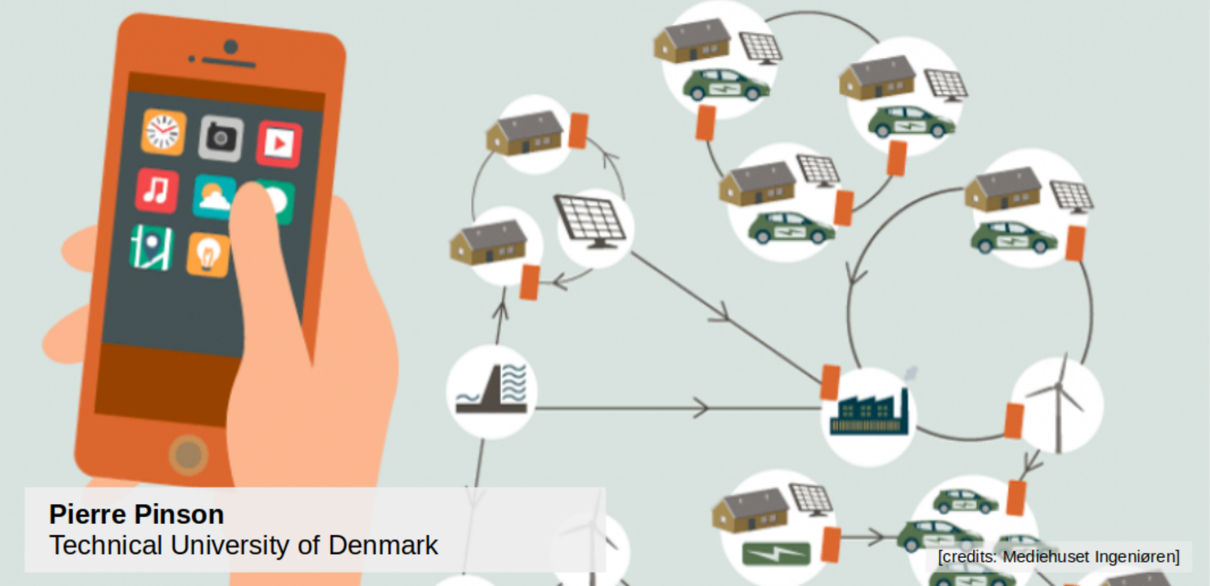
Module 8 – Verification of Renewable Energy Forecasts

8.3 Verification of probabilistic forecasts

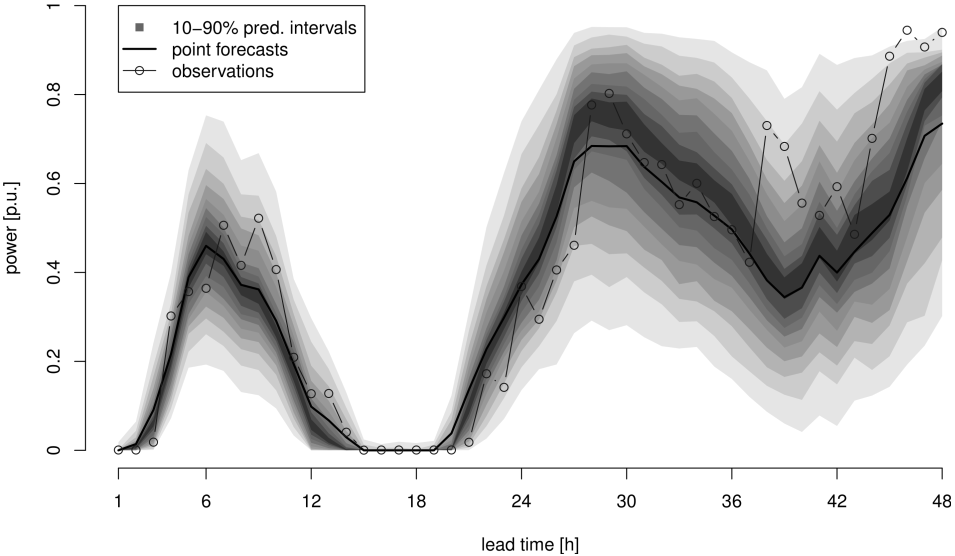


[credits: Mediehuset Ingeniøren]

Pierre Pinson
Technical University of Denmark

Well... it is a bit more difficult

- Evaluating probabilistic forecasts is more involved than evaluating point predictions!
- *Can you tell if this single forecast is good or not?*



Attributes of probabilistic forecast quality

How do you want your forecasts?

- *Reliable?* (also referred to as "probabilistic calibration")
- *Sharp?* (i.e., informative)
- *Skilled?* (all-round performance, and of higher quality than some benchmark)
- Of high *resolution?* (i.e., resolving among situations with various uncertainty levels)
- etc.

Probabilistic calibration

- *Calibration* is about **respecting the probabilistic contract:**
  - for a *quantile forecast* $\hat{q}_{t+k|t}^{(\alpha)}$ with nominal level $\alpha = 0.5$, one expect that the observations y_{t+k} are to be less than $\hat{q}_{t+k|t}^{(\alpha)}$ 50% of the times
  - for an *interval forecast* $\hat{I}_{t+k|t}^{(\beta)}$ with nominal coverage rate $\beta = 0.9$, one expect that the observations y_{t+k} are to be covered by $\hat{I}_{t+k|t}^{(\beta)}$ 90% of the times
  - further than that, since an *interval forecast* $\hat{I}_{t+k|t}^{(\beta)}$ is composed by two quantile forecasts with nominal levels $\underline{\alpha}$ and $\overline{\alpha}$, one evaluates these two quantile forecasts
  - finally for *predictive densities* $\hat{F}_{t+k|t}$, composed by a number $m+1$ of quantile forecasts with nominal levels $\{\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_m\}$, all these quantile forecasts are evaluated, individually
- To do it in practice, we take a *frequentist approach*... **we simply count!**

Assessing calibration

For a given quantile forecast $\hat{q}^{(\alpha)}_{t+k|t}$ and the corresponding observation y_{t+k}, the *indicator variable* $\xi^{(\alpha)}_{t,k}$ is given by

$$\xi^{(\alpha)}_{t,k} = \mathbf{1}\{y_{t+k} < \hat{q}^{(\alpha)}_{t+k|t}\} = \begin{cases} 1, & \text{if} \quad y_{t+k} < \hat{q}^{(\alpha)}_{t+k|t} \quad \textbf{(HIT)} \\ 0, & \text{otherwise} \quad \textbf{(MISS)} \end{cases}$$

- By counting the number of hits over your set of forecasts, one obtains the *empirical level* of these quantile forecasts

The empirical level $a^{(\alpha)}_k$ is given by the mean of $\xi^{(\alpha)}_{t,k}$ over the set of T quantile forecasts,

$$a^{(\alpha)}_k = \frac{n^{(\alpha)}_k}{T}$$

where $n^{(\alpha)}_k$ is the sum of hits:

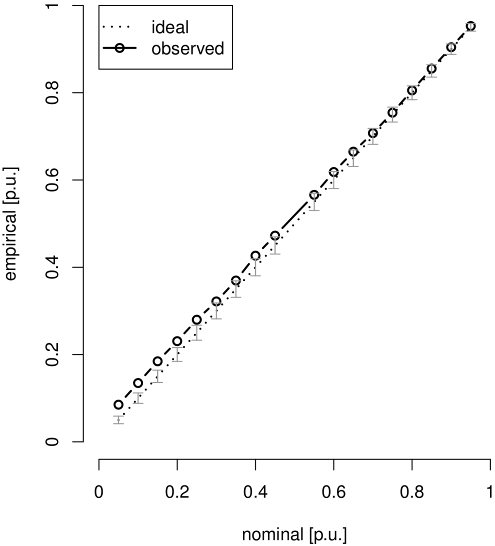
$$n^{(\alpha)}_k = \#\{\xi^{(\alpha)}_{t,k} = 1\} = \sum_{t=1}^{T} \xi^{(\alpha)}_{t,k}$$

Example calibration assessment at Klim with reliability diagrams

- The calibration assessment can be summarized in **reliability diagrams**

- Here example for our probabilistic forecasts at Klim:
    - period: 1.7.2002 - 31.12.2002
    - predictive densities composed by quantile forecasts with nominal levels $\{0.05, 0.1, \ldots, 0.45, 0.55, \ldots, 0.9, 0.95\}$
    - quantile forecasts are evaluated one by one, and their *empirical levels* are reported vs. their *nominal levels*

- **The closest to the diagonal, the better!**



Sharpness

- *Sharpness* is about the **concentration of probability**
- A perfect probabilistic forecast gives a probability of 100% on a single value!
- Consequently, a sharpness assessment boils down to evaluating *how tight the predictive densities are...*

The width of a given interval forecast $\hat{I}^{(\beta)}_{t+k|t}$ is given by the distance between its two bounds

$$\delta^{(\beta)}_{t,k} = \hat{q}^{(\overline{\alpha})}_{t+k|t} - \hat{q}^{(\underline{\alpha})}_{t+k|t}$$

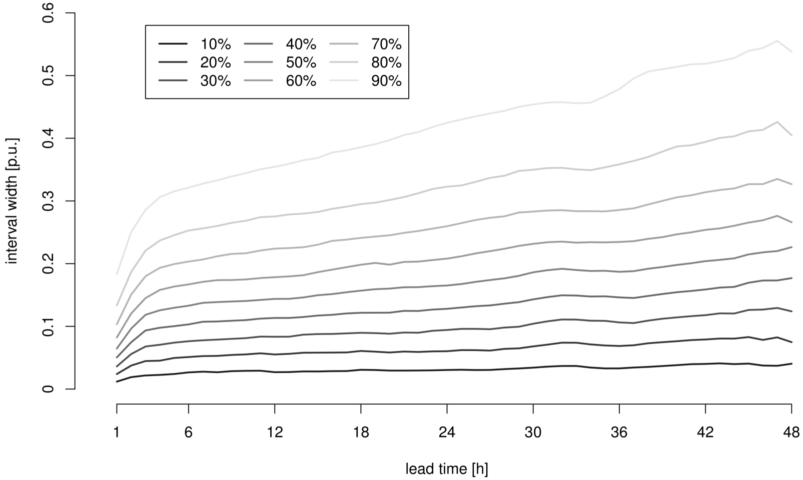
The *sharpness* of these interval forecasts is obtained by calculating their average width over the evaluation period:

$$\bar{\delta}^{(\beta)}(k) = \frac{1}{T} \sum_{t=1}^{T} \delta^{(\beta)}_{t,k}$$

This is done for all the intervals composing the predictive densities

Example: sharpness evaluation at Klim

- Period: 1.7.2012 - 31.12.2012
- Predictive densities are composed by interval forecasts with nominal coverage rates $\beta = 0.1, 0.2, \ldots, 0.9$



- The interval width increases with the lead time, reflecting higher forecast uncertainty

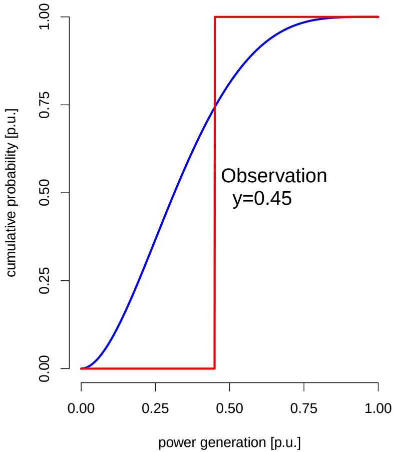
Overall skill assessment

- The *skill* of probabilistic forecasts can be assessed by scores, like MAE and RMSE for the deterministic forecasts.

- The most common *skill score* for predictive densities is the **Continuous Ranked Probability Score (CRPS)**

- For a given predictive density $\hat{F}_{t+k|t}$ and corresponding observation y_{t+k},

$$\text{CRPS}_{t,k} = \int_y \left(\hat{F}_{t+k|t}(y) - \mathbf{1}\{y_{t+k} \leq y\} \right)^2 dy$$

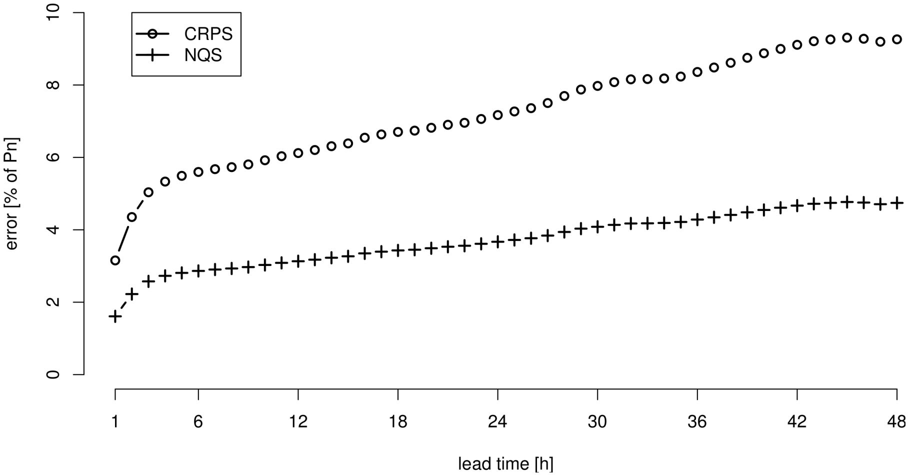


The *CRPS score value* is then given by taking its average for each of the predictive densities and corresponding observation over the evaluation period:

$$\text{CRPS}(k) = \frac{1}{T} \sum_{t=1}^{T} \text{CRPS}_{t,k}$$

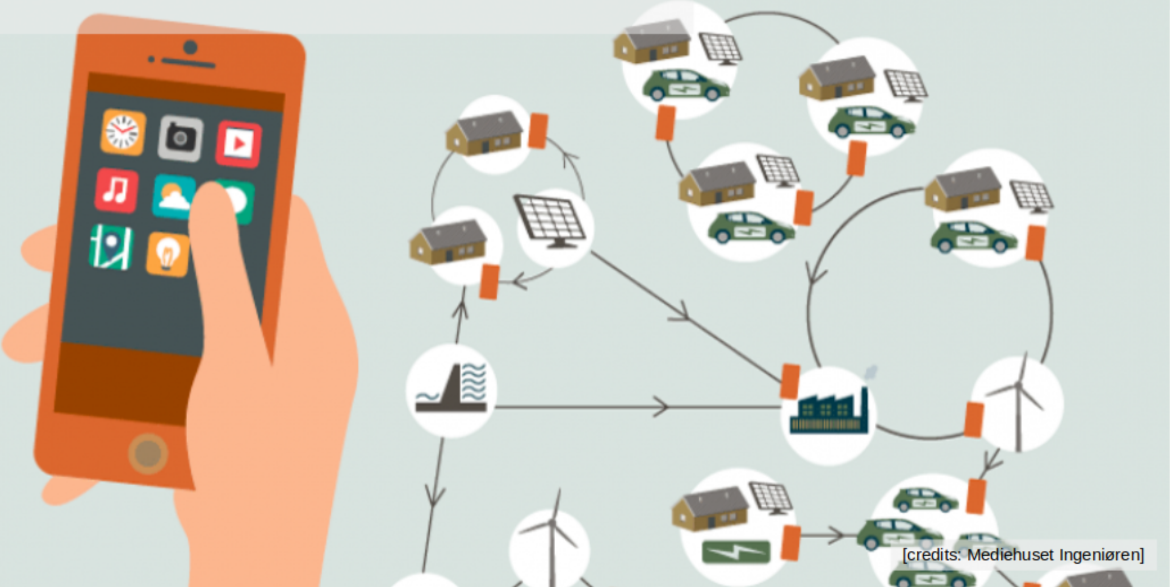
Example: CRPS evaluation at Klim

- Period: 1.7.2012 - 31.12.2012
- Probabilistic forecast quality also degrades with further lead times



- For instance, for 24-ahead forecasts, CRPS is equal to 7% of nominal capacity
- CRPS and MAE (for deterministic forecasts) can be directly compared... This **CRPS value of 7% is better than the MAE value of 8%** in the previous example for deterministic forecasts

Use the self-assessment quizz to check your understanding!



[credits: Mediehuset Ingeniøren]