

Mixtures of (skewed) Gaussian distributions for statistical post-processing

Maxime Taillardat

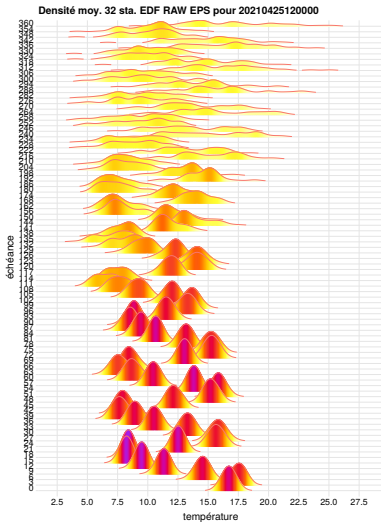
Météo-France / CNRM UMR 3589

maxime.taillardat@meteo.fr

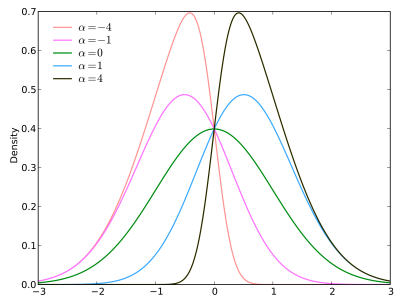
EGU Meeting 2021 - 4/28/2021



Why mixtures/skewed distributions for PP ?



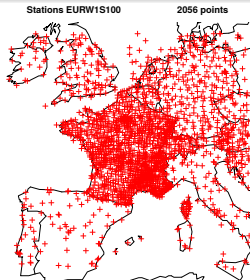
Skew-Normal Distribution



$$f(x) = 2\phi(x)\Phi(\alpha x)$$

Why ?

- ▶ Few EPS predictors available : Only statistics on the T2m distribution (ctrl, mean, std. dev., skewness)
- ▶ A lot of lead times / stations to calibrate
- ▶ A "light" method compared to data-driven techniques
- ▶ A well-documented technique, a lot of work already done on parametric distribution, optimization, inference, training periods...



Densities

- ▶ Normal :

$$\mathcal{N}(\mu, \sigma) \longrightarrow \frac{1}{\sigma} \phi\left(\frac{x - \mu}{\sigma}\right)$$

- ▶ Skew-Normal :

$$\mathcal{SKN}(\xi, s, \alpha) \longrightarrow \frac{2}{s} \phi\left(\frac{x - \xi}{s}\right) \Phi\left(\alpha \left(\frac{x - \xi}{s}\right)\right)$$

- ▶ Normal+Normal mixture :

$$\omega \mathcal{N}(\mu_{\overline{ENS}}, \sigma_{ENS}) + (1 - \omega) \mathcal{N}(\mu_{CTRL}, \sigma)$$

- ▶ Skew-Normal + Normal mixture :

$$\omega \mathcal{N}(\mu_{\overline{ENS}}, \sigma_{ENS}) + (1 - \omega) \mathcal{SKN}(\xi_{CTRL}, s, \alpha_{ENS})$$

Algorithms trained with LogS and/or CRPS if possible.

- ▶ analytical formulas of CRPS available : SKN, N, N+N Mix, or for any type of mixture CDFs F and G :

$$CRPS(\omega F + (1-\omega)G, y) = \omega^2 CRPS(F, y) + (1-\omega)^2 CRPS(G, y) + 2\omega(1-\omega) \int (\mathbf{1}_{\{x > y\}} - F)(\mathbf{1}_{\{x > y\}} - G) dx$$

- ▶ No closed-form expression of max. likelihood for SKN, SKN+N
- ▶ Constrained parametrization ?
- ▶ LogS : bias vs. variance overfit ?

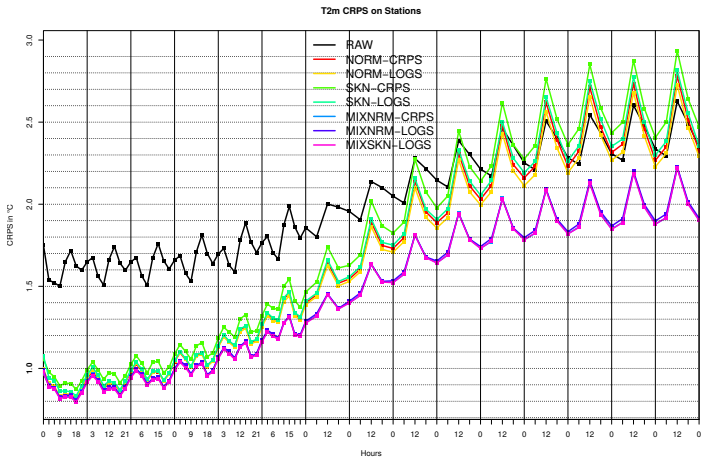
Use of link functions (for std. dev. or weight), e.g. :

$$\sigma_{ENS} = \frac{0.1 + 2e^a}{1 + e^a} + \widehat{sd_{ENS}} \frac{0.8 + 1.3e^b}{1 + e^b}$$

- ▶ Observations available on ~ 2000 stations locations across Western Europe
- ▶ Raw model : EPS from 08/2017 to 07/2020, interpolated on stations
- ▶ 2-fold cross validation
- ▶ training by season : DJF, MAM, JJA, SON.

Next, results are shown for the initialization time 0Z.

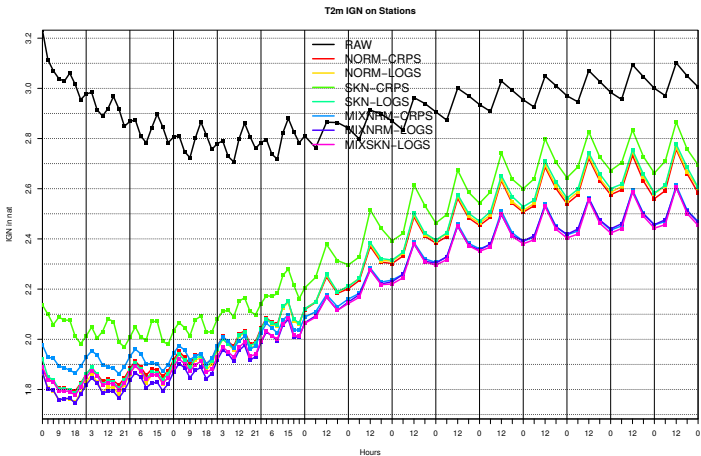
Performance on stations : CRPS averages



Superiority (in $\overline{CRPS}$) of MIX distributions. No significant gain of unimodal PP after day 9. After day 10, unimodal PP -LOGS seems to outperform CRPS optimization.

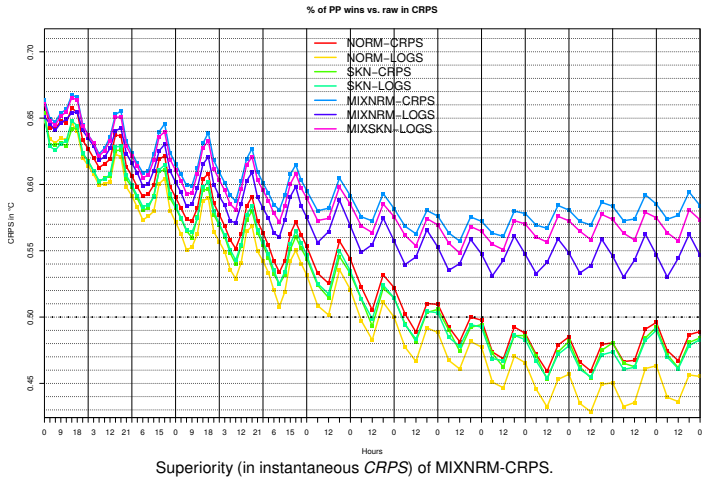
Performance on stations : LogS averages

LogS : Derivation of "Fractional Ignorance" of Roulston & Smith (2002)

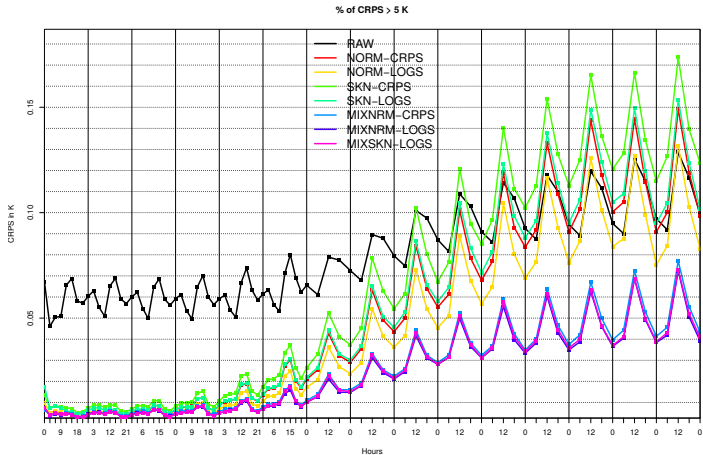


Superiority (in $\overline{\text{LogS}}$) of MIX distributions. Strange shape of RAW on days 2 to 8. SKN-CRPS the worst PP here.
Under-performance of MIXNRM-CRPS up to day 3.

Performance on stations : CRPS dominance

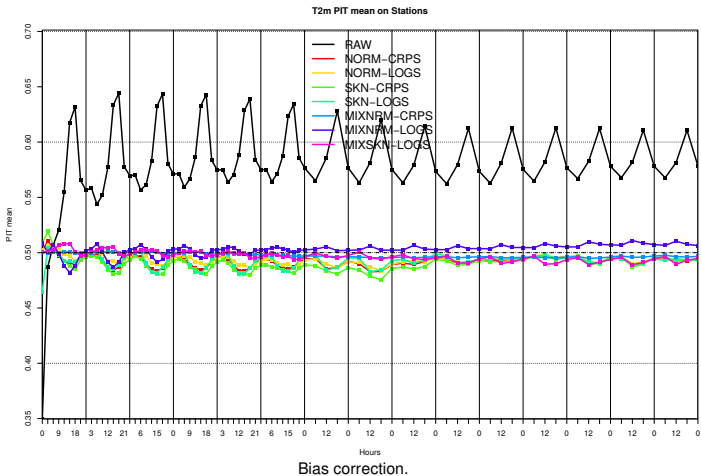


Performance on stations : "big fail" frequency

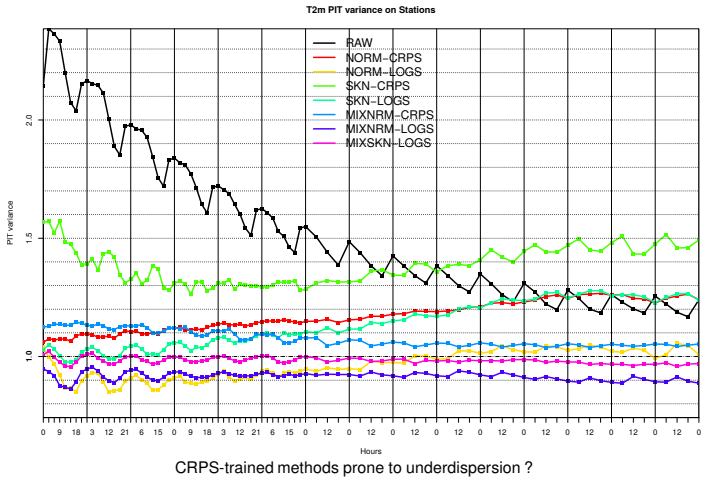


Headline index of the ECMWF. Same comments than $\overline{CRPS}$.

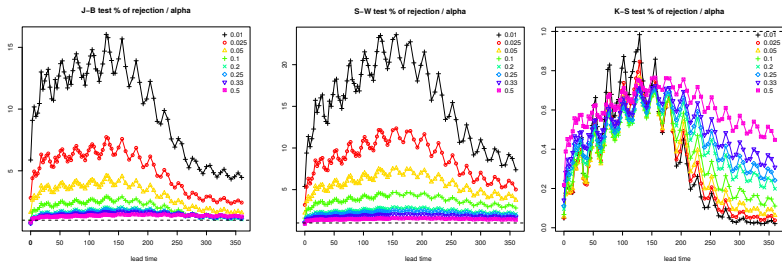
Performance on stations : PIT mean



Performance on stations : PIT variance



Normality tests on RAW distributions



Jarque-Bera (left), Shapiro-Wilk (center), Kolmogorov-Smirnov (right) ratios of rejected tests w.r.t. type I error. Is there a link with RAW Ignorance Score $\overline{LogS}$?

Summary

Why mixtures/skewed distributions for EMOS PP ?

- ▶ Few EPS predictors available : Only statistics on the T2m distribution (ctrl, mean, std. dev., skewness)
- ▶ A "light" method compared to data-driven techniques for a "shallow" dataset
- ▶ Try to keep modalities between *CTRL* and $\overline{ENS}$

Densities benchmarked with LogS or/and CRPS optimization

- ▶ Skew-Normal (Normal iff $\alpha = 0$) :

$$SKN(\xi, s, \alpha) \longrightarrow \frac{2}{s} \phi\left(\frac{x - \xi}{s}\right) \Phi\left(\alpha \left(\frac{x - \xi}{s}\right)\right)$$

- ▶ Normal+Normal mixture :

$$\omega \mathcal{N}(\mu_{\overline{ENS}}, \sigma_{ENS}) + (1 - \omega) \mathcal{N}(\mu_{CTRL}, \sigma)$$

- ▶ Skew-Normal + Normal mixture :

$$\omega SKN(\mu_{\overline{ENS}}, \sigma_{ENS}) + (1 - \omega) SKN(\xi_{CTRL}, s, \alpha_{ENS})$$

Messages

- ▶ Closed-form expressions of CRPS have been derived
- ▶ Superiority of mixture distributions vs. unimodal vs. raw ; increasing w.r.t. lead time
- ▶ Algorithms prone to overfit : use of pseudo-sigmoid link functions
- ▶ Sensitivity of scoring rule used for optim. vs. scoring rule used for verification
- ▶ Skew-Normal : probably better for variables such as wind gusts, precipitation...
- ▶ Article to come (soon I hope...)