natESM

DKRZ DEUTSCHES KLIMARECHENZENTRUM

JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

DKK Deutsches Klima Konsortium

Federal Ministry of Education and Research

Sprint 7

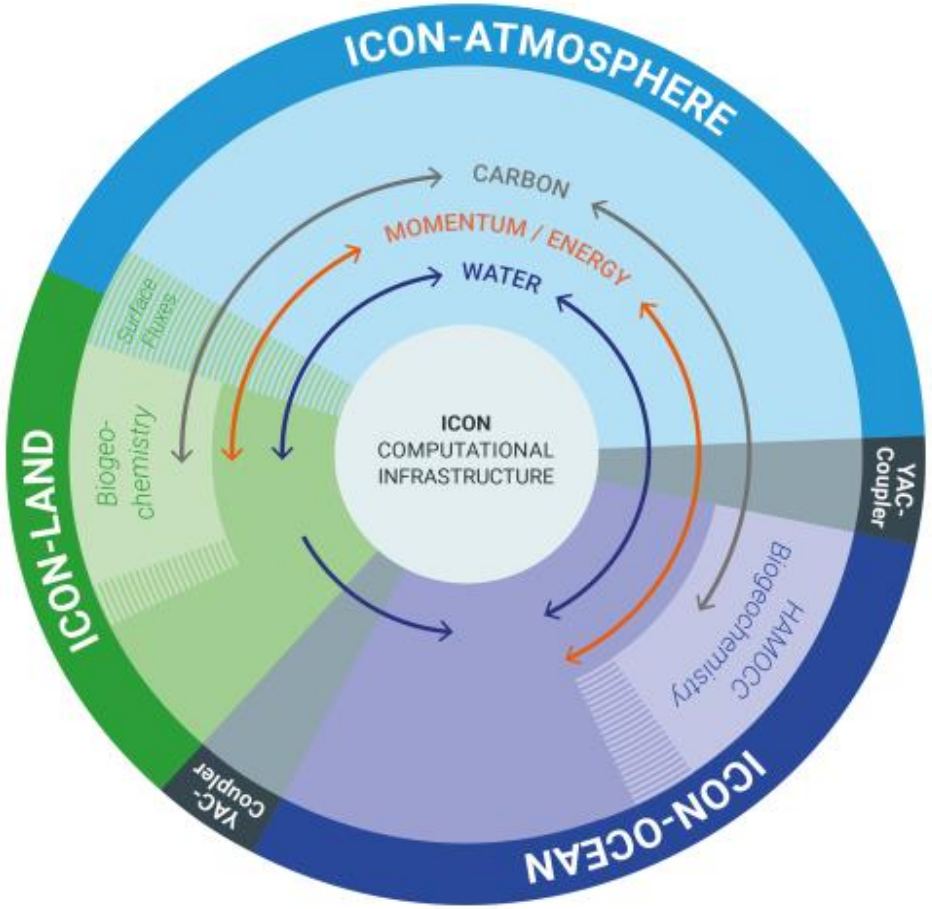
Challenges and results experienced during the HAMOCC sprint

Fatemeh Chegini (MPI-M), Enrico Degregori (DKRZ), Tatiana Ilyina (Uni Hamburg)

HAMOCC: Ocean Biogeochemistry component in ICON

HAMOCC:

- Extended NPZD model
- 20+ tracers in the water column
- Transport of tracers are the most computationally expensive part

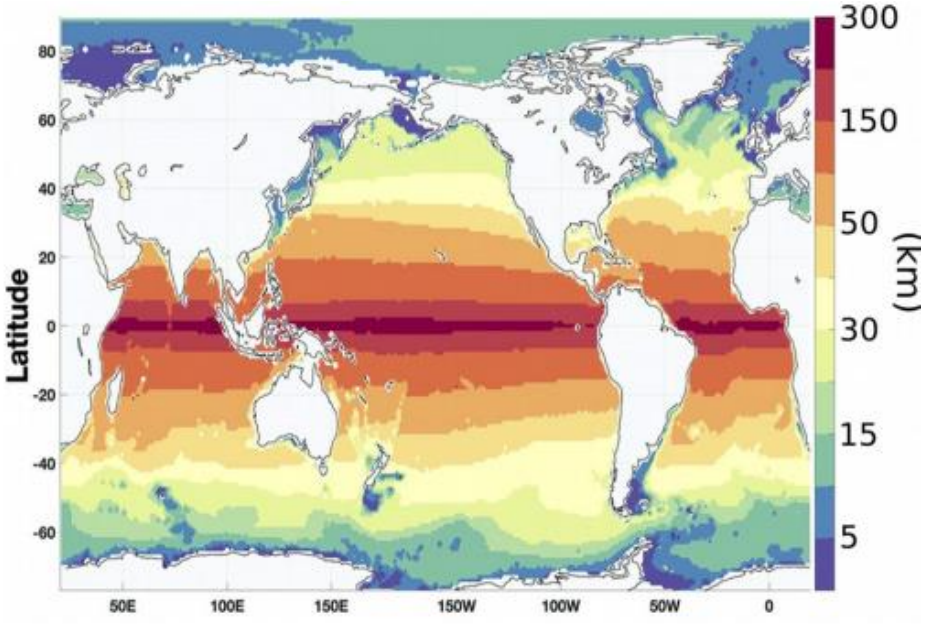


ICON Earth System Model

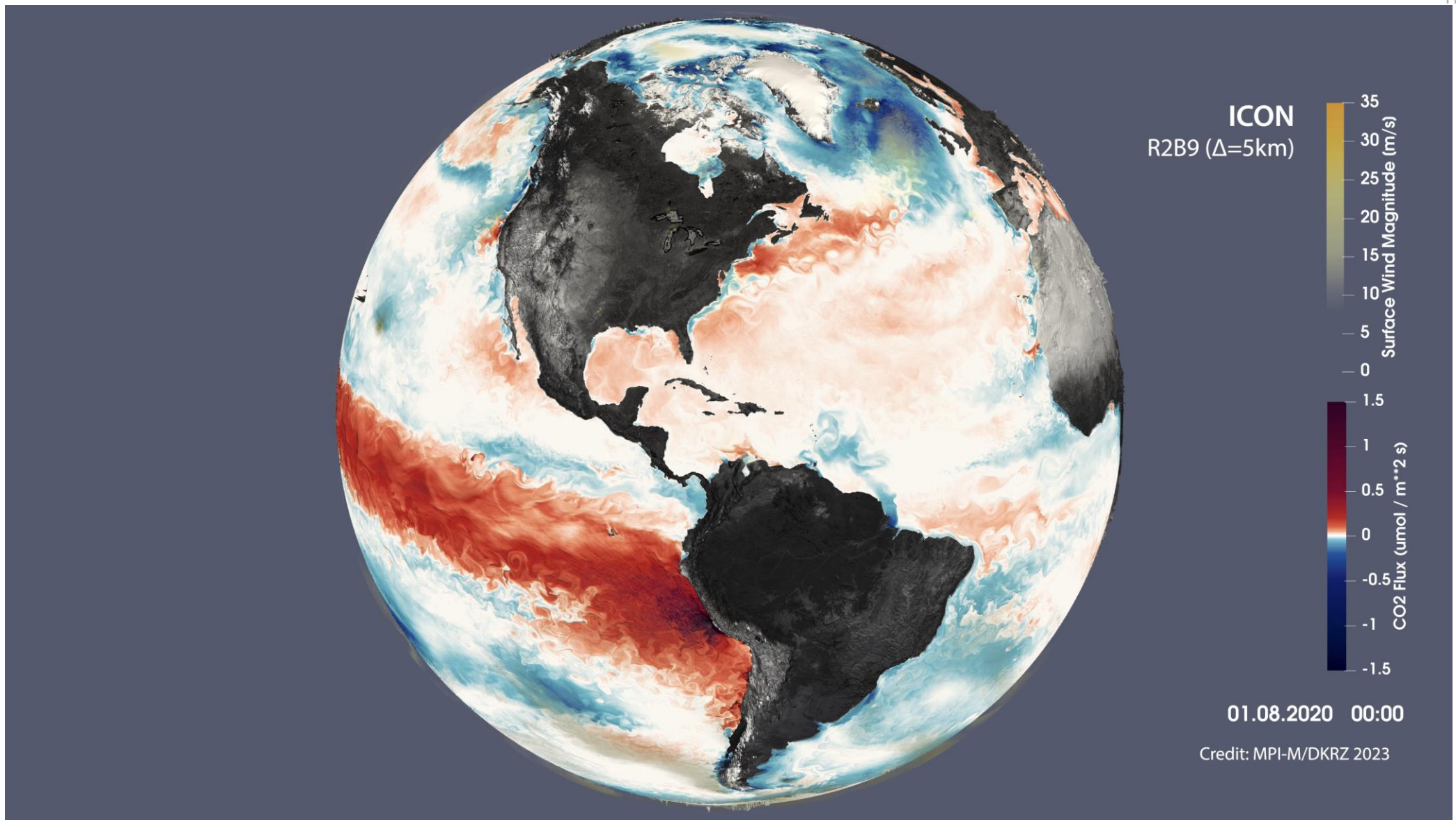
Why high resolution Earth System Models? from the ocean carbon cycle perspective

- Resolving mesoscale eddies in the ocean and convective storms in the atmosphere can directly affect ocean carbon uptake

- Resolving ocean mesoscale eddies impact:
  - upwelling of nutrients
  - local carbon export production (up to 50%) (Harrison et al 2018)
  - primary production -> climate system through the feedback of phytoplankton on ocean light absorption



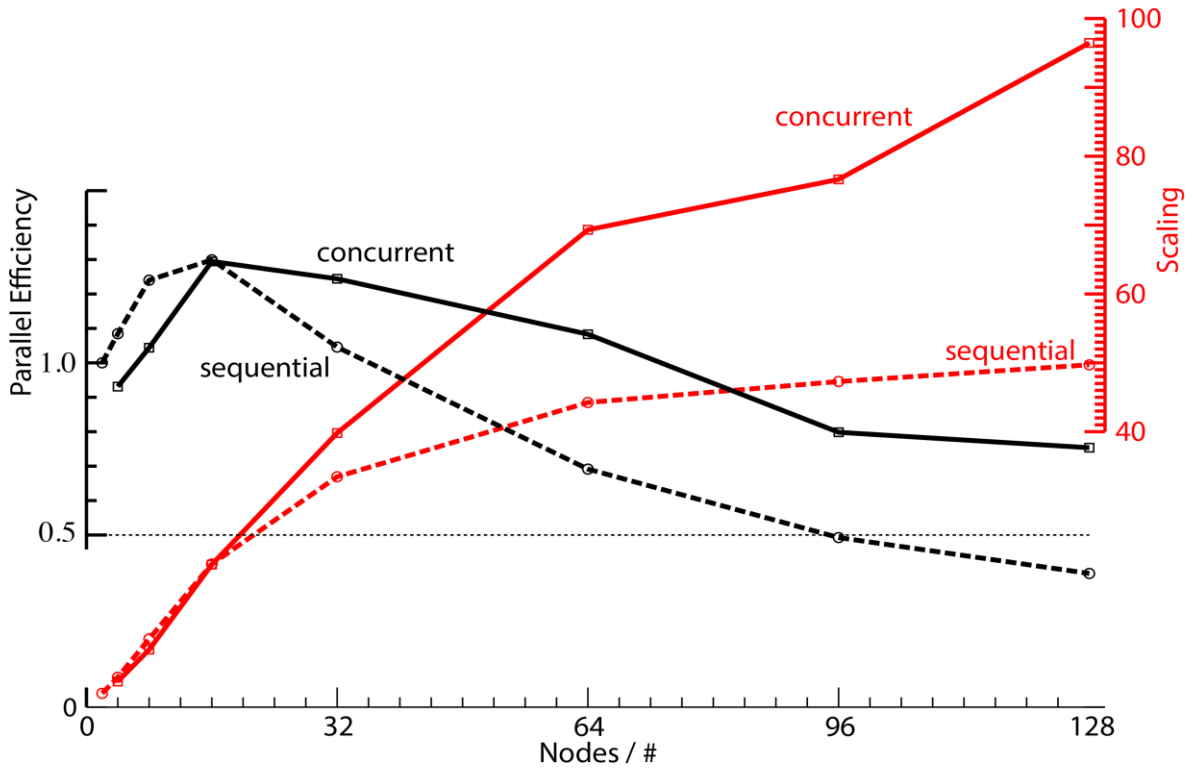
Required horizontal resolution to resolve ocean mesoscale eddies, LaCasce & Groeskamp (2020)


ICON
R2B9 (Δ=5km)
35
30
25
20
15
10
5
0
Surface Wind Magnitude (m/s)
1.5
1
0.5
0
-0.5
-1
-1.5
CO2 Flux (umol / m**2 s)
01.08.2020 00:00
Credit: MPI-M/DKRZ 2023

High Level Concurrency

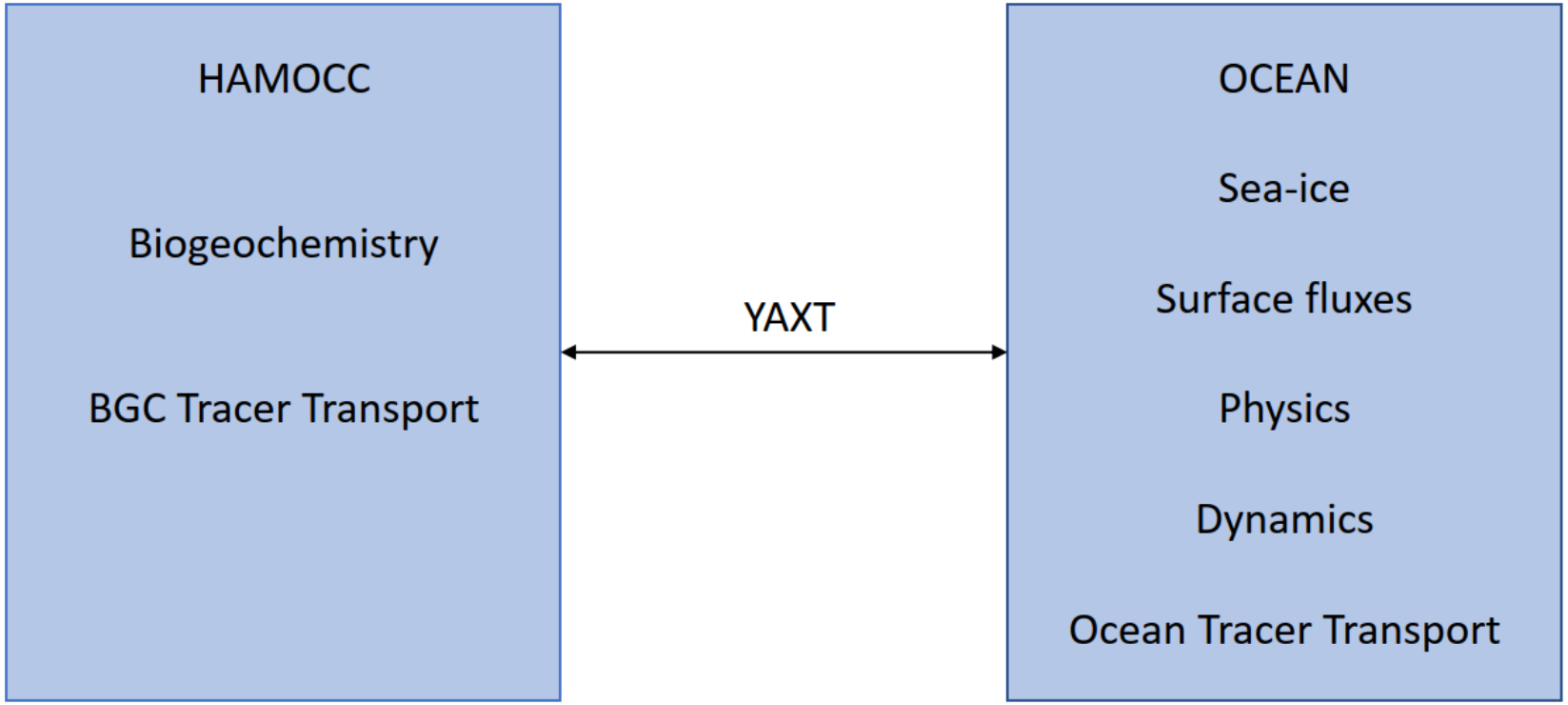
- Different components running asynchronously
- Scale the model beyond classical domain decomposition

Sequential vs Concurrent HAMOCC: Parallel Efficiency and Scaling in ICON-O-HAMOCC (40km res)



Linardakis et al 2022

HAMOCC (Object Parallelism Approach)



YAXT library: communication between two model components with different domain decompositions

Objectives of sprint

- GPU porting of missing subroutines in tracer transport and HAMOCC with further optimizations (using OPENACC)
- Investigate performance and scaling of the concurrent heterogenous setup at different resolutions (up to R2B8).

Porting to GPUs and Optimization

- Avoid temporary arrays initialization to zero in order to get rid of some unnecessary kernels
- Use an asynchronous queue in the HAMOCC model since no MPI communication is involved
- Collapse loops without dependencies in order to achieve a better scaling on GPUs

```
!$ACC PARALLEL LOOP GANG VECTOR COLLAPSE(2) DEFAULT(PRESENT) ASYNC(1) IF(lacc)
DO k = 1, max_klevs
   DO j = start_idx, end_idx
```

Porting to GPUs and Optimization

Monitoring in ICON:

- Computes global variables such as mean SST, NPP, …
- Involves **MPI reductions** every time step ➔ performs poorly on GPU

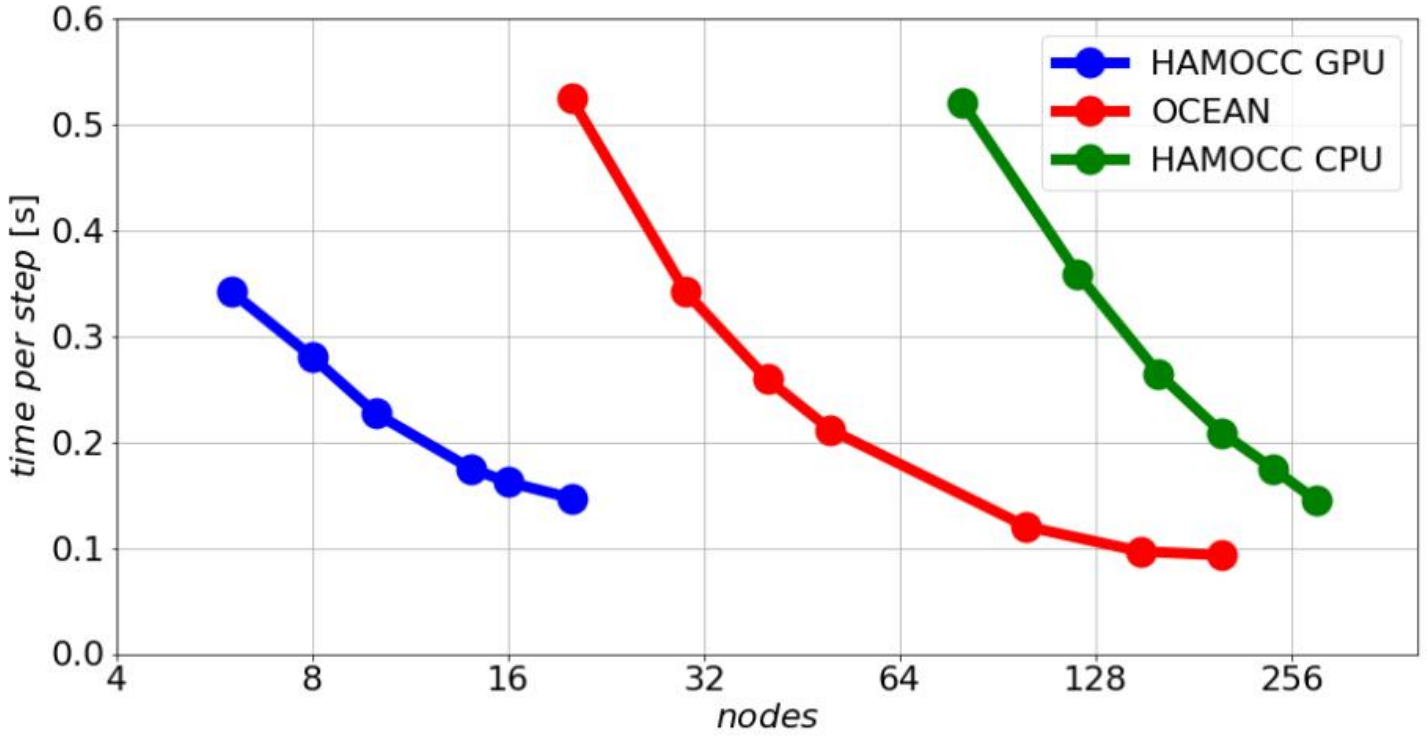
Optimizing Monitoring by implementing a 2 stage procedure:

- At each time step monitoring variables are calculated on the local partition
- At the output step the global monitoring variables are evaluated, involving MPI reductions.

➢ This implementation should also improve the scaling on a CPU system and the same approach can be applied to other model components (i.e. ocean model or atmosphere)

Results

- HAMOCC on GPU shows reasonable scaling:
  - Ratio between CPU nodes and GPU nodes is ~10-20 to 1
- At high resolutions the GPU implementation is:
  - More energy efficient
  - HAMOCC can be run at even higher resolutions and/or with more processes



Scaling plots for the 10km resolution (R2B8L128)

Challenges

- Increasing resolution lead to exponential increase in YAXT initialization:

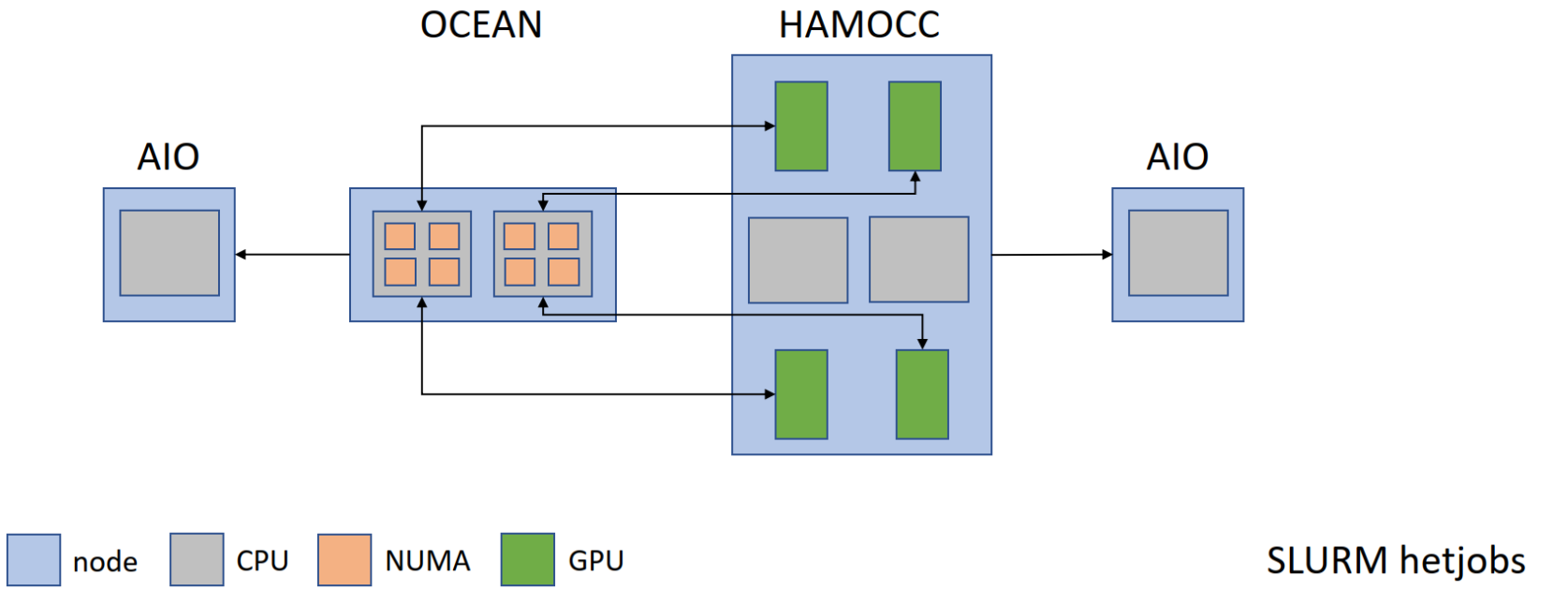
➔ A workaround was implemented by YAXT developers

- YAXT implementation is heavily based on MPI Datatypes which perform poorly on GPUs

➔ A new **exchanger** was implemented by the YAXT developers in the backend which packs/unpacks the data into a buffer before/after the send/recv call

Challenges

- Communication between ocean and HAMOCC is currently the main bottleneck of the heterogeneous setup
  - High amount of data (3D fields) is exchanged at every time step
  - High ratio of MPI processes involved in the exchange (50:1)
- Possibilities to reduce the communication time need to be explored in the future



Summary

- The sprint opens new possibilities to:
  - Run HAMOCC at higher spatial resolutions and in coupled ocean-atmosphere configurations
  - Include more processes (tracers) in HAMOCC without compromising throughput
- The sprint was useful to show:
  - Some flaws of the YAXT exchange library
  - The possibility to easily exchange data between different components on different architectures
- Possibilities to reduce the communication time need to be explored in the future in order to be able to run concurrent heterogeneous setups in production.

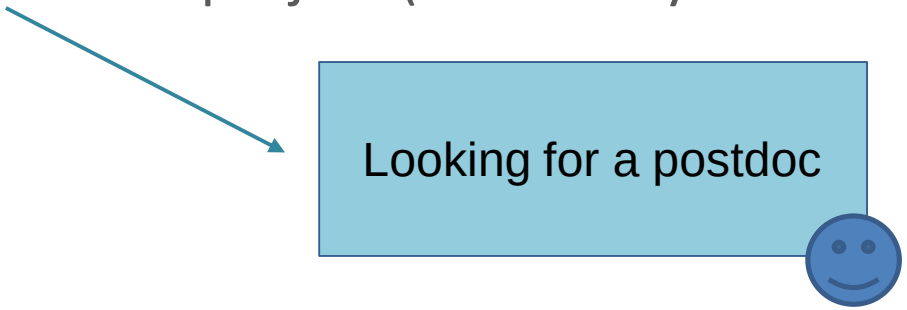
Outlook & open questions

- Scaling of HAMOCC on GPUs on Juwels-Booster (Scalexa project)

  Does the communication improve on Juwels?
- Setup a production run for ICON-ESM R2B9/R2B9 with HAMOCC on GPUs

  ➔ Next sprint?
- Develop emission-driven ICON-ESM with interactive carbon cycle

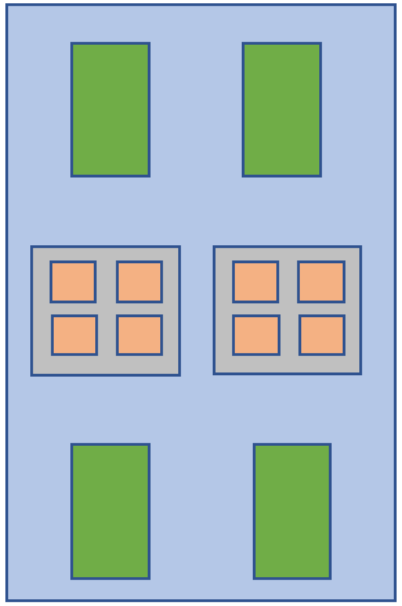
  ➔ICON-4C4M-O project (Funded by Extramural Funding program of DWD)



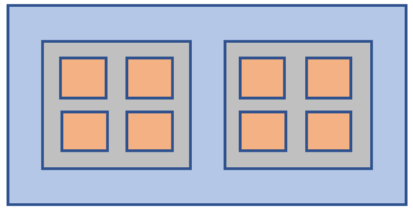
Thank you!

Complementary slides

HAMOCC + OCEAN



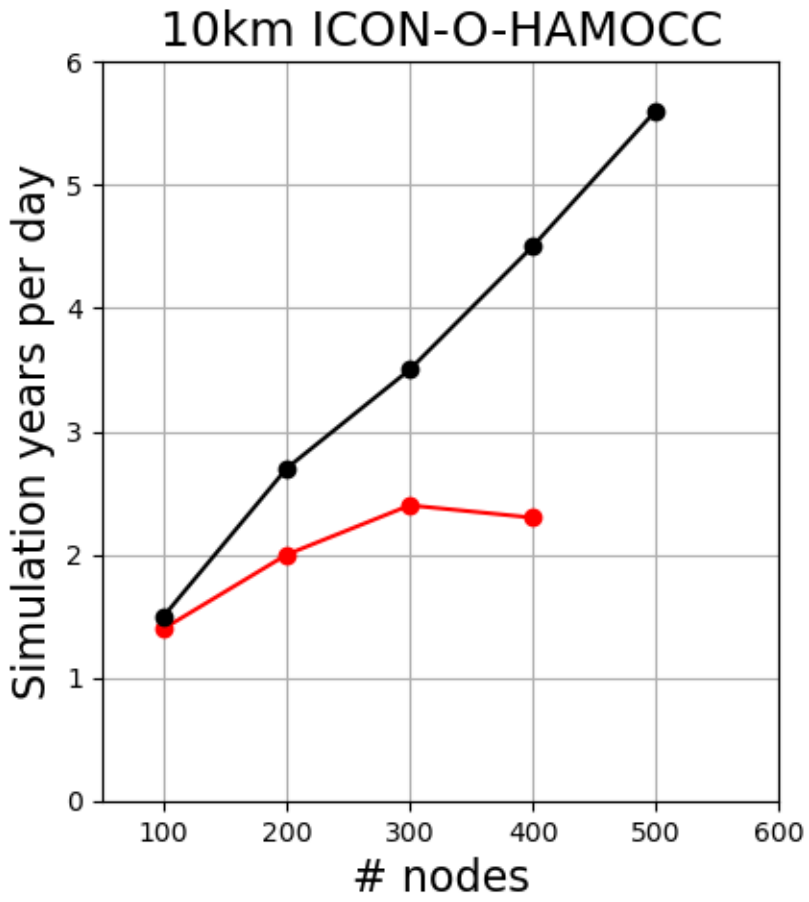
OCEAN

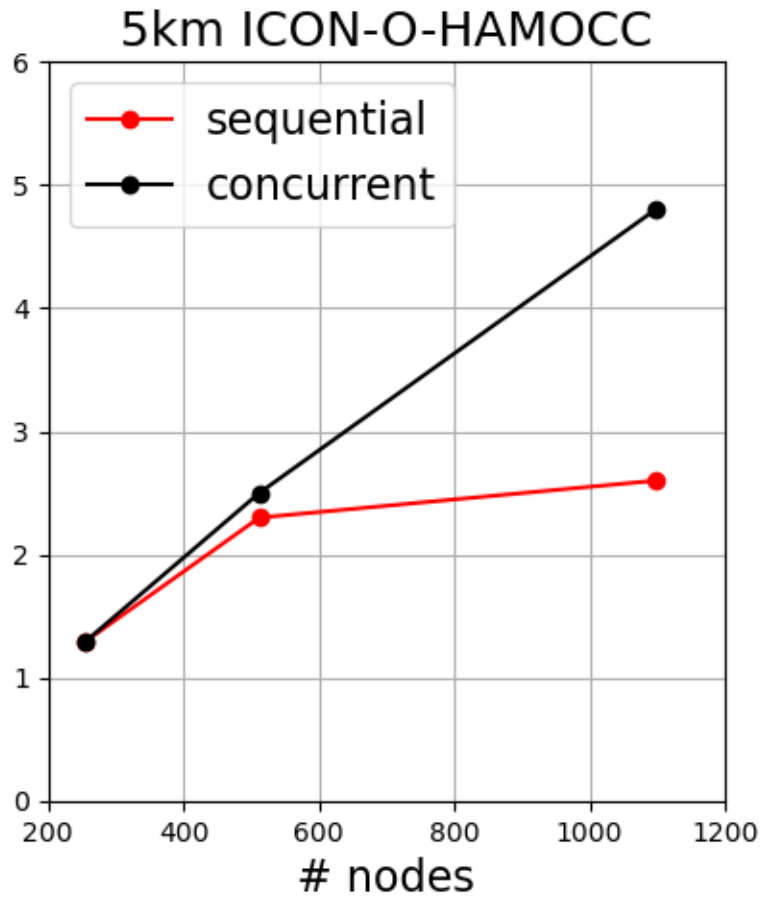


- HAMOCC runs on 4 MPI procs per node and 4 GPUs
- OCEAN runs on 28 MPI procs per node and 4 OpenMP threads
- Hetjobs to achieve load balancing

node CPU NUMA GPU

MPMD + SLURM hetjobs


10km ICON-O-HAMOCC
Simulation years per day
nodes
0
1
2
3
4
5
6
100
200
300
400
500
600


5km ICON-O-HAMOCC
sequential
concurrent
nodes
0
1
2
3
4
5
6
200
400
600
800
1000
1200