HPC trends and the scalability of atmosphere and ocean models

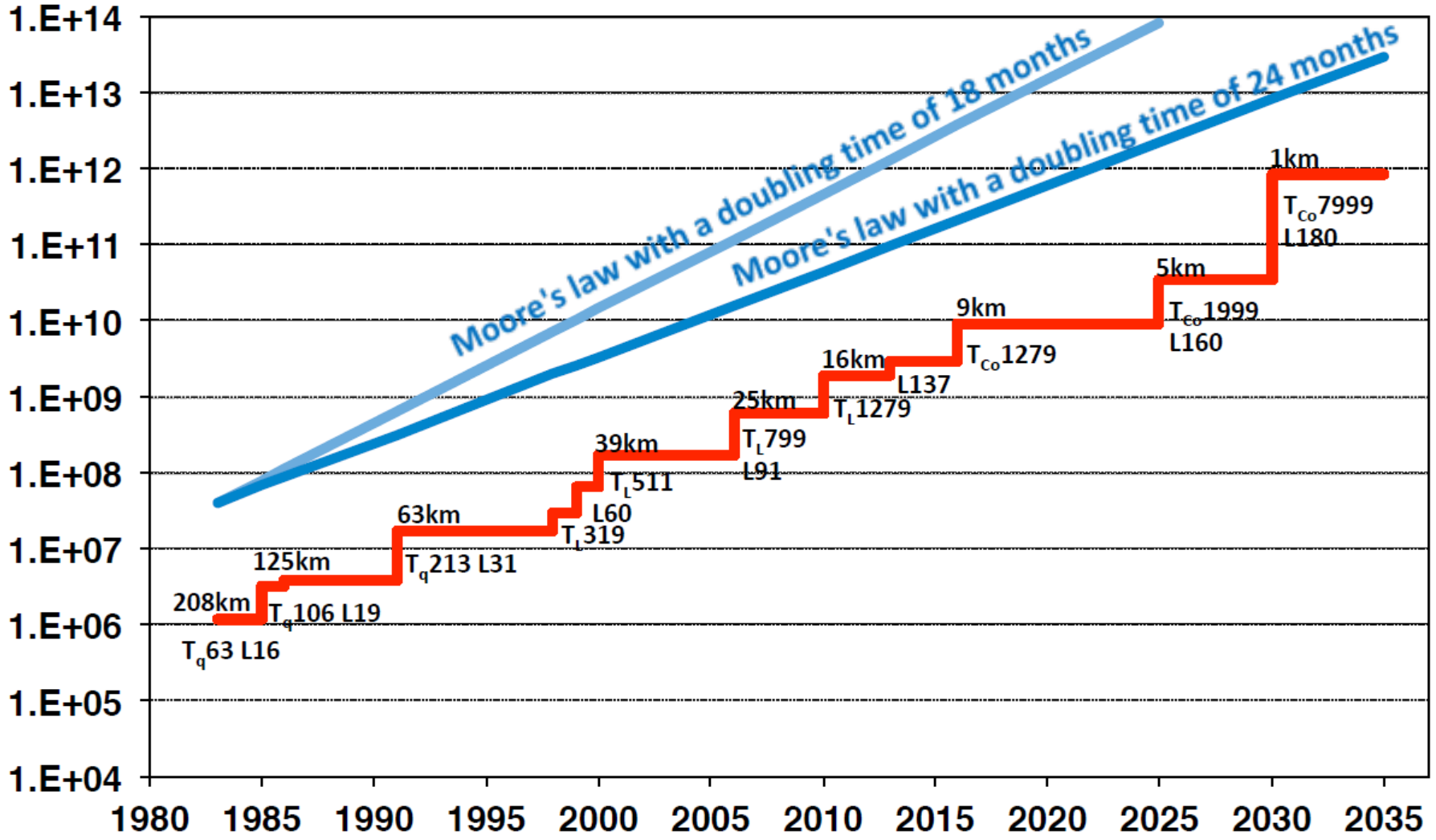
Nils P. Wedi, European Centre for Medium-Range Weather Forecasts (ECMWF)

Many thanks to: ESCAPE partners, Kristian Mogensen, Phil Jones, Silvia Mocavero, Eric Maisonnave, Mike Bell, Alan Wallcraft, V. Balaji, Peter Bauer, Marshall Ward, Simon Marsland, Michel Rixen, …

Outline

- Numerical weather prediction & climate, a brief (HPC) history
- Emerging constraints for ensemble-based assimilation and forecasts of Weather & Climate with increasing complexity
- An intermediate goal: globally uniform weather & climate modelling at 1 km horizontal resolution for both atmosphere and ocean/sea-ice
- HPC trends
- The current state-of-the-art and issues raised

Computational power drives spatial resolution

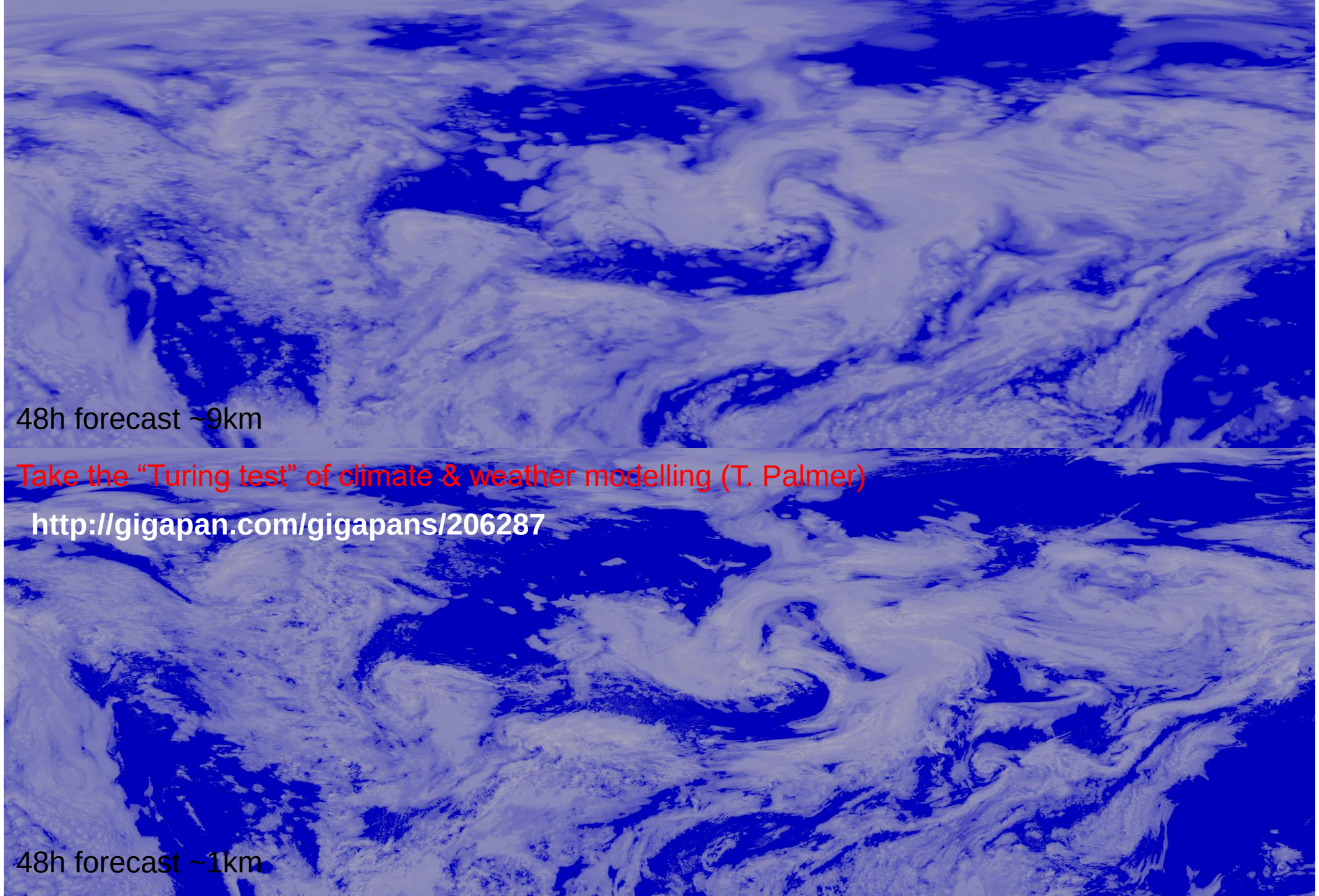


Gap of sustained and peak performance

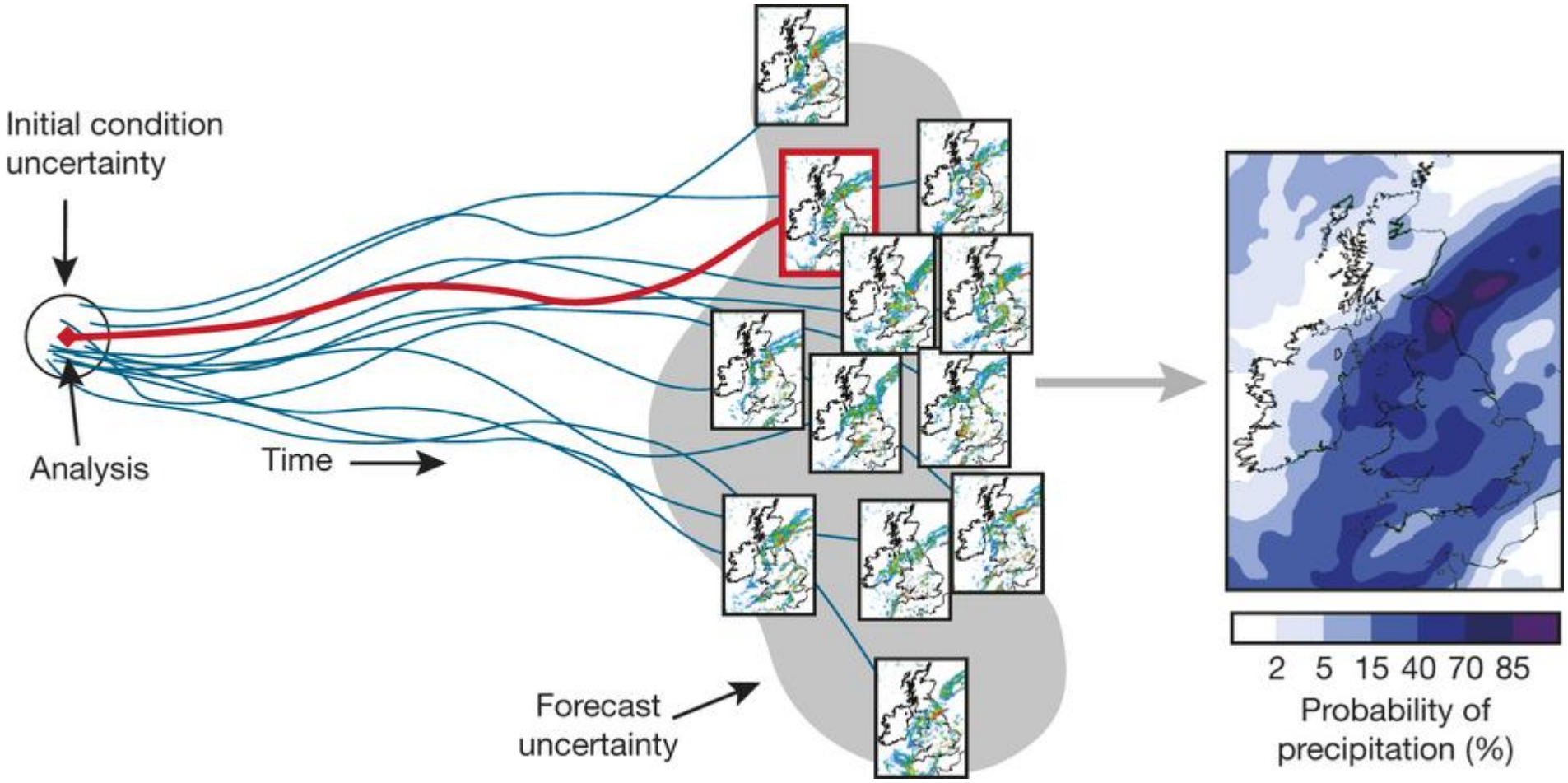
Steepness of gradient from 10km to 1km

(Schulthess et al, 2018)

ECMWF's progress in degrees of freedom
(levels x grid columns x prognostic variables)

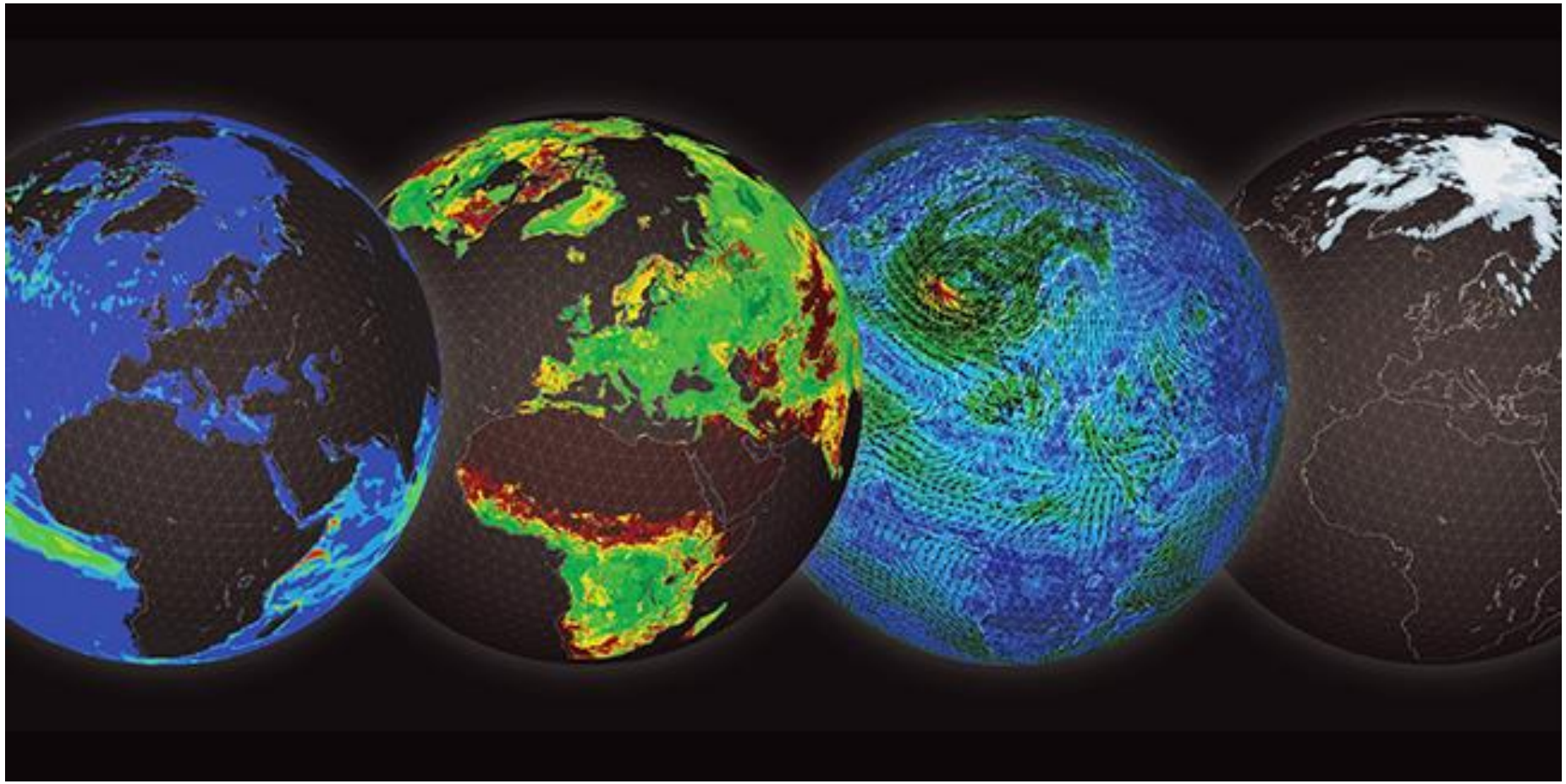

48h forecast ~9km
Take the “Turing test” of climate & weather modelling (T. Palmer)
http://gigapan.com/gigapans/206287
48h forecast ~1km

Ensemble of assimilations and forecasts

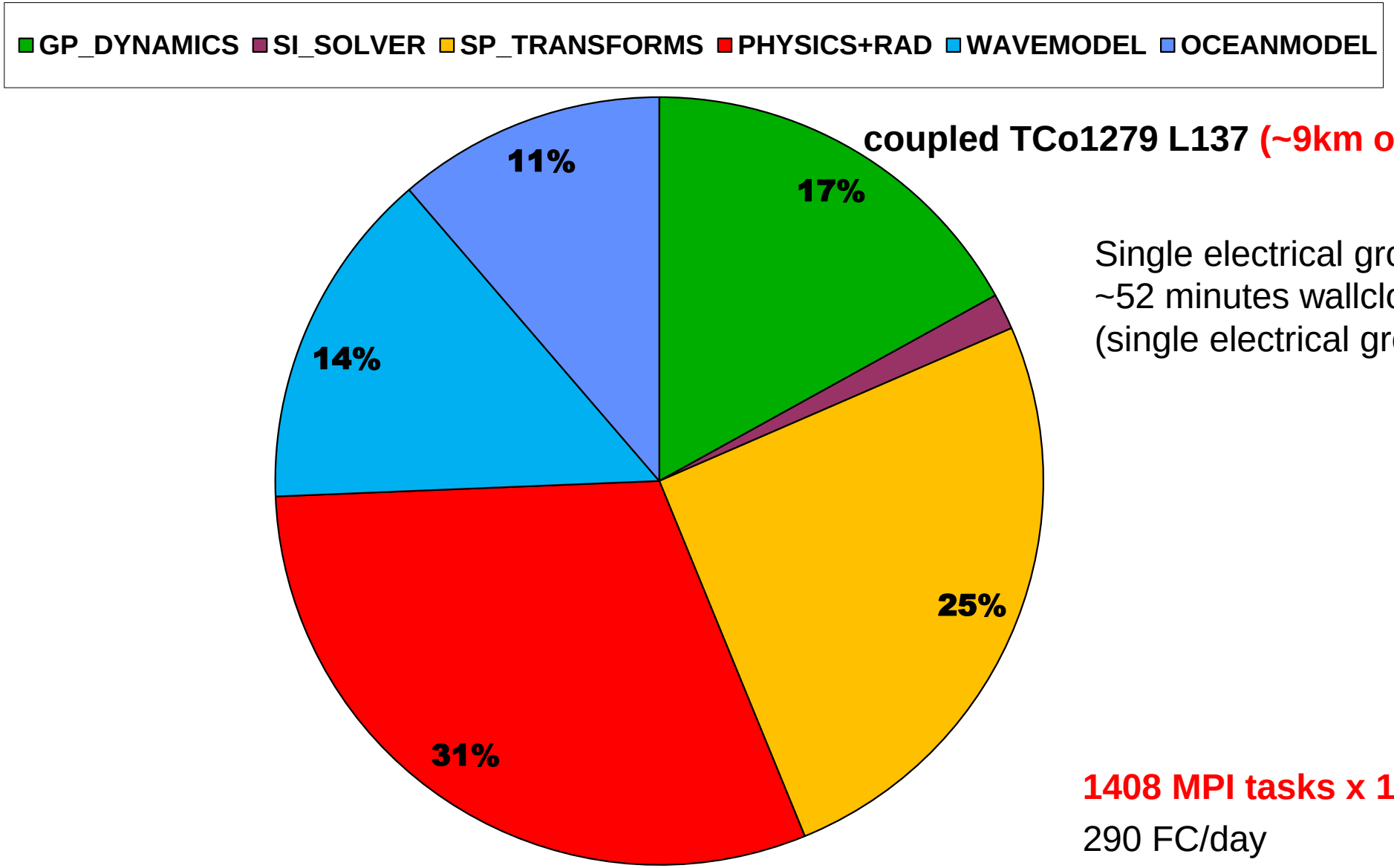


(Bauer et al, 2015)

Ocean – Land – Atmosphere – Sea ice



Where do we spend the time ? Cycle 45r1



coupled TCo1279 L137 (~9km operational) run

Single electrical group:
~52 minutes wallclock time
(single electrical group==384 nodes)

1408 MPI tasks x 18 threads

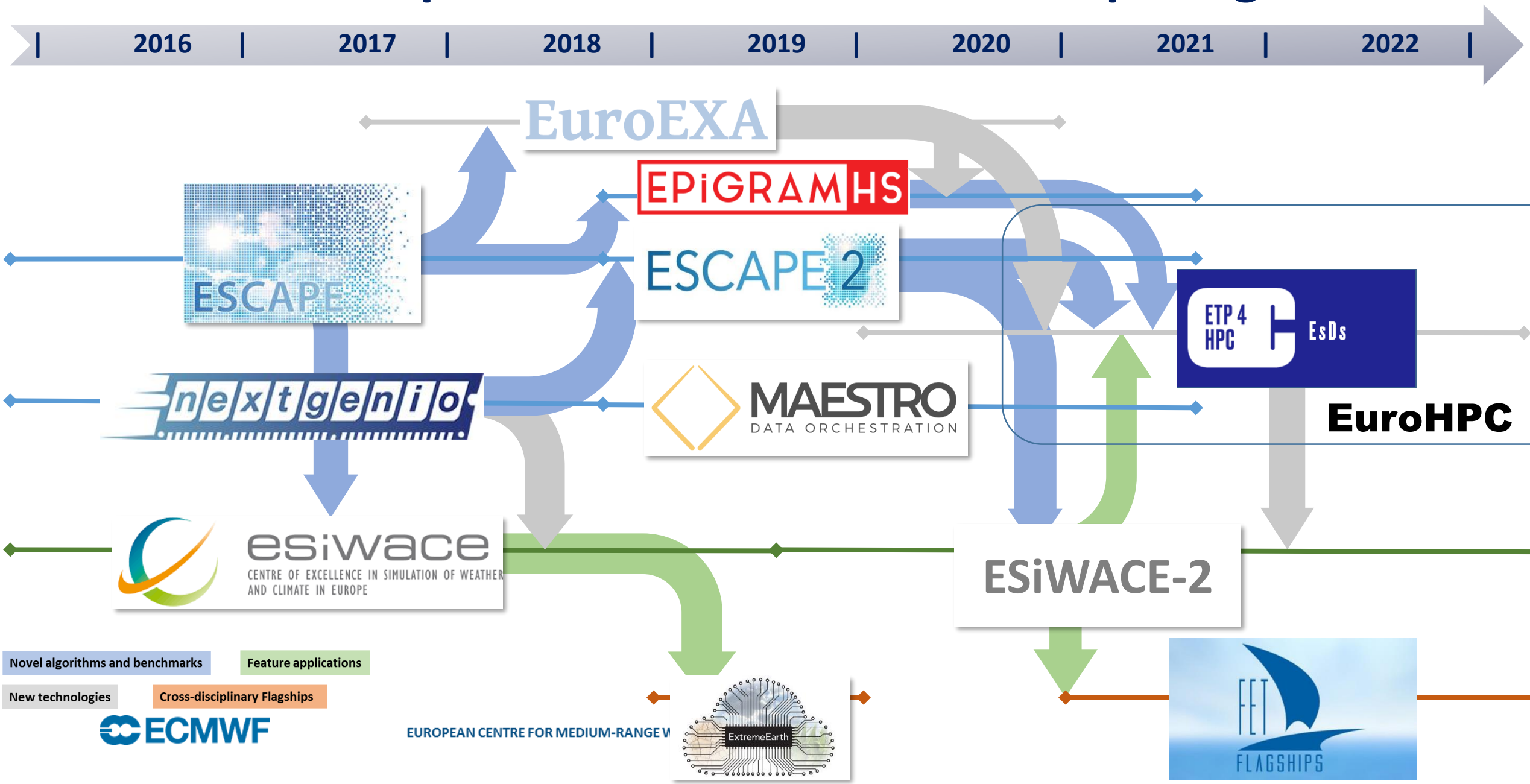
290 FC/day

High Performance Computing (HPC) trends

- Use of manycore CPUs possibly combined with accelerators such as GPUs
- Arrival of open instruction set architectures (FPGA, ARM, RISC-V, …)
- Exascale race driven by a concern for the energy footprint and physical distances between processors (e.g. low-power processors, memory hierarchies, liquid cooling, etc)
- Machine learning, both driving specific processor development (eg. Google TensorFlow) and application development (e.g. physical parametrizations, feature detection in satellite observations)
- Cloud computing and storage (e.g. access to HPC from anywhere, simpler install in embedded virtual environments, data processing near large meteorological archives, etc)

See ECMWF's 18th Workshop on High Performance Computing in Meteorology (www.ecmwf.int)

Roadmap for weather & climate computing



ExtremeEarth

Co-Design

Science:

- Climate prediction
- Weather forecasting
- Earthquake prediction

Impact:

- Hydrology and water
- Energy
- Food and agriculture
- Geo-engineering
- Disasters and risks

- Advanced mathematics & algorithms
- Multi-scale/multi-physics models
- Portable and performant science code
- Domain specific computing framework
- End-to-end demonstrators

Technology:

- Numerical modelling
- Data assimilation and fusion
- Deep learning
- Programming models
- Extreme and cloud computing
- Extreme data handling and storage
- Workflows and visualization

Ultra high-resolution, integrated Earth-system & impact modelling capability

Integrated exascale Earth-system data analytics & management capability

Earth-system HPC technology and exascale capability

Integrated Earth-system information system capability

Selection of other ongoing initiatives

- Europe
  - See slide
  - ICON developments (DWD/MPI)
  - Gung-Ho/LFRic (UKMO)
  - NEMO/SI3 developments (NEMO consortium)
  - Unstructured ocean and sea-ice modelling (eg. AWI)
- US
  - FV3 (NASA/GMAO; NCEP/NWS)
  - CANGA, E3SM focused developments based on higher-order SE, various projects (DOE)
  - NOAA ESRL developments towards Exascale
  - Caltech ESM2.0 (incl the use of ML)
  - CICE; HYCOM, MOM6, MPAS-ocean developments
  - NRL, high resolution ocean, higher-order CG, cloud computing
- Canada
  - Focus on strongly coupled DA
- Japan
  - Post-K (arm-based, very high resolution NICAM+LETKF DA)
- China
  - Exascale by 2020; various model developments ?
- Korea
  - KIAPS new model and DA developments, high-order CG/DG
- Many more ….

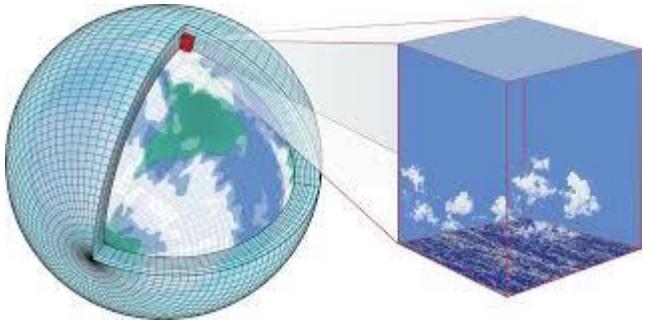
Energy-efficient SCalable Algorithms for weather Prediction at Exascale

Funded by the European Union

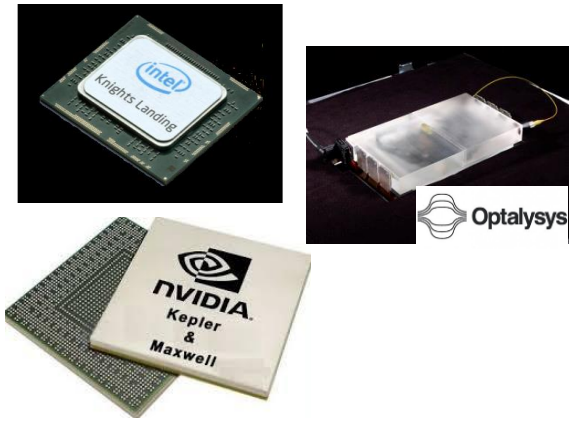
- Pioneering approaches for refactoring society critical legacy codes
- Energy-efficient accelerator use in global weather & climate prediction
- Co-development of novel mathematical algorithms & hardware adaptation
- Defining and encapsulating the fundamental algorithmic building blocks ("Weather and Climate Dwarfs")
- Reviewing the need for precision
- Pioneering algorithm development with hardware adaptation using DSL toolchains
- A HPCW benchmark and cross-disciplinary Verification, Validation, and Uncertainty Quantification (VVUQ)
- Application resilience

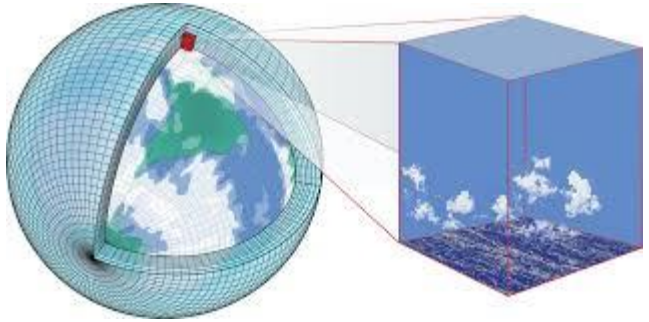
Weather & Climate Dwarfs

(hpc-escape.eu)

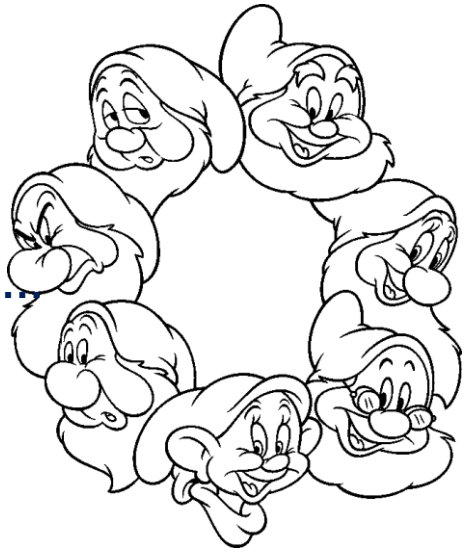


… hardware adaptation …

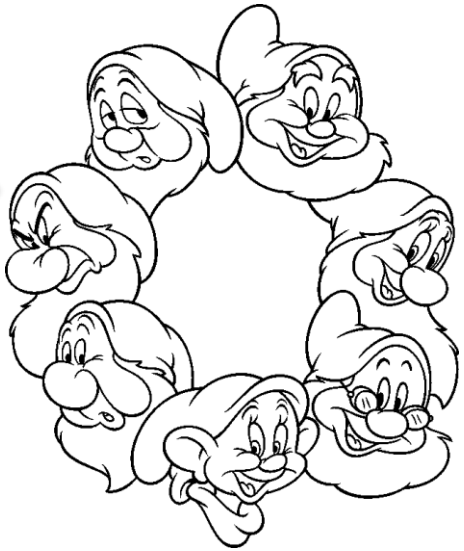




Extract model dwarfs…



… explore alternative numerical algorithms …

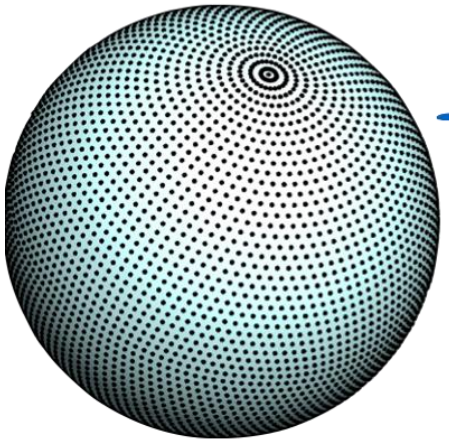


… reassemble model and benchmark

ECMWF · MeteoSwiss · DMI · RMI · NVIDIA · METEO FRANCE Toujours un temps d'avance · ICHEC · Deutscher Wetterdienst Wetter und Klima aus einer Hand · Optalysys · Bull · PSNC · Loughborough University

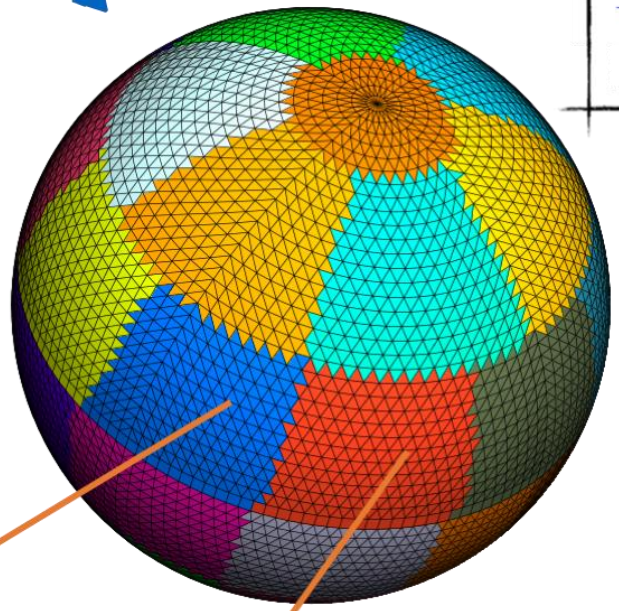
Atlas: a library for NWP and climate modelling

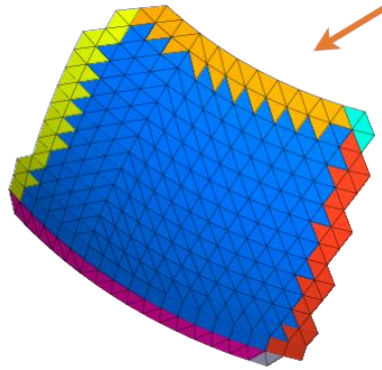
https://github.com/ecmwf

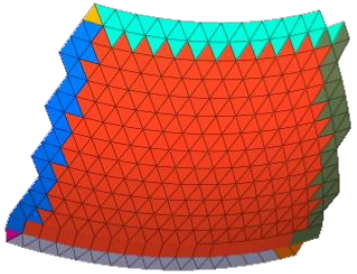


Grid

Mesh

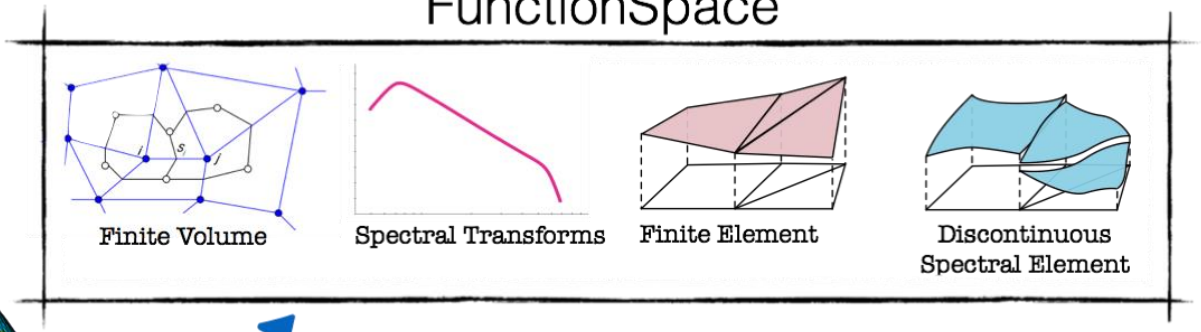


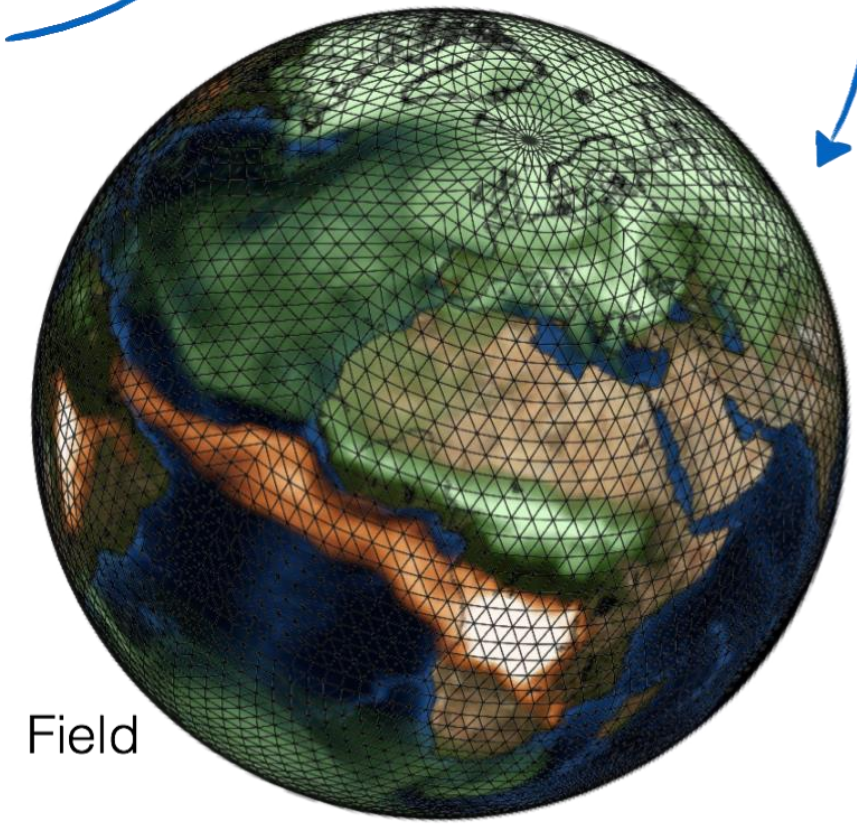




Partitions with halos

FunctionSpace





Field

Deconinck et al. 2017

Domain-specific language toolchain

Funded by the European Union

$$F(\Psi_L, \Psi_R, U) \equiv [U]^+ \Psi_L + [U]^- \Psi_R \quad (3a)$$

$$U \equiv \frac{u\delta t}{\delta x}, \quad [U]^+ \equiv 0.5(U+|U|), \quad [U]^- \equiv 0.5($$

Advection (MPDATA)

```
template <uint_t Color> struct upwind_flux {
  using flux = accessor<0, enumtype::inout, icosah
  using pD =
      in_accessor<1, icosahedral_topology_t::verti
  using vn = in_accessor<2, icosahedral_topology_t

  typedef boost::mpl::vector<flux, pD, vn> arg_lis

  template <typename Evaluation> static void Do(Ev

    constexpr auto neighbors_offsets =
        connectivity<edges, vertices, Color>::offs
    constexpr auto ip0 = neighbors_offsets[0];
    constexpr auto ip1 = neighbors_offsets[1];

    float_type pos = math::max(eval(vn()), (float_
    float_type neg = math::min(eval(vn()), (float_

    eval(flux()) = eval(pos * pD(ip0) + neg * pD(i
  }
};
```

```
ute_upwind_flux(this,pflux,pD,pVn)
e), intent(inout) :: this
t(out) :: pflux(:,:)
t(in)  :: pVn(:,:), pD(:,:)
s, zneg
dges
evels
ge, jlev, ip1, ip2

debug('compute_upwind_flux')

s%dimensions%nb_edges
s%dimensions%nb_levels

DO SCHEDULE(STATIC) PRIVATE(jedge,jlev,ip1,ip2,zpos,zneg)
_edges
ode(1,jedge)
ode(2,jedge)
o_levels
        = max(0._wp,pVn(jlev,jedge))
        = min(0._wp,pVn(jlev,jedge))
jedge)  = pD(jlev,ip1)*zpos+pD(jlev,ip2)*zneg
enddo
!$OMP END PARALLEL DO
end subroutine compute_upwind_flux
```

Complementary skills of CLAW, GridTools (MeteoSwiss) and Atlas (ECMWF)

ECMWF · MeteoSwiss · DMI · RMI · NVIDIA · METEO FRANCE Toujours un temps d'avance · ICHEC Irish Centre for High-End Computing · Deutscher Wetterdienst Wetter und Klima aus einer Hand · Optalysys · Bull atos technologies · PSNC · Loughborough University

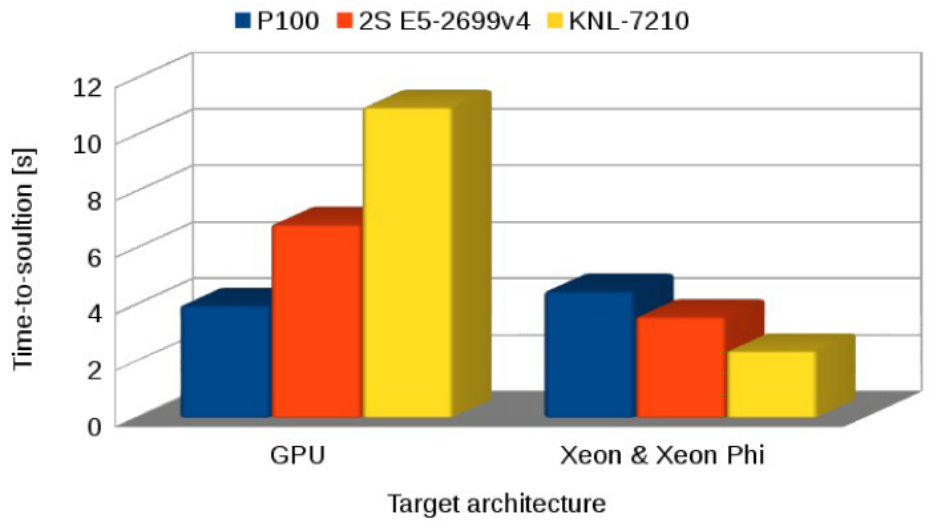
Weather & Climate Dwarfs

Funded by the European Union

Dwarf	prototype implemented	docu-mented	based on Atlas	MPI	Open MP	Open ACC	DSL	Optalysys
D - spectral transform - SH	✅	✅	✅	✅	✅	✅		
D - spectral transform - biFFT	✅	✅		✅	✅	✅		✅
D - advection - MPDATA	✅	✅	✅	✅	✅	✅	✅	
D - advection - semi-Lagrangian	✅	✅	✅	✅				
D - elliptic solver - GCR	✅	✅	✅	✅				
P - cloud microphysics - CloudSC	✅	✅		✅				
P - radiation scheme - ACRANEB2	✅	🚶	🚶	✅				
I - LAITRI (3d interpol. algorithm)	✅	✅						
planned next:								
D - advection - discontinousGalerkin	●	●	●	●				
D - elliptic solver - multigridPrecon	●	●	●	●				

✅: first version running
🚶: in progress
●: planned

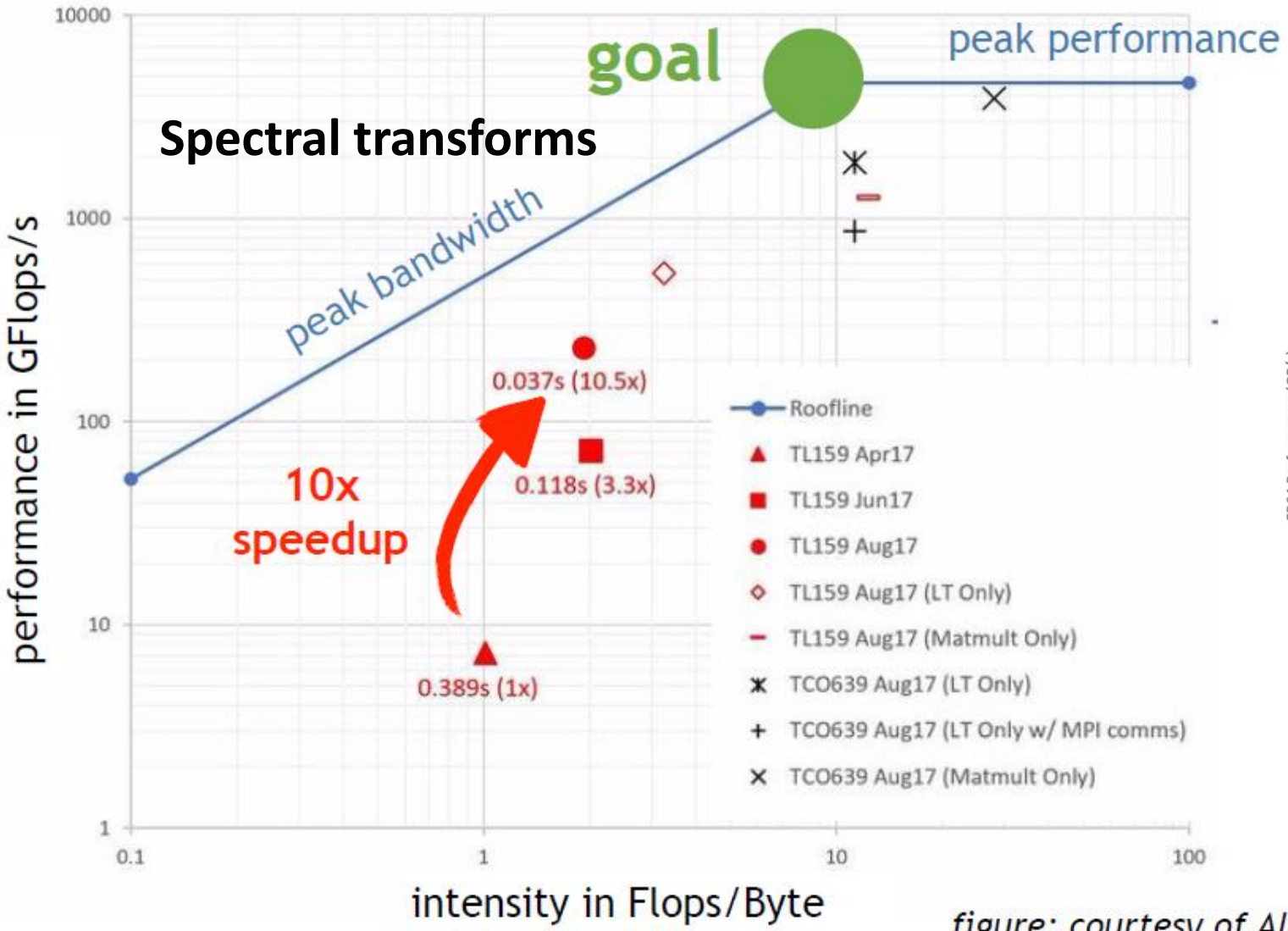
Comparison of software optimized for GPU and Xeon processors

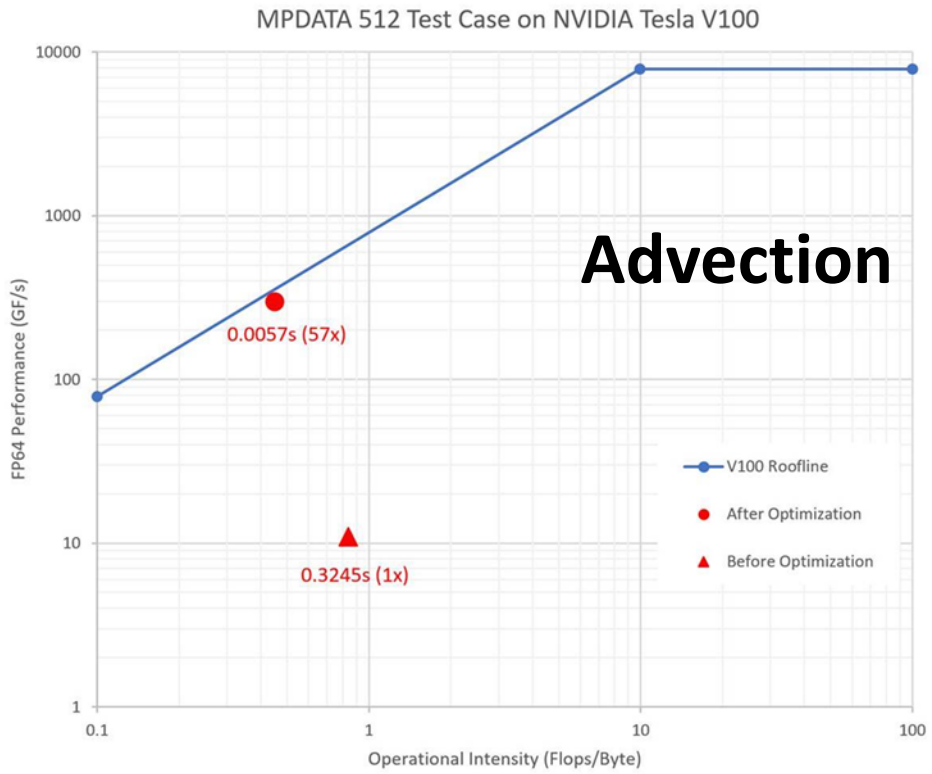


Poulsen & Berg (2017)

Analysing performance with roofline plots

Funded by the European Union

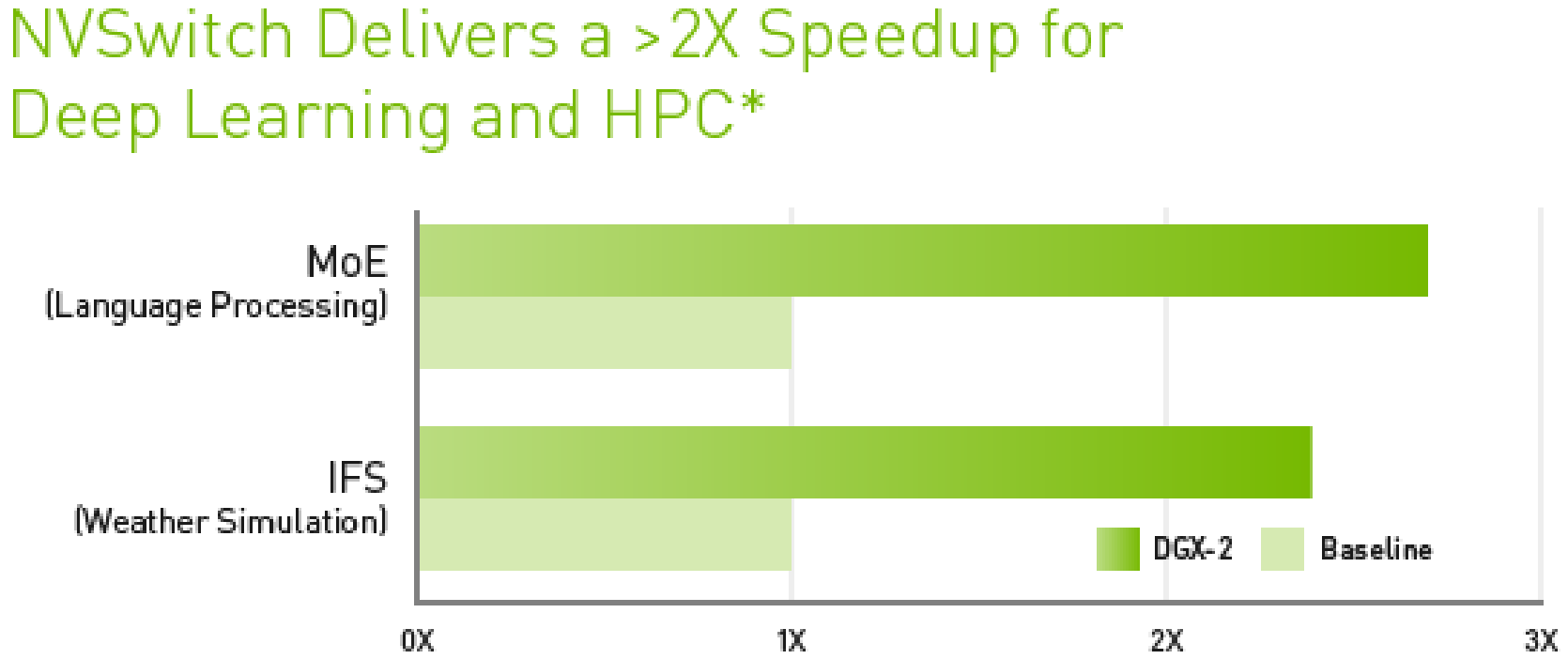




GPU speed-up

figure: courtesy of Alan Gray, Peter Messmer (NVIDIA)

Will Deep Learning influence algorithmic choices for weather & climate ?

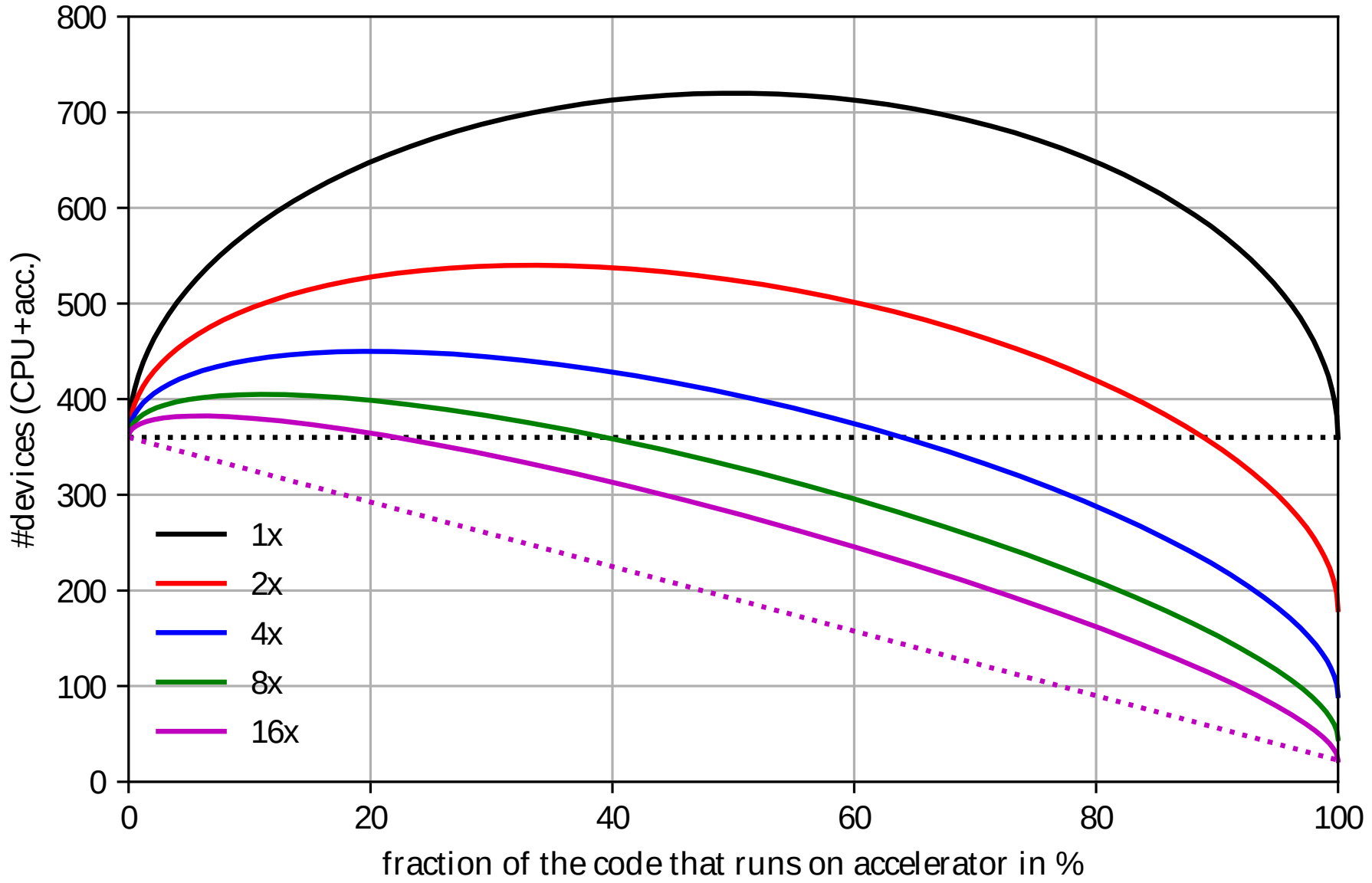


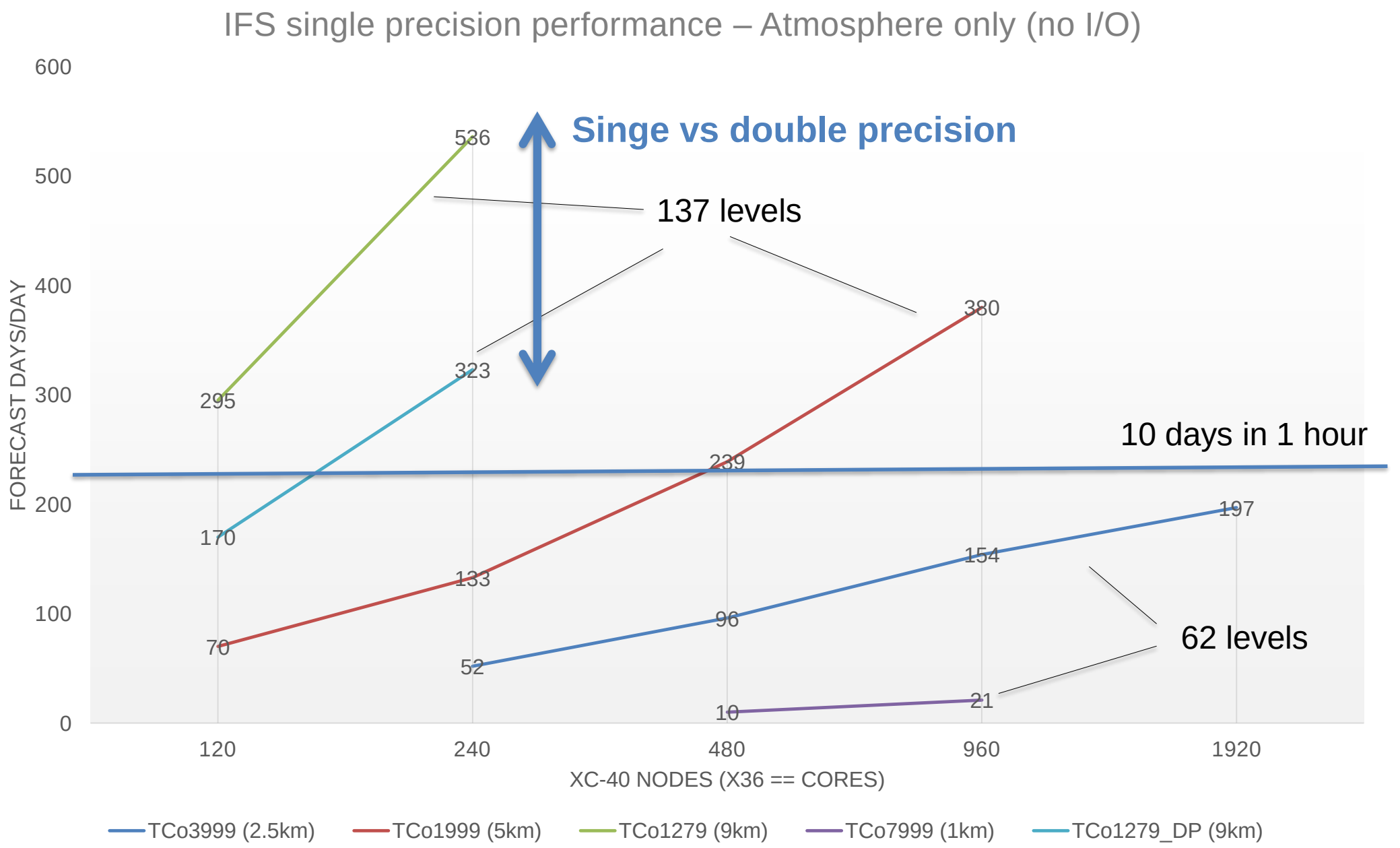
System Configs: Each of the two DGX-1 servers have dual-socket Xeon E5 2690v4 Processor, 8 x V100 GPUs; servers connected via a 4 EDR (100Gb) InfiniBand connections. DGX-2 server has dual-socket Xeon Scalable Processor Platinum 8168 Processors, 16 x Tesla V100 GPUs.

https://news.developer.nvidia.com/nvswitch-leveraging-nvlink-to-maximum-effect/

Funded by the European Union

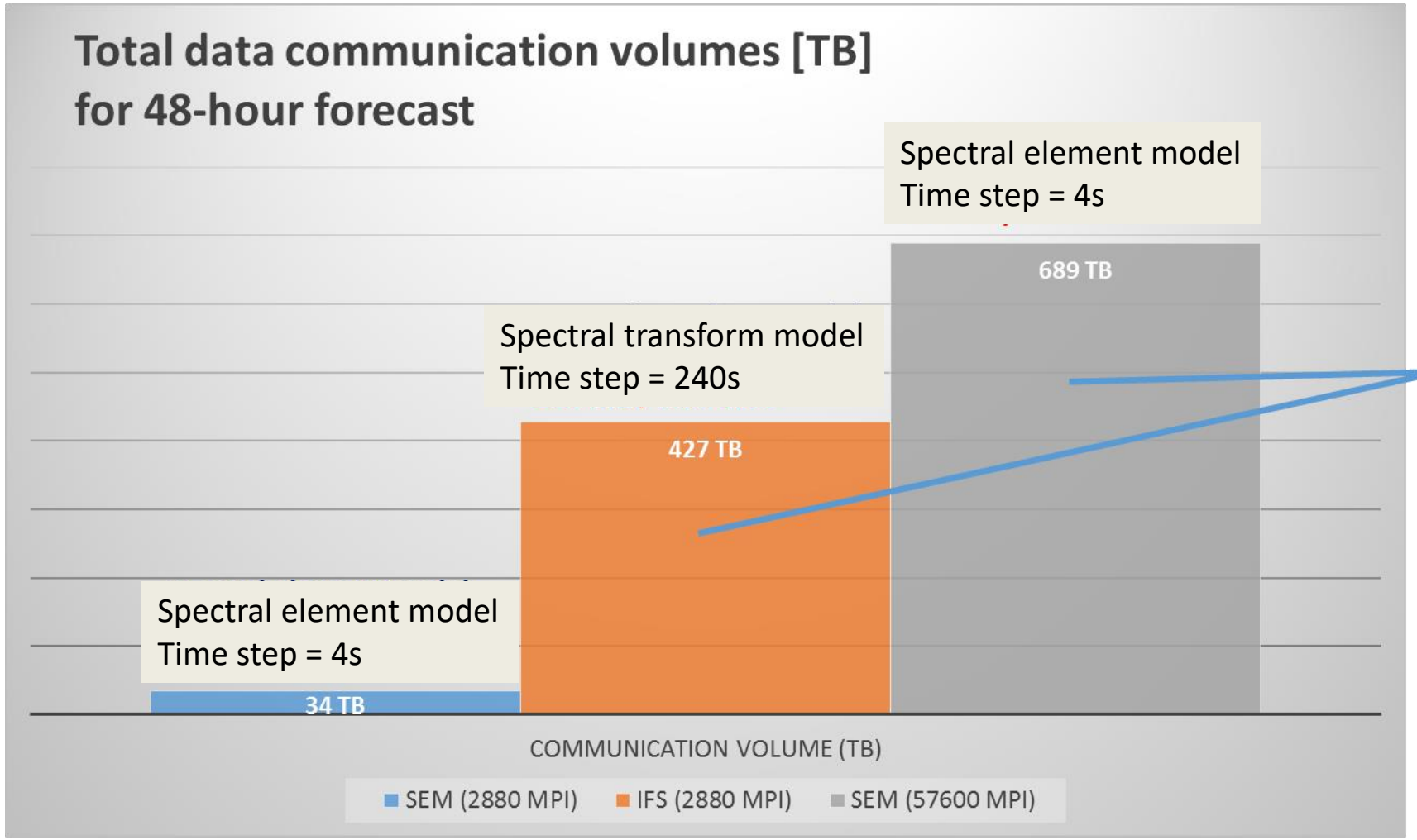
Benefit of accelerators – theoretical model
number of devices (acc.+CPU) at MétéoFrance





(Vana, Dueben et al 2017)

Communication is bad – small time steps are worse



Same time to solution!
Energy efficiency?

Data movement x100 (x1000) more expensive than computations in time (energy)!

[Shalf et al. 2011]

The state of atmospheric models

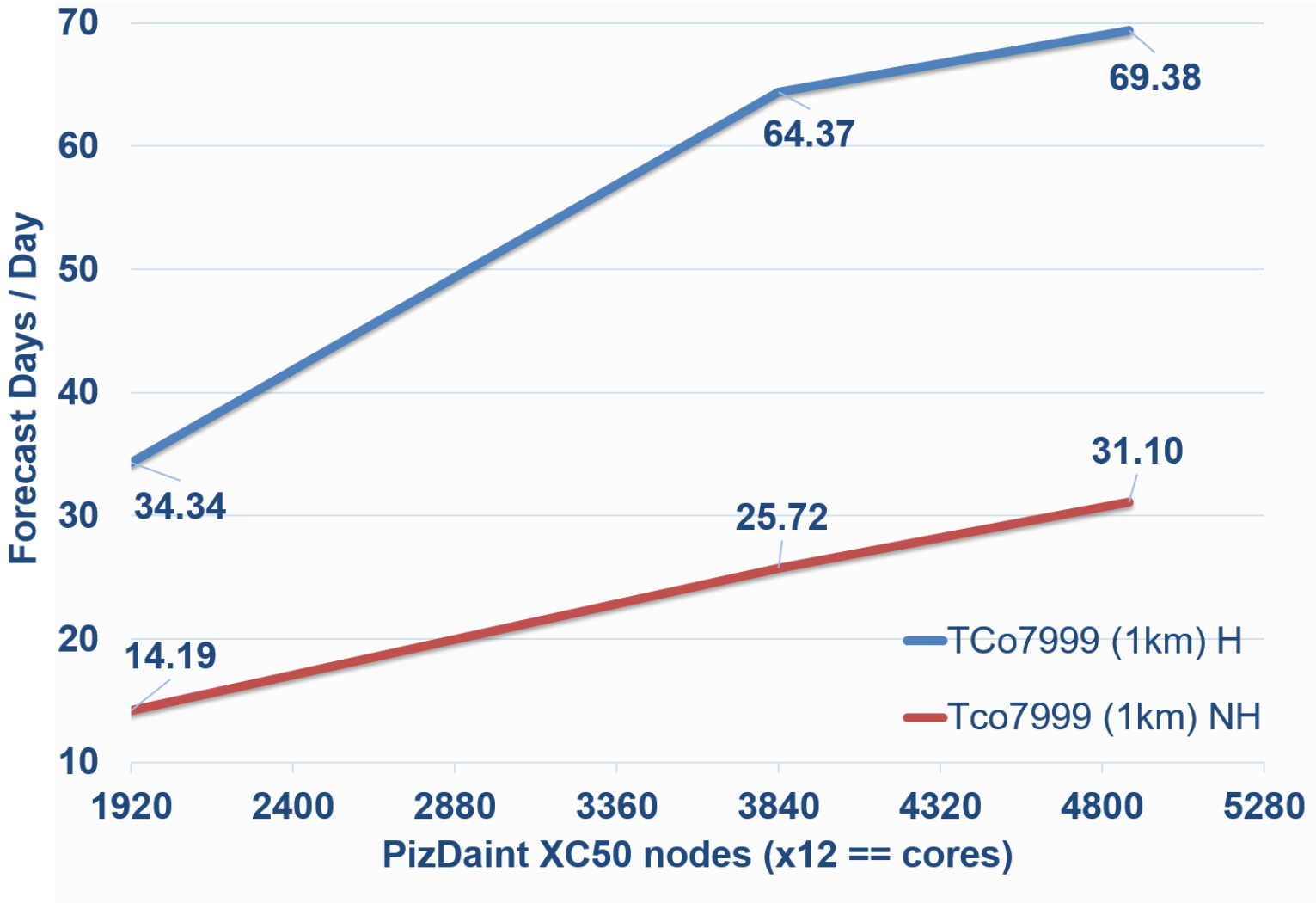
- 1) Is there a scalability / performance / code adaptation problem for global atmospheric models and if so for which models do you consider this a problem and why ?
  - Trend away from spectral models to FV or higher-order FE/SE/DG; however global spectral models still highly competitive; reverse trend due to machine learning ?
- 2) What in the structure or in scientific choices for atmosphere models (or other non-atmosphere components) are particular scalability challenges ?
  - Large-timestep global (semi-implicit) solvers; standard lat-lon grids on the sphere; Ocean/sea-ice/wave ; strong coupling in DA and code suitability for DA frameworks and their scalability; chemistry and biogeochemistry; unstructured grids
- 3) Which global atmosphere model developments are leading efforts on improving performance / adaptation for future HPC ?
  - See separate slide
- 4) Which complete global atmosphere model demonstrated to run at km-scale and ~90-200 vertical levels ?
  - Arpege/IFS, FV3, ICON, NICAM, MPAS, COSMO (near global), see also DYAMOND project
- 5) What computational resources are required to achieve this ? State computational performance if possible (forecast days / day; number of cores ; programming model [e.g. MPI OpenMP hybrid], state MPI task / thread ratio if applicable)
  - See separate slide

The state of atmospheric models

- 6) What do you consider the most scalable ocean model today, e.g. providing the fastest time-to-solution, providing the best cost-benefit ratio for the global ocean/sea-ice problem ? Is this view based on an actual intercomparison of computational performance ?
  - NGGPS intercomparison; scalability vs time-to-solution; adaptation to accelerators ?
  - Hydrostatic global spectral still highly competitve
- 7) What do you think should be or is already done to improve the performance of global atmospheric models ?
  - Development of novel numerical methods that maximise time-stepping size and spatial (and/or temporal) scalability
  - Development of code adaptation toolchains
  - Machine learning tools to replace parts of the code
  - Tools for convenient overlapping of computation and compute including task-based parallel programming
- 8) The very important aspect of coupling is considered separately, but if you have any comments on the performance / best practice relevant to the computational performance of coupled simulations that you consider important please state.
  - Code refactoring for strong ESM coupling

IFS 1km: strong scaling on PizDaint

Goal ~1 year / Day



Hybrid MPI/OpenMP
4880x12

Many thanks to
Thomas Schulthess &
Maria Grazia Giuffreda !

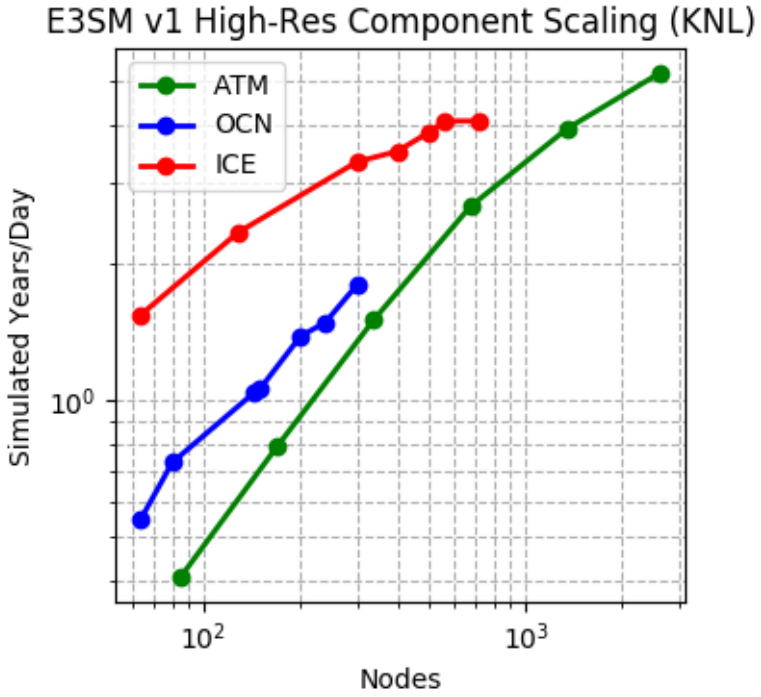
Questionnaire on the state of ocean/sea-ice models

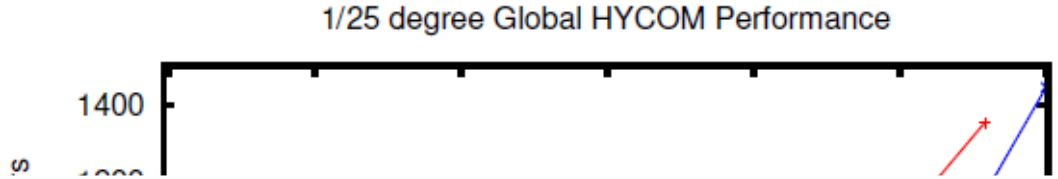
- 1) Is there a scalability / performance / code adaptation problem for global ocean models and if so for which models do you consider this a problem and why ?
  - All models primary limiters: 1) sea-ice and barotropic mode fast mode coupling (not an issue up to ~6000 cores ?), 2) stencil operators with very low flow-to-memory ratio, limited by memory bandwidth, lack of efficient cache/memory access patterns 3) load imbalance (in particular sea –ice ?)
  - NEMO: grid choice (north folding issue on tripolar grid leading to load imbalance)
  - Demonstrated good scaling up to about ~500 nodes (x24-36 cores)
- 2) What in the structure or in scientific choices for ocean and sea-ice models (or other non-atmosphere components) are particular scalability challenges ?
  - 2d solution algorithms for fast barotropic mode and sea-ice dynamics with frequent comms and load imbalance issues, insufficient work to “hide” by overlapping; use of collectives in controlling the simulation/diagnostics; grid choice
  - Barotropic mode less problematic in unstructured code framework ?
- 3) Which global ocean model developments are leading efforts on improving performance / adaptation for future HPC ?
  - Hybrid vertical coordinate (ALE) accepted standard (HYCOM/MOM6); reduces need for vertical levels
  - E3SM: GPU adaptation (factor 10x ?); new algorithms for barotropic mode and solvers; new programming models and higher level abstractions (kokkos; Legion; task-based programming)
  - NEMO: time-stepping; solvers; ALE; mixed-precision; MPI+OpenMP/OpenACC; DSLs; XIOS on I/O

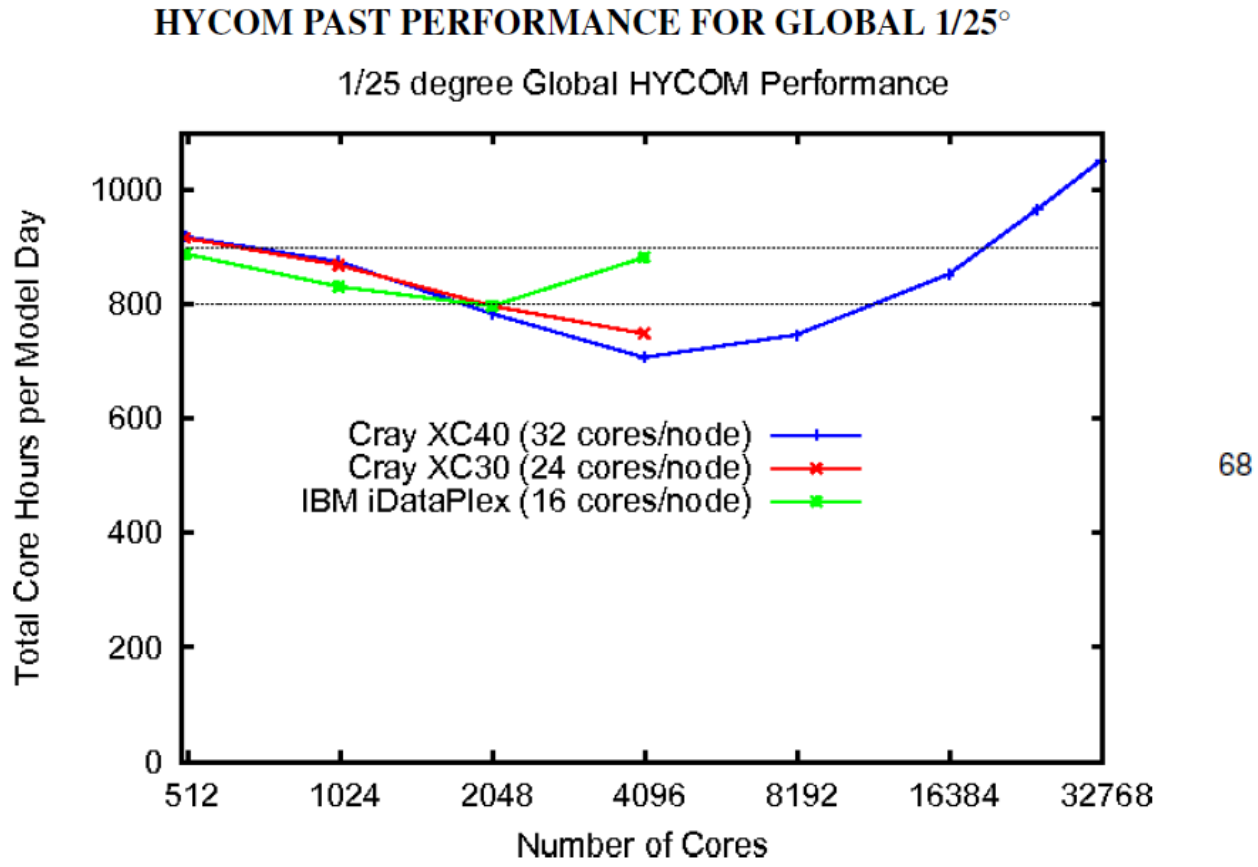
Questionnaire on the state of ocean/sea-ice models

- 4) Which global ocean models are able to run eddy-resolving, say greater or equal 1/36 degree, global problems (state also number of vertical levels) ?
  - Throughput good enough for climate (1SYPD): none
  - OK for climate at ~1/12 degree: MPAS-OCE/HYCOM/MOM6/NEMO
  - HYCOM + MOM6 run from 2019 global at 1/25
- 5) What computational resources are required to achieve this ? State computational performance if possible (ocean only, forecast days / day; number of cores ; programming model [e.g. MPI OpenMP hybrid], state MPI task / thread ratio if applicable)
  - See separate slide
- 6) What do you consider the most scalable ocean model today, e.g. providing the fastest time-to-solution, providing the best cost-benefit ratio for the global ocean/sea-ice problem ? Is this view based on an actual intercomparison of computational performance ?
  - No intercomparison exists, see above.

Ocean model scalability

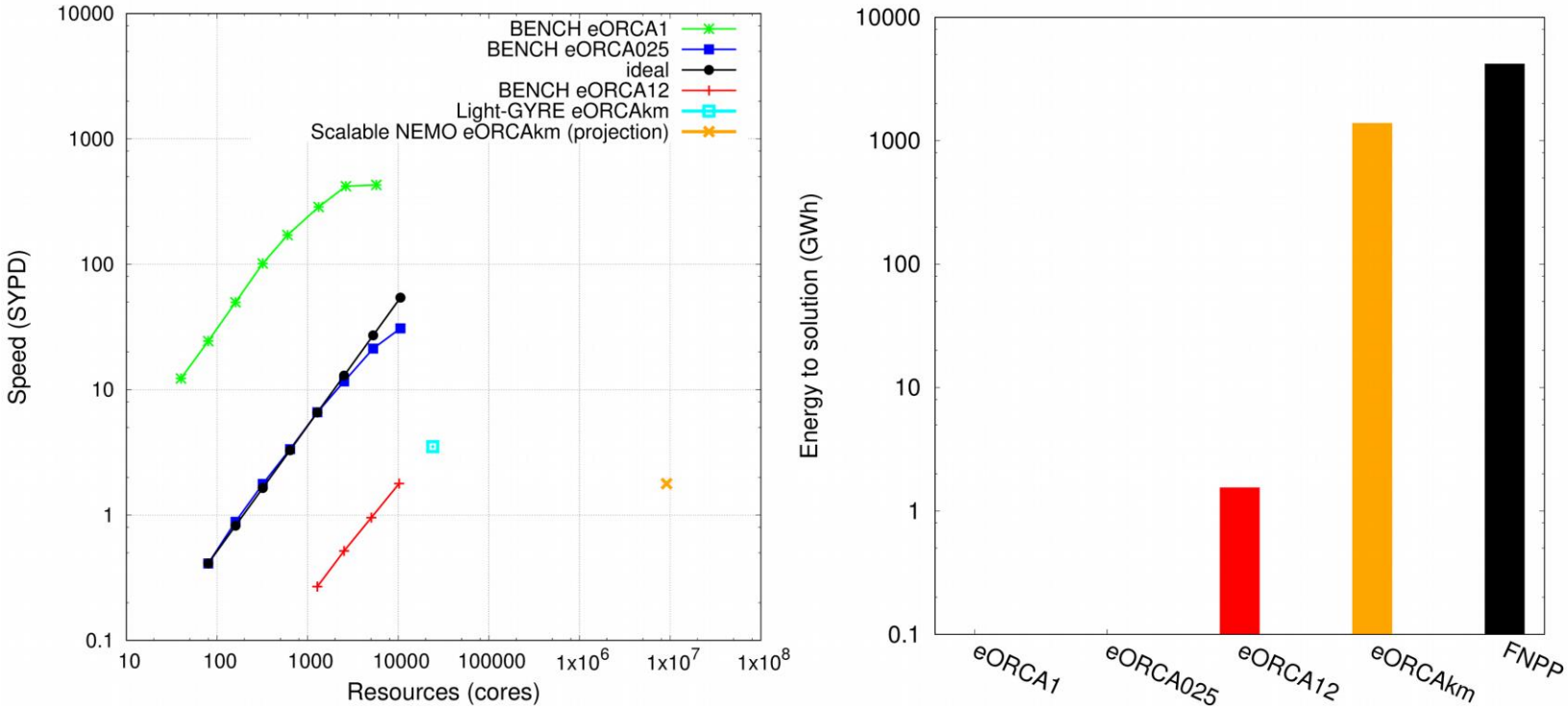






68

continued



Scalability (left) and energy consumption (right) of several NEMO 3.6 model configurations, measured on `beaufix2` Météo-France supercomputer, Intel Broadwell processors. BENCH and GYRE configurations (I) have 75 vertical levels, (ii) exclude realistic bathymetry (no effect on performance), sea-ice, bio-geo-chemistry and output but (iii) include TOP tracers, appropriate physics at each resolution and polar grid folding (BENCH only). Horizontal resolution varies from 1 degree (eORCA1) to 1 km (eORCAkm). Scalable NEMO eORCAkm performances are extrapolated from measurements of a simplified GYRE km scale configuration, assuming a perfect scalability until a 10 million MPI subdomain decomposition. Right figure compares the total production of one reactor of a power plant (FNPP) similar to Fessenheim, France and the energy consumption of a 1,000 year long simulation led with the four NEMO configurations (eORCAkm: projection) at maximum scalability, approximated as suggested in Balaji et al. 2017, assuming `beaufix2` consumption E = 2.15e12 J/month and total capacity A = 5.2e7 CH/month

Questionnaire on the state of ocean/sea-ice models

- 7) What do you think should be or is already done to improve the performance of global ocean models ?
  - Any factor beyond 2-3 will require programming model and algorithmic change, not just code adaptation
  - Test at high core counts to push limits
  - I/O
  - Reduce memory footprint
  - Measure energy consumption
  - Expose key algorithmic motives (dwarves) from ocean and sea-ice to a wider community (atmosphere, academia, vendors)
- 8) The very important aspect of coupling is considered separately, but if you have any comments on the performance / best practice relevant to the computational performance of coupled simulations that you consider important please state.
  - Sub-cycling and overlapping of components; coupling frequency and synchronisation of components
  - Load imbalance
  - Coupling processes rather than ESM components (task parallelism)

Science: Some key issues for modelling weather & climate

- Numerical methods
  - Observed trends: FV to replace spectral-transform; FD/FV in oceans
  - Alternatives: higher-order SE/FE/DG; some outstanding issues, potential for ocean/sea-ice
  - Research on time-stepping methods; parallel-in-time methods
- Coupling strategies
  - Framework developments for coupled data assimilation and code refactoring to fit these
  - Machine learning algorithms integrated into models
  - ESM component strong coupling of fast processes (diurnal cycle)
- Parametrization development
  - Convection parametrization with very high resolution simulations (km-scale)
  - Eddy permitting/resolving in oceans
  - Machine learning to improve or for speed-up
  - Increasing Earth-system component complexity (carbon cycle; hydrology; biogeochemistry)

HPC: Key issues for modelling weather & climate

- Efficient use of emerging energy-efficient hardware
  - Accelerator use; overlapping computation and compute; hierarchical memory; task-based compute; …
- Hardware agnostic approaches to coding
  - Defining and encapsulating the fundamental algorithmic building blocks ("Weather and Climate Dwarfs")
  - Pioneering algorithm development with hardware adaptation using DSL toolchains (GridTools; Atlas; Kokkos; PsyClone …)
  - Verification, Validation, and Uncertainty Quantification (VVUQ) framework
- Trade-off stability, accuracy, resilience and computational performance
  - High (spatial) scalability combined with large time-step solutions
  - Reduce selectively numerical precision
- Machine learning approaches to replace part or all of the model codes
  - Utilising high-resolution observing systems, targeted LES, climate data, …
- Harnessing data flows; cloud computing

Conclusions

- Add questionnaire more specific on sea-ice, chemistry/aerosol, biogeochemistry, regional modelling ?
- Establish a closer connection and sharing of key algorithms across ESM components (atmosphere, ocean, sea-ice, chemistry, bio-geochemistry etc.)
  - First step to make easily runnable, optimizable, verifiable dwarfs of key algorithmic motives from ALL the different ESM components and share with academia, vendors, partners … (ESCAPE/ESCAPE-2/ESRL/…)
- Embrace the use of tools and libraries for code parsing, refactoring, and/ code generation
- Embrace new programming paradigms (ML, task-based parallelism, overlapping computations, etc)
- Assess in detail the need for precision of all algorithms and data

Additional slides