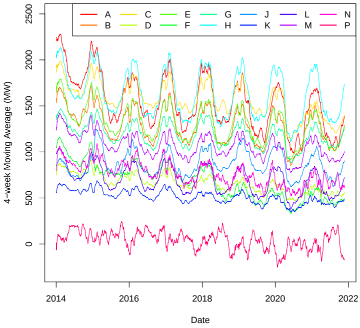
Adaptive Probabilistic Forecasting of Electricity Net-Load

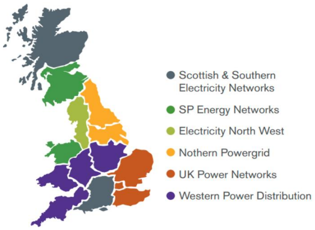
Joseph de Vilmarest,
Jethro Browell, Matteo Fasiolo, Yannig Goude, Olivier Wintenberger

INREC, Essen - September 27-28, 2022

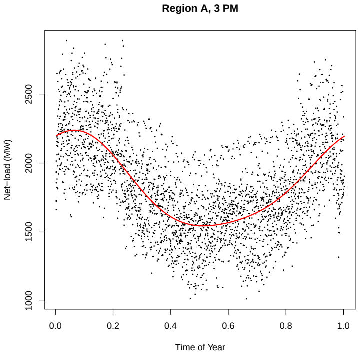
Regional Net-load Forecasting

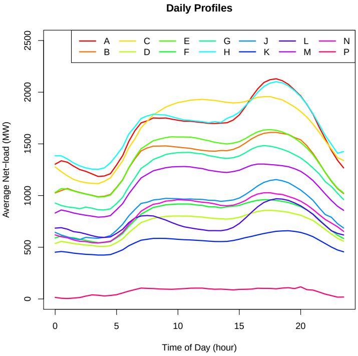
We forecast $y_t \in \mathbb{R}$. Our setting: 14 time series.
net-load = consumption - *intermittent* production.



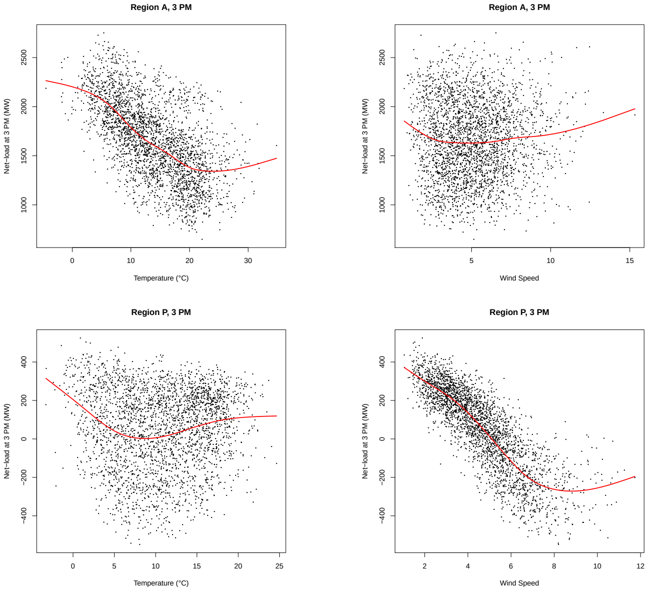


Explanatory Variables: Calendar





Explanatory Variables: Meteorology



Objective

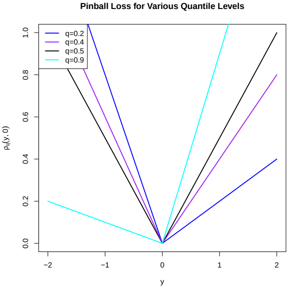
We forecast y_t given x_t. In what sense ?

- **Mean** forecast: $\hat{y}_t = \mathbb{E}[y_t \mid x_t]$.
  Equivalent to the minimum of $\mathbb{E}[(y_t - \hat{y}_t)^2 \mid x_t]$.

Objective

We forecast y_t given x_t. In what sense ?

- **Mean** forecast: $\hat{y}_t = \mathbb{E}[y_t \mid x_t]$.
  Equivalent to the minimum of $\mathbb{E}[(y_t - \hat{y}_t)^2 \mid x_t]$.
- **Probabilistic** forecast: estimation of $\mathcal{L}(y_t \mid x_t)$.
  For $0 < q < 1$, we find $\hat{y}_{t,q}$ such that $\mathbb{P}(y_t \leq \hat{y}_{t,q} \mid x_t) = q$.
  Equivalent to the minimum of $\mathbb{E}[\rho_q(y_t, \hat{y}_t) \mid x_t]$:



Offline vs Online

- **Offline / Batch**: $\hat{y}_t = f_{\hat{\theta}}(x_t)$.
  Example: Empirical Risk Minimizer

$$\hat{\theta} \in \arg\min \sum_{t \in \mathcal{T}} \ell(y_t, f_{\hat{\theta}}(x_t)) \,.$$

Offline vs Online

- **Offline / Batch**: $\hat{y}_t = f_{\hat{\theta}}(x_t)$.
  Example: Empirical Risk Minimizer
  $$\hat{\theta} \in \arg\min \sum_{t \in \mathcal{T}} \ell(y_t, f_{\hat{\theta}}(x_t))\,.$$
- **Online / Adaptive**: $\hat{y}_t = f_{\hat{\theta}_t}(x_t)$ with $\hat{\theta}_{t+1} = \Phi(\hat{\theta}_t, x_t, y_t)$.
  Example: Online Gradient Descent
  $$\hat{\theta}_{t+1} = \hat{\theta}_t - \gamma_t \frac{\partial \ell(y_t, f_\theta(x_t))}{\partial \theta}\bigg|_{\hat{\theta}_t}\,.$$

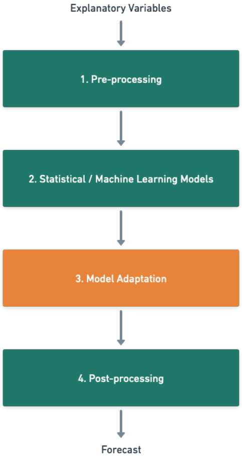
Offline vs Online

- **Offline / Batch**: $\hat{y}_t = f_{\hat{\theta}}(x_t)$.
  Example: Empirical Risk Minimizer

$$\hat{\theta} \in \arg\min \sum_{t \in \mathcal{T}} \ell(y_t, f_{\hat{\theta}}(x_t)) \,.$$

- **Online / Adaptive**: $\hat{y}_t = f_{\hat{\theta}_t}(x_t)$ with $\hat{\theta}_{t+1} = \Phi(\hat{\theta}_t, x_t, y_t)$.
  Example: Online Gradient Descent

$$\hat{\theta}_{t+1} = \hat{\theta}_t - \gamma_t \frac{\partial \ell(y_t, f_\theta(x_t))}{\partial \theta}\bigg|_{\hat{\theta}_t} \,.$$



Offline Model in Two Steps[1]

- Generalized Additive Model with Gaussian distribution for **mean** forecasting:

$$y_t = f_1(x_{t,1}) + \ldots + f_d(x_{t,d}) + \varepsilon_t\,, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma^2)\,.$$

$f_1, \ldots, f_d$: decomposed on spline basis:

$$f_j(x) = \sum_{k=1}^{m_j} \beta_{j,k} B_{j,k}(x)\,.$$

[1] J. Browell and M. Fasiolo (2021), Probabilistic Forecasting of Regional Net-load with Conditional Extremes and Gridded NWP, *IEEE Transactions on Smart Grid*

Offline Model in Two Steps[1]

- Generalized Additive Model with Gaussian distribution for **mean** forecasting:

$$y_t = f_1(x_{t,1}) + \ldots + f_d(x_{t,d}) + \varepsilon_t\,, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma^2)\,.$$

$f_1, \ldots, f_d$: decomposed on spline basis:

$$f_j(x) = \sum_{k=1}^{m_j} \beta_{j,k} B_{j,k}(x)\,.$$

- **Probabilistic** forecasting: quantile regressions on the residuals because the Gaussian assumption is not satisfied in practice:

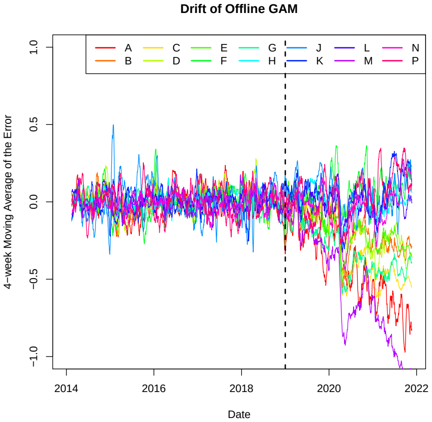
$$\beta_q \in \arg\min_{\beta \in \mathbb{R}^{d_0}} \sum_{t \in \mathcal{T}} \rho_q(y_t - \hat{y}_t, \beta^\top z_t)\,,$$

$$\rho_q(y, \hat{y}_q) = (\mathbb{1}_{y < \hat{y}_q} - q)\,(\hat{y}_q - y)\,.$$

[1] J. Browell and M. Fasiolo (2021), Probabilistic Forecasting of Regional Net-load with Conditional Extremes and Gridded NWP, *IEEE Transactions on Smart Grid*

Motivation for Adaptation

Train: 2014-2018. Test: 2019-2021.



Introduction

Mean Forecast

Probabilistic Forecast

Linear Gaussian State-Space Model

- GAM:

$$y_t - \mathbf{1}^\top f(x_t) \sim \mathcal{N}(0, \sigma^2)\,.$$

Linear Gaussian State-Space Model

- GAM:

$$y_t - \mathbf{1}^\top f(x_t) \sim \mathcal{N}(0, \sigma^2)\,.$$

- State-Space Model

$$\begin{aligned} y_t - {\theta_t}^\top f(x_t) &\sim \mathcal{N}(0, \sigma_t^2)\,, \\ \theta_t - \theta_{t-1} &\sim \mathcal{N}(0, Q_t)\,. \end{aligned}$$

Linear Gaussian State-Space Model

- GAM:

$$y_t - \mathbf{1}^\top f(x_t) \sim \mathcal{N}(0, \sigma^2)\,.$$

- State-Space Model

$$\begin{aligned} y_t - \theta_t^\top f(x_t) &\sim \mathcal{N}(0, \sigma_t^2)\,, \\ \theta_t - \theta_{t-1} &\sim \mathcal{N}(0, Q_t)\,. \end{aligned}$$

Theorem (R. Kalman and R. Bucy, 1961)

If the state-space model is well-specified for known variances, and if $\theta_1 \sim \mathcal{N}(\hat\theta_1, P_1)$, *then* $\theta_{t+1} \mid (x_s, y_s)_{s \le t} \sim \mathcal{N}(\hat\theta_{t+1}, P_{t+1})$ *with*

$$P_{t|t} = P_t - \frac{P_t f(x_t) f(x_t)^\top P_t}{f(x_t)^\top P_t f(x_t) + \sigma_t^2}\,, \qquad P_{t+1} = P_{t|t} + Q_{t+1}\,,$$

$$\hat\theta_{t+1} = \hat\theta_t - \frac{P_{t|t}}{\sigma_t^2}\Big(f(x_t)(\hat\theta_t^\top f(x_t) - y_t)\Big)\,.$$

The Kalman Filter, a Gradient Algorithm

$$P_{t|t} = P_t - \frac{P_t f(x_t) f(x_t)^\top P_t}{f(x_t)^\top P_t f(x_t) + \sigma_t^2}\,, \qquad P_{t+1} = P_{t|t} + Q_{t+1}\,,$$
$$\hat{\theta}_{t+1} = \hat{\theta}_t - \frac{P_{t|t}}{\sigma_t^2}\Big(f(x_t)(\hat{\theta}_t^\top f(x_t) - y_t)\Big)\,.$$

1. **Static**[2]: $Q_t = 0, \sigma_t^2 = 1$.
   $\to P_{t|t} = O(1/t)$.
2. **Dynamic** with constant variances: $Q_t = Q, \sigma_t^2 = \sigma^2$.
   $\to P_{t|t} = O(1)$. Comparable to Adam, AdaGrad.
3. **Variance Tracking**: dynamic with adaptive variances[3].

[2] J. de Vilmarest, O. Wintenberger (2021), Stochastic Online Optimization using Kalman Recursion. *Journal of Machine Learning Research*

[3] J. de Vilmarest, O. Wintenberger (2021), Viking: Variational Bayesian Variance Tracking, *arXiv:2104.10777*

Constant Variances

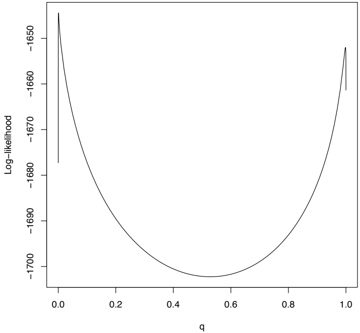
$$y_t - \theta_t^\top f(x_t) \sim \mathcal{N}(0, \sigma^2)\,,$$
$$\theta_t - \theta_{t-1} \sim \mathcal{N}(0, Q)\,.$$

[3] D. Obst, J. de Vilmarest, Y. Goude (2021), Adaptive methods for short-term electricity load forecasting during COVID-19 lockdown in France, *IEEE Transactions on Power Systems*

Constant Variances

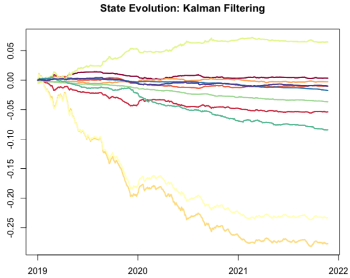
$$y_t - \theta_t^\top f(x_t) \sim \mathcal{N}(0, \sigma^2)\,,$$
$$\theta_t - \theta_{t-1} \sim \mathcal{N}(0, Q)\,.$$

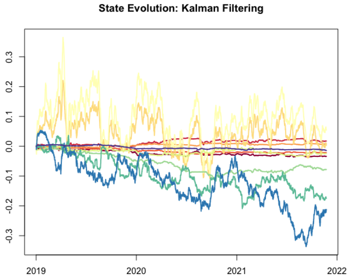
- Non convex log-likelihood.
  No guarantee of optimality.
- Diagonal Covariance Matrix Q.
  Optimization with *iterative grid search*[1].



[3]D. Obst, J. de Vilmarest, Y. Goude (2021), Adaptive methods for short-term electricity load forecasting during COVID-19 lockdown in France, *IEEE Transactions on Power Systems*

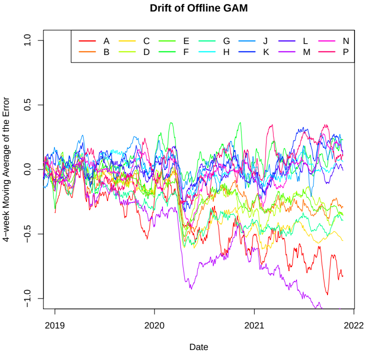
Coefficient Evolution

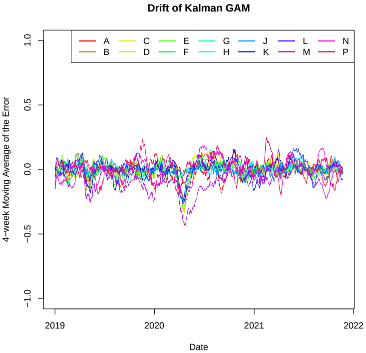




Static setting (left): $\theta_{t+1} = \theta_t$. $P_{t|t} = O(1/t)$.
Dynamic setting (right): $\theta_{t+1} - \theta_t \sim \mathcal{N}(0, Q)$. $P_{t|t} = O(1)$.

Correction of the Drift





Performances

$$RMSE = \sqrt{\frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}}(y_t - \hat{y}_t)^2}\,, \qquad MAE = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}}|y_t - \hat{y}_t|$$

	2019		2020		2021	
Forecast	nRMSE	nMAE	nRMSE	nMAE	nRMSE	nMAE
Persistence (7 days)	0.691	0.589	0.710	0.599	0.737	0.639
Persistence (2 days)	0.767	0.686	0.755	0.668	0.736	0.668
Offline GAM	0.356	0.327	0.485	0.453	0.635	0.601
Incremental offline GAM (yearly)	-	-	0.407	0.376	0.387	0.378
Incremental offline GAM (daily)	0.338	0.307	0.370	0.344	0.377	0.365
Kalman GAM (Static)	0.337	0.307	0.374	0.347	0.380	0.368
Kalman GAM (Dynamic)	**0.324**	**0.292**	**0.328**	**0.301**	**0.332**	**0.307**

Introduction

Mean Forecast

Probabilistic Forecast

Probabilistic Forecast using the Kalman Filter

Under the state-space assumption: $\theta_t \mid (x_s, y_s)_{s<t} \sim \mathcal{N}(\hat{\theta}_t, P_t)$ and $y_t - \theta_t^\top f(x_t) \sim \mathcal{N}(0, \sigma^2)$.

Probabilistic Forecast using the Kalman Filter

Under the state-space assumption: $\theta_t \mid (x_s, y_s)_{s<t} \sim \mathcal{N}(\hat{\theta}_t, P_t)$ and $y_t - \theta_t^\top f(x_t) \sim \mathcal{N}(0, \sigma^2)$.

- **If the model is well-specified**:

$$y_t \sim \mathcal{N}(\hat{\theta}_t^\top f(x_t), \sigma^2 + f(x_t)^\top P_t f(x_t)) .$$

Probabilistic Forecast using the Kalman Filter

Under the state-space assumption: $\theta_t \mid (x_s, y_s)_{s<t} \sim \mathcal{N}(\hat{\theta}_t, P_t)$ and $y_t - \theta_t^\top f(x_t) \sim \mathcal{N}(0, \sigma^2)$.

- **If the model is well-specified**:

$$y_t \sim \mathcal{N}(\hat{\theta}_t^\top f(x_t), \sigma^2 + f(x_t)^\top P_t f(x_t))\,.$$

- **In practice**: mean forecast, then quantile regressions on the residuals $y_t - \hat{\theta}_t^\top f(x_t)$.
  $\rightarrow$ adaptive quantile regression ?

Adaptive Quantile Regression

Offline quantile regression:

$$\beta_q \in \arg\min_{\beta \in \mathbb{R}^{d_0}} \sum_{t \in \mathcal{T}} \rho_q(y_t - \hat{y}_t, \beta^\top z_t)\,.$$

Adaptive Quantile Regression

Offline quantile regression:

$$\beta_q \in \arg\min_{\beta \in \mathbb{R}^{d_0}} \sum_{t \in \mathcal{T}} \rho_q(y_t - \hat{y}_t, \beta^\top z_t) \,.$$

Online Gradient Descent with step size $\alpha > 0$:

$$\beta_{t+1,q} = \beta_{t,q} - \alpha \frac{\partial \rho_q(y_t - \hat{y}_t, \beta^\top z_t)}{\partial \beta}\bigg|_{\beta_{t,q}} \,,$$

where $\frac{\partial \rho_q(y_t - \hat{y}_t, \beta^\top z_t)}{\partial \beta}\Big|_{\beta_{t,q}} = (\mathbb{1}_{y_t < \hat{y}_t + \beta_{t,q}^\top z_t} - q)\; z_t$.

Adaptive Quantile Regression

Offline quantile regression:

$$\beta_q \in \arg\min_{\beta \in \mathbb{R}^{d_0}} \sum_{t \in \mathcal{T}} \rho_q(y_t - \hat{y}_t, \beta^\top z_t)\,.$$

Online Gradient Descent with step size $\alpha > 0$:

$$\beta_{t+1,q} = \beta_{t,q} - \alpha \frac{\partial \rho_q(y_t - \hat{y}_t, \beta^\top z_t)}{\partial \beta}\Big|_{\beta_{t,q}},$$

where $\frac{\partial \rho_q(y_t - \hat{y}_t, \beta^\top z_t)}{\partial \beta}\Big|_{\beta_{t,q}} = (\mathbb{1}_{y_t < \hat{y}_t + \beta_{t,q}^\top z_t} - q)\; z_t$.

$\rightarrow$ choice of α ?

Aggregation of Experts

- We use different step sizes α_k, typically 10^k.

[4] O. Wintenberger (2017), Optimal learning with Bernstein online aggregation, *Machine Learning*

Aggregation of Experts

- We use different step sizes α_k, typically 10^k.
- Experts $\hat{y}_{t,q}^{(k)}$ obtained from α_k.

[4] O. Wintenberger (2017), Optimal learning with Bernstein online aggregation, *Machine Learning*

Aggregation of Experts

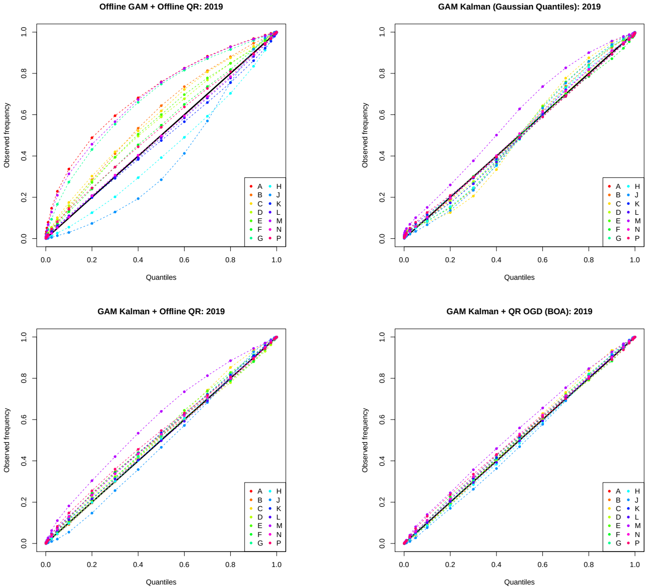
- We use different step sizes α_k, typically 10^k.
- Experts $\hat{y}_{t,q}^{(k)}$ obtained from α_k.
- Aggregation of Experts: Bernstein Online Aggregation[4]:

$$\hat{y}_{t,q} = \sum_k p_t^{(k)} \hat{y}_{t,q}^{(k)} \,,$$

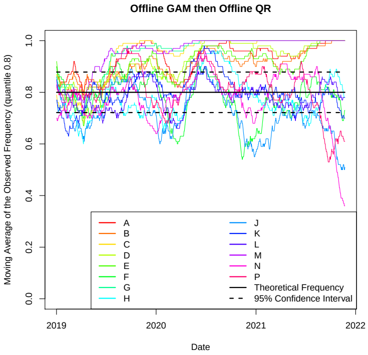
where $p_t^{(k)}$ is obtained sequentially.

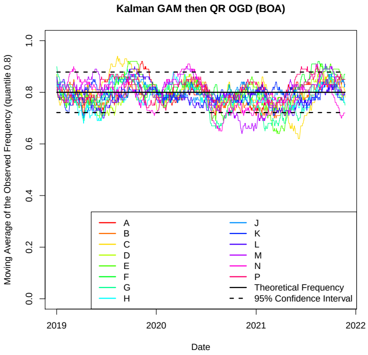
[4] O. Wintenberger (2017), Optimal learning with Bernstein online aggregation, *Machine Learning*

Reliability



Reliability over Time





Evaluation Metric

We use the *continuous ranked probability score*[5]:

$$CRPS(F, y) = \int_{-\infty}^{+\infty} (F(x) - \mathbb{1}_{y \leq x})^2 dx = 2 \int_0^1 \rho_q(y, F^{-1}(q)) dq \,.$$

Discrete variant:

$$RPS\big((\hat{y}_{q_1}, \dots, \hat{y}_{q_I}), y\big) = \sum_{i=1}^{I} \rho_{q_i}(y, \hat{y}_{q_i})(q_{i+1} - q_{i-1}) \,,$$

[5] T. Gneiting and A. E. Raftery (2007), Strictly proper scoring rules, prediction, and estimation, *Journal of the American statistical Association*

Performances

	2019	2020	2021
Offline Method	0.231	0.338	0.454
GAM Kalman (Gaussian Quantiles)	0.212	0.217	0.222
GAM Kalman + Offline QR	**0.206**	**0.214**	**0.217**
Offline GAM + QR OGD (10^{-3})	0.218	0.270	0.293
Offline GAM + QR OGD (10^{-2})	0.207	0.221	0.218
Offline GAM + QR OGD (10^{-1})	0.250	0.248	0.293
Offline GAM + QR OGD (BOA)	0.204	0.211	0.216
GAM Kalman + QR OGD (10^{-2})	0.205	0.204	0.212
GAM Kalman + QR OGD (BOA)	**0.202**	**0.201**	**0.209**

Conclusion

- Linear Gaussian state-space model: an adaptive mean forecaster. Interpretation as a gradient algorithm.
- Similar algorithm for probabilistic forecasting: Online Gradient Descent.

Future work (Viking Conseil):

- Extreme Forecasts Evaluation.
- Definition of covariates: GAM, neural network.
- Choice of the variances (Variance Tracking).