Leveraging AROME reanalysis to develop AI-based regional forecasting systems

Thomas RIEUTORD[1]*, Sara AKODAD[1], Clément BROCHET[1], Benjamin DEVILLERS[1], Mikaël JACQUEMONT[2], Eliott LUMET[1], Olivier PANNEKOUCKE[1], David SALAS Y MELIA[1], Corentin SEZNEC[1], Laure RAYNAUD[1]

[1]Météo-France – CNRM
[2]Bull SAS

*Correspondance : thomas.rieutord@meteo.fr

April 2026

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

METEO FRANCE
À VOS CÔTÉS, DANS UN CLIMAT QUI CHANGE

Table of contents

Introduction

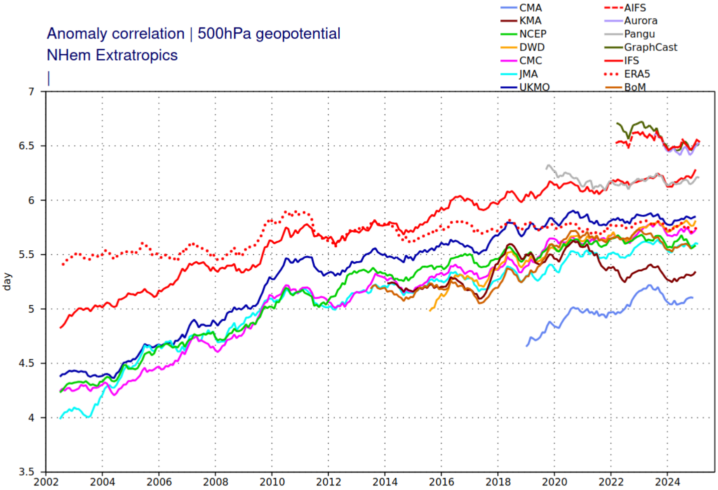


Figure 21: Lead time at which the 500 hPa anomaly correlation in the northern extratropics drops to 86% (specific threshold chosen such that lead time does not exceed forecast range of any centre). Comparison of physics-based forecasts from other global centres, IFS-CF (red), AIFS-Single (red dashes), ERA5 (red dots) for reference, and other ML models run at ECMWF.

Credit: ECMWF, *Evaluation of ECMWF forecasts.* [Haiden et al., 2025]

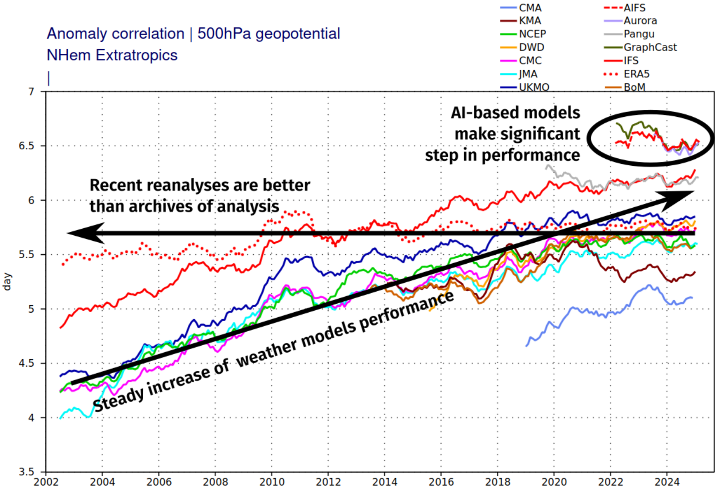
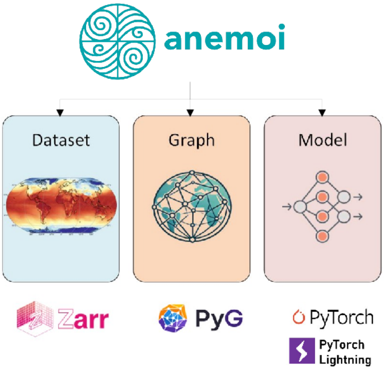


Figure 21: Lead time at which the 500 hPa anomaly correlation in the northern extratropics drops to 86% (specific threshold chosen such that lead time does not exceed forecast range of any centre). Comparison of physics-based forecasts from other global centres, IFS-CF (red), AIFS-Single (red dashes), ERA5 (red dots) for reference, and other ML models run at ECMWF.

Credit: 🅭🅯 ECMWF, *Evaluation of ECMWF forecasts.* [Haiden et al., 2025]

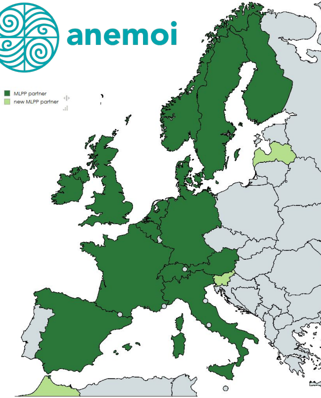
Context

- Within the ECMWF **Machine Learning Pilot Project**, Météo-France is developing AI-based models for regional weather prediction.
- We investigate **auto-regression and downscaling** approaches.
- All development is made thanks to the open-source **Anemoi framework**.
- In parallel, Météo-France is producing a 60-year long **regional reanalysis**: ARRA.



The ECMWF Machine Learning Pilot Project (MLPP)

anemoi



Goal: Integration of **machine learned weather forecasting** into operational value chain of **national weather services,** based on the **Anemoi** software ecosystem.

Duration: mid-2024 - mid-2029 = 5 years

Partners: 18 (17 ECMWF Member/Co-operating States)

Funded by ECMWF, but **driven** by the partners

Covers the full forecasting chain: ML models, ensembles, data assimilation, operations and infrastructure

MLPP is not a single ML lab or model — it is a shared European effort to build ML into the forecasting system.

Questions of this presentation

Can we compare trainings on reanalysis and on analysis?

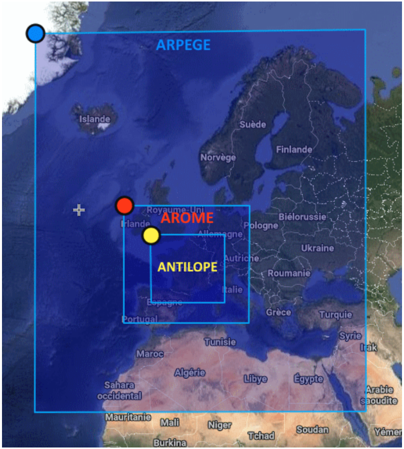
What is the performance of our AI-based model compared to the physics-based?

What is the performance of the downscaling model?

Datasets: TITAN and ARRA

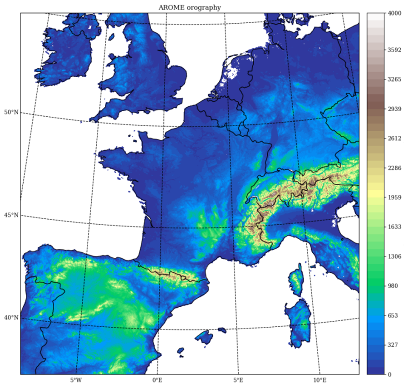
TITAN: Training Inputs and Targets from Arome for Neural networks

- **Type:** archive of operational analyses + radar data.
- **Resolution:** $\Delta x = 2.5$km / 1.3km, Δt = 1h, Δz : 24 levels
- **Time span:** 2020–2024 (5 years)
- **Variables:** 278 (11 aloft, 14 surface)
- **Assimilation:** 3DEnVar [Brousseau et al., 2025]
- **Documentation:** [Berthomier et al., 2024]



ARRA: ARome Re-Analysis

- **Type:** regional reanalysis
- **Resolution:** $\Delta x = 1.3$km, Δt = 1h, Δz : 90 levels
- **Time span:** 1962–2020 (59 years)
- **Variables:** >1000 (everything AROME outputs)
- **Assimilation:** Surface only, every 3h + IAU. Precipitations from MESCAN [Soci et al., 2016]
- **Documentation:** [Bazile et al., 2025]



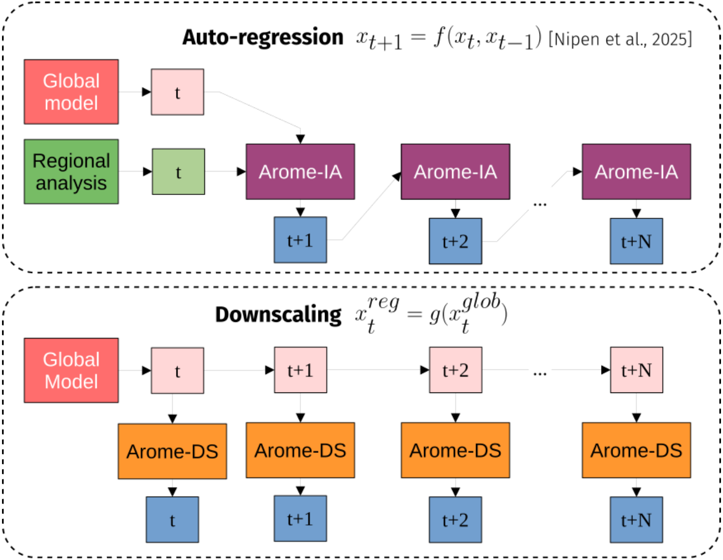
Comparing datasets

	TITAN (analyses)	**ARRA** (reanalysis)
Strengths	More obs assimilated. Most recent cycles. Live coupling	Long time span. Fixed cycle
Weaknesses	Changes in cycles. Short time span	Less obs assimilated. Coupled with other reanalyses

- Reanalyses are expected to be better for training on **large number of years**
- Analyses are expected to be better for **finetuning** on the latest version of the model

Methods

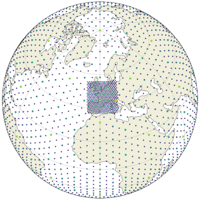
Two approaches for regional forecasting



ANITI cnrs METEO FRANCE anemoi

Courtesy: Eliott Lumet

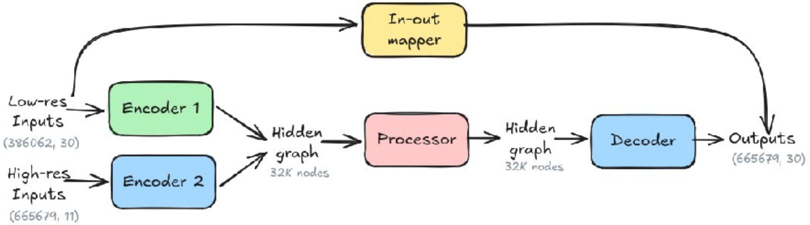
Our auto-regressive model



Stretched-grid deterministic model for regional prediction, as in [Nipen et al., 2024].

- **Resolution:** $\Delta x = 2.5$ km (regional) / 0.25° (global), Δz : 13 levels, $\Delta t = 6$ h
- **Variables:** 72
    - Surface (7 pronostic) : $t_2, u_{10}, v_{10}, p_{msl}, p_{surf}, t_{surf}, td_2$
    - Surface (1 diagnostic) : total precipitations
    - Altitude (5) : t, u, v, z, q
- **Model:** GraphTransformer (1024 channels, 246M parameters)
- **Training strategy:** Transfer learning from an AIFS checkpoint, finetuning on a regional dataset.
- **Inference:** ARPEGE for global, AROME for regional. Real time production 4 times a day.

Our downscaling model



- **Resolution:** Δx = from 10 km (global) to 2.5 km (regional), Δz : 5 levels, $\Delta t = 6$ h
- **Variables:** 30
  - Surface: $t_2, r_2, u_{10}, v_{10}, p_{msl}$
  - Altitude (5) : t, r, u, v, z
- **Model:** GraphTransformer (512 channels, 60M parameters)
- **Training strategy:** EDM diffusion on residuals [Mardani et al., 2025]

Our methods of evaluations

Can we compare trainings on reanalysis and on analysis?

- We train the model twice: on TITAN (2020-2024) and on ARRA 2016-2020: exact **same length and settings but different period**.
- We **evaluate both models quantitatively** against observations and compare them to AROME on June 2024.

What is the performance of our AI-based model compared to the physics-based?

- **Quantitative** evaluation on Oct. 2024 and June 2024.
- **Qualitative** evaluation in the storm Kirk.

What is the performance of the downscaling model?

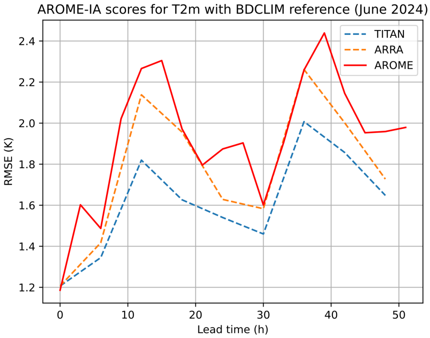
- **Qualitative** evaluation of a few cases.

Results

Comparing trainings on analyses and on reanalysis

Clear advantage of TITAN over ARRA

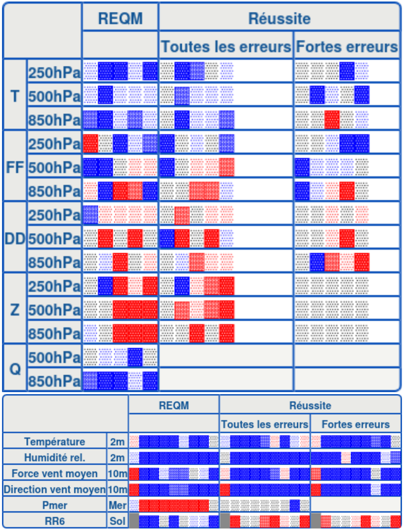
- **Assimilation** in TITAN (oper analyses) include more obs than in ARRA
- **Coupling** in TITAN (with ARPEGE) is better than in ARRA (with CERRA)
- **Inferences** for evaluation (from analyses) favor TITAN



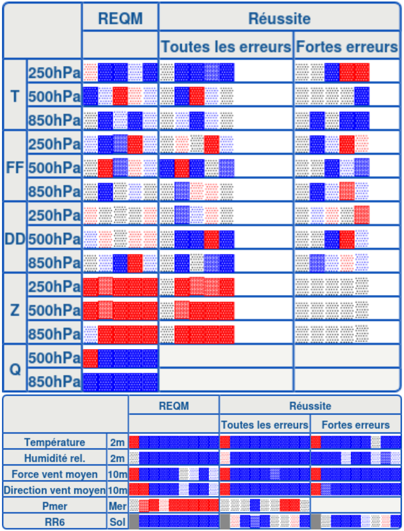
In the following of the presentation, we focus on the evaluation of the model **trained with TITAN**, called "AROME-IA".

Score cards of the auto-regressive model

October 2024



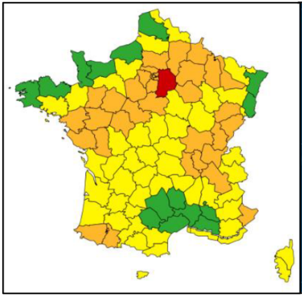
June 2024

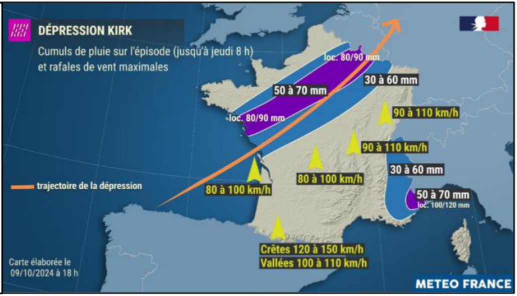


- Colors: AROME-IA better/worse than AROME.
- **Better at the surface** than aloft. Strongly **depend on variables**

The storm Kirk

The storm **Kirk** is an ex-hurricane causing heavy rainfall on the North-West of France and strong wind on the Pyrenees and the Center-East between the 9th and 10 of October 2024.

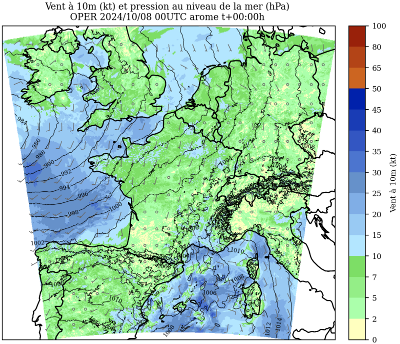




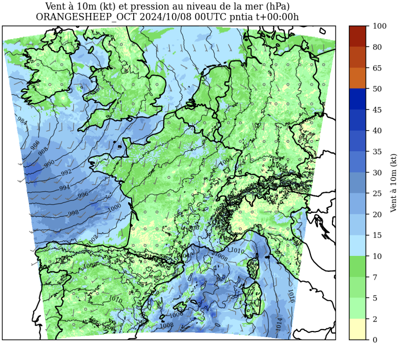
- **1 red warning** for river flooding (Seine et Marne)
- **35 departements** in orange warning for wind et rainfall

Kirk : 10-m wind speed and MSLP

AROME

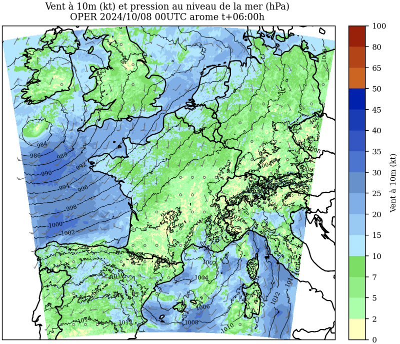


AROME-IA

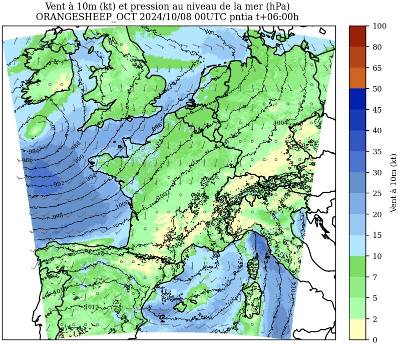


Kirk : 10-m wind speed and MSLP

AROME

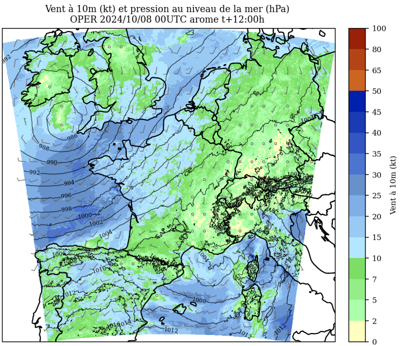


AROME-IA

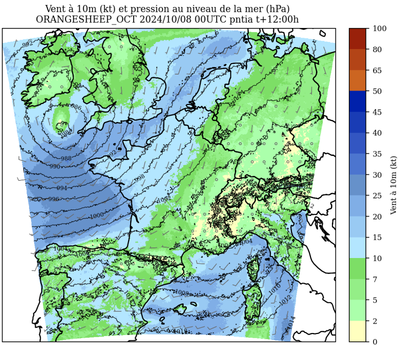


Kirk : 10-m wind speed and MSLP

AROME

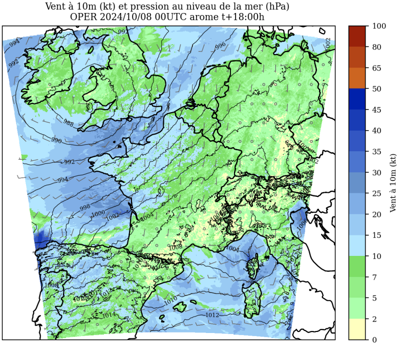


AROME-IA

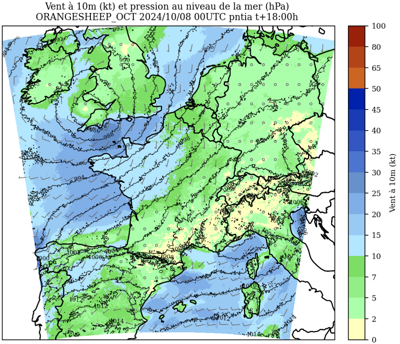


Kirk : 10-m wind speed and MSLP

AROME

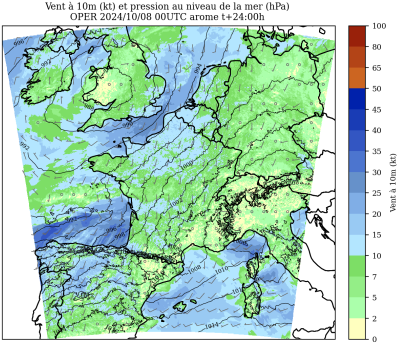


AROME-IA

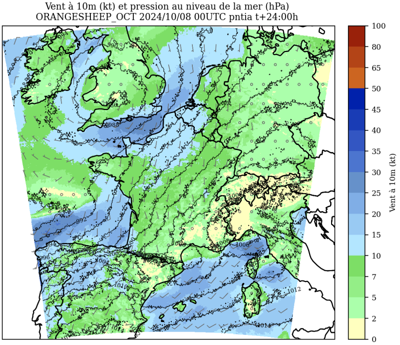


Kirk : 10-m wind speed and MSLP

AROME

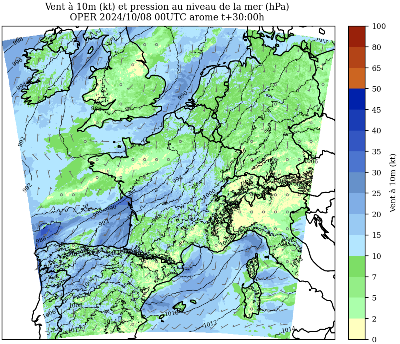


AROME-IA

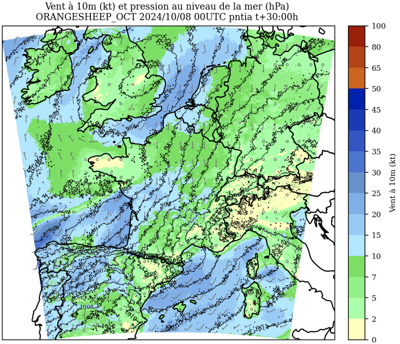


Kirk : 10-m wind speed and MSLP

AROME

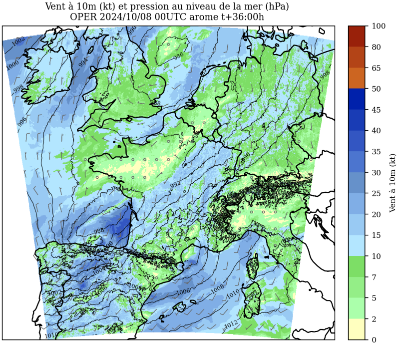


AROME-IA

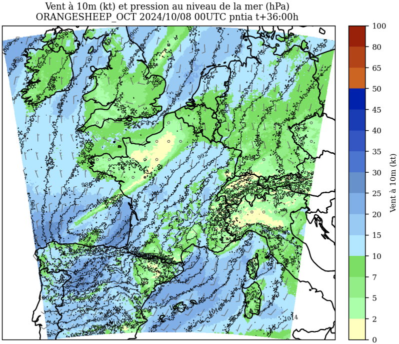


Kirk : 10-m wind speed and MSLP

AROME

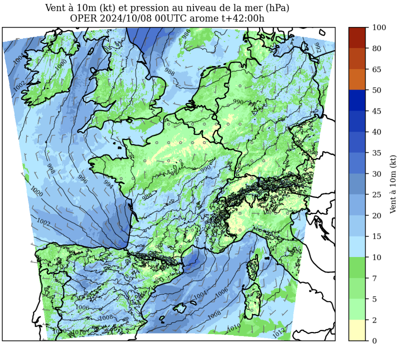


AROME-IA

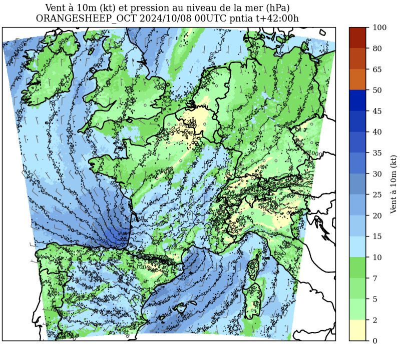


Kirk : 10-m wind speed and MSLP

AROME

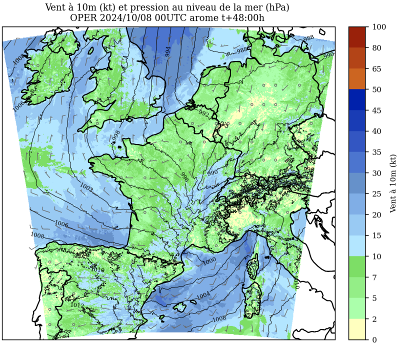


AROME-IA

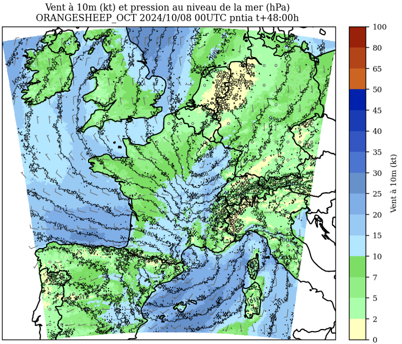


Kirk : 10-m wind speed and MSLP

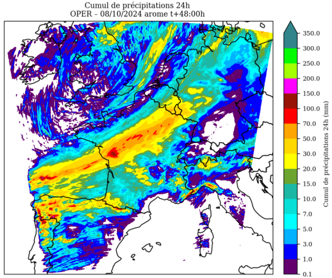
AROME

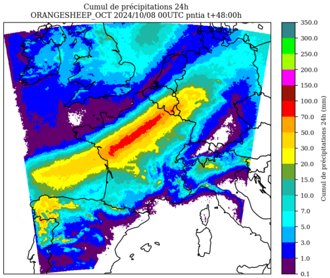


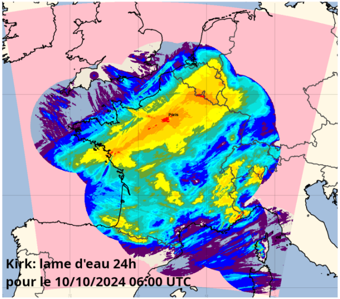
AROME-IA



Kirk : 24-h accumulation of precipitations over the event





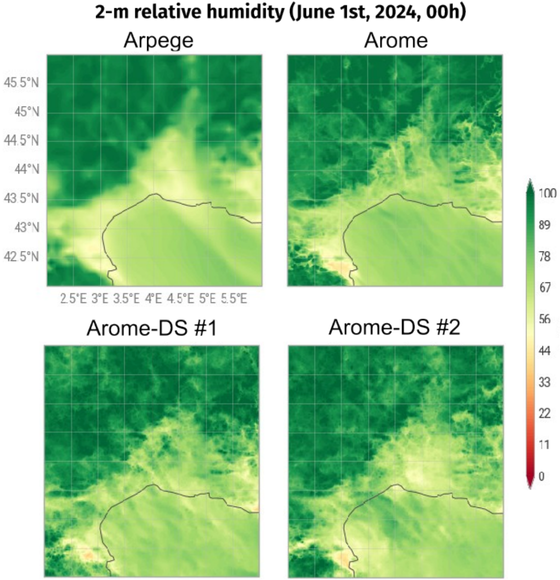


Position and timeline of precipitation are correctly forecast

Intensity of rainfall inaccurate
- too strong on Center
- too weak on Alps and South West

Weird field texture,
both blurry and noisy

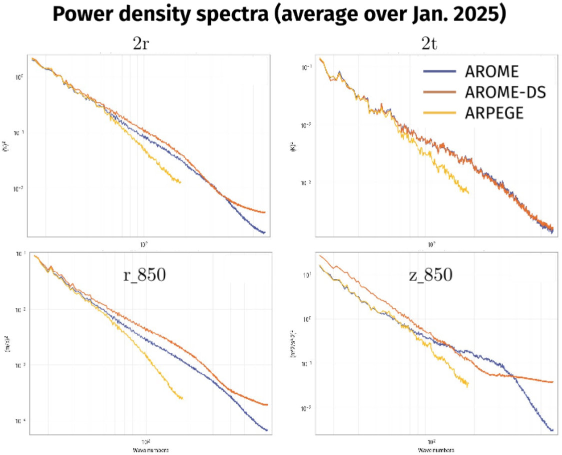
Downscaling of relative humidity fields



Courtesy: Eliott Lumet

- Recover well the **fine-scale** details from AROME
- Learns the correlation with **orography**
- Significant inter-sample **variability**

Downscaling: spectra of a few variables



Courtesy: Eliott Lumet

- Overall good for **surface** variables
- Tendency to overestimate the **fine-scale** energy → consistent with noisy aspect
- Worse for geopotential and **higher levels**
- Need for longer **training**

Optimisation of training: LoRA in Anemoi?

LoRA (Low Rank Adaptation) injects trainable rank decomposition matrices into each layer of the Transformer to replace specific linear layers.

➕ reduces trainable parameters, computational and memory load

➖ degrades performances of the model

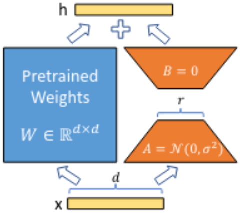


Figure 1: Our reparametrization. We only train A and B.

Credit: [Hu et al., 2021]

- First tests on the AROME-IA configuration:
  - trainable parameters: **246M** $\rightarrow$ **0.67M** (-99.7%)
  - compute speed: **0.23 step/s** $\rightarrow$ **0.25 step/s** (+8.7%)
  - GPU RAM load: **344GB** $\rightarrow$ **280GB** (-18.6%)
- Contribution to come in Anemoi:
  - Fork: `https://github.com/AI4SIM/anemoi-core/`
  - Branch: `feat/lora-adapter`
  - Commit: `805659f89d48564220f46963942780d440286cd2`

Prospects and conclusions

Limitations and prospects

Make better use of ARRA

- Train on increasing length periods (5, 15, 50 years) to remove artefacts and improve intensity.
- Add more variables to meet end-users requirements.
- Focus on extreme events, with smart sampling or specific loss.

Improve downscaling

- Scale up the training: reach convergence, adapt noise level, add variables
- Change in the architecture: one-encoder, with residuals
- Temporal consistency: convergence with auto-regression?

Conclusions

Can we compare trainings on reanalysis and on analysis?

- **Cautiously** because datasets have different strengths and weaknesses
- This experiment is a **first step** towards larger use of ARRA

What is the performance of our AI-based model compared to the physics-based?

- For some variables, **scores are comparable to AROME**
- The **timing and positioning** of events are good.
- The **texture of the fields** is strange (both blurry and noisy), presence of **artefacts**...

What is the performance of the downscaling model?

- Qualitatively **good for some variables**, not for some others.
- Provide ensembles with **satisfying spread**.
- Training has **not reached convergence** yet.

References i

Bazile et al. (2025).
Preliminary evaluation of the ARRA re-analysis over France : focus on extreme precipitation.
Technical Report EMS2025-648, Copernicus Meetings.

Berthomier et al. (2024).
Titan: Training inputs and targets from arome for neural networks.
https://github.com/meteofrance/py4cast/blob/main/doc/titan.md.

Brousseau et al. (2025).
The operational 3DEnVar data assimilation scheme for the Meteo-France convective scale model AROME-France.
EGUsphere, pages 1–31.

Haiden et al. (2025).
Evaluation of ECMWF forecasts.
Technical Memorandum no. 931.

Hu et al. (2021).
LoRA: Low-Rank Adaptation of Large Language Models.
arXiv:2106.09685 [cs].

Mardani et al. (2025).
Residual corrective diffusion modeling for km-scale atmospheric downscaling.
Communications Earth & Environment, 6(1):124.

Nipen et al. (2024).
Regional data-driven weather modeling with a global stretched-grid.
http://arxiv.org/abs/2409.02891.

Soci et al. (2016).
High-resolution precipitation re-analysis system for climatological purposes.
Tellus A: Dynamic Meteorology and Oceanography, 68.