Diverse Aspects of Computing

at DWD

Ulrich Schättler, Marek Jacob, Florian Prill
Department for Numerical Modeling (FE1)
Deutscher Wetterdienst, Germany

With contributions from many more colleagues!

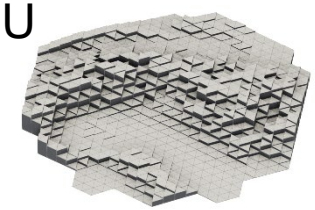
Contents

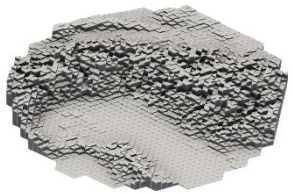
Latest Upgrades of DWD´s Operational Workload

During Phase 1 of SX-Aurora

Upgrades affecting HPC resources

- Sept. 2021: Operational Pollen Forecast System for Middle Europe and the Mediterranean
- Nov. 2022: Resolution upgrade in ICON (global) and ICON-EU



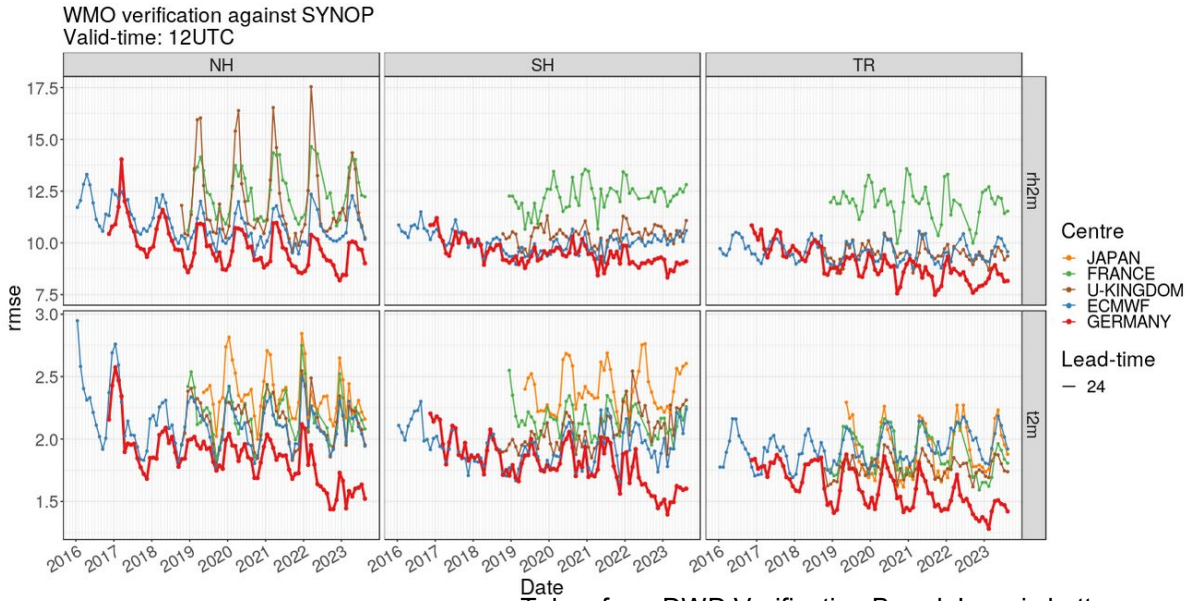


Several upgrades to improve model and data assimilation, e.g.

- extension of model-DA coupling for optimizing T_2M and RH_2M scores
- extension of 10m wind assimilation

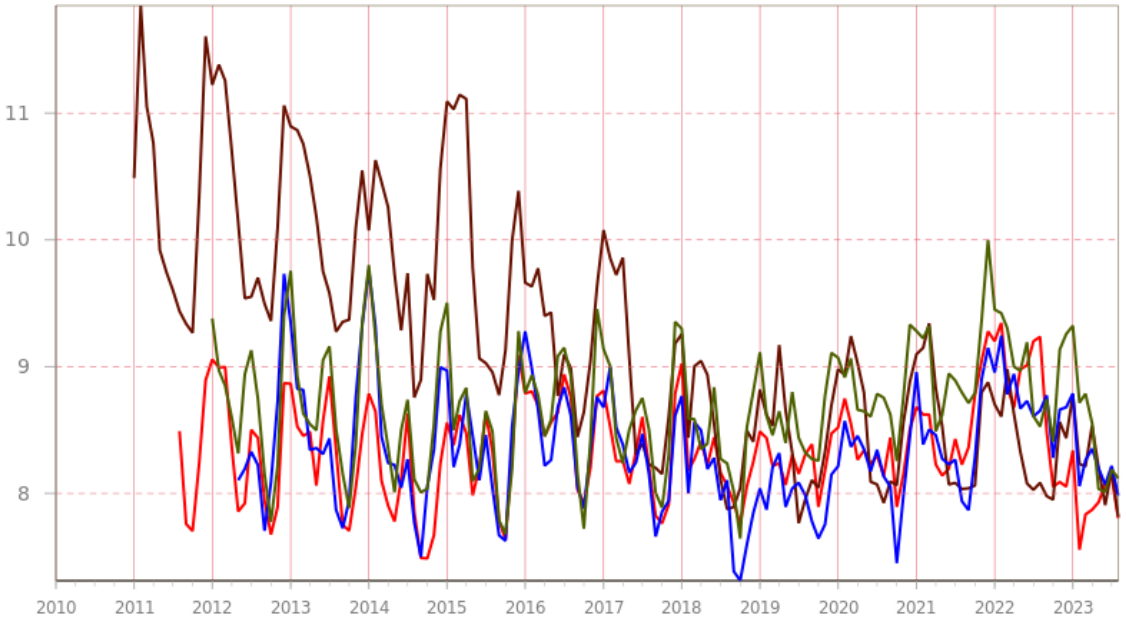
Forecast data available on https://opendata.dwd.de

		Horizontal (km)		Vertical (# levels)	
		Old	New	Old	New
Det.	ICON	13		90	120
	ICON-EU	6.5		60	74
EPS	ICON	40	26	90	120
	ICON-EU	20	13	60	74



Taken from DWD Verification Board. Less is better.

Also: Scores in the Atmosphere (... to tell the whole truth)

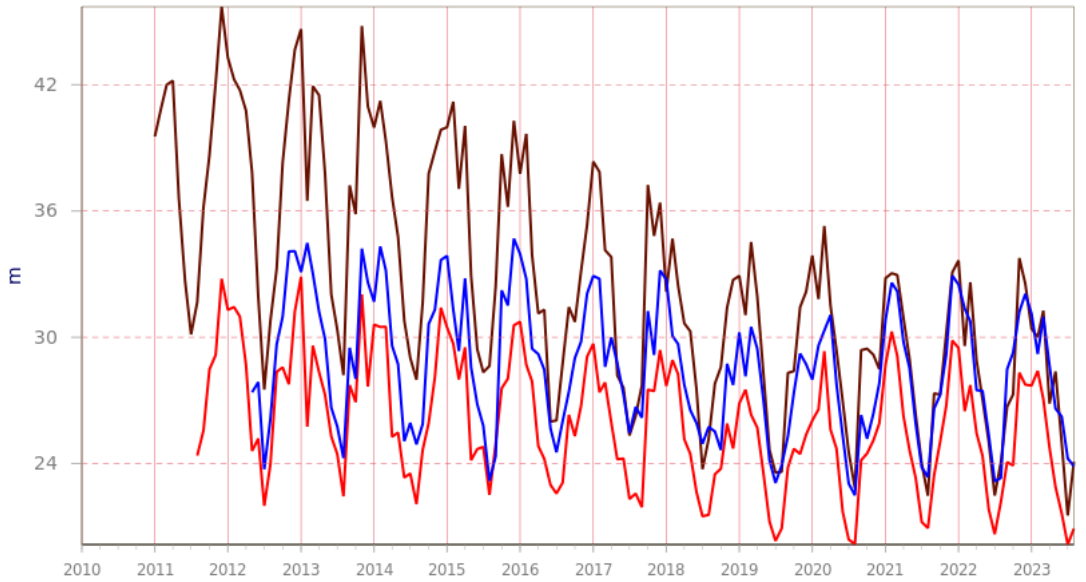


24 hour leadtime: short range

Geopotential on Northern Hemisphere, 850 hPa

- ➔ RMSE
- ➔ Verification against observations
- ➔ 12 UTC forecast

120 hour leadtime: medium range



ECMWF 12 DWD 12 MetOffice 12 Meteo-France 12

taken from https://wmolcdnv.ecmwf.int/scores/time_series/...

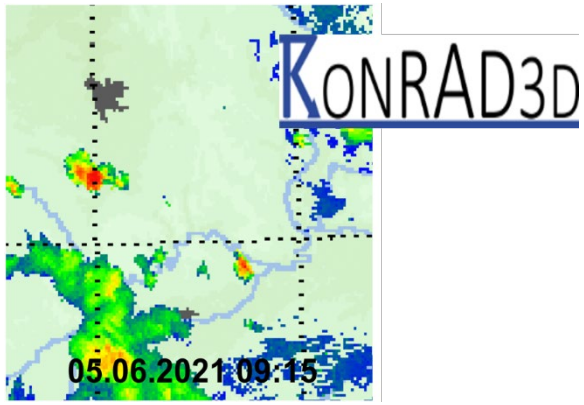
„Seamless INtegrated FOrecastiNg System: SINFONY

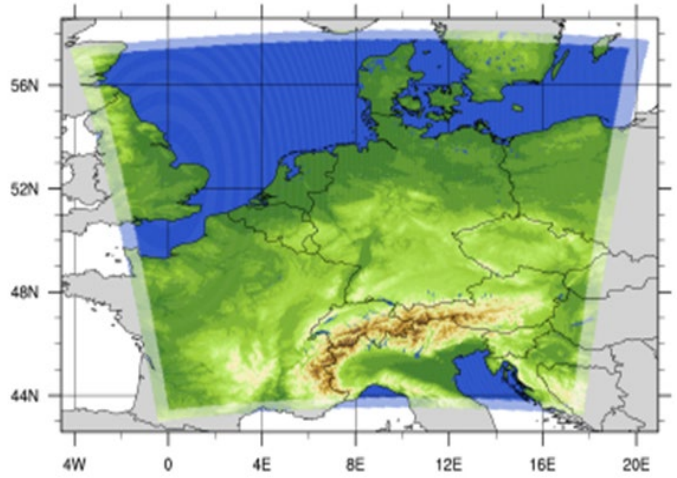
Goal: seamless prediction of the atmospheric state, in particular for the prediction of small-scale, severe weather events.

A new prediction system is developed in the project and implemented as a Rapid Update Cycle (RUC). It is rather similar to the existing ICON-D2 system, but:

- ➔ Hourly forecasts with a range up to +14 hours
- ➔ Data assimilation with KENDA-LETKF using additional data (cell objects from radar observations, lightnings).
- ➔ ICON-LAM EPS with 2 km resolution and adapted model physics (2 moment cloud microphysics scheme).
- ➔ Output every 5 minutes
- ➔ Products are created as a blend of nowcasting and very short-range forecasts.

(see: https://www.dwd.de/EN/research/researchprogramme/sinfony_iafe/sinfony_en_node.html)



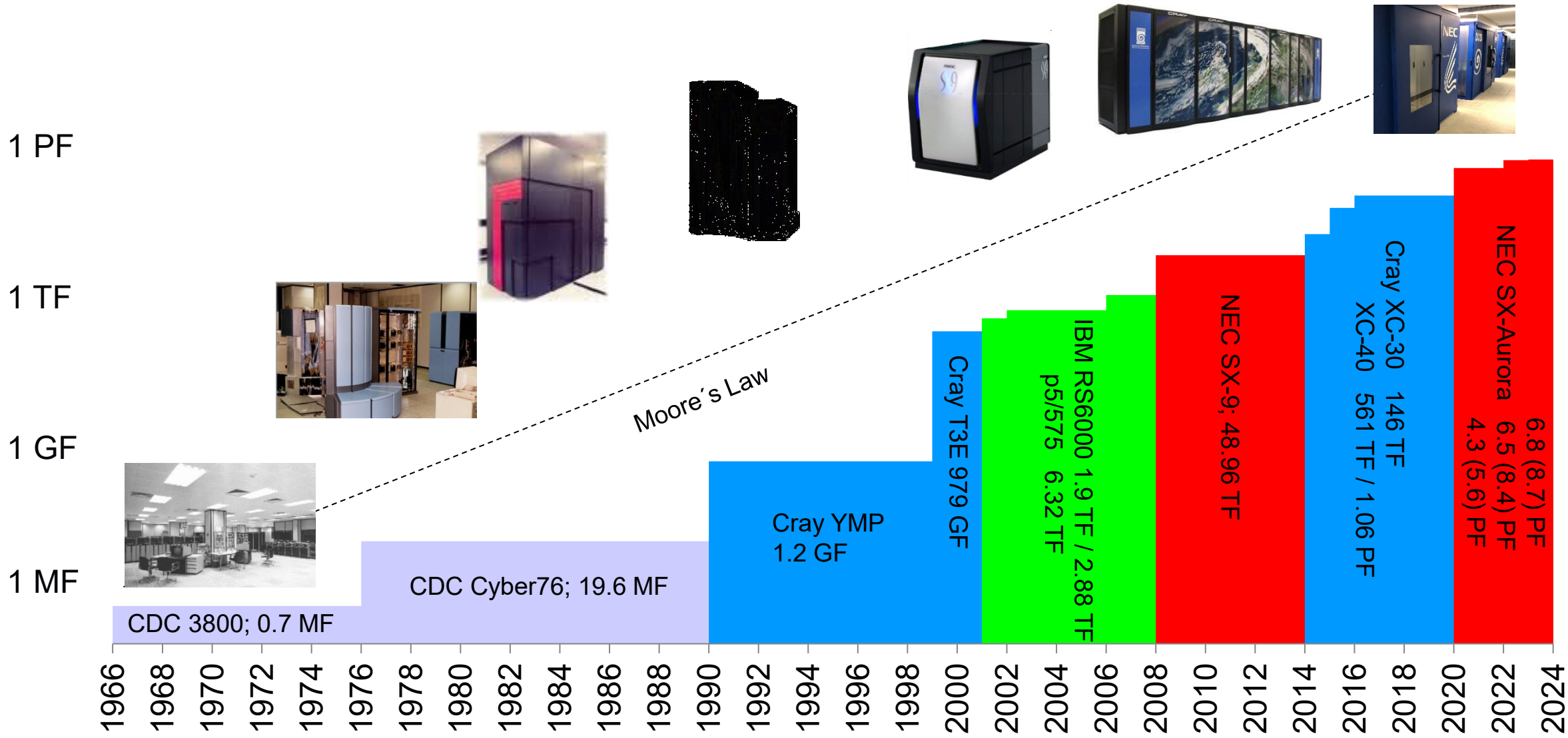


It is planned that part of the HPC will be dedicated to this system!

⇒ This needs Phase 2 of our HPC system!

Increasing Diversity (Heterogeneity) of the HPC System

DWD Computers 1966-2024



Phase 2: Upgrade of NEC SX-Aurora Tsubasa

Originally the upgrade was planned for end of 2022, but has been split in two steps:

- ➔ 2a) Upgrade with existing Aurora 10AE nodes in autumn 2022, but did not reach desired end performance.
- ➔ 2b) Upgrade with new Aurora 30B nodes: delayed because of well-known reasons.

Phase	Operational				Research				Available
	VH	VE	Cores	Size EPS	VH	VE	Cores	Size EPS	
0	178	1424	11392	49	232	1856	14848	64	12/2019
1	224	1792	14336	161	292	2336	18688	210	12/2020
2a	339	2712	21696	243	440	3520	28160	316	09/2022
2b	391	3128	28352	324	508	4064	36864	421	09/2023

Size EPS:
ensemble members to check performance upgrade factor.
Committed factors are:

Phase 0: 1
Phase 1: 3
Phase 2: 6

- ➔ For Phase 2b Aurora 30B nodes were added:

	52	416	6656	80	68	544	8704	105	09/2023

Differences Aurora 10 / 30

The building block of the NEC vector system is a node, which consists of

- ➔ a scalar vector host (VH): 24-core AMD Rome; 2.8 GHz; 256 GB memory)
- ➔ 8 vector engines (VE): SX-Aurora TSUBASA, which are of type

Vector Engine	Aurora 10AE	Aurora 30B
Frequency (GHz)	1.584	1.600
# cores	8	16
Theoretical peak / core DP (GF/s)	304	307
Theoretical peak / core SP (GF/s)	608	614
Memory Bandwidth (TB/s)	1.35	2.45
Cache Capacity (MB)	16	64
Memory Capacity (GB)	48	96

Some Timings

1 day global forecast (R02B07N08; 20 km global resolution with a nest over Europe with 10 km resolution)

#nodes / cores	Batch timings	Compu tations	Model init
4 / 512	978	935.0	32.7
8 / 1024	496	450.2	36.1
16 / 2048	269	221.2	35.2

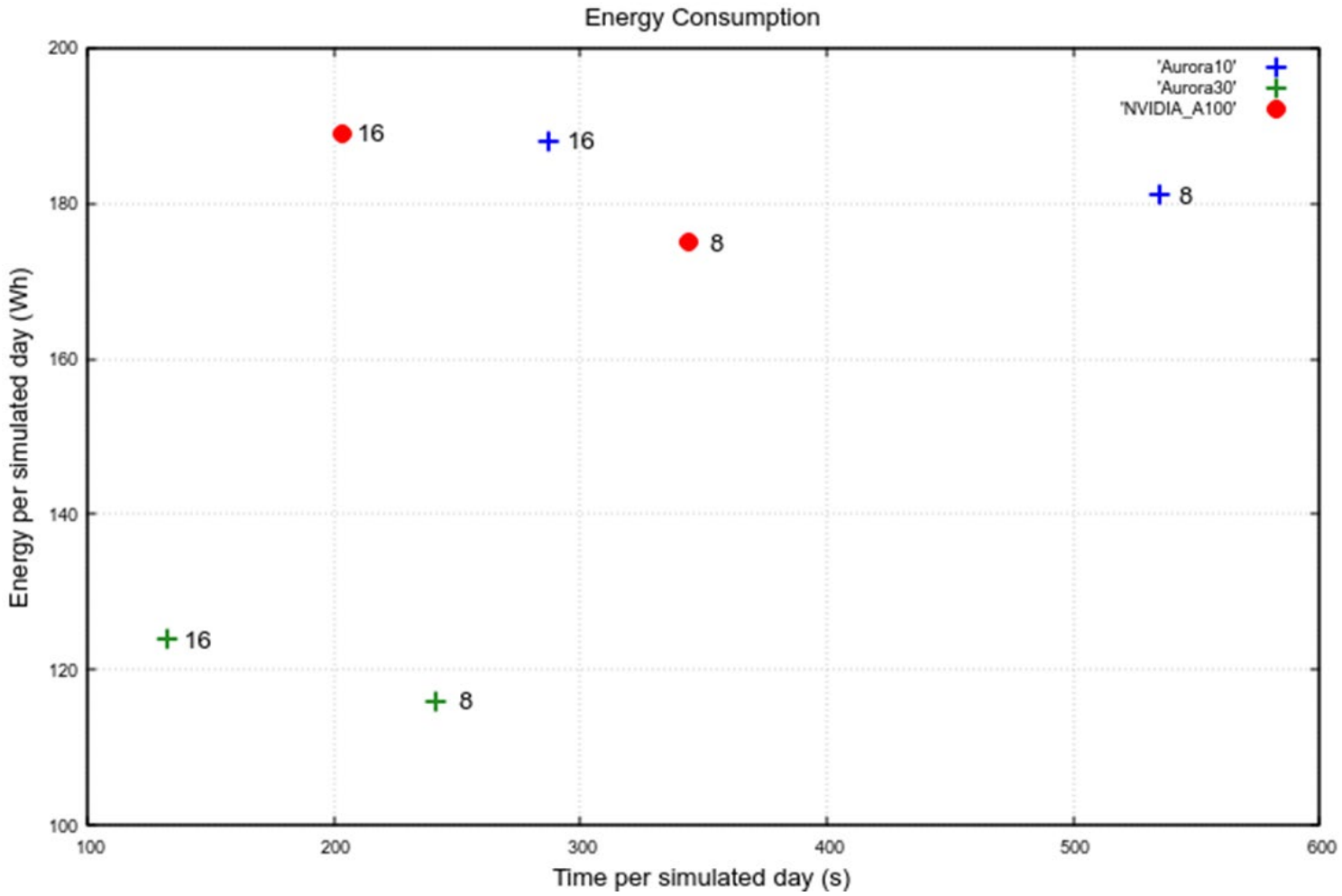
ATOS@ECMWF (AMD EPYC Rome)

#nodes / VEs / cores	Batch timings	Compu tations	Model init
1 / 8 / 64	591.9	546.1	45.3
2 / 16 / 128	333.7	293.2	39.7
1 / 8 / 128	282.6	244.0	37.1
2 / 16 / 256	170.2	132.7	35.8

SX-Aurora Tsubasa @DWD (Aurora10/30)

Every run uses 1 scalar core
from the VH for input.

Time and Energy Measurements



Same experiment as last slide (R02B07N08)

- ➔ 8 / 16 Vector Engines or GPUs, resp.
- ➔ which means 1 or 2 nodes.

- ➔ VE10 and A100 about equal in energy required, but A100 finishes in about 27-35% less time.
- ➔ VE30 saves 35% of energy and finishes faster.

Green 500 Ranking

Phase 1 (from Green500, June 2021)

Rank in Green 500	Rank in TOP 500	#cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)	Efficiency (GFlops/W)
49	105	18,688	3.87	5.61	648	5.972
50	124	14,336	3.25	4.28	565	5.752

Phase 2a: Upgrade with existing Aurora 10 nodes in autumn 2022 (from Green500, June 2023)

Rank in Green 500	Rank in TOP 500	#cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)	Efficiency (GFlops/W)
76	99	28,160	6.43	8.41	954	6.746
79	124	21,696	5.05	6.54	834	6.055

Phase 2b: Watch out for the November 2023 list!

New GPU Partition

2 GPU nodes have been added to the HPC in 2021:

- 2x AMD EPYC Milan 7713, 64 cores
- 8x A100 NVIDIA SM4 80 GB GPUs
- 8x Mellanox InfiniBand HDR100 HCA

[1]

Used for Project „Met4Airports“

- Forecasting operation disruptions at airports due to meteorological conditions using AI
- Combination of weather and airport operations data using AI
- Try to forecast delays and capacity values

Are also used for developing and testing ICON GPU version.

[1] Winter at Munich Airport https://www.munich-airport.de/winterdienst-am-airport-3258037

What Is So Diverse About That System?

One MPI job can run accross the whole system (requiring a proper system configuration).

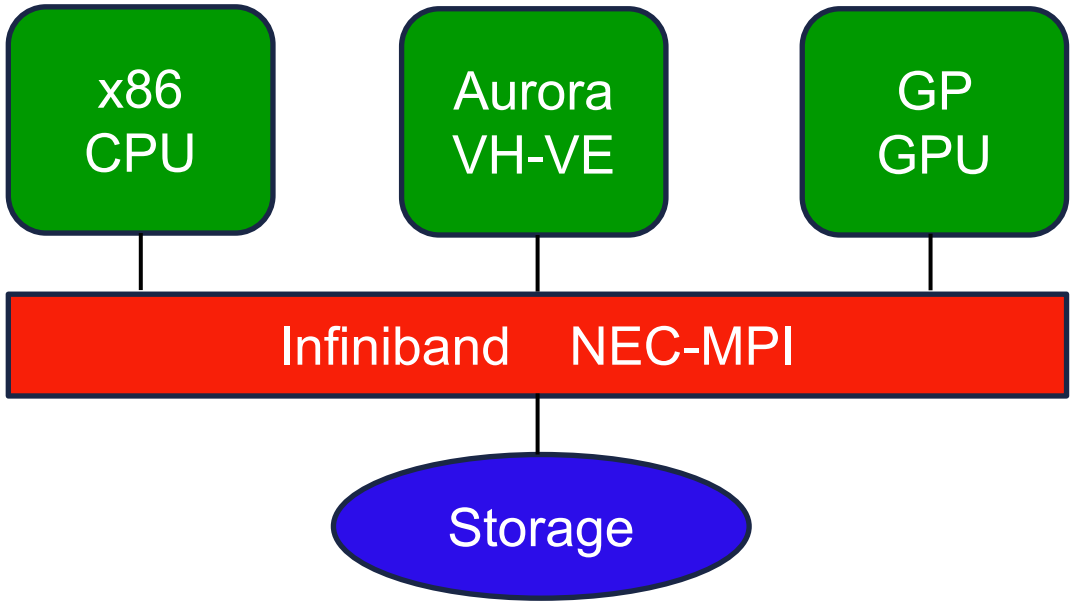
- ➔ Several binaries can be used (but all have to be linked with NEC-MPI).
- ➔ Using task parallelism parts of a job can run on different hardware.

On the Aurora you can run different binaries on the host and the engines.

When not using task parallelism it is possible to offload tasks from the VH to the VE and vice versa (using NEC proprietary tools).

DWD Jobs put I/O tasks to the scalar vector host cores:

```
mpirun -v -vh -node 0    -np 1  <scalar binary> : \
      -venode -node 0-1 -np 16 <vector binary>
```



How can you split your jobs to exploit such a system?

Outlook: Phase 3

The contract between DWD and NEC terminates end of 2024.

- ➔ There is an option to extend it for two more years.
- ➔ Thanks to the energy efficiency of the Aurora technology, DWD plans to negotiate a further performance upgrade within the existing power budget. A performance increase of up to 40% is targeted.

A replacement of the system then is planned for end of 2026.

- ➔ The procurement for the new system might already be started in 2025.

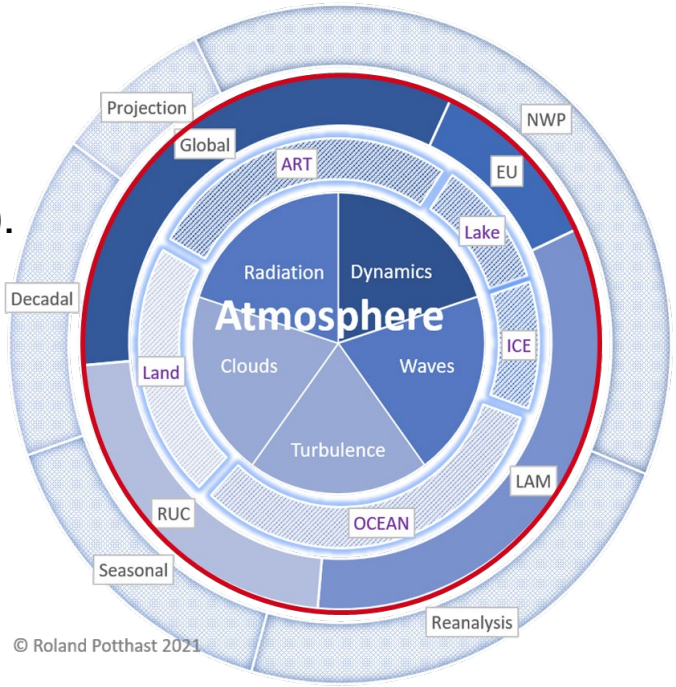
ICON

ICON - A Community Model

Core (development) partners are:

➔ DWD, MPI-M, DKRZ, KIT, ETH and MCH.

There are many more partners using ICON (some also contributing to development).



Available components:

➔ NWP, climate, ocean, land, ART.

➔ Runs in global and limited-area mode.

Possible Platforms:

➔ Standard x86 CPUs: full model.

➔ Vector processors: in principle full model, perhaps not all parts fully optimized for vectors.

➔ GPGPUs: Selected configurations of the model have been ported to GPUs using OpenACC.

From early 2024 on an Open Source License to use ICON will be available!

ICON - A Diverse Community

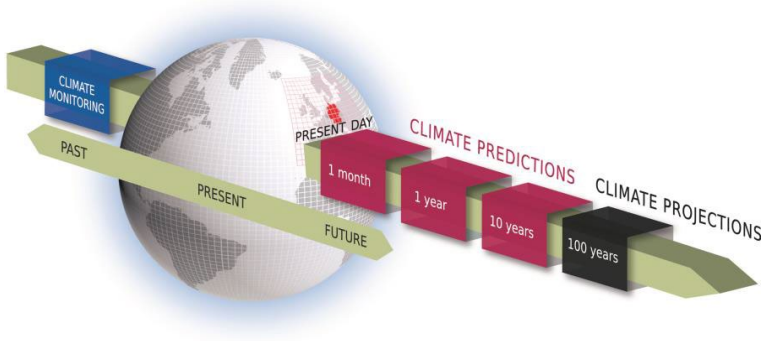
DWD	NWP	Daily operational production
MPI-M	Climate	Research
DKRZ	Technical infrastructure	Research
ART	Aerosols and reactive trace gases	Research
ETH	Climate	Research
MeteoSwiss	NWP	Daily operational production

- Partners taking part in research projects are strongly interested in Open Source Release.
- MPI-M cannot take care anymore for „quasi-operational“ climate predictions (e.g. contributions to IPCC). DWD will take over soon.
  - ICON Seamless
- Climate partners are interested in exascale computing:
  - WarmWorld
  - EXCLAIM
- Several partners are interested to plug in own code / components:
  - ComIn
- And there is interest in a cloud solution:
  - ICONIC

ICON Seamless

Uniform model and data assimilation for:

- ➔ Numerical Weather Prediction (NWP)
- ➔ Climate Prediction (seasonal, decadal)
- ➔ Climate Projections (global and regional)



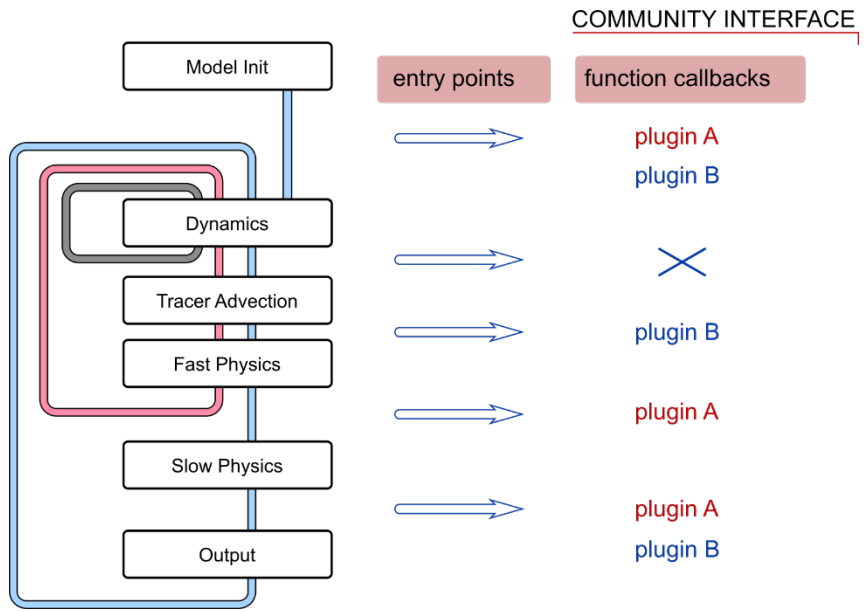
In collaboration with DWD, MPI-M, KIT, DKRZ, MPI-BGC and many more.

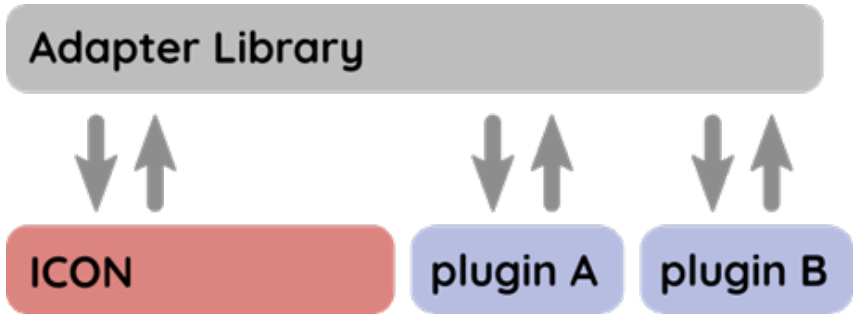
Timeframe: 2025+

Summary of developments: “Combine the best out of original NWP- and climate physics packages”.

What are the aims of ComIn?

- ➡ Providing a standardized **public interface** for third party codes (‘**plugins**’) coupled to ICON
- ➡ Significantly **reduced maintenance** for ICON as well as for third party code developers
- ➡ Plugins **easier to migrate** to new ICON releases
- ➡ Establishing ICON as the core model for applications ranging from **NWP** to **ESM**
- ➡ Enables **multi-language support** (Fortran, C/C++, Python)





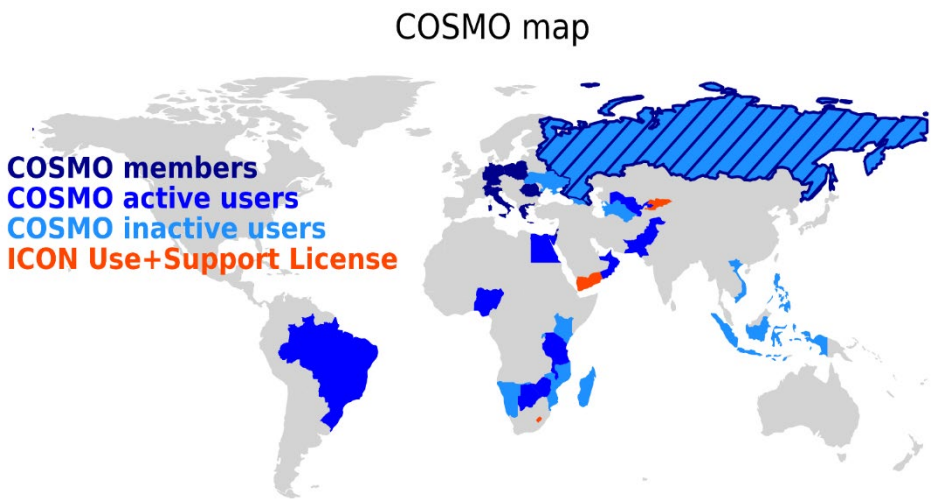
How does ComIn work in a nutshell?

- ➡ ComIn organizes the **data exchange** and **simulation events** between the ICON model and multiple plugins.
- ➡ **ComIn Callback Register**: Subroutines of the plugins are called at pre-defined events during the ICON simulation.
- ➡ The **ComIn Adapter Library** is included by ICON and the plugins. It contains descriptive data structures and regulates the access and the creation of model variables.

ICON - More Activities

Migrating COSMO Users to ICON-LAM

- Most of our partner-weather services still use the COSMO-Model (support available until 2025).
- They should migrate to ICON-LAM in the next two years.
- Currently we are preparing this migration process (new license contracts, support, trainings).



ICONIC - ICON in the Cloud

- A pilot project financed by the World Bank in 2021/22.
- Demonstrating that a regional numerical weather forecast can be run over the internet („in the cloud“). Using a pre-operational setup for central asia.
- And using the abstraction „cloud“ to handle the diversity of hardware at least to some extent.
- Due to the need of several COSMO partners we will continue with this work using COSMO funding.

Summary

Summary

- Improvement of ICON and the data assimilation is going on.
- To run the operational workload, DWD runs a powerful HPC using vector processors.
- The new processors Aurora 30 are even more powerful and energy-efficient.
- Together with our partners ICON is further developed to suite ~~all~~ most needs.

It is difficult
to predict...

...especially
the future.

Thank you very much
for your attention!