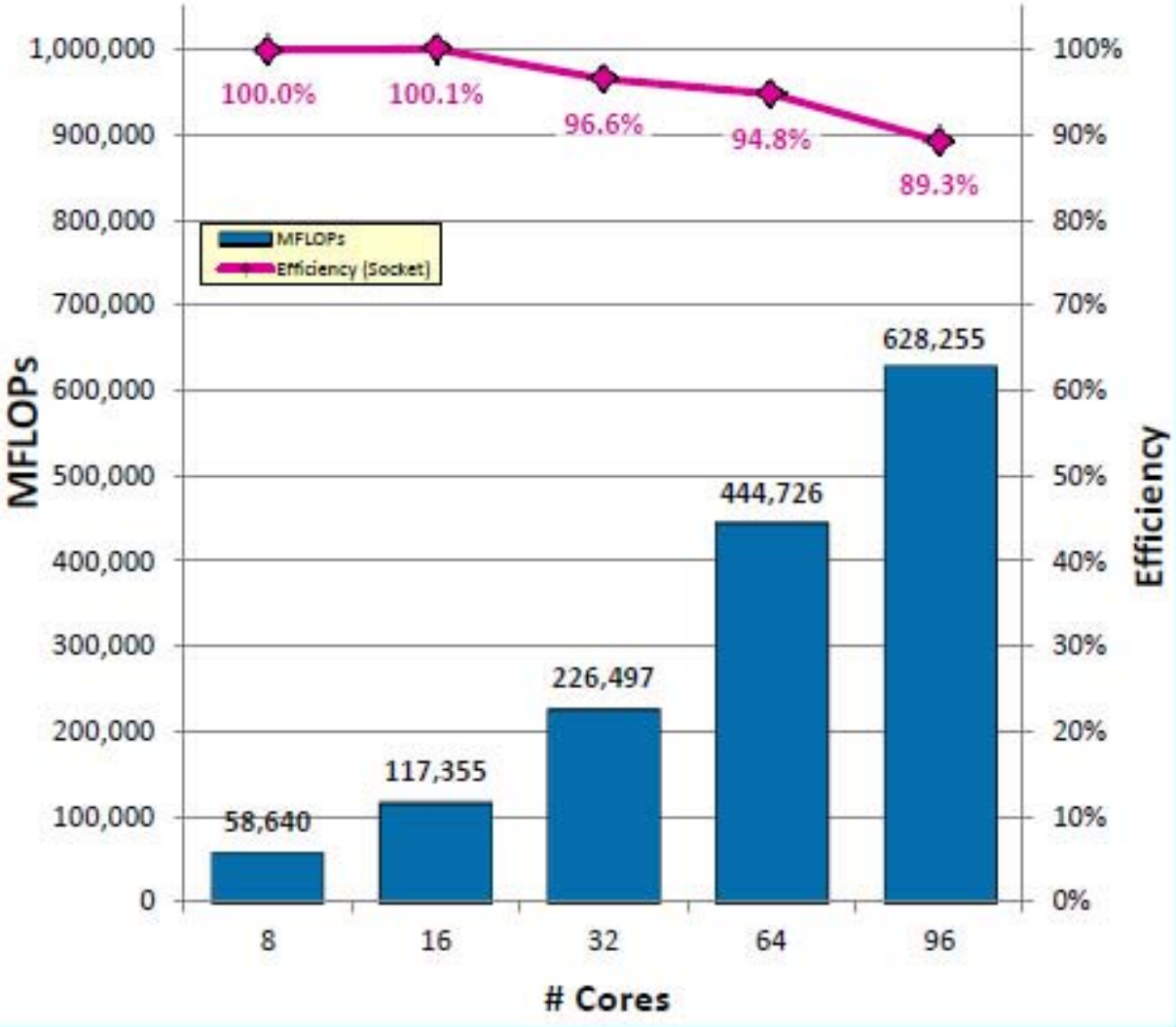
| Nodes | 512 |
|---|---|
| Storage Cap | 1 PB |
| I/O Bandwidth | 1 TB/s |
| Random IOPS | 100 Million |
| Compute Power | 104 TF |
| Network Bisect. | 512 GB/s |

… scale it like BG/Q.

# OPENMP PARALLELIZATION (3)

Last Update: 1/18/2011

## DGEMM (INTEL MKL) – MATRIX SIZE: 25,000 X 25,000 - INTEL NEHALEM EX



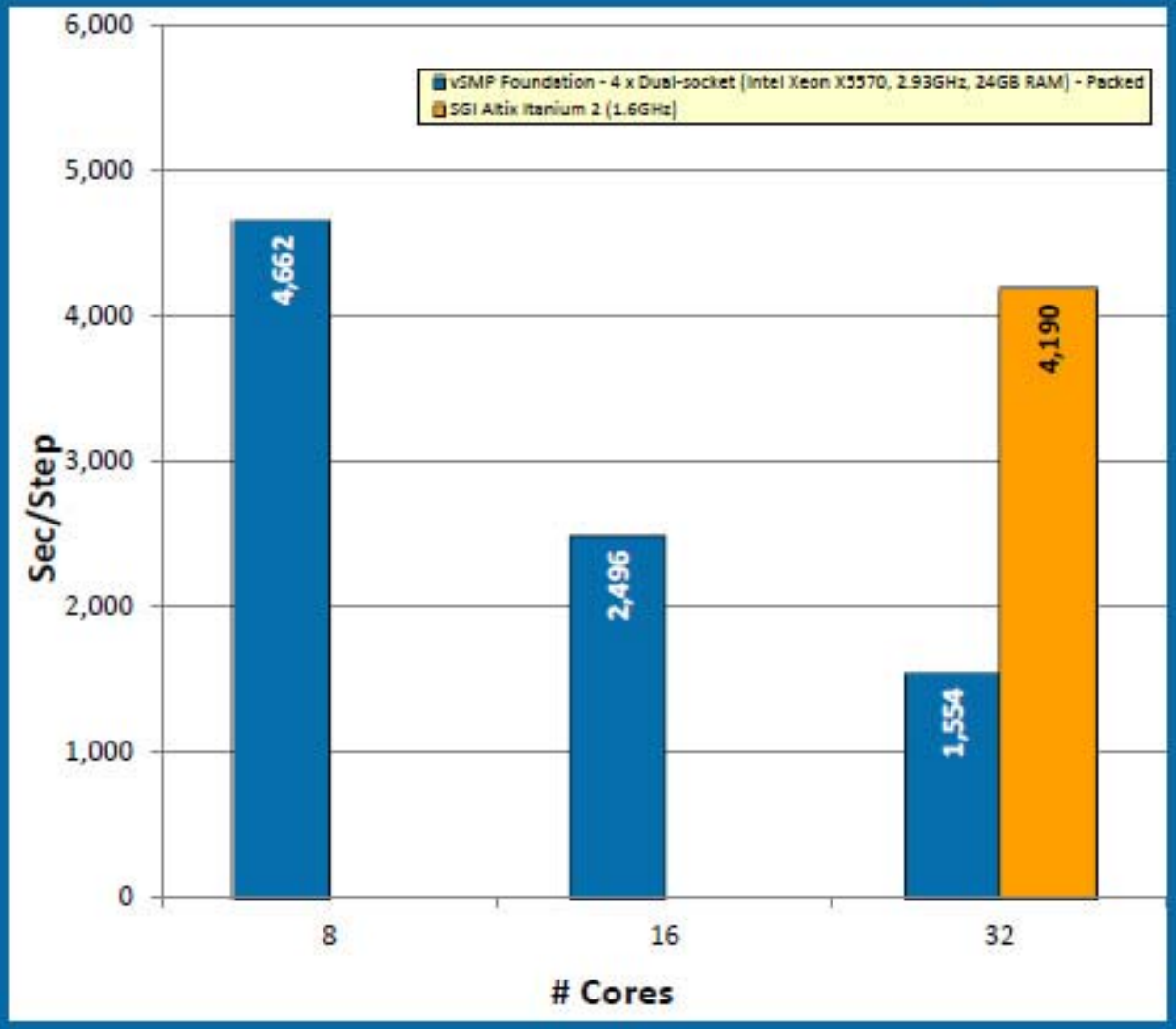

- MKL is Intel's Math Kernel Library, which is using threads for parallelization and is the corner stone for many applications.
- DGEMM is the Matrix Multiply function which is the base for many numerical algorithms.
- vSMP Foundation demonstrates about 90% efficiency scaling across 12 sockets.
- Intra-board efficient (8 to 32 cores) is lower than inter-board efficiency when using 8-cores system (previous chart)
- System configuration:
  - 3 X Quad-socket servers (Intel Xeon X7550 @ 2.00 GHz, 128 GB RAM), HT off, Turbo Boost not enabled

Threaded

# MM5

Last Update:
9/30/2010

## CUSTOMER BENCHMARK: 3KM RESOLUTION (X=283 , Y=253, LEVELS=31)




- Performance comparison of:
  - vSMP Foundation: 4 nodes
  - SGI Altix
- vSMP Foundation demonstrates:
  - 75% efficiency with 32 cores
  - 2.75 X faster than SGI Altix
- System configuration:
  - vSMP Foundation: 4 X Dual-socket servers (Intel Xeon X5570 @ 2.93 GHz, 24 GB RAM)
  - SGI Altix: Itanium 2 @ 1.6 GHZ

Throughput / MPI

# Simulation vs. Archive

- **Simulation**

  Huge datasets are required physically close to the HPC systems for assimilation of observational data and storage & analysis of simulation outputs during and shortly after simulation runs

  e.g. HECToR, ARCHER, ICE-CSE

- **Archive**

  Huge datasets are required at one or more data centres for long-term archive, retrieval and analysis

  e.g. BADC, BODC

## These are separable requirements

# European Exascale Software Initiative (EESI)

Recommendation from WG4.4 on Scientific Software Engineering (chair MA)


www.eesi-project.eu

- Development of a flexible generic I/O layer that can be used by applications to interface with either the storage system or the data analysis system. This layer should then be extended with advanced data reduction techniques to carry out **in- situ** domain-specific **data reduction and feature extraction**

# Thank you for your attention

Mike Ashworth

mike.ashworth@stfc.ac.uk