Local forecast of probability distributions by statistical adaptation

S. FARGES (serge.farges@meteo.fr)
DPREVI/GCRI
April 05 2012

METEO FRANCE
Toujours un temps d'avance

Introduction

In an atmospheric model, coasts, orography, soil type and vegetation status are not perfectly well represented. In addition for computation purposes that space is discretized ; by a way of consequence **a statistical post-processing is necessary to account for the local climatology of the site (station)**.

This is called **statistical adaptation** of numerical model outputs (SA). Meteo-France currently produces approximately **3.5 billion forecasts using SA per day** (for a station, a parameter and time range given).

Uncertainty is present since the start of the forecast covered time range. **We want to increase the use of a probabilistic approach**, including, if possible, by forecasting the distributions of the parameters. **We have several methods available to us**, as we shall see...

Forecast of a probabilistic distribution : different ways

1. Dynamic approach : the ensemble forecast

- The ensemblist systems (varEPS, NCEP, PEARP...) :
  - Advantages : realizations of trajectories and maps, extreme phenomena forecasting (if enough members).
  - Disadvantage : cost ($\Rightarrow$ limited resolution).
- The multi-models systems (IFS + ARPEGE + ...) :
  - Advantages : realizations of trajectories and maps, quality of the deterministic forecast in the short time range.
  - Disadvantages : more time to collect the data from the various producers and limited number of ensemble members.
- The multi-ensemblists systems (TIGGE) :
  - Advantages and disadvantages of previously described approaches.

Often require post-processing statistics (because the produced probabilities are often unreliable) : Ensemble Dressing, Bayesian Model Averaging (BMA), Nonhomogenous Gaussian Regression (NGR) or Ensemble Regression.

2. Statistical approach : probabilistic statistical adaptation of an atmospheric model

- Discrimination models (LDA, logistic regression, neural networks...) :
  - Advantages : cost and robustness of linear methods.
  - Disadvantage : production of occurrence probabilities but no production of distribution.
- Generalized Linear Models :
  - Advantages : cost and possibility to use the underlying probability distributions.
  - Disadvantages : limited number of supported distributions (exponential family) not always perfectly calibrated after fitting.

Forecast of a probabilistic distribution : different ways

Proposed new approach : *the generalized linear regression by Gaussian anamorphosis*

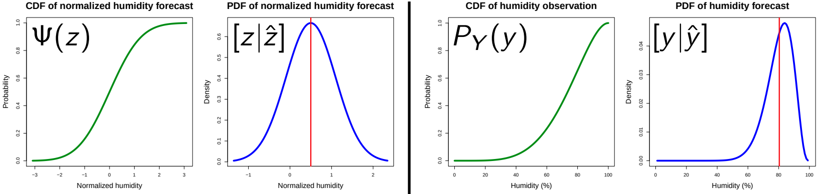
- Advantages : cost, robustness (linear model), no constraint on probabilistic distributions and strongly calibrated (in theory). Can be used for the postprocessing of an ensemble forecast coupled with a BMA.
- Disadvantage : generally no direct computable formulation of deterministic forecast (expectation).

Remark : statistical methods have the disadvantage of not being able to easily produce realistic realizations of trajectories or map.

1. Methodology (beta distribution for humidity)

Normalized space (z) | Initial space (y)

$\Phi(y)$



$\Phi^{-1}(z)$

$$\hat{z} = \sum_{p=1}^{P} a_p x_p \quad (\epsilon_{\hat{z}} = \hat{z} - z)$$

$$z|\hat{z} \sim \mathcal{N}(\hat{z}, \sigma_{\epsilon_{\hat{z}}})$$

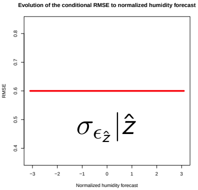
$$\hat{y} = \int_{-\infty}^{+\infty} \Phi^{-1}(z)\varphi(z, \hat{z}, \sigma_{\epsilon_{\hat{z}}})dz$$

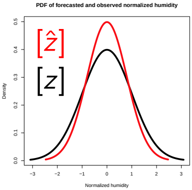
$$[y|\hat{y}] = [y]\frac{\varphi(\Phi(y), \hat{z}, \sigma_{\epsilon_{\hat{z}}})}{\varphi(\Phi(y))}$$

$$\Phi(y) = \Psi^{-1} \circ P_Y(y)$$

Generalized linear regression by Gaussian anamorphosis

Normalized space (z)



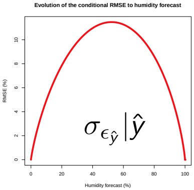


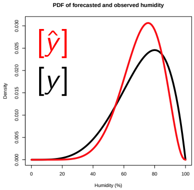
$$\sigma_z^2 = \sigma_{\hat{z}}^2 + \sigma_{\epsilon_{\hat{z}}}^2$$

$$\text{corr}(\hat{z}, \epsilon_{\hat{z}}) = 0 \text{ et } \mathrm{E}(\epsilon_{\hat{z}} | \hat{z}) = 0$$

- Independance between $\hat{z}$ and $\epsilon_{\hat{z}}$
- Homoscedasticity
- Probabilistic and marginal calibration of $[z|\hat{z}]$

Initial space (y)





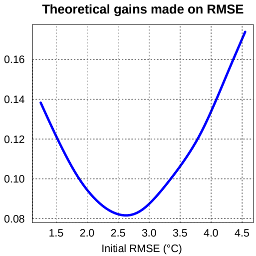
$$\sigma_y^2 = \sigma_{\hat{y}}^2 + \sigma_{\epsilon_{\hat{y}}}^2$$

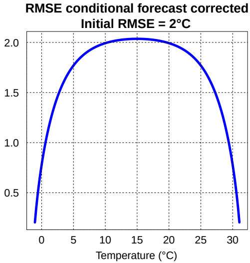
$$\text{corr}(\hat{y}, \epsilon_{\hat{y}}) = 0 \text{ et } \mathrm{E}(\epsilon_{\hat{y}} | \hat{y}) = 0$$

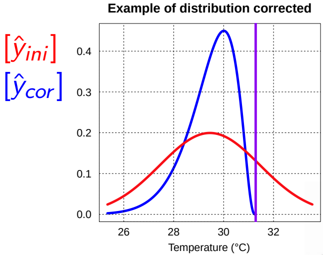
- Only linear independance between $\hat{y}$ and $\epsilon_{\hat{y}}$
- No homoscedasticity
- Probabilistic and marginal calibration of $[y|\hat{y}]$

3. Application examples

Using a truncated normal distribution for a better forecasting of extreme temperatures (theoretical simulation)





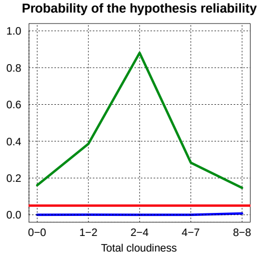


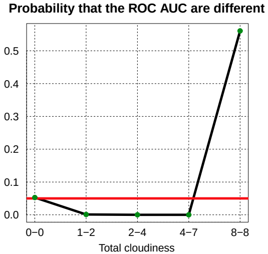
Record value

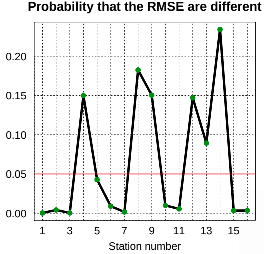
Generalized linear regression by Gaussian anamorphosis

Forecast of the cloudiness : comparison of the qualities of the model against those of logistic regressions (for probabilities forecast) and linear regression (for deterministic forecast)

- New approach
- Classical approach





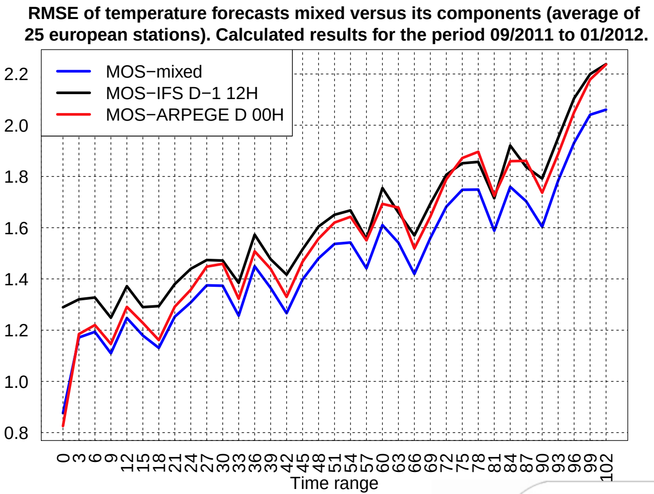


- New approach is better
- Classical approach is better

Probabilistic mixed statistical adaptation

This is a statistically post-processed multi-model system.

Advantage for the deterministic forecast for the short time range :



Probabilistic mixed statistical adaptation

The goal of the probabilistic mixed SA is to fit a linear regression model to each model of an ensemble set, using a BMA.

We have K models M_k. z distribution forms (in the normalized space) given by the BMA :

$$p(z|(M_k)_{k=1,\cdots,K}, \Theta, z^T) = \sum_{k=1}^{K} \alpha_k p(z|M_k, \theta_k, z^T)$$

$$z|M_k, \theta_k, z^T \sim \mathcal{N}(\sum_{i=1}^{q} a_i^k x_i^k, \sigma_k)$$

$z^T = (z_1, \cdots, z_n)$, $\theta_k = ((a_i^k)_{i=1,\cdots,q}, \sigma_k)$ et $\Theta = ((\alpha_k, \theta_k)_{k=1,\cdots,K})$

We estimate the Θ parameters using the maximum likelihood with the EM algorithm.

Step E (expectation) : we estimate the probabilities $p(M_k|\Theta, z^T)$

$$p(M_k|\Theta^g, z^T) = \frac{\alpha_k^{g-1} p(z^T|M_k, \theta_k^{g-1})}{\sum_{l=1}^{K} \alpha_l^{g-1} p(z^T|M_l, \theta_l^{g-1})}$$

METEO FRANCE
Toujours un temps d'avance

Probabilistic mixed statistical adaptation

Step M (maximisation) : parameters Θ are iterativaly estimated

$$\alpha_k^g = \frac{1}{n}\sum_{j=1}^{n} p(M_k|\Theta^g, z_j)$$

$$(a_i^k)_{i=1,\cdots,q}^g = (X_k' P_k^g X_k)^{-1} X_k' P_k^g z^T$$

X_k matrix of model predictors k

$$P_k^g = \begin{pmatrix} p(M_k|\Theta^g, z_1) & 0 & \cdots & 0 \\ 0 & p(M_k|\Theta^g, z_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p(M_k|\Theta^g, z_n) \end{pmatrix}$$

$$\sigma_k^g = \sqrt{\frac{1}{n\alpha_k^g}\sum_{j=1}^{n} p(M_k|\Theta^g, z_j)(\hat{z}_{kj}^{\,g} - z_j)^2}$$

METEO FRANCE
Toujours un temps d'avance

Conclusion

- Among the various ways to produce probability distributions, the *generalized linear regression by Gaussian anamorphosis* has the advantage of being **inexpensive**, **robust**, **to overcome discrimination techniques** requiring a statistical model by class, and **to issue reliable probabilistic forecasts** (if the observations distribution is adjusted properly and the number of predictors sufficiently high).
- The deterministic forecast produced by this model **has good properties** and may be better than the one obtained by a linear regression.
- Coupled with BMA, the Gaussian anamorphosis can be used to improve the quality of an ensemble forecast. This is a statistical-dynamical system **particularly promising**.

Prospects : probabilistic approach and spatialization

- Multi-parameters probabilistic mixed SA (**course in progress**).
- **Probabilistic spatialization of SA on a regular grid**. It can be facilitated in a normalized space by anamorphosis. Idea of a procedure :
  1. Normalization of all the explanatories variables before regression.
  2. Report of the regressions based on a spatial classification of the parameter analyzed by AROME model.
  3. Evaluation of the observations distributions on the grid based on the observed data of the stations and analyzed by AROME (suggestion : interpolation of the Hermite's polynomials coefficients with adjustement cubic splines).
- **Exploiting the forecaster's expertise in probabilistic forecasting**, especially in the case of events strongly bi-modal (effect of the presence or not of low clouds on the temperature).

The end

METEO FRANCE
Toujours un temps d'avance