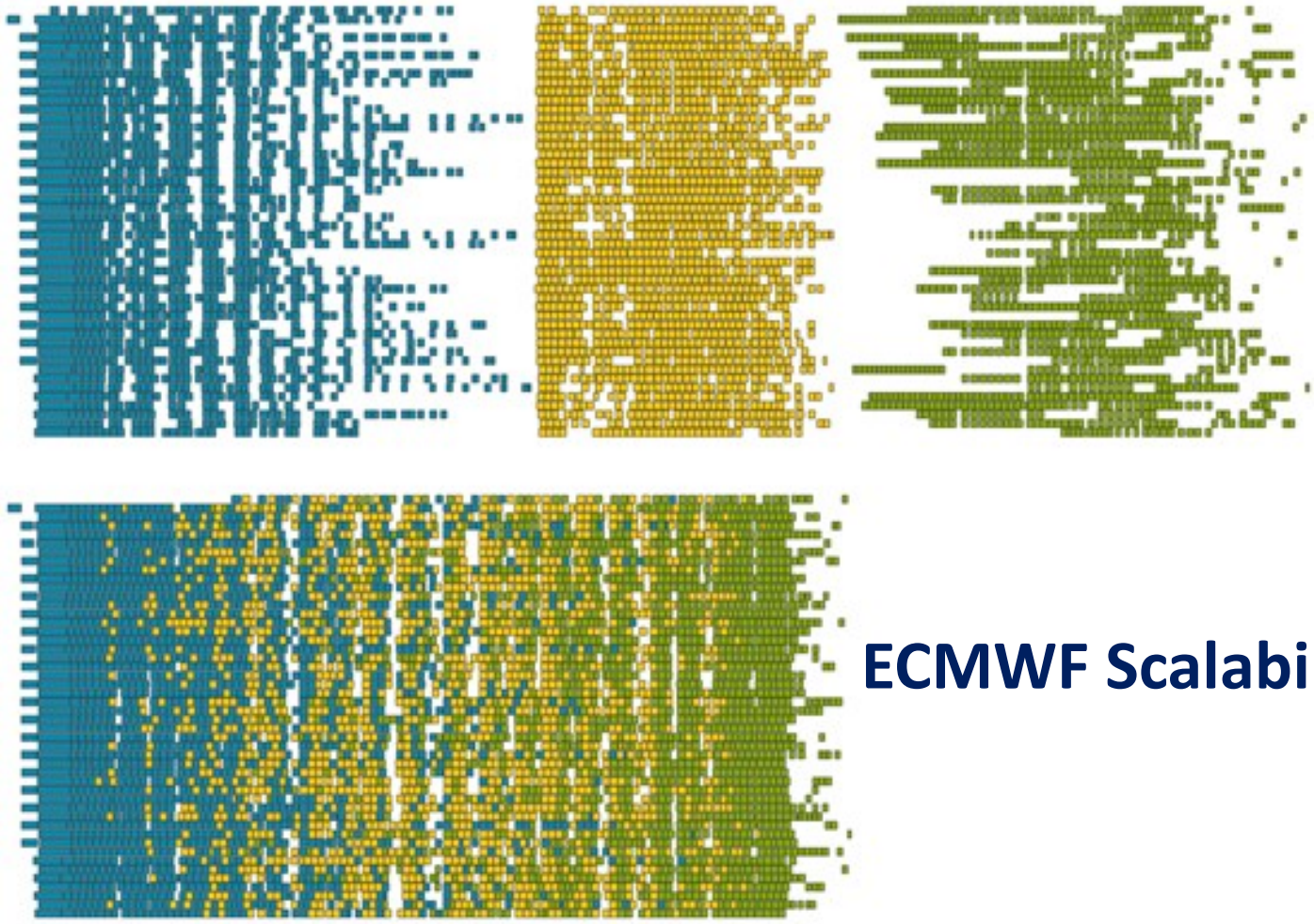


Picture: Stan Tomov, ICL, University of Tennessee, Knoxville

ECMWF Scalability Programme

Peter Bauer
Willem Deconinck, Mike Hawkins,
Kristian Mogensen, George Mozdzynski,
Tiago Quintino, Deborah Salmond,
Yannick Trémolet, Nils Wedi

Scenario 2020+: Compute and Archive

Compute (communication):

- model time step of 30 seconds
- 10 day forecast
- model on 4,000,000 cores
- max. 1 hour wall clock
- 1 step needs to run in under 0.125 seconds
- by using 32 threads per task with 128,000 MPI tasks:
  - a simple MPI_SEND from 1 task to all other 128k tasks will take an estimated 128k x 1 µsec = 0.128 seconds

→ **Global communications** (+ memory limitations)**?**

Archive*:

- EC-Earth at 25km with 10 years/day on 10,000 cores
- 25 member ensemble x 4 for e.g. calibration:
  - 1,000,000 core experiment
  - 25-year run over 2.5 days produces 60,000,000 core hours
  - 250 Gbyte per compute month per member
  - 6 Pbyte per day = 0.5 Tbit per second

→ **Data I/O rates, reliable management on disks for post-processing and dissemination?**

(*Example courtesy Bryan Lawrence U Reading)

Scenario 2015: Compute

Operational application run = wall clock time x number of cores … **x 1.0/scaling factor**

Today's ECMWF ensemble (50M T639 L91 legA):

= 2 hours x 12,000 cores … x 1.0

Tomorrow's ECMWF ensemble (50M T1023 L91 legA):

= 2 hours x 3 x 12,000 cores … x 1.0

= 2 hours x 36,000 cores … x 1.0

= 2 hours x 48,000 cores … x 1.0/0.75

= 1 hours x 64,000 cores … x 1.0/0.75/0.75

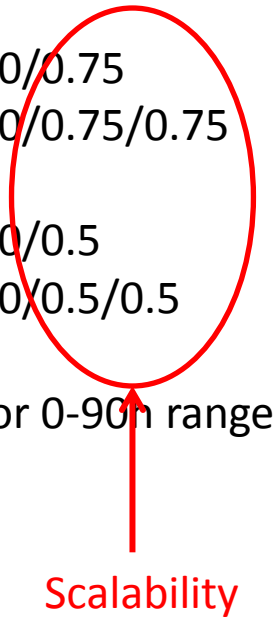
= 2 hours x 72,000 cores … x 1.0/0.5

= 1 hours x 144,000 cores … x 1.0/0.5/0.5

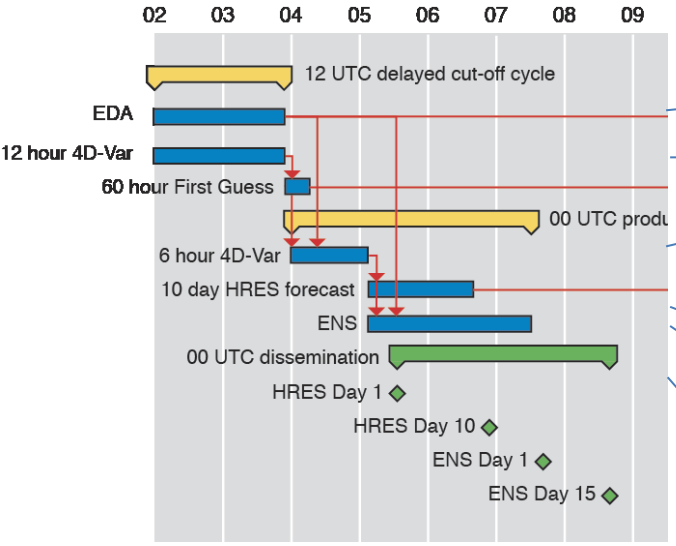
Tomorrow's ensemble output: 4 runs per day with 1-hourly ML output for 0-90h range

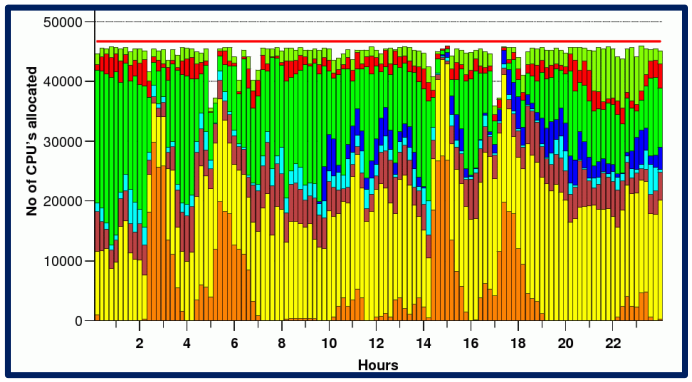
The day after tomorrow?

- Global, convection resolving scales HRES O(1-2 km), ENS O(5 km)?
- Aerosols, trace gases, ocean, waves, sea-ice coupling?



ECMWF production workflow





Data assimilation (obs. pre-processing):

- **12h EDA**: 25 members, 2 outer loops, inner loops w/ iterations, 6h integrations, low resolution
- **6/12h 4DVAR**: 3 outer loops, inner loops w/ iterations, 6/12h integrations, high/low resolution, wave coupling
- Observation DB incl. feedback, ML and PL output

Model integration:

- **10d HRES**: 10d integrations, high resolution (radiation low resolution), wave coupling
- ML and PL output
- **10d/32d ENS**: 10d/32d integrations, lower resolution (radiation low resolution), ocean-wave coupling,
- (2 t-steps ML and) PL output

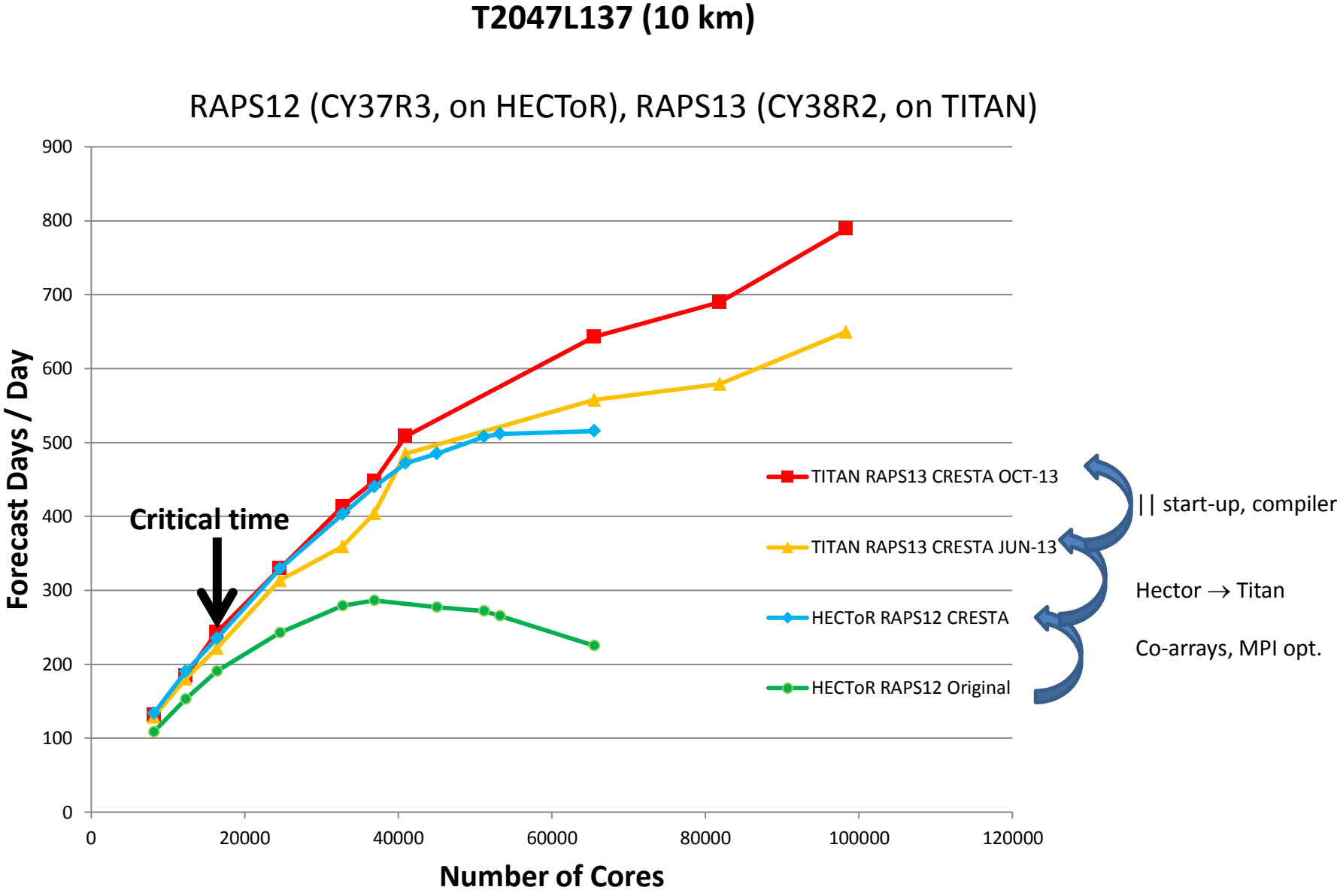
Data processing/archiving/dissemination:

- Data management
- Dissemination via RMDCN

Issues:

Sequential algorithms/data access, non-local communication, memory limits, load imbalance, redundant file access, coupling barriers etc.

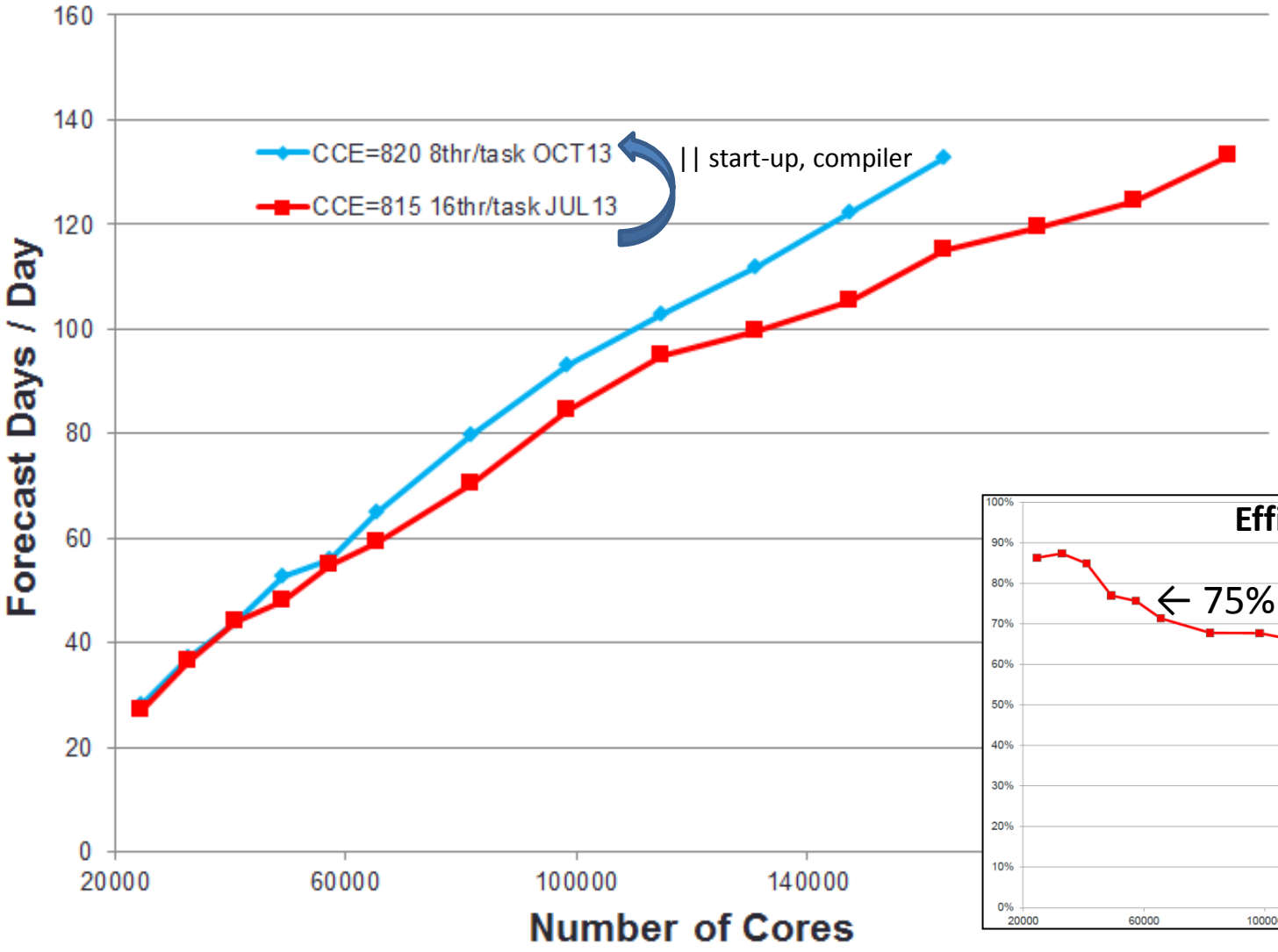
Scalability of computing: The good …

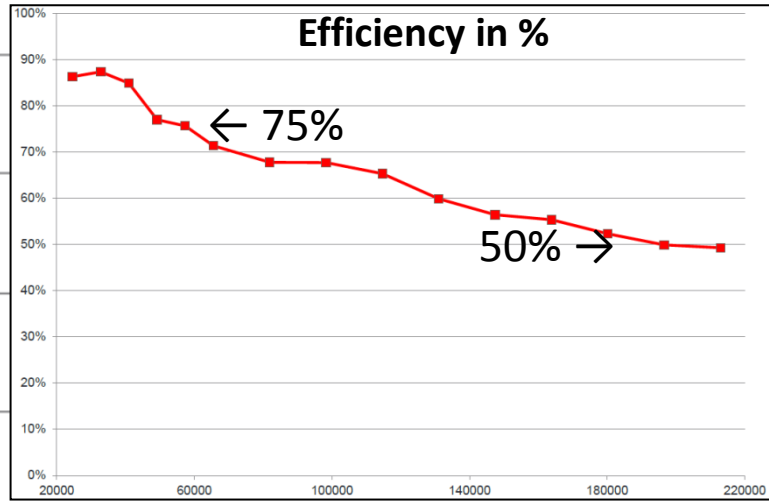


Scalability of computing: … the bad

Critical time

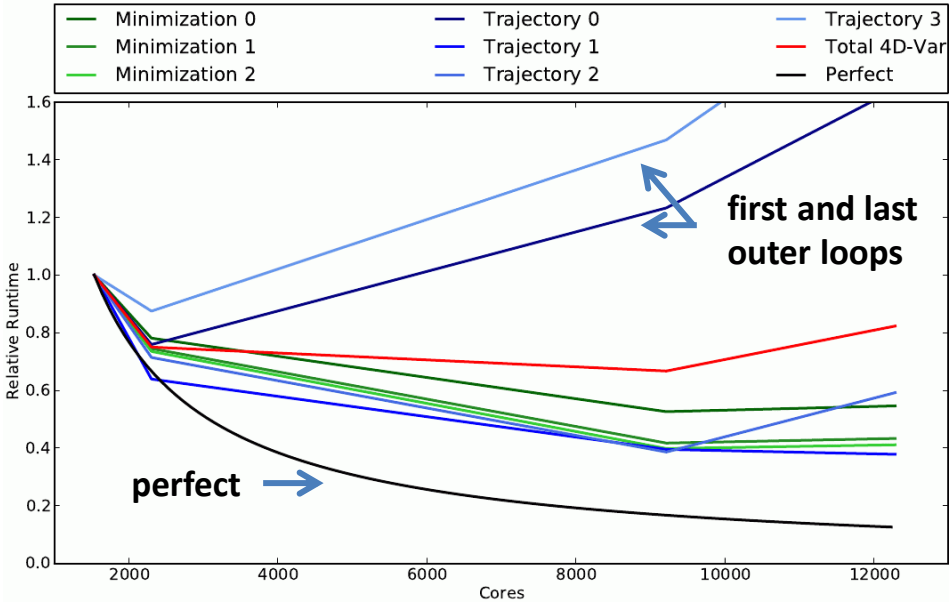
T3999L137 (5 km)

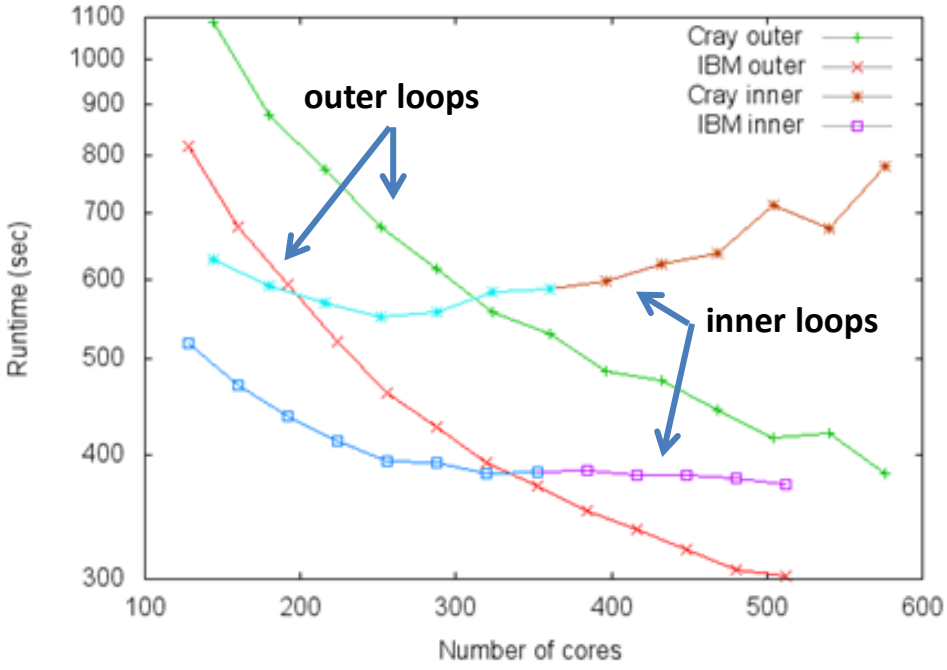




Scalability of computing: … and the ugly

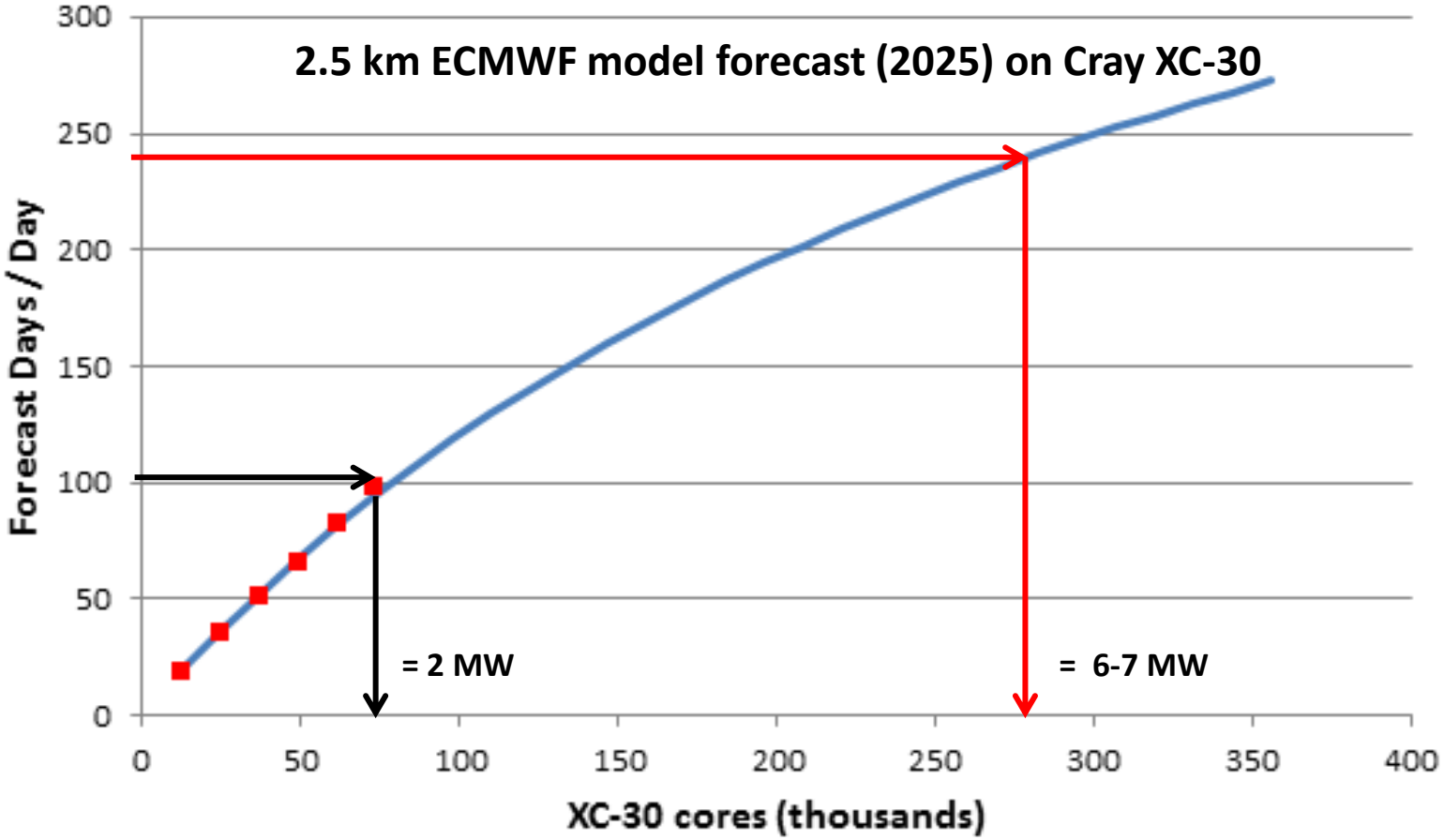
4DVAR (T1279c outer loops,
T255/T399/T399l inner loops,
12-hour window) →
(only a few thousand cores)





← NEMOVAR (1/4 degree, 5-day window)
(only a few hundred cores)

Power: Simplified



→ 100 days/day require 84,000 cores = 2 MW

→ 240 days/day require 280,000 cores = 6-7 MW

→ x 10 = 60-70 MW?

(scaling from HRES to full HPC incl. other suites, RD experimentation, MS quota)

Scalability Programme: Objectives

New capabilities :

- An integrated forecasting system (IFS) combining a flexible framework for scientific choices to be made with maximum achievable parallelism.
- Portable code structures ensuring efficiency and code readability exploiting a range of expected future technologies.
- Metrics and framework for code testing allowing quantitative assessment of scalability.

… given operational schedule constraint

Success metrics: - efficiency gains in Watts/FLOP or Watts
- efficiency gains in Gbyte/s and Pbyte

→ The Scalability Programme provides science **support**!

Scalability Programme: Structure

Board:

Chair: Florence Rabier (FD), Members: Erland Källén (RD), Adrian Wander (CD), Sinead McAtavie (AD), Andy Brown (Met Office), Alain Joly (Météo-France), Jeannette Onvlee (KNMI), Piet Termonia (RMI)

Projects:

Observation processing
(COPE)
- Workflow
- Pre-processing
- Filtering
- Monitoring

Data assimilation
(OOPS)
- Control structure
- Algorithms
- Coupling
- Scripts

Numerical methods
(PolyMitos)
- Data structures
- Discretization
- Concurrency
- Coupling

Data processing
(HERMES)
- Workflow
- Grids
- Interpolation
- Data formats
- Compression

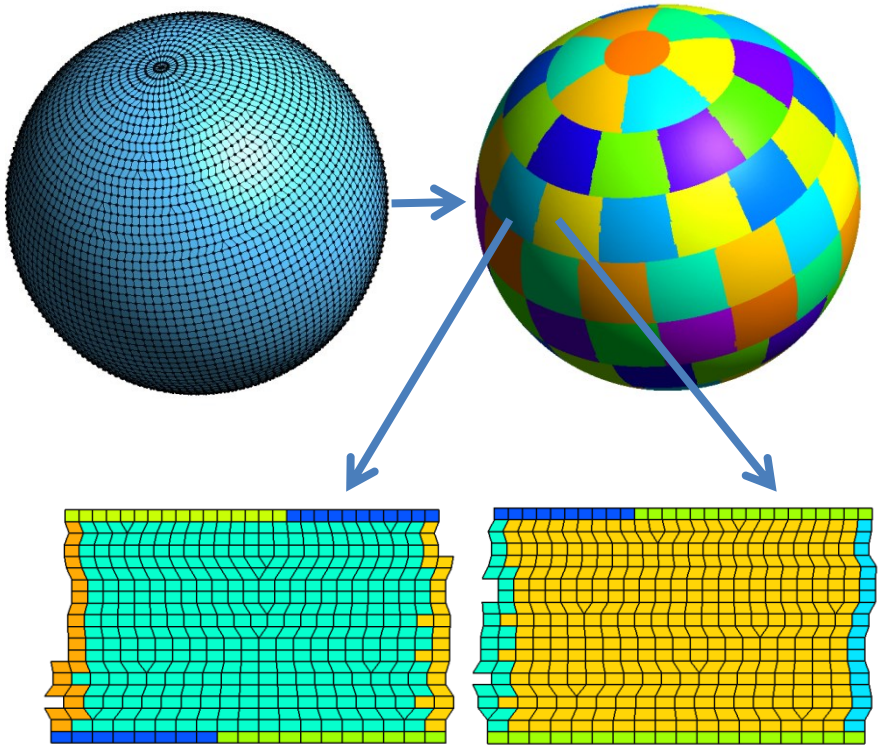
IFS code adaptation
(OAFS)
- Workflow
- Refactoring
- Benchmarking
- Optimization
- Migration

Computer architecture support
(CAS)
- Cray phase-2 (CPU, accelerators)
- RAPS benchmarking
- I/O benchmarking
- Migration

Scalability & Science example #1: PolyMitos & PantaRhei

PolyMitos

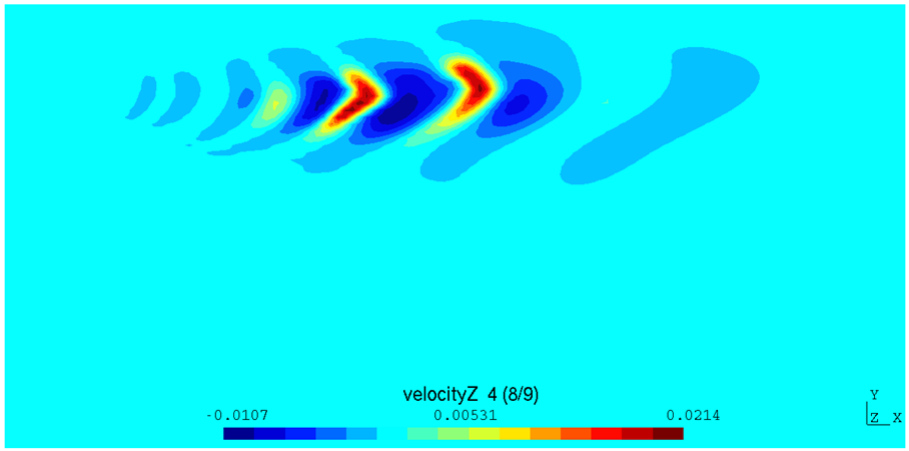
- ATLAS data structure framework
- Unstructured mesh / structured grid support
- Compact stencil space discretisation
- Nearest neighbour communication



- … multiple grids, I/O, spectral transforms

PantaRhei

- Research on equations:
  - fully compressible
- Conservative, monotone transport
- 3D Helmholtz solver, pre-conditioning

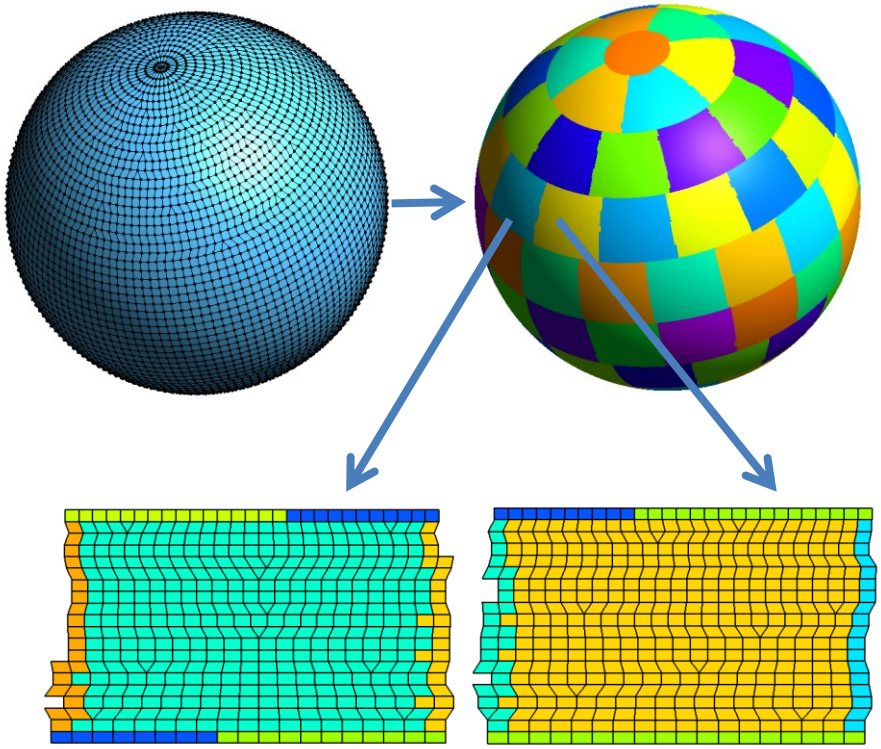


- … hybridization with IFS, physics-dynamics

Scalability & Science example #1: PolyMitos & Model

PolyMitos

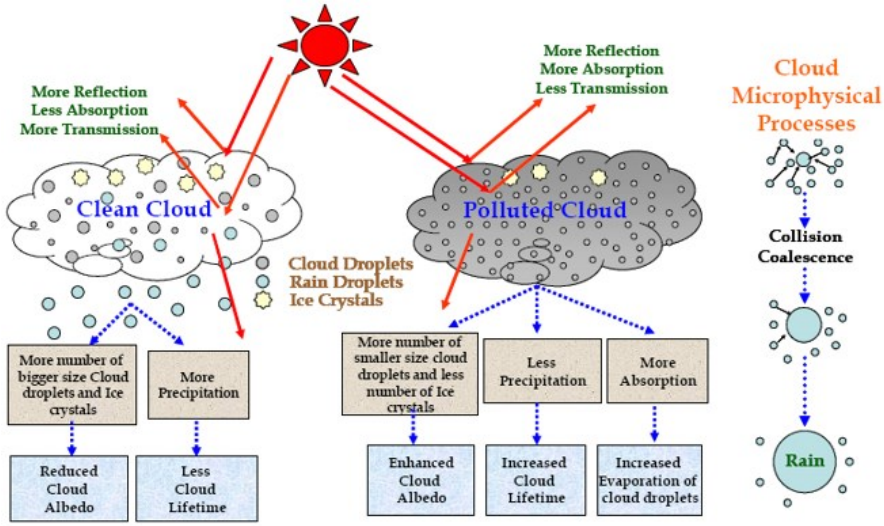
- ATLAS data structure framework
- Unstructured mesh / structured grid support
- Compact stencil space discretisation
- Nearest neighbour communication



- … multiple grids, I/O, spectral transforms

Model physics

- New prognostic variables:
  - Aerosols
  - Trace gases
  - Higher-moment cloud schemes
- High-resolution radiation calculations

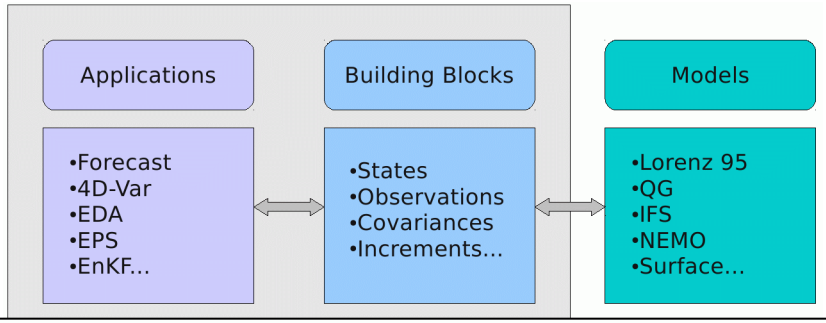


… also with accelerators (CAS, OAFS)

Scalability & Science example #2: OOPS & 4D-Var

OOPS

- Object-oriented design and call structure
- Top-level control level with abstract building blocks
- Classes can be flexibly assembled

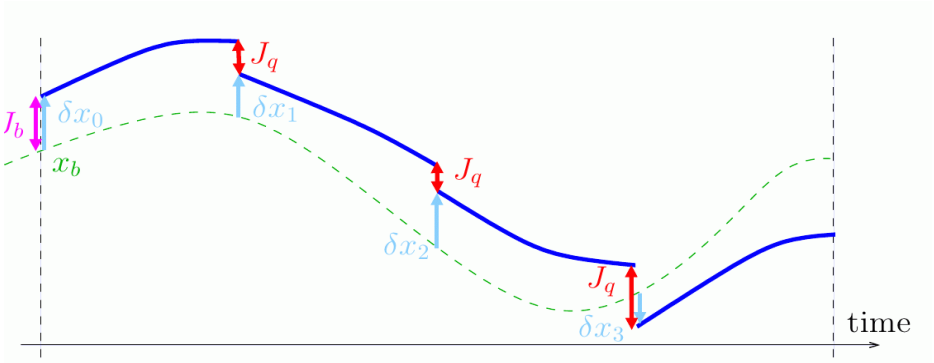


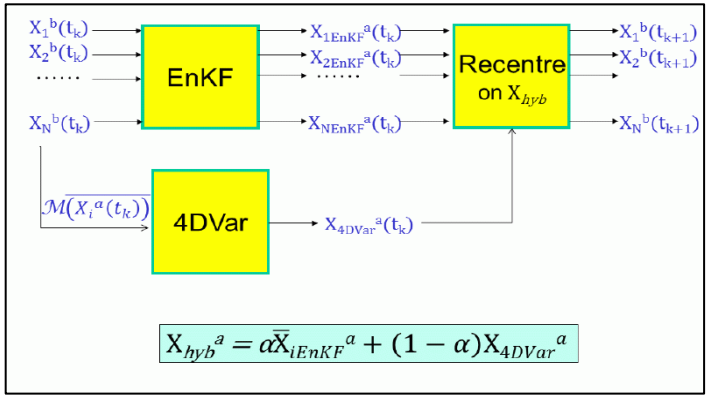
- In model space:
  1. Geometry
  2. State
  3. Increment
  4. ModelConfiguration
  5. LinearModel (Trajectory)
- In observation space:
  6. ObsOperator
  7. ObsAuxControl;
  8. ObsAuxIncrement;
  9. ObsVector
  10. ObsOperatorTrajectory;
- To make the link:
  11. Locations
  12. ModelAtLocations
- Covariance matrices (if generic ones are not used):
  13. Model space (**B** and **Q**)
  14. Observation space (**R**)
  15. Localization (4D-Ens-Var)

- … use of PolyMitos data structures, coupled data assimilation

(Hybrid) 4D-Var

- Model error formulation
- Saddle-point algorithm
- Hybrid ensemble – variational formulation



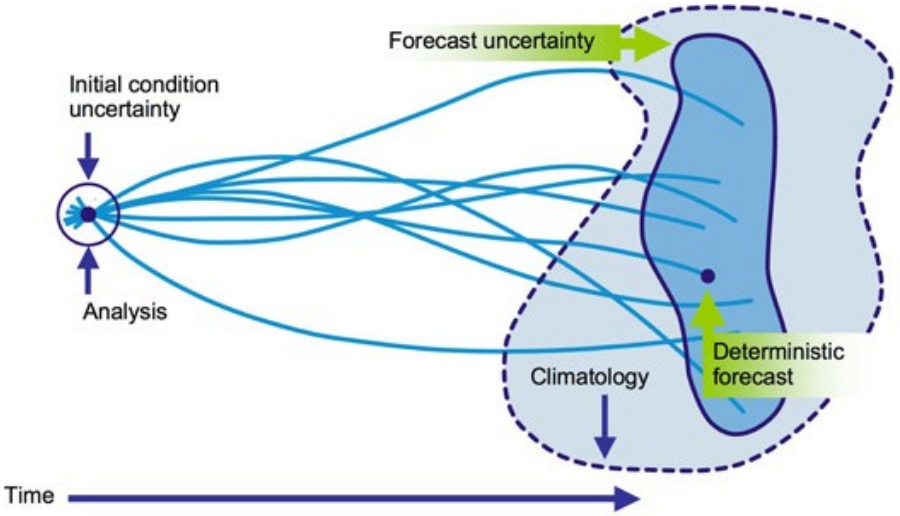


- … long-window 4D-Var, EnKF, coupled data assimilation

Scalability & Science example #2: ENS & HERMES

ENS

- High-resolution output on model levels
- On-the fly diagnostics of information content
- Reduction of degrees of freedom

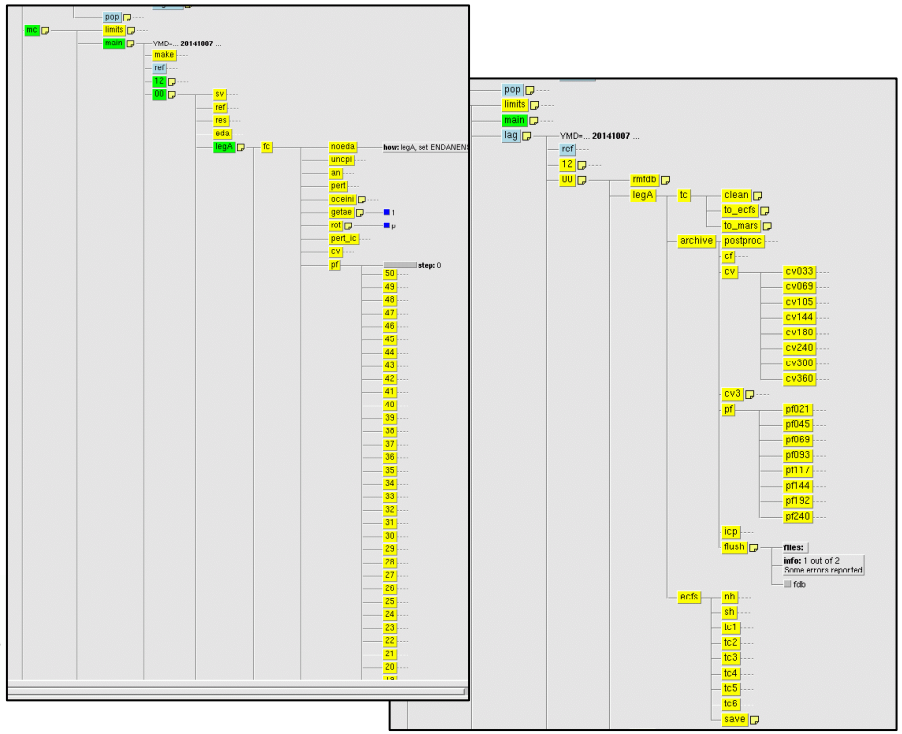


$$P(p, \Pi_0) \geq P(p_G^*, \Pi_0) = \sum_{i=1}^{M} P(p_i^*, \Pi_{0,i}) = \sum_{i=1}^{M} \frac{1}{2}[\ln \gamma_i^{-1} + \gamma_i - 1] + \sum_{i=1}^{M} \frac{1}{2}[\overline{z}_i^2].$$

- … advanced ensemble products

HERMES

- On-the-fly task configuration
- Parallelised post-processing
- Archiving away from single parameter – global field, high-level pdf information



- … possible restart configurations for enhancing resilience

Scalability Programme: Expected Outcome

Analysis (data assimilation)

- C++ layer, control structure
- Object oriented call of building blocks (model, operators, error stats, coupling)
- Full *flexibility* with respect to algorithms (variational, ensemble)
- Scalability through parallelisation in time

Forecast (model, ensembles)

- Data structures allowing any localization (→ assimilation/data management)
- Full *flexibility* with respect to equations, solvers, coupling
- Full *flexibility* with respect to meshes/grids (→ assimilation/data management)

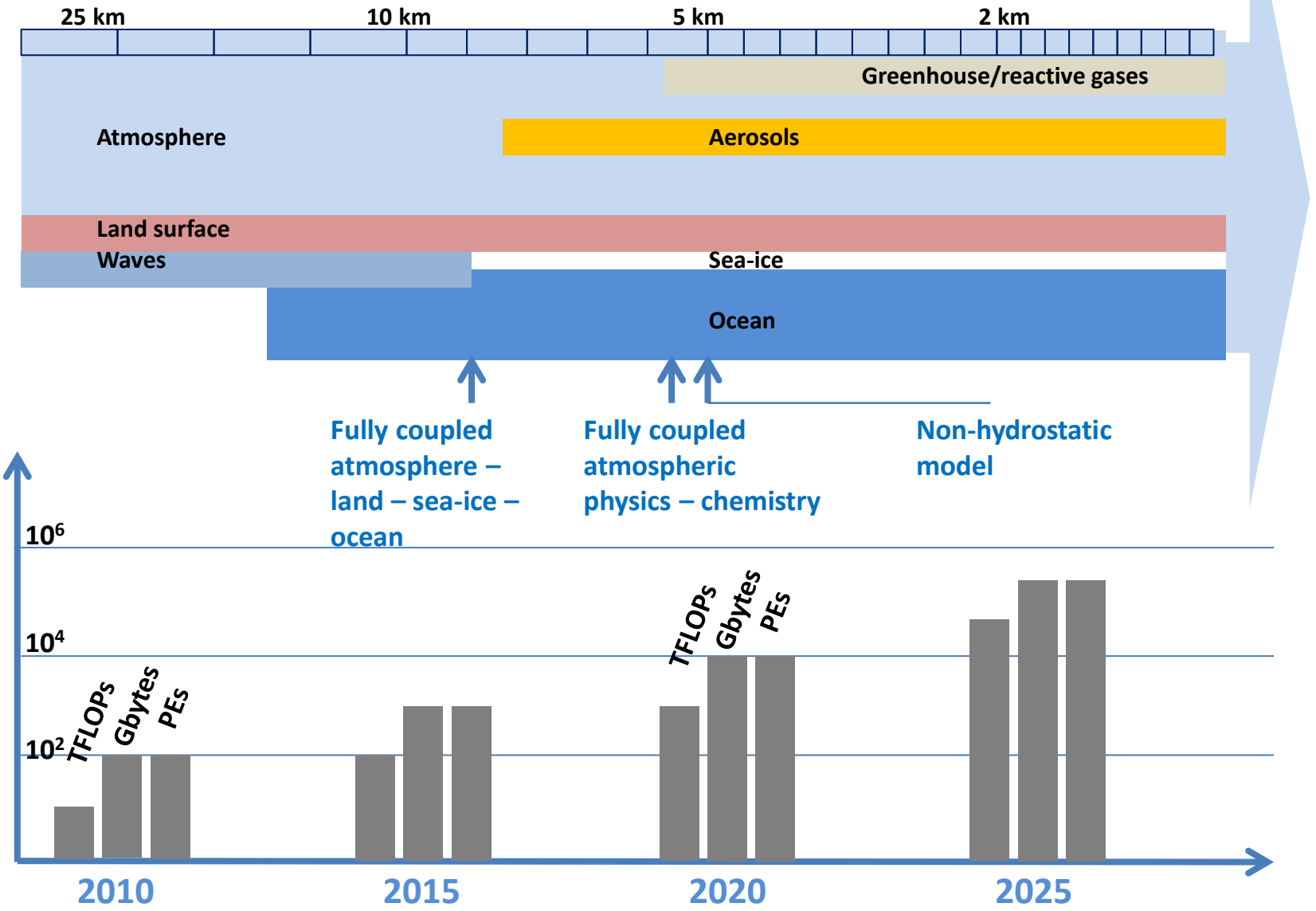
Pre-/post-processing

- Front-end data analysis to define workflow and tasks (delegation)
- Streaming interface to pipeline I/O between tasks
- Back-end data processing for different hardware options

Hardware support/code adaptation

- Benchmarking simulators (compute, data)
- Specialist DSL libraries (numerics, physics, math, solvers etc.)
- Interface with vendors (compilers, standards), portability

Model evolution with Scalability Programme



Model evolution without Scalability Programme