A B C/J geoverbund aachen bonn köln jülich

HPSC TerrSys

JÜLICH FORSCHUNGSZENTRUM

Some JUQUEEN geoscience usage examples from SimLab TerrSys

February 02, 2015 | Klaus Görgen[1,2,4], Fabian Gasper[3,4], Jessica Keune[2,4], Sebastian Knist[1,2,4],
Lukas Poorthuis[1], Jehan Rihani[2], Bernd Schalge[2], Prabhakar Shrestha[2],
Mauro Sulis[2], Stefan Kollet[3,4]

[1] SimLab TerrSys, Jülich Supercomputing Centre, Jülich Research Centre
[2] Meteorological Institute, University of Bonn
[3] Agrosphere (IBG-3), Jülich Research Centre
[4] Centre for High Performance Scientific Computing in Geosciences (Geoverbund ABC/J)

SimLab TerrSys

Terrestrial Systems Modelling Platform

TerrSysMP porting and tuning on JUQUEEN

Convection permitting continental WRF RCM

Conclusion

SimLab TerrSys

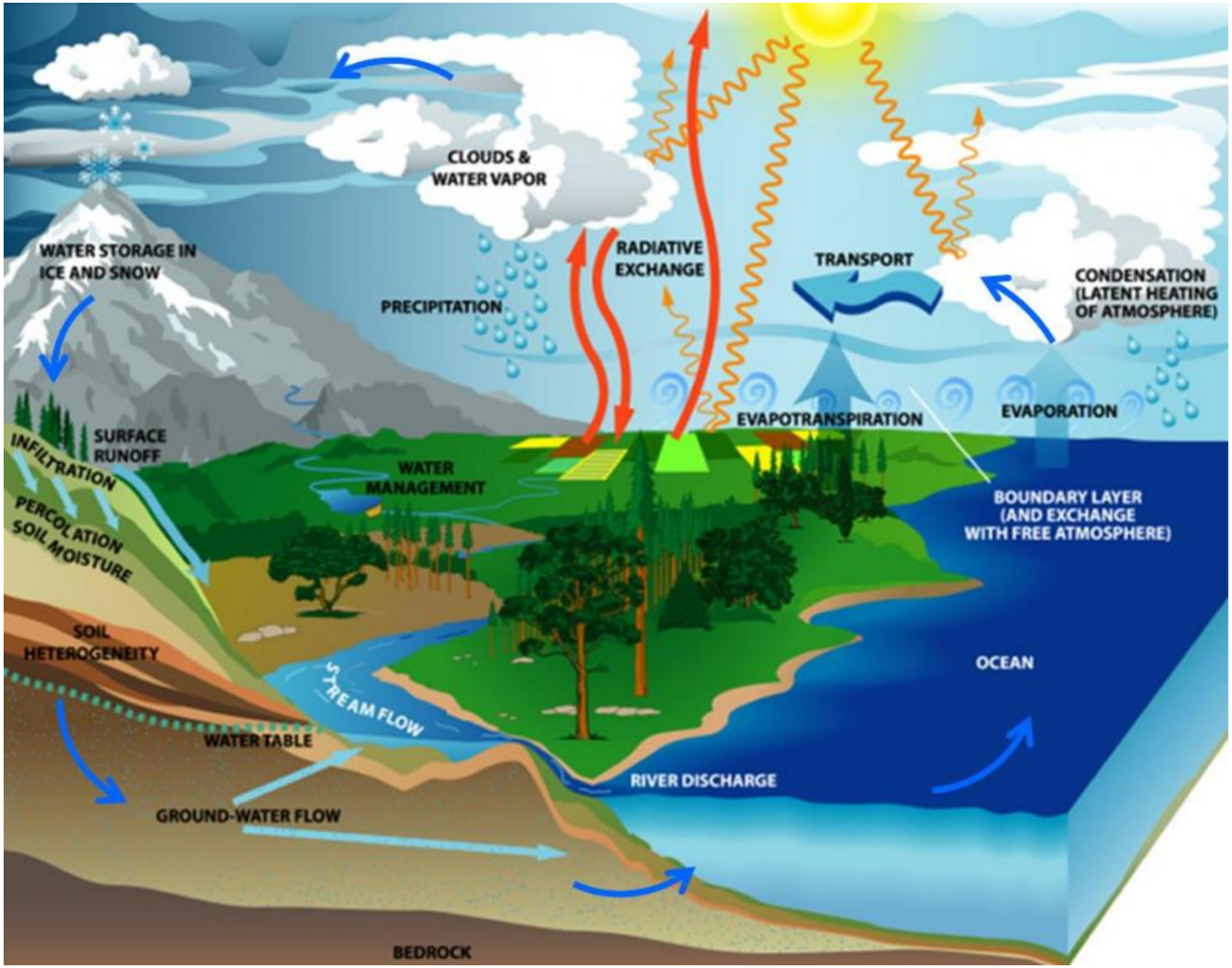
Terrestrial Systems Modelling Platform

TerrSysMP porting and tuning on JUQUEEN

Convection permitting continental WRF RCM

Conclusion

Terrestrial system
Integral part of geo-ecosystems



http://www.esrl.noaa.gov

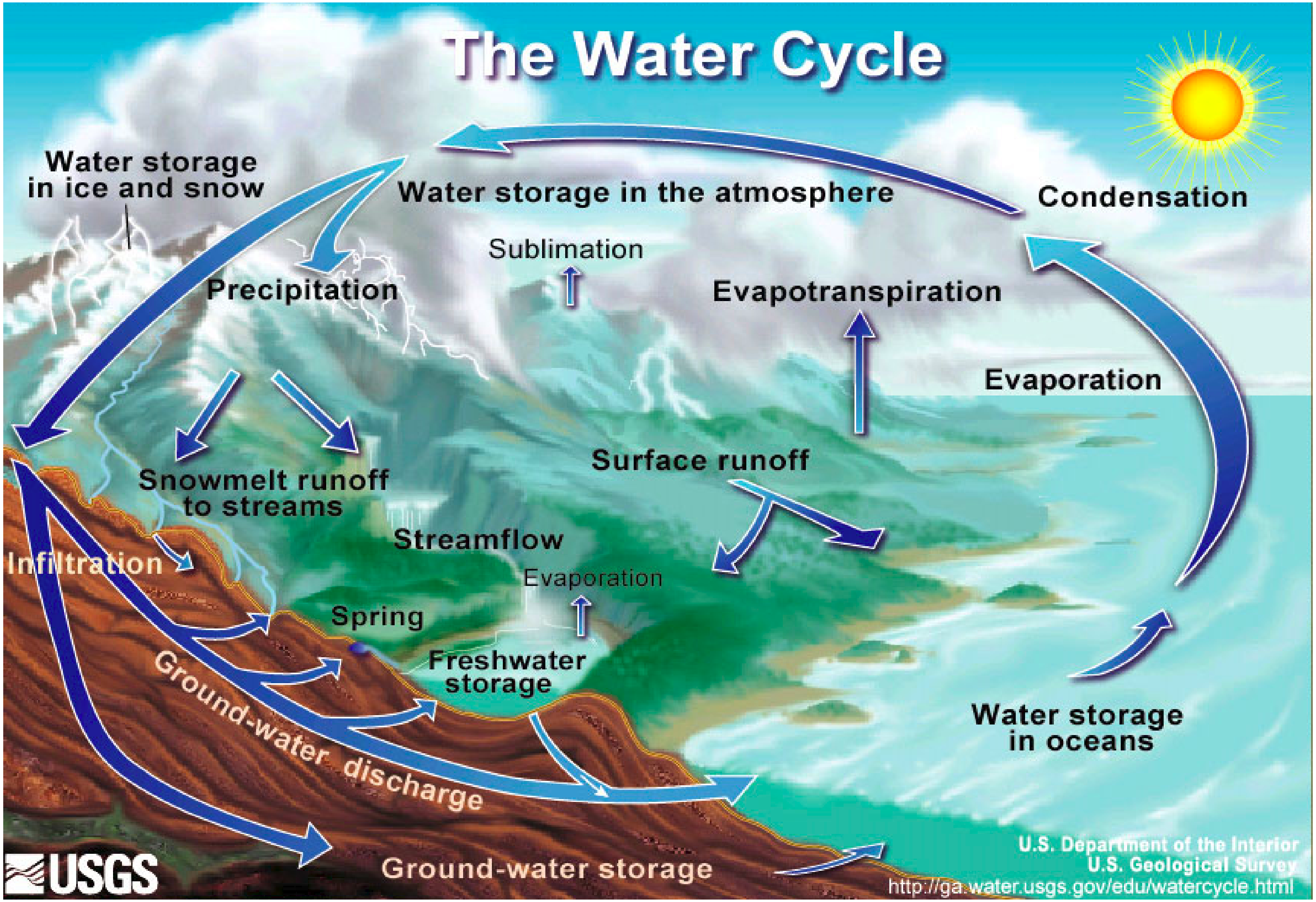
Complex interactions and feedbacks between various sub-systems of the coupled geo-ecosystem (e.g. pedo-, bio-, hydro- or atmosphere)

Linkages through through energy, mass and momentum transfers

Multitude of spatio-temporal scales

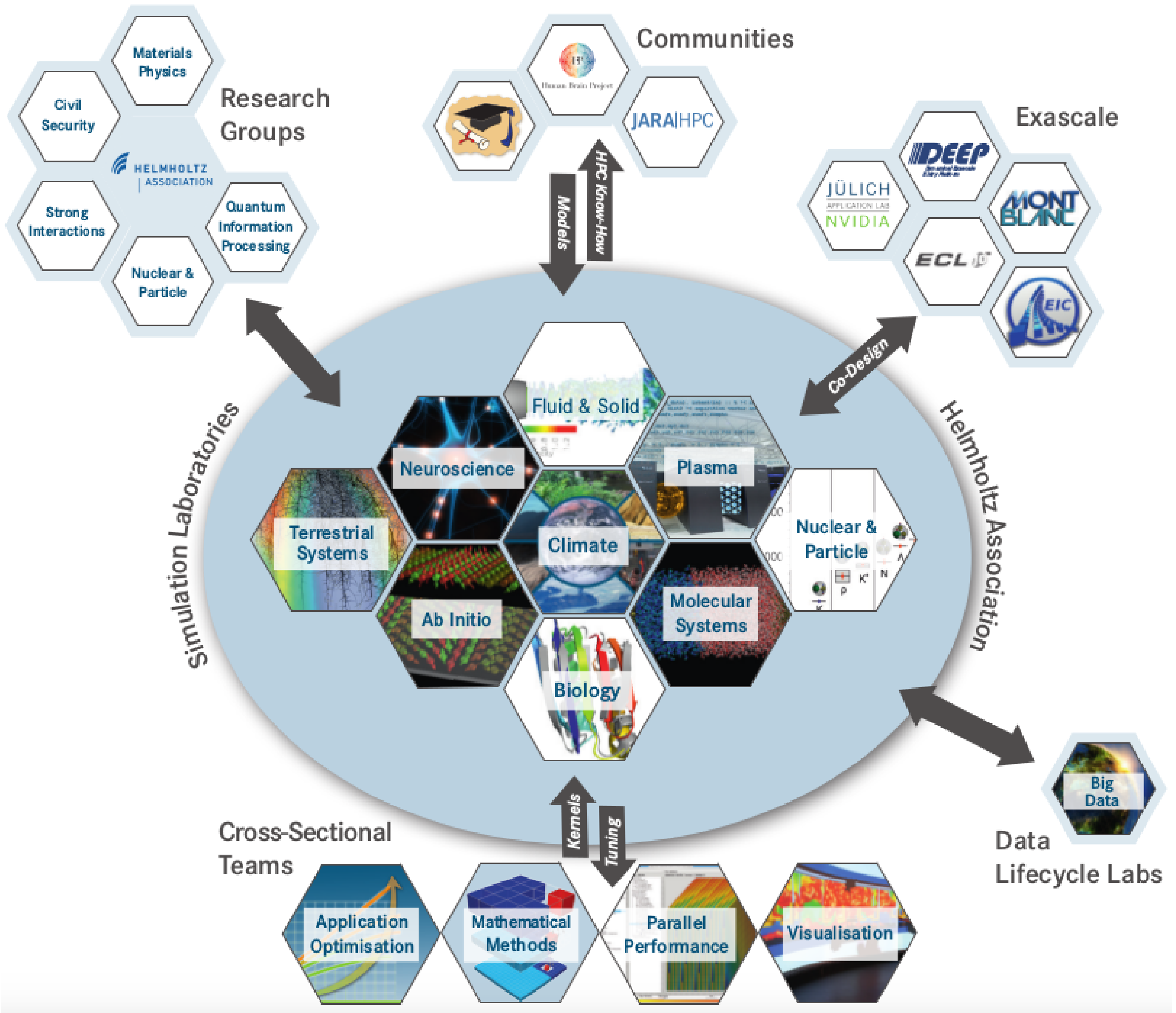
Anthropogenic physical (climate) system changes modify land surface and ecosystem processes with impacts on may sectors (water management, farming, etc.)

Current research focus: Water cycle
Functioning, variability, climate change, ...



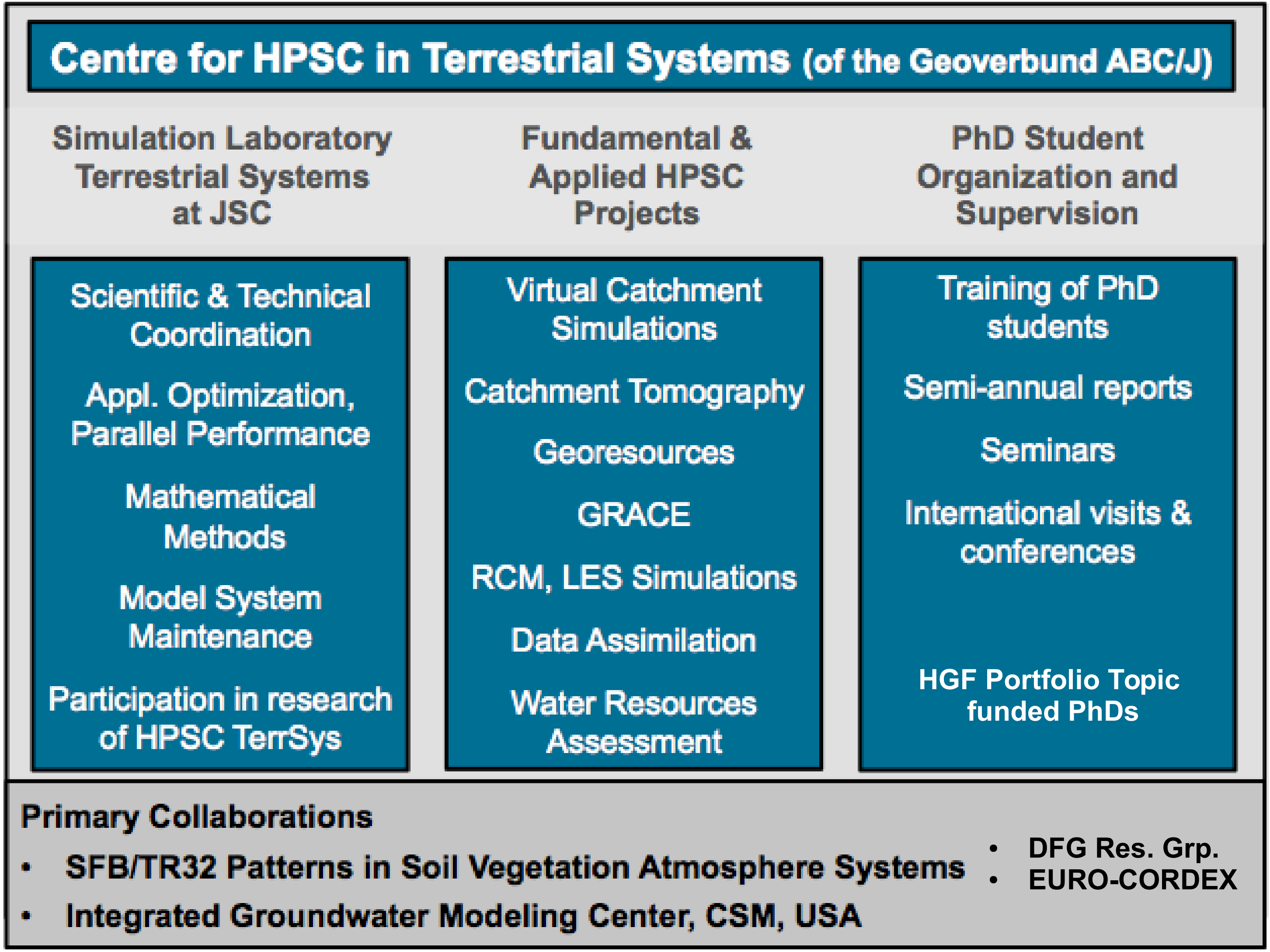
U.S. Department of the Interior
U.S. Geological Survey
http://ga.water.usgs.gov/edu/watercycle.html

SimLab TerrSys is part of community driven support infrastructure



SimLab TerrSys (since Oct 2012)
... of HPSC TerrSys (Geoverbund ABC/J) at JSC

Supercomputing in Geosciences



Centre for HPSC in Terrestrial Systems (of the Geoverbund ABC/J)

Simulation Laboratory Terrestrial Systems at JSC

- Scientific & Technical Coordination
- Appl. Optimization, Parallel Performance
- Mathematical Methods
- Model System Maintenance
- Participation in research of HPSC TerrSys

Fundamental & Applied HPSC Projects

- Virtual Catchment Simulations
- Catchment Tomography
- Georesources
- GRACE
- RCM, LES Simulations
- Data Assimilation
- Water Resources Assessment

PhD Student Organization and Supervision

- Training of PhD students
- Semi-annual reports
- Seminars
- International visits & conferences
- HGF Portfolio Topic funded PhDs

Primary Collaborations

- SFB/TR32 Patterns in Soil Vegetation Atmosphere Systems
- Integrated Groundwater Modeling Center, CSM, USA
- DFG Res. Grp.
- EURO-CORDEX

SimLab TerrSys
Supercomputing in Geosciences

Background and motivation

- Global (environmental) change is affecting terrestrial systems at all scales
- Physical (climate) system changes will be accompanied by major state changes of land surfaces and ecosystems and of the services provided by them with multiple (socioeconomic) impacts
- Demanding models (complex nonlinear processes, interactions, feedbacks across scales, long runtimes, ensembles) and expertise in Geoverbund ABC/J area

Objectives and tasks (balance between support and research)

- Interactions and feedbacks in **coupled geo-ecosystems**
- Modelling of transport processes **across scales**
- Advancement of complex **coupled model systems** (towards exascale), challanges: big data, parallel I/O, booster architectures, hybrid code, high-resolution earth system models at regional scale
- **Porting and scaling of applications**, standard implementations, tools, performance optimizations, processing chains (technical support)
- **HPC support** for the terrestrial systems community
- **Coordination** (e.g. compute time grants), migration of applications to JSC
- Research activities on **HPSC in the geosciences**, ntl. and int'l. collaborations

Interface between geo-science community and JSC resources

Some science questions of TR32, FOR2131, IBG-3, MIUB projects, [...]

- **Higher spatial resolution** leads to a more realistic reproduction of small-scale heterogeneities with multiple impacts via e.g. the surface atmosphere coupling, less influence of parameterisations
- **Integrated simulation platforms** of terrestrial systems are capable of reproducing all relevant feedbacks and interactions on **multiple scales**; they are useful tools in predicting hydrologic responses of catchments/continents
- **Enhanced representation of the hydrological cycle**: water resources from the bedrock into the atmosphere, lateral surface and 3D subsurface flow of water
- **Groundwater hydrodynamics** impact residence times of water molecules in the subsurface leading to power-law type scaling characteristics
- What is the influence of **variability in water and energy fluxes and states at the land surface** on this scaling behavior from the catchment to the continental scale?
- How does **global change** (water resources use, land use, climate,…) **impact** this scaling behavior and thus cycling of matter and energy?
- Performance analysis and scaling behavior in large-scale supercomputing environments
- Modular and least invasive coupling designs
- Efficient application of tools like TerrSysMP in massively parallel computer environments; future HPC architectures (**exascale**)?

SimLab TerrSys

Terrestrial Systems Modelling Platform

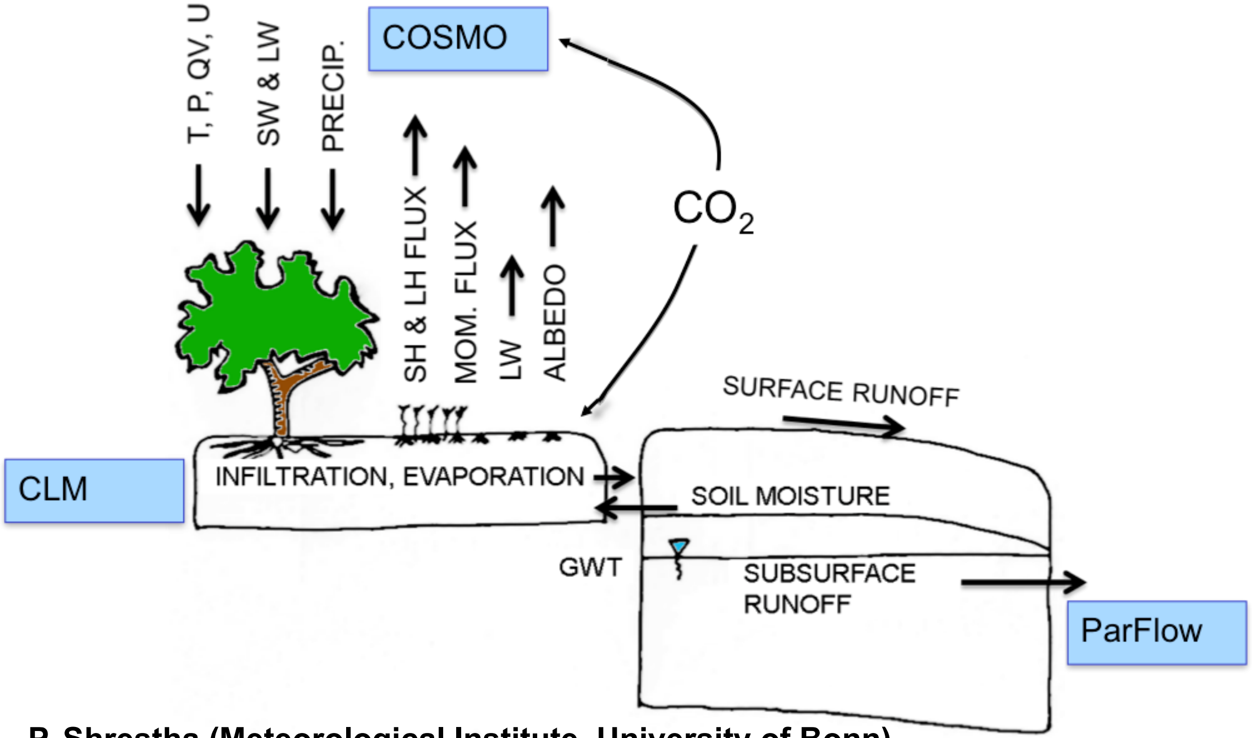
TerrSysMP porting and tuning on JUQUEEN

Convection permitting continental WRF RCM

Conclusion

TerrSysMP objectives
Coupled energy, mass, momentum transfers

- A scale-consistent highly modular fully integrated soil-vegetation-atmosphere modelling system (TerrSysMP) (currently) using COSMO, Community Land Model and ParFlow
- Physically-based representation of transport processes across scales down to sub-km resolutions, explicit feedbacks between compartments (terrestrial hydrological cycle)
- Towards earth system modelling at regional scale



COSMO

T, P, QV, U

SW & LW

PRECIP.

SH & LH FLUX

MOM. FLUX

LW

ALBEDO

CO_2

CLM

INFILTRATION, EVAPORATION

SURFACE RUNOFF

SOIL MOISTURE

GWT

SUBSURFACE RUNOFF

ParFlow

P. Shrestha (Meteorological Institute, University of Bonn)

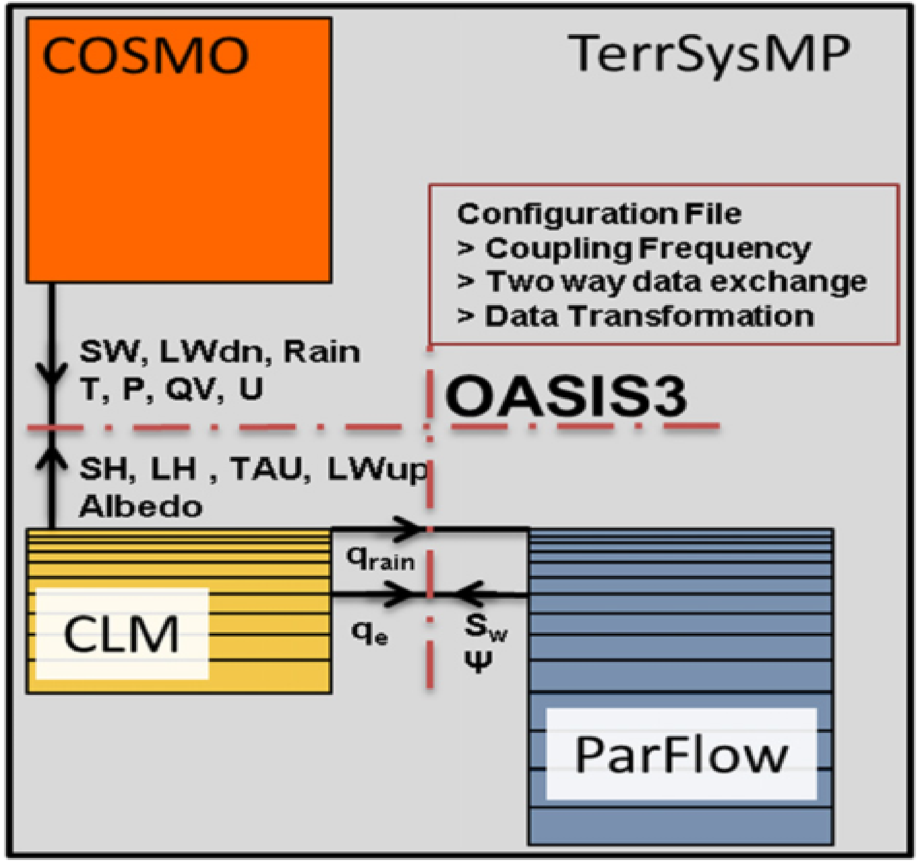
Development:

Transregional Collaborative Research Centre 32 (TR32, “Patterns in Soil-Vegetation-Atmosphere-Systems”)

TerrSysMP
Component models and features

Coupling interface: OASIS3 / OASIS3-MCT

- Uses MPMD execution model
- Suitable for independently developed codes
- Implementation is less code-intrusive
- Component Models can have different spatio-temporal resolution
- Sub-cycling, temporal averaging, grid interpolation possible
- Downscaling option also implemented
- MPI-1 and MPI-2 possible
- OASIS3 creates MPI_COMM_WORLD
- Various configuration options (component models standalone and combinations)



Shrestha et al. (2014, Mon Weather Rev)

SimLab TerrSys

Terrestrial Systems Modelling Platform

TerrSysMP porting and tuning on JUQUEEN

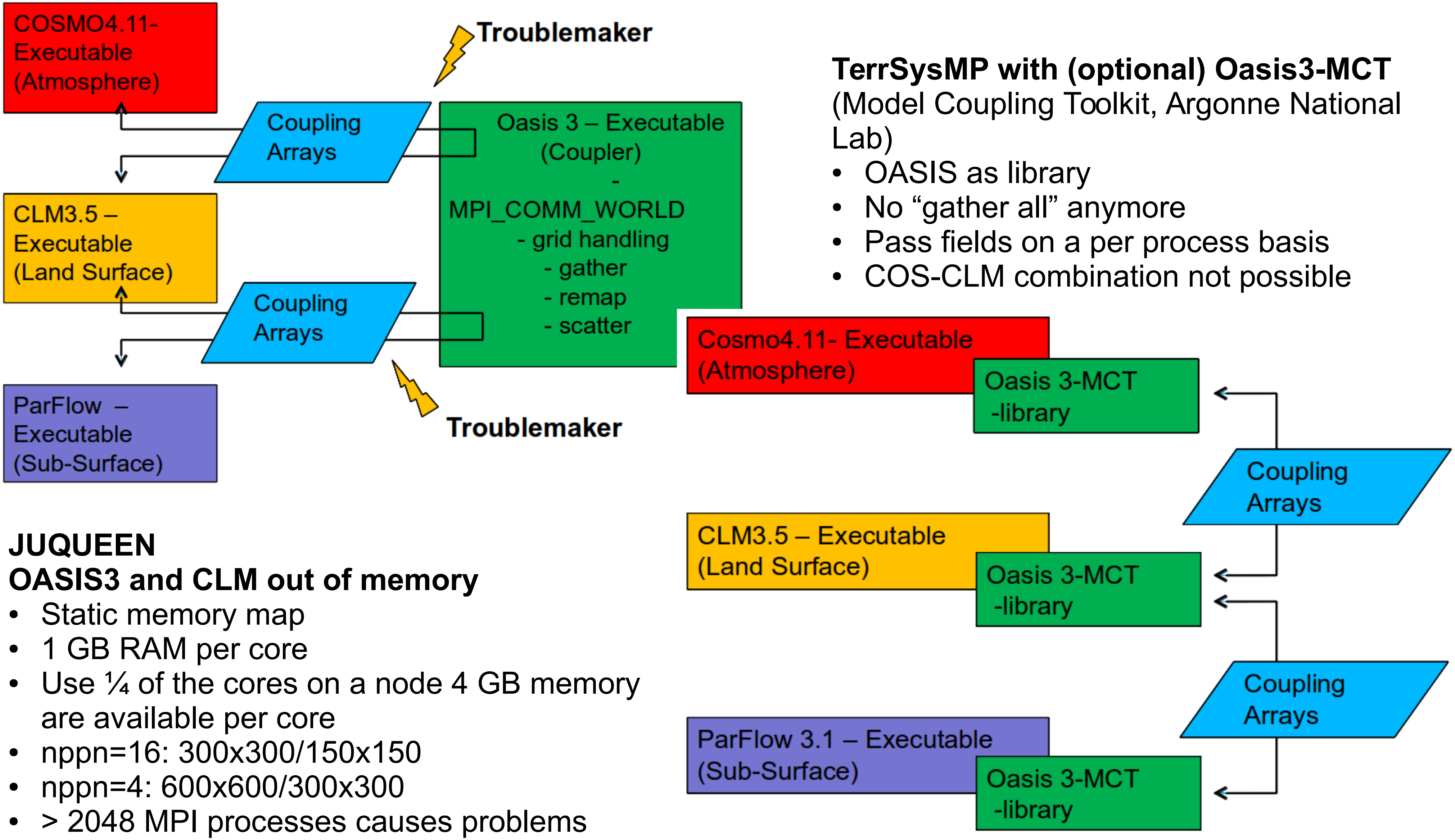
Convection permitting continental WRF RCM

Conclusion

Some TerrSysMP (technical) developments during 2013/14 towards high scalability / large domains

- Works very well on standard Linux clusters, but resource needs required JUQUEEN porting
- **Porting** (MPMD, xl compilers, libs, mapping)
- **Optimisation** (Scalasca profiling of component models and coupled code, optimized model settings (x2), static load balancing (x1.12), compiler optimizations (x1.7), RAM profiling)
- Functionality enhancements (patches JUQ + OASIS + idealized runs, multi-step jobs, restarts, map-file generation, versioning and logging, optional Scalasca instrumentation, file-handling, etc.)
- **Refactoring coupling (optionally OASIS3 or OASIS3-MCT, x1.25)**
- Reference implementations, standard reference test data (real (NRW domain) and idealized datasets)
- Improved pre-processing, forcing data generation (OASIS3 rmp files, etc.)
- Bugfixes (parallel CLM) (before: just ranks per node = 4 and 1 process for CLM)
- Documentation, tutorials, regular user and developer meetings
- Code distribution incl. patches from central MIUB git repository
- Very close collaboration with developers at MIUB (TR32 Z4, Shrestha and Sulis)
- Scaling / HPC related developments are at JSC/IBG-3, physics development etc. at MIUB

Refactoring of coupling scheme
OASIS3 -> OASIS3-MCT, overcome RAM bottleneck


COSMO4.11- Executable (Atmosphere)
Troublemaker
Coupling Arrays
Oasis 3 – Executable (Coupler)
-
MPI_COMM_WORLD
- grid handling
- gather
- remap
- scatter
CLM3.5 – Executable (Land Surface)
Coupling Arrays
ParFlow – Executable (Sub-Surface)
Troublemaker
Cosmo4.11- Executable (Atmosphere)
Oasis 3-MCT -library
Coupling Arrays
CLM3.5 – Executable (Land Surface)
Oasis 3-MCT -library
Coupling Arrays
ParFlow 3.1 – Executable (Sub-Surface)
Oasis 3-MCT -library

TerrSysMP with (optional) Oasis3-MCT
(Model Coupling Toolkit, Argonne National Lab)

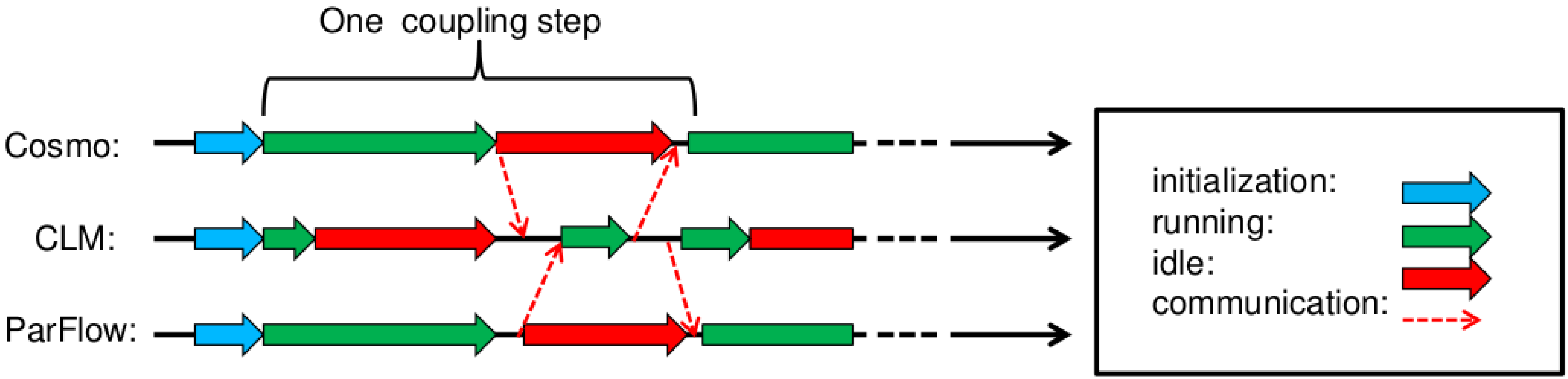
- OASIS as library
- No “gather all” anymore
- Pass fields on a per process basis
- COS-CLM combination not possible

JUQUEEN
OASIS3 and CLM out of memory

- Static memory map
- 1 GB RAM per core
- Use ¼ of the cores on a node 4 GB memory are available per core
- nppn=16: 300x300/150x150
- nppn=4: 600x600/300x300
- > 2048 MPI processes causes problems

“Load balancing” for MPMD using Scalasca How many processes per component model?

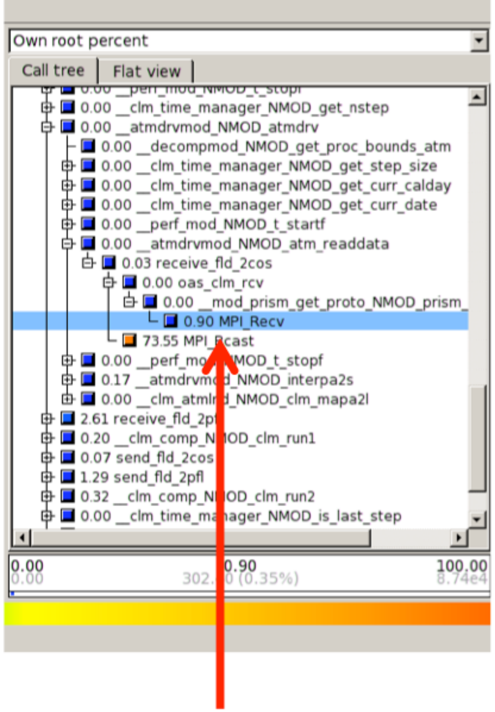
- Important for TerrSysMP: highly varying runtimes among component models
- Potentially highly inefficient, avoid wait states in coupler
- Each new setup (model domains, time steps, etc.) requires new load balancing
- Has to be done before model run (mapping on JUQUEEN hardware)
- **Use of Scalsca with TerrSysMP (MPMD execution model) to determine optimum resource distribution**
- No optimized mapping on machine network topology



F. Gasper (IBG-3, FZJ)

“Load balancing” for MPMD using Scalasca
How many processes per component model?

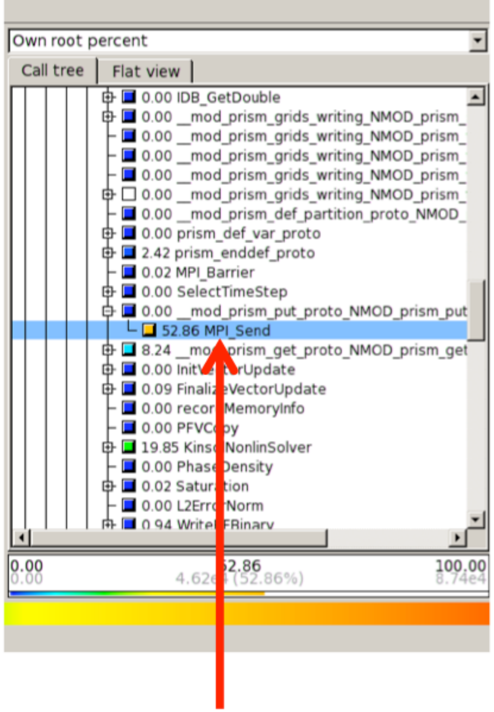
- Let the Model run for a representative number of timesteps
- Interrogating the profile leads to in-depth knowledge of waits/code bottlenecks
- With this method we were able to reduce the runtime by 16% compared to a balancing based on “hand written” timings



```
0.00 __perf_mod_NMOD_t_stopf
0.00 __clm_time_manager_NMOD_get_nstep
0.00 __atmdrvmod_NMOD_atmdrv
  0.00 __decompmod_NMOD_get_proc_bounds_atm
  0.00 __clm_time_manager_NMOD_get_step_size
  0.00 __clm_time_manager_NMOD_get_curr_calday
  0.00 __clm_time_manager_NMOD_get_curr_date
  0.00 __perf_mod_NMOD_t_startf
  0.00 __atmdrvmod_NMOD_atm_readdata
    0.03 receive_fld_2cos
      0.00 oas_clm_rcv
        0.00 __mod_prism_get_proto_NMOD_prism_
          0.90 MPI_Recv
      73.55 MPI_Bcast
  0.00 __perf_mo MOD_t_stopf
  0.17 __atmdrvmo d_NMOD_interpa2s
  0.00 __clm_atmlr d_NMOD_clm_mapa2l
2.61 receive_fld_2pf
0.20 __clm_comp_N OD_clm_run1
0.07 send_fld_2cos
1.29 send_fld_2pfl
0.32 __clm_comp_N OD_clm_run2
0.00 __clm_time_ma ager_NMOD_is_last_step

0.00                0.90              100.00
0.00         302.  0 (0.35%)          8.74e4
```

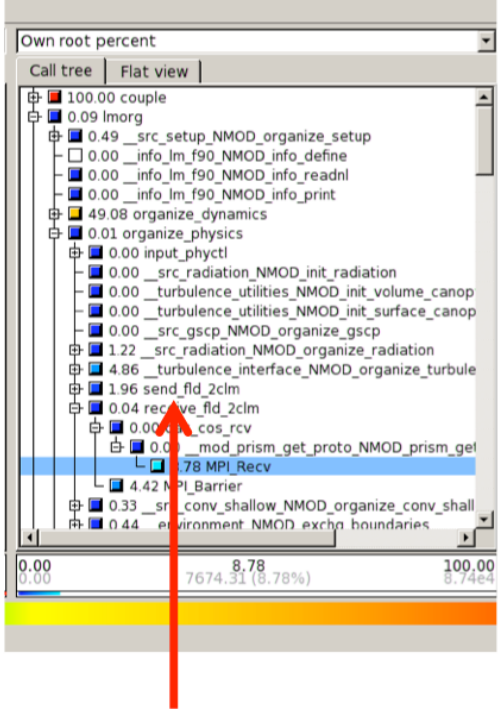
CLM waits 73% of the time for Cosmo



```
0.00 IDB_GetDouble
0.00 __mod_prism_grids_writing_NMOD_prism_
0.00 __mod_prism_grids_writing_NMOD_prism_
0.00 __mod_prism_grids_writing_NMOD_prism_
0.00 __mod_prism_grids_writing_NMOD_prism_
0.00 __mod_prism_grids_writing_NMOD_prism_
0.00 __mod_prism_def_partition_proto_NMOD_
0.00 prism_def_var_proto
2.42 prism_enddef_proto
0.02 MPI_Barrier
0.00 SelectTimeStep
0.00 __mod_prism_put_proto_NMOD_prism_put
  52.86 MPI_Send
8.24 __mo  prism_get_proto_NMOD_prism_get
0.00 InitV  rUpdate
0.09 Finaliz eVectorUpdate
0.00 recor  MemoryInfo
0.00 PFVC py
19.85 Kins  NonlinSolver
0.00 Phase Density
0.02 Satur tion
0.00 L2Erro Norm
0.94 Write  FBinary

0.00                52.86             100.00
0.00         4.62e 4 (52.86%)         8.74e4
```

Parflow waits 53% of the time for CLM (and CLM waits for Cosmo)

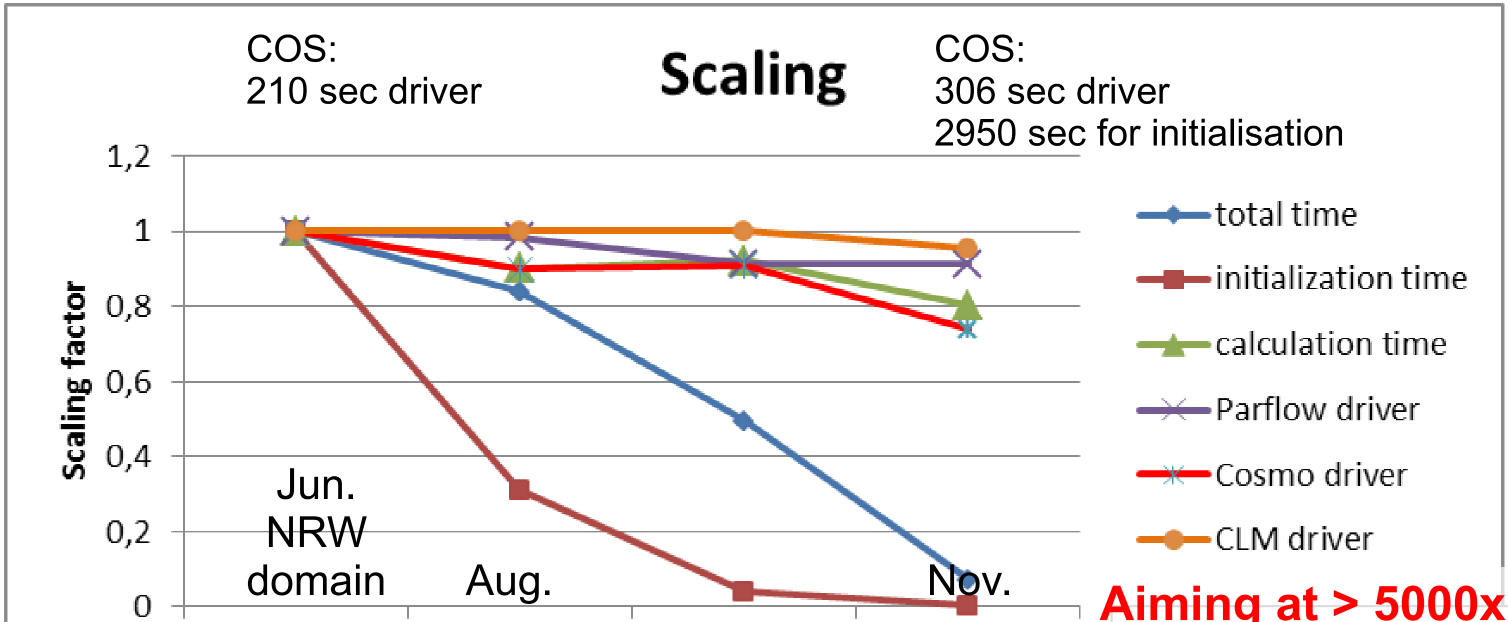


```
100.00 couple
0.09 lmorg
  0.49 __src_setup_NMOD_organize_setup
  0.00 __info_lm_f90_NMOD_info_define
  0.00 __info_lm_f90_NMOD_info_readnl
  0.00 __info_lm_f90_NMOD_info_print
  49.08 organize_dynamics
  0.01 organize_physics
    0.00 input_phyctl
    0.00 __src_radiation_NMOD_init_radiation
    0.00 __turbulence_utilities_NMOD_init_volume_canop
    0.00 __turbulence_utilities_NMOD_init_surface_canop
    0.00 __src_gscp_NMOD_organize_gscp
    1.22 __src_radiation_NMOD_organize_radiation
    4.86 __turbulence_interface_NMOD_organize_turbule
    1.96 send_fld_2clm
    0.04 rec ive_fld_2clm
      0.00   cos_rcv
        0.0  __mod_prism_get_proto_NMOD_prism_get
          .78 MPI_Recv
      4.42 M I_Barrier
    0.33 __sr _conv_shallow_NMOD_organize_conv_shall
    0.44 __e vironment_NMOD_exchg_boundaries

0.00                8.78              100.00
0.00         7674.31 (8.78%)          8.74e4
```

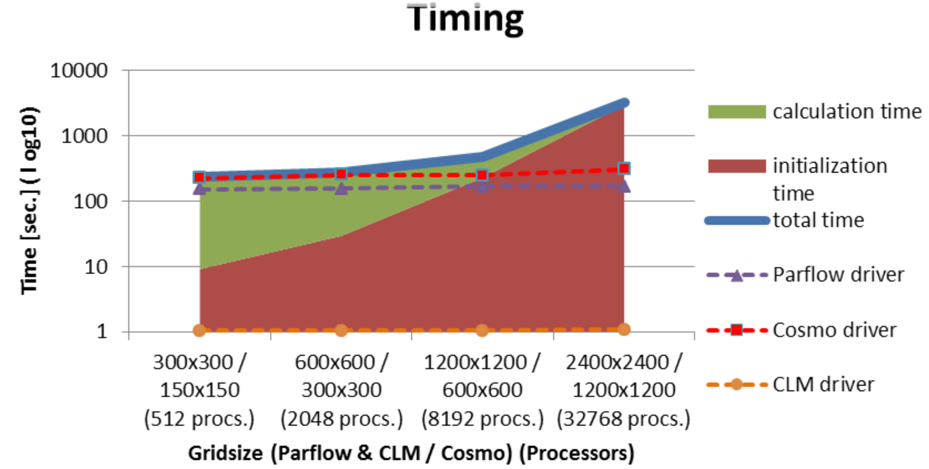
Cosmo does not have to wait (slowest)

Weak scaling results, TerrSysMP fully coupled 512 cores end of Feb, 32k cores beginning of Dec



- Idealized test case
- 3 h simulation time
- 1 km COS
  0.5 km CLM, PFL
- Weak scaling
- Optimally balanced
- Increase of domain size by factor of 4 (64x overall wrt our unit size)
- No I/O for largest runs

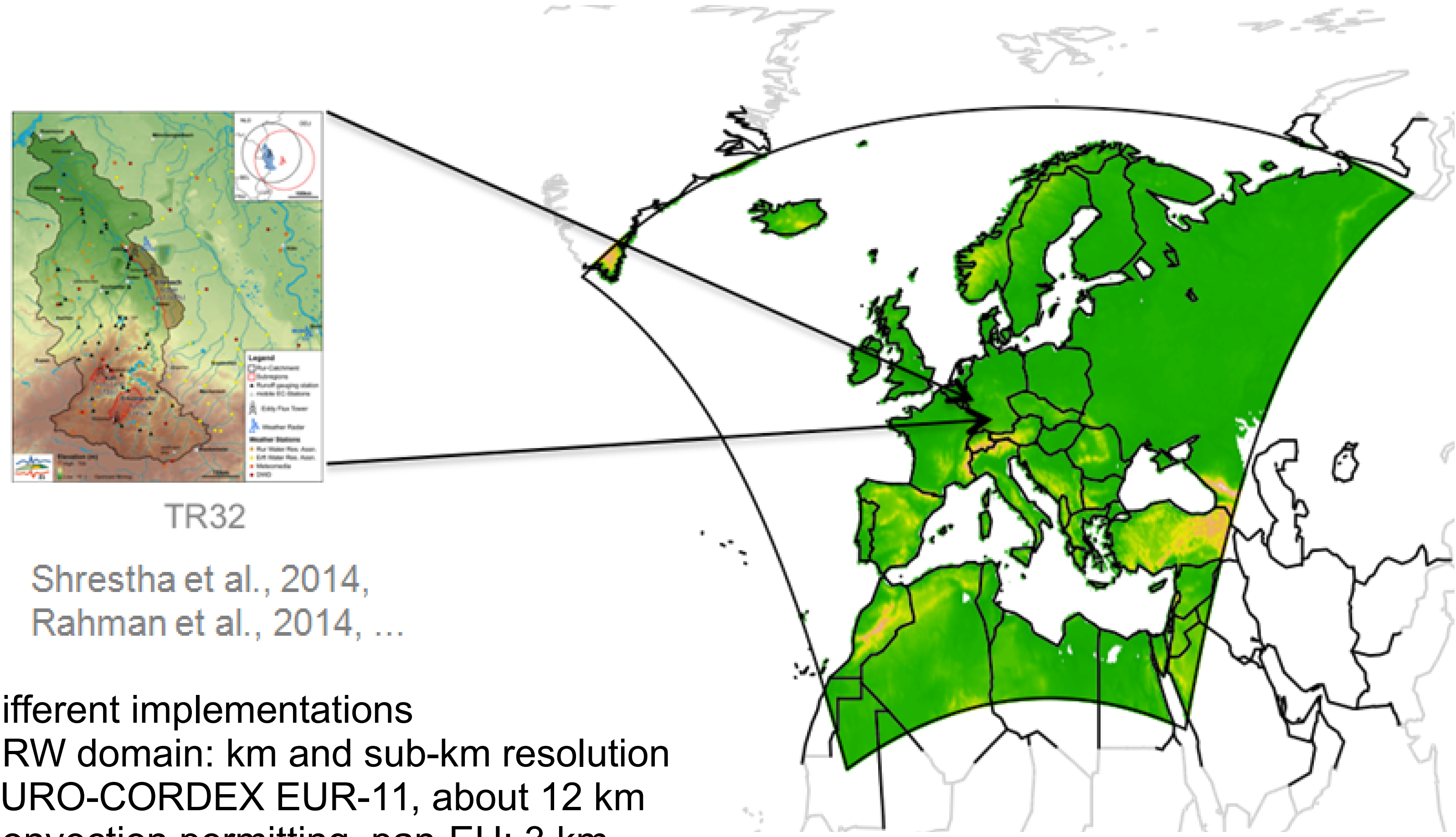
Aiming at > 5000x5000



- OASIS3-MCT coupler
- COS 22x16, CLM 80, PFL 8x8 (basic resource assignment for 300x300/150x150 from Scalasca)
- SMT1, ranks per node 16
- 900 sec coupling frequency

Encouraging weak scaling behaviour, 1 month simulation time = 1 day wallclock

Model domains
From catchment to continental scales



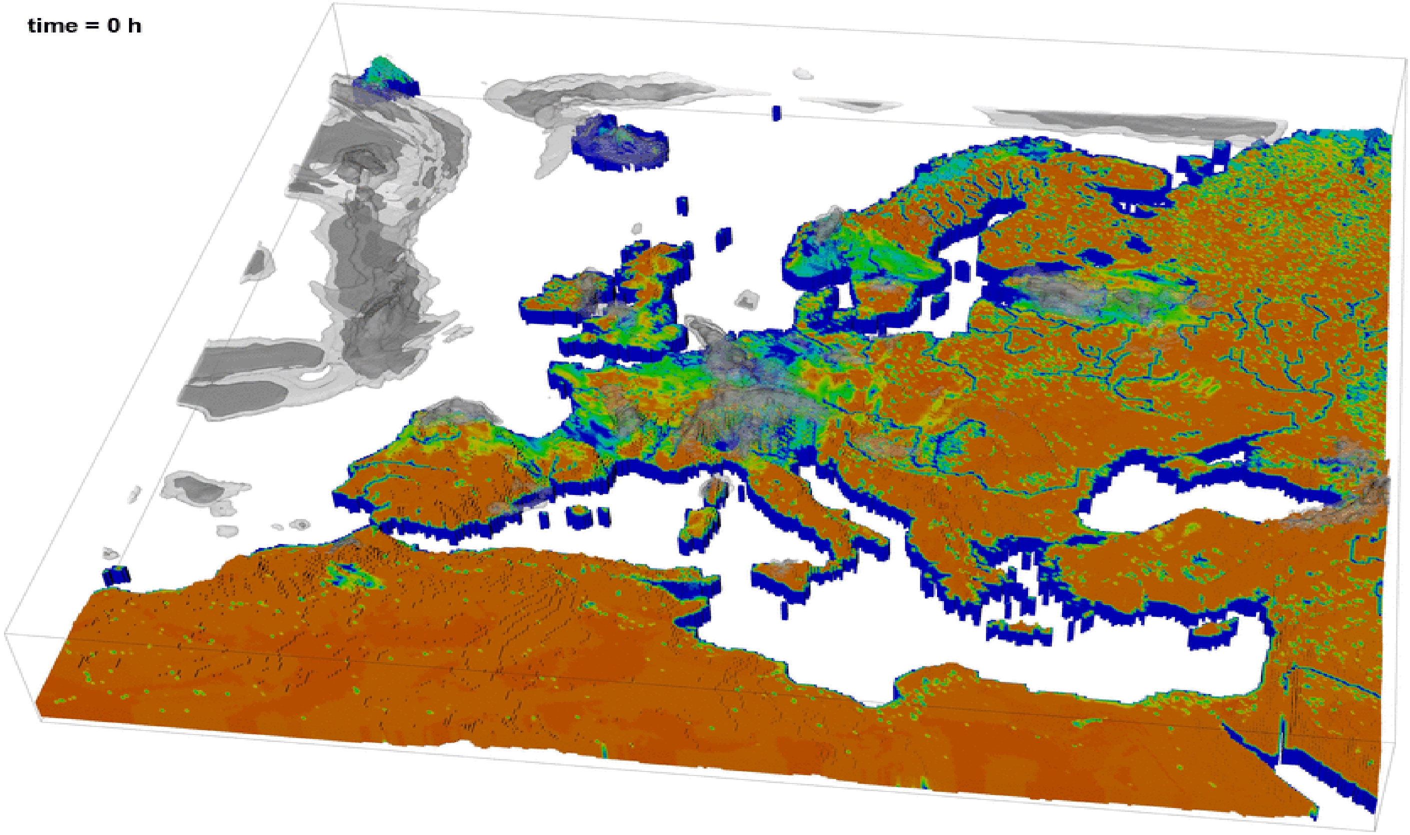
TR32

Shrestha et al., 2014,
Rahman et al., 2014, …

Different implementations
NRW domain: km and sub-km resolution
EURO-CORDEX EUR-11, about 12 km
Convection permitting, pan-EU: 3 km

J. Keune (Meteorological Institute, University of Bonn)

Validation runs
Fully coupled simulations at 12 km resolution

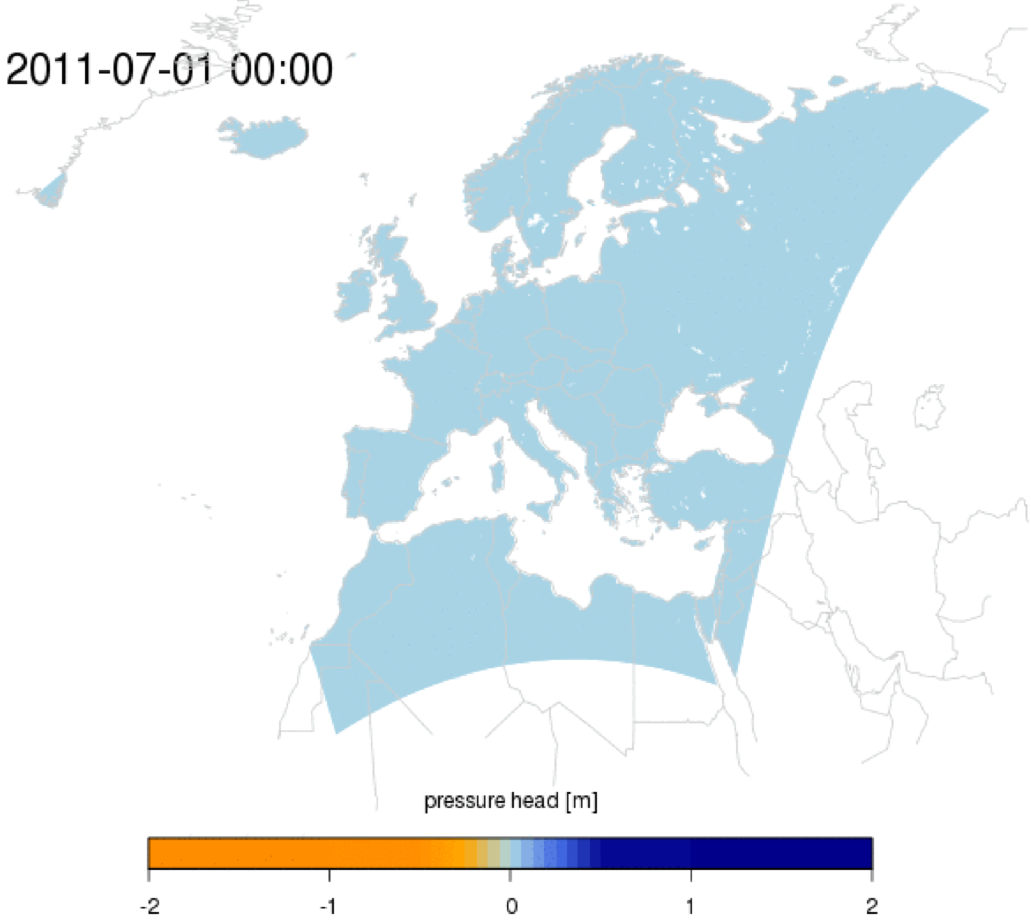


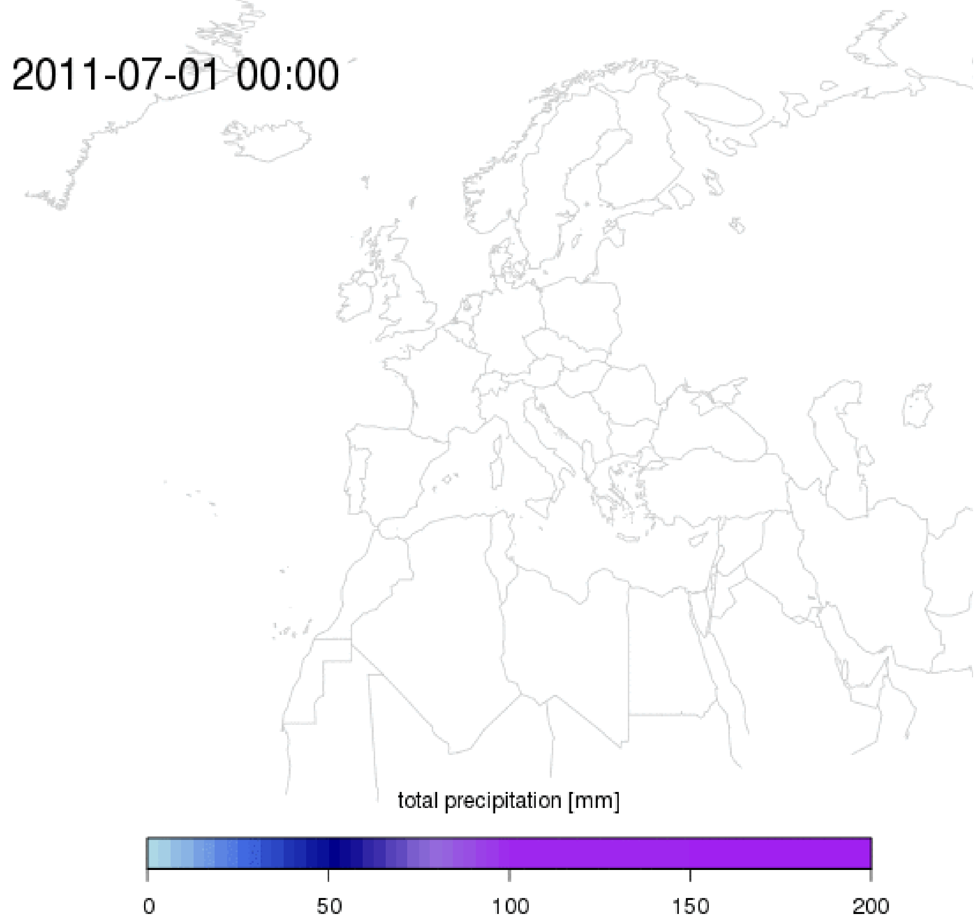
3 day snapshot, Elbe flood event 2013; end of May, beginning of June 2013
Cloud liquid water and ice (COSMO), soil mositure (ParFlow)
JUQUEEN

S. Kollet (IBG-3, FZJ)

Use of TerrSysMP to reproduce river flow
Surface runoff directly linked to meteorology

Test run, fully coupled TerrSysMP, EUR-11 domain (12 km), JUQUEEN
Total precipitation (COSMO), pressure head (ParFlow)





J. Keune (Meteorological Institute, University of Bonn)

SimLab TerrSys

Terrestrial Systems Modelling Platform

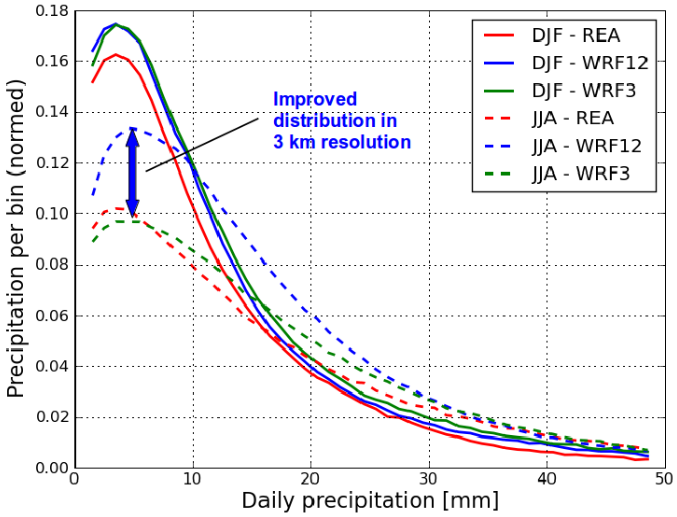
TerrSysMP porting and tuning on JUQUEEN

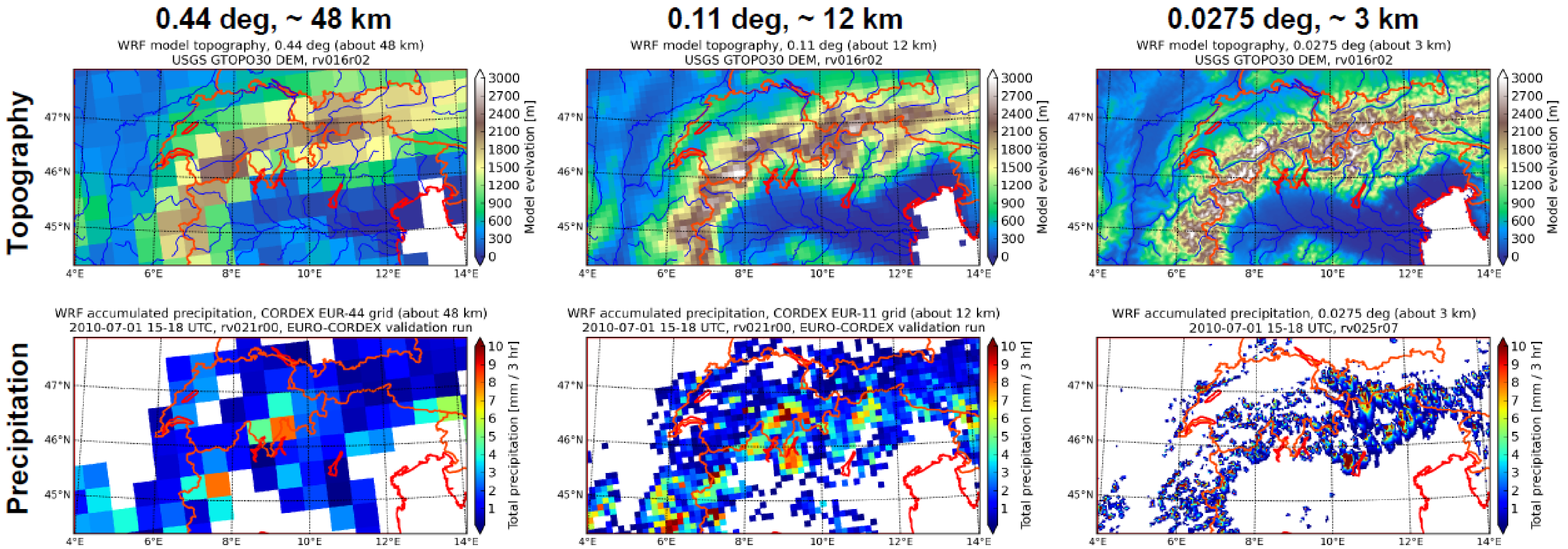
Convection permitting continental WRF RCM

Conclusion

Added value through very high resolution

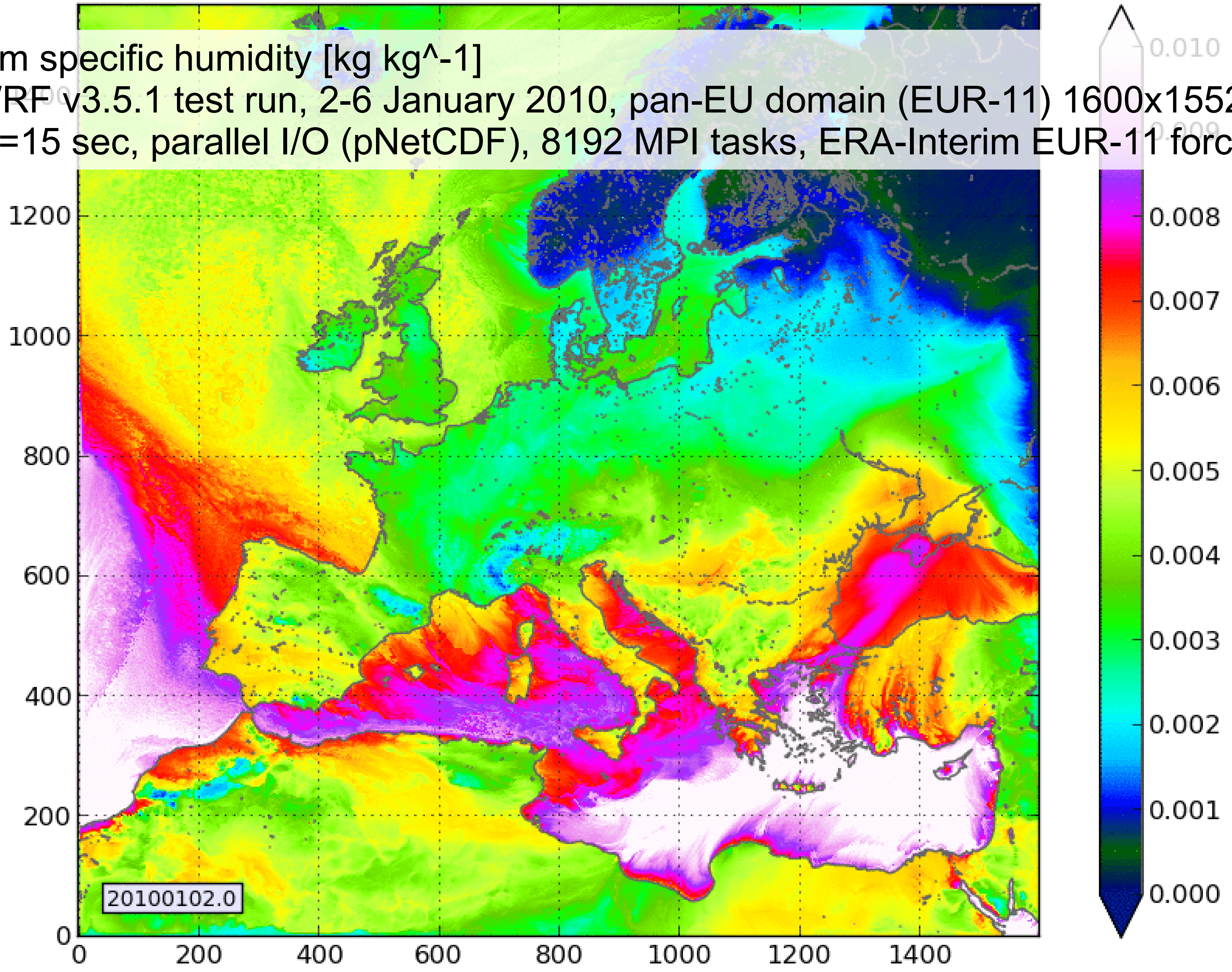
- Representation of surface and subsurface heterogeneities
- Identification of patterns accross scales
- Large potential for multiscale process studies
- Atmospheric dynamics better represented
- Improved precipitation intensities, diurnal cycle, distribution
- Part of contribution to WCRP EURO-CORDEX (JUROPA-2)





0.44 deg, ~ 48 km
0.11 deg, ~ 12 km
0.0275 deg, ~ 3 km
Topography
Precipitation
WRF model topography, 0.44 deg (about 48 km)
USGS GTOPO30 DEM, rv016r02
WRF model topography, 0.11 deg (about 12 km)
USGS GTOPO30 DEM, rv016r02
WRF model topography, 0.0275 deg (about 3 km)
USGS GTOPO30 DEM, rv016r02
WRF accumulated precipitation, CORDEX EUR-44 grid (about 48 km)
2010-07-01 15-18 UTC, rv021r00, EURO-CORDEX validation run
WRF accumulated precipitation, CORDEX EUR-11 grid (about 12 km)
2010-07-01 15-18 UTC, rv021r00, EURO-CORDEX validation run
WRF accumulated precipitation, 0.0275 deg (about 3 km)
2010-07-01 15-18 UTC, rv025r07
47°N
46°N
45°N
4°E
6°E
8°E
10°E
12°E
14°E
3000
2700
2400
2100
1800
1500
1200
900
600
300
0
Model evelvation [m]
10
9
8
7
6
5
4
3
2
1
Total precipitation [mm / 3 hr]

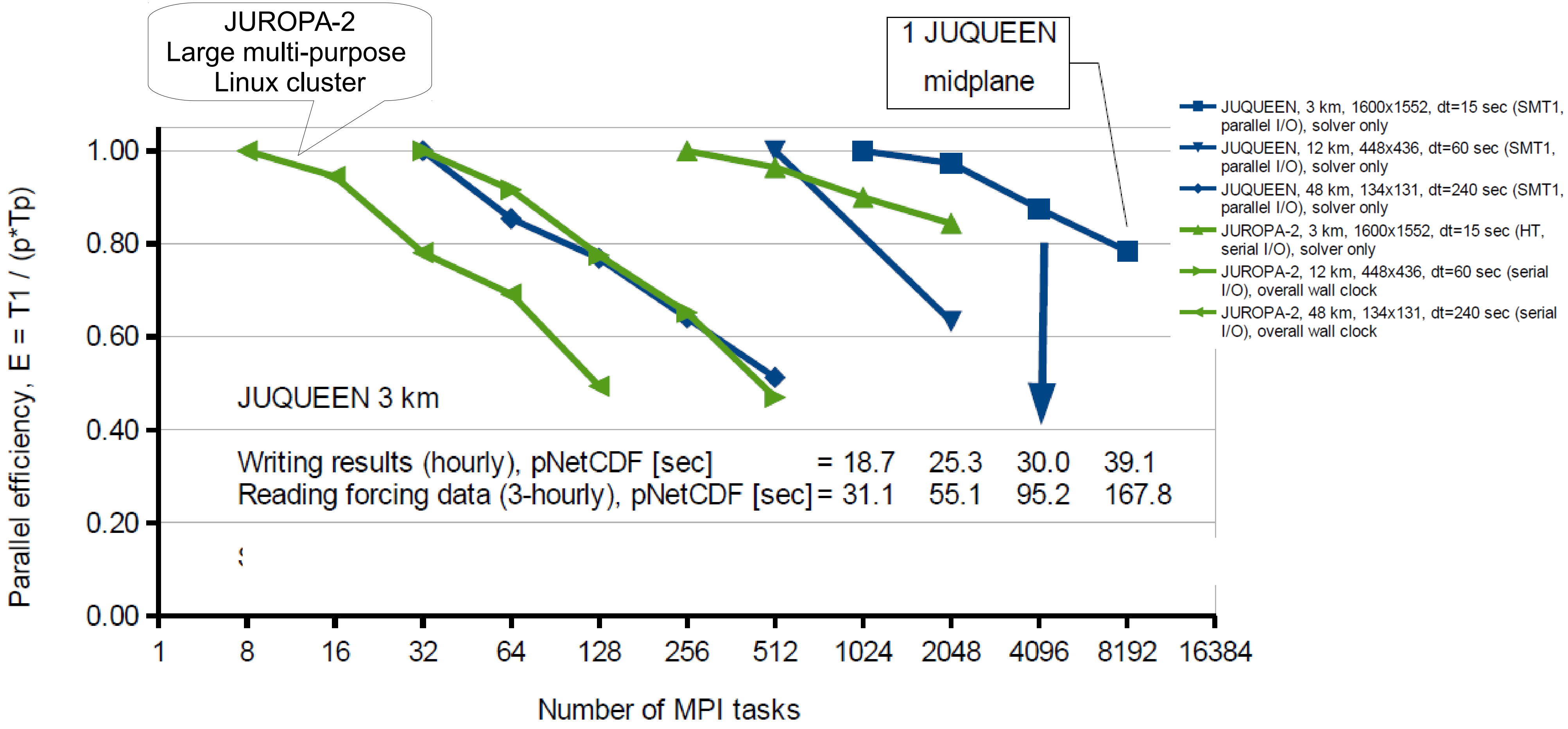
WRF 3 km implementation on JUQUEEN
Continental convection permitting runs still rare



2 m specific humidity [kg kg^-1]
WRF v3.5.1 test run, 2-6 January 2010, pan-EU domain (EUR-11) 1600x1552x50, dt=15 sec, parallel I/O (pNetCDF), 8192 MPI tasks, ERA-Interim EUR-11 forcing

min=0.00001, max=0.01439, mean=0.00462, median=0.00455, std. dev.=0.00257

WRF RCM strong scaling
pan-EU domains (EUR-44, EUR-11, 3 km), real data cases


JUROPA-2
Large multi-purpose
Linux cluster

1 JUQUEEN
midplane

JUQUEEN 3 km

Writing results (hourly), pNetCDF [sec] = 18.7 25.3 30.0 39.1
Reading forcing data (3-hourly), pNetCDF [sec] = 31.1 55.1 95.2 167.8

Using systems like JUQUEEN may be challenging.
It seems not ideal for small problem sizes but well suited for continental domains.
Combination of JUROPA/JURECA and JUQUEEN.

SimLab TerrSys

Terrestrial Systems Modelling Platform

TerrSysMP porting and tuning on JUQUEEN

Convection permitting continental WRF RCM

Conclusion

Conclusions
TerrSysMP, WRF on JUQUEEN

- To use JUQUEEN efficiently for typical geoscience problems (climate modelling) needs a number of **adjustments** and the **suitable problem size** (ensemble members, resolution, model domain)
- Possibility of “farming”, e.g. via **MPMD**
- Large “**legacy**” geoscience codes may hardly use all performance features of JUQUEEN
- **Little RAM per node** and static memory map, static builds
- MPMD: ideal for **coupled model systems**; optimisation of mapping on interconnect 5D Torus network topology not done: weight for specific shape, outweighed by shorter queuing time
- Slow CPUs, use of 4-way simultaneous multithreading, have minimum SMT2
- **Need many ranks per node** for good parallel efficiency
- Program structure must avoid many “MPI_ALLGATHER” operations
- Hybrid code (MPI+OpenMP)
- Parallel I/O (e.g. pNetCDF), big data frameworks needed (I/O, *in-situ*, compressions, etc.)
- Profiling crucial of RAM, I/O (Darshan), communication (Scalasca)

SimLab TerrSys

of the JSC and Centre for HPSC in Terrestrial Systems (Geoverbund ABC/J)

http://www.fz-juelich.de/ias/jsc/slts

http://www.hpsc-terrsys.de

http://www.geoverbund-abcj.de