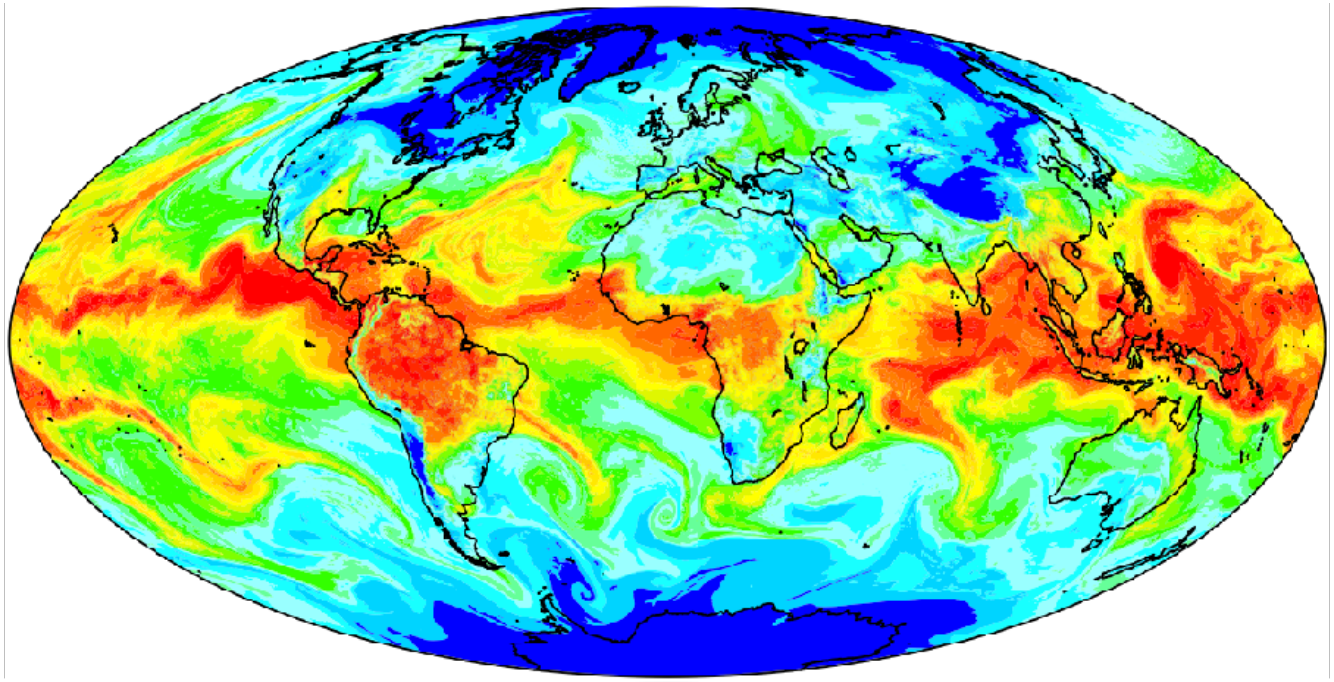
Variable-resolution modelling and extreme scaling applications with the Model for Prediction Across Scales (MPAS)

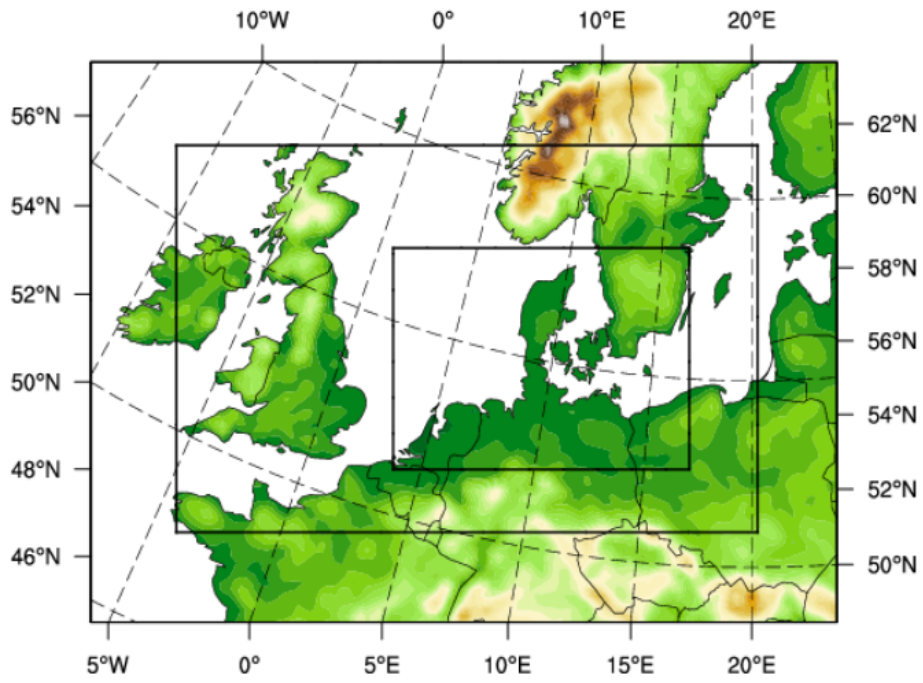


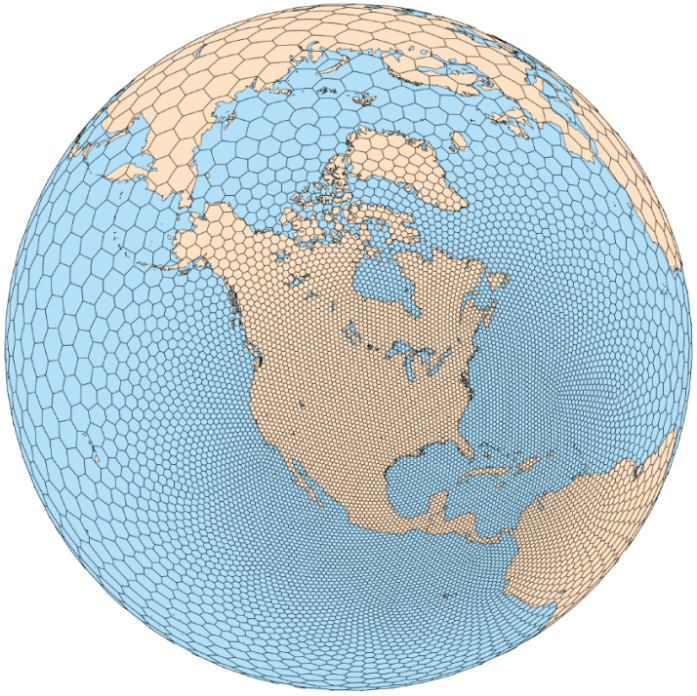
Dominikus Heinzeller[1,2], Michael Duda[3], Matthijs Kramer[4,5], Hugo Hartmann[4], Wim van den Berg[4], Gert-Jan Steeneveld[5], Harald Kunstmann[1,2]

[1] KARLSRUHE INSTITUTE OF TECHNOLOGY, INSTITUTE OF METEOROLOGY AND CLIMATE RESEARCH
[2] AUGSBURG UNIVERSITY, INSTITUTE OF GEOGRAPHY, CHAIR OF REGIONAL CLIMATE AND HYDROLOGY
[3] NATIONAL CENTER FOR ATMOSPHERIC RESEARCH, MESOSCALE AND MICROSCALE METEOROLOGY LABORATORY
[4] METEOGROUP, RESEARCH DEPARTMENT
[5] WAGENINGEN UNIVERSITY, METEOROLOGY AND AIR QUALITY SECTION

Going global and variable - more bang for the buck?

MPAS
Model for Prediction Across Scales





WRF

- Regional, nested modelling
- Artefacts at edges (reflection of waves)
- Boundary forcing from external model
- Multiple domains for global applications
- Highly customisable and widely used

MPAS

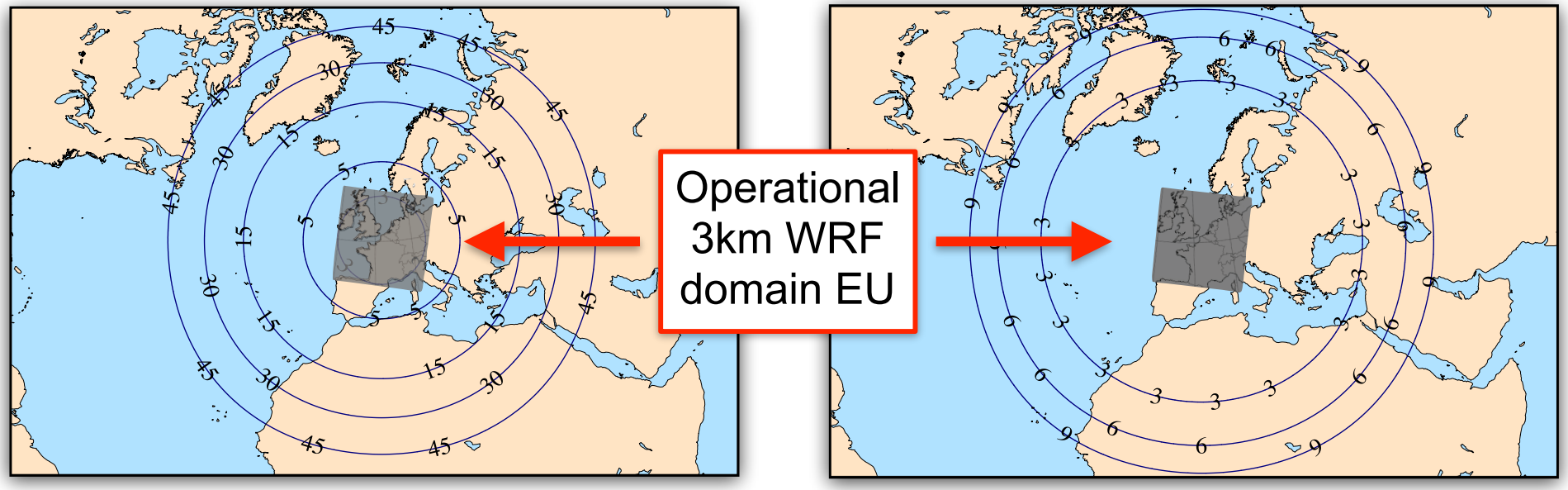
- Global, (ir)regular Voronoi grid
- Smooth transitions (local filters)
- No boundary conditions
- Multiple high-res areas in one domain
- Techniques and schemes from WRF

An NWP study of three selected events over Europe

MPAS
Model for Prediction Across Scales

- 72h and 84h short-term forecasts
- Variable-resolution meshes transitioning the grey zone, using the Grell & Freitas (2014) cu scheme
- Uniform 3km mesh as reference model (4 x 72h fcst)
- Validation against operational WRF configuration at MeteoGroup, with Wageningen University

Mesh	nCells	Conv
60-3km	835,586	GF
30-3km	1,294,335	GF
15-3km	6,488,066	GF
3km	65,536,002	GF/off
WRF 3km	612 x 612	off



60-3km mesh with high-resolution area centred over Europe (sim. for 30-3km)

15-3km mesh with high-resolution area centred over Europe

Storm, foehn, hail - three selected synoptic situations

MPAS
Model for Prediction Across Scales

- **Convective storm (front) case, Netherlands, October 2013**

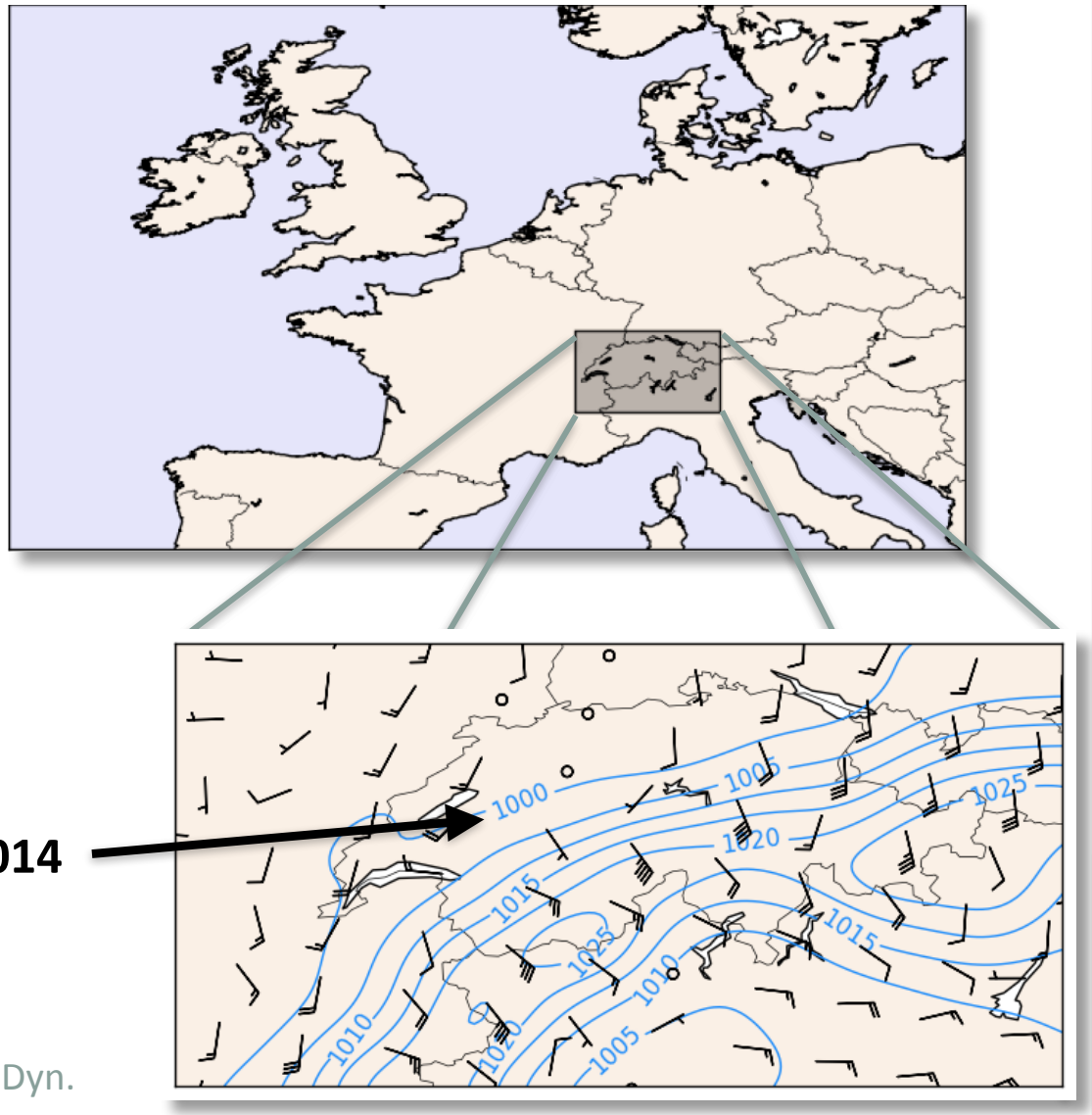
  strong depression (980 hPa),
  multiple frontal systems
  caused heavy winds
  and casualties/damage

- **Hail case, Netherlands, December 2015**

  stationary warm front,
  heavy thunderstorms,
  hail up to 6 cm and winds up to 34 m/s

- **Foehn case, Switzerland, November 2014**

  high winds, topographic lifting



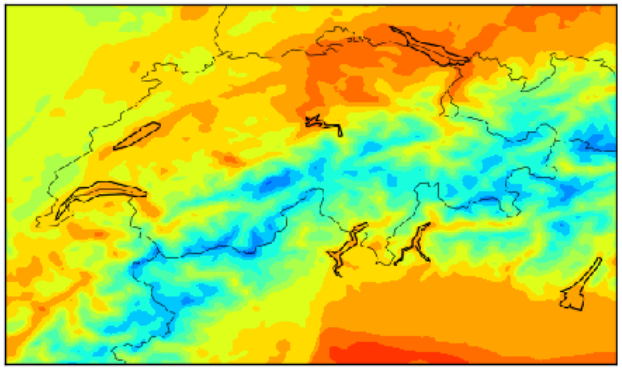
Kramer et al., in preparation for special issue in Clim. Dyn.

Foehn case, Switzerland, Nov. 2014: model results I

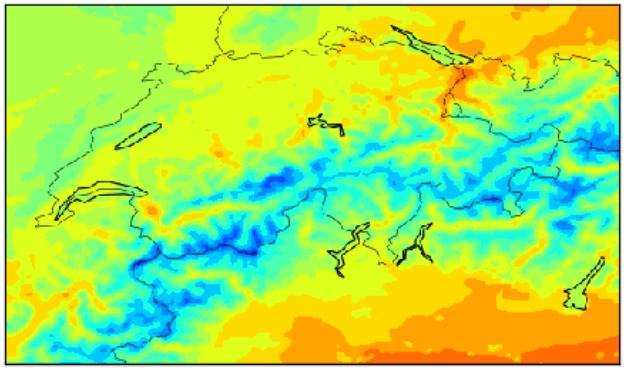
MPAS
Model for Prediction Across Scales

2m temperature [°C], Nov 4 10:00UTC

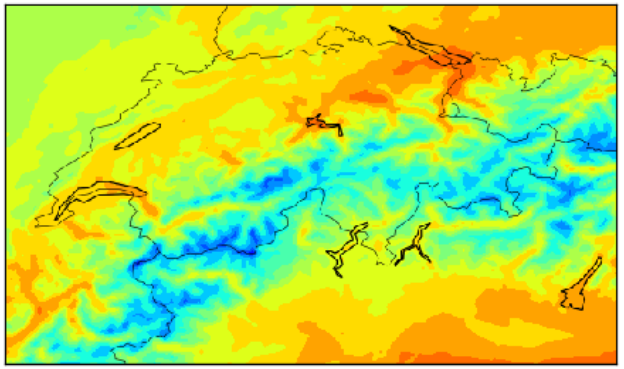
WRF 3km 03_00



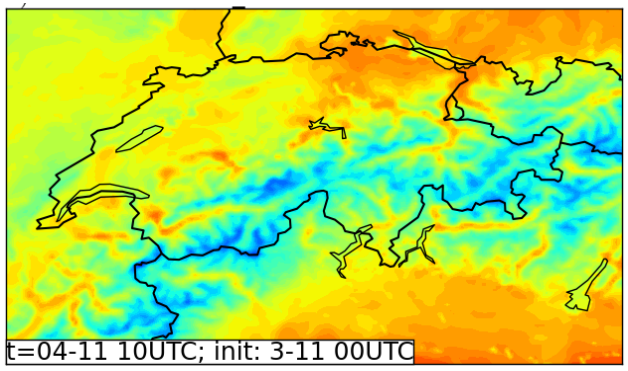
MPAS 60-3km 03_00



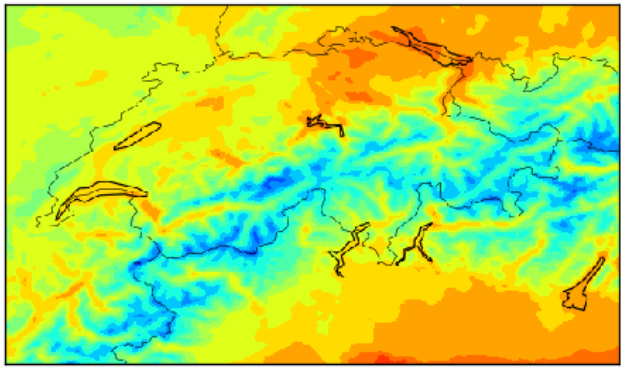
MPAS 60-3km 03_12

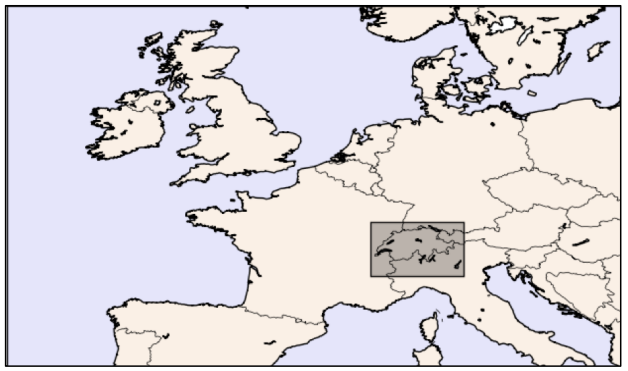


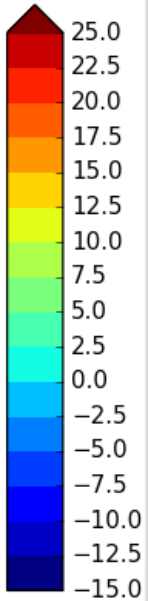
MPAS 30-3km 03_00



MPAS 3km 03_12





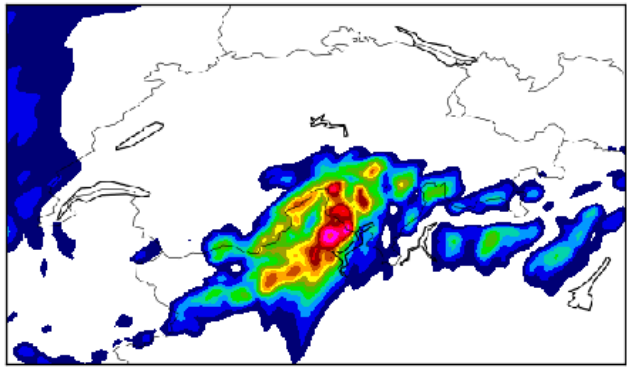


Foehn case, Switzerland, Nov. 2014: model results II

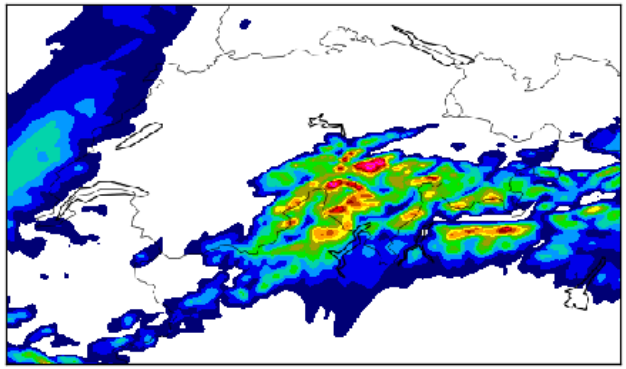
MPAS
Model for Prediction Across Scales

Cumulative precipitation [mm] since initialisation, Nov 6 00:00UTC

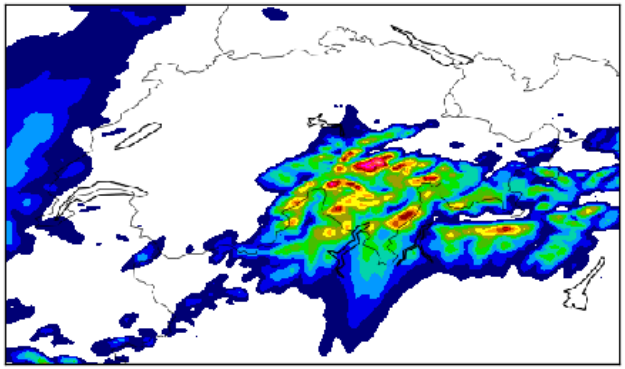
WRF 3km 03_00



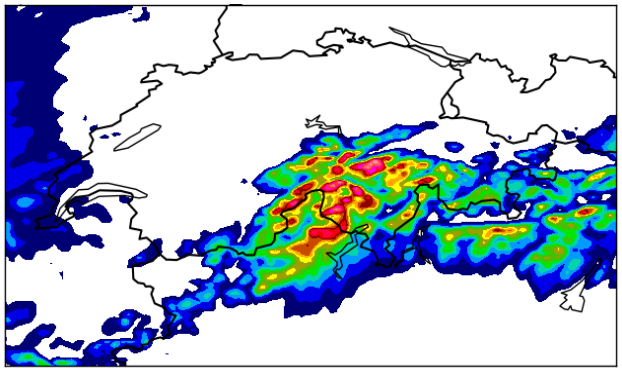
MPAS 60-3km 03_00



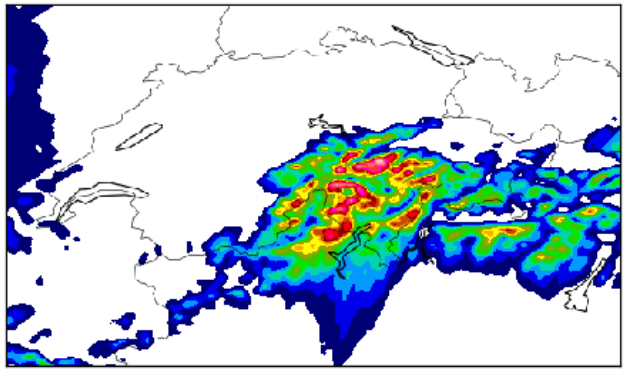
MPAS 60-3km 03_12

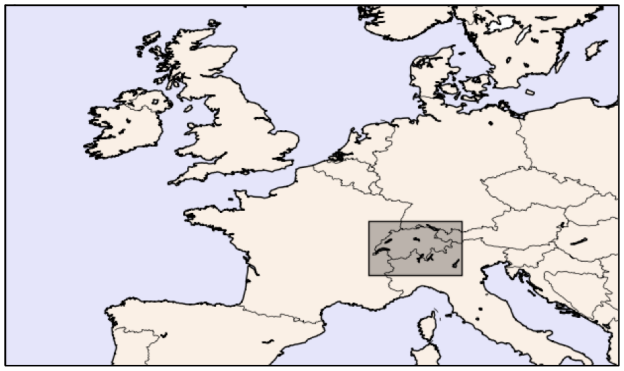


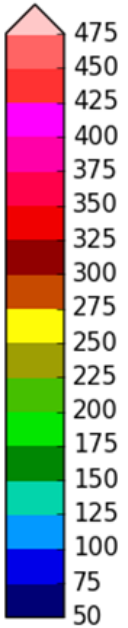
MPAS 30-3km 03_00



MPAS 3km 03_12

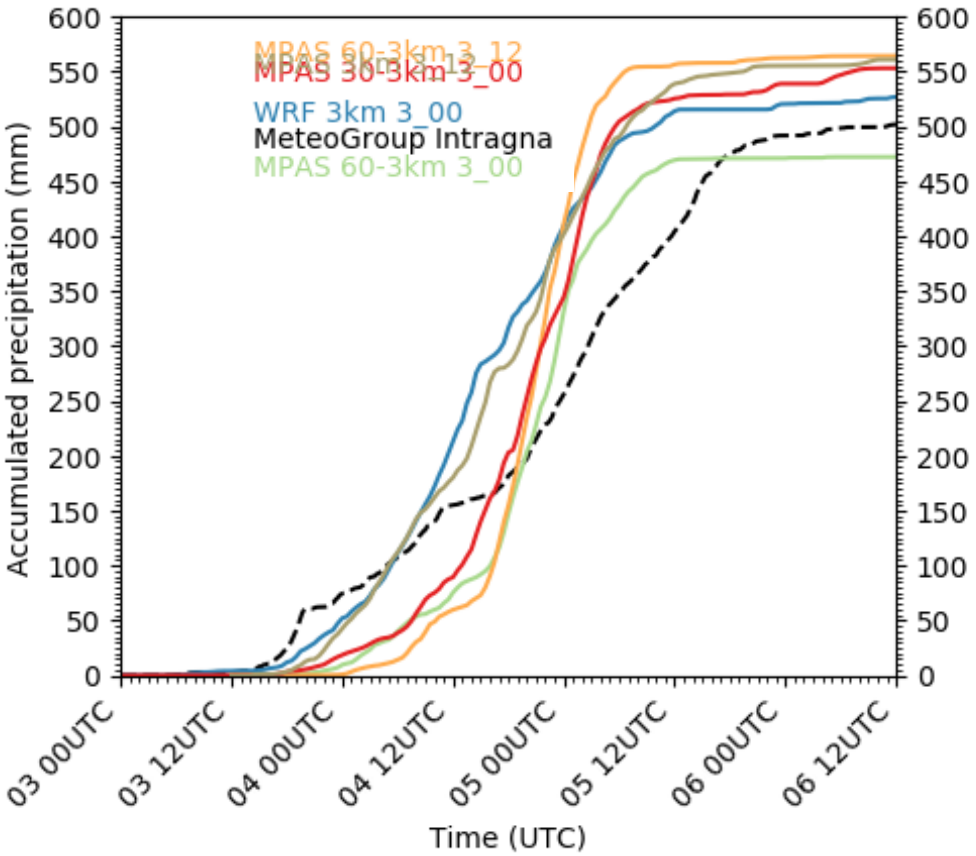


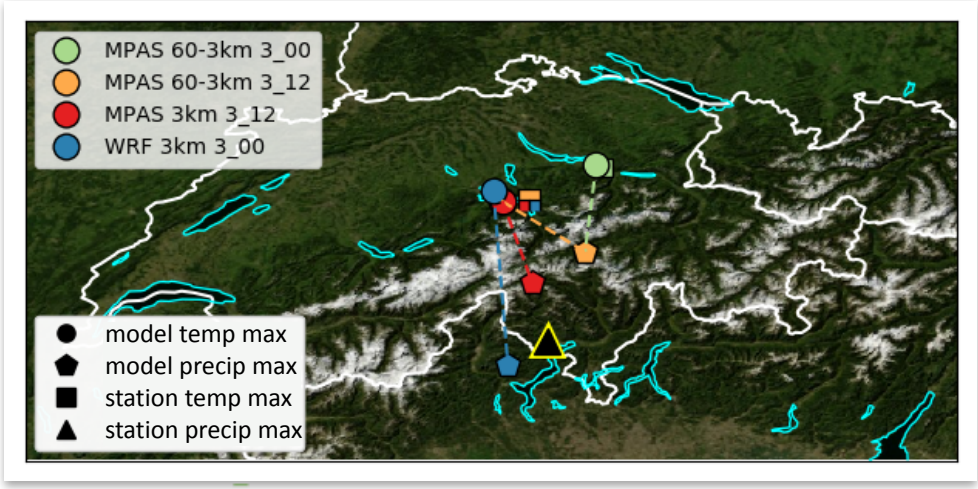




Foehn case, Switzerland, Nov. 2014: model results III

Time series of accumulated precipitation at location of maximum precipitation





- Selecting single location for model and observations to focus on extreme character
- Onset late for all runs, best for WRF 3km/MPAS 30-3km (startup too close to event?)
- 15-3km runs work in progress

Exascaling is projected for the end of this decade

MPAS
Model for Prediction Across Scales

The leading operational NWP centres (ECMWF, NOAA) are pushing towards convection-permitting, global forecasts.

- ECMWF is working on the FVM (Finite Volume Model)
- NOAA has chosen the GFDL FV3 model to replace GFS

As individual CPU cores are not getting any faster, next-generation HPC systems will scale out to even larger numbers of cores and resort to accelerators and many-core chips.

K computer,
Kobe, Japan

Show stoppers on next generation HPC environments

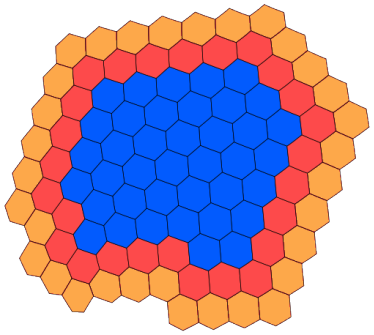
MPAS
Model for Prediction Across Scales

Disk I/O bottleneck

Highly scalable and fast I/O libraries such as SIONlib for internal (and external) data.

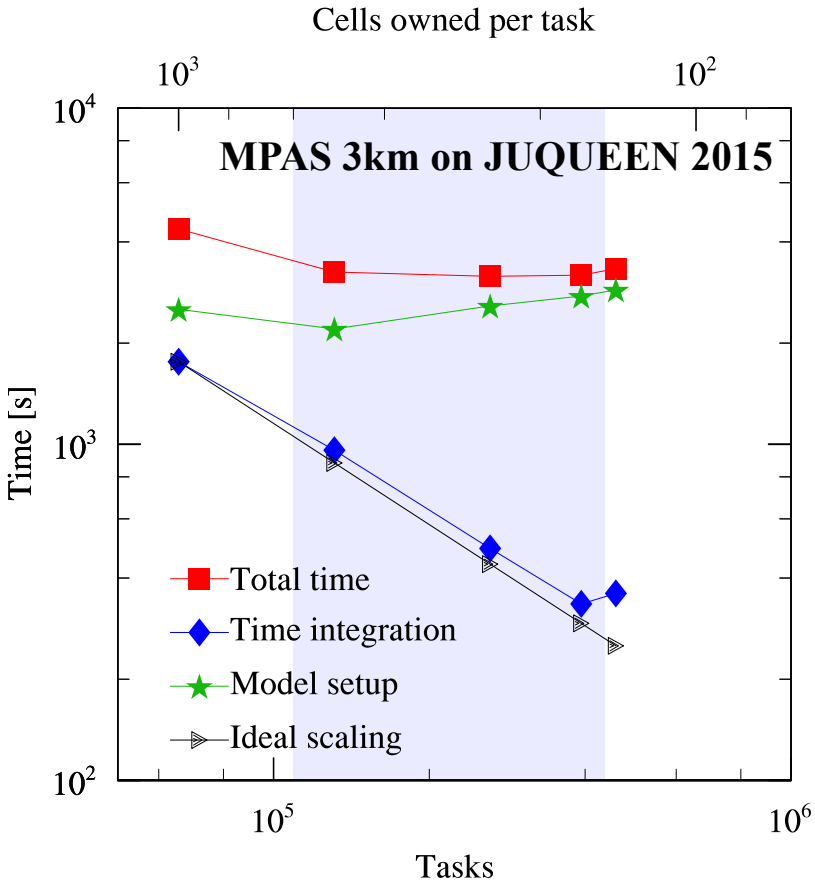
MPI communication overhead

Hybrid MPI+OpenMP: fewer tasks sending data (key for porting to Knights Landing)



Block of owned cells N_o and 2 layers of halo cells N_h

$$N_h = \pi\left(\sqrt{\frac{N_o}{\pi}} + 2\right)^2 - N_o$$



KONWIHR (Competence Network for Scientific High Performance Computing) grant for the *Improvement of I/O layer and hybrid parallelisation of the Model for Prediction Across Scales*.

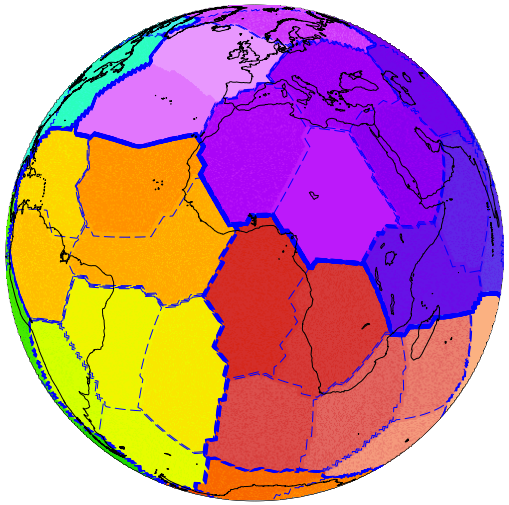
Implementation of SIONlib I/O layer in MPAS v4.0+

SIONlib - scalable I/O library for (massively) parallel access to task-local files (Wolfgang Frings, FZJ)

Addition of SIONlib I/O layer to currently supported I/O formats (netCDF, pnetCDF, netCDF4 through PIO), mimicking netCDF file structure (metadata, data).

Reading/writing in SIONlib format requires to use the same number of tasks and the same graph partition.

Information encoded in SIONlib data can be used to skip parts of the bootstrapping at model startup.



http://www.fz-juelich.de/ias/jsc/EN/Expertise/Support/Software/SIONlib/_node.html

```
<stream name="history"
        type="output"
        io_type="pnetcdf,cdf5"
        filename_template="history.nc"
        output_interval="03:00:00">
    <file name="stream_list.history"/>
</stream>
```

```
<stream name="diagnostics"
        type="output"
        io_type="sionlib"
        filename_template="diagnostics.sl"
        output_interval="00:15:00">
    <file name="stream_list.diagnostics/>
</stream>
```

Global 2km mesh on LRZ SuperMUC

MPAS
Model for Prediction Across Scales

Uniform 2km mesh with 147 Mio grid cells,
2048 nodes x 16 MPI x 1 OpenMP tasks

```
-rw-r--r-- 1 di73bim2 pr94mi 120G Jan 19 23:40 diag.2013-10-27_12.00.00.nc
-rw-r--r-- 1 di73bim2 pr94mi 120G Jan 19 23:54 diag.2013-10-27_12.05.00.nc
-rw-r--r-- 1 di73bim2 pr94mi 120G Jan 20 00:13 diag.2013-10-27_12.10.00.nc
-rw-r--r-- 1 di73bim2 pr94mi 3.0T Jan 19 23:39 history.2013-10-27_12.00.00.nc
-rw-r--r-- 1 di73bim2 pr94mi 3.0T Jan 20 00:12 history.2013-10-27_12.10.00.nc
```

pnetcdf, cdf5

	timer name	total
1	total time	3585
2	initialise	1176
3	bootstrapping	540
3	stream input	612
2	time integration	1580
2	stream output	818

sionlib

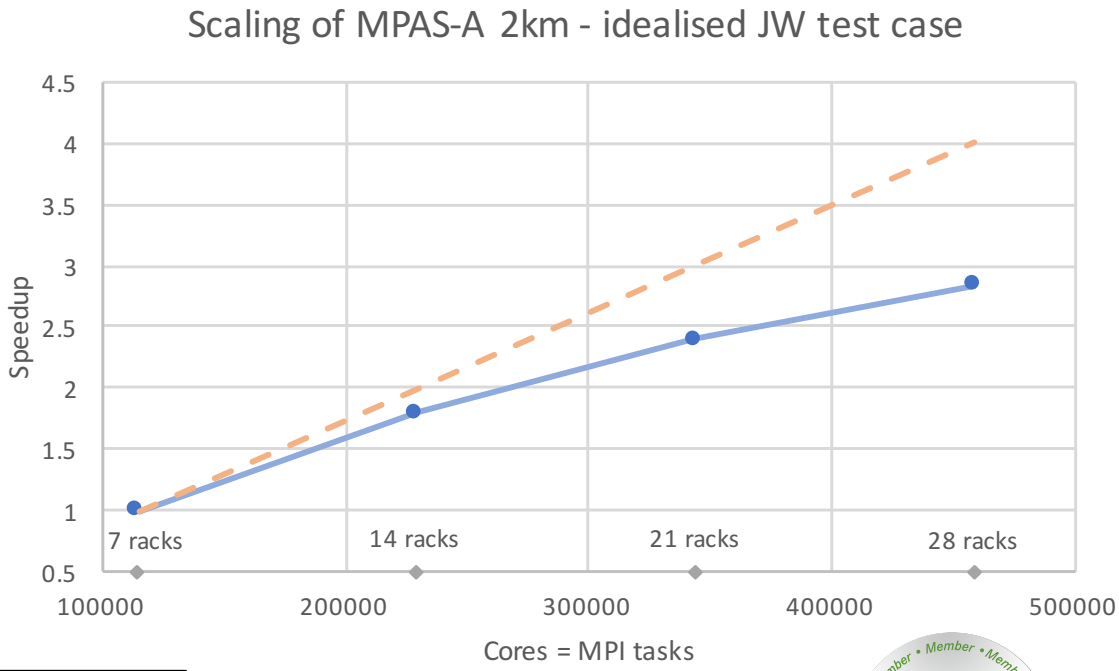
	timer name	total
1	total time	2117
2	initialise	244
3	bootstrapping	168
3	stream input	52
2	time integration	1658
2	stream output	204

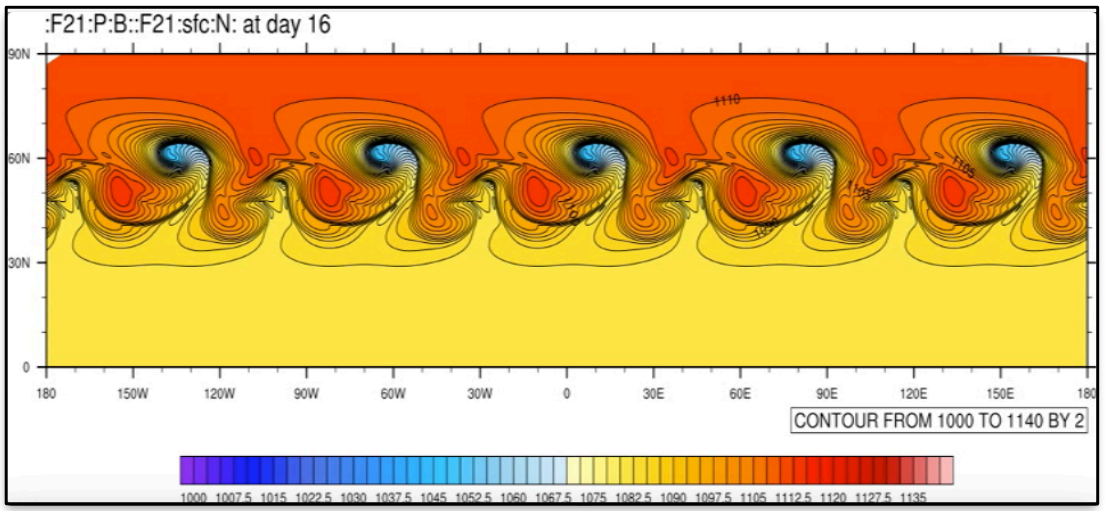
Extreme scaling 2017: 2km, 147 Mio cells on JUQUEEN

MPAS
Model for Prediction Across Scales

Idealised test case: Jablonowski and Williamson (2006) baroclinic wave

- Regular 2km mesh, 147 Mio cells
- Initial conditions file 1.8TB SIONlib
- 10min model integration
- Disk output (SIONlib): 4TB in total
- Dynamics + I/O, no physics: more stringent test of dynamical solver





Conclusions & Outlook

MPAS
Model for Prediction Across Scales

Variable-resolution models can address limitations of nested models at reasonable cost. Scale-aware physics parameterisations need further attention, and further performance improvements are required to make MPAS run as fast or faster than WRF (per grid cell).

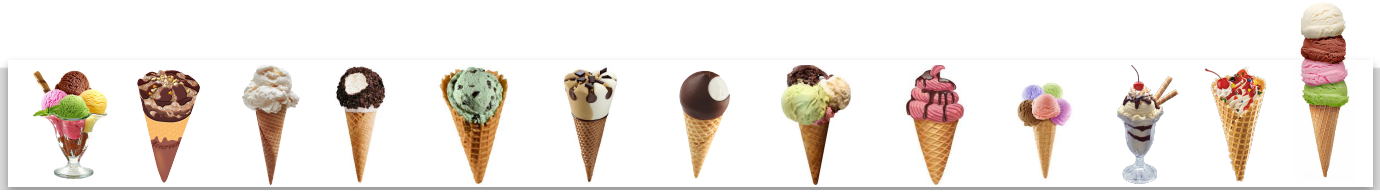
Kramer et al., in prep.

Global, convection-resolving atmospheric simulations are within reach of current and next-generation HPC systems. Keys to success are an efficient I/O layer for massively parallel applications and an adaptation for many-core processors and accelerators.

Heinzeller et al., in prep.

What's next? Stay tuned for MPAS 1km on HLRS Hazel Hen (Cray CX40)!
580 million grid cells, minimum 500TB memory, initial conditions approx. 12 TB

Karlsruhe Institute of Technology
Los Alamos NATIONAL LABORATORY
KONWIHR
WAGENINGEN UR
Universität Augsburg University
lrz
SCC Steinbuch Centre for Computing
KIT-Campus Alpin IMK-IFU: Atmospheric Environmental Research
NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH
MeteoGroup



Bonus features

Runtime performance MPAS versus WRF

MPAS
Model for Prediction Across Scales

MPAS 60-3km: scaling of the time integration (dynamics + physics + file output) as function of tasks (bottom) or cells per task (top).

Using a scalable I/O library, the performance of the time-integration depends only on the number of cells per task (mesh-independent).

WRF setup optimised on MeteoGroup cluster, MPAS out of the box on SuperMUC.

Configuration	WRF	MPAS
Vertical levels	42	55
Microphysics	WSM6	Thompson
Cumulus	-	Grell-Freitas
Boundary layer	MYNN	MYNN
Longwave radiation	RRTMG	RRTMG
Shortwave radiation	RRTMG	RRTMG
Surface layer	MYNN	MYNN
Land-surface	NOAH	NOAH

36h forecast scenario		Mesh	nCells	Cores required	Core-hours fcst
Integration time [h]	36	3km	65,536,002	62,832	251,328
Time to solution [h]	4	15-3km	386	6176	24,704
Cells per task	1050	30-3km	77	1232	4928
Cores per node	16	60-3km	50	800	3200
Cores per node (WRF)	10	WRF 3km	7	70	280

Optimisation of hybrid MPI+OpenMP parallelisation

MPAS
Model for Prediction Across Scales

In MPAS-Atmosphere, threading using OpenMP inside MPI is implemented for the time-integration routine only. A simple change to the hybrid implementation and threading of one additional routine yield impressive performance improvements.

Original code

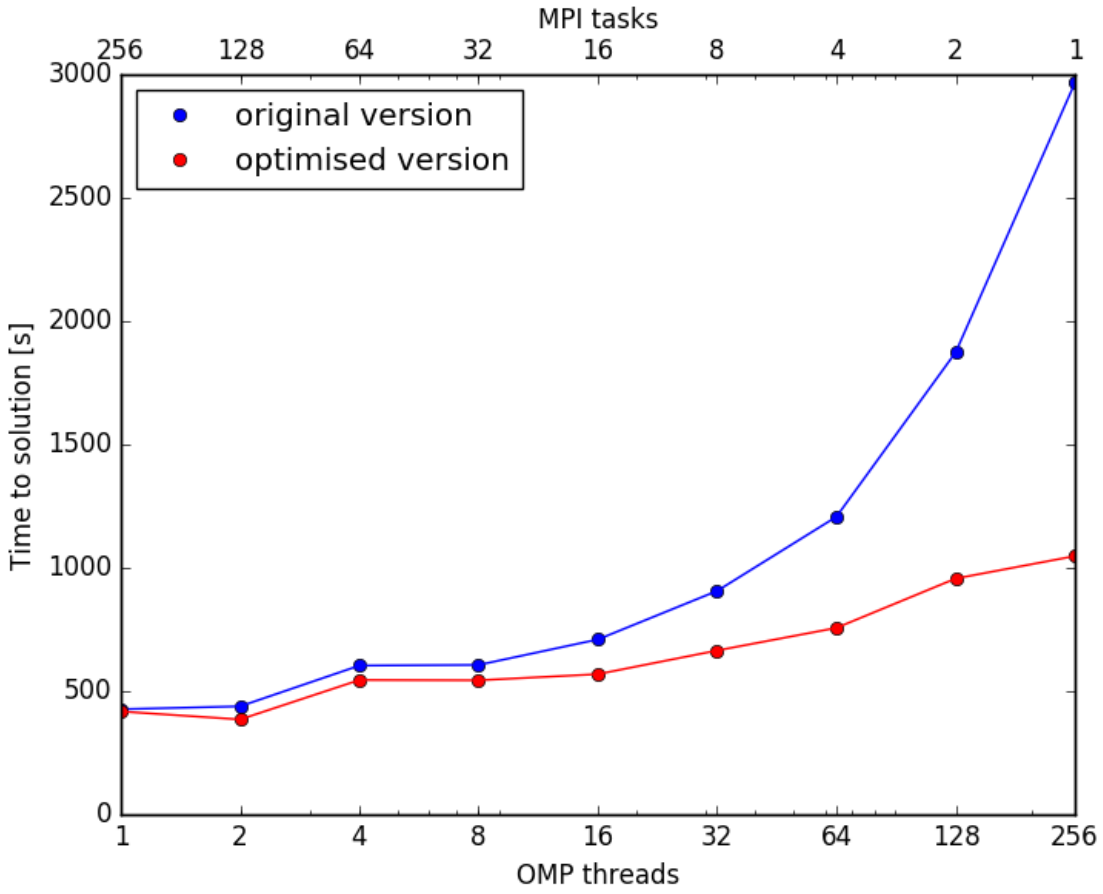
```
! call non-threaded function
...
!$OMP PARALLEL DO
do thread=1,nThreads
   ! call threaded function
end do
!$OMP END PARALLEL DO
...
!$OMP PARALLEL DO
do thread=1,nThreads
   ! call threaded function
end do
!$OMP END PARALLEL DO
```

Optimised code

```
!$OMP parallel
...
!$OMP do schedule(static,1) private(thread)
do thread=1,nThreads
   ! call threaded function
end do
!$OMP end do
...
!$OMP master
! call non-threaded function
!$OMP end master
...
!$OMP end parallel
```

Optimisation of hybrid MPI+OpenMP parallelisation

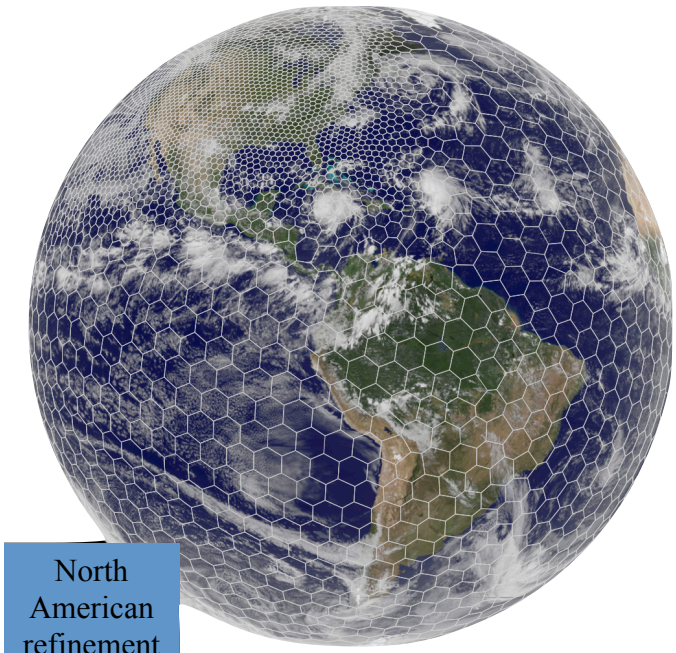
In MPAS-Atmosphere, threading using OpenMP inside MPI is implemented for the time-integration routine only. A simple change to the hybrid implementation and threading of one additional routine yield impressive performance improvements.

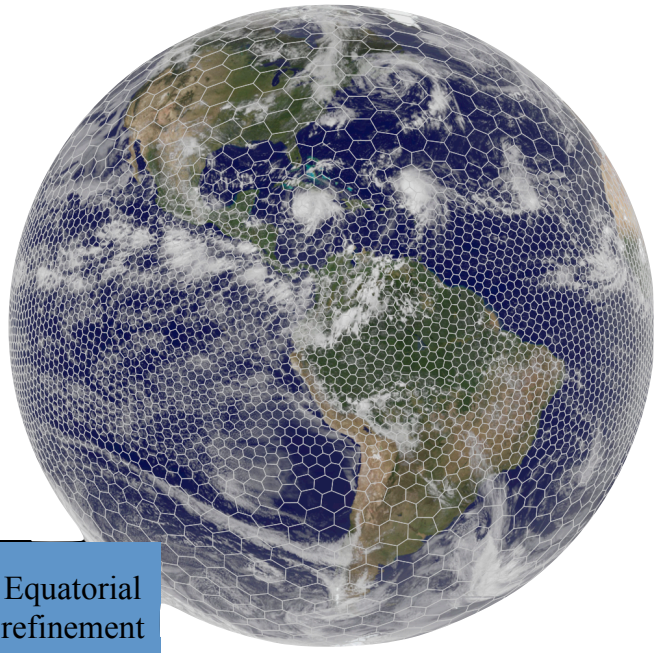


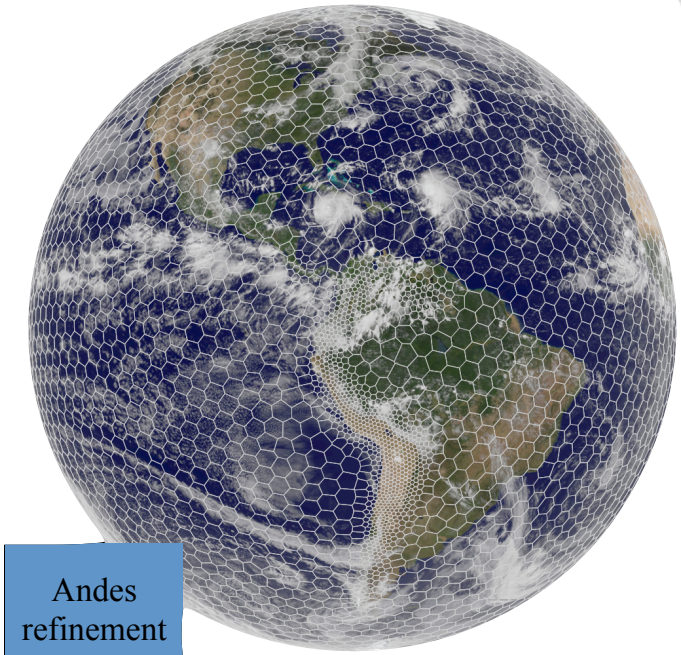
Performance improvements of hybrid code on Intel Xeon Phi Knights Landing for a uniform 240km mesh (10242 cells).

Note 1: With optimised compiler flags for the Intel KNL, these performance improvements are smaller (appr. 20% “only”).

Note 2: regrouping MPI halo exchanges gives 10-20% speedup







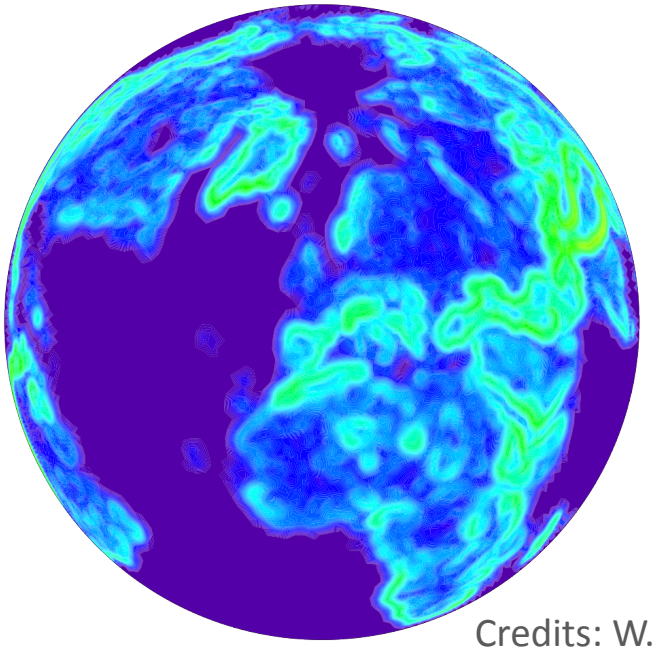
To boldly go where no man has gone before …

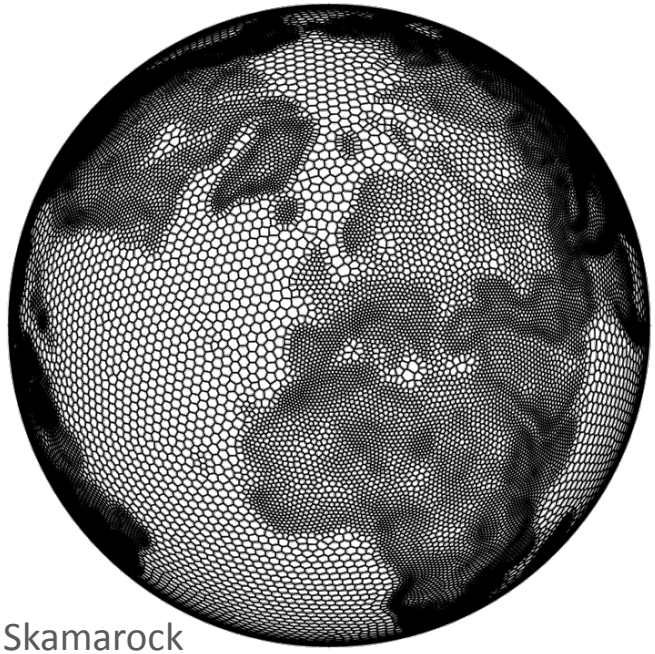
Topographic mesh refinement

Density function proportional to magnitude of topographic gradient

left: 4^{th} root of density function
right: resulting SCVT mesh

step towards scale- and climate-aware meshes





Credits: W. Skamarock