Deutscher Wetterdienst
Wetter und Klima aus einer Hand

AICON – Introducing ML-based weather forecasting at DWD

anemoi

Florian Prill, Marek Jacob & DWD AICON Team
ECMWF Workshop on HPC | 17 September 2025

Warm Up: Context & Goals

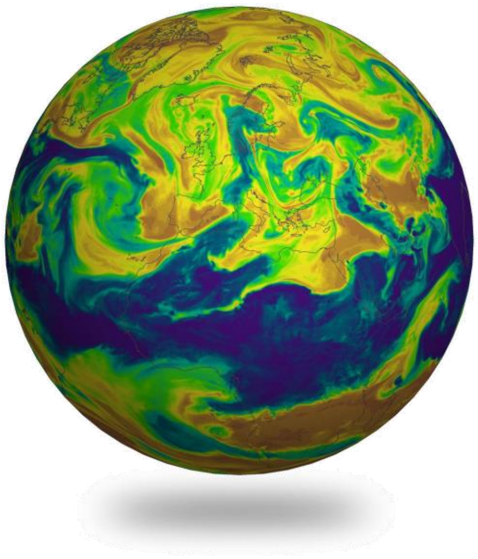
Substantial progress in the realm of data-driven weather forecasting: *FourCastNet*, *Pangu-Weather*, *GraphCast*, *GenCast*, *Aurora* (NVIDIA/Huawei/Google Deepmind, Microsoft), …

DWD numerical weather prediction: ML-based forecasts as a complement to global and regional ICON model.

Operational NWP System AICON-Global

Current status (2025-09-03): AI-based forecasts provided for evaluation, research, and training. Complements but does not replace the current operational model, forecast skill is still evolving.

Example of an AICON-Global forecast for specific humidity at the respective ICON vertical level 101



Source: 2025-09-03, Operationelles NWV-System Änderungsmitteilung

The Anemoi Framework

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

Anemoi

Python-based toolkit for data-driven models
Github: https://github.com/ecmwf/anemoi-core

PyTorch PyTorch Lightning PyG
Zarr xarray earthkit

https://events.ecmwf.int/event/410

ECMWF

Lang, S., Alexe, M., Chantry, M., Dramsch, J., Pinault, F., Raoult, B., Clare, M. C. A., Lessig, C., Maier-Gerber, M., Magnusson, L., Ben Bouallègue, Z., Prieto Nemesio, A., Dueben, P. D., Brown, A., Pappenberger, F., & Rabier, F. (2024). AIFS -- ECMWF's data-driven forecasting system. arXiv. https://doi.org/10.48550/arXiv.2406.01465

Nipen, T. N., Haugen, H. H., Ingstad, M. S., Nordhagen, E. M., Salihi, A. F., Tedesco, P., Seierstad, I. A., Kristiansen, J., Lang, S., Alexe, M., Dramsch, J., Raoult, B., Mertes, G., & Chantry, M. (2024). Regional data-driven weather modeling with a global stretched-grid [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2409.02891

The Anemoi Framework

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

Collaborative European initiative

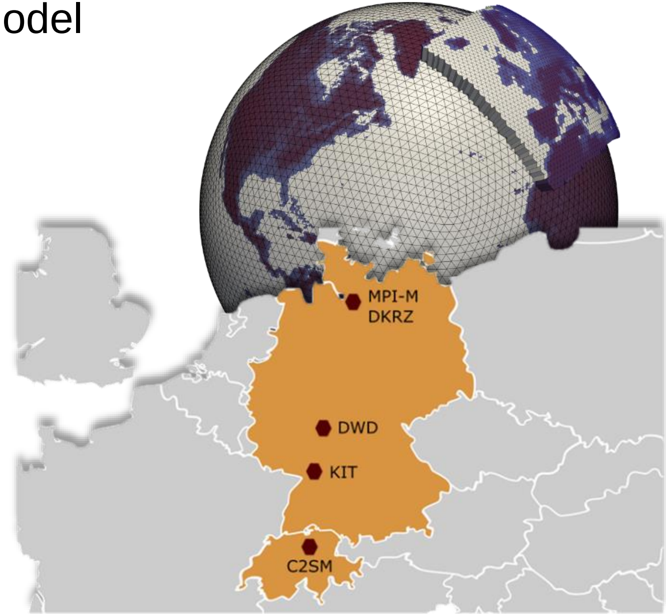
- Anemoi plays a key role in the development of multiple ML-powered weather models:
  AIFS (Artificial Intelligence Forecasting System, ECMWF), **Bris** (MetNorway, extends the AIFS) and **AICON** (DWD).
- DWD abandoned its in-house development in June 2024 in favor of the shared Anemoi codebase.
- The Anemoi Framework received the EMS Technology Achievement Award 2025.
- Development related to EUMETNET E-AI: Artificial Intelligence and Machine Learning in Weather, Climate and Environmental Application

ICON Model

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

ICOsahedral Nonhydrostatic: global and regional grid point model

- numerical grid point model, developed 2004 – today
- in operational production at DWD since 2015
- applicability on a wide range of scales from ~100km to ~100m, local mass conservation, tracer air-mass consistency



Duration of a deterministic 180h ICON run: 50 min
48 VE nodes NEC SX Aurora 1

see the ICON talk by Uli Schättler, Thursday, 18 September!

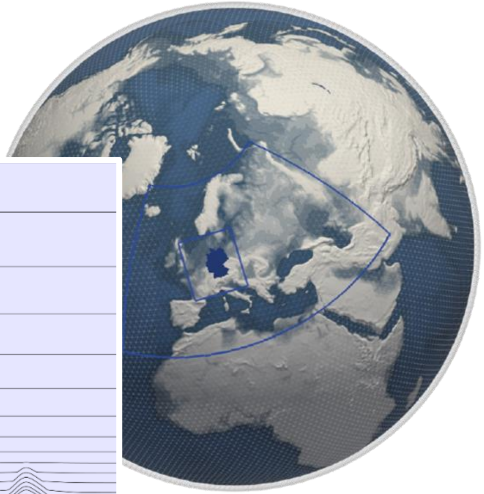
ICON Reanalysis Dataset

The data-driven AICON-Global is based on the ICON reanalysis dataset

ICON DREAM = Dual resolution Reanalysis for Emulators, Applications and Monitoring

- 13 km mesh, 2,949,120 horizontal data points
- (current) reanalysis time range: 2010-01 until 2025-04
- storage size of dataset: 6.45 PB
- 2 km regional reanalysis: currently work in progress
- no release yet; a service is planned for 2026

Operational ICON domains (global/regional)



ICON Smooth Level Vertical (SLEVE) coordinate;
Note: different from ERA5 reanalysis dataset with (hybrid sigma-)pressure levels!

Ref.: ICON-DREAM: A new dual resolution reanalysis from DWD. 6th WCRP International Conference on Reanalysis. (2024). https://confit.atlas.jp/guide/event/icr6/subject/OR2-01/detail (A. Valmassoi et al., DWD)

AICON Model Architecture

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

Reduced-level Zarr dataset: 29.5 TB

prognostic

```
PS, P[:], T[:], U[:], V[:], QV[:], T_2M, U_10M, V_10M, QV_S, RELHUM_2M, T_G, ALB_RAD,
H_SNOW, SMI[0/1], T_SO[0/1],
```

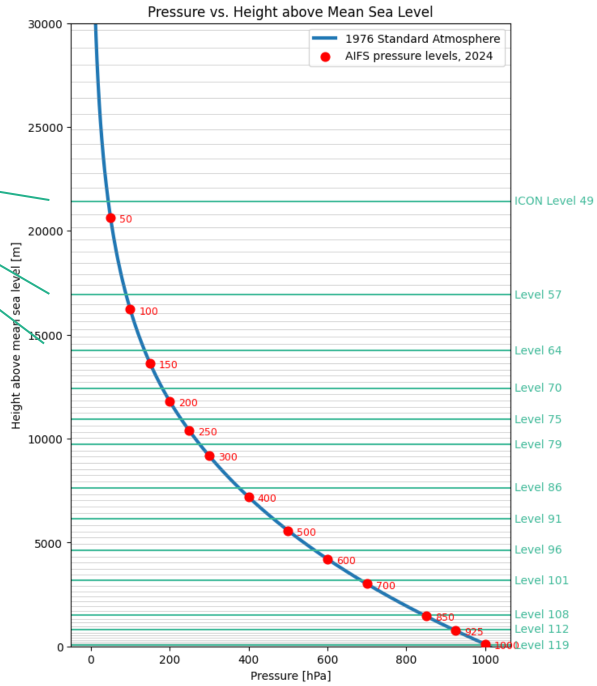
forcings:

```
HSURF, FR_LAND, Z0, FR_LAKE, EMIS_RAD, SSO_STDH, cos/sin_latitude/longitude,
cos/sin_julian_day, cos/sin_local_time, insolation
```

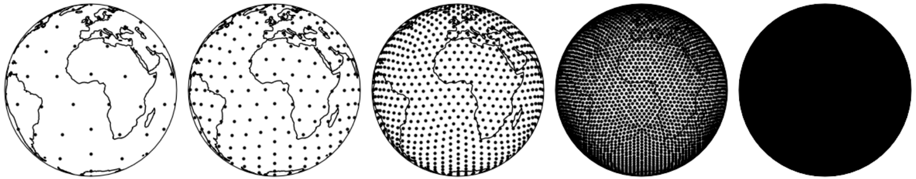
diagnostic:

```
TOT_PREC
```

13 ICON levels (top-down ordering):
49, 57, 64, 70, 75, 79, 86, 91, 96,
101, 108, 112, 119



ICON data locations are based on cell centers of a triangular grid.
Grid generation inherently defines hierarchical decompositions:

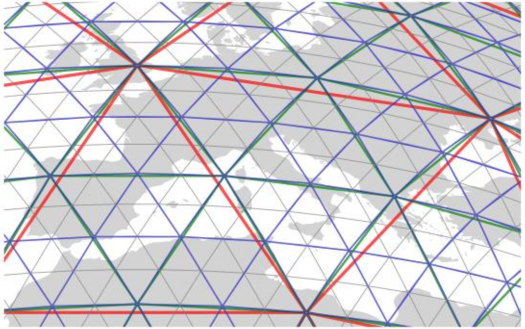


AICON Model Architecture (cont'd)

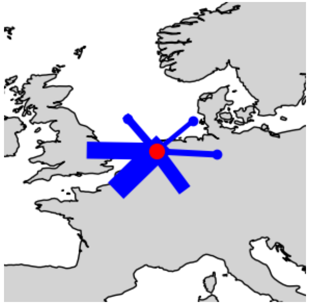
GraphCast-like encoder-processor-decoder architecture.

encoder:	maps onto a hidden state
processor:	contains latent information
decoder:	takes final hidden state

- Graph construction directly based on ICON's triangular meshes.
- Graph-Transformer GNN: Message passing is done via a multi-head attention mechanism. Each node's new features become a weighted average of its neighbors' features.



Processor: Different levels of refinement facilitate communication at different scales.



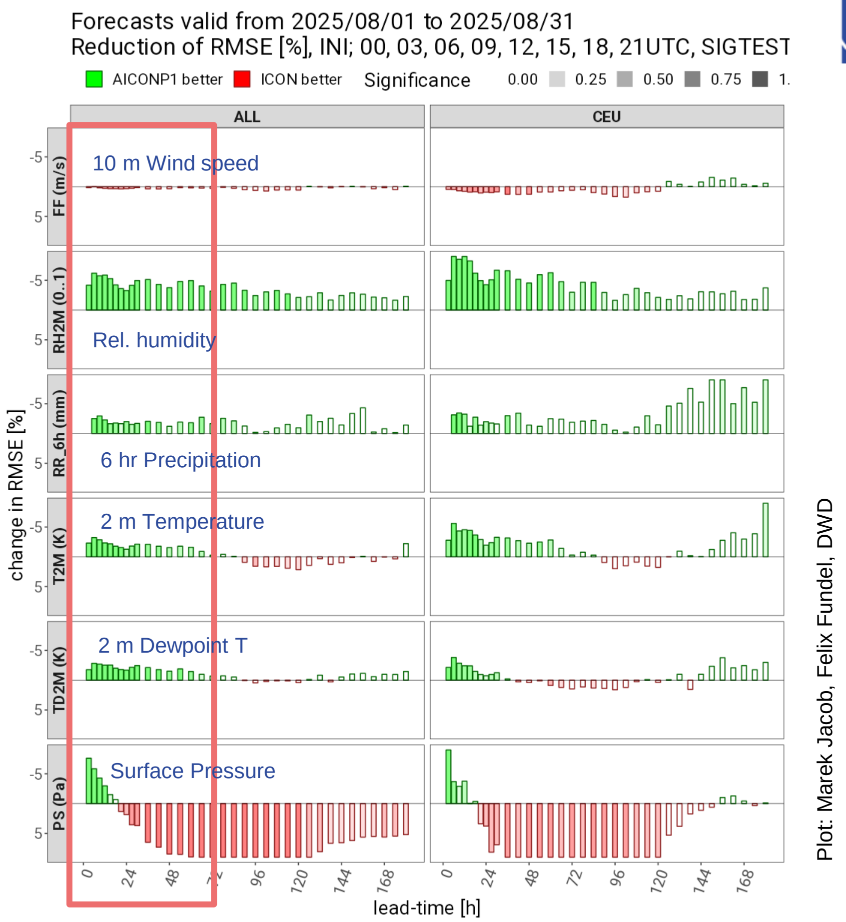
Ex.: Attention graph, relative importance of a neighbor to a target node.

Verification Results

- AICON at 13 km icosahedral grid

  training.start: 2010-01-01T00:00:00
  training.end: 2023-12-31T21:00:00
  frequency: 3h, training without rollout

- Verification against SYNOP observations
- AICON vs ICON. Color coding:
  - Green: AICON better
  - Red: ICON better
- Good near-surface forecast skill for short range (0 – 72 hrs)

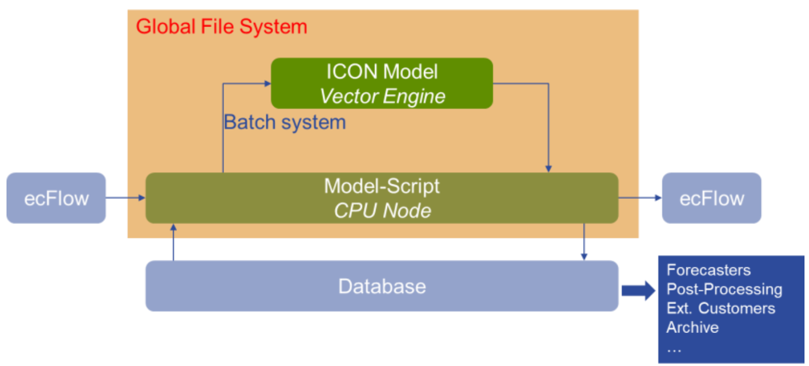


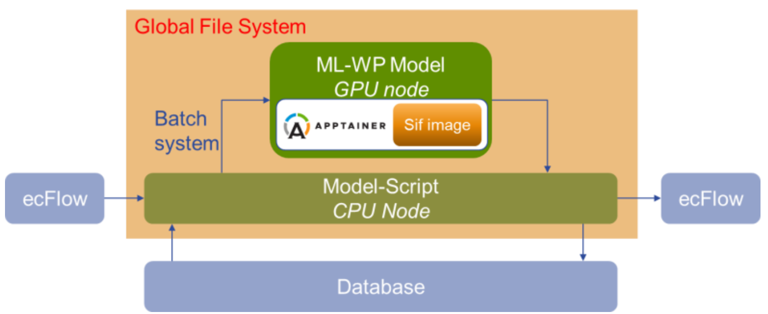
Anemoi version: anemoi-models==0.5.0
inference checkpoint 89887843679d47deac1b82be5ef7ffc3

Production Environment

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

AICON fits seamlessly into the existing process chain of 24/7 numerical weather forecasting.





(M. Jacob, DWD)

Containerized multi-GPU inference

- straight deployment into production environment: base layer for Pytorch, Eccodes
- incl. pre-processing (e.g. SMI), post-processing (lat-lon interpolation, clipping)

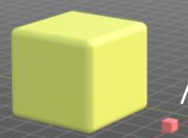
Energy Statistics

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

Duration/energy estimate of a deterministic ICON global run (13 km)

deterministic ICON global run
energy consumption: 60.24 kWh
~ 21.87 kg CO_2 emission[1]

AICON inference
energy consumption: 0.13 kWh
~ 48.28 g CO_2 emission



[1] CO_2 emissions based on German electricity mix (as of 2023/2024)

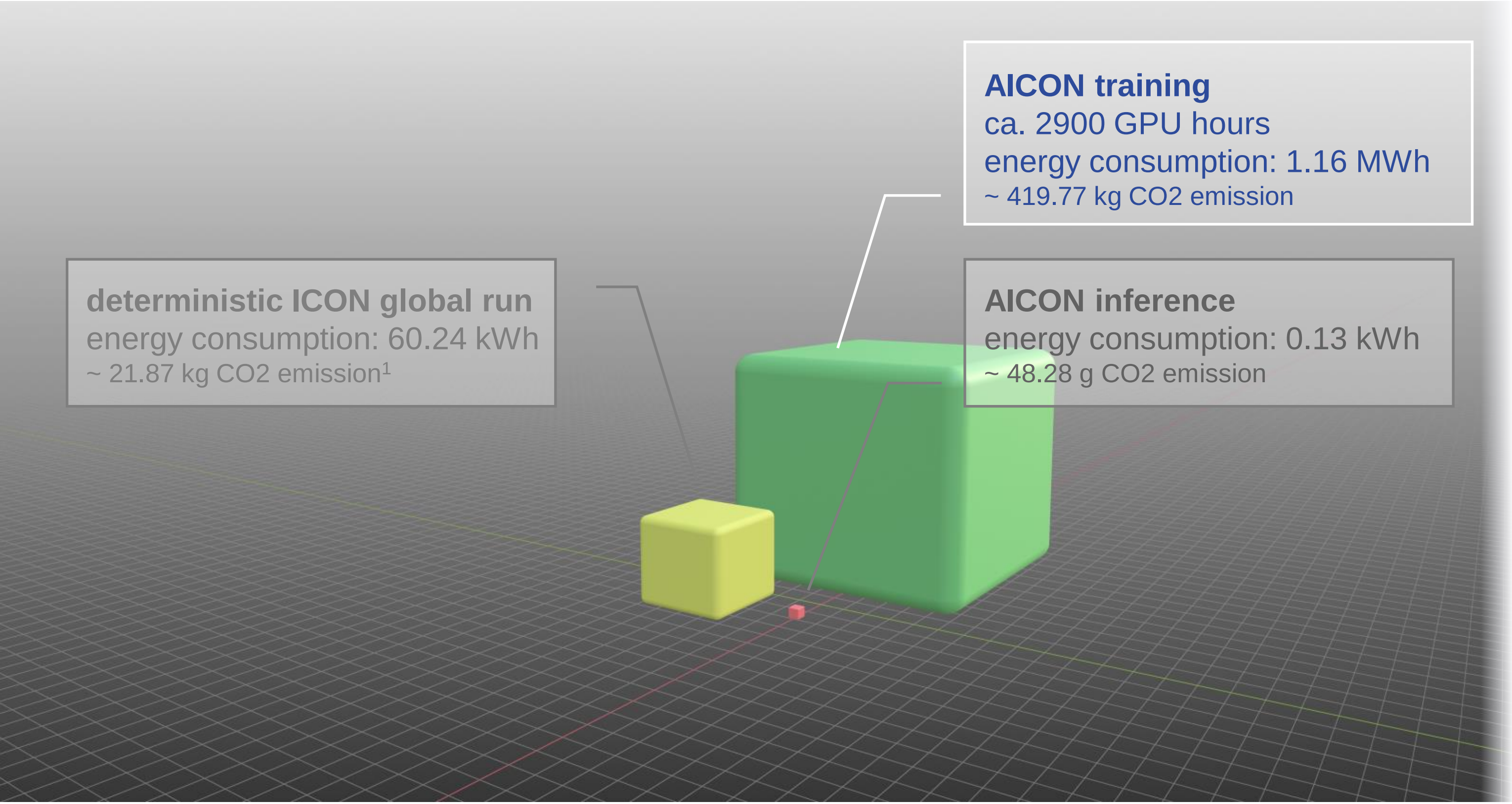
Details

180h lead time

GPU: A100-80, 400W

SX Aurora 1: avg. 1480W

Energy Statistics



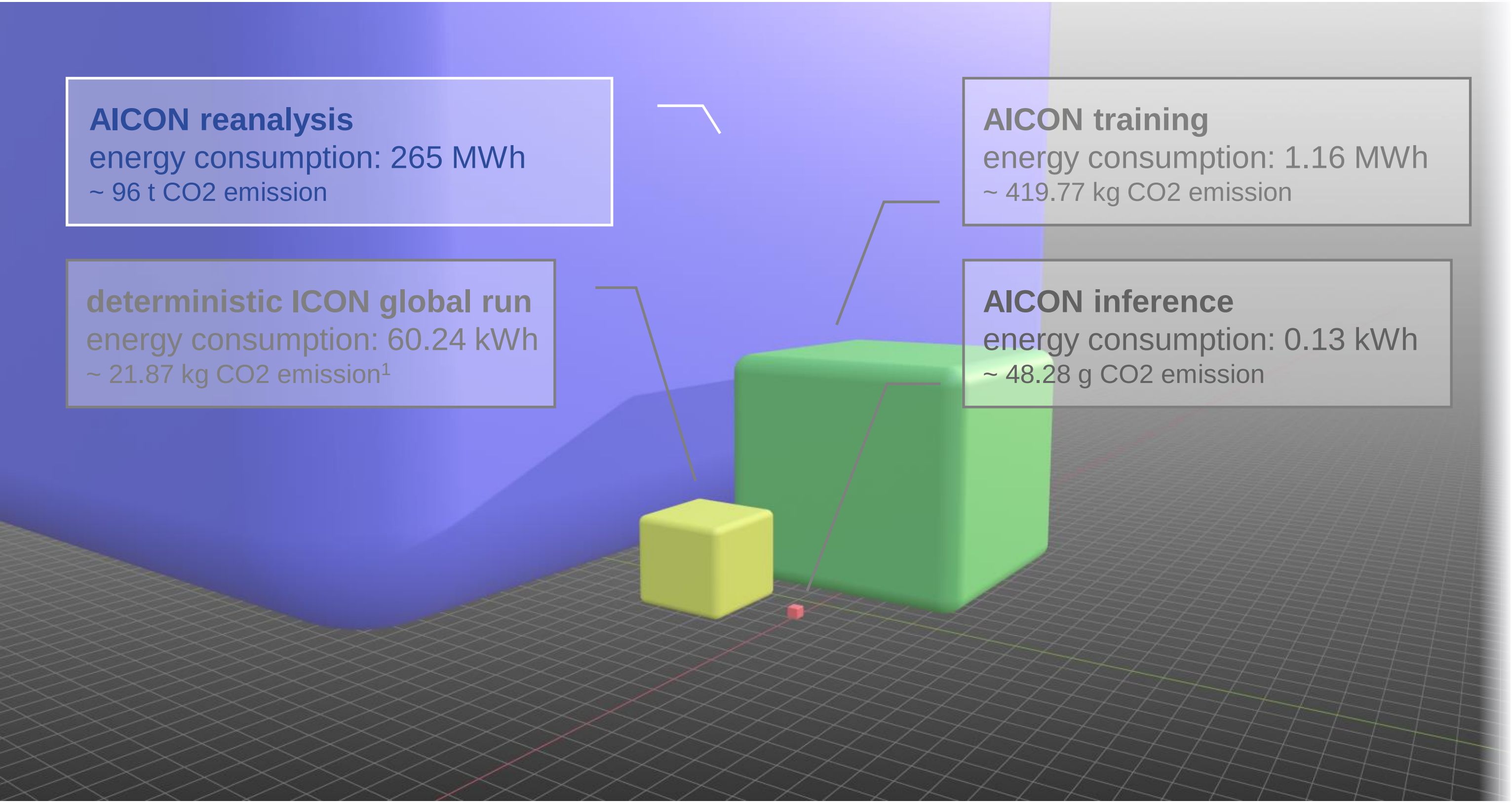
Model has a relatively low training cost.

<u>For comparison:
Llama 2-7B training</u>

184,320 GPU hours, hardware: A100-80G

(https://arxiv.org/abs/2307.09288)

Energy Statistics

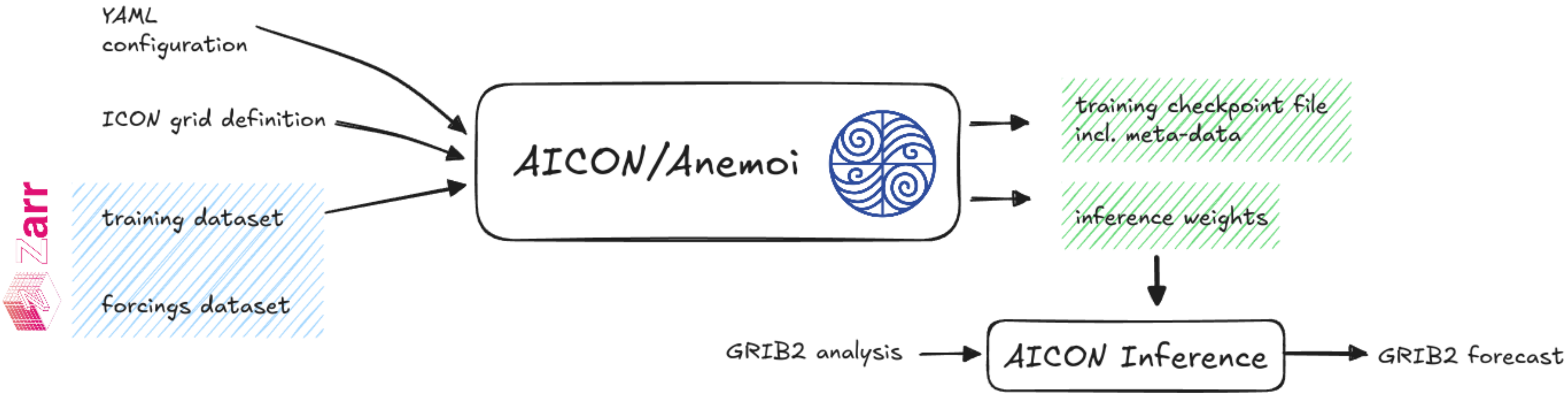


Details

Reanalysis
deterministic, global,
2010-01 until 2025-04

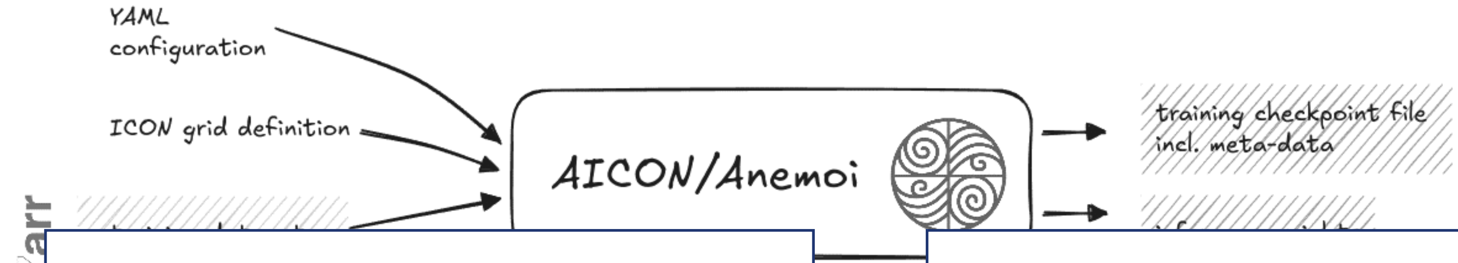
(currently on-going,
including back-extension)

Training Environment



Transfer learning between mesh resolutions: organization of a cost-efficient training schedule as long as higher resolution training graphs preserve the essential structural and statistical properties: 53 km → 26 km → 13 km

Training Environment

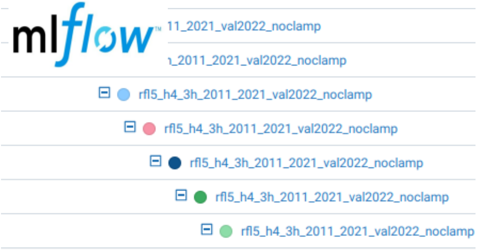


Training on HoreKa A100-40G

HoreKa supercomputer at KIT (Photo: S. Raffeiner, KIT/SCC)

Experiment tracking MLFlow

`mlflow.dwd.de`



“New HPC”

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

Training & inference: contrast to current (almost symmetric) hardware configuration.

- NEC SX-Aurora TSUBASA Vector Engine,
- Top500 list June 2025 theoretical peak performance
  position 113: 16.43 Rpeak (PFlop/s) /
  position 156: 12.22

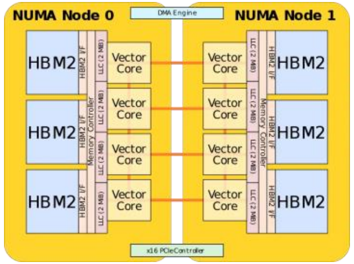
asymmetry ratio = 1.34

NEC SX-Aurora rack mount model

Is there a way to make use of existing hardware, originally targeted at grid point models?

Inference: Utilization of the existing vector HPC?

SX-Aurora: Vector processors on PCIe-based accelerator cards with high-bandwidth memory on-chip



https://en.wikichip.org/wiki/nec/microarchitectures/sx-aurora

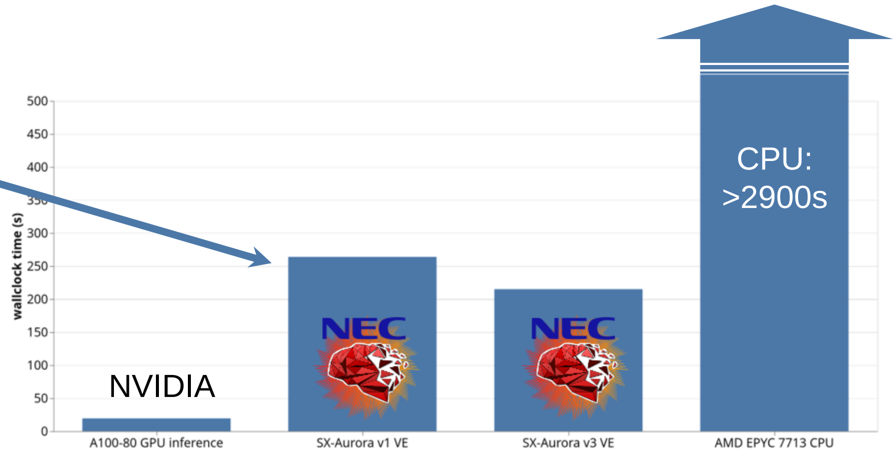
AI Compiler NEC SOL

Proof-of-concept: AICON runs on NEC Vector Engine (VE)!

- Only very minor adjustments to the inference code are necessary (essentially `import sol; sol.device.set('ve', 0)`).
- Currently: VE uses 16 cores, but only 1 VE.

But: GPU has a significant(10x) advantage!

- only needs to read 50% of the memory for the model parameters and can still use its TensorCores for `torch.float16`.
- GPU can make use of its atomic read-modify-write operations



Acknowledgments:
Many thanks to
Nicolas Weber @ NEC

Work in Progress

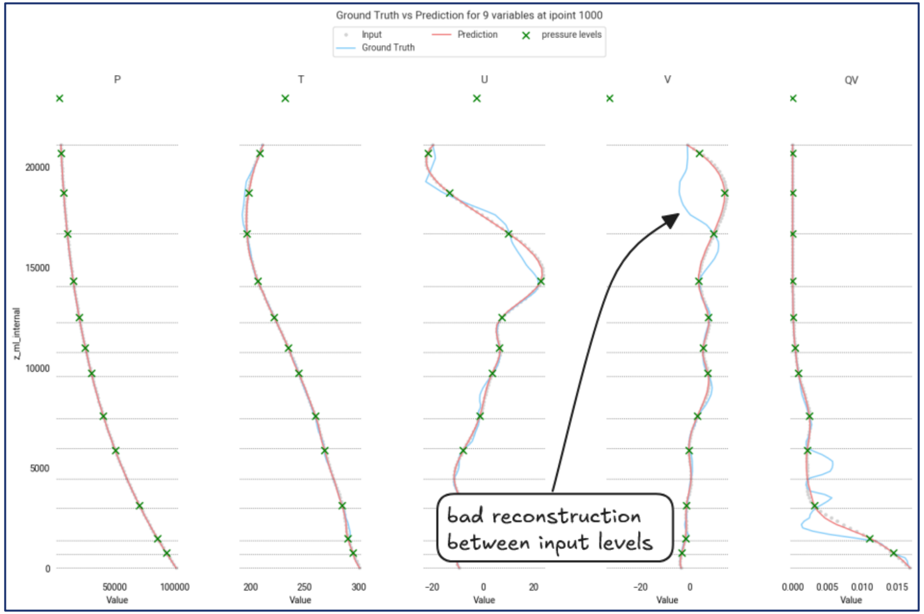
1D Vertical Super-Resolution

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

AICON predicts a comparably small number of vertical levels.

Column-wise offline technique:

- pre-upsampling super-resolution model: the low-resolution model output is first upscaled to the target high-resolution size using a traditional interpolation method.
- Convolutional neural network then learns to improve and refine the upsampled data.



Ex.: ICON data, ML-based reconstruction of vertical profile.

da Silva Rodrigues, J. D., & Morcrette, C. J. (2025). Improving vertical detail in simulated temperature and humidity data using machine learning. *Atmospheric Science Letters*, 26(2), e1288. https://doi.org/10.1002/asl.1288

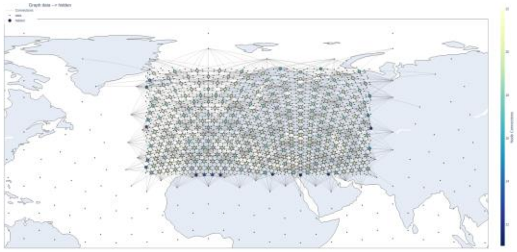
AICON Limited Area Model

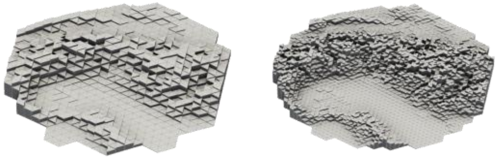
Work in progress (Sabrina Wahl et al., DWD)

Merge global and regional input datasets.

- Regional reanalysis ~6.5 km EU nest.
- Overlay the regional dataset t+0h with the global dataset t+3h.
- Global input at boundaries and inside the domain, encoder edges from the sets of global and regional data nodes.
- Processor: multi-mesh over the LAM region.

… different from “stretched-grid model approach” where global grid points that overlap with the regional model are removed (Nipen et al. 2024: https://arxiv.org/abs/2409.01891).





R03B07, 13 km grid R03B08, 6.5 km

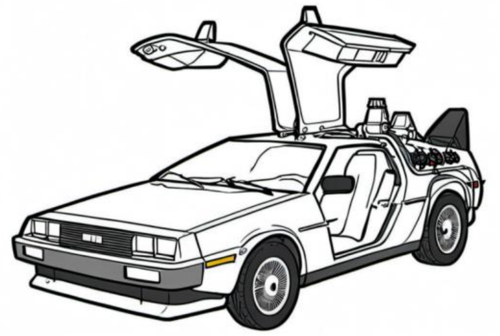
ICON, alpine region, ASTER orography

Preparing for Operational Rollout

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

Plans

- Q3 2025 : AICON (13 km global) technically operational for validation by the DWD forecast center ✓
- Q2 2026 AICON-LAM (6.5 km EU) in NWP operation (technically operational)
- Q1 2027 AICON-LAM (2 km DE) in NWP operation (technically operational)



- Pain Points and Challenges: physical inconsistencies, clipping needed; higher temporal resolution (1h), adaptability important, …

- AI-driven data assimilation (AIDA) – not covered here:

Keller, Jan & Potthast, Roland (2024). AI-based data assimilation: Learning the functional of analysis estimation. 10.48550/arXiv.2406.00390.

Florian Prill

Met. Analyse und Modellierung
Deutscher Wetterdienst

e-mail: Florian.Prill@dwd.de

The authors gratefully acknowledge the computing time provided on the high-performance computer HoreKa by the National High-Performance Computing Center at KIT (NHR@KIT). This center is jointly supported by the Federal Ministry of Education and Research and the Ministry of Science, Research and the Arts of Baden-Württemberg, as part of the National High-Performance Computing (NHR) joint funding program (https://www.nhr-verein.de/en/our-partners). HoreKa is partly funded by the German Research Foundation (DFG).