Correcting and using uncertainty quantification through statistical post-processing

Maxime Taillardat

Météo-France / CNRM UMR 3589
maxime.taillardat@meteo.fr

13/11/2025

METEO FRANCE

CNRM

Outline

Ensemble forecasts in a nutshell

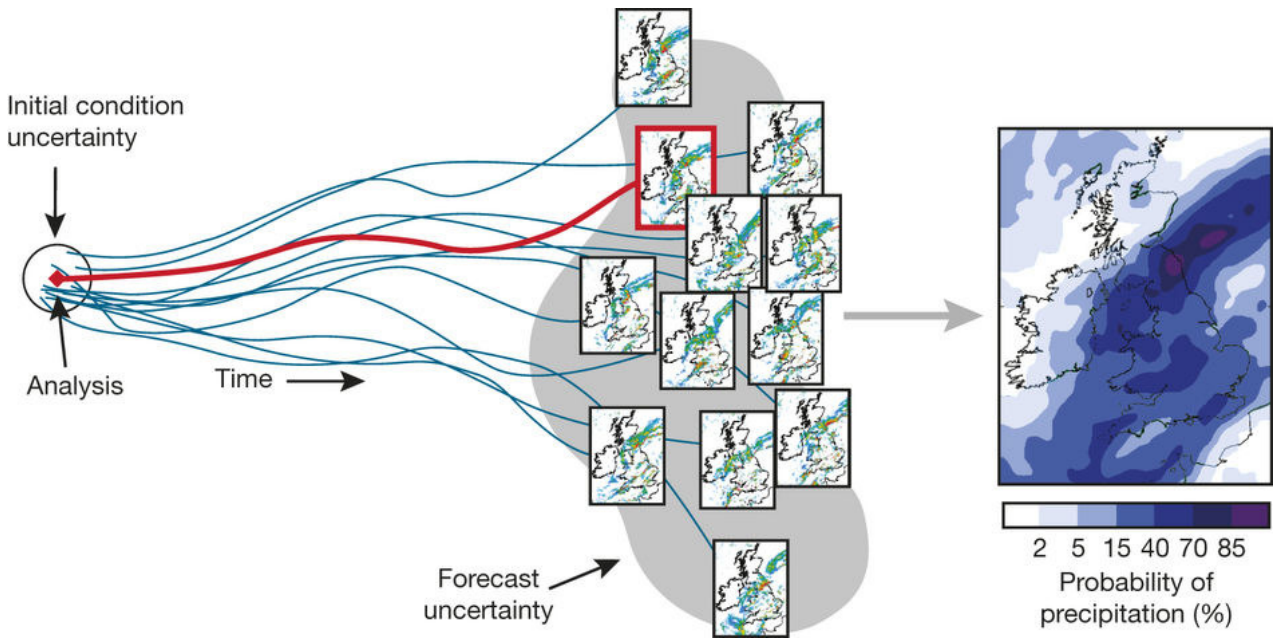
	Initial state perfectly known	Uncertainty in initial state
Equations known	Laplace's daemon	Chaos (Poincaré-Lorenz)
Uncertainty in equations	Hawkmoth effect	« We are here»

Unlimited calculations :

Fokker-Planck

Limited calculations :

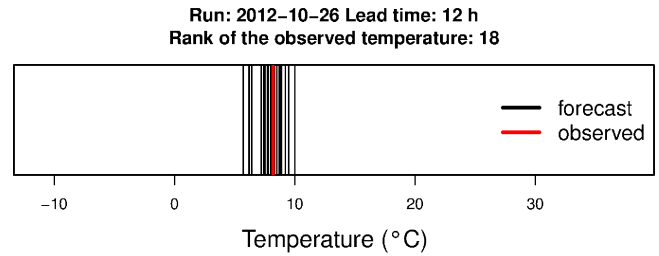
Ensemble prediction

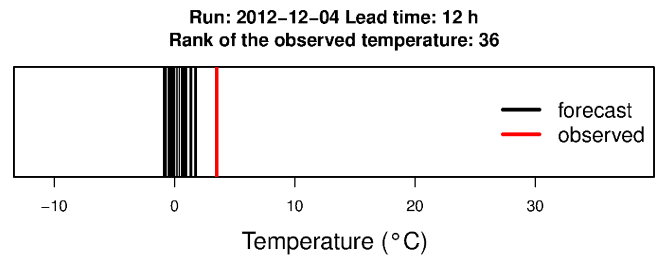


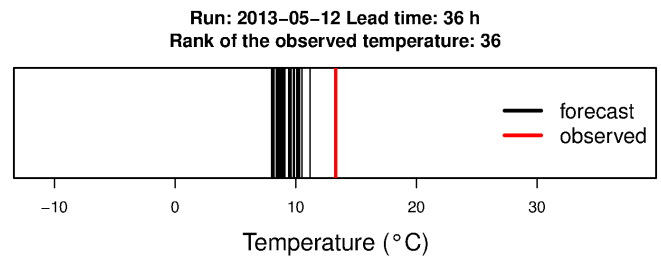
Motivation for post-processing

Météo-France's 35-members global ensemble system (PEARP), 5 km resolution over France.

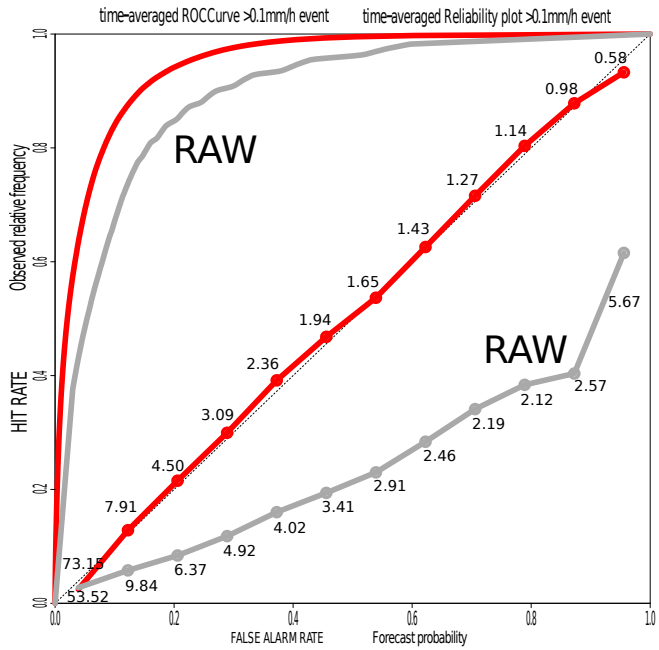
Observations and forecasts of 2-m temperature (T2m) at Lyon-Bron for the run of 1800UTC (different lead times)

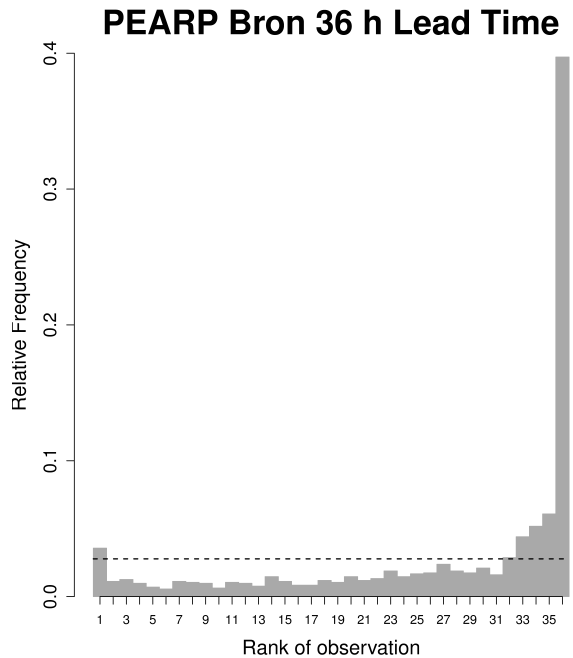






Motivation for post-processing





Possible sources of mismatch

- Vertical / horizontal resolution
- Physical processes
- Approximations / design in data assimilation
- Coupling

How to verify/evaluate ensemble forecasts ?

- Differ from point forecasts
- A "point" (eg. for one day) verification is a nonsense
- It has to be **statistical**

Several attributes sought

- **Reliability/Calibration**
  - Accordance between forecasted probabilities and observed frequencies of an event and/or exchangeability between observations and ensemble members
- **Resolution/Discrimination/Information**
  - Ability to differ from a climatological forecast
- **Sharpness**
  - Getting the least dispersed forecast, subject to calibration

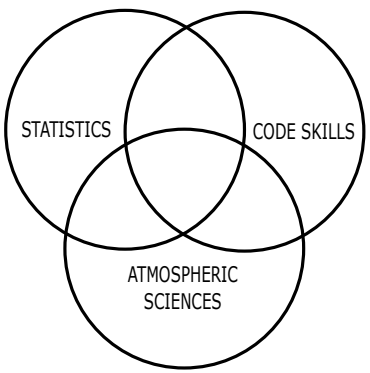
Types of statistical post-processing

Escape from the model land...

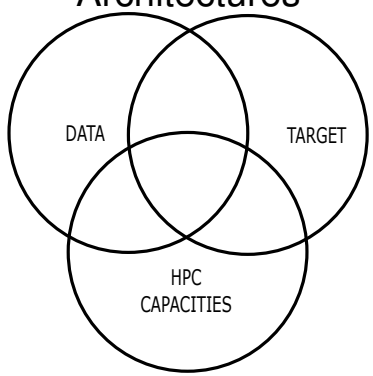
- Deterministic PP: bias correction of a deterministic model
- Ensemble PP (calibration): distribution correction of an EPS
- Blending: deterministic or probabilistic agregation of "experts"
- Statistical learning between past forecasts and observations
- Dynamical structure of NWP is not altered

A review of techniques: Vannitsem, S., et al. "Statistical Postprocessing for Weather Forecasts: Review, Challenges, and Avenues in a Big Data World." Bulletin of the American Meteorological Society 102.3 (2021): E681-E699.

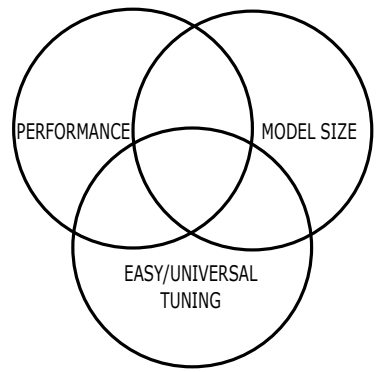
People



Architectures



Methods



Current post-processing chain at Météo-France

Model XXX	Domain	Resolution	Variables	Type of Obs.
AROME	France	1km	T, Tn, Tx	Stations
AROME	France	1km	Q, Qn, Qx	Stations
AROME-FOS	FOS	2.5km	FF, FX	Stations + Analysis
AROME-FOS	Tahiti	2.5km	RR1	Stations + Analysis
AROME-IFS	France	1km	T, Tn, Tx	Stations
ARPEGE	France	1km	T, Tn, Tx	Stations
ARPEGE	France	10km	FF, FX	Stations + Analysis
ARPEGE	France	10km	TCC	Satellite
ARPEGE	World	50km	TCC	Satellite
ALL	France	1km	T, Tn, Tx	Stations
IFS	France	1km	T, Tn, Tx	Stations
IFS	France	1km	Q, Qn, Qx	Stations
IFS	Tahiti	10km	RR, RRn, RRx	Stations + Analysis
IFS	World	25km	TCC	Satellite
AROME-EPS	France	2.5km	RR1	Radar
AROME-EPS	Europe	2.5km	FF, FX	Stations + Analysis
ARPEGE-EPS	France	1km	T, Tn, Tx	Stations
ARPEGE-EPS	France	10km	FF, FX	Stations + Analysis
ARPEGE-EPS	France	10km	RR3	Radar
EPS	France	1km	T, Tn, Tx	Stations

EMOS (Q)RF BOA CNN Boosting MLR

Benefits of non-parametric post-processing

No assumptions on the weather variable you deal with

- Raw ensemble: 

EMOS

Non-parametric

Benefits of non-parametric post-processing

No assumptions on the weather variable you deal with

- Raw ensemble:

EMOS

- EMOS post-processing:

Non-parametric

Benefits of non-parametric post-processing

No assumptions on the weather variable you deal with

- Raw ensemble:

EMOS

- EMOS post-processing:

Non-parametric

- Post-processing possible results:

Benefits of non-parametric post-processing

No assumptions on the weather variable you deal with

- Raw ensemble:

EMOS

- EMOS post-processing:

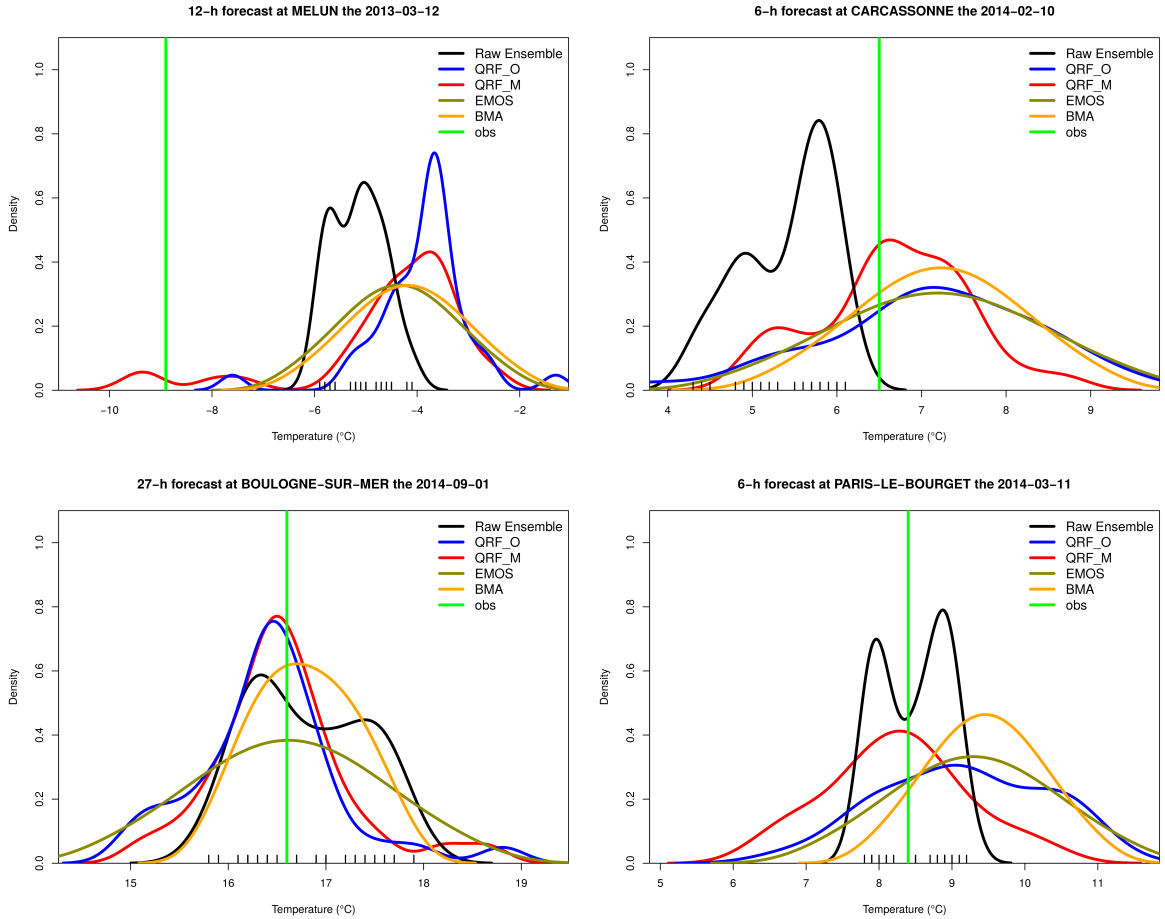
Non-parametric

- Post-processing possible results:





Best-of post-processing possible outputs



Meinshausen, N. (2006). Quantile regression forests. Journal of Machine Learning Research, 7(Jun), 983-999.
Taillardat, M., Mestre, O., Zamo, M., & Naveau, P. (2016). Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. Monthly Weather Review, 144(6), 2375-2393.

Benefits of non-parametric post-processing

No assumptions on the weather variable you deal with

- Raw ensemble:

EMOS

- EMOS post-processing:

Non-parametric

- Post-processing possible results:





Benefits of non-parametric post-processing

No assumptions on the weather variable you deal with

- Raw ensemble:

EMOS

- EMOS post-processing:

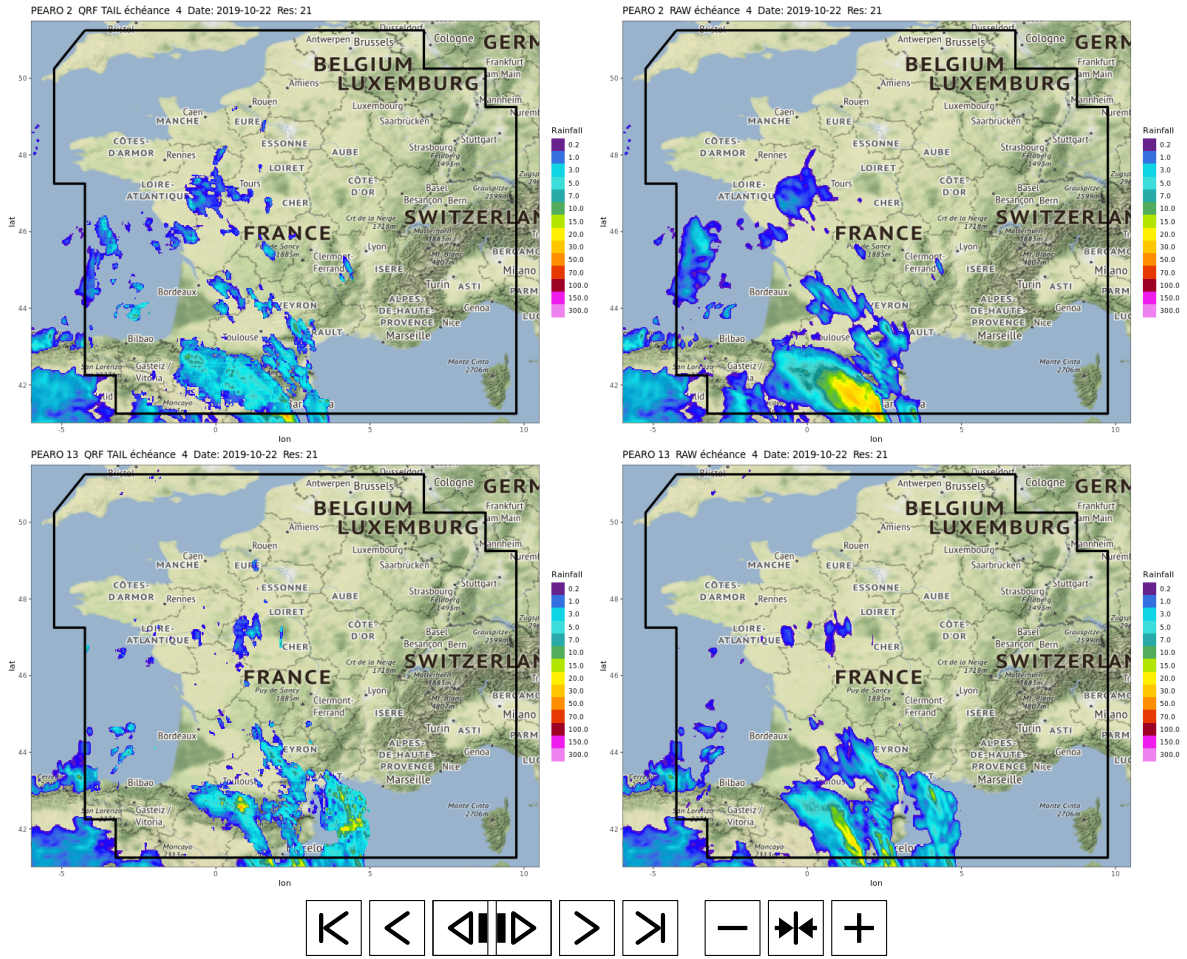
Non-parametric

- Post-processing possible results:





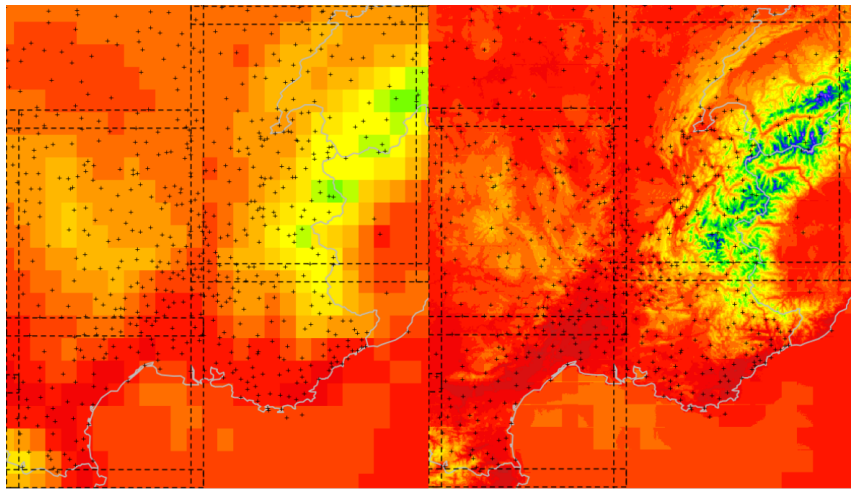
ECC + post-processing visualization



2 PP members (left) with their associated raw members (right)

Taillardat, M., Fougères, A. L., Naveau, P., & Mestre, O. (2019). Forest-based and semiparametric methods for the postprocessing of rainfall ensemble forecasting. Weather and Forecasting, 34(3), 617-634.

Downscaling & Spatialization - Example on EPS



Let k the kth station of a sub-domain D, i the associated target grid point (0.01°), and j the associated raw grid point (0.25°).
Let v be the validity time and S the season. For a given init. time and lead time :

- First (offline), compute a linear projection between each of 625 grid points T_i (coming from AROME) vs. their nearest raw grid point T_j:

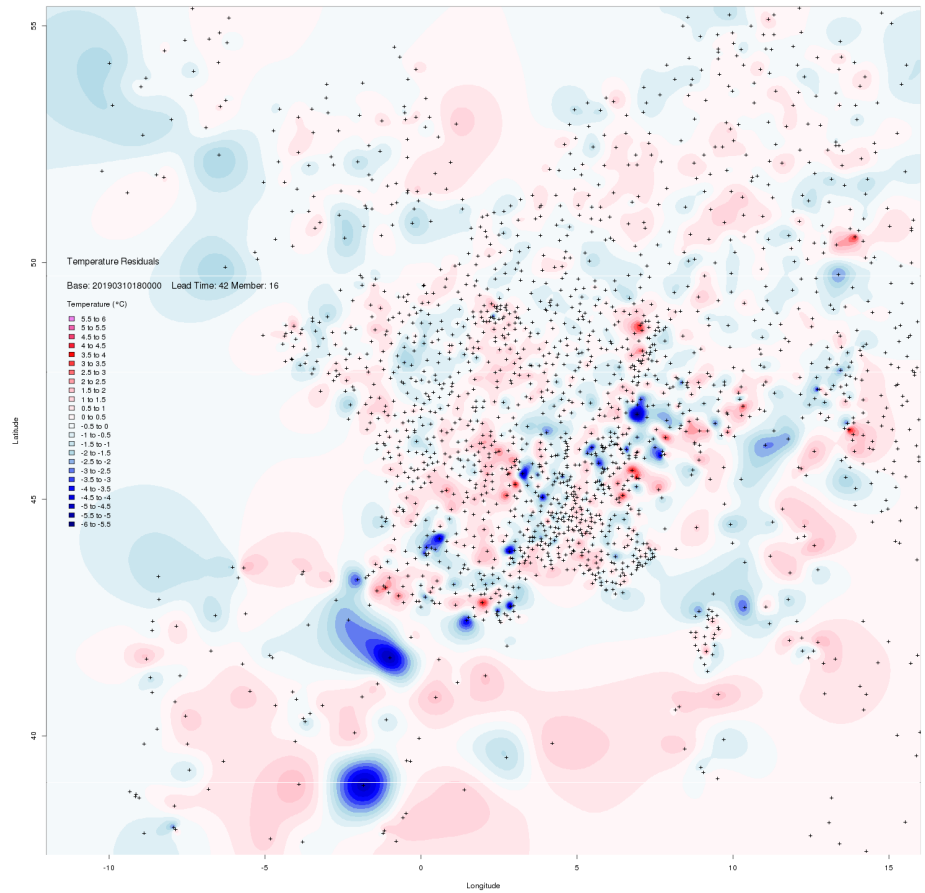
$$T_i = \gamma_{ijvS} + \eta_{ijvS} T_j$$

- Spatial trend estimation: Determine the coefficients (OLS with forward AIC) of the regression between the PP temperatures T_k and the "projected temperatures" T_i:

$$\begin{aligned} T_k &= \beta 0_D + \beta 1_D(\gamma_{ijvS} + \eta_{ijvS} T_j) \\ &+ \beta 2_D \text{alti}_i + \beta 3_D(\text{alti}_i - a*_D)\mathbf{1}\{\text{alti}_i > a*_D\} \\ &+ \beta 4_D \text{d2s}_i \\ &+ \alpha 1_D \text{PC1}_i + \alpha 2_D \text{PC2}_i + \alpha 3_D \text{PC3}_i + \alpha 4_D \text{PC4}_i \\ &+ \epsilon_k \end{aligned}$$

- Interpolation of the residuals ϵ_k using multiresolution B-splines

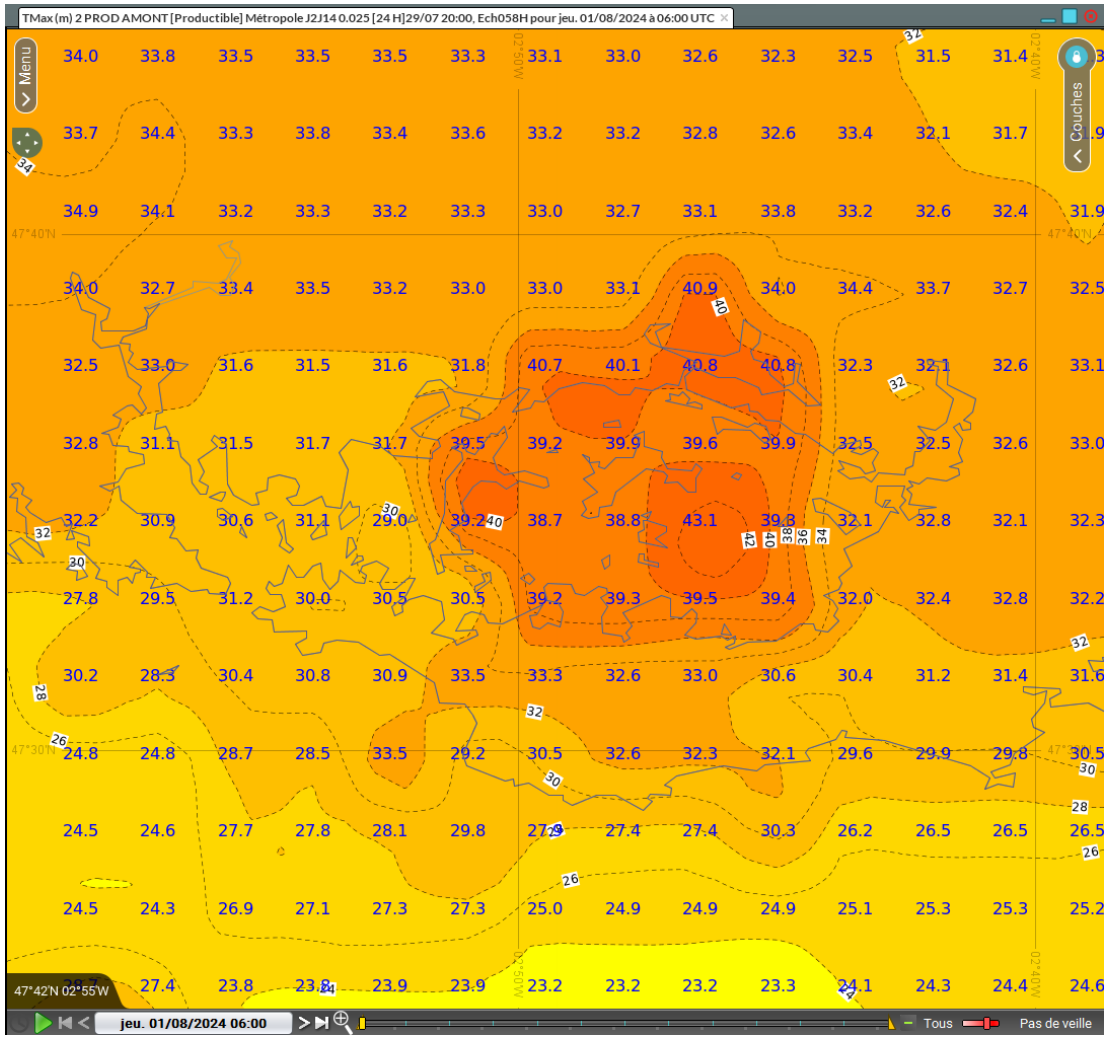
Downscaling & Spatialization - Example on EPS



Taillardat, M. and Mestre, O. (2020). From research to applications – examples of operational ensemble post-processing in France using machine learning, Nonlin. Processes Geophys., 27, 329–347.

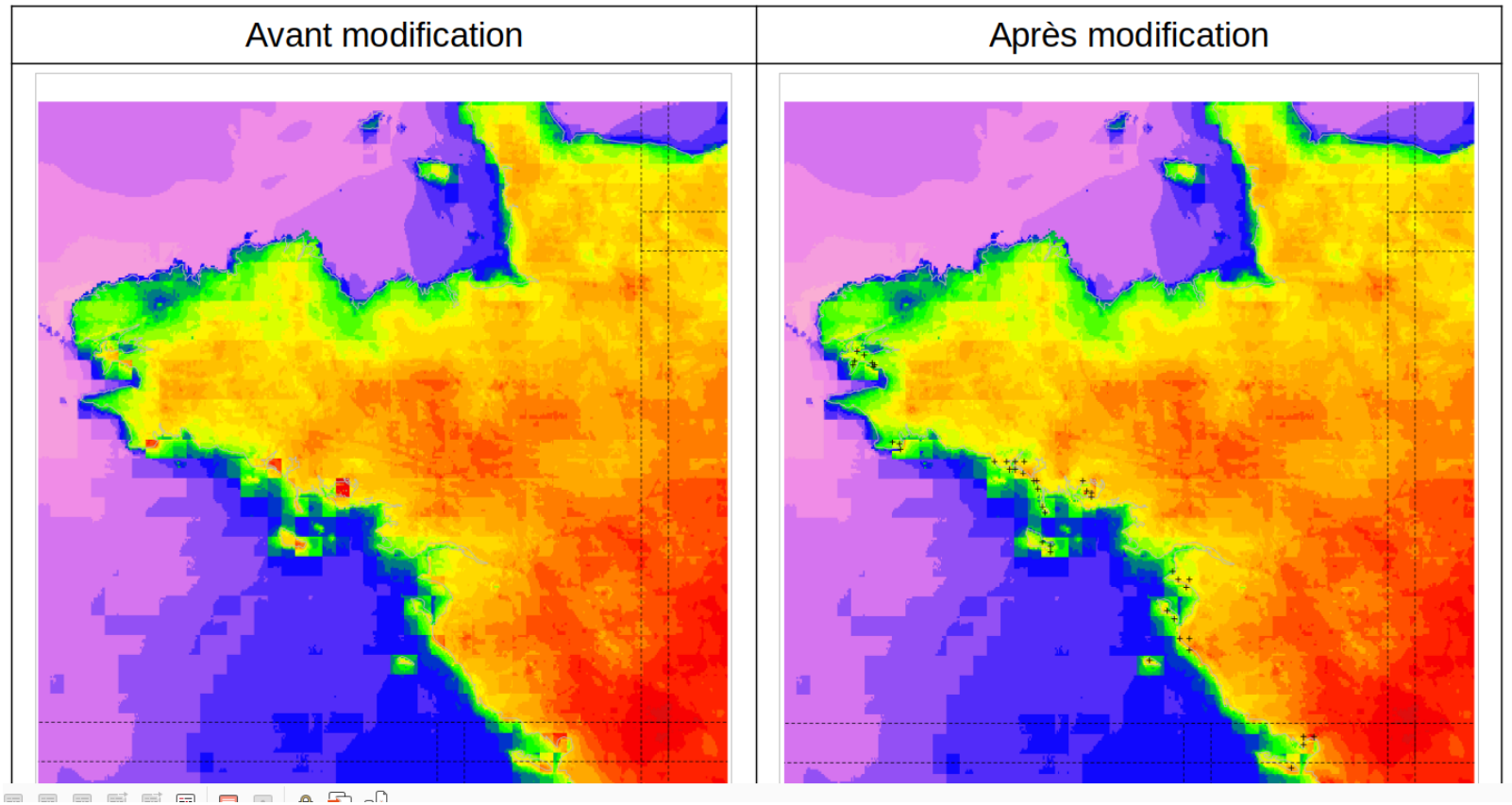
Coarse to fine grids : downscaling

43 °C in Brittany in summer ?



Coarse to fine grids : downscaling

Sea/land mask conflict



Use of predictable / easy to maintain techniques matters a lot.

PP via U-Net

- Widely-used architecture: see e.g. Grönquist et. al, Dai and Hemri (cGAN, G: UNet) (2021) ; Horat and Lerch, Hu et al. (2023) ; Ben Bouallègue et al. (2024).

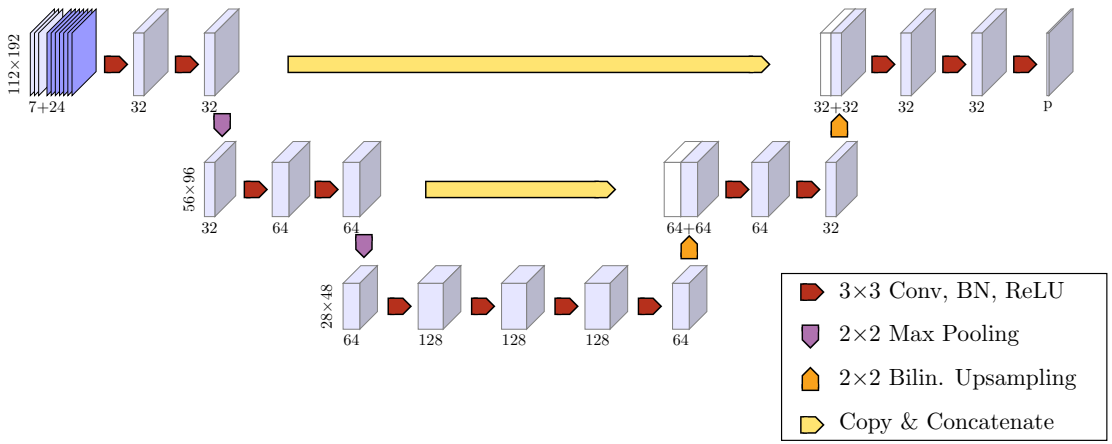


Figure: Architecture of distributional regression U-Nets. *Conv* stands for convolution, *BN* stands for batch normalization, *ReLU* stands for rectified linear unit and *Bilin. Upsampling* stands for bilinear upsampling. p is the number of distribution parameters: for GTCND and CSGD, $p = 3$.

$$F^{\text{gtcnd}}_{L,\mu,\sigma}(z) = \begin{cases} L + \frac{1-L}{1-\Phi(-\mu/\sigma)}\left(\Phi(\frac{z-\mu}{\sigma}) - \Phi(-\mu/\sigma)\right) & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases},$$

Dataset structure

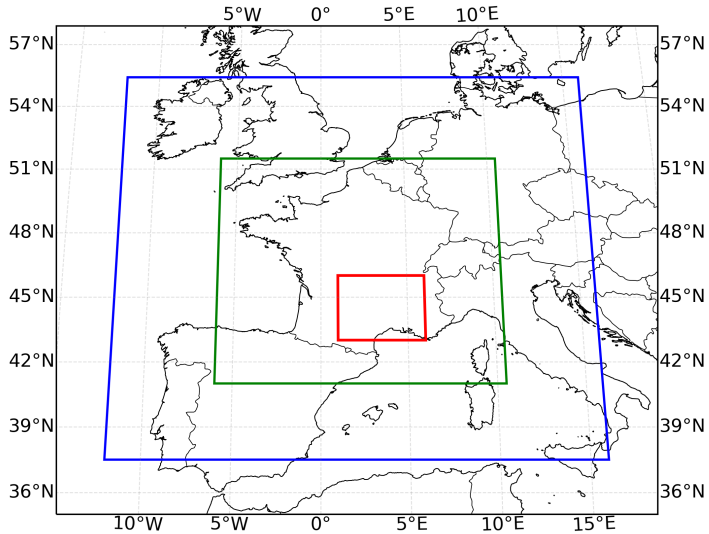


Figure: Domains covered by AROME-EPS (blue), ANTILOPE radar reference (green) and the region of interest (red).

Variable	Value	Description
d	31	number of predictors
H	112	height (in grid points) of the region of interest (latitude)
W	192	width (in grid points) of the region of interest (longitude)
$n_{trainval}$	1091	# of days in the training/validation dataset
n_{test}	365	# of days in the test dataset

Table: Dimensions of the dataset used in this study.

Predictors & methods

target: RR3, init. time 15Z, lead time 21h.

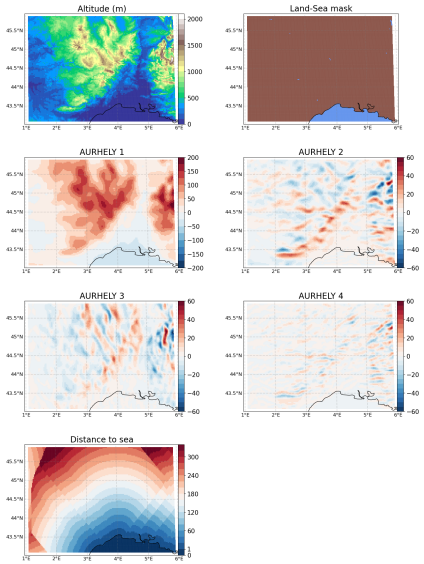


Figure: Constant fields.

Predictors (min,max, moy., sd.):

- RR3
- RFLCTVT MAX
- CAPE_INS
- TPW 850hPa
- HU 700hPa
- AROME convection index

Methods

- QRF / **TQRF** (Taillardat et al. ; 2016, 2019) : Oper Method. *Per* grid point. 15.3B parameters (data + tree splits). Package R ranger. 20 minutes on 32 of 128 AMD Rome (2.2GHZ) CPUs.
- DRU : Distributional U-Net. Whole grid. 1M parameters. Tensorflow/Keras. 10x7 minutes on 1 NVIDIA V100 GPU and 128 AMD Rome (2.2 GHz) CPUs

Event results (w/o. neighborhood)

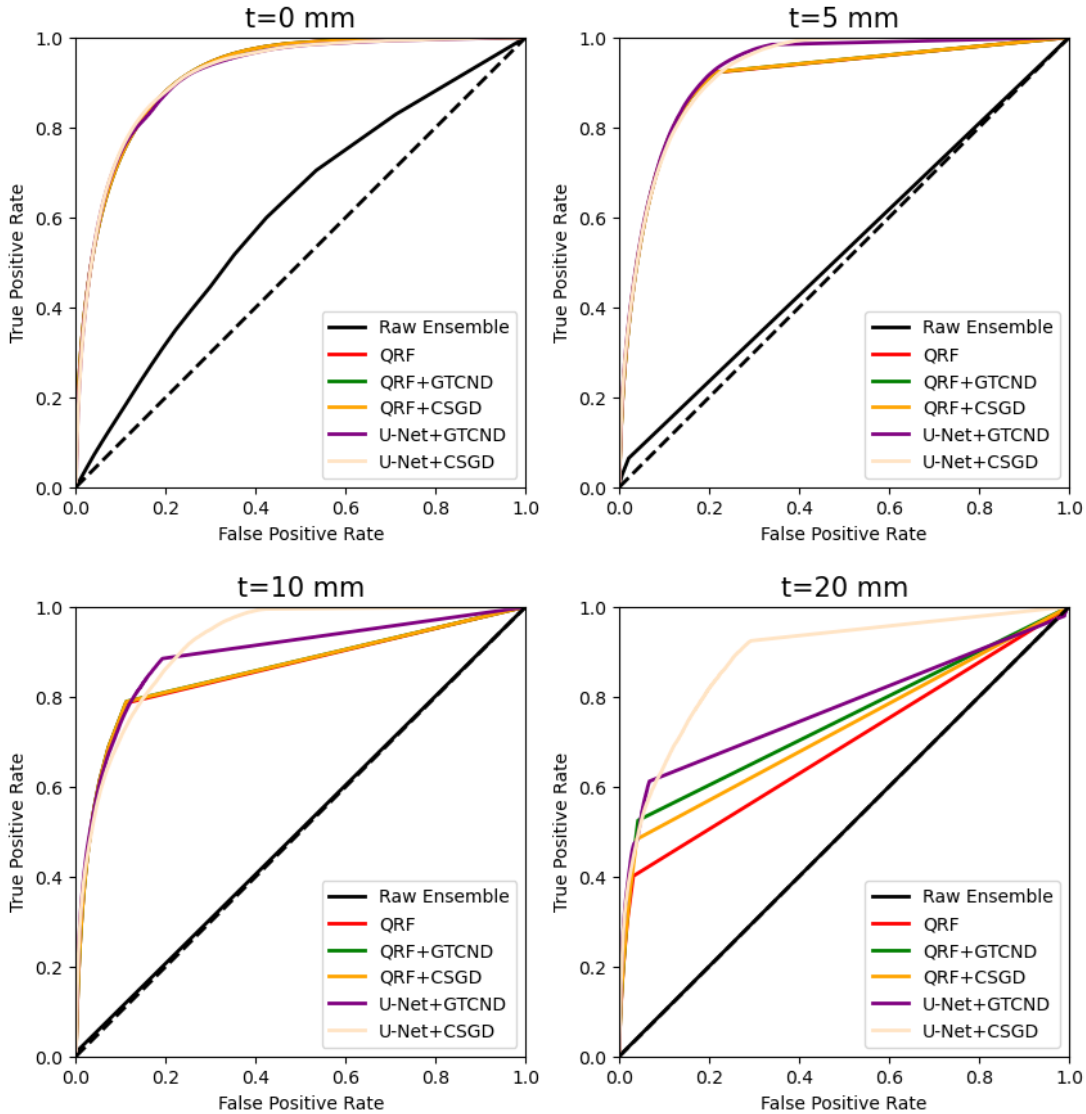
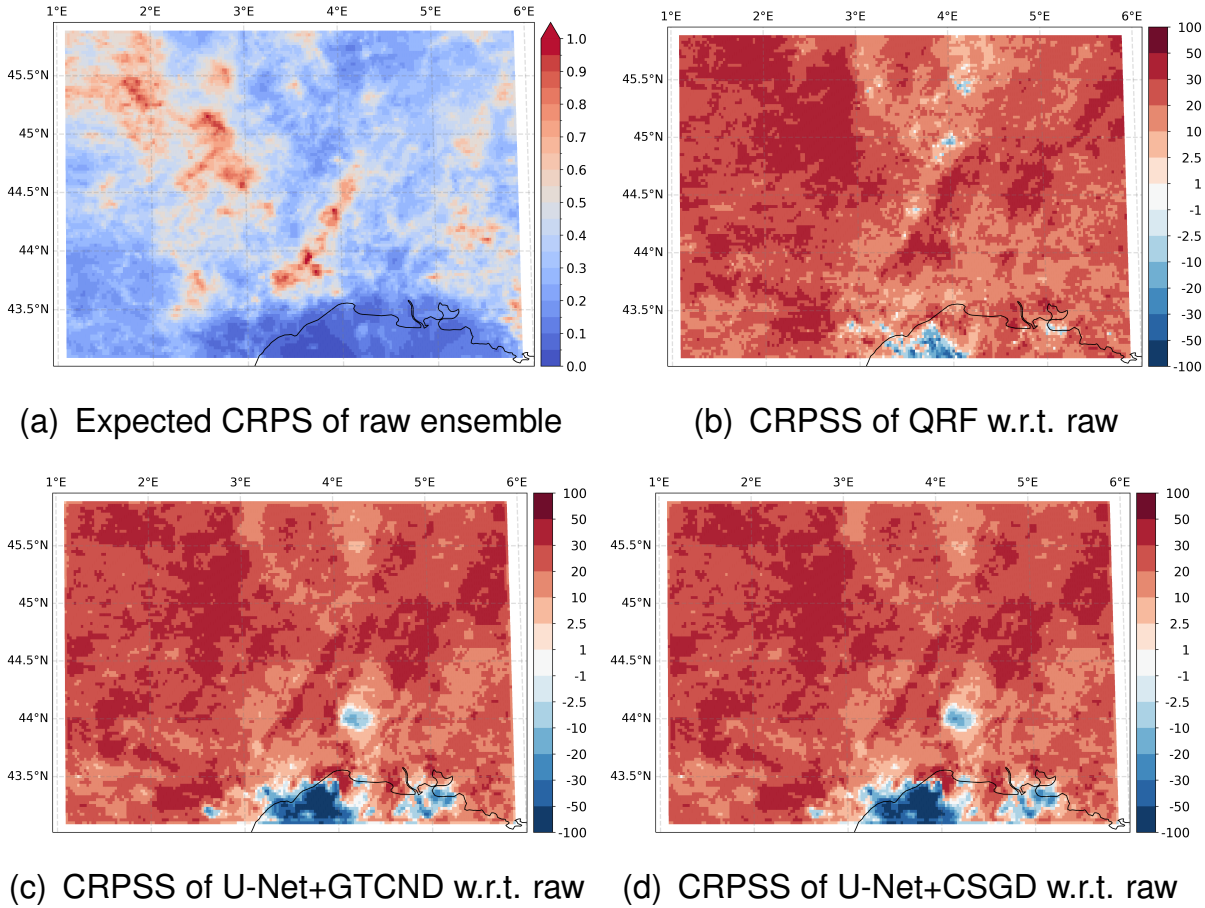


Figure: Receiver operating characteristic (ROC) curves of binary events corresponding to the exceedance of a threshold $t \in \{0, 5, 10, 20\}$ (in mm of precipitation).

Global results



(a) Expected CRPS of raw ensemble

(b) CRPSS of QRF w.r.t. raw

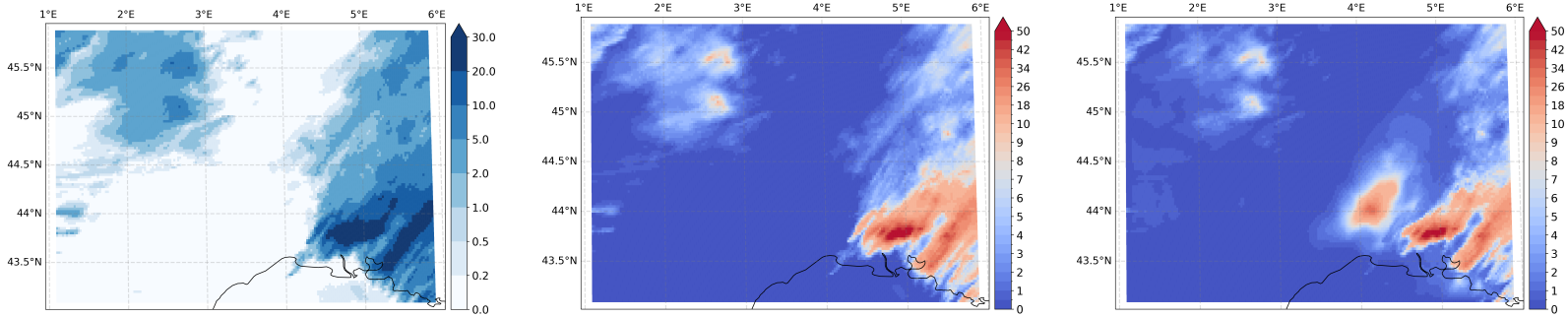
(c) CRPSS of U-Net+GTCND w.r.t. raw

(d) CRPSS of U-Net+CSGD w.r.t. raw

Figure: Predictive performance of the benchmark methods in terms of CRPS. (a) Expected CRPS of the raw ensemble, (b) CRPSS of QRF w.r.t. the raw ensemble and CRPSS w.r.t. RAW of (c) DRU+GTCND and (d) DRU+CSGD. CRPSS differences (T)QRF vs. DRU: 0.5%

$$CRPS(F, y) = \int_{-\infty}^{\infty} (F(x) - \mathbf{1}\{x \geq y\})^2 \, \mathrm{d}x = \mathbb{E}_F|X - y| - \frac{1}{2}\mathbb{E}_F|X - X'|$$

Numerical instability...



(a) Precipitation (mm)

(b) CRPS of raw ensemble

(c) CRPS of U-Net+GTCND

(d) Predicted L

(e) Predicted μ

(f) Predicted σ

Figure: Example of a numerical instability of U-Net+GTCND for a forecast valid on November 3, 2022 at 1500 UTC.

Conclusions

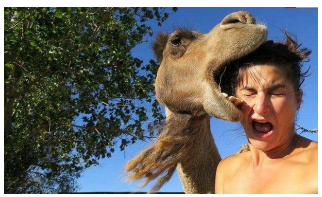
- Memory use/storage gain: x500
- Equal to current method w.r.t. performance
- ... But numerical issues ?
- Not enough data for the method ?

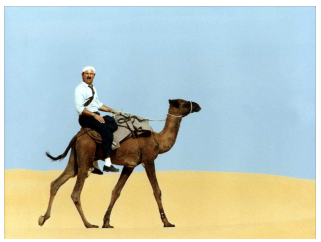
Pic, R., Dombry, C., Naveau, P., & Taillardat, M. (2025). Distributional regression U-Nets for the postprocessing of precipitation ensemble forecasts. Artificial Intelligence for the Earth Systems, 4(4), 240067.

Outline

How to make a point forecast from a distribution ?

It may depend on the reluctance to use probabilistic forecasts.





- Let's do forecast elicitation

What is elicitation ?

Elicitation

In Statistics, elicitation is the action of choosing a functional from a distribution F.

Fissler, T., & Ziegel, J. F. (2016). Higher order elicitability and Osband's principle. The Annals of Statistics, 44(4), 1680.

A F-functional is elicitable if there is a consistent loss function minimized by it

- The mean elicits the MSE : $\mathbb{E}_F[Y] = \text{argmin}_\theta \left(\mathbb{E}_F[\,(Y-\theta)^2\,]\right)$,
- The median elicits the MAE : $q^F_{0.5} = \text{argmin}_\theta \left(\mathbb{E}_F[\,|Y-\theta|\,]\right)$,
- The α-quantile of F elicits the Quantile Score$_\alpha$:
  $q^F_\alpha = \text{argmin}_\theta \left(\mathbb{E}_F[|\mathbf{1}_{\{\theta > Y\}} - \alpha||\theta - Y|\,]\right)$

Well-known quantities are not directly elicitable : variance, mode, expected shortfall...
In **weather forecasting**, ensemble forecast mean must minimize the MSE for example.

Rule (of thumb ?) for RRx (after D4)

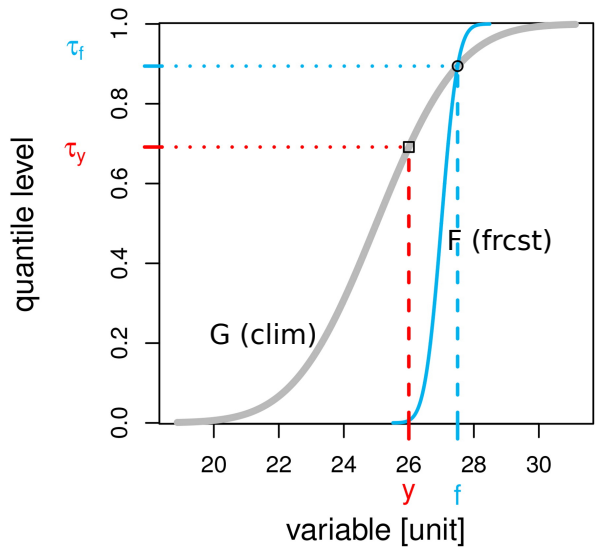
- If weather type "rain" under 50% among members : forecast = $0mm/xh$
- If not : Q70 of "wet" members.

We would like to rationale this rule with :

- an adaptative functional
- using the whole ensemble information
- minimizing a non-trivial score
- and weather-situtation dependent

What is needed ?

- The ensemble forecast (a conditional distrubtion)
- Its climate (the unconditional distribution)



The CPF τ_f is the point as $\tau_f = G(f) = F(f)$.
The CPF $T_G : F \to \tau_f$ is elicitable for the Diagonal Score...
(Closest point of the topleft point in ROC curves, for all events).

The Diagonal Score

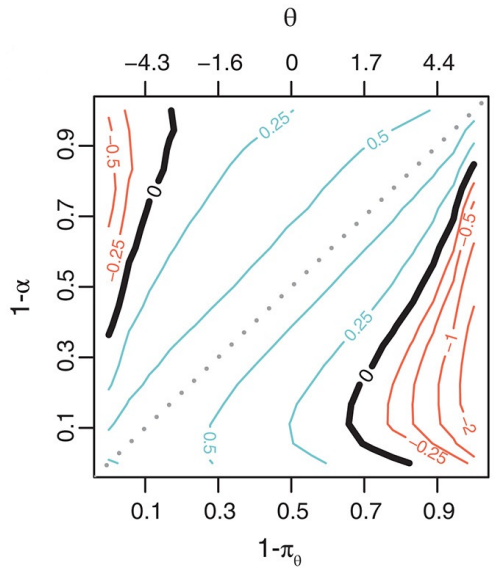
$T_G : F \to \tau_f$ is elicitable for the Diagonal Score (DS) : $\tau_f = \text{argmin}_\theta \left(\mathbb{E}_F[\, D_G(\theta, y)\,]\right)$

$$D_G(\theta, y) = 2\left[(\mathbf{1}_{\{\theta > G(y)\}} - G(y))^2 - (\mathbf{1}_{\{\theta > G(y)\}} - \theta)^2\right]$$

DS can be written with "elementary" scores

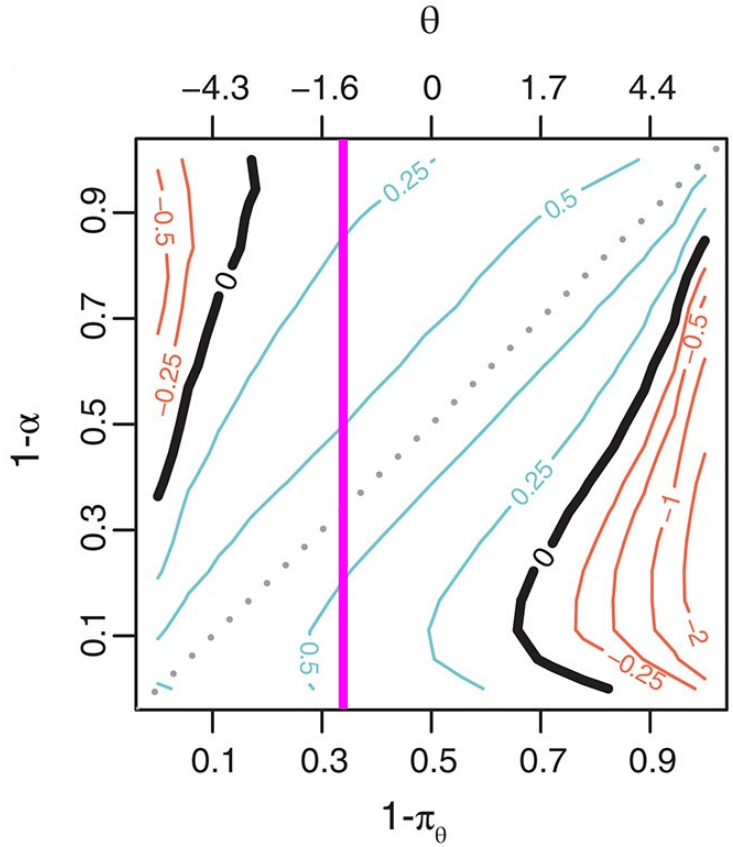
$$s_{\alpha,\theta}(F^{-1}(1-\alpha), y) = [1-\alpha]\mathbf{1}_{\{y > \theta \geq F^{-1}(1-\alpha)\}} + [\alpha]\mathbf{1}_{\{y \leq \theta < F^{-1}(1-\alpha)\}}$$

Forecast Skill Card



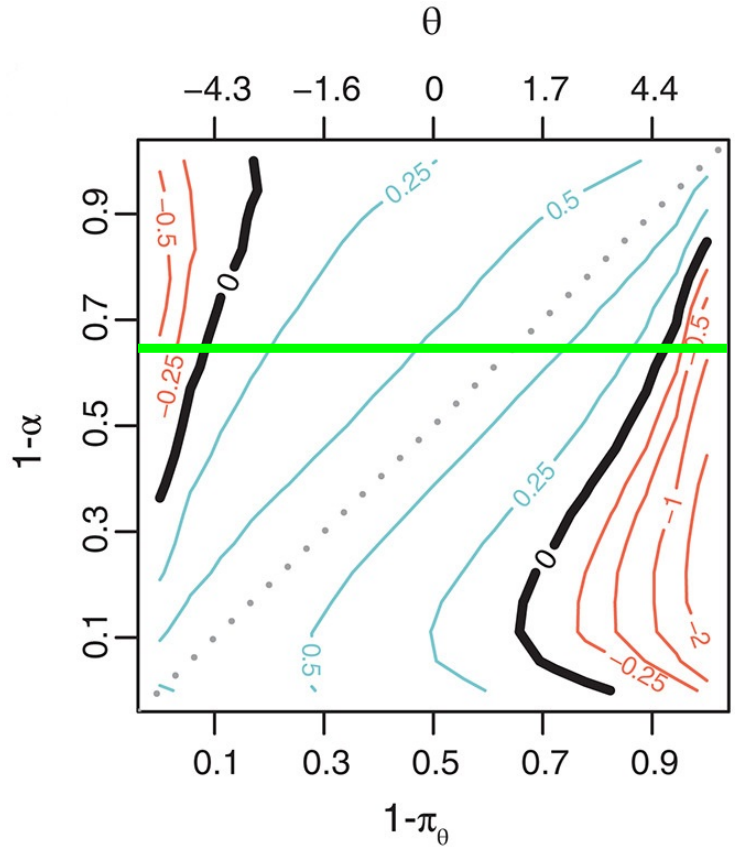
$$\pi_\theta = 1 - G(\theta)$$

Forecast Skill Card



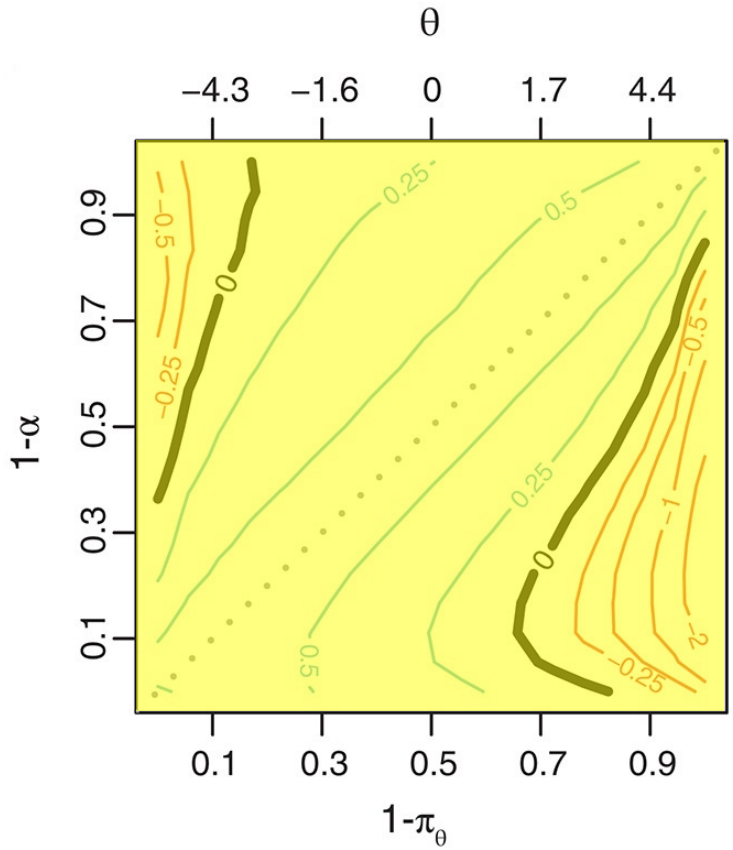
$$BS_{\pi_\theta}(F, y) = 2 \int_0^1 s_{\alpha,\theta}(F^{-1}(1-\alpha), y)\, \mathrm{d}\alpha$$

Forecast Skill Card



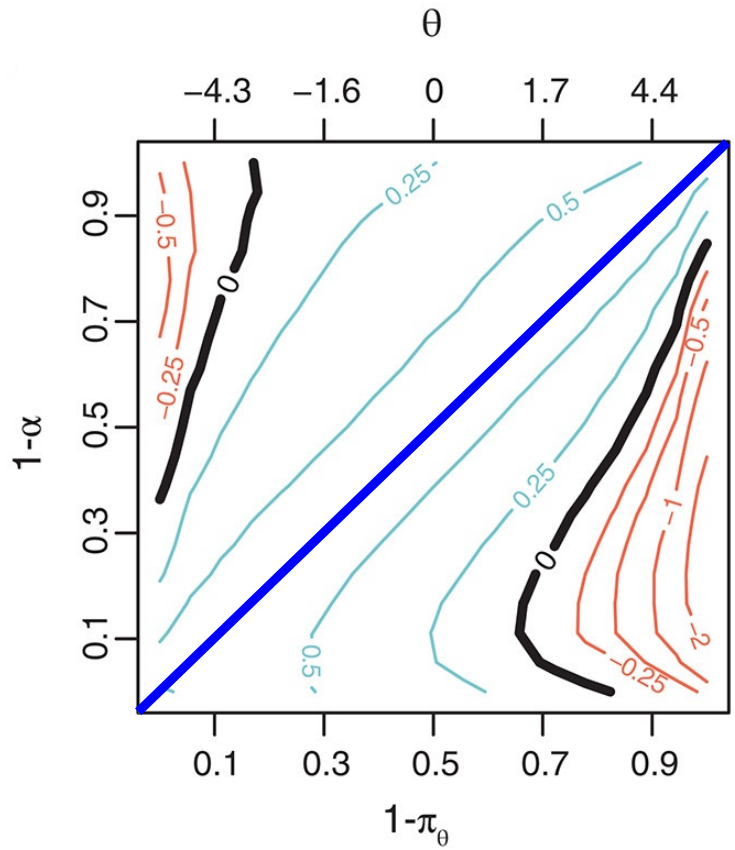
$$QS_\alpha(F, y) = \int_{\mathbb{R}} s_{\alpha,\theta}(F^{-1}(1-\alpha), y)\, \mathrm{d}\theta$$

Forecast Skill Card



$$CRPS(F, y) = 2 \int_{\mathbb{R}} \int_0^1 s_{\alpha,\theta}(F^{-1}(1-\alpha), y)\, \mathrm{d}\alpha\, \mathrm{d}\theta$$

Forecast Skill Card



$$DS(F, y) = 2 \int_{\mathbb{R}} \int_0^1 s_{\alpha,\theta}(F^{-1}(1-\alpha), y) \mathbf{1}_{\{\pi_\theta = \alpha\}} \, \mathrm{d}\alpha \, \mathrm{d}\theta$$

Forecast Skill Card

$$DS(F, y) = \int_0^1 s_{\alpha, G^{-1}(1-\alpha)}(F^{-1}(1-\alpha), y)\, \mathrm{d}\alpha$$

but :

$$\mathbb{E}_t[s_{\alpha, G^{-1}(1-\alpha)}(F^{-1}(1-\alpha), y)] = \alpha(1-\alpha)[1 - H_\alpha + F_\alpha]$$

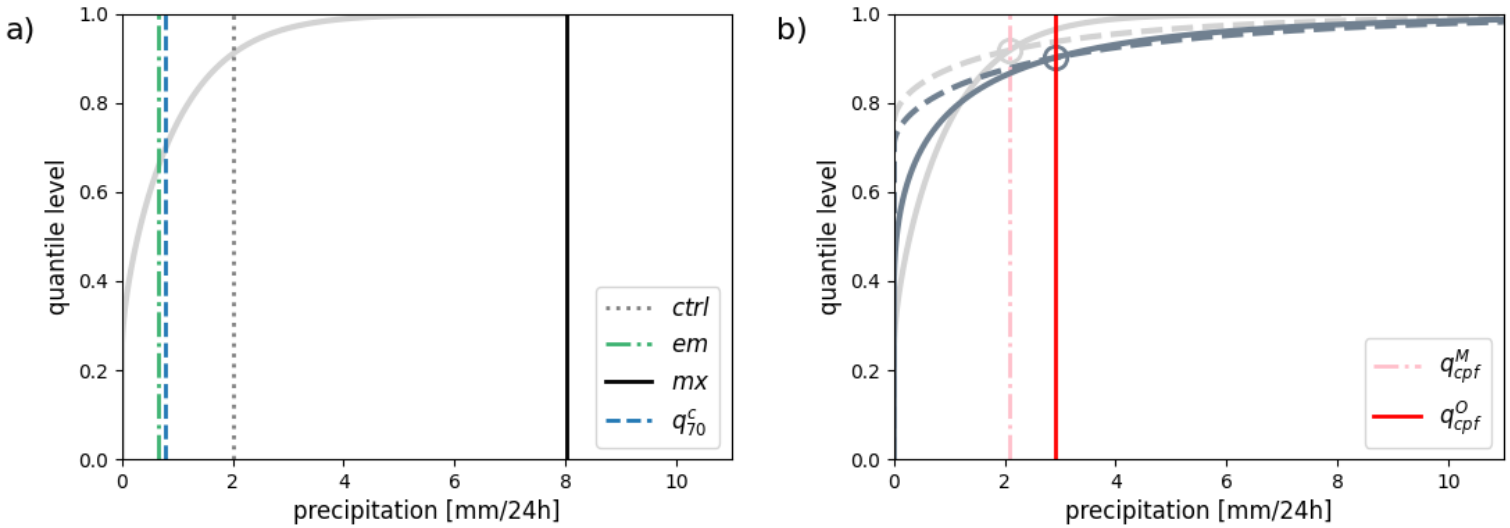
and

$$\mathbb{E}_t[s_{\alpha, G^{-1}(1-\alpha)}(G^{-1}(1-\alpha), y)] = \alpha(1-\alpha)$$

So the elementary diagonal skill score is the PSS for the event of base rate α :

$$K_\alpha(F, G) = 1 - \frac{\mathbb{E}_t[s_{\alpha, G^{-1}(1-\alpha)}(F^{-1}(1-\alpha), y)]}{\mathbb{E}_t[s_{\alpha, G^{-1}(1-\alpha)}(G^{-1}(1-\alpha), y)]} = H_\alpha - F_\alpha$$

Qualitative evaluation



Benchmark

- ctrl : deterministic forecast IFS
- mx : maximum of EPS
- em : mean of EPS
- $q^c{}_{70}$: operational forecast
- $q^M{}_{cpf}$: $G^{-1}(CPF)$ with $G = G_{EPS}$ (dashed light grey)
- $q^O{}_{cpf}$: $G^{-1}(CPF)$ with $G = G_{obs}$ (dashed dark grey) and EPS post-processed (after calibration M-climate = obs.)

G_{obs} : climate : COMEPHORE radar + rain gauge product in RR24 2000-2020.

Illustration : August 14, 2024

Crossing-point quantiles (q^{O}_{cpf}) for consecutive runs

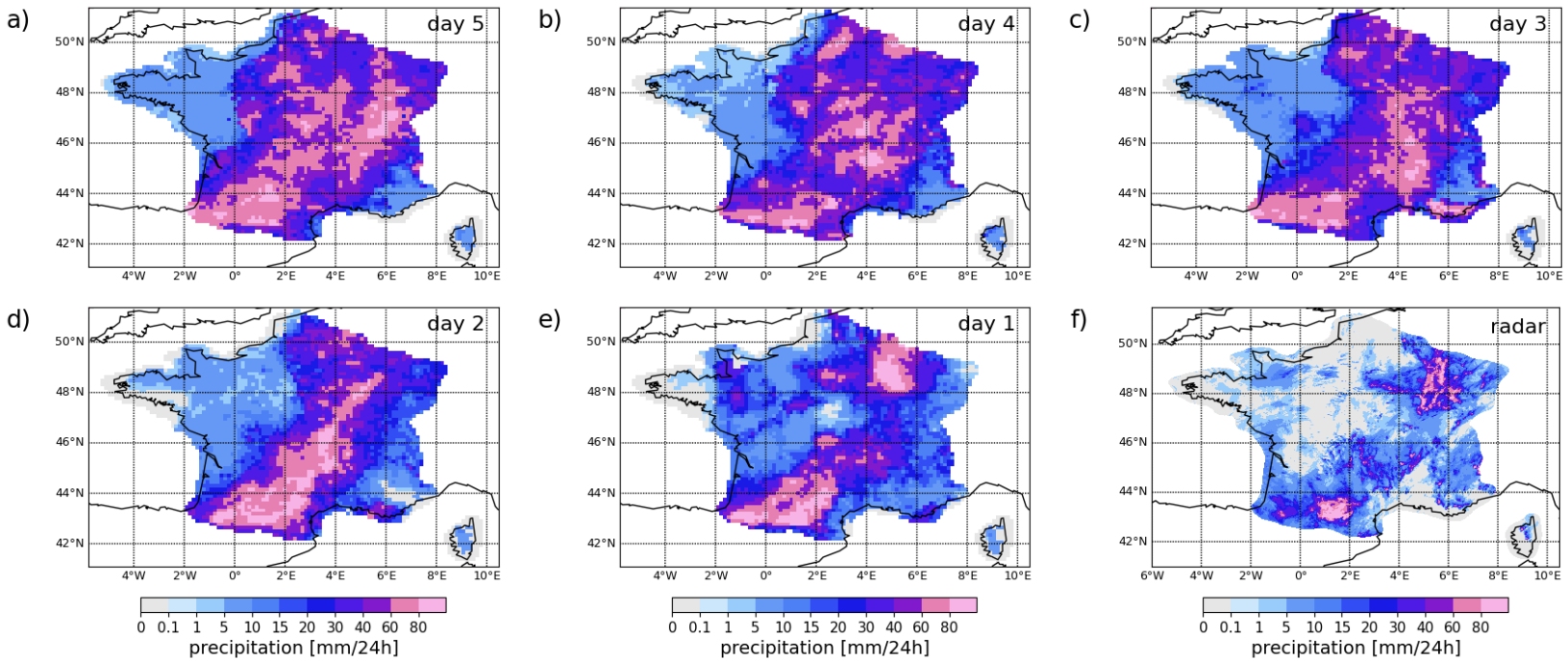


Illustration : August 14, 2024

Point-forecasts at day 5

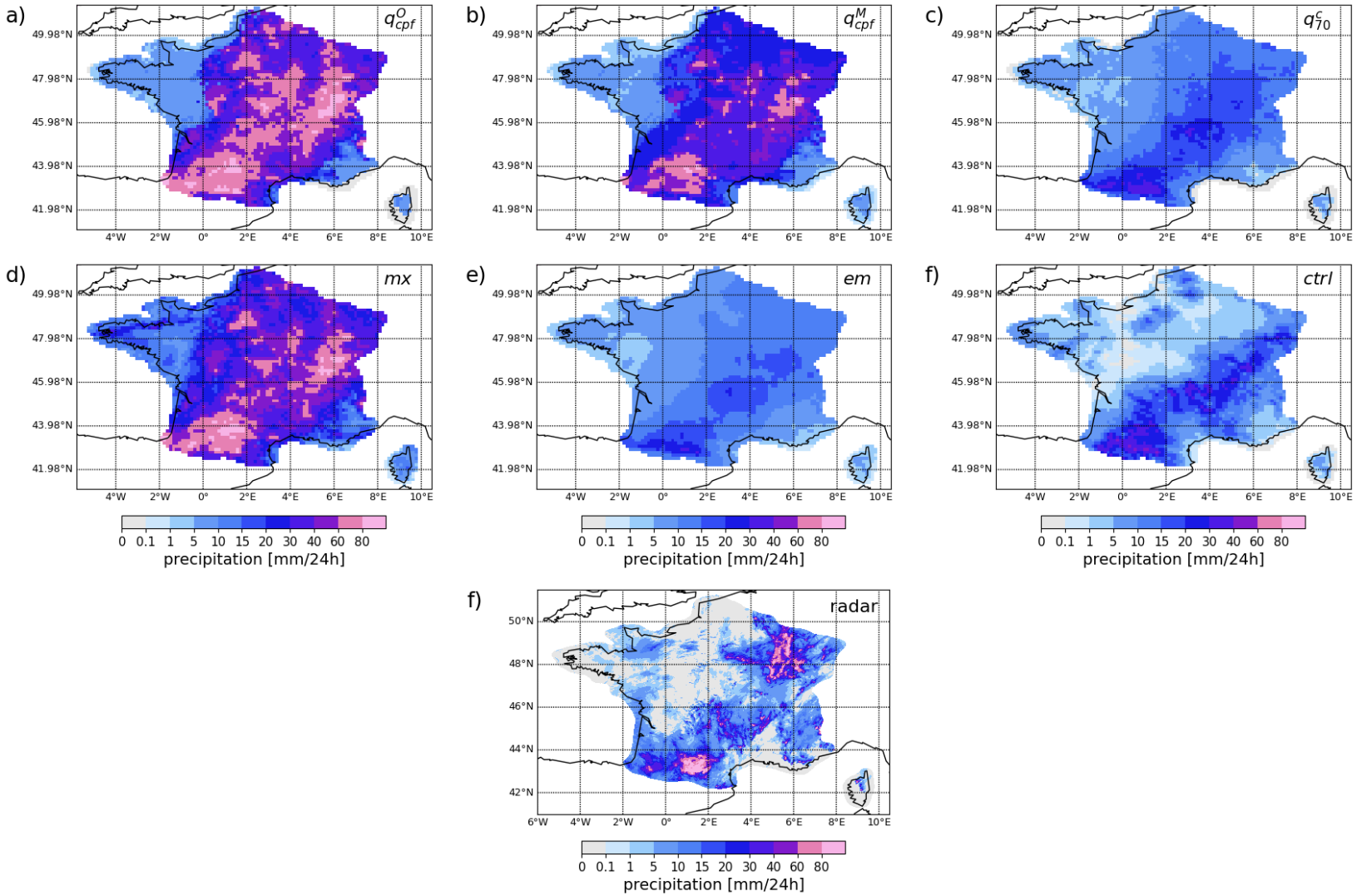


Illustration : July 26, 2024

Point-forecasts at day 5

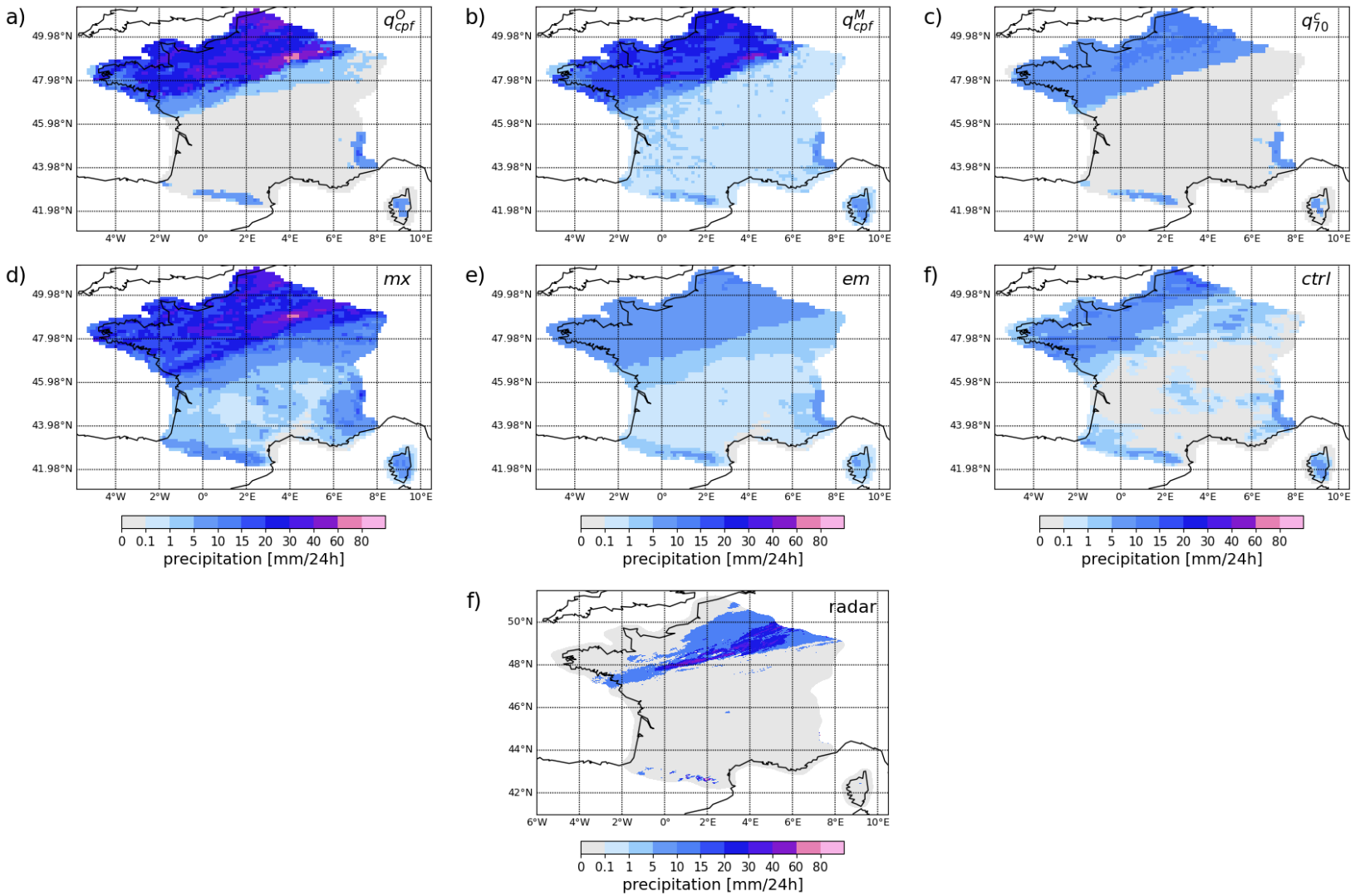
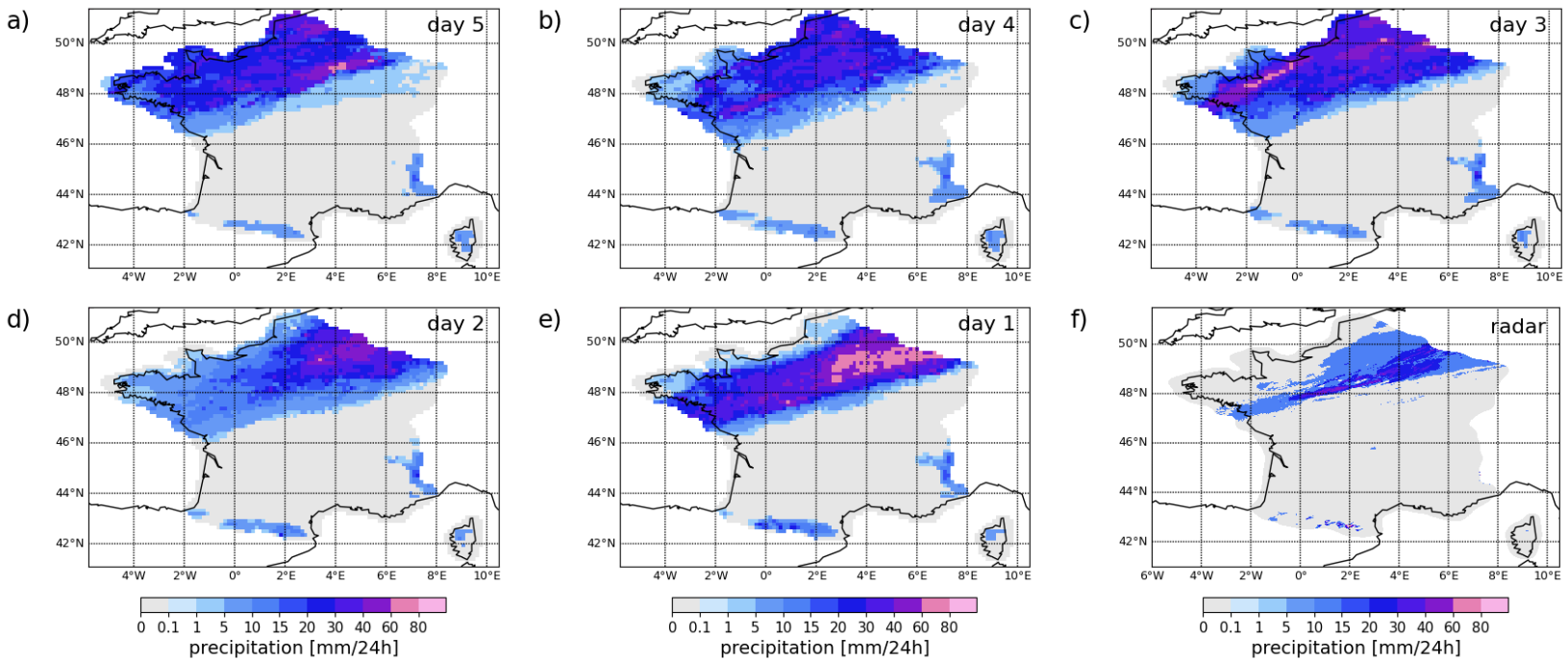
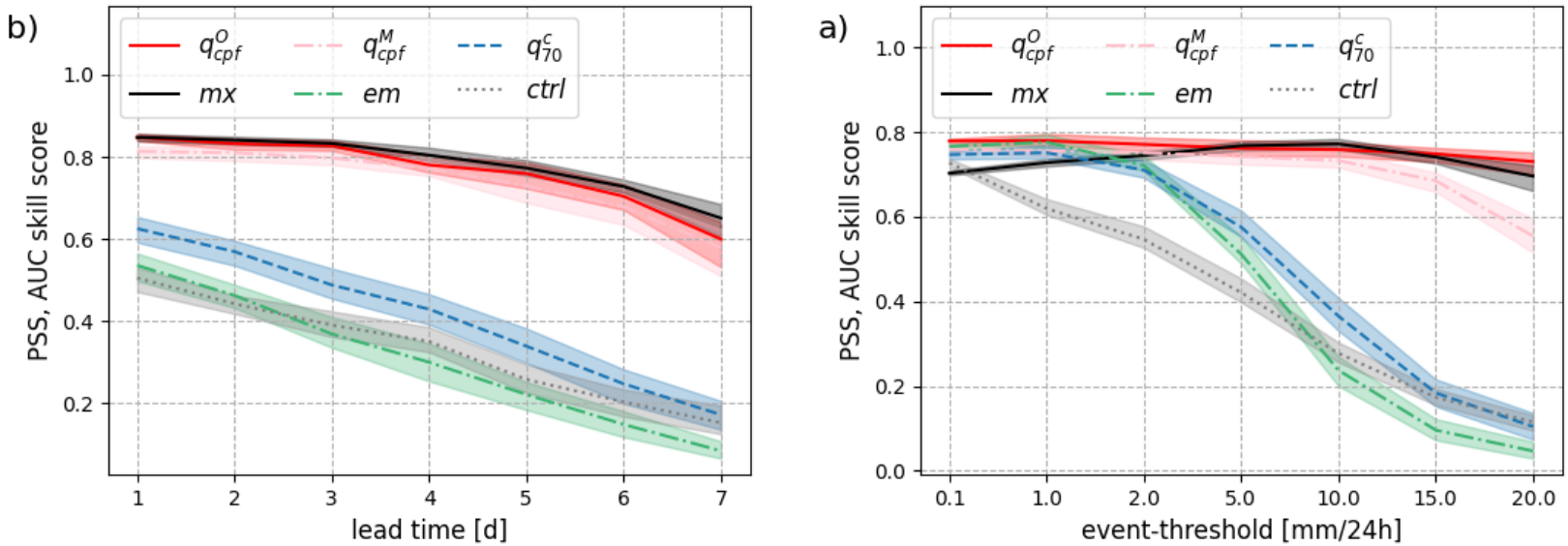


Illustration : July 26, 2024

Crossing-point quantiles for consecutive runs

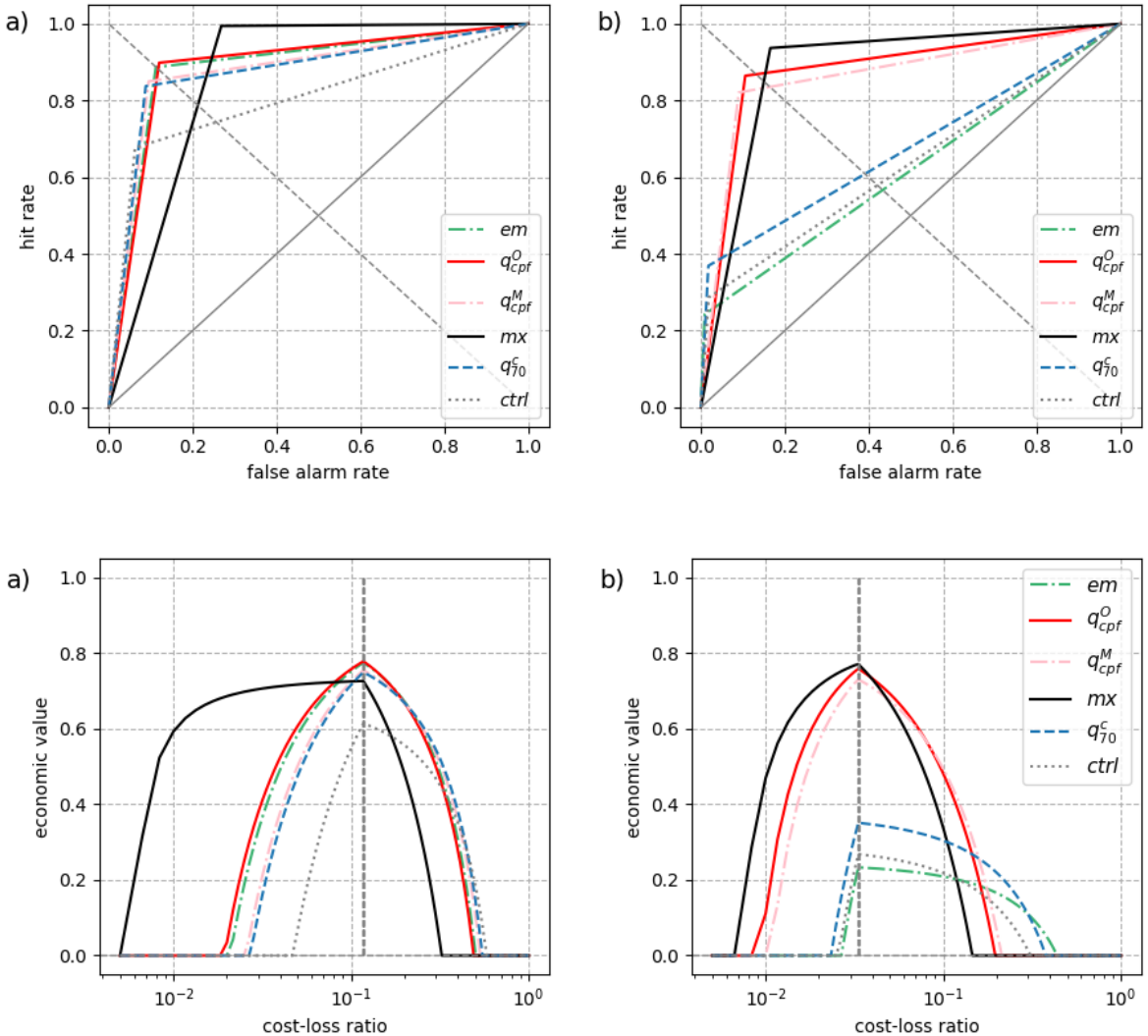


Numerical Evaluation (summer 2022 to 2024)



Left : PSS $10mm/24h$; Right : PSS at D5.

ROCs and potential economic value



Left : $1mm/24h$; Right : $10mm/24h$.

Conclusions

- CPF quality depends on the quality of the underlying EPS
- Self-adaptive-quantile are well-suited for users whose C/L ratio is unfocused
- But also for "low probability, high-impact" events
- Diagonal Score could be used as loss function for "non Gaussian" point estimates

Bouallègue, Z. B., & Taillardat, M. (2025). Self-Adaptive Quantiles for Precipitation Forecasting. Tellus A, 77(1), 160-172.

Conclusions & thoughts

- PP is squeezed between NWP and automation/final production ; both mostly gridded
- A research/operational job in National Weather Centers going to be "data provider" or "companion" of Universities or IT.
- UQ is a day-to-day task
- Even if point estimates are still needed...
- Forecast consistency is as important as forecast quality.
- Do we have enough (homogeneous) data in PP (vs. AI-NWP) for the new algorithms to come ?
- PP appears to be necessary even for AI-NWP w.r.t. local effects and extreme events.