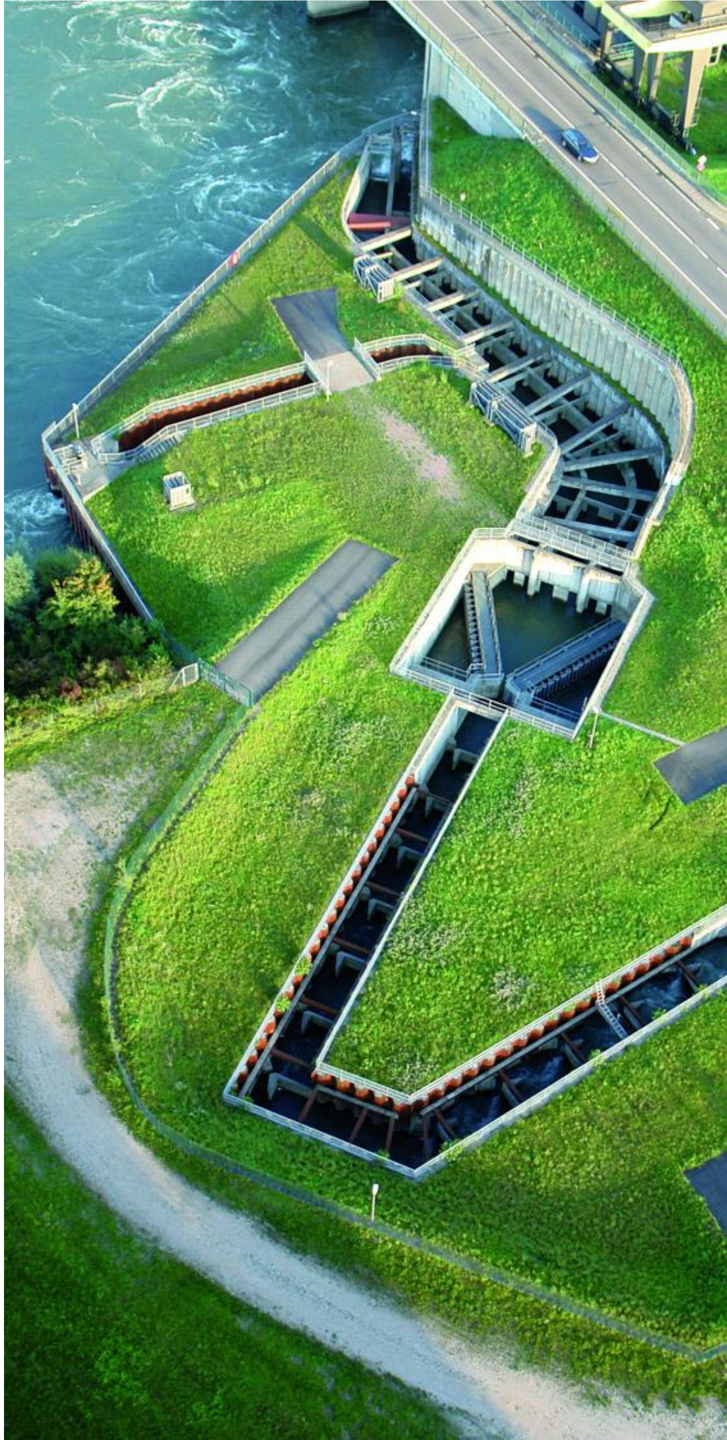
SOME STATISTICAL MODELING PROBLEMS AT EDF

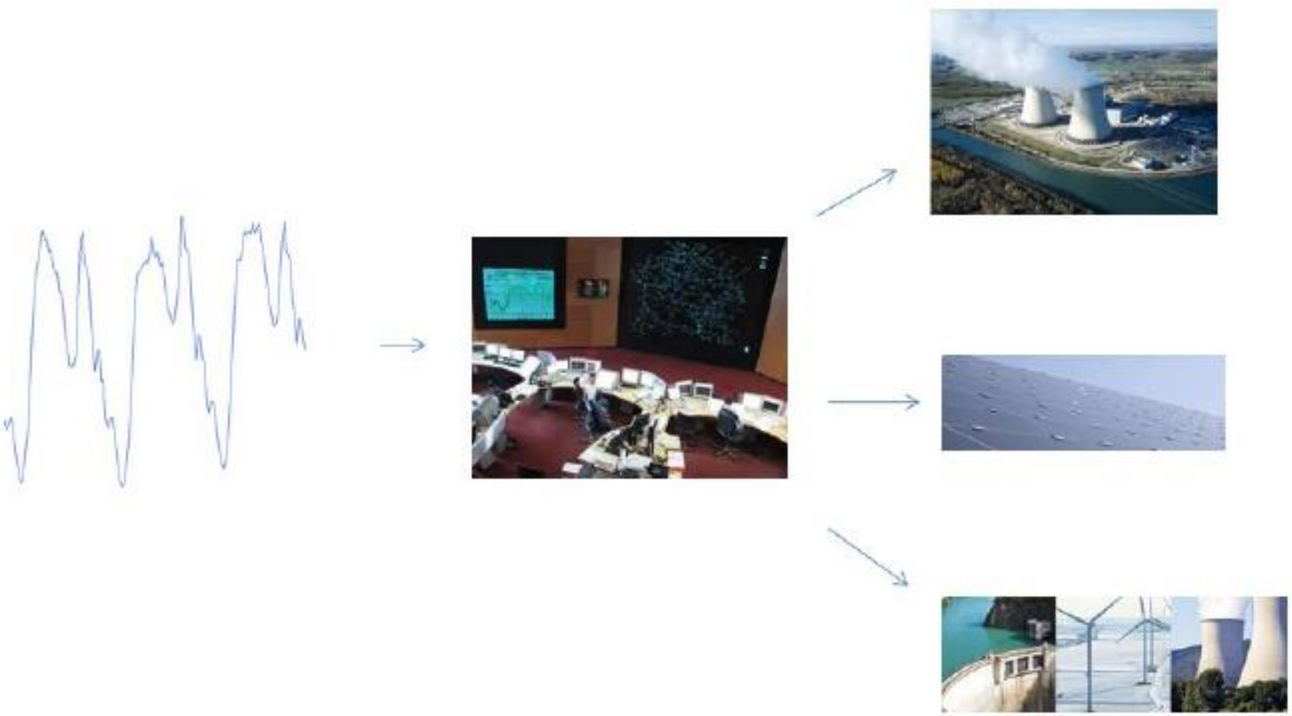
ITT, Bath, 2nd June 2015

Yannig Goude, EDF R&D-Université PSUD
Raphael Nédellec, EDF R&D
Georges Hebrail, EDF R&D



ELECTRICITY LOAD FORECASTING

- **Electricity consumption is the main entry for optimising the production units and managing the grid**



NEW INDUSTRIAL CHALLENGES IN ELECTRICITY

- **Smart grids**
  - More and more « real time » data (ex: linky, 1million meter in 2016)
  - Demand response (new tariffs, real time pricing…)
  - New communication tools with customers (webservice….)
- **Renewables energy development**
  - A more and more probabilistic context: need probabilistic forecasts as enter for risk optimisation tools
- **Opening of the electricity market:**
  - Losses/gains of customers
- **Sensors data:**
  - Production/consumption sites
  - Smart home, internet of things
- **New usages/tariffs:**
  - Electric cars
  - Heat pumps, smart phones, battery charge, computers, flat screens….
  - Demand response

STATISCAL CHALLENGES

- **Big data**
  - Parallelizing statistical algorithms
  - Complex data analysing: heteregonous spatial/temporal sampling, different sources/nature of data
  - Sequential data treatment (data flow, CEP)
  - Heterogenous data treatment
  - Functional data analysis
- **Adaptivity**
  - Non-parametric models
  - Model selection, data driven penalty…
- **Sequential estimation**
  - Break detection, on-line update
  - Aggregation of experts with on-line weights
- **Spatio-temporal**
  - Spatial correlation modeling/simulation
- **Multi-scale models**
  - Multi-horizon models
  - Multi level data on the grid
- **Data mining for time series**
  - Machine learning for time series
- **Probabilistic forecasts**
  - Density forecast
  - Quantile models
- **Large scale simulations**
  - Simulation platform, parallel processing
  - Complex systems dynamics

TOPIC OF THAT TALK

- **Probabilistic forecasts**
  - Quantile models
- **Adaptivity**
  - Non-parametric models
- **Sequential estimation**
  - Aggregation of experts with on-line weights

Probablistic load forecasting on the distribution grid

Probabilistic price forecasting

GEFCOM 14 data (similar than EDF ones)

- **Spatio-temporal**
  - Spatial correlation modeling/simulation

Solar radiation modeling for PV production

PROBABILISTIC FORECAST: A HOT TOPIC

- **A special issue of International Journal of Forecasting**

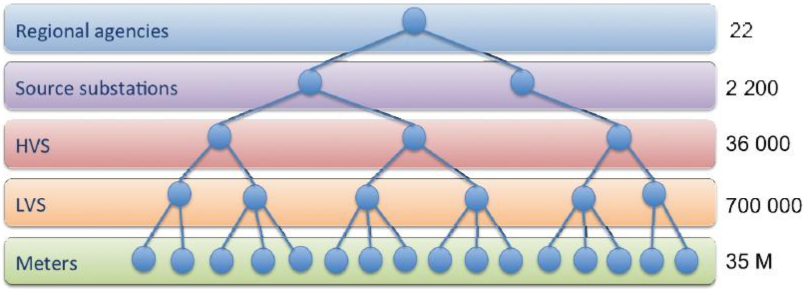
- **GEFcom 2014 competition, sponsored by IEEE Power and Energy Society**

- Participation (nb of teams): Load (333), Price (250), Wind (208), Solar (218)

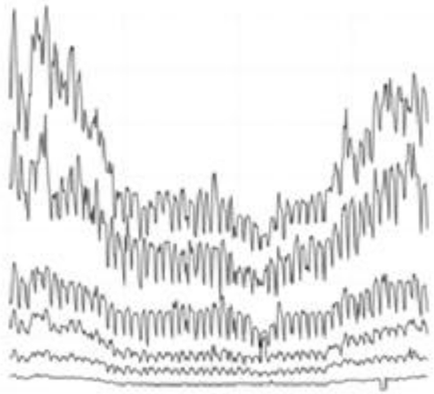
PROBABILISTIC FORECAST: INDUSTRIAL MOTIVATION

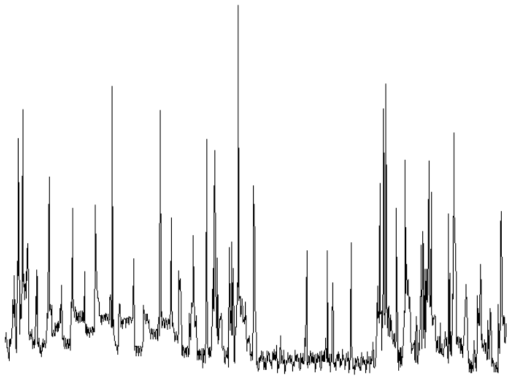
- **Renewables energy development**
  - A « probabilistic world »: *Wind power generation is a direct function of the meteorological conditions, which we humans have no control of, and is hence highly fluctuating* Pr. P.Pinson

- **Opening of the electricity market:**
  - Losses/gains of customers:
    - scenario based forecasts, adaptivity, time varying parameters
    - Bottom up forecasts



- **Local forecasting**
  - Low agregation level
  - Wide variety of consumers
  - Covariate selection
  - Noisy data





GAM MODELS FOR LOAD FORECASTING

- **A good trade-off complexity/adaptivity**

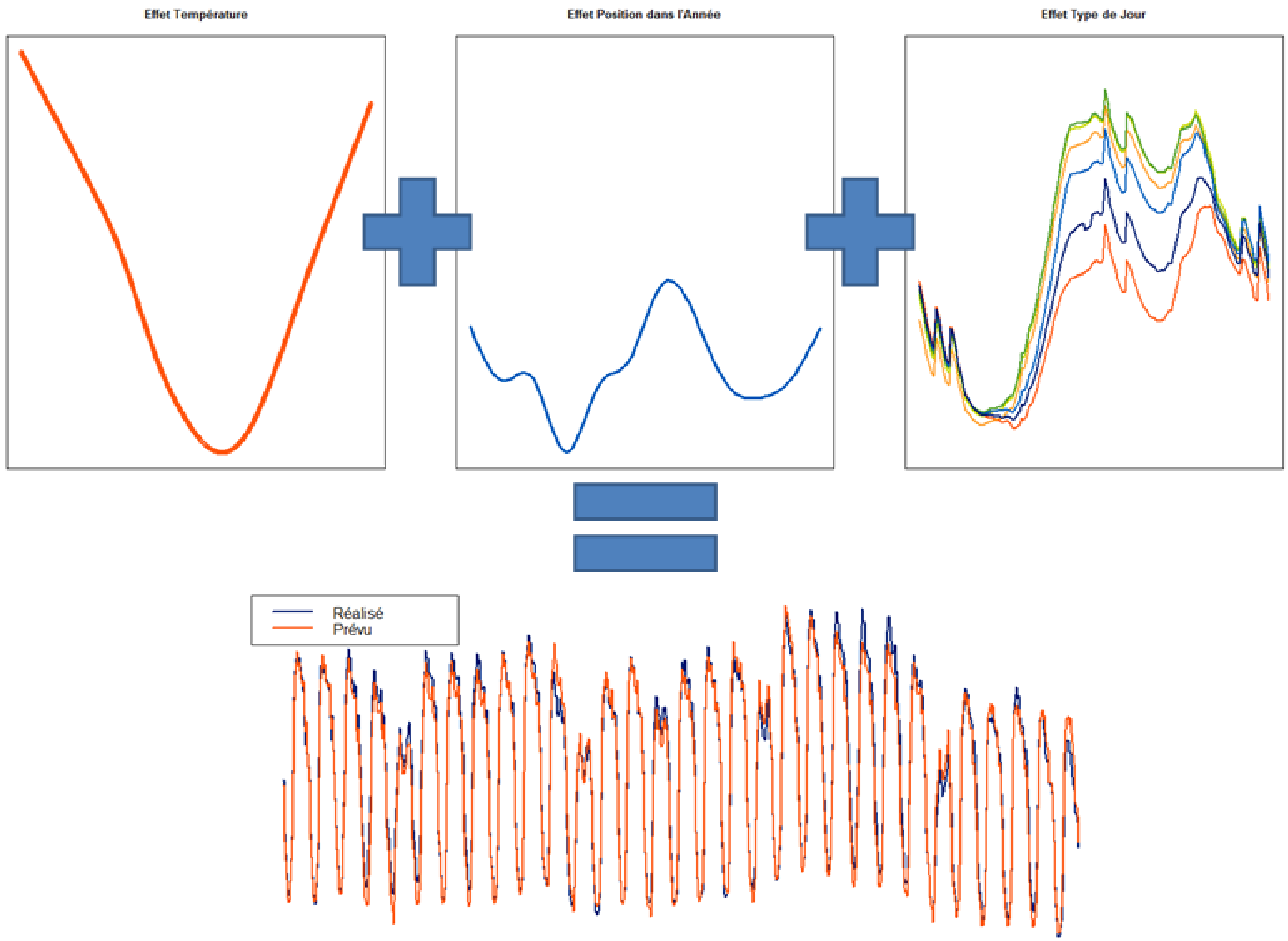
$$y_t = f_1(x_t^1) + f_2(x_t^2) + \ldots + f(x_t^3, x_t^4) + \ldots + \varepsilon_t$$

$$min_{\beta, f_j} ||y - f_1(x_1) - f_2(x_2) - \ldots||^2 + \lambda_1 \int f_1''(x)^2 dx + \lambda_2 \int f_2''(x)^2 dx + \ldots$$

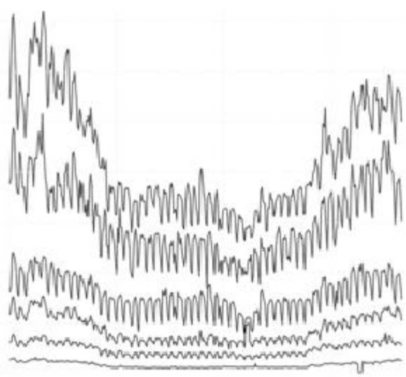
- **Publications**
  - Application on load forecasting
    - A. Pierrot and Y. Goude, Short-Term Electricity Load Forecasting With Generalized Additive Models **Proceedings of ISAP power**, pp 593-600, 2011.
    - R. Nédellec, J. Cugliari and Y. Goude, GEFCom2012: Electricity Load Forecasting and Backcasting with Semi-Parametric Models, International Journal of Forecasting , 2014, 30, 375 - 381.
  - GAM parallel for big datasets
    - S.N. Wood, Goude, Y. and S. Shaw, Generalized additive models for large datasets, to appear in **Journal of Royal Statistical Society-C**.
  - Adaptive GAM (forgeting factor)
    - A. Ba, M. Sinn, Y. Goude and P. Pompey, Adaptive Learning of Smoothing Functions: Application to Electricity Load Forecasting Advances in **Neural Information Processing Systems** 25, 2012, 2519-2527.

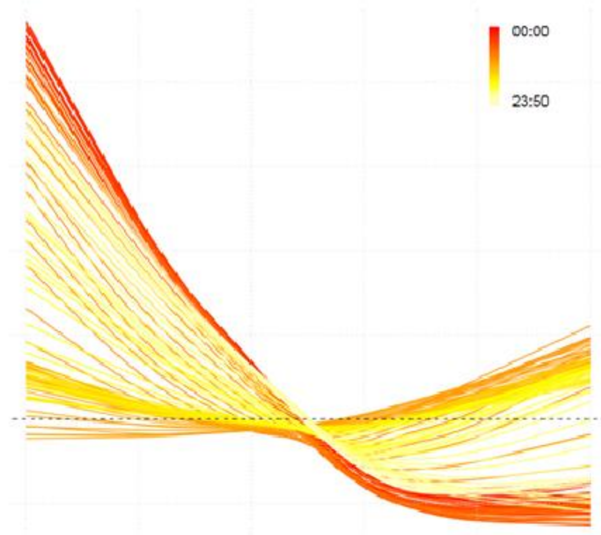
GAM MODELS FOR LOAD FORECASTING

$$y_t = f_1(T_t) + f_2(I_t) + f_3(H_t) + \varepsilon_t$$



GAM MODELS FOR LOCAL LOAD FORECASTING
APPLICATION ON ERDF SUBSTATIONS





Fit 2200 on 10 min electricity data models automatically

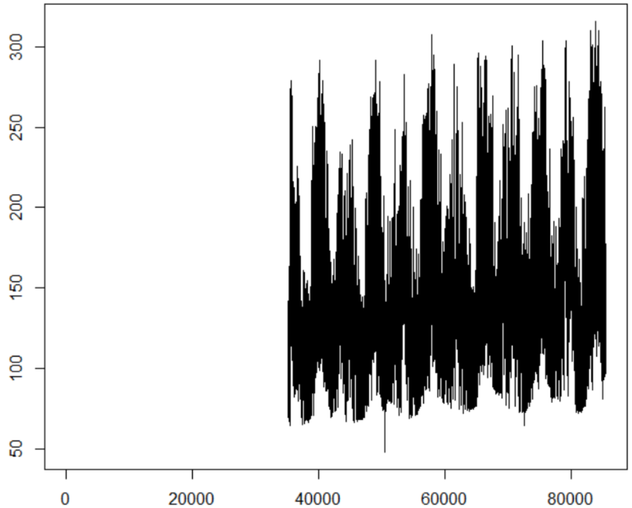
Goude, Y., Nédellec, R. and Kong, N., Local Short and Middle term Electricity Load Forecasting with semi-parametric additive models to appear in IEEE transactions on smart grid, 2013, 5 , Issue: 1, 440 - 446

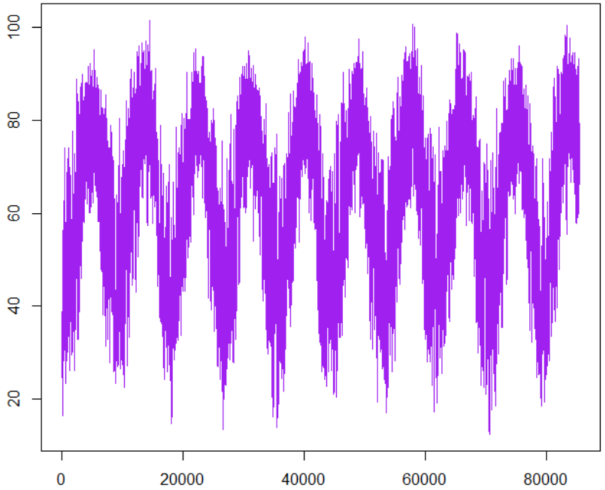
GEFCOM 14 LOAD FORECASTING TRACK

- **hourly electricity load data (somewhere in US), 6 years from January 1, 2006 to December 31, 2011**

- **Temperature data (25 stations) ,11 years**

- **GEFcom 2014 competition: each week from september 2014 to december 2014 produce a monthly forecast of each month of 2012 for the 0.01, 0.02, …,0.99 quantile of the load.**



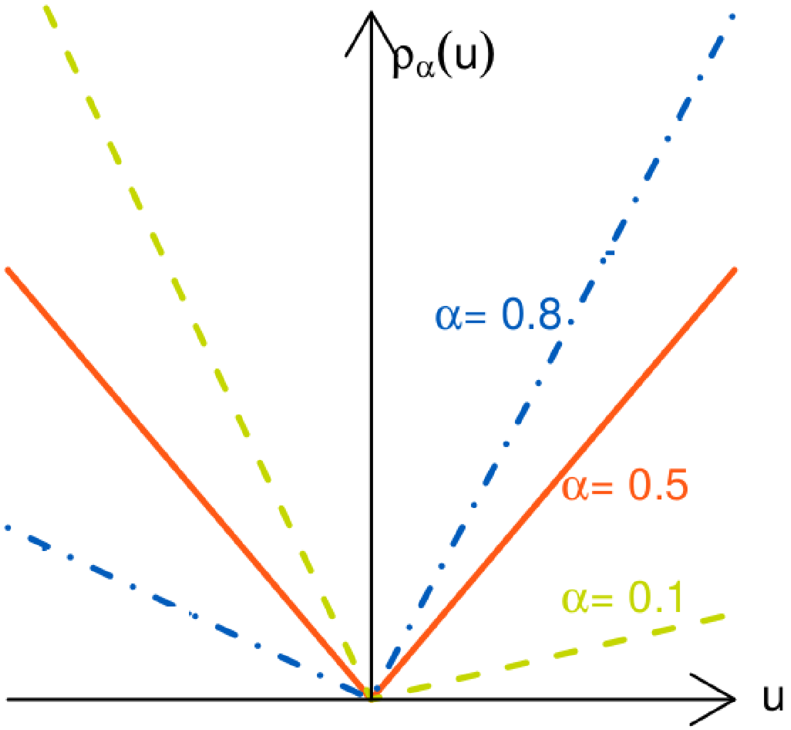


- **Our team:**
  - Tololo: Pierre Gaillard, Raphael Nédellec, Yannig Goude

GEFCOM 14 LOAD FORECASTING TRACK

- **competition GEFCOM 2014**
  - Performance evaluation: pin-ball loss

$$L(q_a, y) = \begin{cases} (1 - a/100)(q_a - y), & \text{if } y < q_a; \\ a/100(y - q_a), & \text{if } y \geq q_a; \end{cases}$$



GEFCOM 14 LOAD FORECASTING TRACK

- **Our original idea**
  - Quantile non-linear regression: GAM for each quantile

$$q_\alpha(Y|X) = F^{-1}_{Y|X}(\alpha) = \inf\{y \in \mathbb{R}, F_{Y|X}(y) \geq \alpha\}$$

$$q_\alpha(Y|X) \in \arg\min_g \mathbb{E}[\rho_\alpha(Y - g(X))|X]$$

$$\rho_\alpha(u) = \alpha(u - \mathrm{I}(u < 0))$$

$$\boxed{f_1(X_{t,1}) + f_2(X_{t,2}) + f_3(X_{t,3}, X_{t,4}) + \ldots}$$

Koenker, R., 2013. quantreg: Quantile Regression. R package version 5.05. URL http://CRAN.R-project.org/package=quantreg

Koenker, R. W., Bassett, G. W., 1978. Regression quantiles. Econometrica 46 (1), 33–50.

Issue: numerical problems with non-linear effects, time consuming

GEFCOM 14 LOAD FORECASTING TRACK

- **Our approach:**
  - Fitting a GAM to model the conditional mean:

$$Y_t = f_1(Toy_t) + f_2(t) + f_3(T_t) + h(DayType_t) + \varepsilon_t$$

  - Fitting a GAM to model the conditional variance:

$$\left(Y_t - \widehat{Y_t}\right)^2 = g_1(Toy_t) + g_2(T_t) + \varepsilon_t ,$$

  - Linear quantile regression on:

$$Y_t = Z_t\beta + \varepsilon_t$$

$$\mathbf{Z}_t = (\widehat{f_1}(Toy_t), \ \widehat{f_2}(t), \ \ldots , \ \widehat{g_1}(Toy_t), \ \widehat{g_2}(T_t))$$

Rk: this is conditional to temperature, same approach is conducted for the temperature (on Toy) to model the meteorological randomness, then plug in

GEFCOM 14 LOAD FORECASTING TRACK

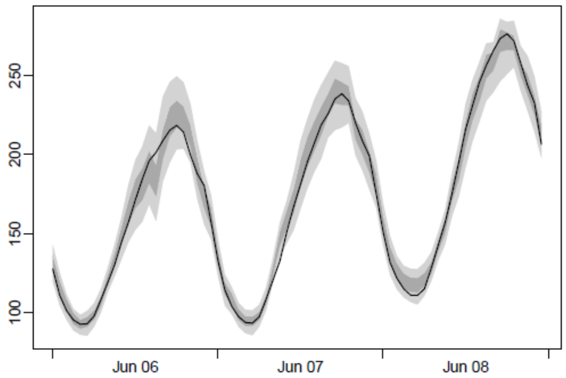


Figure 4: Medium-term forecasted distribution of the load from June 6, 2011 to June 8, 2011 by using the real values of the temperature T_t. As in all the following plots of probabilistic distributions, we only plot the 50% confidence interval in dark gray and the 90% confidence interval in light gray.

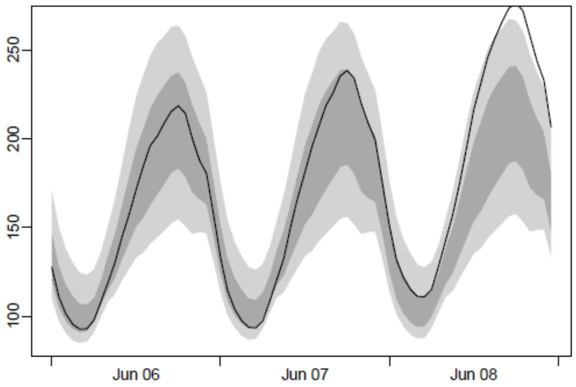


Figure 5: Medium-term forecasted distribution of the load from June 6, 2011 to June 8, 2011 obtained by averaging the forecasted distributions of the load over the forecasted distribution of the temperature.

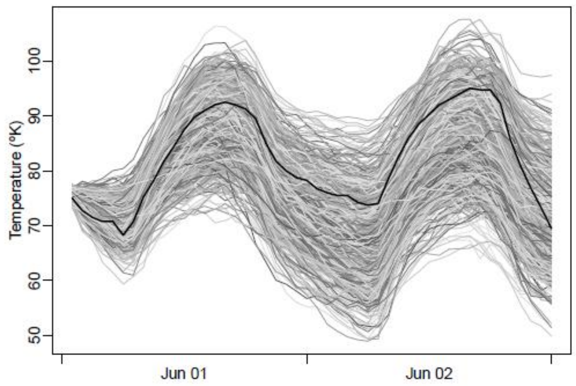


Figure 6: 800 temperature scenarios (T_s) generated for June 1, 2011 to June 2, 2011. The line in black depicts the observed temperature.

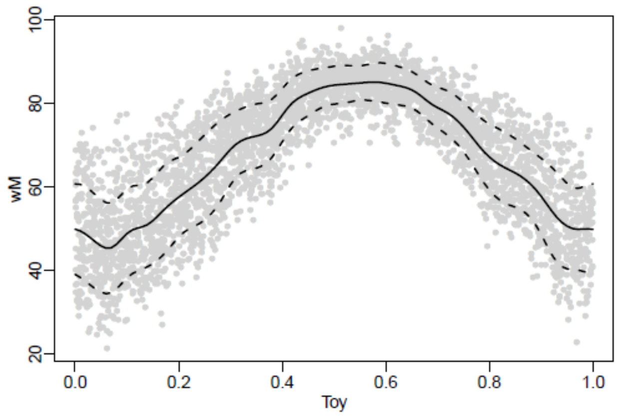
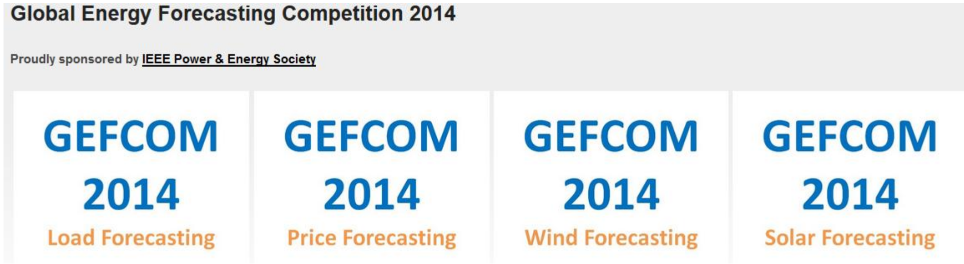


Figure 3: Observed values of T_t together with the smooth functions $\widehat{f_1}$ and $\widehat{f_1} \pm \widehat{g_1}$ fitted by Models (8) and (9).

PROB. ELECTRICITY PRICE FORECASTING

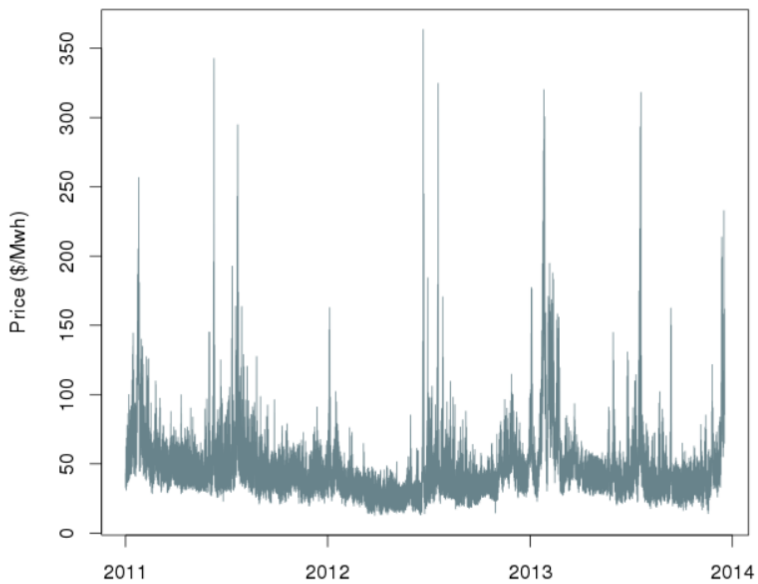
- **competition GEFCOM 2014,** sponsored by IEEE Power and Energy Society
  - september 2014-december 2014
  - Probabilistic forecast (quantile 1%,…,99%) of hourly electricity prices in US based on:
    - Zonal/total electricity load forecast
    - Past prices
  - Online forecasting of 15 days
  - Performance evaluation: pin-ball loss

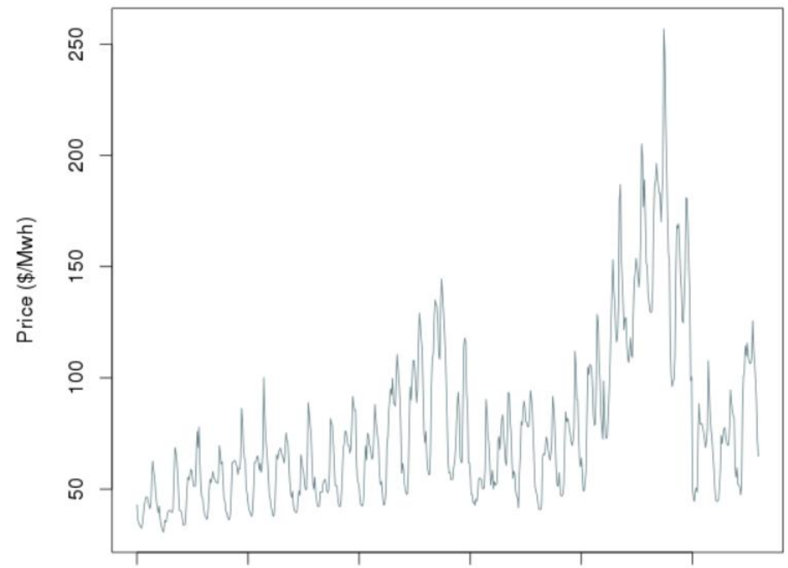


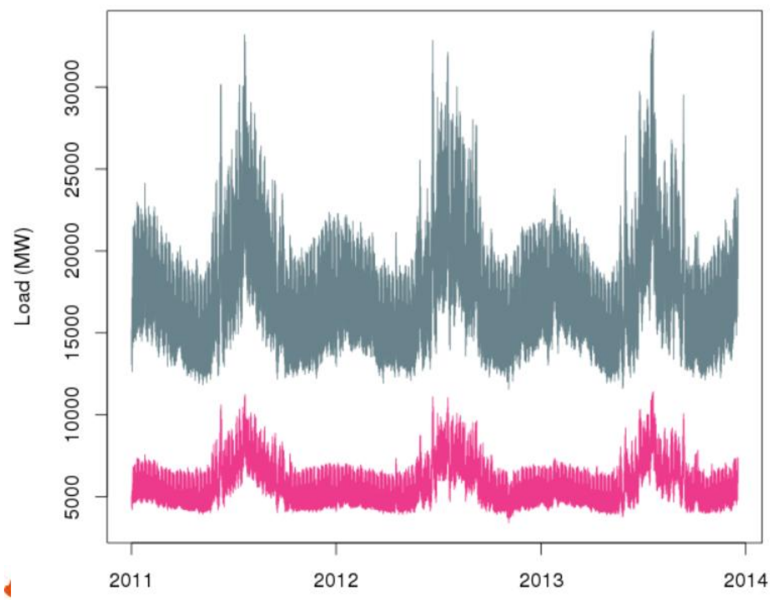
  - Participation (nb of teams): Load (333), Price (250), Wind (208), Solar (218)

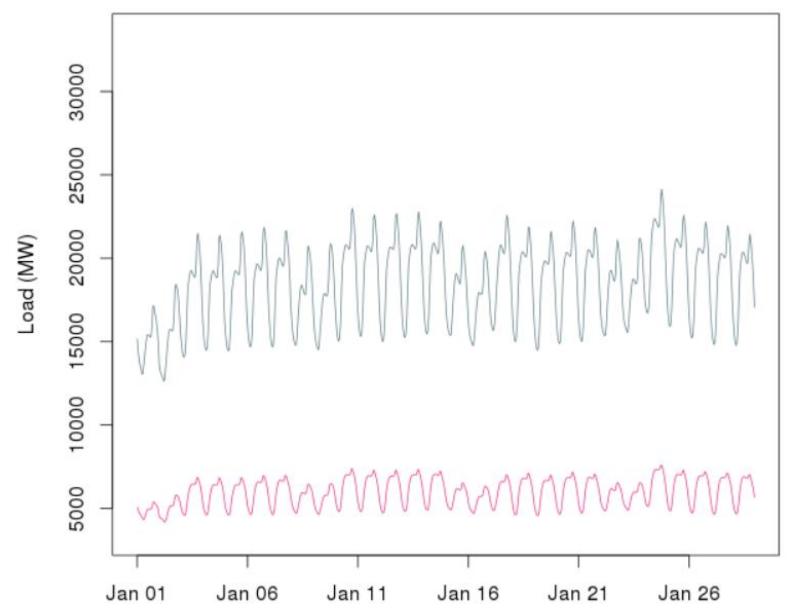
Focus on h+1, …, h+24h forecasts

ELECTRICITY PRICE DATA









PROB. ELECTRICITY PRICE FORECASTING

GEFCOM14

- **Aggregation of 13 experts:**
  - autoregressive model (AR)

$$\log(\mathrm{P}_t) = \alpha_1 \log(\mathrm{P}_{t-24}) + \alpha_2 \log(\mathrm{P}_{t-48}) + \alpha_3 \log(\mathrm{P}_{t-168}) + \alpha_4 \log(\mathrm{P.min}_{t-24}) + h(\mathrm{DayType}_t) + \varepsilon_t$$

  - An autoregressive model with forecasted electric loads as additional covariates (ARX).
  - A threshold autoregressive model TAR defined as an extension of AR to two regimes depending of the variation of the mean price between a day and eight days ago.
  - TARX the extension of ARX to the two regimes model.
  - Spike pre-processed autoregressive model PAR
  - PARX similar to PAR, but ARX is fitted with pre-processed prices.

  inspire from Weron, R., Misiorek, A., 2008. *Forecasting spot electricity prices: A comparison of parametric and semiparametric time series models*. **International Journal of Forecasting** 24 (4), 744 – 763

  - 2 linear regressions
  - 2 GAMS
  - 2 random forests
  - GBM

$$\log(\mathrm{P}_t) = \alpha_1 \log(\mathrm{P}_{t-24}) + \alpha_2 \log(\mathrm{P}_{t-48}) + \alpha_3 \log(\mathrm{P}_{t-168}) + \alpha_4 \log(\mathrm{P.max}_t) + \alpha_5 \mathrm{FZL}_t^{(0.95)} + \alpha_6 \mathrm{FTL}_t^{(0.95)} + \alpha_7 \mathrm{FZL}_t^{(0.8)} + \alpha_8 \mathrm{FTL}_t^{(0.8)} + h(\mathrm{DayType}_t) + \varepsilon_t$$

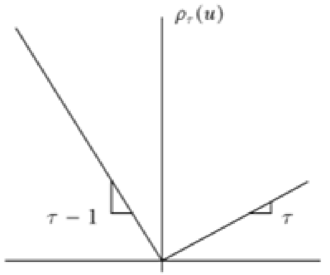
PROB. ELECTRICITY PRICE FORECASTING

GEFCOM14

- **(Convex) Aggregation with pin-ball loss:**

$$\widehat{y}_t = \sum_{i=1}^{N} \widehat{p}_{i,t} x_{i,t} \qquad \widehat{p}_{k,t} = \frac{e^{-\eta \sum_{s=1}^{t-1} \ell_s(x_{k,s})}}{\sum_{i=1}^{K} e^{-\eta \sum_{s=1}^{t-1} \ell_s(x_{i,s})}}$$

$$l_t(x_{k,t}) = \rho_\tau(y_t - x_{k,t}) \longrightarrow$$



- **Extension to linear aggregation:**
  - substitute to original experts $\beta x_{1,t}, \ldots, \beta x_{K,t}, -\beta x_{1,t}, \ldots, -\beta x_{K,t}$

SEQUENTIAL AGGREGATION OF EXPERTS

Each instance t

- Each expert suggests a prediction $x_{i,t}$ of the consumption y_t
- We assign weight to each expert and we predict

$$\widehat{y}_t = \widehat{\boldsymbol{p}}_t \cdot \boldsymbol{x}_t \quad \left(= \sum_{i=1}^{N} \widehat{p}_{i,t} x_{i,t} \right)$$

Our goal is to minimize our cumulative loss

$$\underbrace{\sum_{t=1}^{T} (\widehat{y}_t - y_t)^2}_{\text{Our loss}} = \underbrace{\min_{\boldsymbol{q} \in \Delta_N} \sum_{t=1}^{T} (\boldsymbol{q} \cdot \boldsymbol{x}_t - y_t)^2}_{\text{Loss of the best convex combination}} + \underbrace{R_T}_{\text{Estimation error}}$$

Good set of experts
As varied as possible

Good aggregating algorithm

EXPONENTIALLY WEIGHTED AVERAGE FORECASTER (EWA)

Each instance t

- Each expert suggests a prediction $x_{i,t}$ of the consumption y_t
- We assign to expert i the weight

$$\widehat{p}_{i,t} = \frac{exp\left(-\eta \sum_{s=1}^{t\ \ 1}(x_{i,s} - y_s)^2\right)}{\sum_{j=1}^{N} exp\left(-\eta \sum_{s=1}^{t\ \ 1}(x_{j,s} - y_s)^2\right)}$$

- and we predict $\widehat{y}_t = \sum_{i=1}^{N} \widehat{p}_{i,t} x_{i,t}$

Our cumulated loss is upper bounded by

$$\underbrace{\sum_{t=1}^{T}(\widehat{y}_t - y_t)^2}_{\text{Our loss}} \leqslant \underbrace{\min_{i=1,\dots,d} \sum_{t=1}^{T}(x_{i,t} - y_t)^2}_{\text{Loss of the best expert}} + \underbrace{\square\sqrt{T \log N}}_{\text{Estimation error}}$$

EXPONENTIATED GRADIENT FORECASTER (EG)

Each instance t

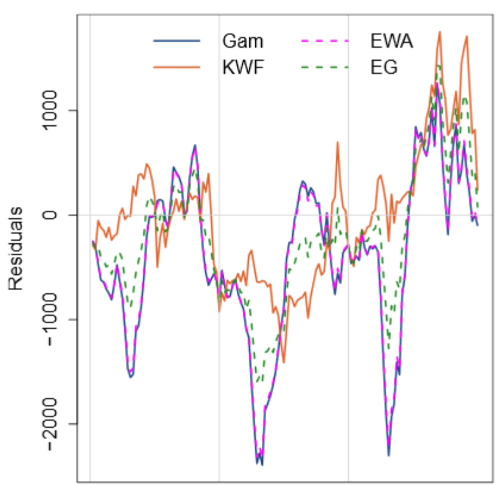
- Each expert suggests a prediction $x_{i,t}$ of the consumption y_t
- We assign to expert i the weight

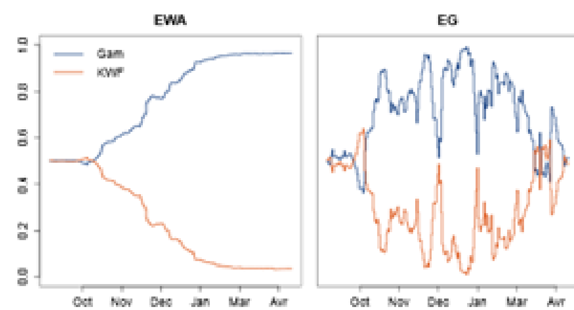
$$\widehat{p}_{i,t} \propto exp\left(-\eta \sum_{s=1}^{t-1} \ell_{i,s}\right) \qquad \text{where } \ell_{i,s} = 2(\widehat{y}_s - y_s)x_{i,s}$$

- and we predict $\widehat{y}_t = \sum_{i=1}^{N} \widehat{p}_{i,t} x_{i,t}$

Our cumulated loss is then bounded as follow

$$\underbrace{\sum_{t=1}^{T}(\widehat{y}_t - y_t)^2}_{\text{Our loss}} \leqslant \underbrace{\min_{\boldsymbol{q}\in\Delta_N} \sum_{t=1}^{T}(\boldsymbol{q}\cdot\boldsymbol{x}_t - y_t)^2}_{\text{Loss of the best convexe combination}} + \underbrace{\square\sqrt{T\log N}}_{\text{Estimation error}}$$





MULTIPLE LEARNING RATE-POLYNOMIAL, RIDGE

Algorithm 1 The polynomially weighted average forecaster with multiple learning rates (ML-Poly)

Input: $h \geqslant 1$, horizon of prediction
Initialize: For $t \leqslant h$, $\boldsymbol{p}_t = (1/K, \ldots, 1/K)$ and $\boldsymbol{R}_1 = (0, \ldots, 0)$
for each instance $t = 1, 2, \ldots, n-h$ **do**
0. pick the learning rates
$$\eta_{k,t} = 1/\left(1 + \sum_{s=1}^{t} \left(\ell_s(\widehat{y}_s) - \ell_s(x_{k,s})\right)^2\right)$$
where $\ell_s : x \mapsto x(y_s - \widehat{y}_s)$.
1. form the mixture $\widehat{\boldsymbol{p}}_{t+h}$ defined component-wise by
$$\widehat{p}_{k,t+h} = \eta_{k,t}\,(R_{k,t})_+ \;/\; \left[\boldsymbol{\eta}_t \cdot (\boldsymbol{R}_t)_+\right]$$
where $\boldsymbol{x}_+$ denotes the vector of non-negative parts of the components of $\boldsymbol{x}$
2. predict $\widehat{y}_{t+h} = \widehat{\boldsymbol{p}}_{t+h} \cdot \boldsymbol{x}_{t+h}$ and observe y_{t+1}
3. for each expert k update the regret
$$R_{k,t+1} = R_{k,t} + \ell_t(\widehat{y}_{t+1}) - \ell_t(x_{k,t+1})$$
end for

Gaillard, P., Stoltz, G., van Erven, T.: *A second-order bound with excess losses*, **COLT proceedings** (2014).

Gaillard, P. & Goude, Y. *Forecasting electricity consumption by aggregating experts; how to design a good set of experts* to appear **in Lecture Notes in Statistics: Modeling and Stochastic Learning for Forecasting in High Dimension**, 2014

Automatic calibration works well in practice

Fast tuning

Algorithm 2 The ridge regression forecaster (Ridge)

Input: $\lambda > 0$, learning rate; $h \geqslant 1$, horizon
Initialize: for $t \leqslant h$, $\widehat{\boldsymbol{p}}_t = (1/K, \ldots, 1/K)$
for each instance $t = 1, 2, \ldots, n$ **do**
1. form the mixture $\widehat{\boldsymbol{p}}_{t+h}$ defined by
$$\widehat{\boldsymbol{p}}_t = \underset{\boldsymbol{u} \in \mathbb{R}^K}{\operatorname{argmin}} \left\{ \sum_{s=1}^{t} (y_s - \boldsymbol{u} \cdot \boldsymbol{x}_s)^2 + \lambda \left\| \boldsymbol{u} - \widehat{\boldsymbol{p}}_0 \right\|_2^2 \right\}$$
2. output prediction $\widehat{y}_{t+h} = \widehat{\boldsymbol{p}}_{t+h} \cdot \boldsymbol{x}_{t+h}$
end for

Stable weights

OPERA: ONLINE PREDICTION BY EXPERTS AGGREGATION

mixture {opera} R Documentation

Compute an aggregation rule

Description

The function `mixture` performs an aggregation rule chosen by the user. It considers a sequence `y` of observations to be predicted sequentially with the help of experts advices `x`. The forms at each instance `t` a prediction by assigning weight to the experts advices and combining them.

Usage

```
mixture(y, experts, aggregationRule = "MLpol", w0 = NULL, awake = NULL,
  href = 1, period = 1, delay = 0, y.ETR = NULL)
```

Arguments

`y` A vector containing the observations to be predicted.

`experts` A matrix containing the experts forecasts. Each column corresponds to the predictions proposed by an expert to predict `Y`. It has as many columns as there are experts.

`aggregationRule` Either a character string specifying the aggregation rule to use or a list with a component `name` specifying the aggregation rule and any additional parameter needed. Currently available aggregation rules are:

"EWA"

Exponentially weighted average aggregation rule. A positive learning rate **eta** can be chosen by the user. The bigger it is the faster the aggregation rule will learn from observations and experts performances. However too hight values lead to unstable weight vectors and thus unstable predictions. If it is not specified, the learning rate is calibrated online.

"FS"

Fixed-share aggregation rule. As for `ewa`, a learning rate **eta** can be chosen by the user or calibrated online. The main difference with `ewa` aggregation rule rely in the mixing rate **alpha***\in [0,1]* wich considers at each instance a small probability `alpha` to have a rupture in the sequence and that the best expert may change. Fixed-share aggregation rule can thus compete with the best sequence of experts that can change a few times (see `bestShifts`), while `ewa` can only compete with the best fixed expert. The mixing rate is either chosen by the user either calibrated online.

"Ridge"

Ridge regression. It minimizes at each instance a penalized criterion. It forms at each instance linear combination of the experts' forecasts and can assign negative weights that not necessarily sum to one. It is useful if the experts are biased or correlated. It cannot be used with specialized experts. A positive regularization coefficient **lambda** can either be chosen by the user or calibrated online.

"MLpol"

Polynomial Potential aggregation rule with different learning rates for each expert. The learning rates are calibrated using theoretical values. There are similar aggregation rules like "BOA" (Bernstein online Aggregation see [Wintenberger, 2014] "MLewa", and "MLprod" (see [Gaillard, Erven, and Stoltz, 2014])

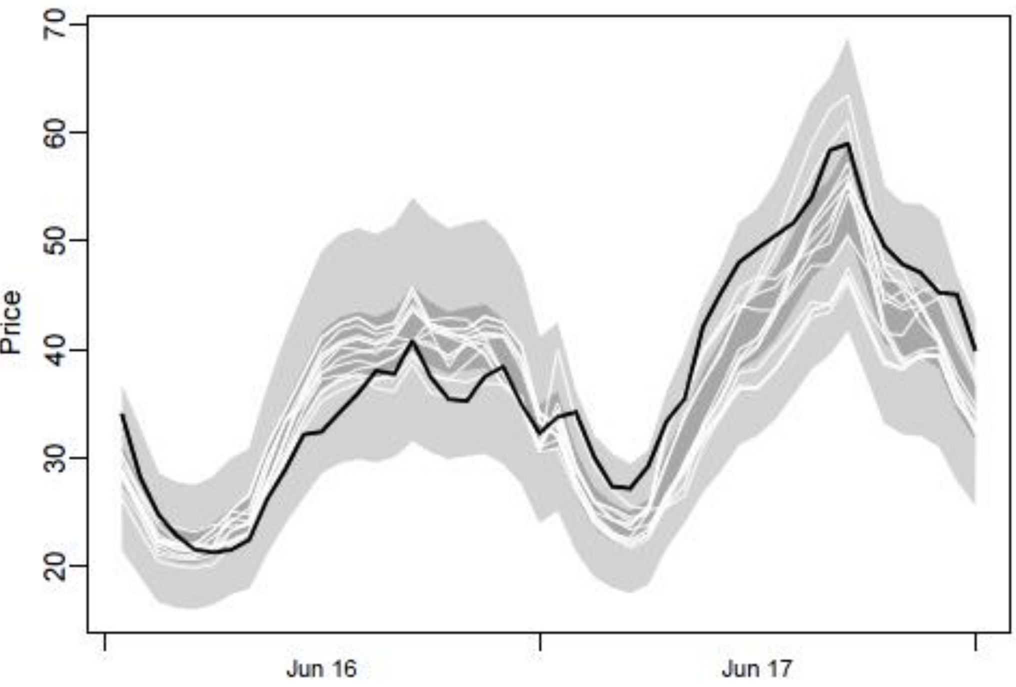
"pinball"

It performs a mixing aggregation rule for quantile regression. At each instance, it forms the mixture by performing a convex minimisation. It chooses the mixture that minimizes among the past a penalized criterion based on cumulated pinball loss.

PROB. ELECTRICITY PRICE FORECASTING

GEFCOM14

- **Results**



Task	TOLOLO	quantGAM	quantMixt	quantGLM
Jun. 06	XX	0.72	0.85	1.87
Jun. 17	1.06	1.15	1.37	0.71
Jun. 24	1.91	1.31	1.58	3.05
Jul. 04	1.71	2.06	1.27	1.59
Jul. 09	1.45	0.99	3.31	1.57
Jul. 13	1.10	2.23	1.20	1.18
Jul. 16	2.01	2.63	2.28	5.02
Jul. 18	9.15	5.13	7.90	11.72
Jul. 19	4.68	4.80	6.45	13.27
Jul. 20	1.59	1.90	2.35	2.80
Jul. 24	0.75	0.75	1.78	1.42
Jul. 25	2.46	2.30	0.84	2.12
Dec. 06	2.96	0.82	1.03	0.86
Dec. 07	1.35	3.63	3.23	3.22
Dec. 17	3.56	3.83	4.26	2.87

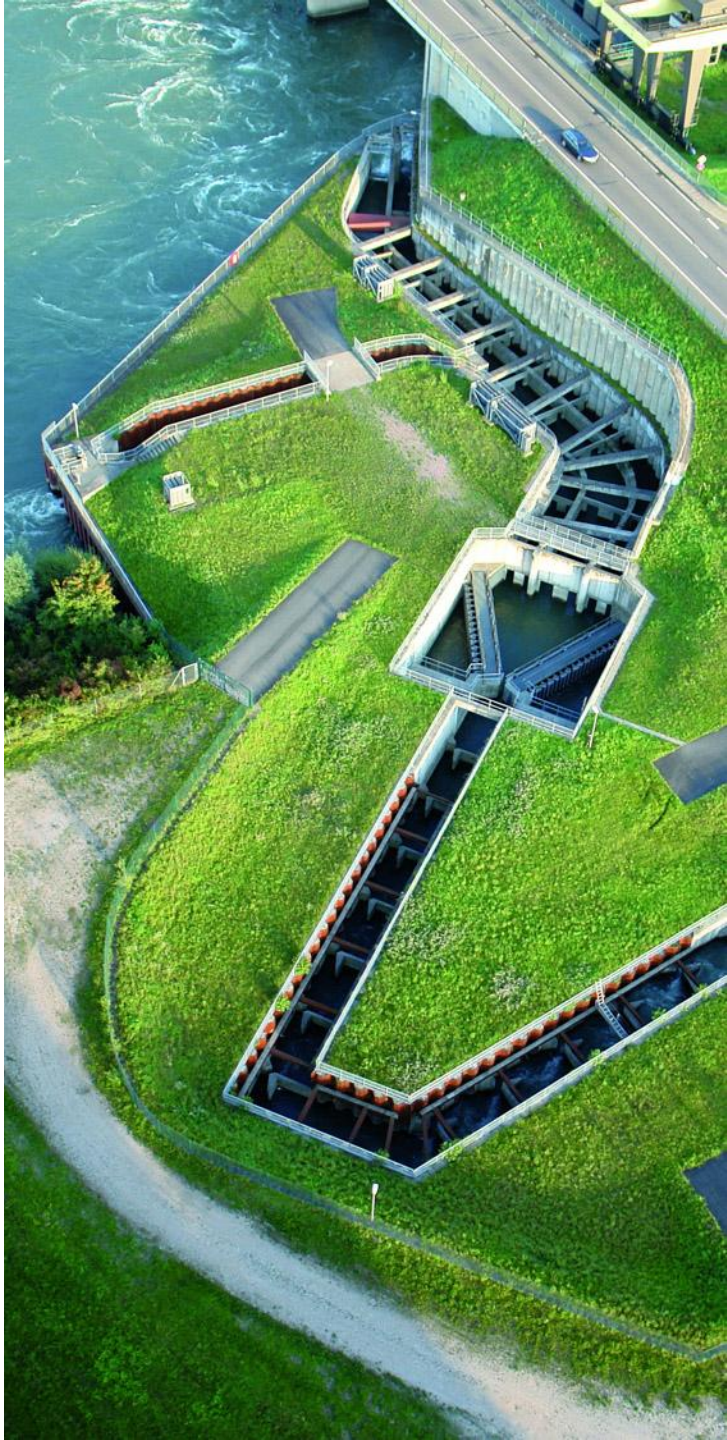
PROB. ELECTRICITY PRICE & LOAD FORECASTING

GEFCOM14

- **Results: 1st rank of the competition for both tracks**

	Load		Price
Ranking	Team	Rating	Team
1	**Tololo**	**50,0%**	**Tololo**
2	**Adada**	**49,0%**	Team Poland
3	Jingrui (Rain) Xie	48,0%	GMD
4	OxMath	47,6%	C3 Green Team
5	E.S. Mangalova	45,4%	pat1

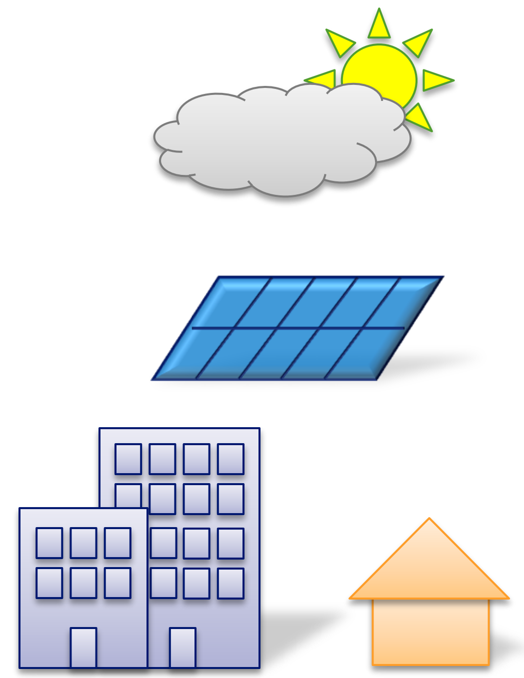
VERY SHORT TERM FORECASTING OF INDIVIDUAL / LOCAL PV GENERATION



PROBLEM DEFINITION

Development of renewable energies

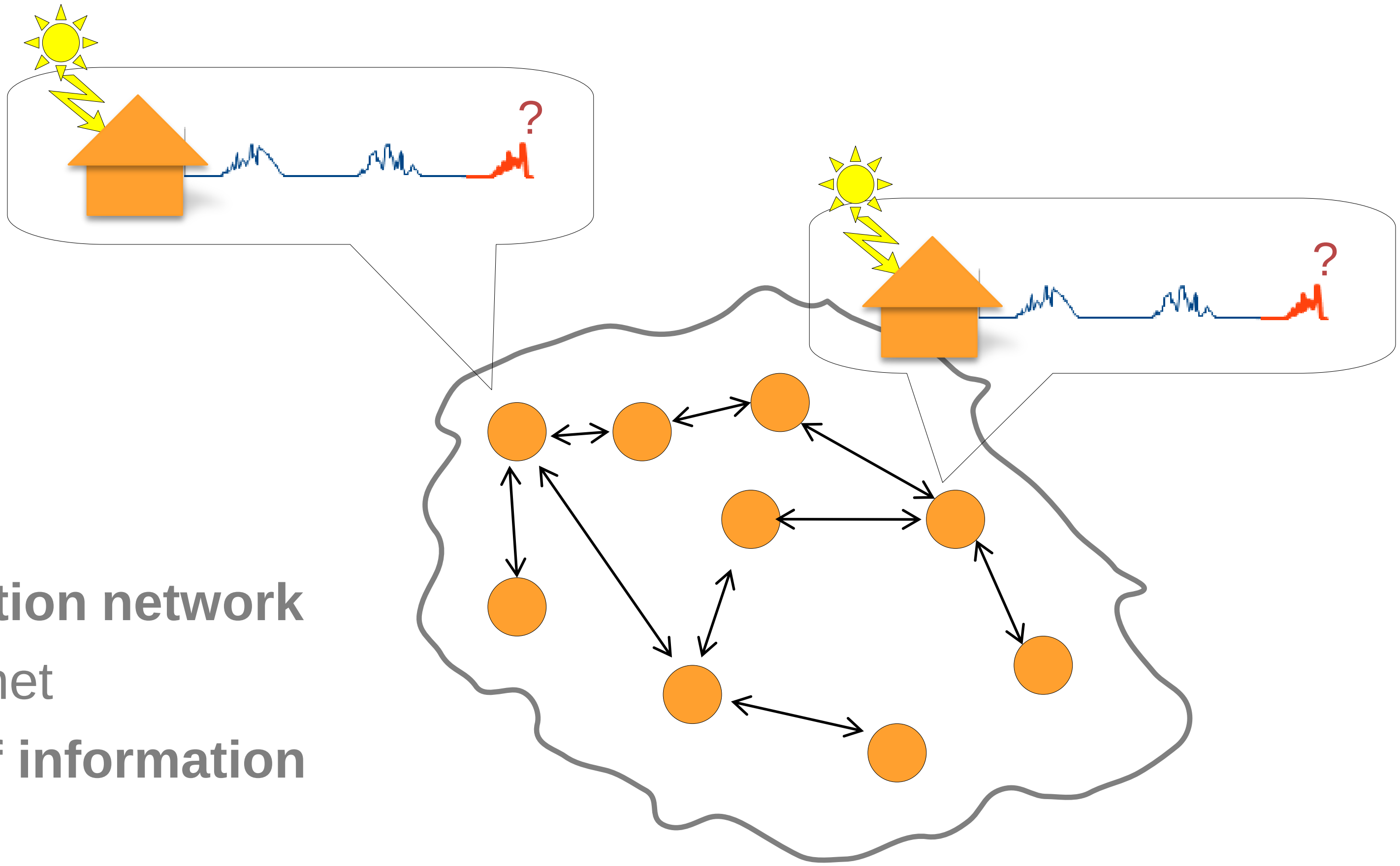
- Need for prediction of production
  - Electric grid operation
  - Local auto-consumption



- Very short / short / medium term prediction
- Aggregated / local / individual prediction

→ We focus on very-short term (10’ to 3h) individual prediction

COLLABORATIVE INDIVIDUAL PREDICTIONS



- **Communication network**
  - Ex. Internet
- **Exchange of information**
  - Ex. P2P
- **Collaborative statistical model**

COLLABORATIVE STATISTICAL MODEL

- **W advection vector**

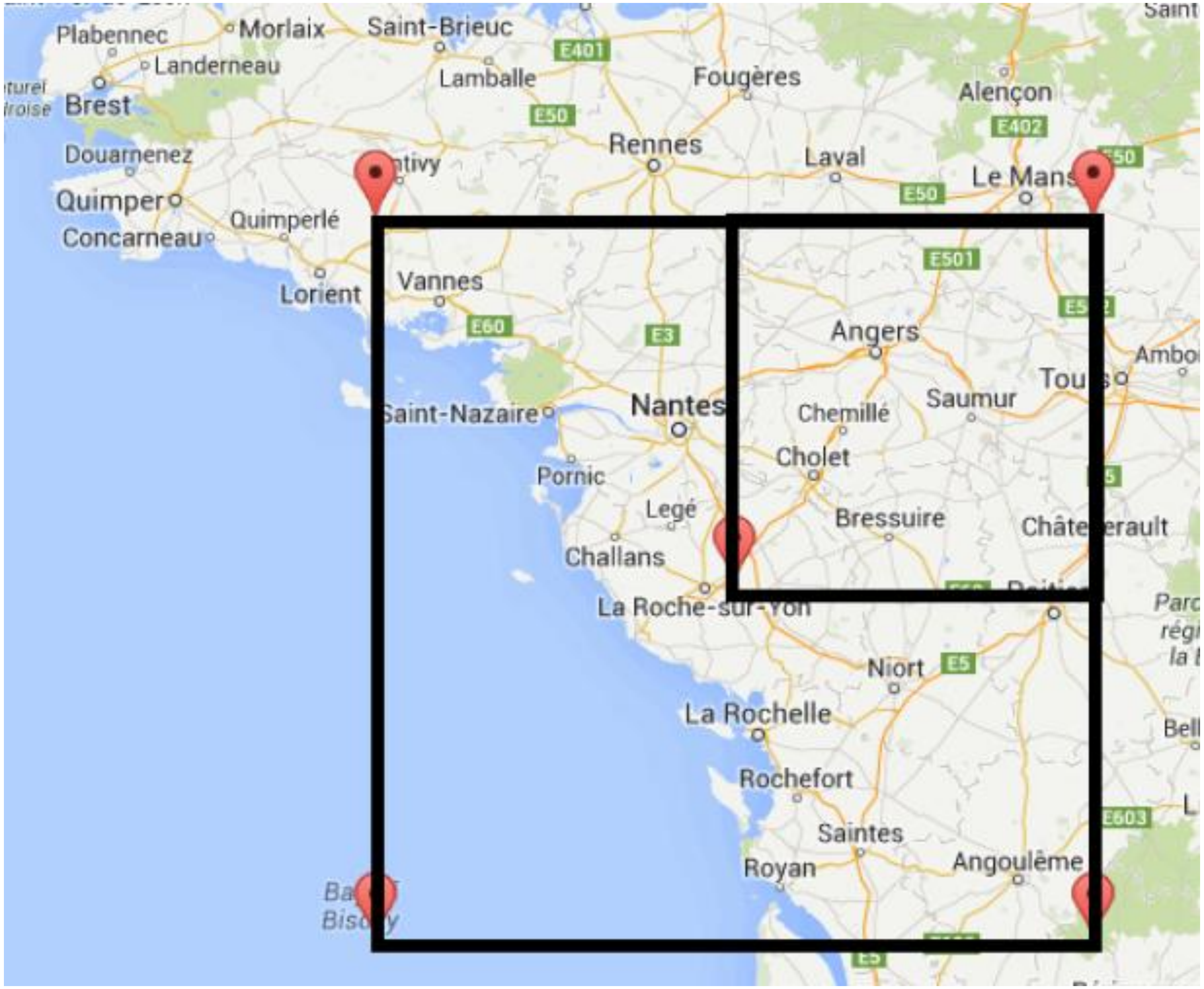
$$\widehat{F}^\star \begin{pmatrix} P \\ t+h \end{pmatrix} = F^\star \begin{pmatrix} P - h\,\vec{W}\begin{pmatrix} P \\ t \end{pmatrix} \\ t \end{pmatrix}$$

- **Estimation of advection vector**

$$\vec{W}\begin{pmatrix} P \\ t \end{pmatrix} = \underset{\vec{W}(P,t)\in\mathcal{V}\left(\vec{W}\begin{pmatrix} P \\ t-\delta t \end{pmatrix}\right)}{\operatorname{argmin}} \left[\sum_{\substack{P'\in\mathcal{V}(P) \\ t'\in\mathcal{V}(t)}} Erreur\left(F^\star \begin{pmatrix} P' - h\vec{W} \\ t' - h \end{pmatrix}, F^\star \begin{pmatrix} P' \\ t' \end{pmatrix} \right) \right]$$

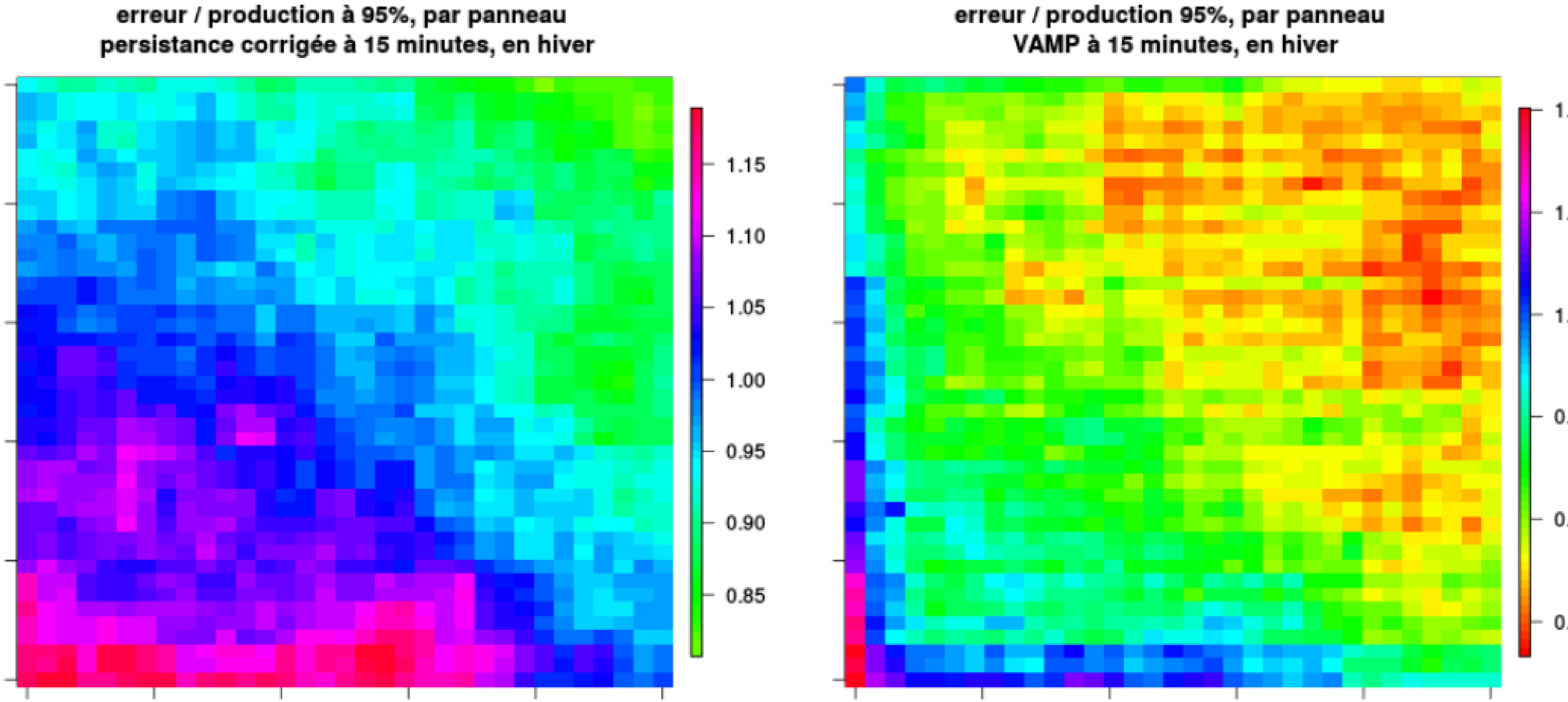
EVALUATION ON CLOUD COVERAGE DATABASE

- Météo France
- Satellite observation
- 43 * 33 pixels
- 1 pixel: 3.3km x 4.5 km
- Cloud type: every 15'
- 2011 - 2012



EVALUATION ON CLOUD COVERAGE DATABASE

- **Comparison with corrected persistence**
- **Slight improvement**



OPEN QUESTIONS

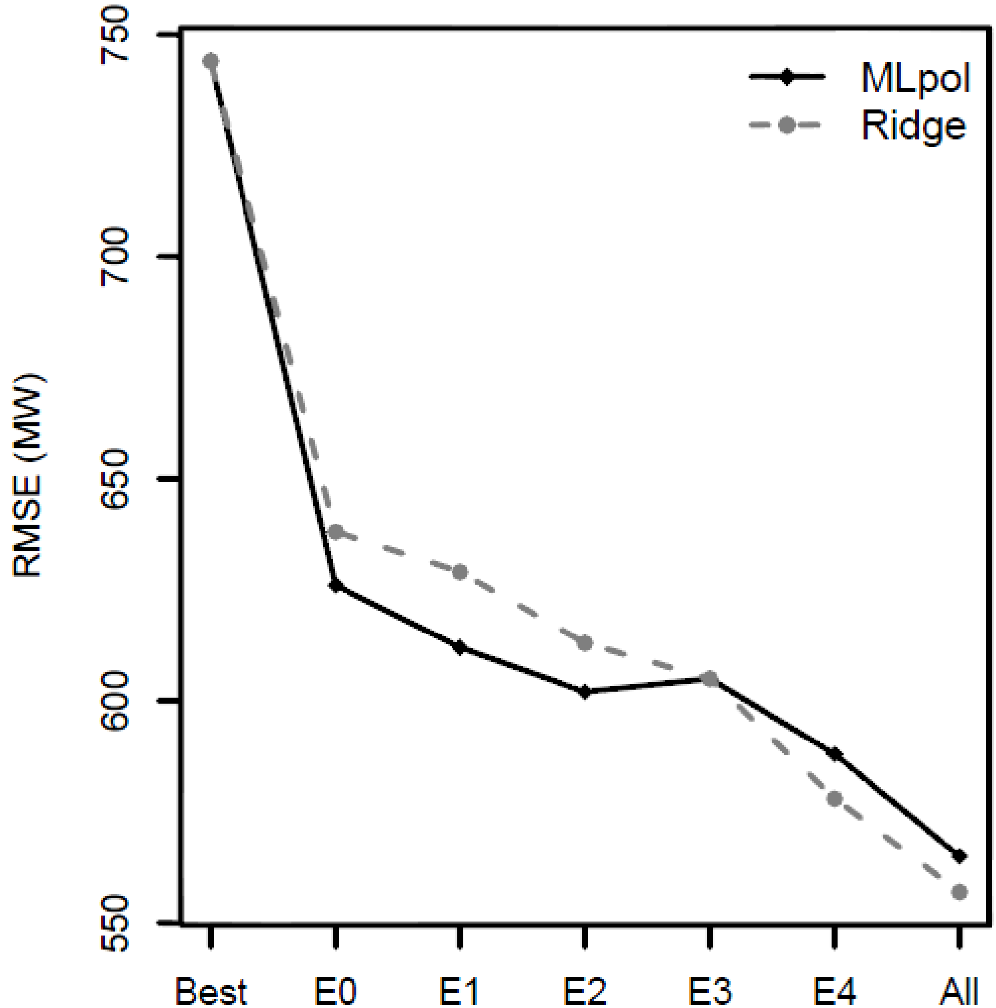
- **Spatio-temporal data models**
  - Spatio-temporal dynamics for PV, Wind power forecasting
  - Statistical meteorological forecasts (T°, solar radiation, wind), specially probabilistic forecasts
- **Other probabilistic approaches:**
  - Price forecasting: non-Gaussian GAMs (GAM lss…)
- **Non-linear quantile regression with additive models:**
  - A need for theoretical work on our 2 step procedure
  - **work on real quantile GAM**, not only linear adjustement of the median model
  - **Adaptive** (with time) forecasts to deal with breaks, data flow (related to BAM)
- **Improvement are coming from agregation of experts (specially on short horizons)**
  - Derive experts from GAMs (bagging, covariate selection, focus on special periods…)
  - machine learning method (Random forest, gradient boosting machine, deep learing)

ANNEXES

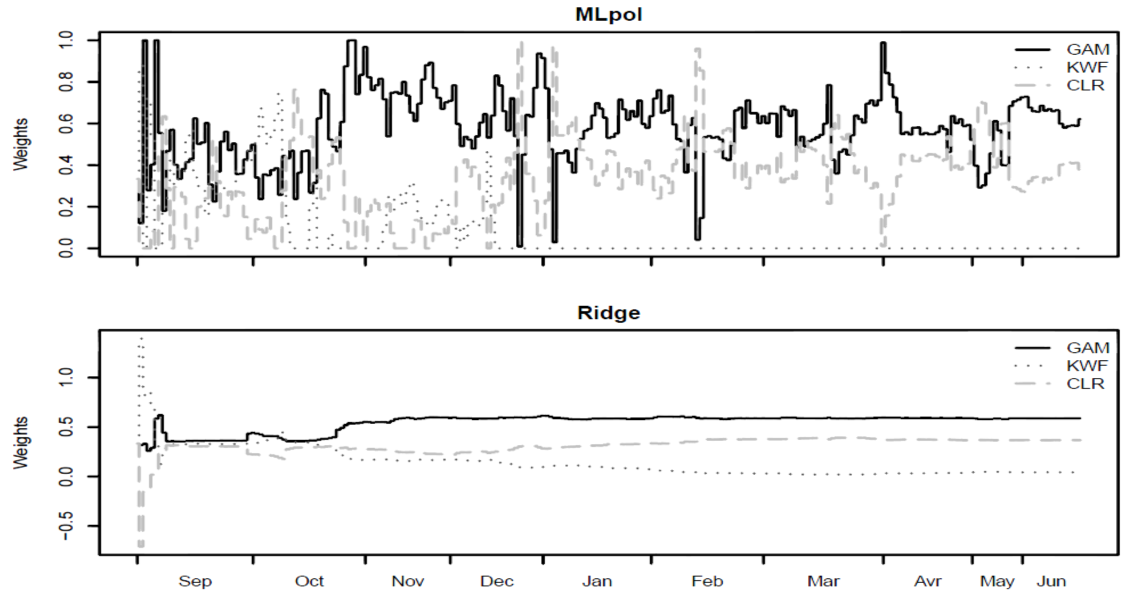
AGGREGATION

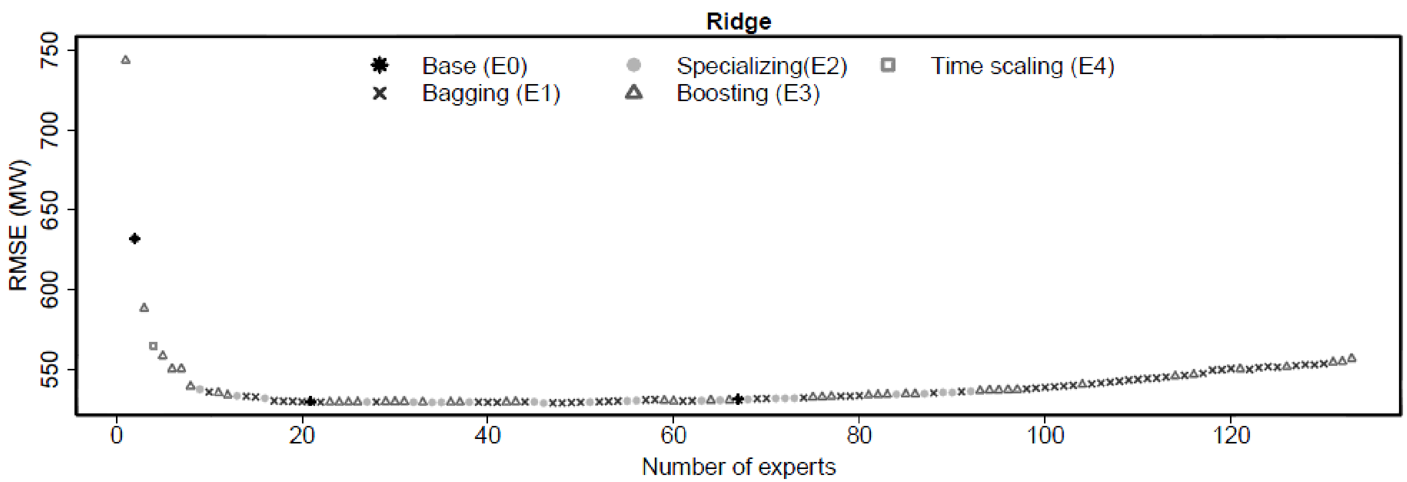
Combining
gain de 110MW

Designing experts
gain de 90MW



AGGREGATION





WEATHER STATION SELECTION

$$Y_t = f_1(Toy_t) + f_2(t) + f_3(T_t) + h(DayType_t) + \varepsilon_t$$

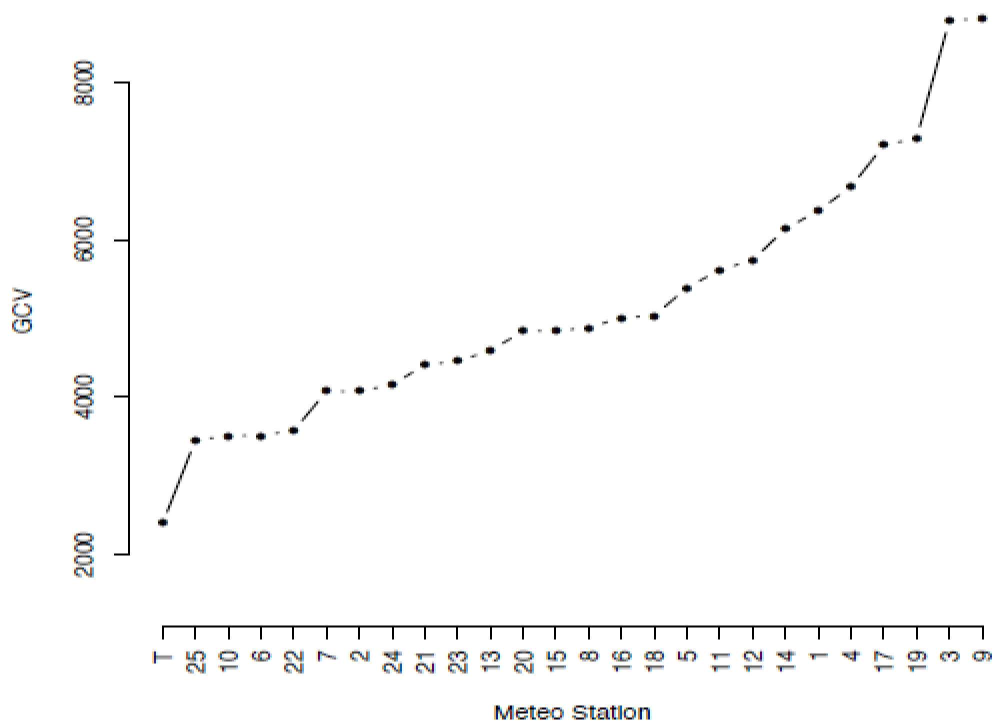


Figure 2: GCV score obtained for each temperature station compared to the one obtained by the average temperature T_t.