Exascale I/O challenges for Numerical Weather Prediction

A view from ECMWF

Tiago Quintino, B. Raoult, S. Smart, A. Bonanni, F. Rathgeber, P. Bauer, N. Wedi

ECMWF

tiago.quintino@ecmwf.int

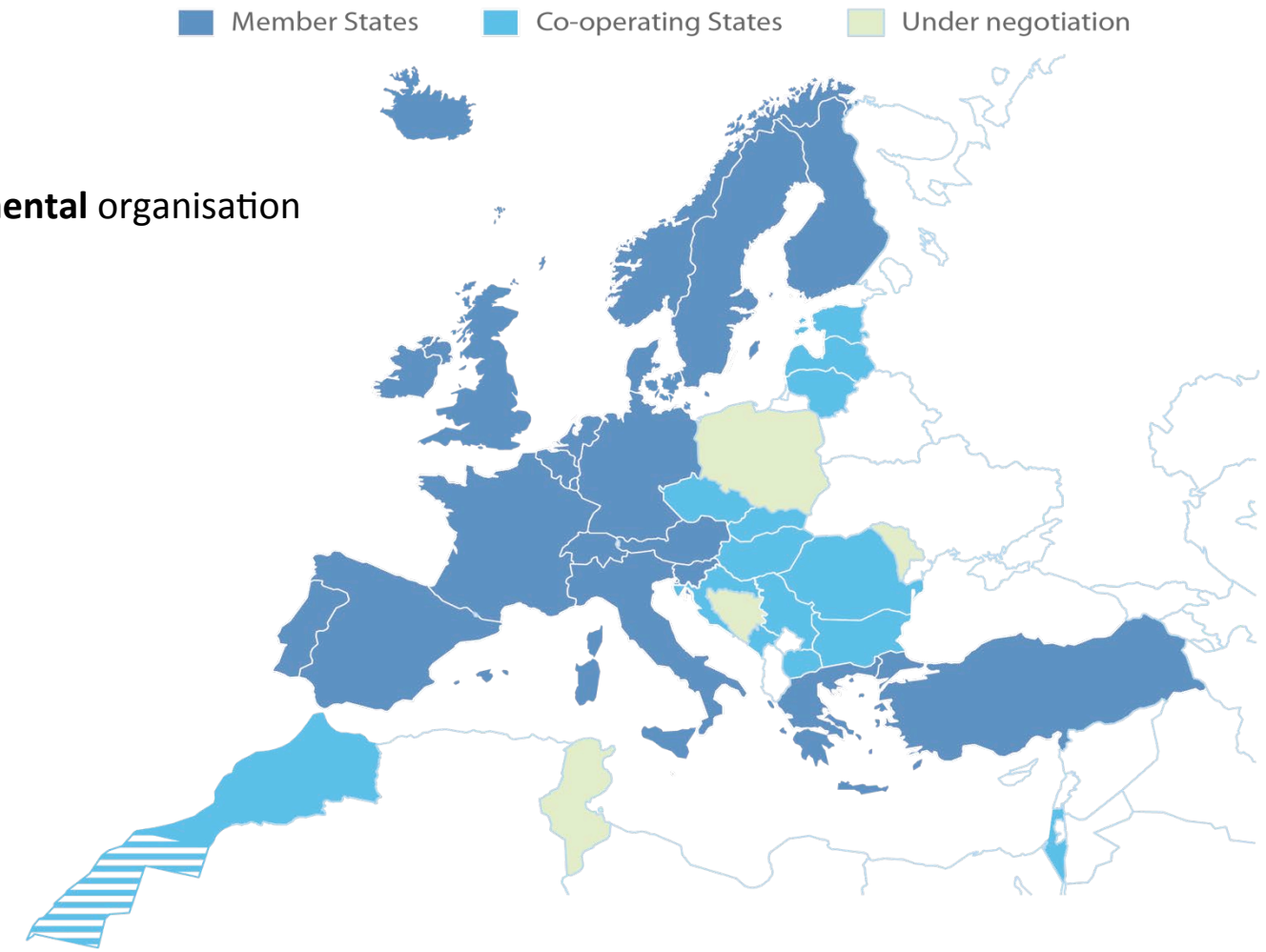
SuperComputing'16, Workshop on Exascale IO Challenges, Innovations and Solutions

ECMWF

ECMWF

An independent **intergovernmental** organisation

21 Member States
13 Co-operating States



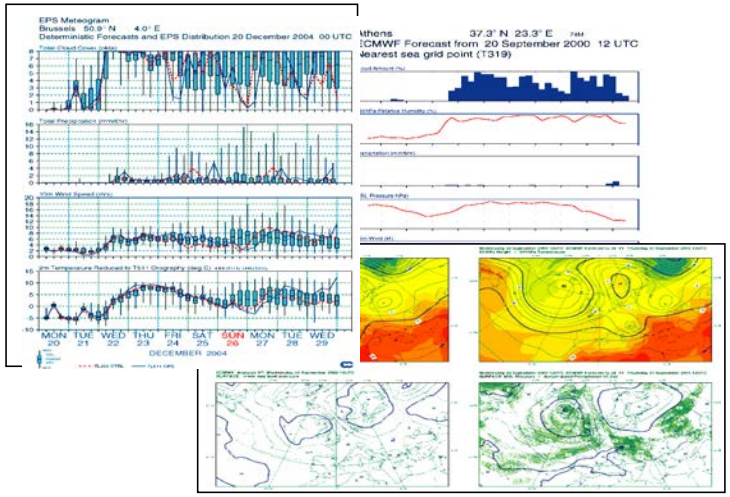
Who are we and what do we do?

- Produce **global weather forecasts**
- Medium-Range, **up to 15 days ahead**
- Also **monthly** and **seasonal** forecasts
- Collect and store meteorological data on perpetual archive

Reading, United Kingdom

What do we have to achieve this?

- 260 staff, specialists and contractors
- State-of-the-art supercomputers and data handling system

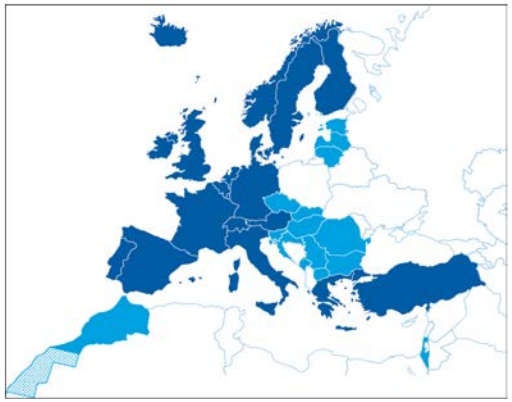


Numerical Weather Prediction @ ECMWF

Global observation system

Global numerical weather forecasts

National weather services



Users

ECMWF's HPC Targets

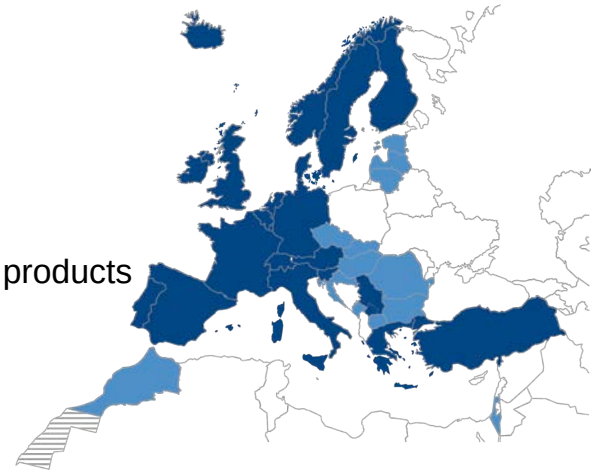
What do we do?

Operations – **Time Critical**

- Operational runs – 2 hours from observation cut-off to deliver forecast products
- 10 day forecast twice per day, 00Z and 12Z
- Boundary Conditions 06Z and 18Z, monthly, seasonal, etc.

Research – **Non Time Critical**

- Improving our models
- Climate reanalysis, etc



HPC Facility Targets

- **Capability**, minimise the time to solution of Model runs
- **Capacity**, maximise the throughput of research jobs per day

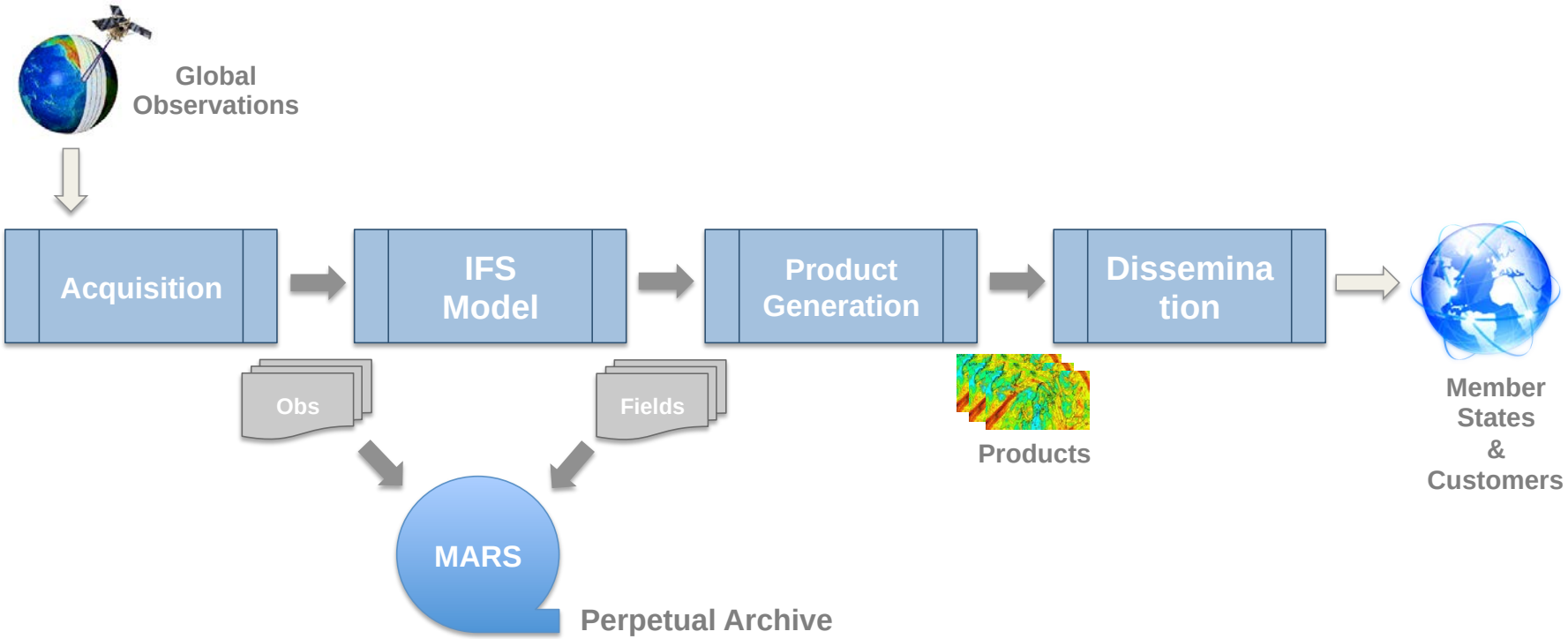
Tension

Time Critical vs. **Non Time Critical**

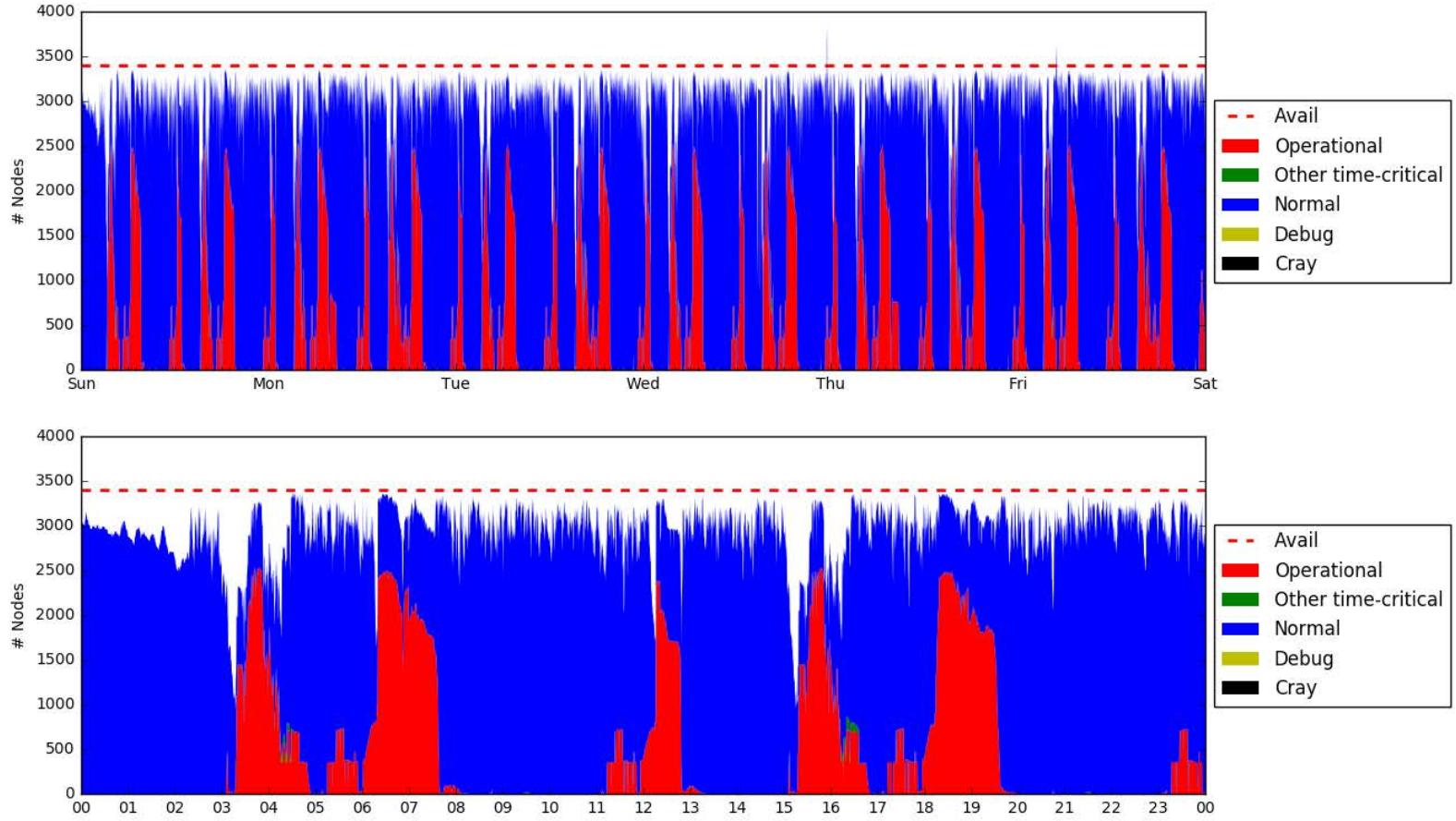
Capacity vs. **Capability**

***Challenge**: design our HPC system to optimise these goals, minimising TCO?*

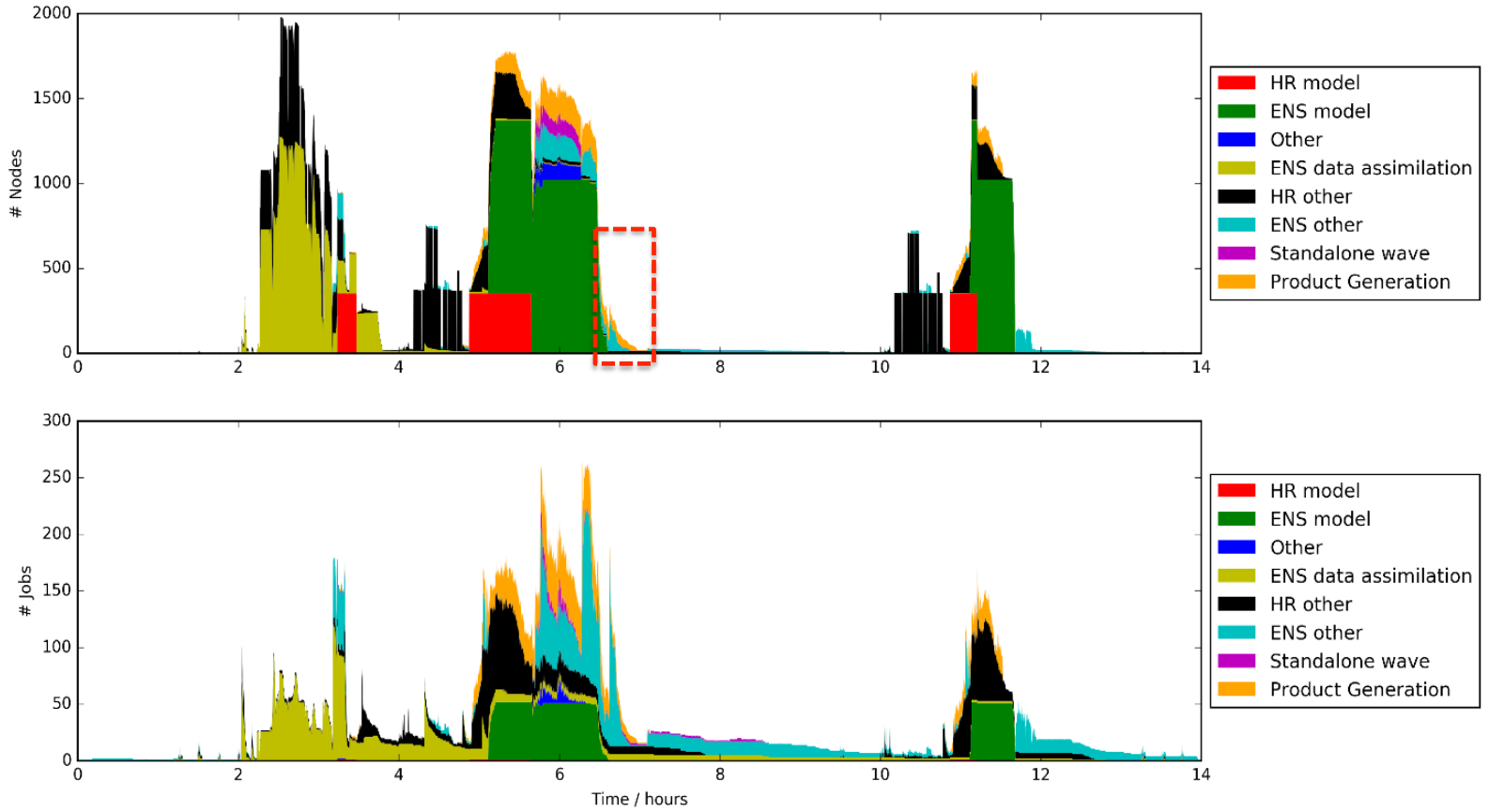
ECMWF's Production Workflow

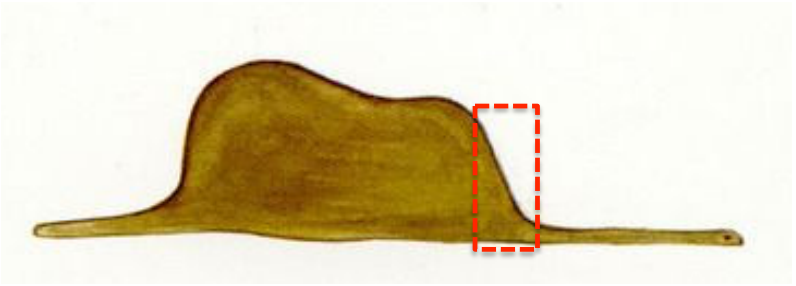


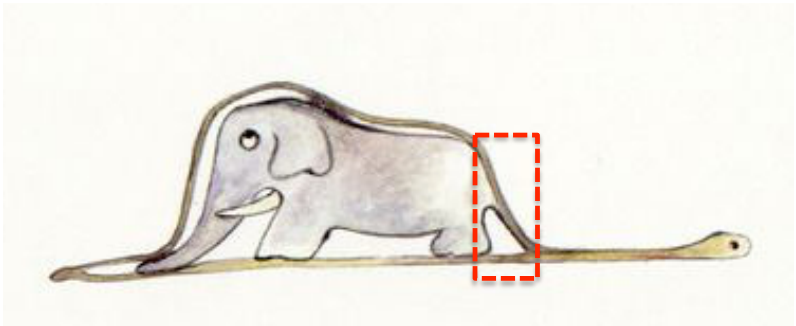
ECMWF HPC Job profile



Operational workload: Job allocation (1 cycle)

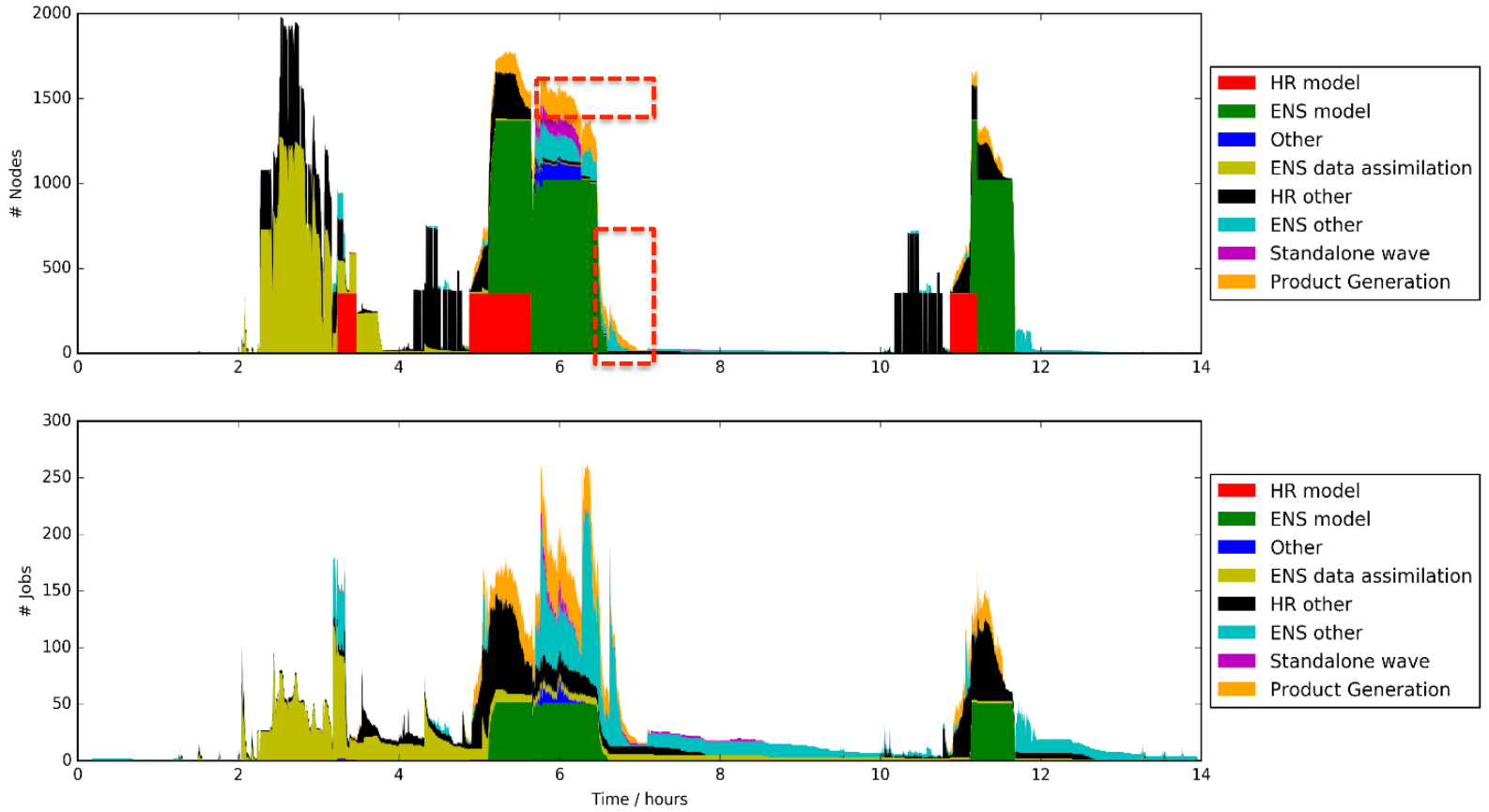




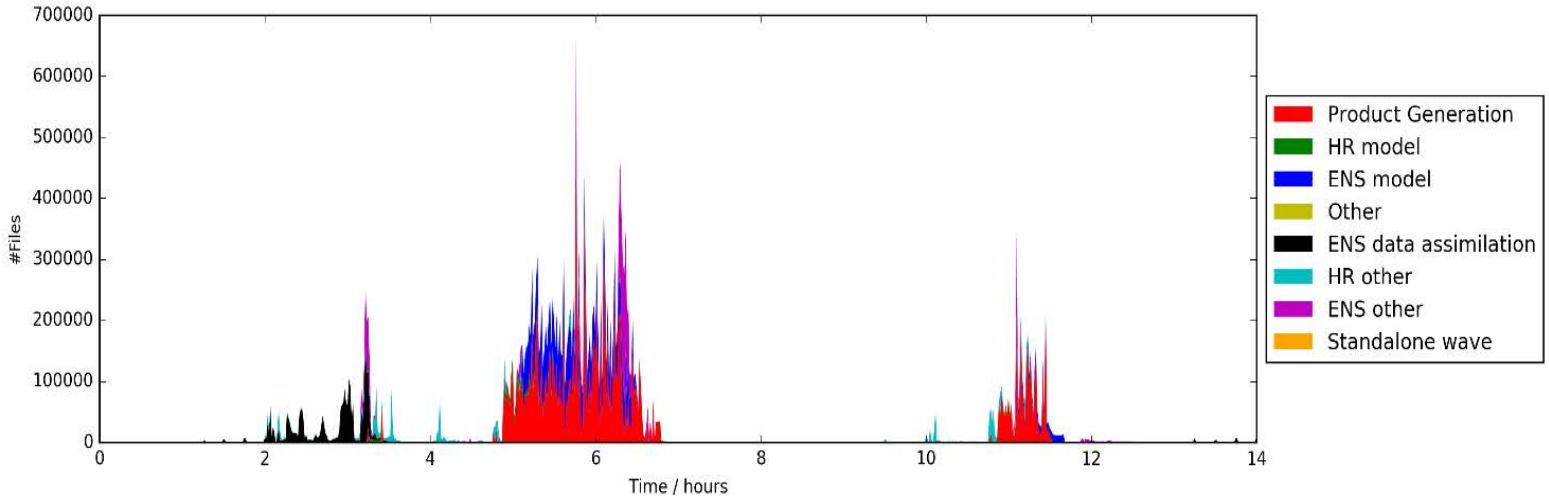


Le Petit Prince, Antoine de Saint-Exupéry

Operational workload: Job allocation (1 cycle)

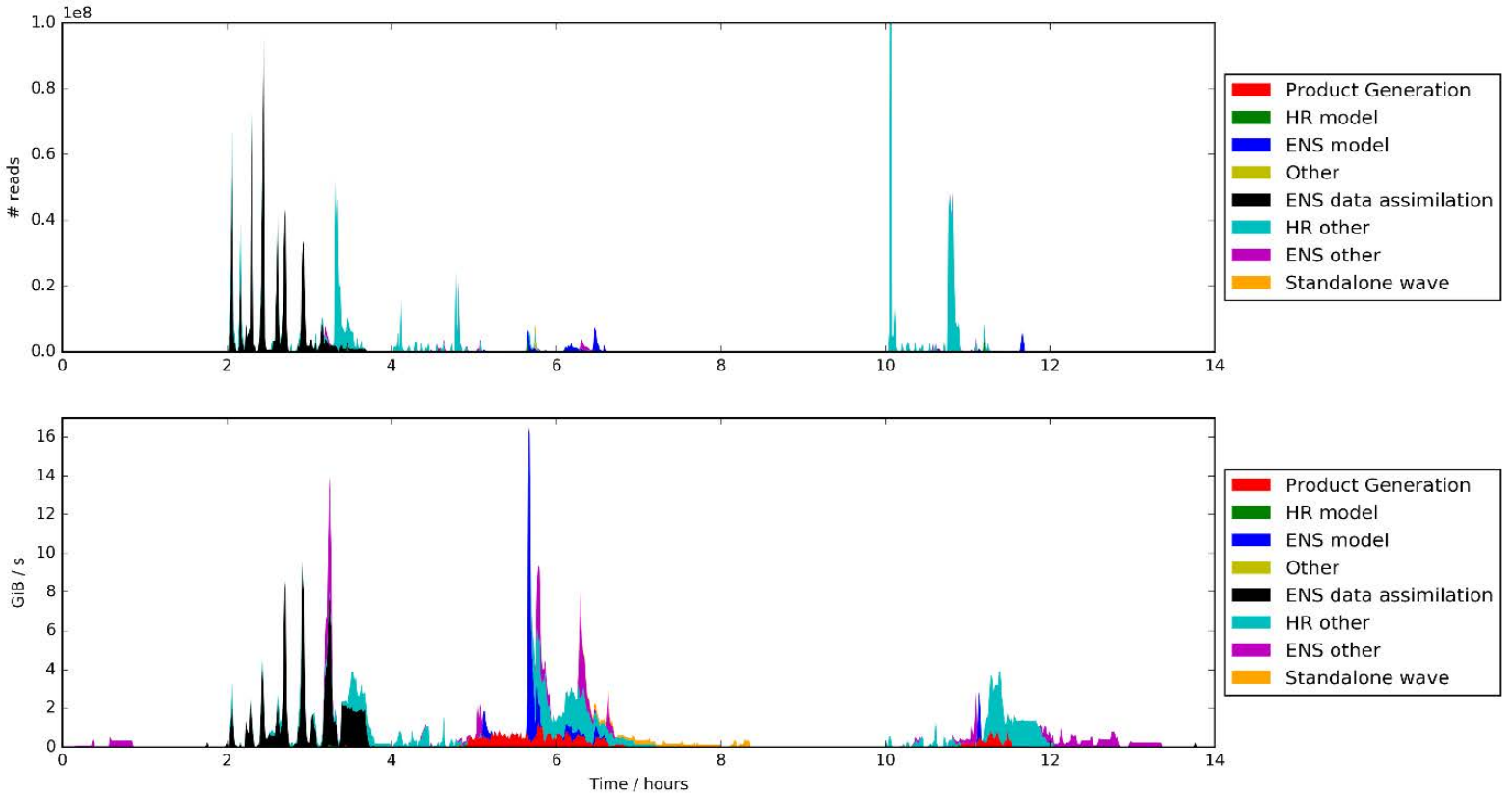


Operational workload: Files opened (1 cycle)

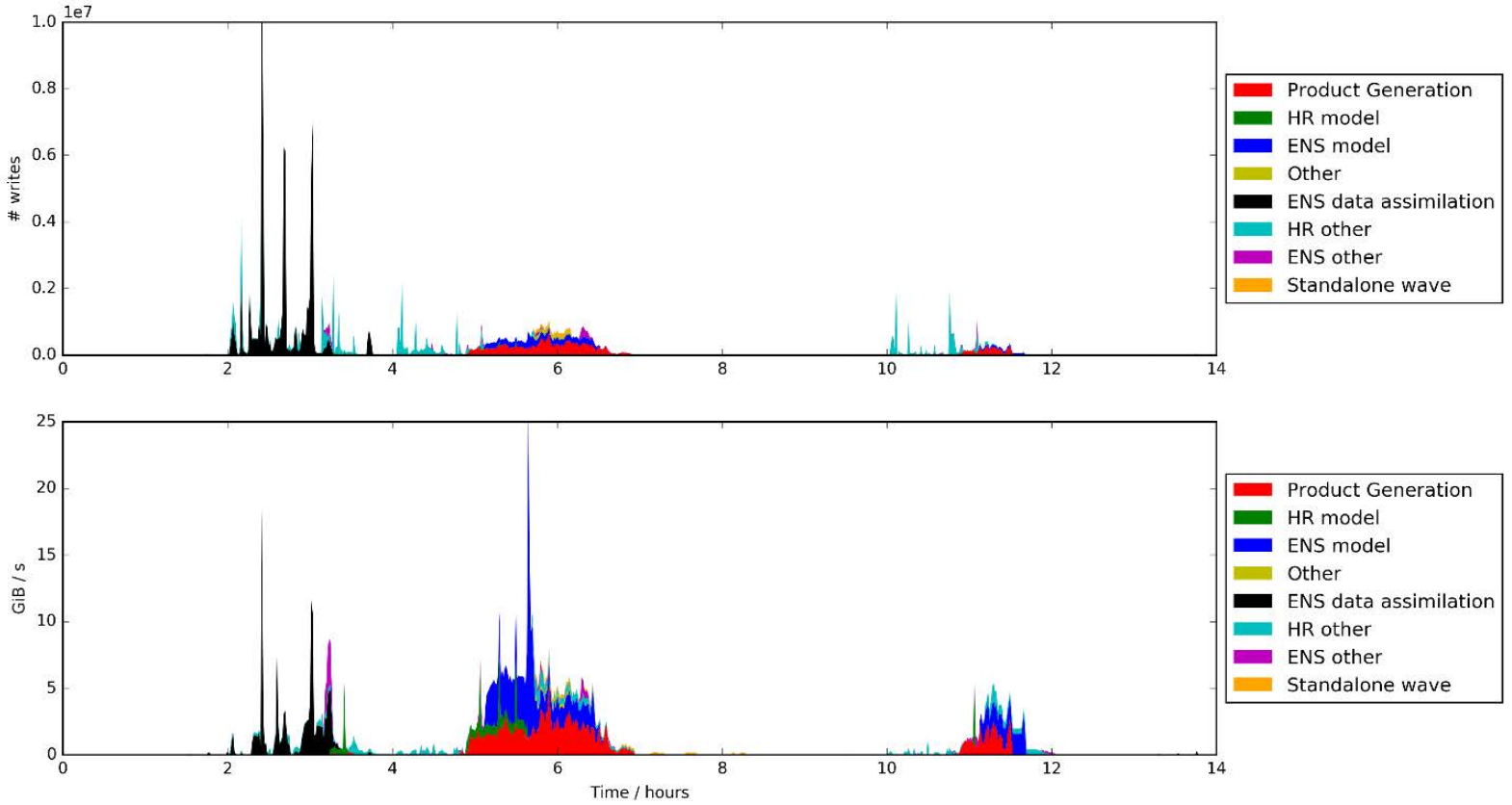


Target Files = # Users x # Steps x # Ranks

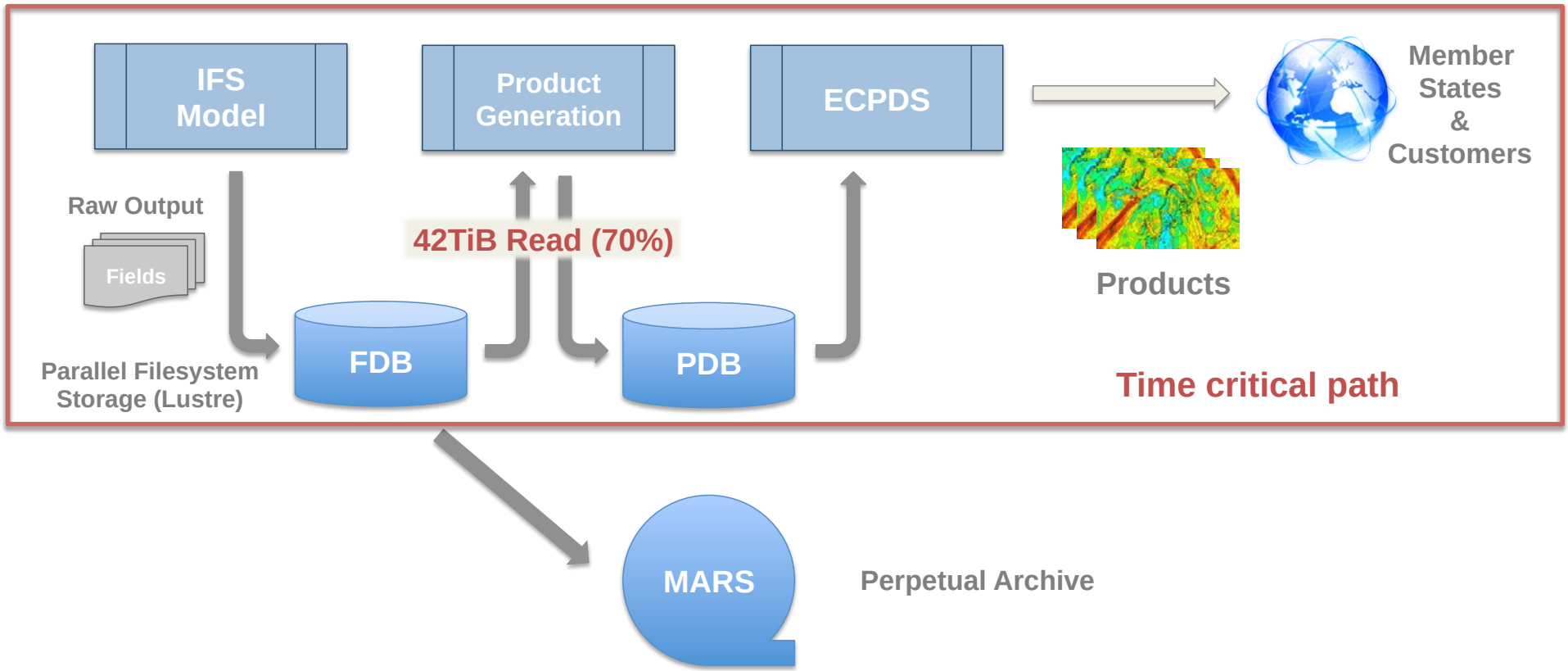
Operational workload: Input Read (1 cycle)



Operations workload: Output written (1 cycle)



ECMWF's Production Workflow



Estimated Growth in Model IO

2015

16km, 137 levels

Time critical

- 21 TB/day written
- 22 Million fields
- 85 Million products
- 11 TB/day send to customers

Non-time critical

- 100 TB/day archived
- 400 research experiments
- 400,000 jobs / day

2020

Increase: 2 horizontal, 1 upper air

Time critical

- 128 TB/day written
- 90 Million fields
- 450 Million products
- 60 TB/day send to customers

Non-time critical

- 1 PB/day archived
- 1000 research experiments

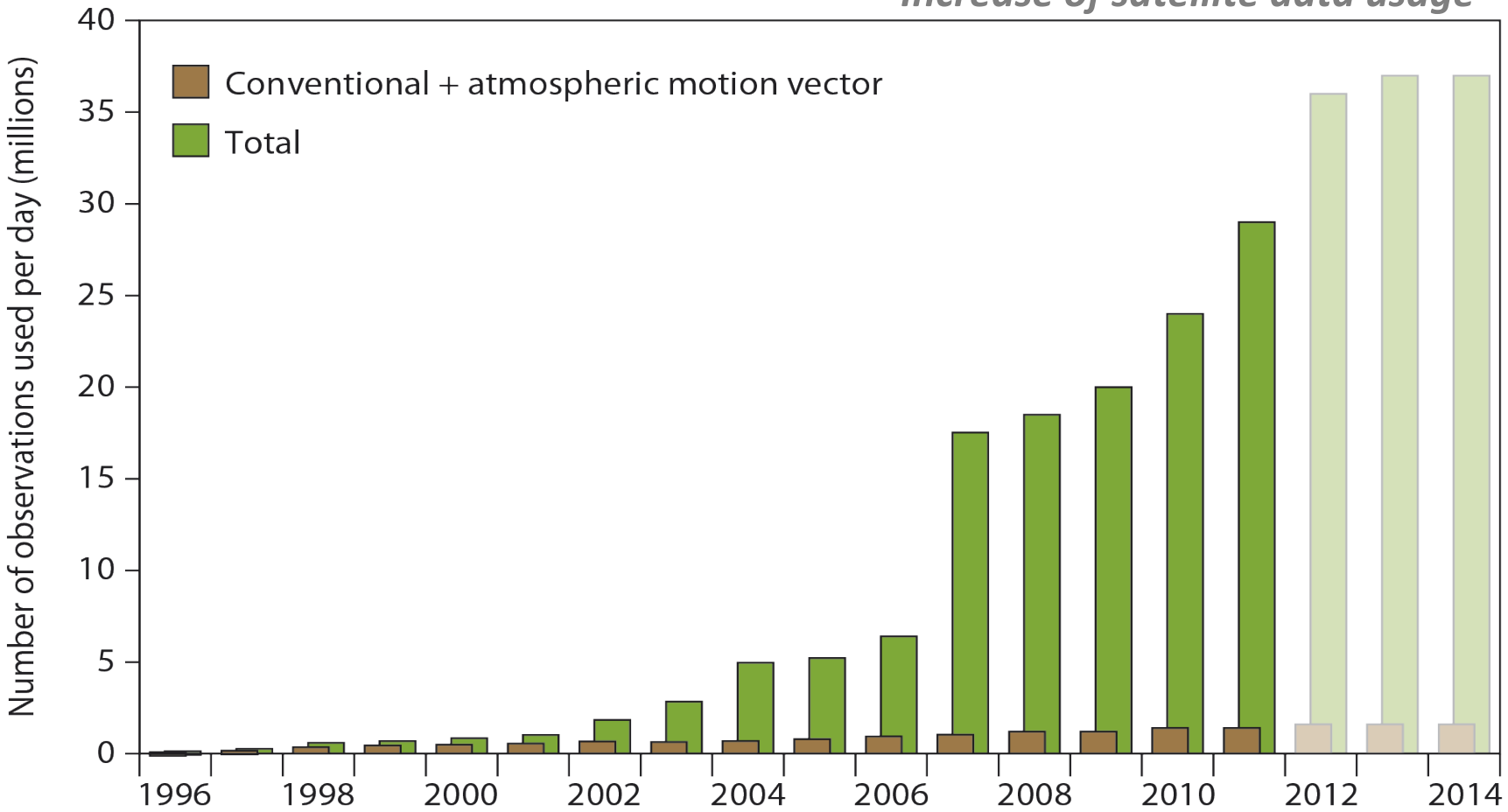
The 3 V's of Big Data

"Big Data is high **volume**, high **velocity**, and/or high **variety** information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization."

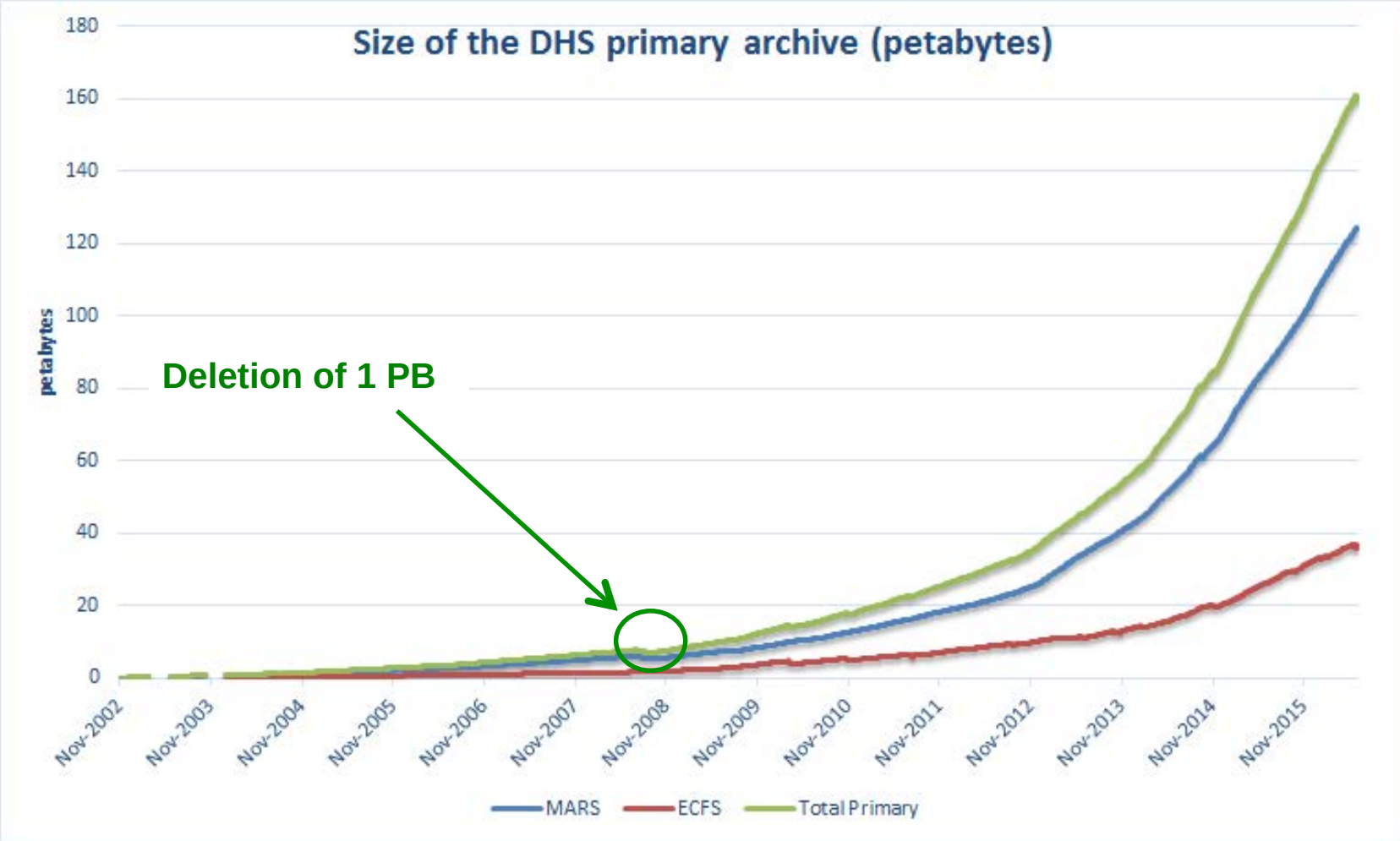
"3D Data Management: Controlling Data Volume, Velocity and Variety", D. Laney, Gartner, 2001

V is for Volume: Observations

Increase of satellite data usage

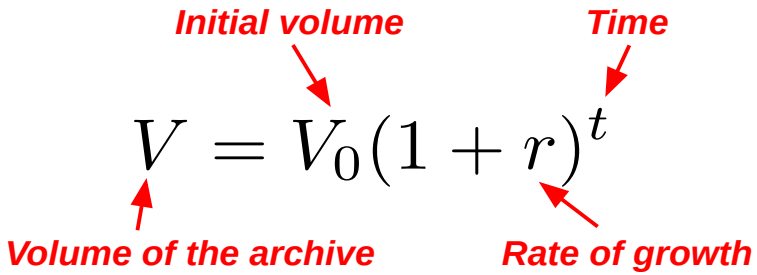


V is for Volume: Archive



V is for Velocity

- ECMWF's archive grows exponentially:

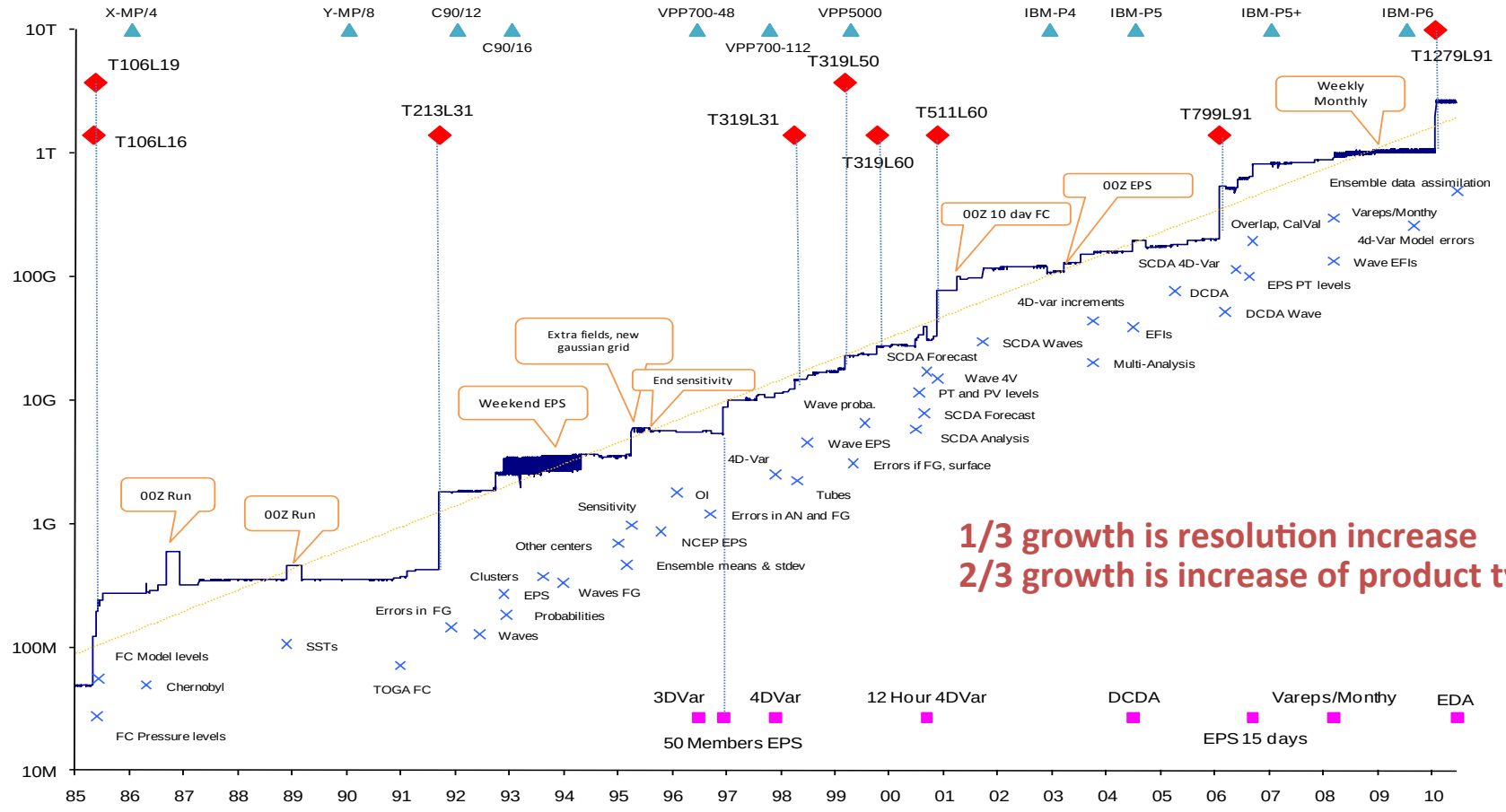


  - r is around 0.5, which is a 50% increase per year
  - The rate of added data also grows exponentially at the same rate!

$$\frac{\partial V_0(1+r)^t}{\partial t} = V_0 log(1+r)(1+r)^t = A_0(1+r)^t$$

- In 1995, the size of the archive was increasing at a rate of **14 TB/year**
- Nov 2016, the size of the archive increases at a rate of **150 TB/day**

V is for Variety



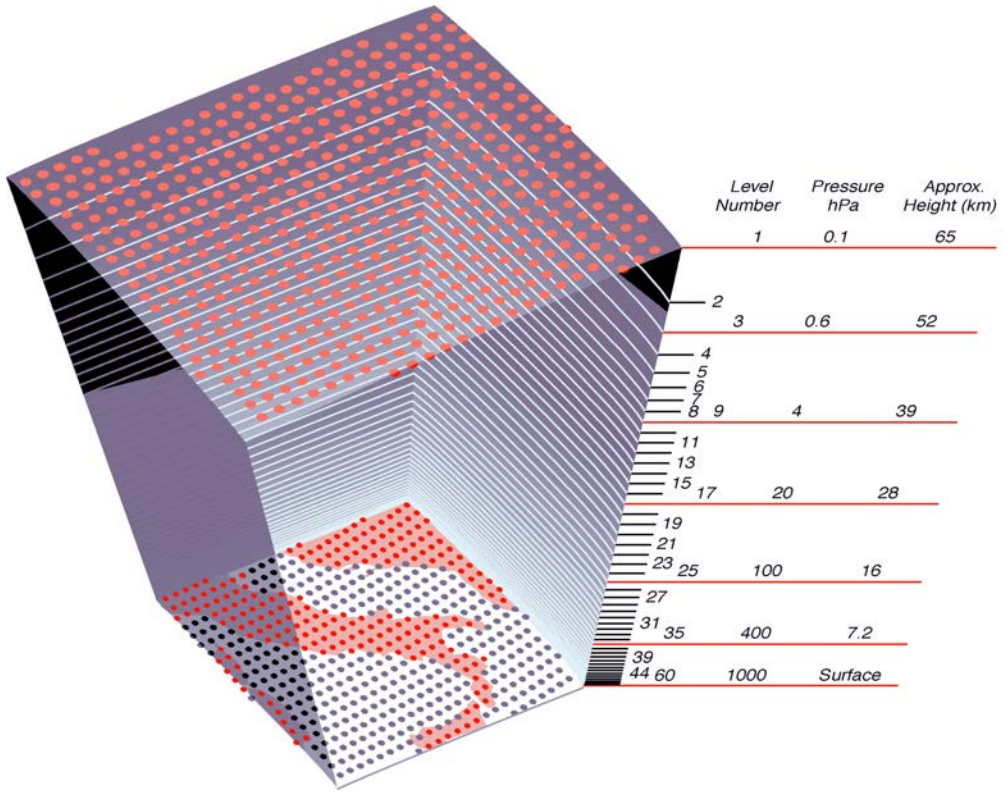
1/3 growth is resolution increase
2/3 growth is increase of product types

Meteorological Fields

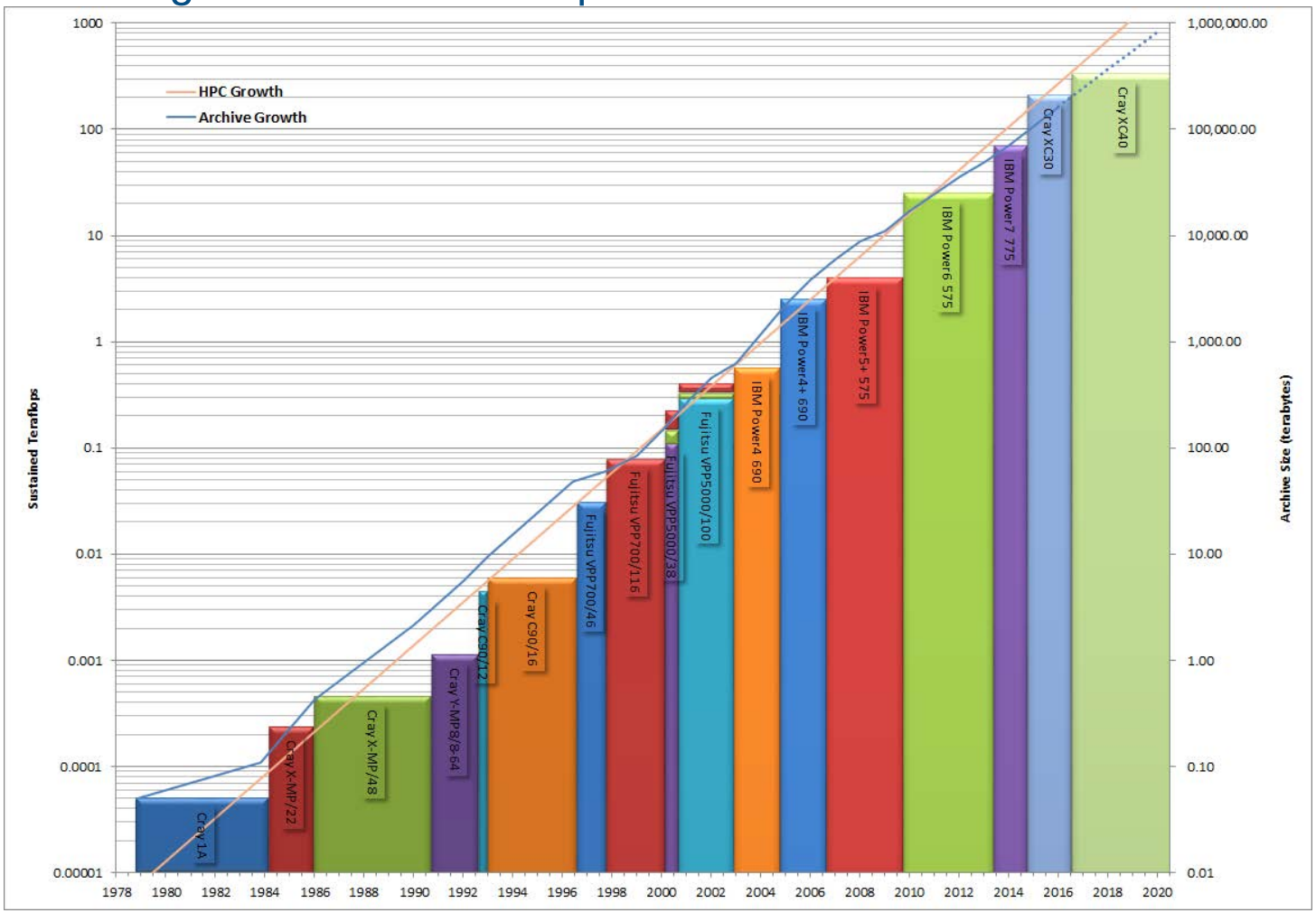
Fields:
- 6.6M grid points x 137 levels
- 904M values, **per variable**

Daily Operational Cycle:
- 18M fields
- 27 TiB output

Time critical window: 1h !!!



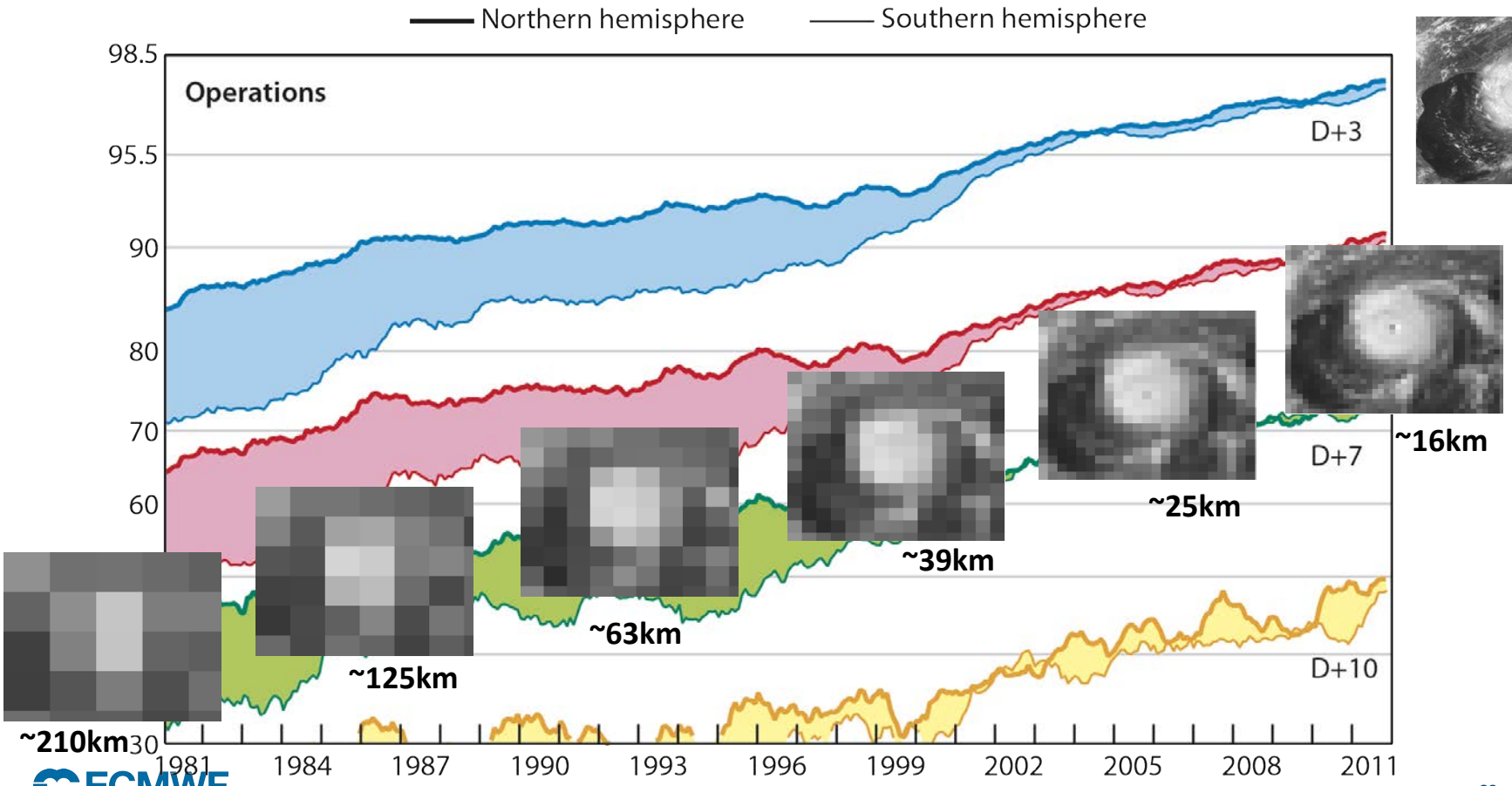
Archive growth versus HPC performances



Log scale

Evolution of Forecast scores, comparison northern and southern hemispheres

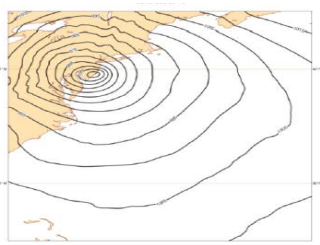
Anomaly correlation of 500 hPa height forecasts



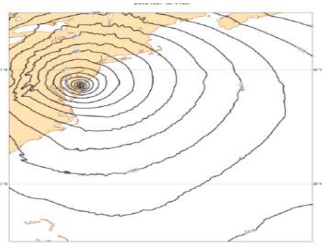
Benefits of High Resolution

Mean sea-level pressure

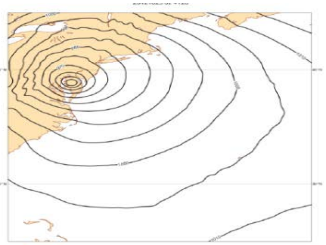
AN 30 Oct



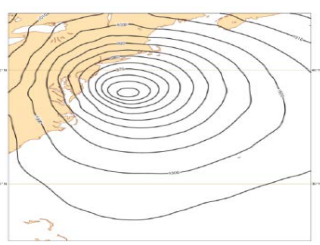
5d FC T3999



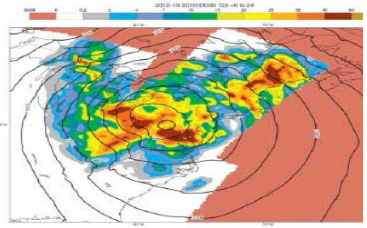
5d FC T1279



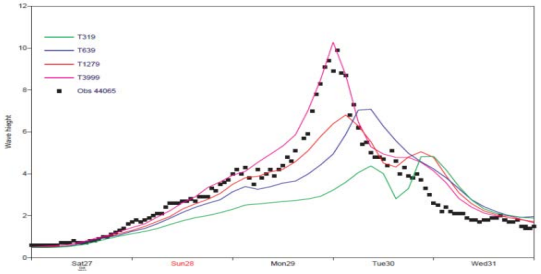
5d FC T639



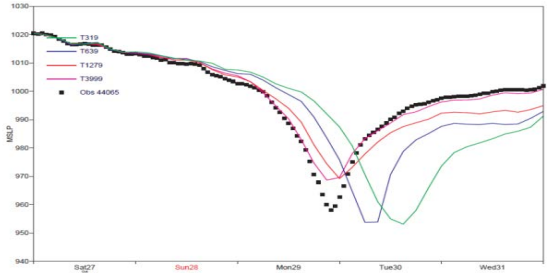
Precipitation: NEXRAD 27 Oct



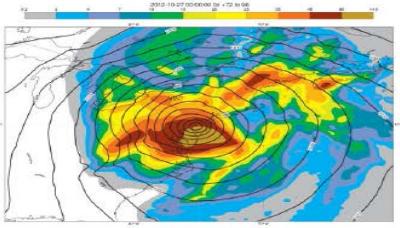
3d FC Wave height



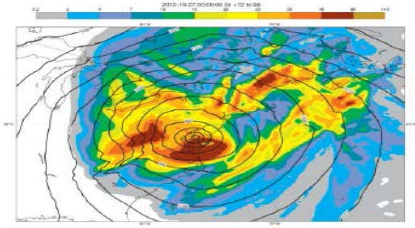
Mean sea-level pressure



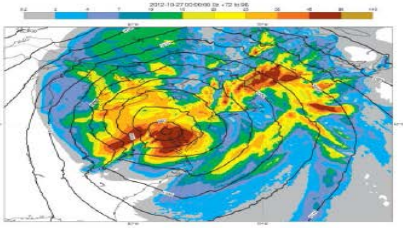
4d FC T639



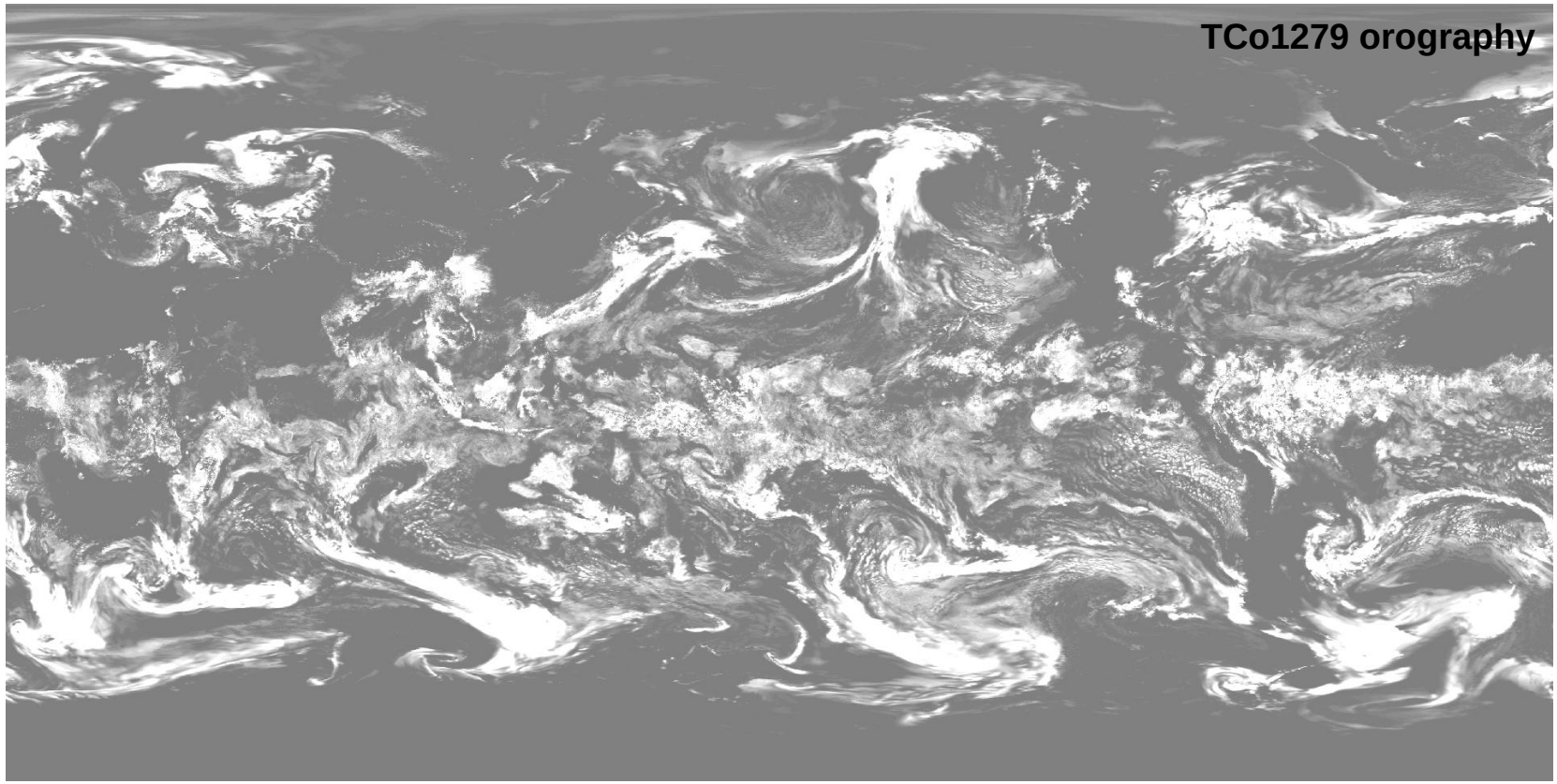
4d FC T1279



4d FC T3999

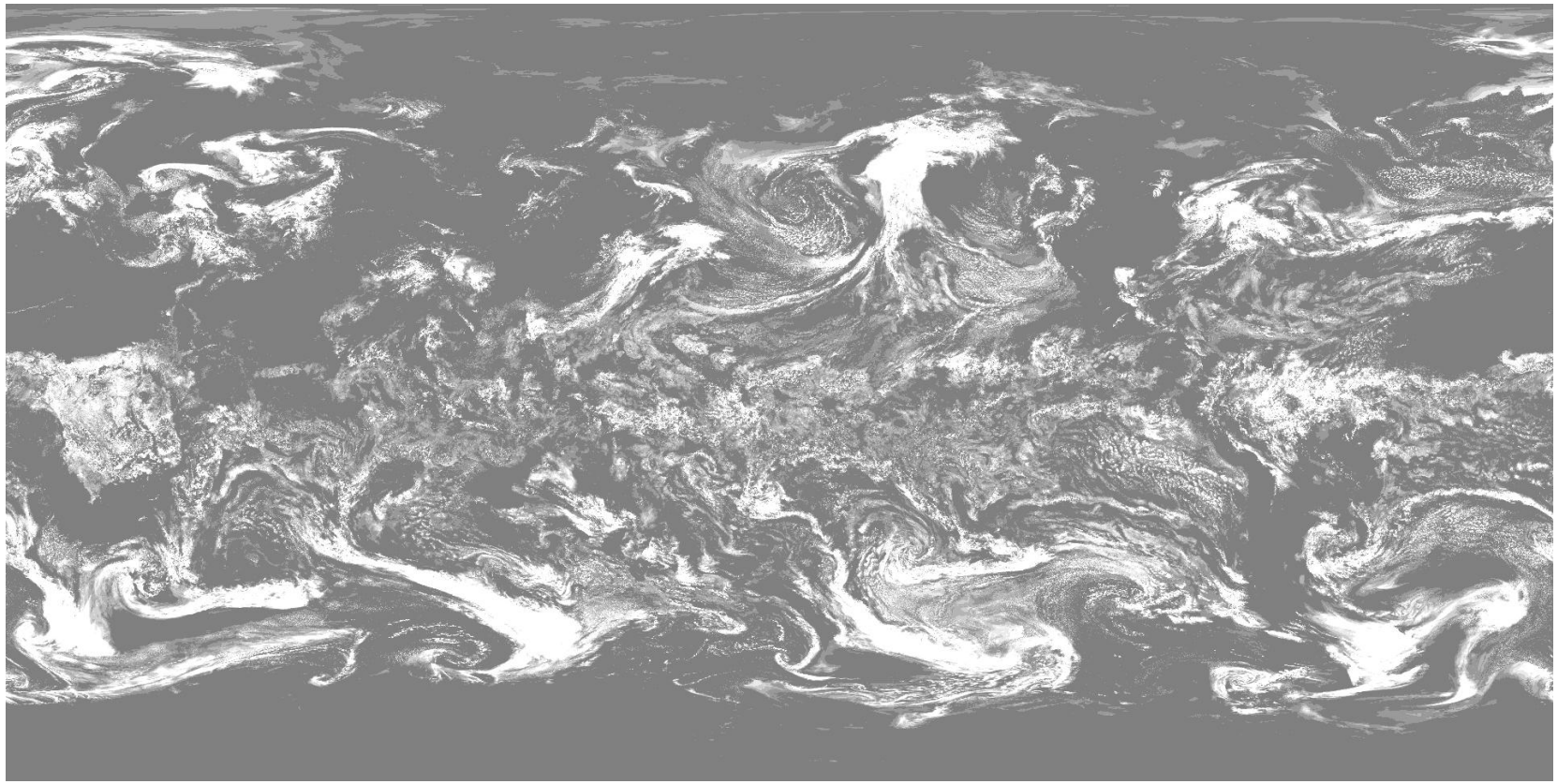


TCo1279 (~9km) world's highest resolution global NWP today



(12h forecast, *hydrostatic*, with deep convection parametrization, 450s time-step, 240 Broadwell nodes, ~0.75s per timestep)

TCo7999 (~1.25km) 256 Megapixel camera with predictive skill!



(12 h forecast, *hydrostatic*, no deep convection parametrization, 120s time-step, 960 Broadwell nodes, ~10s per timestep)

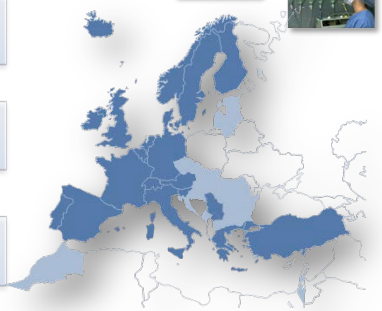
History and Future of Resolution Upgrades

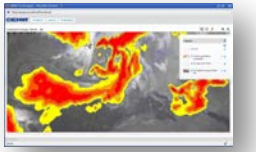
Resolution	Grid size	Grid Points	Field Size (in memory)
T319	62.5 km	204 k	1.6 MB
T511	39 km	524 k	4 MB
T799	25 km	1.2 M	9.6 MB
T1279	16 km	2.1 M	16.8 MB
Tco1279	**9 km**	**6.6 M**	**50.4 MB**
Tco1999	5 km	16.1 M	122.6 MB
Tco3999	2.5 km	64 M	490 MB
Tco7999	*1.25 km*	*256 M*	***1909 MB***

The tendency of memory per core diminishing …

… is likely to have serious implications on the post-processing workflows!

10-Year Challenge





30x more data sent to customers per day in critical path

Dissemination

RMDCN

Internet

Data acquisition

10x more observational data per day

Forecast run

Product generation

Web services

Internet

2000x more model data per time step

25x more forecast product data per day in critical path

100x more data archived per day

Archive

Data Handling System

Conclusions & Questions

- NWP has had I/O **exponential growth** for many years.
- What is different?
  - Moving from **compute centric to data centric** paradigm
  - Minimise data movement and bring compute to data
- Update our **legacy codes and workflows** to this new paradigm
- How to **adapt upcomming technologies** for complex workflows?
  - Burst Buffers
  - NVRAM
  - Storage-side compute
  - Object stores
- Can we move **beyond the filesystem**? How intrusive should that be?
  - Interpreting scientific data as objects
  - Challenges in data modelling and data curation