Report: ESiWACE: Scalability of Earth System Models

Project: 1040
Project title: ESiWACE: Scalability of Earth System Models
Principal Investigators: Florian Ziemen (DKRZ), Daniel Klocke (DWD), Bernadette Fritzsch (AWI), Bjorn Stevens (MPI-M), Julian Kunkel (ECMWF), Joachim Biercamp (DKRZ)
Allocation Period: 2019-01-01 to 2019-12-31

ICON

Publications based on the DYAMOND ICON Simulations from 2018

In the last year, two papers lead by MPI-M authors were published based on the results of the DYAMOND simulations performed in this project in 2018 (Stevens et al., 2019, Hohenegger et al., in press). More publications using this dataset will follow in a special issue of the Journal of the Meteorological Society of Japan[1].

Performance for coupled ocean-atmosphere simulations

High-resolution coupled ocean-atmosphere simulations were investigated in terms of performance. Considerations focussed on the DYAMOND++ set up that has been developed recently by MPI-M and uses a 5km-resolving grid for both ocean and atmosphere, cf. Fig. 1. An overview of selected run times is given in Table 1; details can be found in Neumann and Serradell (2019), that is the Deliverable 2.12 “Implementation of ICON 10km global coupled demonstrator and performance analysis” of the project ESiWACE. As the title already suggests, the original goal of a 10km-10km-coupled set up could already be overcome in this compute project.

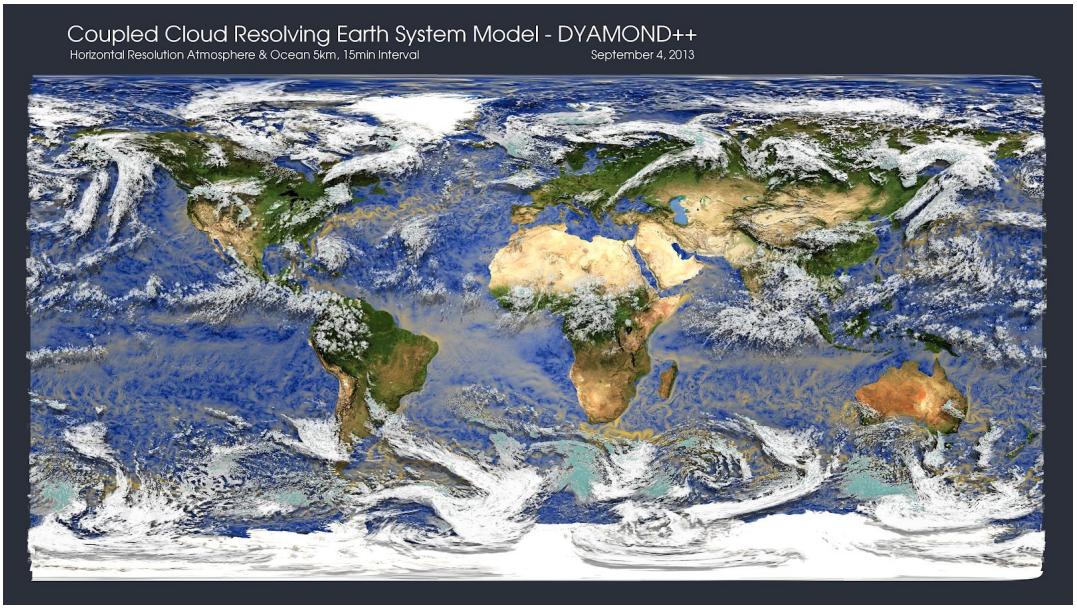

Figure 1: Visualization of vertically integrated cloud water (white) and cloud ice (teal) and ocean currens (yellow) after four days of running the 5km-5km coupled set up. Courtesy by Niklas Röber.

[1] http://jmsj.metsoc.jp/special_issues_editions/DYAMOND_info.html

Nodes	Nodes (atm)	Nodes (oce)	Notes	SDPD
420	300	120	Base line	15.6
420	300	120	nproma=32	16.4
250	150	100	Min. set up	9.8
550	450	100	Add HCOLL option	15.1
420	300	120	Add HCOLL option	14.5
500	400	100	Best throughput	19.4

Table 1: Performance in terms of simulated days per day (SDPD), using different numbers of compute nodes of partition compute2 on Mistral.

The baseline of the experiments was given by the first set up in Table 1. Tuning the parameter nproma in ICON resulted in a slight performance increase by 5%. The minimum set up in terms of compute nodes which still provides enough memory per node was found to be 250 nodes. Scaling beyond 500 nodes was difficult due to errors during MPI communicator instantiation. Although this could be overcome by additional options (HCOLL parameter), this resulted in severe performance penalties. The best throughput (19.4 SDPD) could be achieved using 500 nodes.

Sparse Grid Regression for Performance Prediction

We further investigated the sparse grid regression technique for performance prediction in high-resolution atmosphere-only ICON simulations, as subject of the DYAMOND project. Sparse grids allow to describe and explore high-dimensional spaces under certain smoothness assumptions. Since weather and climate models inherently depend on a multitude of parameters with regard to performance, we employed sparse grids to discretize the corresponding parameter space and employed adaptive mesh refinement for the sparse grids to provide higher accuracy in regions which exhibit higher gradients with regard to parameter variation. We measured the performance of the ICON 5km-resolving model in all combinations spanned by the number of nodes (100/200/300/400), no. of OpenMP threads (1/2/4/6/12/18), no. of vertical levels (60/70/80/90), nproma (2/4/8/16/32). Random sub-sets were chosen out of this set and the average relative prediction error on the remaining sub-set was evaluated, using the sparse grid regressor. This procedure was repeated 10 times, resulting in the shown mean relative error in Fig. 2: only using 10% of the compute samples resulted in a prediction error of less than 10%. We carried out similar experiments also for the ICON climate benchmark V16.0 (aqua-planet experiment)[2]. Findings were published in Neumann (2019).

[2] https://redmine.dkrz.de/projects/icon-benchmark/wiki/Performance_ICON_Benchmark_v160

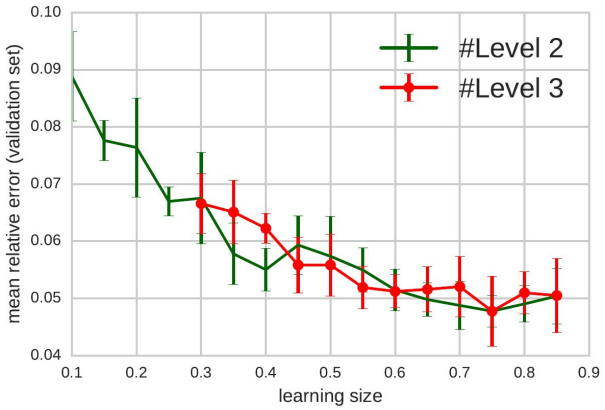

Figure 2: Mean relative error from sparse grid regression for ICON-5km simulations (Neumann, 2019). “Learning size” indicates the percentage of samples used for training the sparse grid. “Level 2” and “Level 3” correspond to two sparse grid variants.

FESOM

During the reporting period, the Finite-volumE Seaice–Ocean circulation Model, Version 2.0 (FESOM2), was further optimized. A methodology for hierarchical mesh partitioning developed at the AWI computing center was validated. The method applies heuristic mesh partitioning preprocessing based on the topology of the targeted computing system. The 3D computational parts scale very well. The 2-D kernels were analyzed and improvements incorporated. This includes overlapping computation with communication in the solver for the SSH part. In order to improve the scalability of the seaice model, a modified elasto-viscous-plastic approach is used (Koldunov et al, 2019b). Extensive Scalability tests were performed at DKRZ, which are described in Koldunov et al. (2019a)

ESDM

High-resolution climate simulations produce large amounts of data. This is also the case for the DYAMOND simulations, with overall outputs per model run in the range of 8-132 TB. The Earth-System Data Middleware (ESDM) is developed in the scope of ESiWACE to improve data reading/writing from such memory-intensive simulations to storage.

During the reporting period, various short benchmark runs have been executed on 10-500 nodes of Mistral and the software has been hardened to allow for production. The results for running the benchmarks on 200 nodes with varying numbers of processes are shown in Figure 3. The figure shows the results for different processes per node (x-axis) considering ten timesteps of 300 GB data each.

As the baseline for exploring the efficiency, we run the IOR benchmark using optimal settings (i.e., large sequential I/O). The graphic shows two IOR results: storing file-per-process (fpp) on Lustre (ior-fpp), as this yielded better performance than the results for shared file access, and storing fpp on local storage (ior-fpptmp).

Mistral has two file systems (Lustre01 and Lustre02) and five configurations with ESDM were tested: storing data only in Lustre02, settings where data are stored on both Lustre file systems concurrently (both), and environments with in-memory storage (local tmpfs). We also explored if fragmenting data into 100MB files and 500MB files is beneficial (the large configurations). Note that the performance achieved on a single file system is slightly faster to the best-case performance achieved with optimal settings using the benchmarks. We conclude that the fragmentation into chunks accelerates the benchmark.

By utilizing the two file systems resembling a heterogeneous environment effectively, we can improve the performance from 150 GB/s to 200 GB/s (133% of a single file system). While this was just a benchmark testing, it shows that we were able to exploit the available performance and thus, we claim ESDM is beneficial for IO intense climate/weather models.

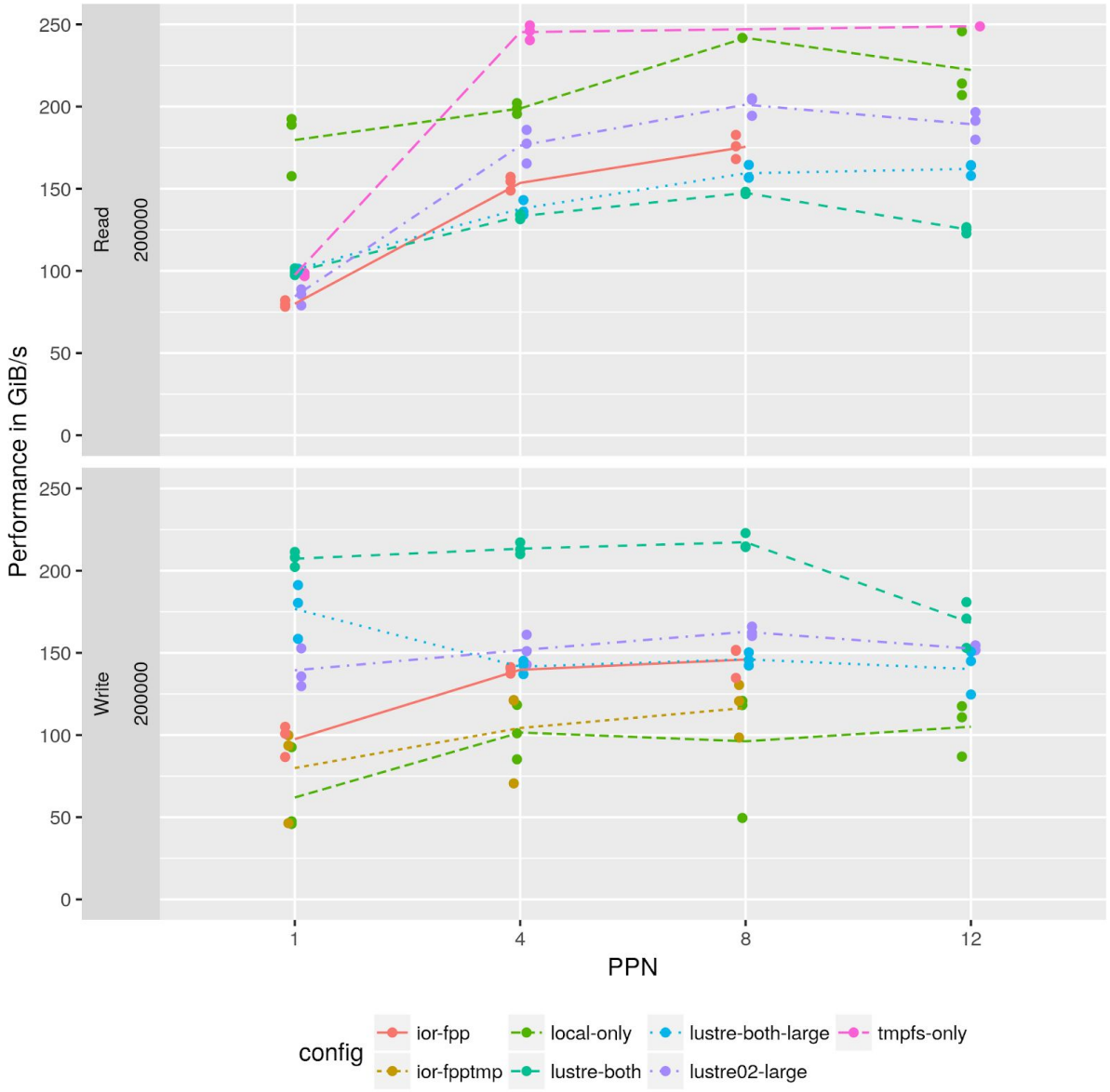

Figure 3: Performance of ESDM and comparison to the IOR benchmark for various configurations on Mistral.

References

- **Hohenegger**, C., Kornblueh, L., Klocke, D., Becker, T., Cioni, G., Engels, J. F., et al.: Climate statistics in global simulations of the atmosphere from 80 to 2.5 km grid spacing. *Journal of the Meteorological Society of Japan, 98* (**in press** for Spec. Ed. on DYAMOND, 2020).
- **Koldunov**, N. , Aizinger, V. , Rakowsky, N. , Scholz, P. , Sidorenko, D. , Danilov, S. and Jung, T.: Scalability and some optimization of the Finite-volumE Sea ice–Ocean Model, Version 2.0 (FESOM2) , Geoscientific Model Development, 12 (9), pp. 3991-4012. doi: 10.5194/gmd-12-3991-2019, **2019a**
- **Koldunov**, N. V., Danilov, S., Sidorenko, D., Hutter, N., Losch, M., Goessling, H., Rakowsky, N., Scholz, P., Sein, D., Wang, Q., and Jung, T.: Fast EVP Solutions in a High-Resolution Sea Ice Model, J. Adv. Model. Earth Syst., 11, 1269–1284, https://doi.org/10.1029/2018MS001485, **2019b**
- **Neumann**, P., K. **Serradell**. Implementation of ICON 10km global coupled demonstrator and performance analysis (D2.12). Deliverable of the project ESiWACE, https://doi.org/10.5281/zenodo.2596977, **2019**
- **Neumann**, P. Sparse Grid Regression for Performance Prediction Using High-Dimensional Run Time Data. Accepted for publication in Euro-Par 2019: Parallel Processing Workshops, **2019**
- **Stevens**, B.; Satoh, M.; Auger, L.; Biercamp, J.; Bretherton, C. S.; Chen, X.; Düben, P.; Judt, F.; Khairoutdinov, M.; Klocke, D.; Kodama, C.; Kornblueh, L.; Lin, S.-J.; Neumann, P.; Putman, W. M.; Röber, N.; Shibuya, R.; Vanniere, B.; Vidale, P. L.; Wedi, N. & Zhou, L.: DYAMOND: the DYnamics of the Atmospheric general circulation Modeled On Non-hydrostatic Domains, *Progress in Earth and Planetary Science,* **2019***, 6*, 61