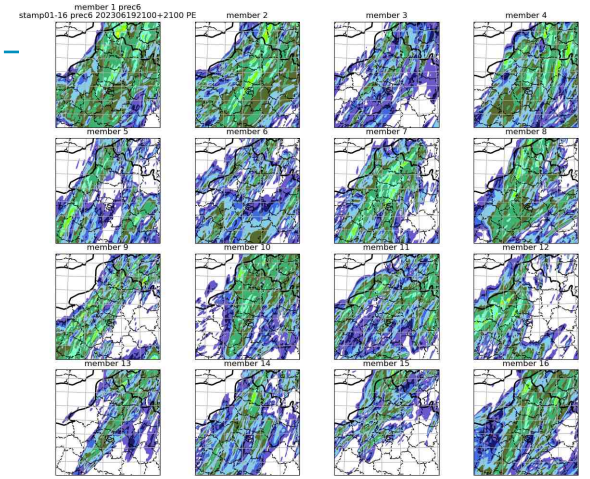
Measuring the value of precipitation forecasts in hi-res NWP

with a user-oriented score

François Bouttier and Hugo Marchal, Météo-France/CNRM

NHESS preprint : https://egusphere.copernicus.org/preprints/2024/egusphere-2023-3111/

DWD hecto workshop, 5-7 Feb 2024



Data : one year of

- deterministic Arome-France, dx=1.4km (+ ongoing work at dx=500m)
- 17-member Arome EPS ensemble dx=1.4km, range 3-48 h
- vs obs: 1-km precip analyses (radar + gauges)

Method : ensemble post-processing

➔ *impact model :* select space & time scales + weight false alarms vs non-detections

➔ *application :* rain threshold exceedance:

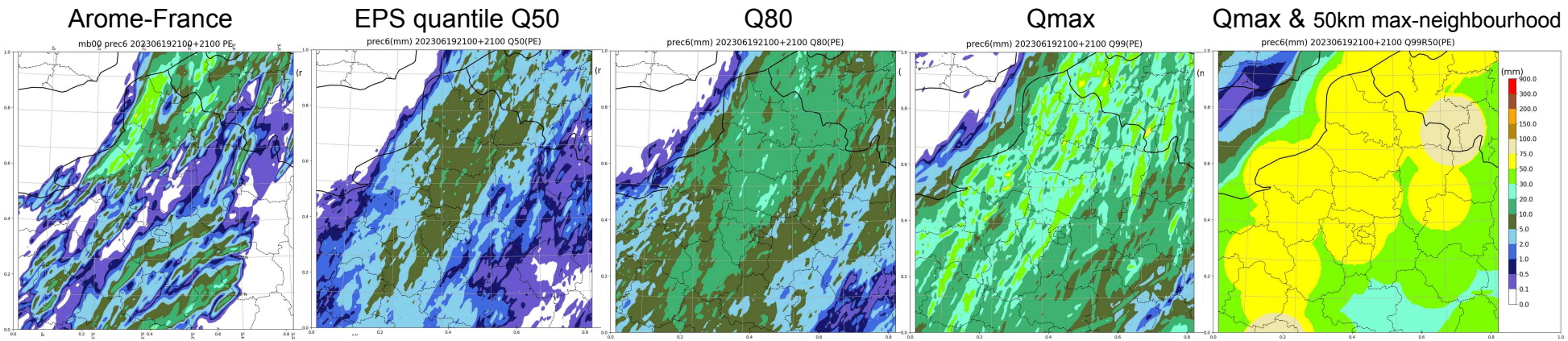
- convert ensemble output into binary forecasts using **calibrated neighbourhood & probability thresholding**
- score for high-impact events: **F2** (allows ~4 x more FAs than NDs)

Drawbacks of usual NWP scores

- deterministic HSS, CSI, ETS, ND, FA... : require tuning a quantile or prob threshold to handle EPS
- CRPS, Brier, etc : too abstract for most end users, do not indicate which prob thresholds to use
- ROC, Economic Value : for users that calibrate the forecasts themselves. Collapse for rare events.

Evaluation metric used here :

1. max-neighbourhood "NMEP" post-processing of ensemble members (+dressing)
2. objective optimization of *Popt* prob threshold on past forecast archive : "optimal forecast scenario"
3. forecast scores vs full resolution observations using ETS and F2 metrics



Popt Probability threshold optimization

- (neighbourhood radius is optimized in parallel)
- includes ensemble calibration

Results on heavy rain events : (6h accum > 40mm)

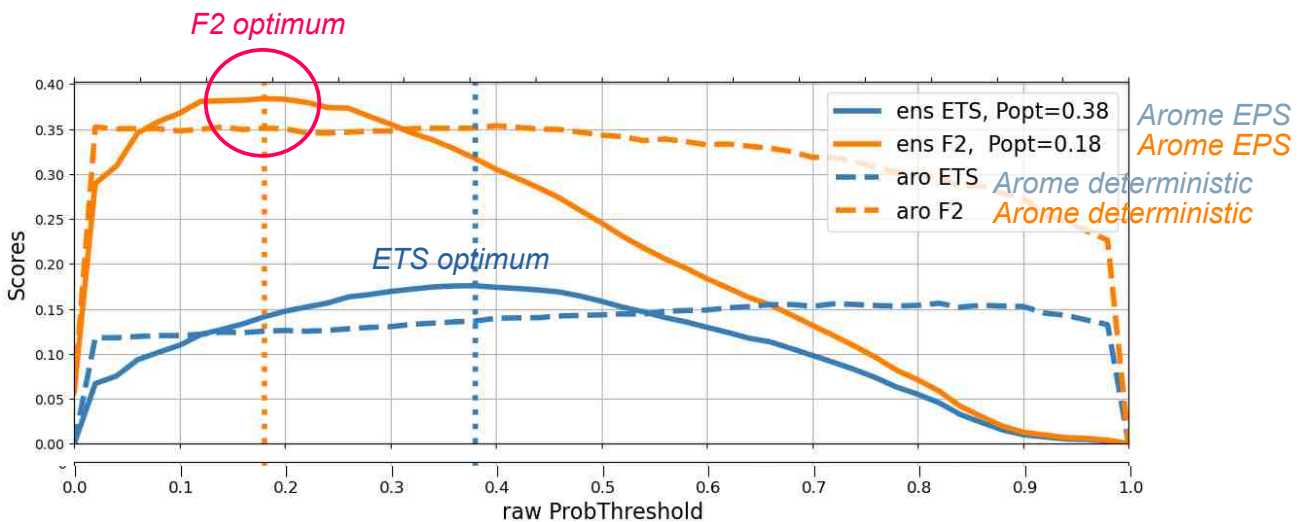
- F2-score is optimum at radius ~30km and prob Popt ~ 0.18
- quite stable wrt intensity so equivalent to using 82%-quantile as reference scenario
- EPS is a bit better than deterministic models (when using near-optimum settings)

BUT Popt is quite different with light rain or other user objectives (ex. BSS, CSI, HSS, ETS)

F2-score :

F2 = a / (a+1/5b + 4/5c)

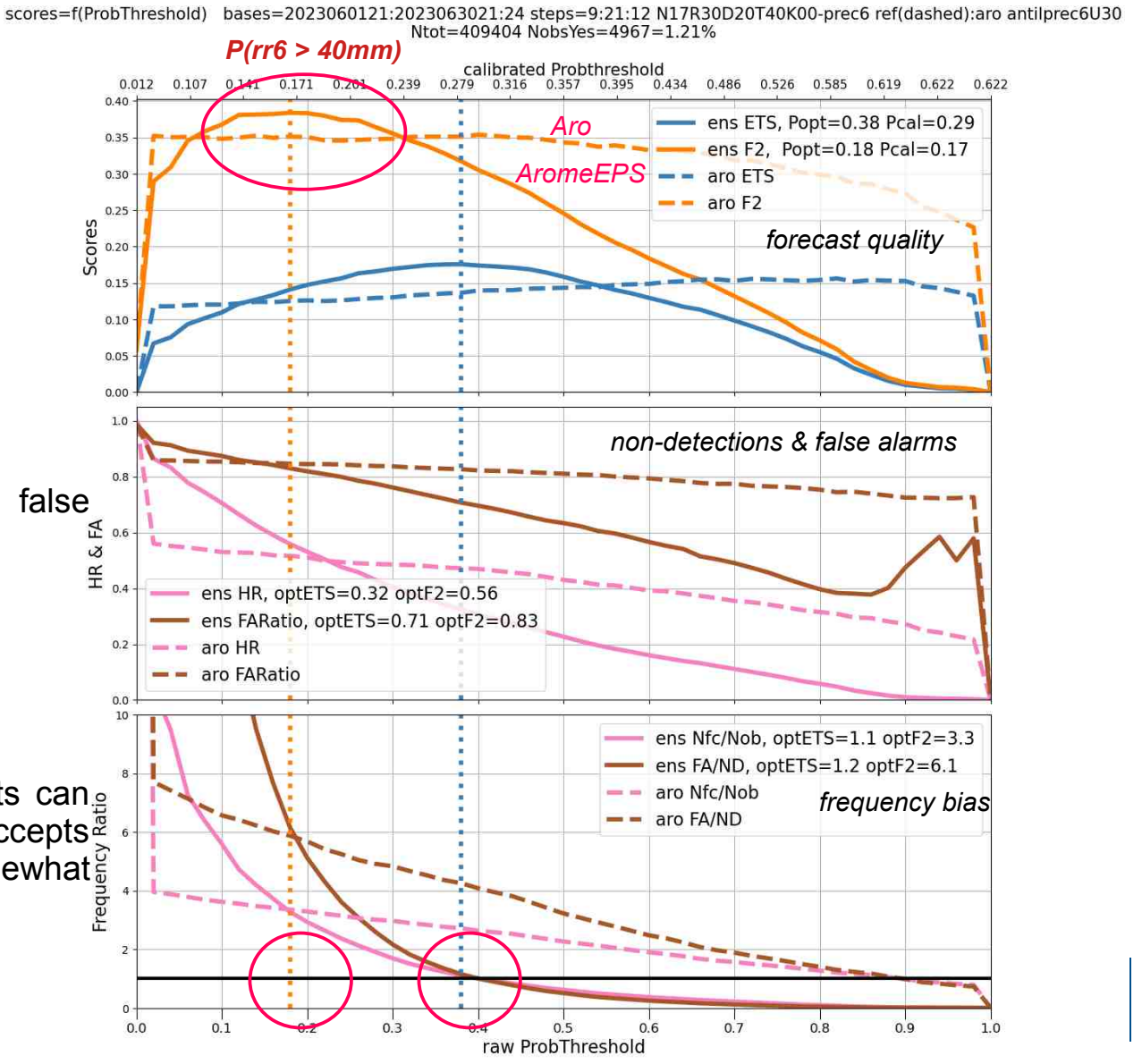
a : correct positives
b : FA
c : ND
in 2x2 confusion (contingency) matrix



Variation of error statistics vs Popt probability threshold

Optimizing F2-score leads to ~4x more false alarms than non-detection.

For infrequent events, probabilistic forecasts can only beat deterministic models if one accepts some **overforecasting bias** (i.e. a somewhat "unphysical" forecast)



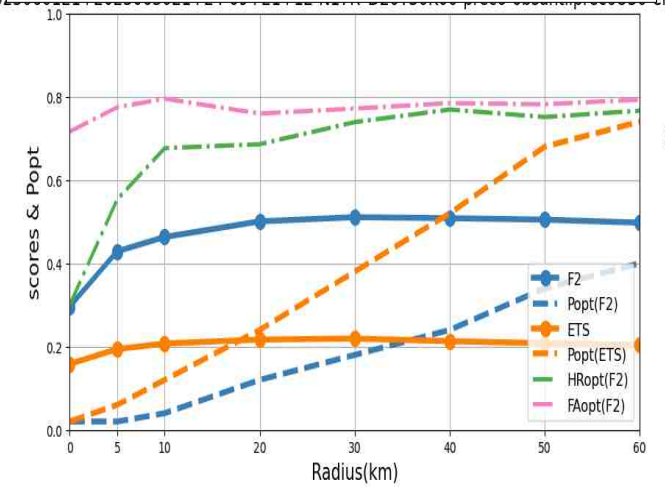
Sensitivity study of Popt tuning and implied forecast performance

forecast range does not matter until 36h

(but the diurnal cycle does)

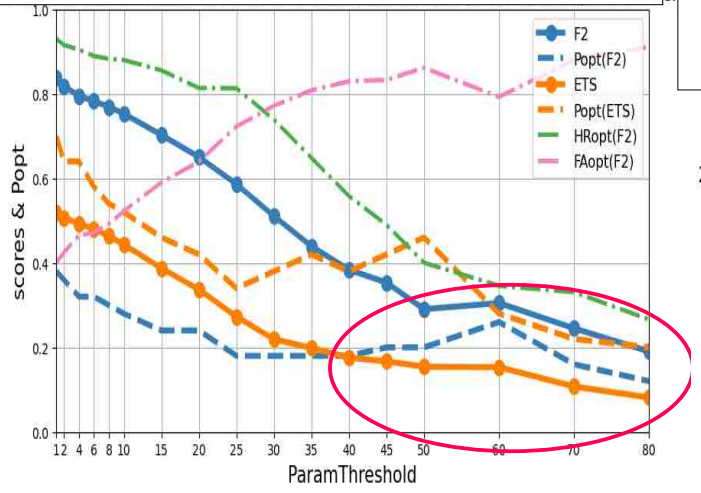
impact of postprocessing **neighbourhood radius** :

- optimum ~ 20 to 50km for heavy rain
- (matters less for light rain)



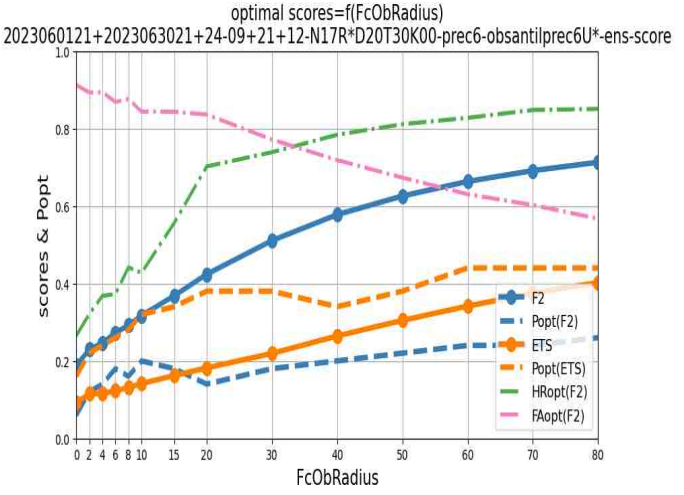
impact of rain intensity :

- optimal quantile levels extrapolate well to extreme events



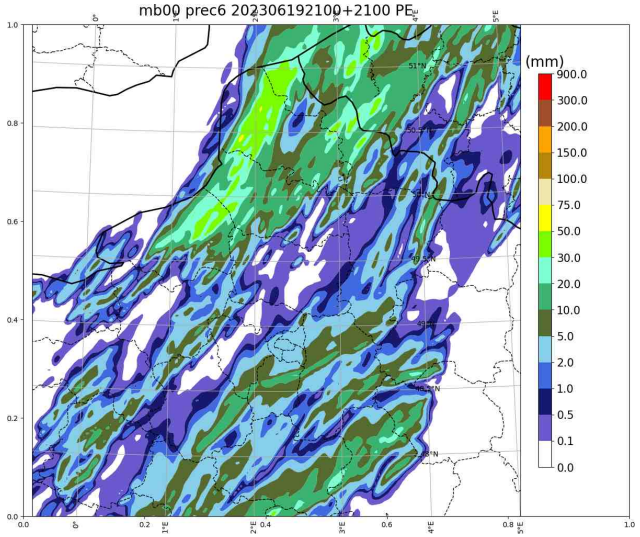
verification scale : *(see also FSS papers)*

- very low predictability of heavy rain under 20km : poor detection, lots of false alarms

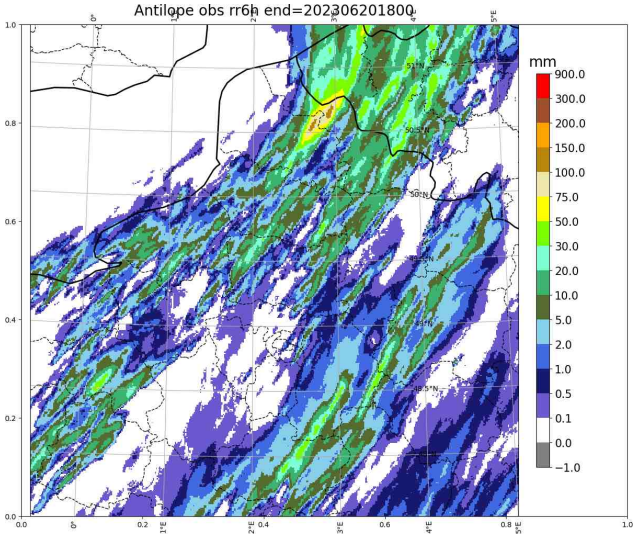


Case of 20 juin 2023 (frontal)

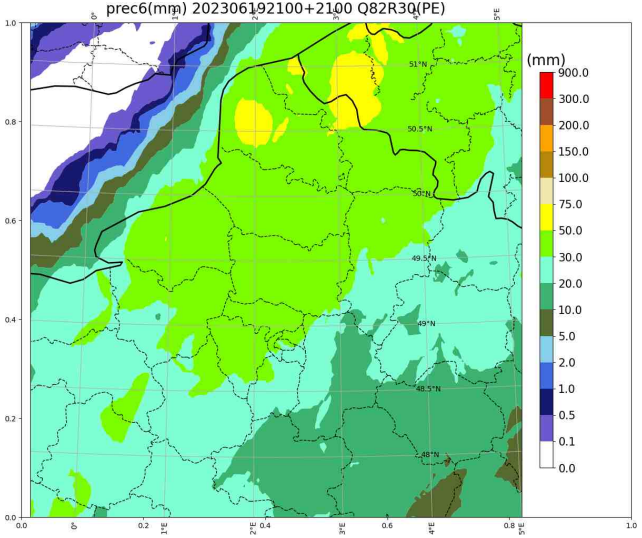
Arome deterministic



obs

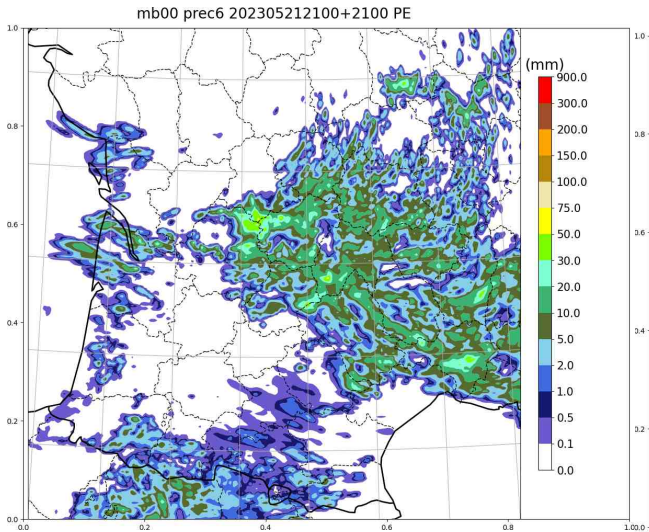


F2-optimal quantile for heavy rain

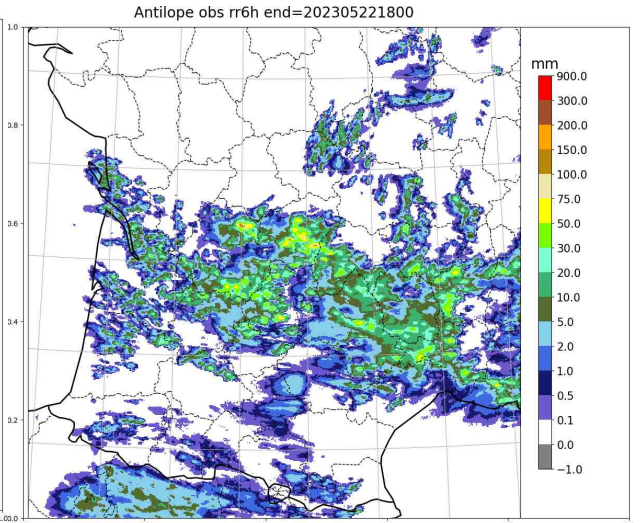


Case of 22 May 2023 (weakly forced thunderstorms)

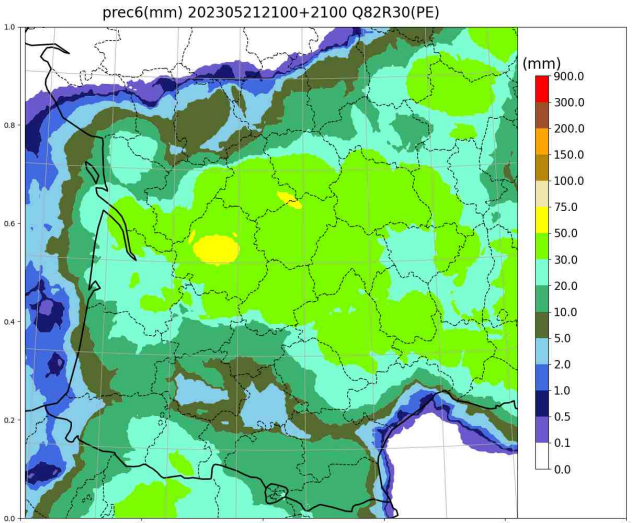
Arome deterministic



obs



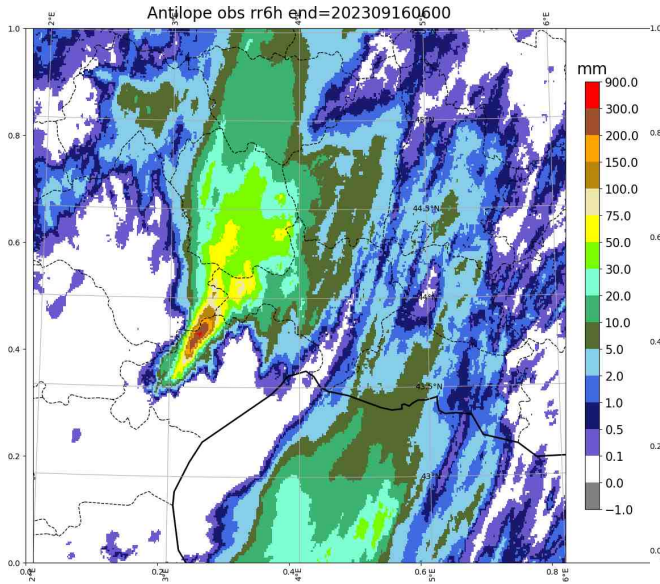
F2-optimal quantile for heavy rain



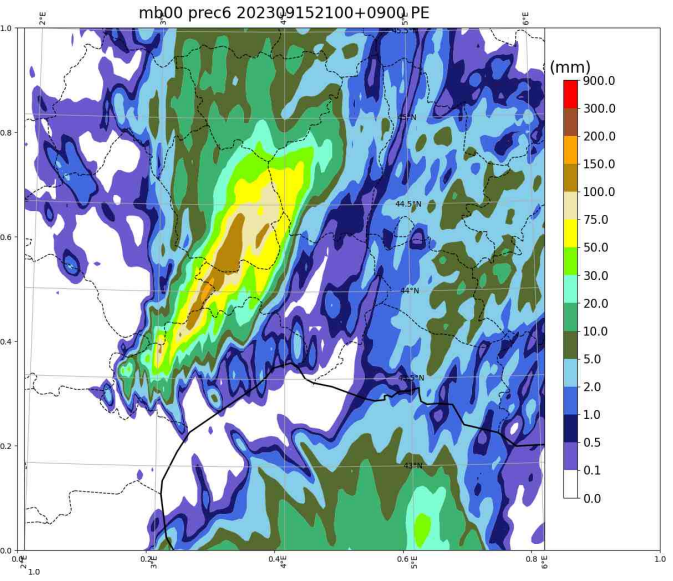
Case of 16 Sep 2023 (orographically forced heavy Mediterranean event)

The globally optimum neighbourhood is too large for this event type
-> need adaptive neighbourhood post-processing

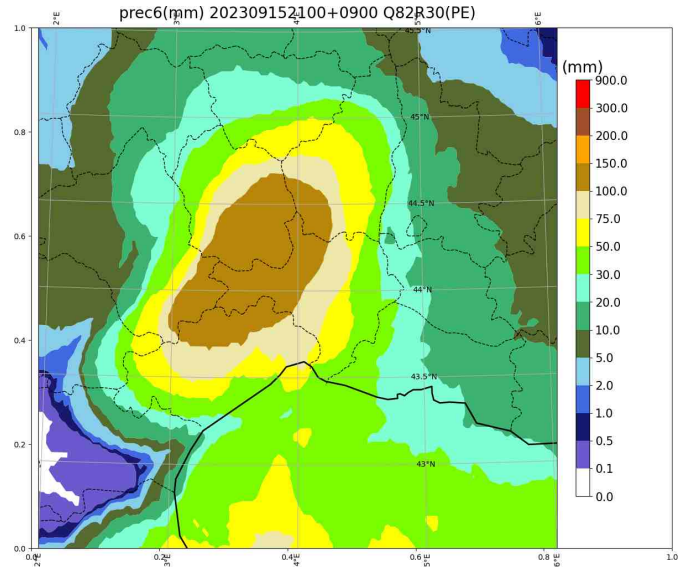
Arome deterministic



obs



F2-optimal quantile for heavy rain



Next steps

Target :

- apply to a 500-m AROME ensemble for nowcasting
- 0-6h range, updated hourly
- emphasis on heavy thunderstorm and flash flood events

Problems :

- demonstrate value over a larger, lower-resolution ensemble
- need assimilation to fit obs/clouds better
- too expensive ! Use some AI downscaler to emulate the 500-m model ?