# DownscaleBench: A benchmark dataset for statistical downscaling of meteorological fields

## Workshop on Large-Scale Deep Learning for the Earth System | Bonn

2024-08-29 I **MICHAEL LANGGUTH[1]**, SEBASTIAN LEHNER[2], PAULA HARDER[3], ANKIT PATNALA[1], IRENE SCHICKER[2], MARKUS DABERNIG[2], MARTIN G. SCHULTZ[1]

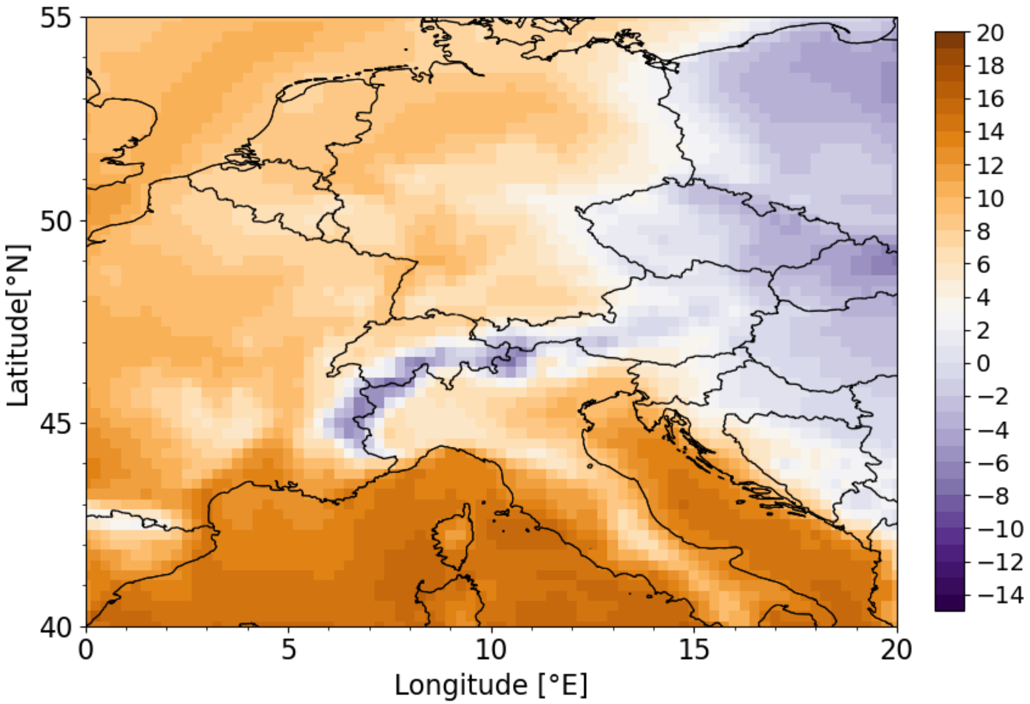
[1] Juelich Supercomputing Centre (JSC), [2] Geosphere Austria, [3] Mila Quebec AI Institute

"The MAELSTROM project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 955513. The JU receives support from the European Union's Horizon 2020 research and innovation programme and United Kingdom, Germany, Italy, Luxembourg, Switzerland, Norway".

# MOTIVATION

## Why do we need a benchmark dataset?

- High-resolved weather data is crucially demanded
- Recent success of deep learning for statistical downscaling (e.g. *Mardani et al., 2023*, *Harder et al., 2023*, *Harris et al., 2022*)
- Intercomparison difficult:
  - Variety of downscaling tasks
  - Different datasets
  - Different evaluation methods
- Benchmark datasets steer progress in AI, e.g. ImageNet (*Deng et al., 2009*), GLUE (*Wang et al., 2018*)
- For meteorological applications, benchmarks such as WeatherBench 2 (*Rasp et al., 2023*) are rare

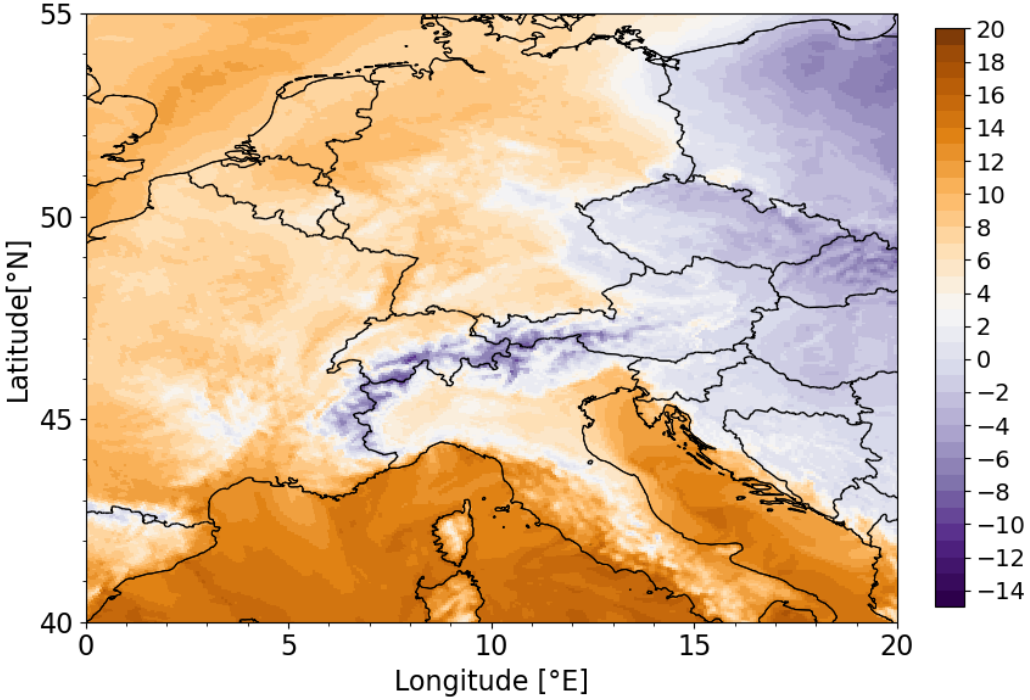



2m temperature field from the ERA5-reanalysis ($\Delta x = 0.25°$) on 2018-12-01,13:00 CET.

# MOTIVATION

## Why do we need a benchmark dataset?

- High-resolved weather data is crucially demanded
- Recent success of deep learning for statistical downscaling (e.g. *Mardani et al., 2023*, *Harder et al., 2023*, *Harris et al., 2022*)
- Intercomparison difficult:
  - Variety of downscaling tasks
  - Different datasets
  - Different evaluation methods
- Benchmark datasets steer progress in AI, e.g. ImageNet (*Deng et al., 2009*), GLUE (*Wang et al., 2018*)
- For meteorological applications, benchmarks such as WeatherBench 2 (*Rasp et al., 2023*) are rare



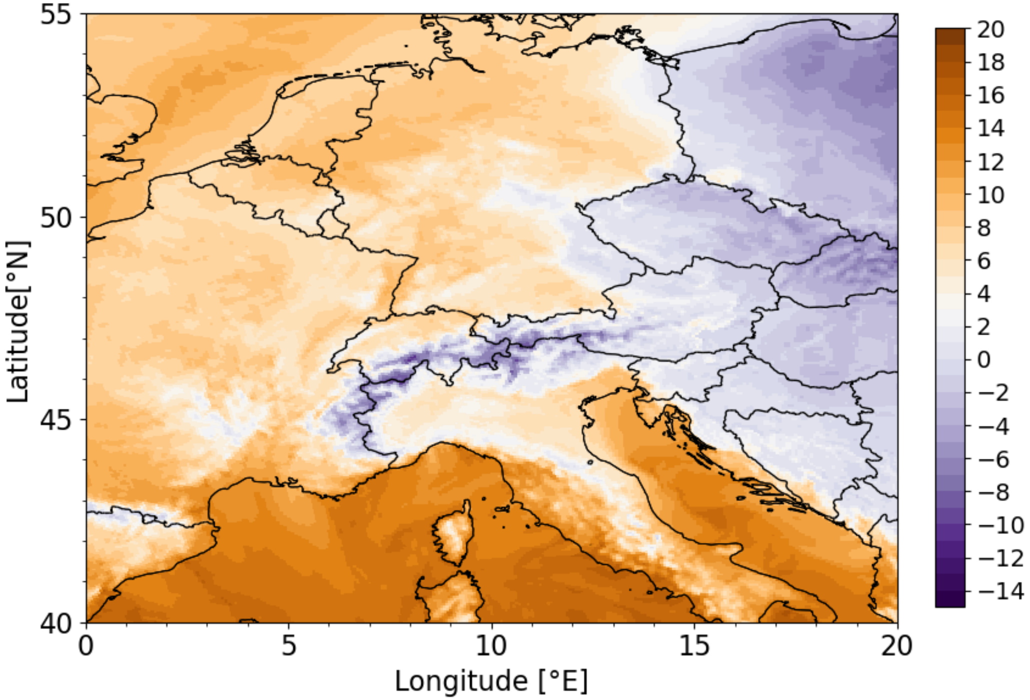

2m temperature field from COSMO-REA6 ($\Delta x = 0.055°$) on 2018-12-01,13:00 CET.

# MOTIVATION

## Why do we need a benchmark dataset?

- High-resolved weather data is crucially demanded
- Recent success of deep learning for statistical downscaling (e.g. *Mardani et al., 2023*, *Harder et al., 2023*, *Harris et al., 2022*)
- Intercomparison difficult:
  - Variety of downscaling t
  - Different dataset
  - Different eval
- Benchmark datas s in AI, e.g. ImageNet (*Deng et* ), GLUE (*Wang et al., 2018*)
- For meteorological a lications, benchmarks such as WeatherBench 2 (*Rasp et al., 2023*) are rare

**A benchmark dataset to steer progress in statistical downscaling**

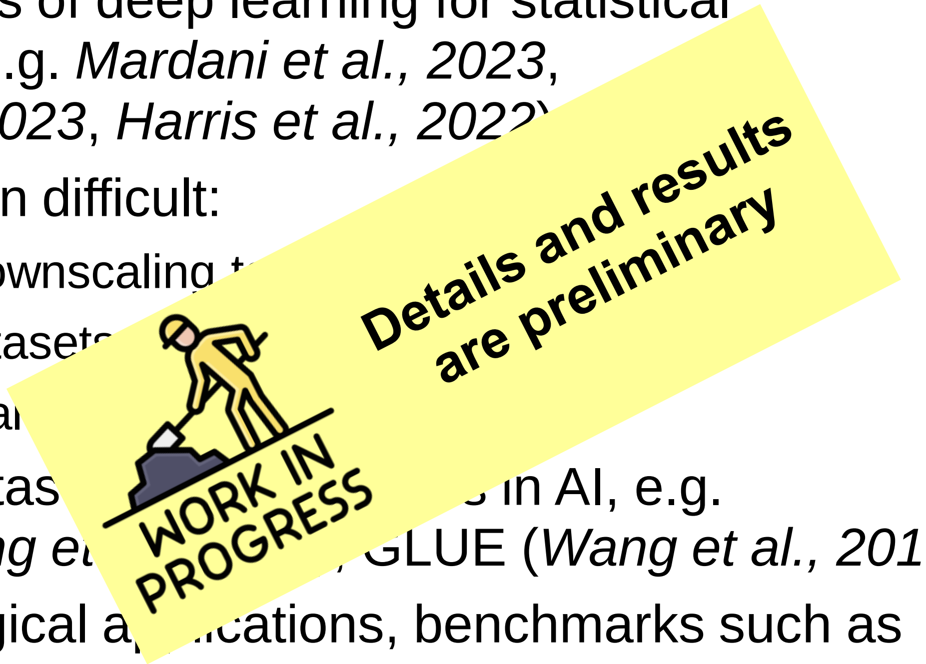


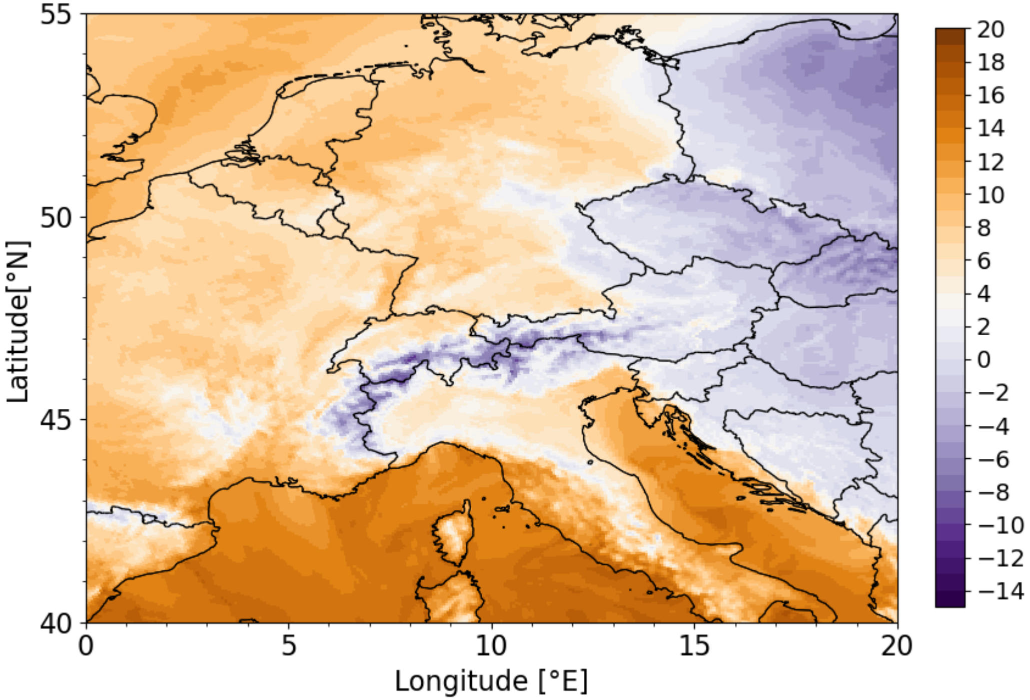

2m temperature field from COSMO-REA6 ($\Delta x = 0.055°$) on 2018-12-01,13:00 CET.

# MOTIVATION

## Why do we need a benchmark dataset?

- High-resolved weather data is crucially demanded
- Recent success of deep learning for statistical downscaling (e.g. *Mardani et al., 2023*, *Harder et al., 2023*, *Harris et al., 2022*)
- Intercomparison difficult:
  - Variety of downscaling t
  - Different datasets
  - Different eval
- Benchmark datas … s in AI, e.g. ImageNet (*Deng et* … , GLUE (*Wang et al., 2018*)
- For meteorological a … ications, benchmarks such as WeatherBench 2 (*Rasp et al., 2023*) are rare




2m temperature field from COSMO-REA6 ($\Delta x = 0.055°$) on 2018-12-01,13:00 CET.

# THE BENCHMARK DATASET

## Design choices and components of the benchmark dataset

- Benchmark dataset closely follows requirements listed in *Dueben et al., 2022*
  1) Clear problem statement for real-life task
  2) Open data provision in high-level programming language
  3) Results and code of baseline competitor models
  4) Evaluation metrics defined
  5) Visualization and diagnostics in code

# THE BENCHMARK DATASET

## Design choices and components of the benchmark dataset

- Benchmark dataset closely follows requirements listed in *Dueben et al., 2022*
  1) Clear problem statement for real-life task
  2) Open data provision in high-level programming language
  3) Results and code of baseline competitor models
  4) Evaluation metrics defined
  5) Visualization and diagnostics in code

**1) Downscaling task(-s):**

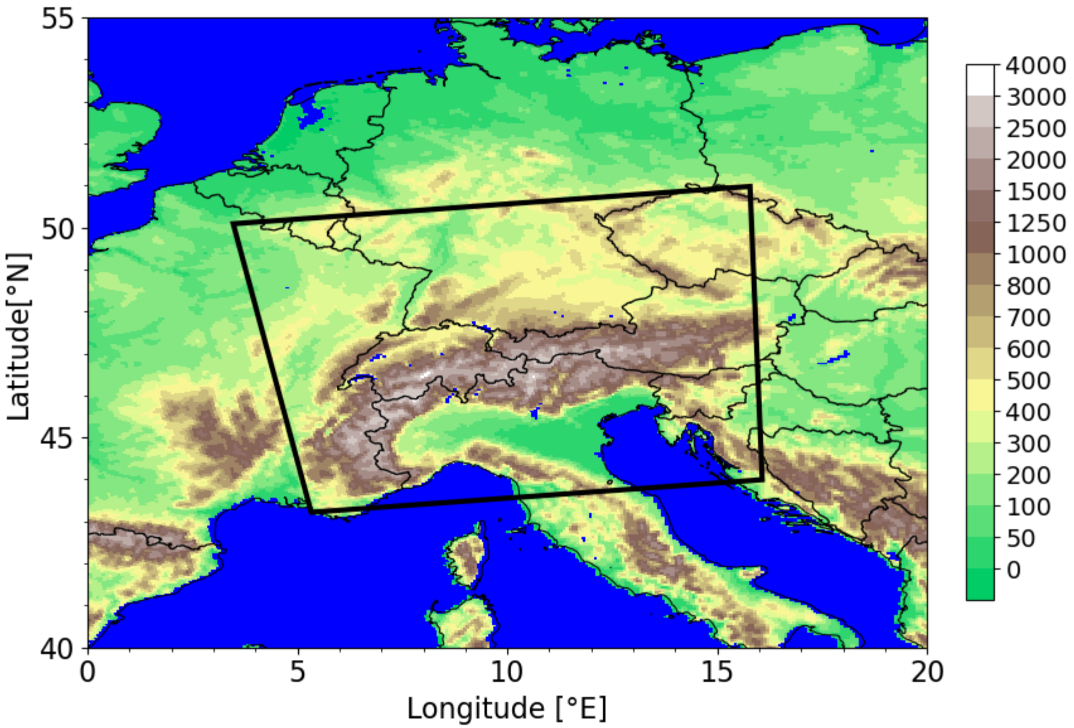
Emulate a high-resolved reanalysis (COSMO REA6) of i) ***2m temperature***, ii) ***100m wind*** and iii) ***global horizontal irradiance*** from the ERA5-reanalysis

# THE BENCHMARK DATASET

## Design choices and components of the benchmark dataset

**2) The data:**

- Predictands from COSMO-REA6 data ($\Delta x^{rot}_{CREA6} = 0.055°$)
- Task-specific set of predictor variables from the ERA5-reanalysis dataset ($\Delta x_{ERA5} = 0.25°$)
- Ready-to-use dataset provided via climetlab-plugin
- Exemplary data pipelines for TensorFlow and PyTorch in code base




Surface topography form COSMO REA6. The target domain of the downscaling benchmark comprises 144x128 grid points and is rendered.

# THE BENCHMARK DATASET

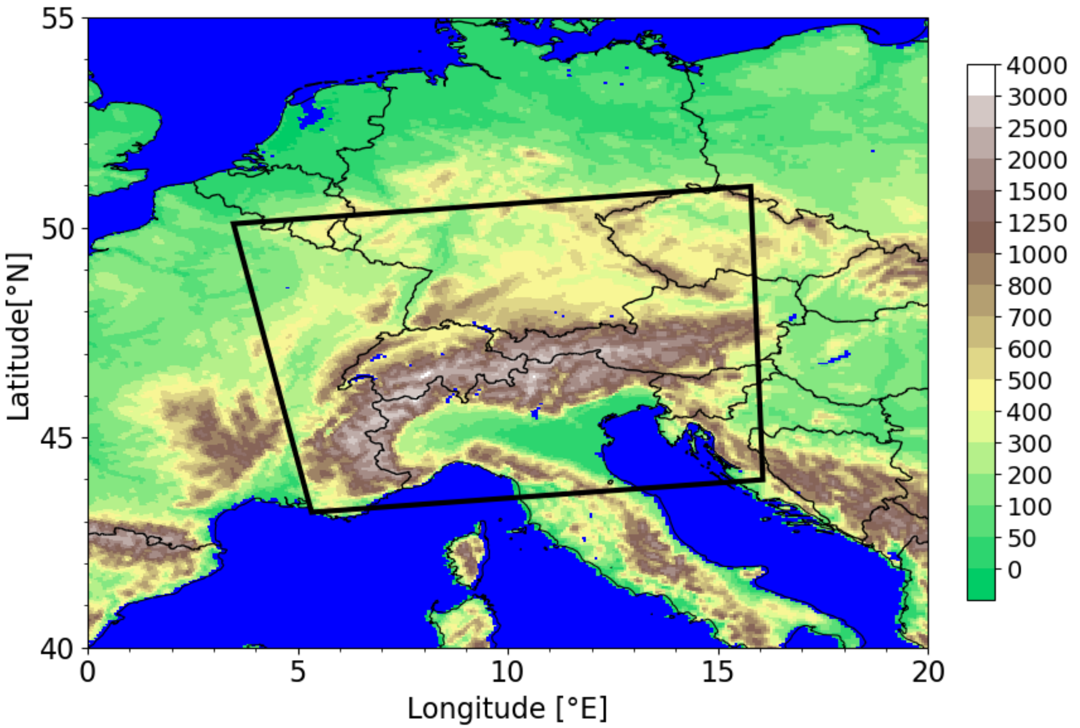
## Design choices and components of the benchmark dataset

**2) The data:**

- Predictands from COSMO-REA6 data ($\Delta x^{rot}_{CREA6} = 0.055°$)
- Task-specific set of predictor variables from the ERA5-reanalysis dataset ($\Delta x_{ERA5} = 0.25°$)
- Ready-to-use dataset provided via climetlab-plugin
- Exemplary data pipelines for TensorFlow and PyTorch in code base

```
!pip install climetlab climetlab_downscaling_benchmark
import climetlab as cml
cml_ds = cml.load_dataset("t2m_downscaling", dataset="validation")

ds = cml_ds.to_xarray()
ds.to_netcdf("downscaling_benchmark_t2m_val.nc")
```




Surface topography form COSMO REA6. The target domain of the downscaling benchmark comprises 144x128 grid points and is rendered.

# THE BENCHMARK DATASET

## Design choices and components of the benchmark dataset

**3) Baseline competitor models**

a) Deep neural networks:

- U-Net by *Sha et al., 2020* (tuned)
- DeepRU (U-Net variant) by *Höhlein et al., 2020*
- WGAN with U-Net by Sha as generator
- WGAN by *Harris et al., 2022*
- SwinIR by *Liang et al., 2021*

b) Classical statistical model:

- Standardized Anomaly MOS (SAMOS) by *Dabernig, 2017*

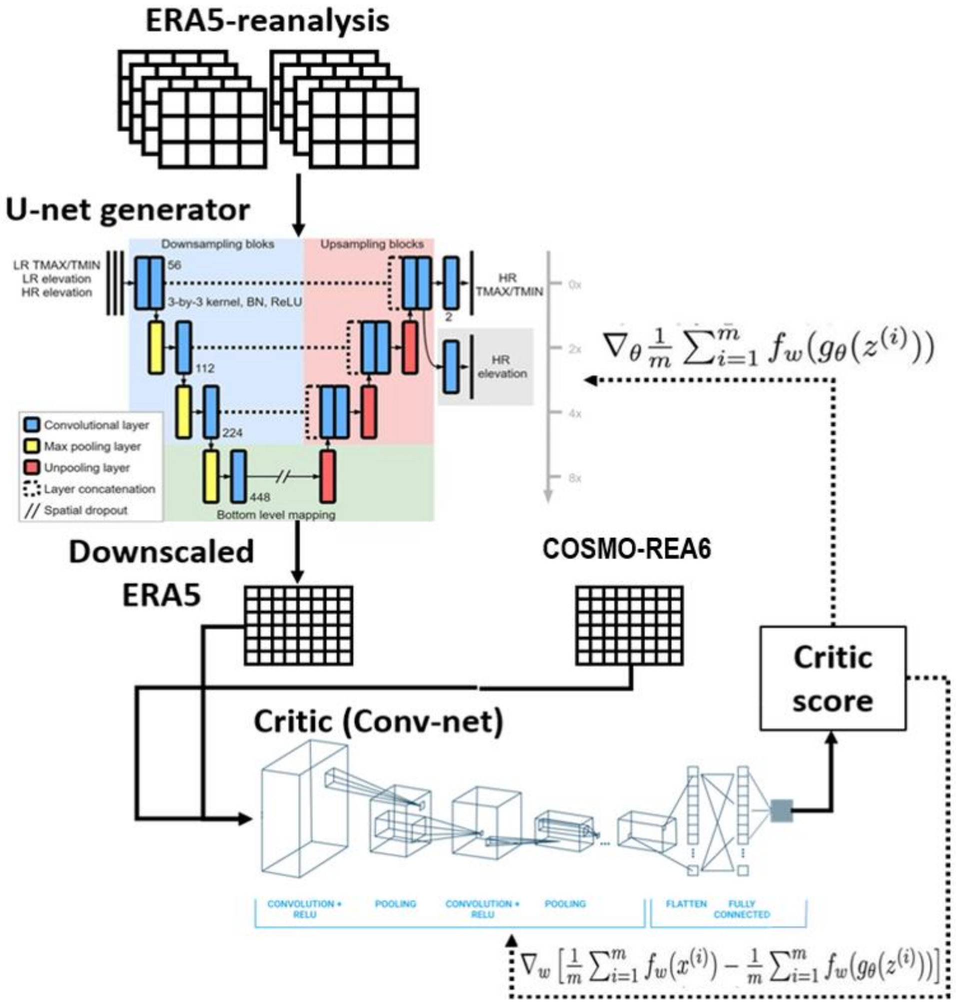



Illustration of the WGAN with the U-Net by Sha et al. (2020) as generator.

# THE BENCHMARK DATASET

## Design choices and components of the benchmark dataset

### 4&5) Evaluation metrics and diagnostics

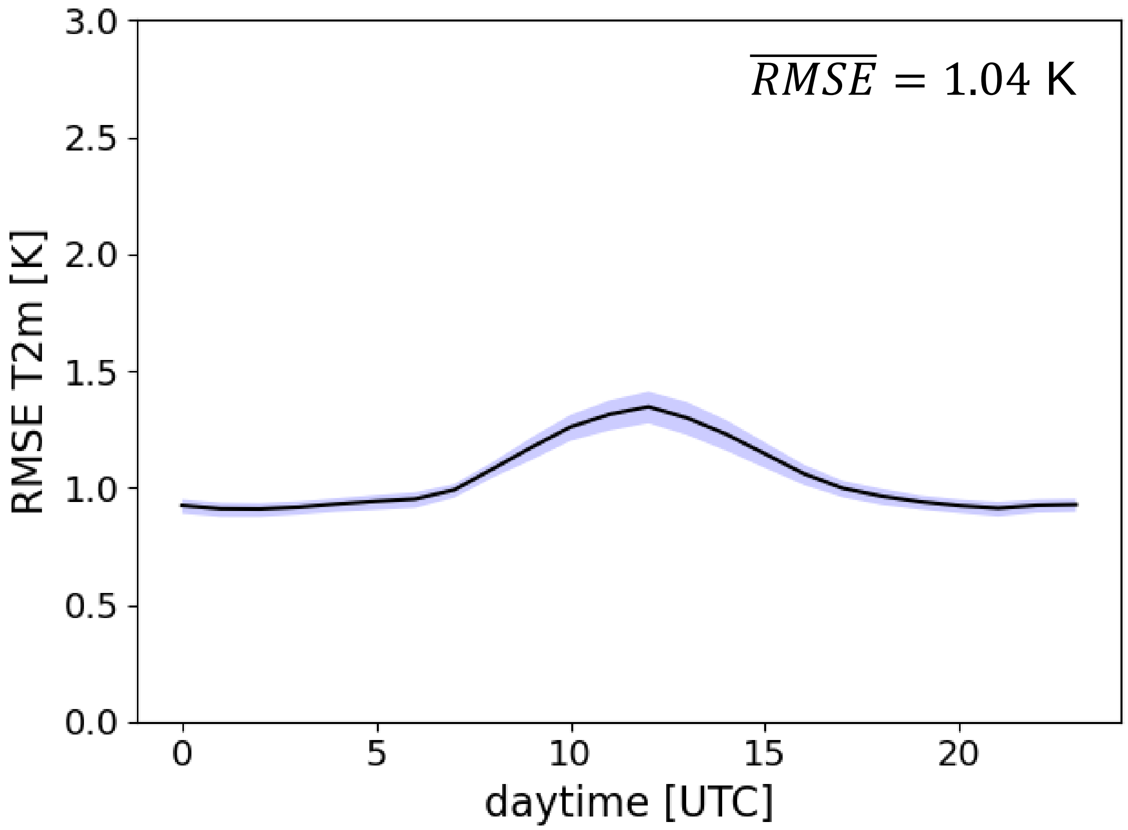
**Aim**: Standardized evaluation framework

- Set of (task-specific) evaluation metrics
- Diagnostics for marginal distribution, e.g. power spectra analysis
- Various plot products for detailed analysis
- Two postprocessing steps:
    1) Single model evaluation (script-based) → prerequisite: results in netCDF-file
    2) Inter-comparison between various models (Jupyter Notebook)

# EVALUATION FRAMEWORK

## Examples from single model evaluation (T2m downscaling)

Results from
**Sha WGAN**



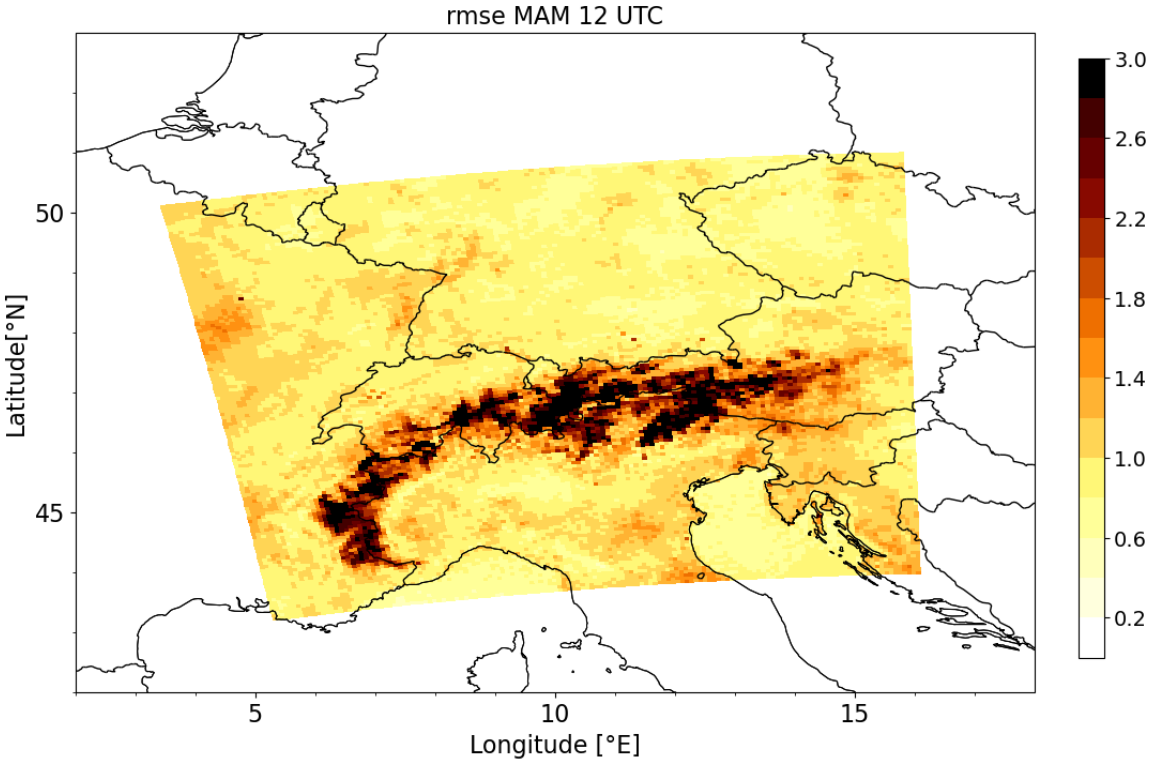

**Better**

Averaged metrics
plotted against daytime
(*seasonal evaluation possible*!)

# EVALUATION FRAMEWORK

## Examples from single model evaluation (T2m downscaling)

Results from
**Sha WGAN**

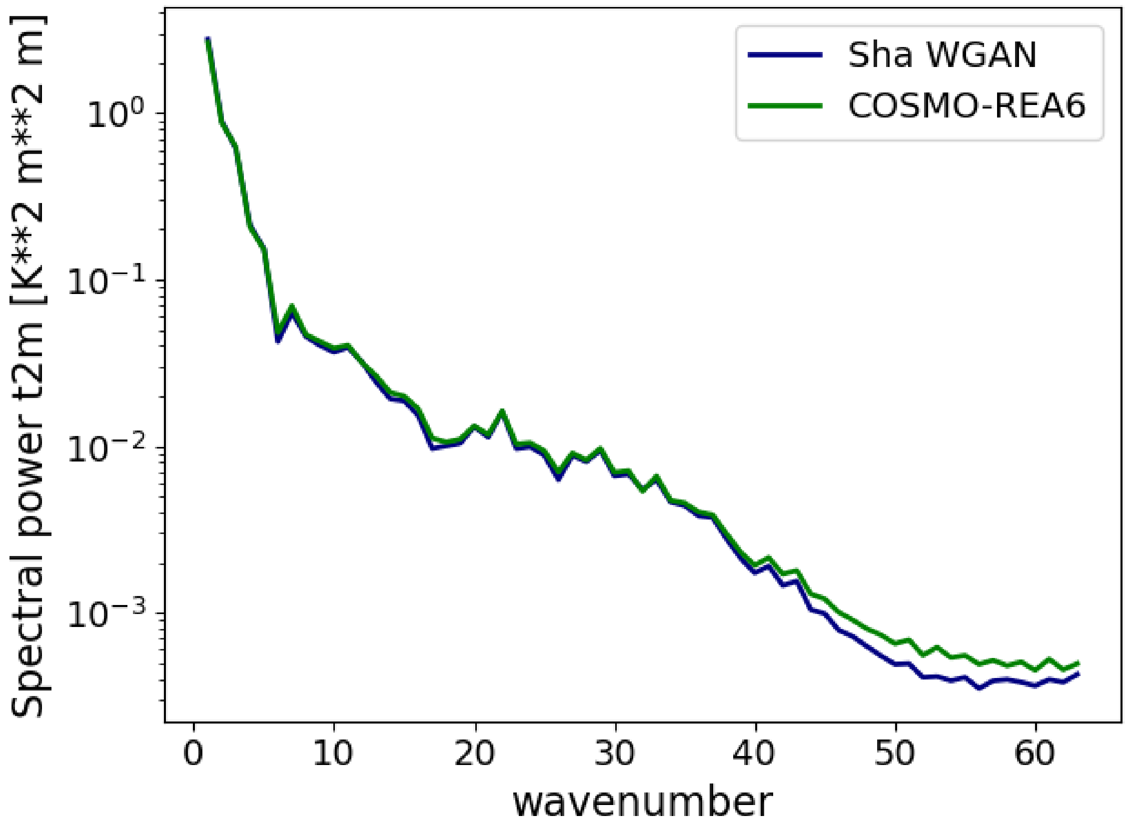



Season-wise spatial evaluation, e.g. RMSE for spring months MAM, or …

# POSTPROCESSING – SINGLE MODEL EVALUATION

## Examples from the 2m temperature downscaling task

Results from
**Sha WGAN**

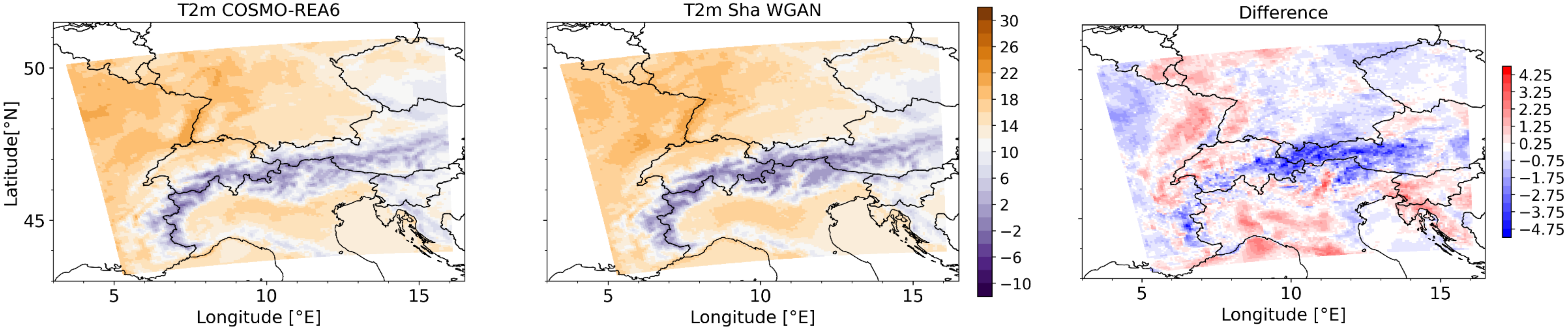



Various options to investigate the spatial variability, e.g. with power spectra …

# EVALUATION FRAMEWORK

## Examples from single model evaluation (T2m downscaling)

Results from
**Sha WGAN**

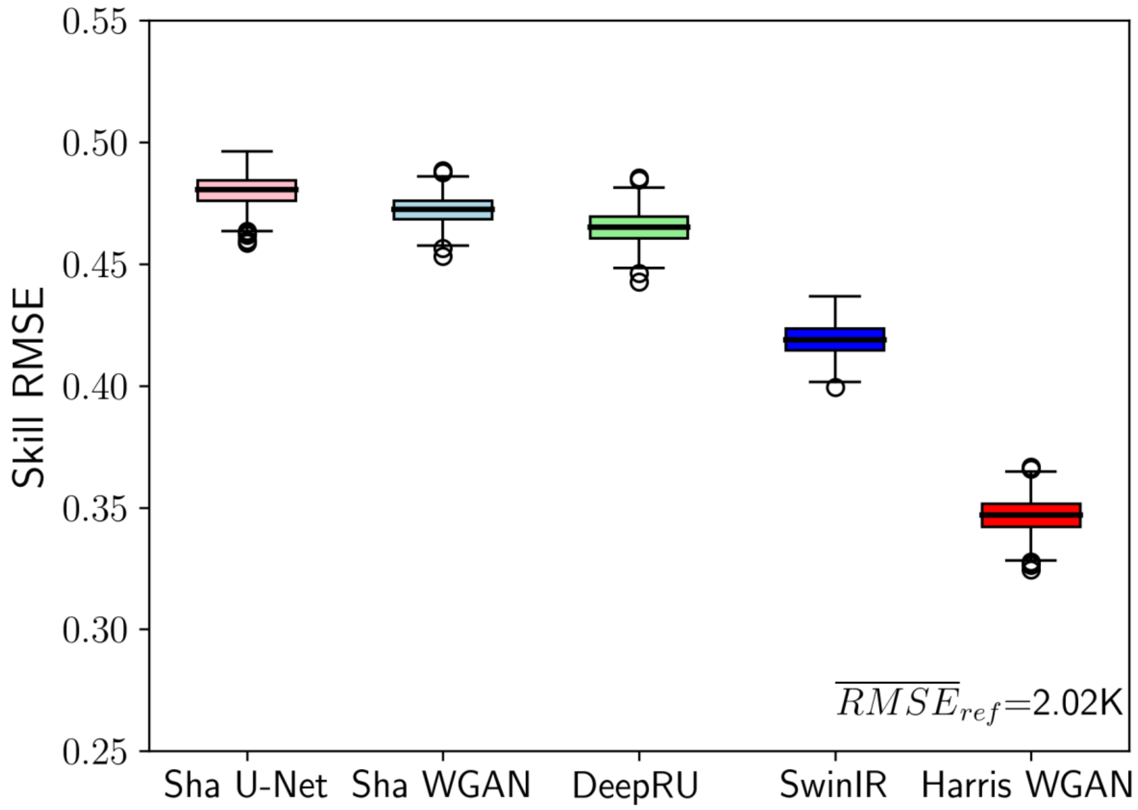



Investigation of specific examples.
Here: downscaling results from 2018-04-07 18 UTC.

# EVALUATION FRAMEWORK

## Examples from model inter-comparison (T2m downscaling)

### Intercomparison




**Better**

Bilinear interpolation as (simple) reference
→ to be replaced by SAMOS

# OUTLOOK

## What's coming next

- Finalize training of baseline models
- Full implementation of metrics and diagnostic tools
- Publication of data via climetlab plug-in
- Publication of code on github
- Accompanying paper

Coming soon

**Future steps:**

- Extension to probabilistic downscaling
- Include precipitation downscaling task

Work initiated by the
**MAELSTROM project**

Check
this out

# REFERENCES


- Dabernig, Markus, et al. "Spatial ensemble post-processing with standardized anomalies." *QJRMS* 143.703 (2017): 909-916.
- Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." *2009* IEEE Inter. Conf. Comp. Vis., IEEE, 2009.
- Dueben, Peter D., et al. "Challenges and benchmark datasets for machine learning in the atmospheric sciences: Definition, status, and outlook." Art. Intell. Earth Sc., 1.3 (2022): e210002.
- Harder, Paula, et al. "Hard-Constrained Deep Learning for Climate Downscaling." *JMLR*, 24.365 (2023): 1-40.
- Harris, Lucy, et al. "A generative deep learning approach to stochastic downscaling of precipitation forecasts." *J. Adv. Model. Earth Sy.*, 14.10 (2022):
- Höhlein, Kevin, et al. "A comparative study of convolutional neural network models for wind field downscaling." *Meteorol. Appl., 27.6* (2020): e1961.
- Liang, Jingyun, et al. "Swinir: Image restoration using swin transformer." *IEEE Inter. Conf. Comp. Vis.* (2021).
- Mardani, Morteza, et al. "Generative residual diffusion modeling for km-scale atmospheric downscaling." *arXiv preprint arXiv:2309.15214* (2023).
- Rasp, Stephan, et al. "WeatherBench 2: A benchmark for the next generation of data-driven global weather models." *arXiv preprint arXiv:2308.15560* (2023).
- Sha, Yingkai, et al. "Deep-learning-based gridded downscaling of surface meteorological variables in complex terrain. Part I: Daily maximum and minimum 2-m temperature." *J. Appl. Meteorol. Clim.,* 59.12 (2020): 2057-2073.
- Wang, Alex, et al. "GLUE: A multi-task benchmark and analysis platform for natural language understanding." *arXiv preprint arXiv:1804.07461* (2018).

# DownscaleBench: A benchmark dataset for statistical downscaling of meteorological fields

## Workshop on Large-Scale Deep Learning for the Earth System | Bonn

2024-08-29 I **MICHAEL LANGGUTH[1]**, SEBASTIAN LEHNER[2], PAULA HARDER[3], ANKIT PATNALA[1], IRENE SCHICKER[2], MARKUS DABERNIG[2], MARTIN G. SCHULTZ[1]

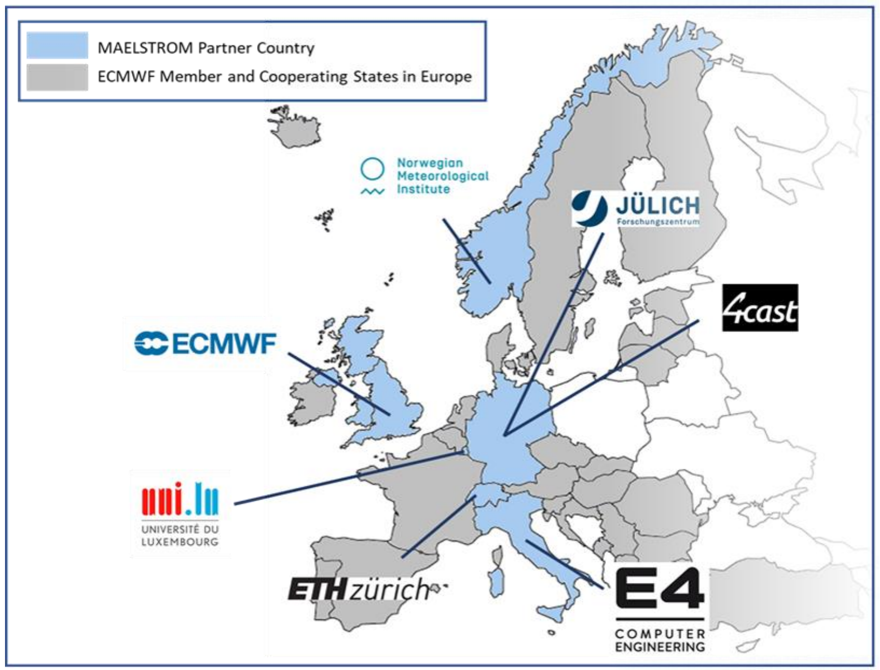
[1] Juelich Supercomputing Centre (JSC), [2] Geosphere Austria, [3] Mila Quebec AI Institute

"The MAELSTROM project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 955513. The JU receives support from the European Union's Horizon 2020 research and innovation programme and United Kingdom, Germany, Italy, Luxembourg, Switzerland, Norway".

# THE MAELSTROM PROJECT

## A EURO-HPC project to foster ML for meteorological applications

- **MA**chin**E** Learning for **S**calable me**T**eo**RO**logy and cli**M**ate
- Euro HPC project coordinated by ECMWF (Apr’21 – Apr’24)
- Main objectives:
  - Develop ML solutions for meteorological applications
  - Enable efficient use of new capacities on supercomputers for the Weather and Climate community
- Collaboration between meteorologists, software developers and HPC specialists
- Six machine learning applications under development
- Benchmark initiative from Application 5



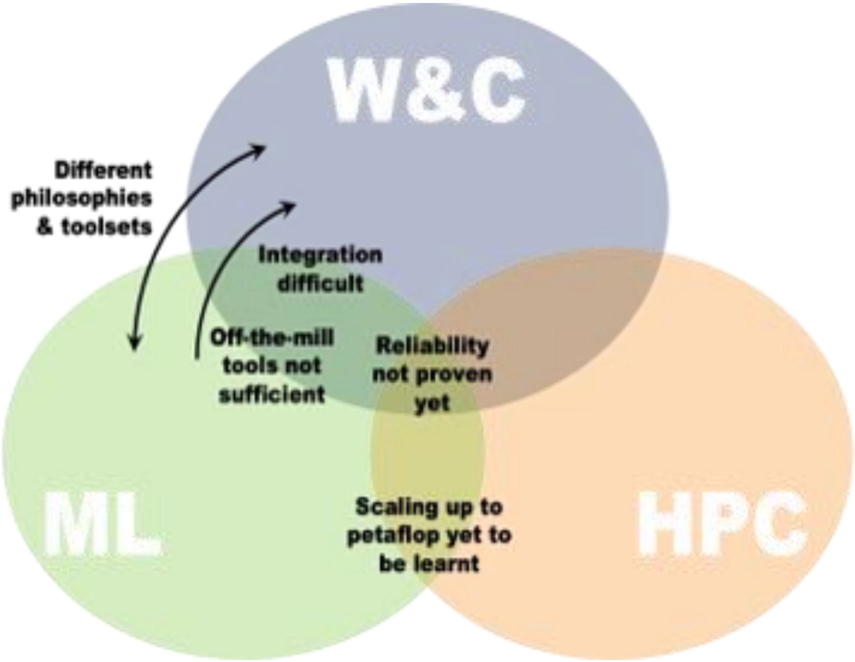

Headquarter location of all partners of the MAELSTROM consortium.

Check this out →

# THE MAELSTROM PROJECT

## A EURO-HPC project to foster ML for meteorological applications

- **MA**chin**E** Learning for **S**calable me**T**eo**RO**logy and cli**M**ate
- Euro HPC project coordinated by ECMWF (Apr’21 – Apr’24)
- Main objectives:
  - Develop ML solutions for meteorological applications
  - Enable efficient use of new capacities on supercomputers for the Weather and Climate community
- Collaboration between meteorologists, software developers and HPC specialists
- Six machine learning applications under development
- Benchmark initiative from Application 5




Co-design cycle in MAELSTROM.

Check this out

# EVALUATION FRAMEWORK

## Examples from the 2m temperature downscaling task

**Dynamic predictors (ERA5 only):**

- 2m temperature + temperature at model levels 137, 135, 131, 127 and 122
- Surface pressure
- 10m (u,v)-wind
- Surface latent and sensible heat fluxes
- Boundary layer height

**Static predictors (ERA5 & COSMO REA6):**

- Surface topography
- Land-sea mask

**Evaluation metrics and diagnostics:**

- RMSE
- Bias
- Mean Error of local standard deviation[1]
- Average gradient amplitude error
- Energy spectra analysis (incl. RALSD-metric)[2]
- Integrated Quadratic Distance[3]
- Conditional Quantile Plots

[1] Zerenner, 2017
[2] Harris et al., 2022
[3] Thorarinsdottir et al., 2013

# DATASETS OF THE DOWNSCALING TASKS

## The 100m (u,v)-wind downscaling task

**Dynamic predictors (ERA5 only):**

- 100m (u,v)-wind
- (u,v)-wind at model levels 135, 133, 131, 127 and 122
- Boundary layer height
- Geopotential height at 500 hPa
- Surface pressure

**Static predictors (ERA5 & COSMO REA6):**

- Surface topography
- Land-sea mask

# DATASETS OF THE DOWNSCALING TASKS

## The 100m (u,v)-wind downscaling task

**Dynamic predictors (ERA5 only):**

- 100m (u,v)-wind
- (u,v)-wind at model levels 135, 133, 131, 127 and 122
- Boundary layer height
- Geopotential height at 500 hPa
- Surface pressure

**Static predictors (ERA5 & COSMO REA6):**

- Surface topography
- Land-sea mask

**Evaluation metrics and diagnostics:**

- MSE and absolute relative error
- Cosine dissimilarity
- Magnitude difference
- Mean Error of local standard deviation
- Kinetic energy spectra

# DATASETS OF THE DOWNSCALING TASKS

## The horizontal global radiation downscaling task

- Post-processed global horizontal irradinace (GHI) from *Frank et al., 2018* as target data (rather than raw COSMO REA6)

**Dynamic predictors (ERA5 only):**

- Surface net solar radiation
- Top net solar radiation
- High, medium and low cloud cover
- Cloud base height
- Total column liquid water
- Surface pressure
- CAPE
- Evaporation

**Static predictors (ERA5 & COSMO REA6):**

- Surface topography
- Land-sea mask
- Slope of sub-grid scale orography

# DATASETS OF THE DOWNSCALING TASKS

## The horizontal global radiation downscaling task

- Post-processed global horizontal irradinace (GHI) from *Frank et al., 2018* as target data (rather than raw COSMO REA6)

**Dynamic predictors (ERA5 only):**

- Surface net solar radiation
- Top net solar radiation
- High, medium and low cloud cover
- Cloud base height
- Total column liquid water
- Surface pressure
- CAPE
- Evaporation

**Static predictors (ERA5 & COSMO REA6):**

- Surface topography
- Land-sea mask
- Slope of sub-grid scale orography

**Evaluation metrics and diagnostics:**

- Relative RMSE and MAE
- Bias
- Mean Error of local standard deviation
- Fraction Skill Score
- Conditional Quantile Plots