Performance Modeling and Scalability for the ICON model

Philipp Neumann, Panagiotis Adamidis, Joachim Biercamp
Deutsches Klimarechenzentrum (DKRZ)

Luis Kornblueh, Matthias Brück
Max-Planck-Institut für Meteorologie

Daniel Klocke
Deutscher Wetterdienst

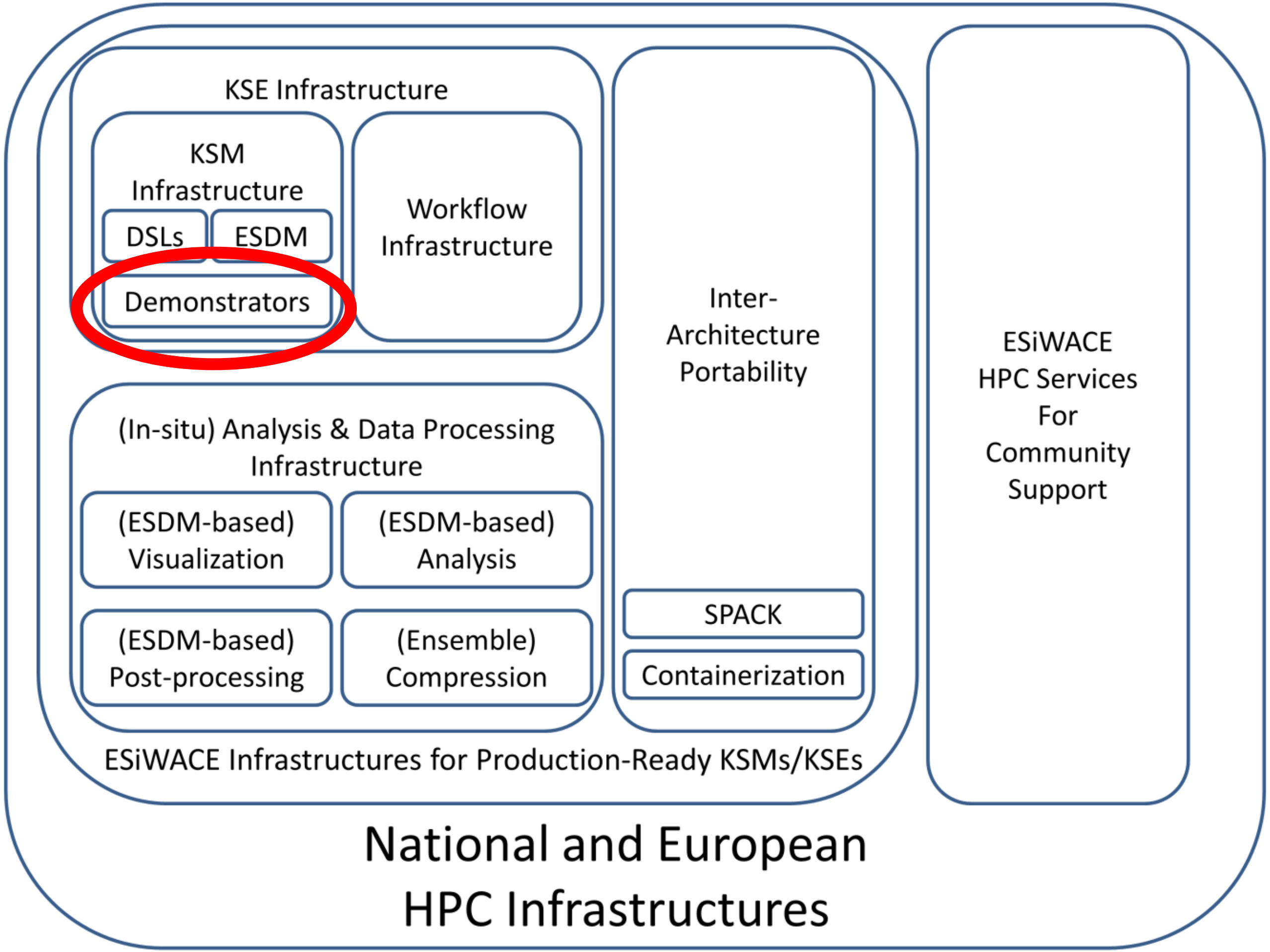
Niklas Röber
DKRZ

…and the members of ESiWACE and DYAMOND

ESiWACE: Centre of Excellence in Simulation of Weather and Climate in Europe

- Goals
  - Substantially improve efficiency and productivity of weather and climate models
  - Prepare models for exascale systems
    → scalability and performance analysis, tuning, …
- ESiWACE: **Kilometre-scale demonstrators (prototypical)**
  → models: ICON, IFS, NEMO, EC-Earth
- ESiWACE2: Towards production-ready models at pre-exascale
- Read more: Website: www.esiwace.eu
  ESiWACE newsletters: www.esiwace.eu/newsletter

ESiWACE



P. Neumann, J. Biercamp. ESiWACE: On European Infrastructure Efforts for Weather and Climate Modeling at Exascale. Submitted, 2019

ICON Aqua-Planet Experiment @1.25km

- **1.25km resolution,**
  335 544 320 horiz. cells,
  45 vert. levels
- **1408 nodes,**
  2MPI x 18 OpenMP
- **Throughput: 1.8 SDPD**,
  no IO
- Benchmark (160km-5km)
  available at:

https://redmine.dkrz.de/projects/icon-benchmark/wiki/Instructions_on_download_execution_and_analysis_ICON_Benchmark_v160

DYAMOND: **Dy**namics of the **A**tmospheric General Circulation **M**odeled **o**n **N**on-hydrostatic **D**omains

- Intercomparison of O(3km) atmospheric global models
  → ICON, NICAM, MPAS, GEOS, FV3, SAM, UM, ARPEGE-NH, IFS-H
- ICON 2.5km throughput: ca. 6 SDPD (540 nodes/19 440 cores)
- Scientific use case of ESiWACE demonstrators
  Read more: www.esiwace.eu/services/dyamond

Outline

1. Accessible resolutions
2. **Semi-analytical performance modeling for hi-res runs**
3. **Performance modeling and prediction with sparse grids**

Goal: Provide means to portable scalability estimates (with exascale and high-dimensionality in mind)

Read more:

1. P. Neumann et al.
   Phil. Trans. R. Soc. A. 377:20180148, 2019
2. T. Schulthess et al.
   IEEE Computing in Science & Engineering 21(1):30-41, 2018
3. B. Stevens et al. DYAMOND (Submitted)

Performance Models for Scalability Predictions at Exascale (ICON-DYAMOND 5km)

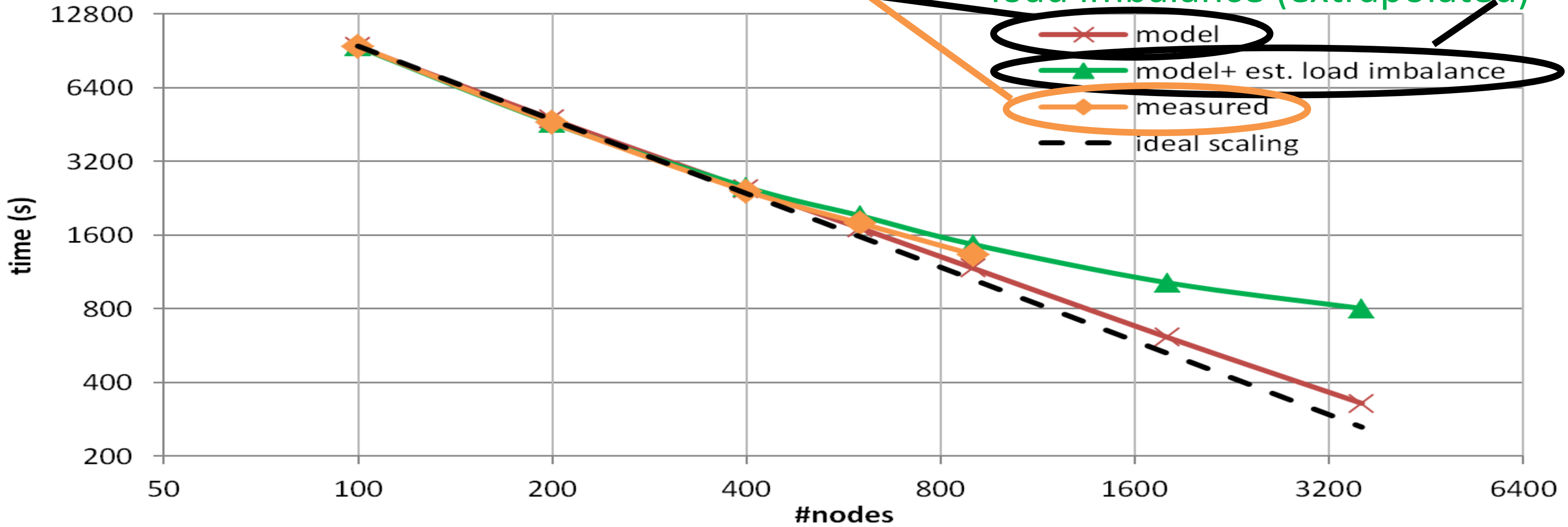
- Modeling of atmosphere-only ICON-DYAMOND 5km setup
  → domain decomposition of unstructured grid
    → load imbalances due to different subdomain sizes (and varying with message sizes)
  → additional load imbalances due to, e.g., cloud cover
  → 39 communication phases (nearest-neighbor exchange of cell/vertex/edge data)
- Model 1: $t(N) := t_{compute}(N) + t_{commun}(N)$ with
  $t_{compute}(N) := t_{compute}(1)/N$,
  $t_{commun}(N)$ modeled using message sizes, network latency and bandwidth
  → fully analytical hardware-aware model, given one measurement $t(1)$
- Model 2: same as model 1, but additionally models
  $t_{compute}(N) := (t_{compute}(1) - t_{imbalance}(1))/N + t_{imbalance}(N)$,
  $t_{imbalance}(N)$ measured/extrapolated
  → semi-analytical hardware-aware model

Performance Models for Scalability Predictions at Exascale (ICON-DYAMOND 5km)

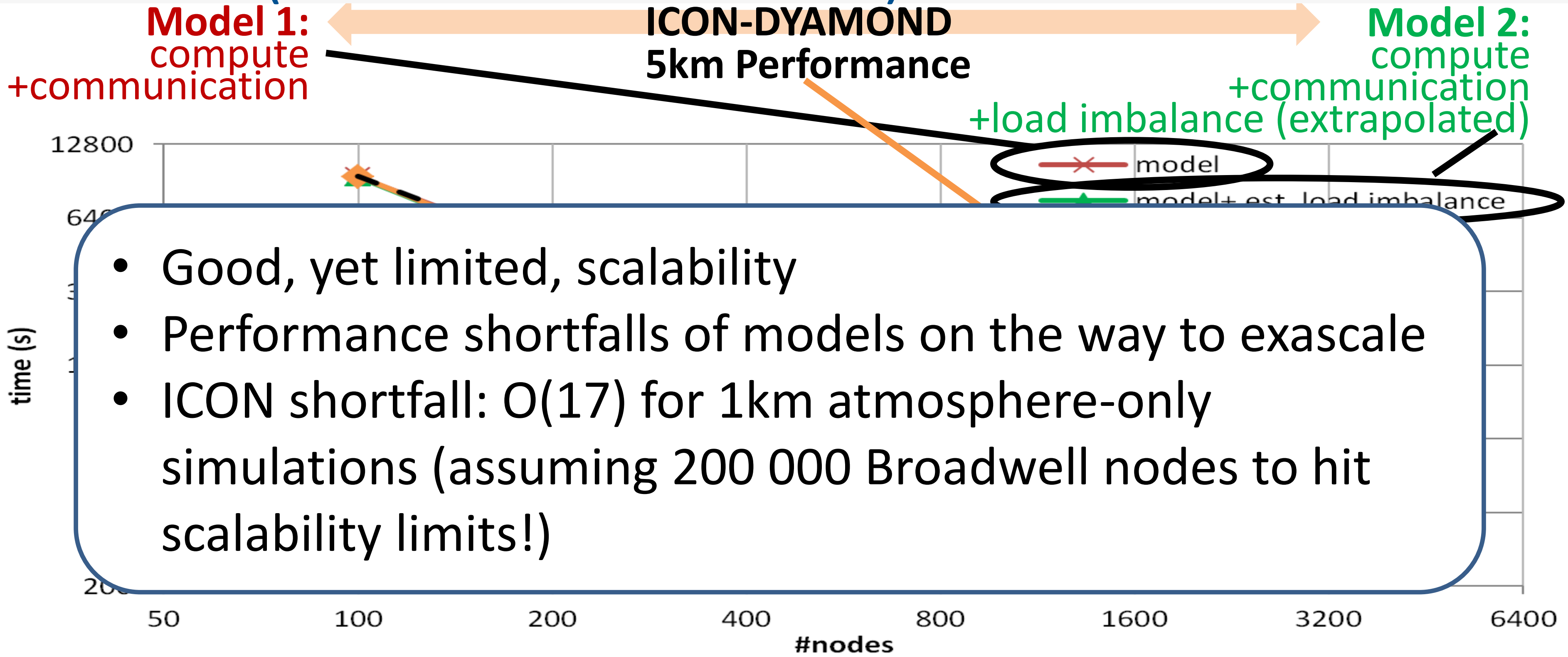
Model 1: compute +communication

ICON-DYAMOND 5km Performance

Model 2: compute +communication +load imbalance (extrapolated)



Performance Models for Scalability Predictions at Exascale (ICON-DYAMOND 5km)



- Good, yet limited, scalability
- Performance shortfalls of models on the way to exascale
- ICON shortfall: O(17) for 1km atmosphere-only simulations (assuming 200 000 Broadwell nodes to hit scalability limits!)

Performance Modeling with Sparse Grids: Objectives

- **Multi-parameter influence on computational performance**
  → **computational:** OpenMP/MPI decomposition, loop-blocking, vector lengths, …
  → **algorithmic:** time step, number of iterations, error control/tolerance,…
  → **all aforementioned categories for every model subcomponent**
  → **high-dimensional parameter space**

- **Objectives: performance estimate for complex ESMs…**
  …to gain insight into (wanted or unwanted) hotspots
  …to improve scheduling (relevant to workflows?)

- **<u>Approach:</u>** **Regression on high-dimensional parameter space via adaptive sparse grids**

Regression on Sparse Grids in a Nutshell[1]

- Define linear hat function per sparse grid point
  → defines function space V_n
- Solve regression problem on run time data y_j, measured for parameter combination x_j:

$$u = \arg\min_{v \in V_n} \left(\frac{1}{M} \sum_{j=1}^{M} (y_j - v(\vec{x}_j))^2 + \lambda C(v) \right)$$

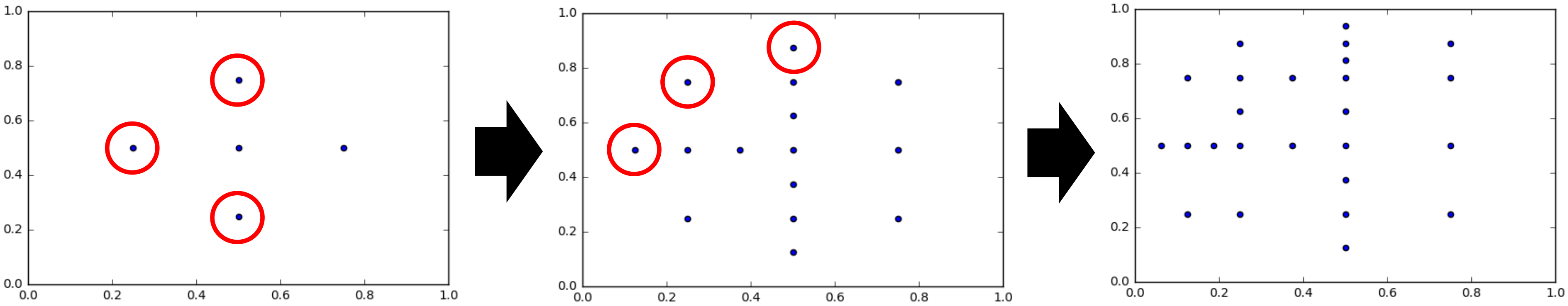
with $v(\vec{x}) := \sum_i \alpha_i \varphi_i(\vec{x})$

- Results in linear system: $\left(\frac{1}{M} B B^\top + \lambda \mathbb{I} \right) \vec{\alpha} = \frac{1}{M} B \vec{y}$

1 D. Pflüger. Dissertation, 2010

Evaluation Procedure

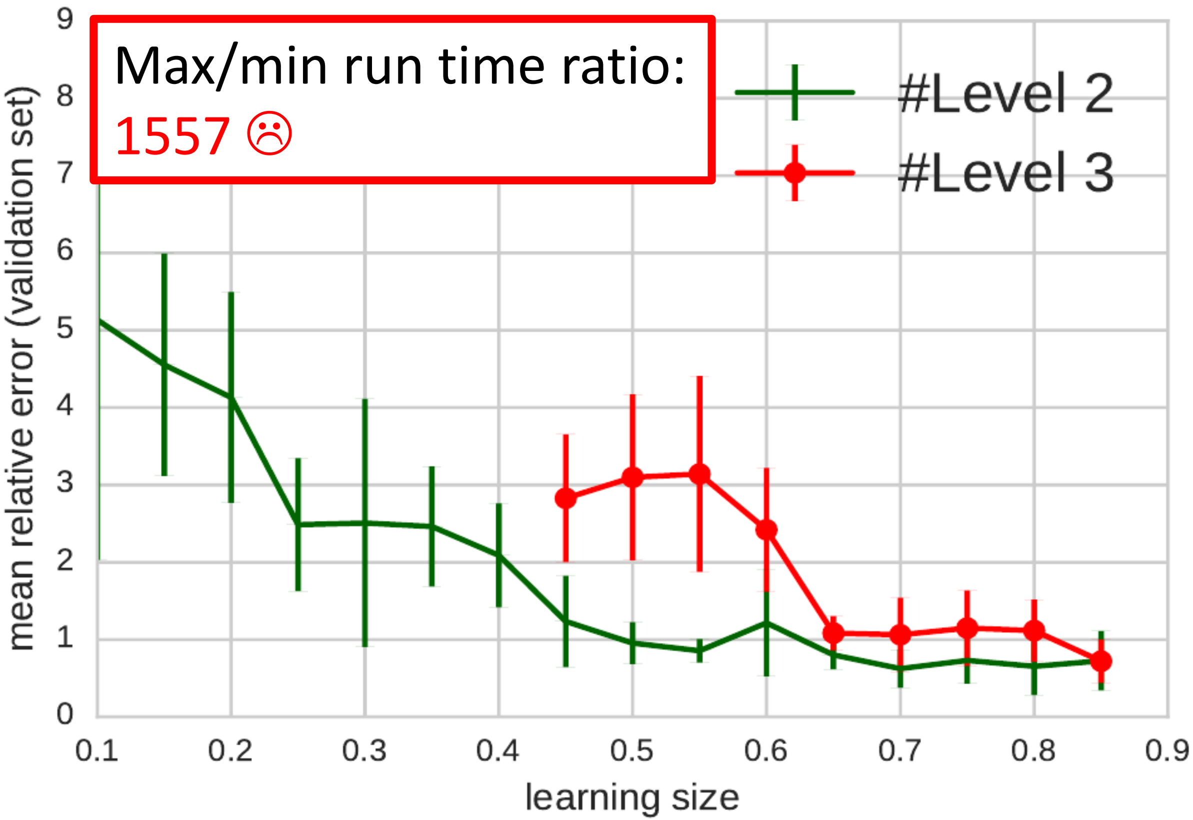
- Data splitting:
  Use s % of data for learning and $1\text{-}s$ % for validation
- Mean relative error:
  - Start from one data split
  - Compute and average relative errors for this data split
  - Repeat this procedure for 10 data splits and average errors
- Consider different initial sparse grid level refinements (level-2 and level-3 grids)
- Apply local refinement ($r=3$ refinement iterations, $m=3$ refinements/it.)



Example: Particle Simulation (5D Parameter Space)

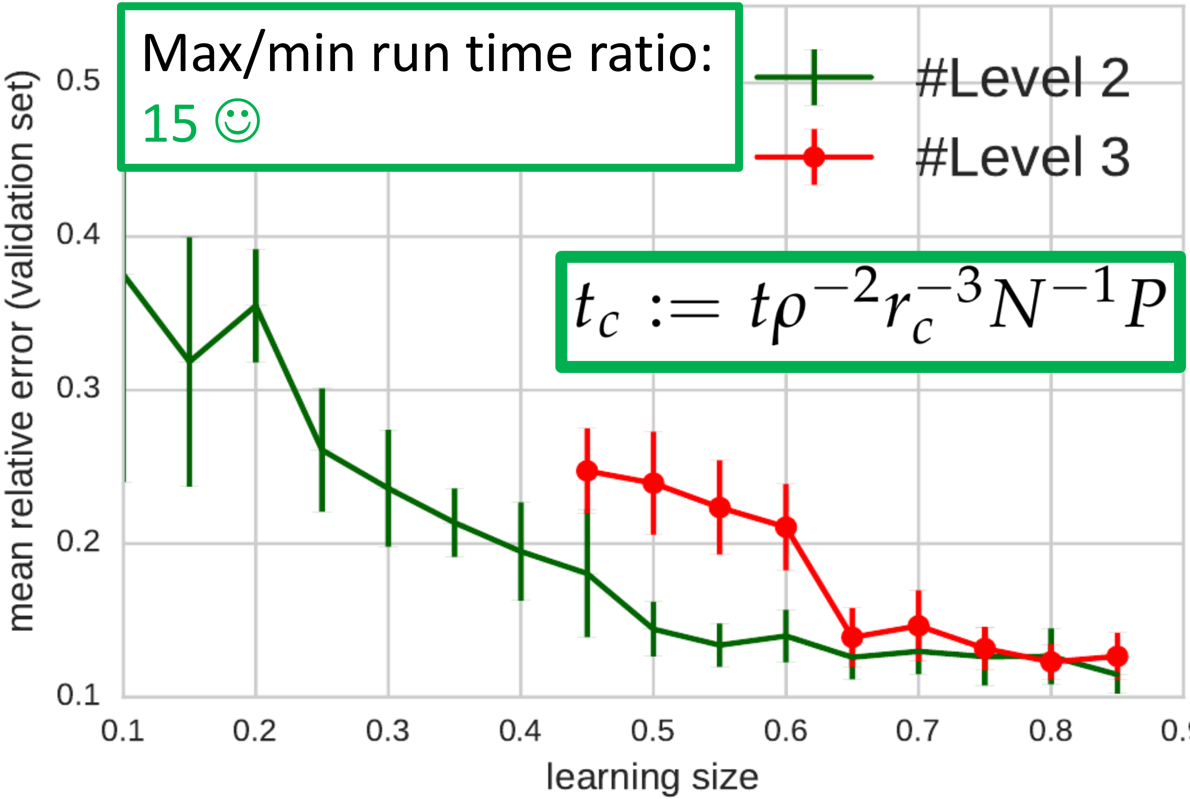
- Parameters:
  particle density ρ,
  number of particles N,
  cut-off radius r_c,
  blocksize,
  no MPI processes P
- Random sampling of run time space
  → 357 samples

Max/min run time ratio:
1557 ☹



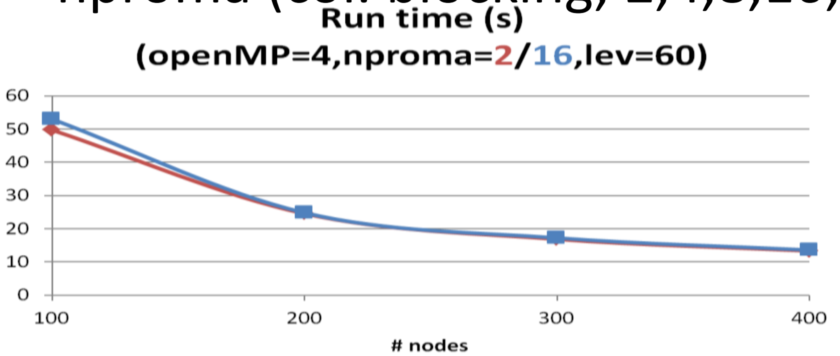
Example: Particle Simulation (5D Parameter Space)

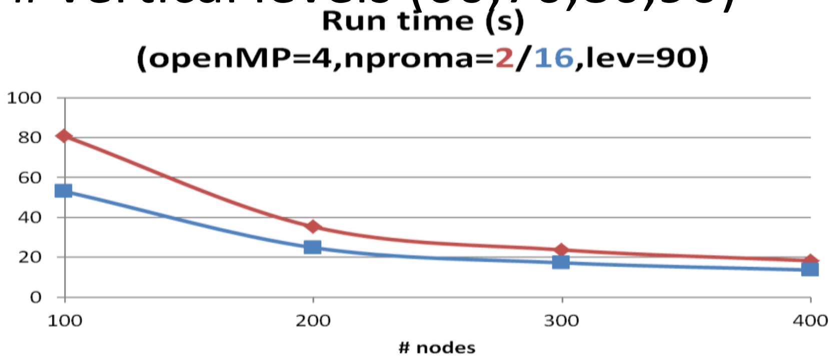
- Parameters:
  particle density ρ,
  number of particles N,
  cut-off radius r_c,
  blocksize,
  no MPI processes P
- Random sampling of run time space
  → 357 samples

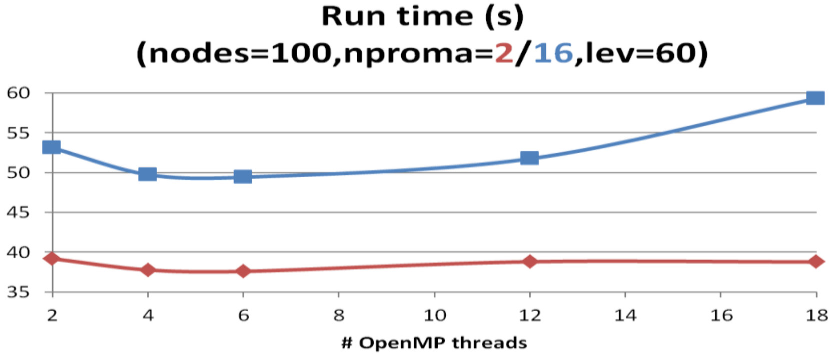


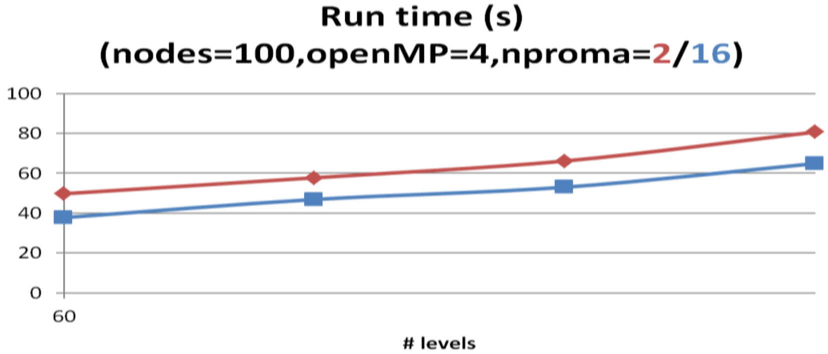
Example: ICON-DYAMOND 5km Runs (4D Space)

- Parameters:
  # OpenMP threads (2,4,6,12,18), # nodes (100,200,300,400),
  nproma (col. blocking; 2,4,8,16,32), # vertical levels (60,70,80,90)



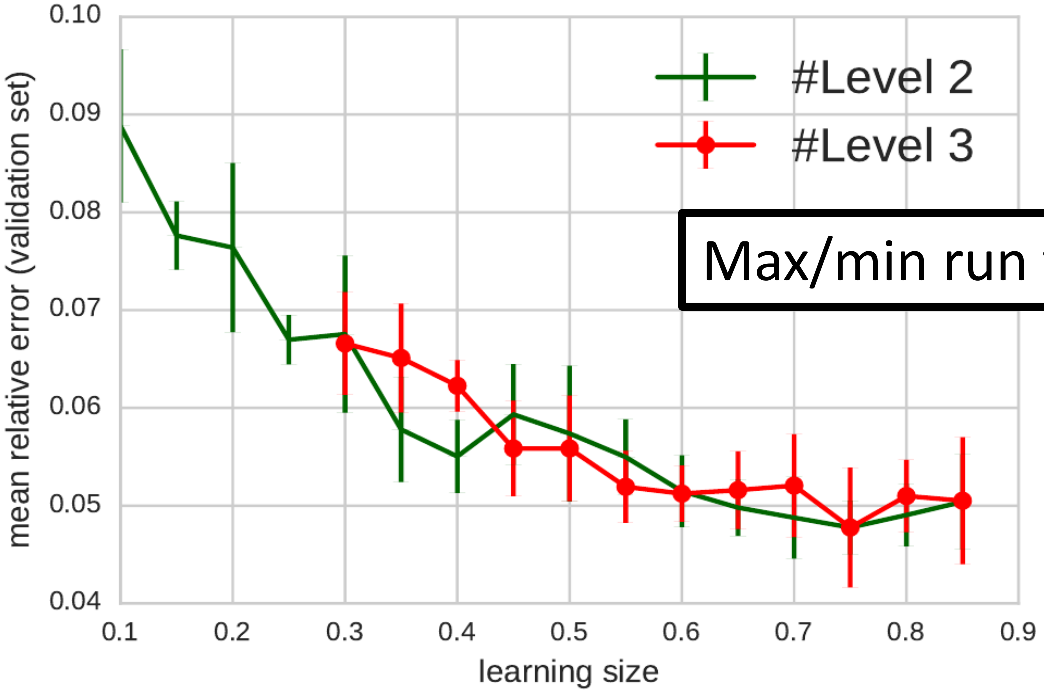






Example: ICON-DYAMOND 5km Runs (4D Space)

- Parameters:
  # OpenMP threads (2,4,6,12,18), # nodes (100,200,300,400), nproma (col. blocking; 2,4,8,16,32), # vertical levels (60,70,80,90)



Max/min run time ratio: 2

Summary

- DYAMOND/ESiWACE: Towards production-ready scalable global hi-res modeling
  → scalability and performance
  → scientific insight and model intercomparison (DYAMOND)
- Performance shortfall of global high-resolution models (still) circumvents (sub)-kilometre-scale simulations
  → factor O(17) for ICON, similar for other models
  → ***this factor is (quasi-)independent from the supercomputer's size!***
- Scalability investigation and prediction via performance modeling
  → ***semi-analytical model for ICON-5km describes model's scaling behaviour well***
- Performance prediction for arbitrary parameters
  → ***sparse grid regression for high-dimensional parameter spaces works well***
  → future: comparison with neural networks, Gaussian processes

P. Neumann acknowledges ESiWACE. ESiWACE has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 675191. This material reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains.