Verification of Renewable Energy Forecasts

Pierre Pinson

Technical University of Denmark

DTU Electrical Engineering - Centre for Electric Power and Energy
mail: ppin@dtu.dk - webpage: www.pierrepinson.com

YEQT Winter School on Energy Systems - 12 December 2017

Learning objectives

Through this lecture and additional study material, it is aimed for the students to be able to:

1. Explain what makes **renewable energy forecasts** of different quality and value
2. Describe how one may **evaluate the quality** of different forms of forecasts
3. Appraise how different **scores** and **diagnostic tools should be used and interpreted**

A few interesting quotes on forecasting

Some of my favorites:

"Prediction is very difficult, especially if it's about the future"

–Nils Bohr, Nobel laureate in Physics

"Forecasting is the art of saying what will happen, and then explaining why it didn't!"

–Anonymous

"It is far better to foresee even without certainty than not to foresee at all"

–Henri Poincaré

A good sample is gathered at:

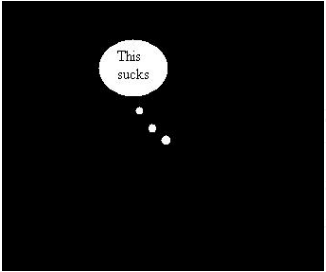
Exeter University - famous forecasting quotes

Let's accept it...

- **Forecasts are always wrong!**
- Bad forecasts translate to **consequences** - these may be:

,

- *security issues* in, e.g., offshore wind farm maintenance
- *financial losses* for those participating in the markets
- *overall decrease in social welfare*



- *blackouts!* (well, hopefully not)
- ... but definitely, *harsh criticism on using renewables for supplying us with electricity*

Outline

1. **What makes a good forecast?**
2. **Test case and general considerations**
3. **Verification of point (deterministic) forecasts**
    - scores
    - diagnostic tools
4. **Verification of probabilistic forecasts**
    - attributes of forecast quality
    - scores
    - diagnostic tools

1. **What makes a good forecast?**

The nature of "goodness" in forecasting

- Following Murphy (ref. and link below), the *nature of "goodness"* in weather forecasting (same goes for other types of forecasts) consists in:

The nature of "goodness" in forecasting

- Following Murphy (ref. and link below), the *nature of "goodness"* in weather forecasting (same goes for other types of forecasts) consists in:

  - **Forecast consistency**:

    > *"Forecasts should correspond to the forecaster's best judgement on future events, based on the knoweldge available at the time of issuing the forecasts"*

The nature of "goodness" in forecasting

- Following Murphy (ref. and link below), the *nature of "goodness"* in weather forecasting (same goes for other types of forecasts) consists in:

  - **Forecast consistency**:

    "Forecasts should correspond to the forecaster's best judgement on future events, based on the knoweldge available at the time of issuing the forecasts"

  - **Forecast quality**:

    "Forecasts should describe future events as good as possible, regardless of what these forecasts may be used for"

The nature of "goodness" in forecasting

- Following Murphy (ref. and link below), the *nature of "goodness"* in weather forecasting (same goes for other types of forecasts) consists in:

  - **Forecast consistency**:

    "Forecasts should correspond to the forecaster's best judgement on future events, based on the knoweldge available at the time of issuing the forecasts"

  - **Forecast quality**:

    "Forecasts should describe future events as good as possible, regardless of what these forecasts may be used for"

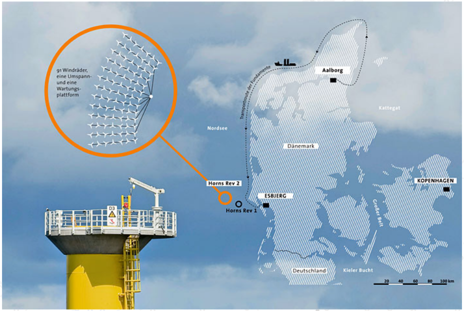
  - **Forecast value**:

    "Forecasts should bring additional benefits (monetary or others) when used as input to decision-making"

[Extra reading:
AH Murphy (1993). What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting* **8**: 281–293 (pdf)]

Illustrative example (1)

- You are in charge of **optimal maintenance planning at Horns Rev**, and have booked both a vessel and an helicopter for onsite service (for a cost of 100.000€)



- The conditions for this to happen at time $t+k$ are
  - wind speed: $u_{t+k} \leq 15$ m.s^{-1}
  - wave height: $h_{t+k} \leq 1.8$ m

- 24 hours before service (time t), this is your last chance to cancel before huge financial penalties (another 100.000€)

- Your two forecasters (*Foresight* and *Blindspot*) tell you that:

	Foresight	*Blindspot*
$\hat{u}_{t+k\|t}$	12.6 m.s^{-1}	3.4 m.s^{-1}
$\hat{h}_{t+k\|t}$	1.6 m	0.2 m

- In both cases, you go ahead with the planned service...

Illustrative example (1, continued)

- At time $t+k$, this is what actually happened:

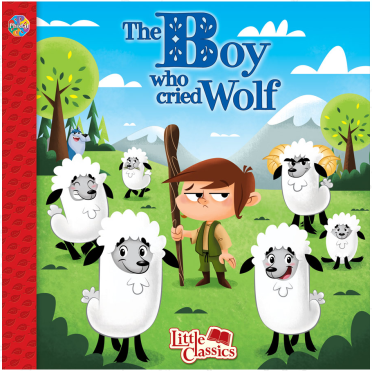
	Foresight	Blindspot
$\hat{u}_{t+k\|t}$	12.6 m.s^{-1}	3.4 m.s^{-1}
$\hat{h}_{t+k\|t}$	1.6 m	0.2 m
u_{t+k}	12.3 m.s^{-1}	
h_{t+k}	1.45 m	

- In both cases, your overall cost is 100.000€,
- Both *Foresight* and *Blindspot* served their purpose, since you made the right decision... **Forecast value is good**
- You might want to have a chat with *Blindspot*, since its **forecast quality appears to be far from good!**

Illustrative example (2)

The boy who cried wolf (Tale from Ancient Greece) - revisited.

- ROGUE TRADING® made huge losses last year, due to expensive upregulation events...
- It is therefore decided to get a new forecaster that would be good at predicting them
- *Foresight* and *Blindspot* are in competition for the job



- The score is simple:

$$\mathrm{Sc} = 100 \cdot \frac{\#\{\text{events leading to upregulation predicted}\}}{\#\{\text{events leading to upregulation}\}}$$

- the higher the better! (0 is worst, 100 is best)

Illustrative example (2, continued)

If you were *Foresight* and *Blindspot*, what would you do?

Illustrative example (2, continued)

If you were *Foresight* and *Blindspot*, what would you do?

- The two competitors have sharpened their strategy:

	Foresight	*Blindspot*
Strategy	Always predict need for upregulation!	Do your best to find when upregulation will occur...

- The results on the benchmarking exercise are such that:
  - $\#\{\text{market time units}\} = 8760$
  - $\#\{\text{events leading to upregulation}\} = 3237$
  - $\#\{\text{events leading to upregulation predicted by Foresight}\} = 3237$
  - $\#\{\text{events leading to upregulation predicted by Blindspot}\} = 2500$

- Their scores:

	Foresight	*Blindspot*
Sc	100%	77.2%

Illustrative example (2, continued)

If you were *Foresight* and *Blindspot*, what would you do?

- The two competitors have sharpened their strategy:

	Foresight	*Blindspot*
Strategy	Always predict need for upregulation!	Do your best to find when upregulation will occur...

- The results on the benchmarking exercise are such that:
  - $\#\{\text{market time units}\} = 8760$
  - $\#\{\text{events leading to upregulation}\} = 3237$
  - $\#\{\text{events leading to upregulation predicted by Foresight}\} = 3237$
  - $\#\{\text{events leading to upregulation predicted by Blindspot}\} = 2500$

- Their scores:

	Foresight	*Blindspot*
Sc	100%	77.2%

- *Foresight* **gets the job!**

Illustrative example (2, continued)

- The consequences are:
  - even though never missing on upregulation events, ROGUE TRADING® will always miss the down regulation ones
  - eventually, the financial loss may still be there... and possibly much higher than expected

Illustrative example (2, continued)

- The consequences are:
  - even though never missing on upregulation events, Rogue Trading® will always miss the down regulation ones
  - eventually, the financial loss may still be there... and possibly much higher than expected

- A more **consistent** way to evaluate these forecasters would be to consider:

	event *happens*	*no* event
event *predicted*	**HIT**	**FALSE ALARM**
event *not predicted*	**MISS**	**CORRECT REJECTION**

- And a *proper* score, ensuring forecast consistency, is:

$$\text{Sc} = 100 \cdot \frac{\#\{\text{hits}\}}{\#\{\text{hits}\} + \#\{\text{misses}\} + \#\{\text{false alarms}\}}$$

- The higher the better! (0 is worst, 100 is best)

  (This score is called the *Threat Score* (TS))

Illustrative example (2, continued)

- In the present case:

	Foresight	*Blindspot*
#{hits}	3237	2320
#{misses}	0	917
#{false alarms}	5523	180
#{correct rejections}	0	5343

- The resulting *Threat Score* (TS) values are:

	Foresight	*Blindspot*
TS	36.9%	67.9%

- **Conclusions:** if using a proper score...
  - *Blindspot* should have gotten the job!
  - I can promise that ROGUE TRADING® would have lower financial losses

2 Test case and general considerations

Test case: the Klim wind farm

- The wind farm:
  - *full name*: Klim Fjordholme
  - *onshore/offshore*: onshore
  - *year of commissioning*: 1996

  - *nominal capacity* (P_n): 21 MW
  - *number of turbines in farm*: 35
  - *average annual electricity generation*: 49 GWh

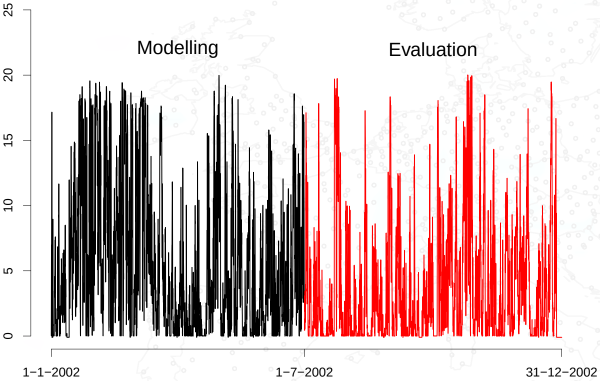
  - *data available*: 1999-2003 (for some researchers)
  - *temporal resolution*: 5 mins, and hourly averages
  - *forecasts*: deterministic and probabilistic

- A link to the online description:
  Vattenfall's Klim wind farm

- The wind farm has been recommissioned recently:
  NordJyske online article

Splitting of available data

- Forecasting is about
  - being able to predict future events, in new situations
  - not only explain what happen in the past...
- **One need to verify forecasts on data that has not been used for the modelling!**



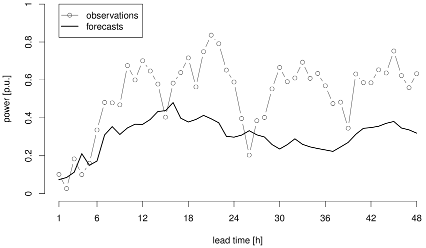
- Here we will focus on the last 6 months of 2002, while giving examples for some other periods

3. **Verification of point (deterministic) forecasts**

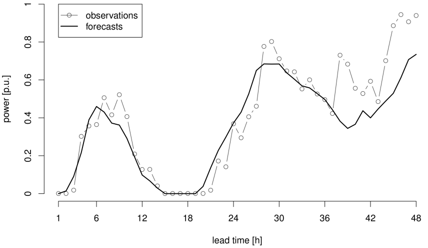
Visual inspection of forecasts

- Visual inspection allows you to develop susbtantial insight on forecast quality...
- This comprises a **qualitative** analysis only

- *What do you think of these two?*
  Are they good or bad?



Forecast issued on 16 November 2001 (18:00)

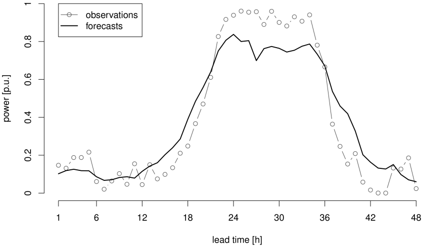


Forecast issued on 23 December 2003 (12:00)

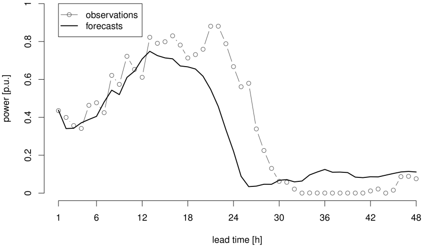
Various types of forecast error patterns

- Errors in renewable energy generation (but also load, price, etc.) are most often driven by weather forecasts errors

- Typical error patterns are:
  - **amplitude errors** (left, below)
  - **phase errors** (right, below)



Forecast issued on 29 March 2003 (12:00)



Forecast issued on 6 November 2002 (00:00)

Quantitative analysis and the forecast error

- For continuous variables such as renewable energy generation (but also electricity prices or electric load for instance)
  - **qualitative** analysis ought to be complemented by a **quantitative** analysis
  - these are based on *scores* and *diagnostic tools*

The base concept is that of the **forecast error**:

$$\varepsilon_{t+k|t} = y_{t+k} - \hat{y}_{t+k|t}, \qquad -\mathrm{P}_n \leq \varepsilon_{t+k|t} \leq \mathrm{P}_n$$

where

- $\hat{y}_{t+k|t}$ is the forecast issued at time t for time $t+k$
- y_{t+k} is the observation at time $t+k$
- P_n is the nominal capacity of the wind farm

- It can be calculated
  - directly for the quantity of interest
  - as a *normalized* version, for instance by dividing by the nominal capacity of the wind farm if evaluating wind power forecasts:

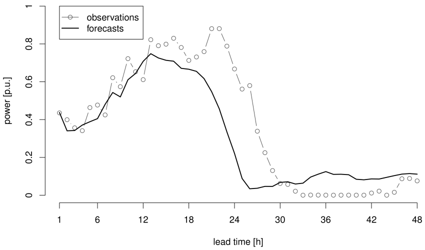
$$\varepsilon_{t+k|t} = \frac{y_{t+k} - \hat{y}_{t+k|t}}{\mathrm{P}_n}, \quad -1 \leq \varepsilon_{t+k|t} \leq 1$$

Forecast error: examples

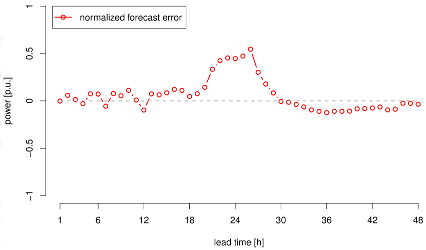
Example 1: If the 24-ahead prediction for Klim is of 18 MW, while the observation is 15.5MW

- $\varepsilon_{t+k|t} = -2.5$MW (if not normalized)
- $\varepsilon_{t+k|t} = -0.119$ (or, -11.9%, if normalized)

Example 2: forecast issued on the 6 November 2002 (00:00)



Forecast and observations



Corresponding forecast errors

(Note that we prefer to work with normalized errors from now on...)

Scores for point forecast verification

- One cannot look at all forecasts, observations, and forecasts errors over a long period of time
- *Scores* are to be used to summarize aspects of forecast accuracy...

The most common scores include, as function of the lead time k:

- **bias** (or Nbias, for the normalized version)

$$\text{bias}(k) = \frac{1}{T} \sum_{t=1}^{T} \varepsilon_{t+k|t}$$

- **Mean Absolute Error (MAE)** (or NMAE, for the normalized version)

$$\text{MAE}(k) = \frac{1}{T} \sum_{t=1}^{T} |\varepsilon_{t+k|t}|$$

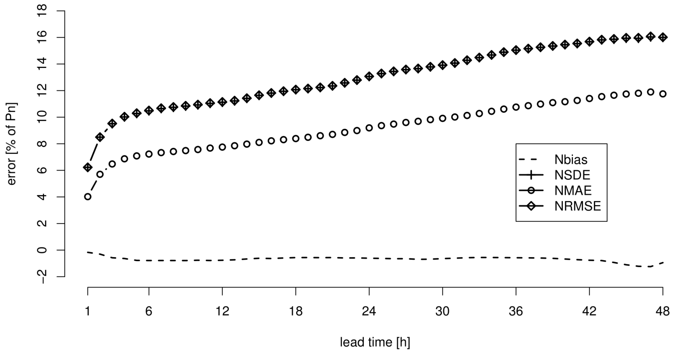
- **Root Mean Square Error (RMSE)** (or NRMSE, for the normalized version)

$$\text{RMSE}(k) = \left[\frac{1}{T} \sum_{t=1}^{T} \varepsilon_{t+k|t}^2 \right]^{\frac{1}{2}}$$

- MAE and RMSE are *negatively-oriented* (the lower, the better)
- Let us illustrate their advantages and drawbacks... (black board illustration)

Example: calculating a few scores at Klim

- Period: 1.7.2012 - 31.12.2012
- Forecats quality necessarily degrades with further lead times



- For instance, for 24-ahead forecasts:
  - bias is close to 0, while NMAE and NRMSE are of 8% and 12%, respectively
  - on average, there is ± 1.68 MW between forecasts and measurements

Comparing against benchmark approaches

- Forecasts from advanced methods are expected to outperform simple benchmarks!
- Two typical benchmarks are (to be further discussed in Lecture 11):
  - **Persistence** ("what you see is what you get"):

    $$\hat{y}_{t+k|t} = y_t, \quad k = 1, 2, \ldots$$

  - **Climatology** (the "once and for all" strategy):

    $$\hat{y}_{t+k|t} = \bar{y}_t, \quad k = 1, 2, \ldots$$

    where $\bar{y}_t$ is the average of all measurements available up to time t

A *skill score* informs of the relative quality of a method vs. a relevant benchmark, for a given lead time k:

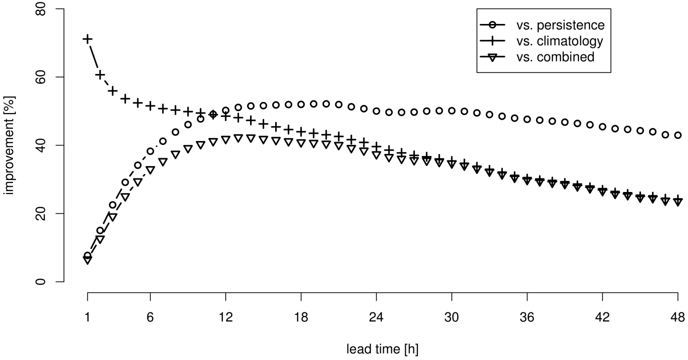
$$\text{SSc}(k) = 1 - \frac{\text{Sc}_{\text{adv}}(k)}{\text{Sc}_{\text{ref}}(k)}, \quad \text{SSc} \leq 1 \qquad \text{(possibly expressed in \%)}$$

where

- 'Sc' can be MAE, RMSE, etc.,
- 'Sc_{adv}' is score value for the advanced method, and
- 'Sc_{ref}' is for the benchmark

Example: benchmarking at Klim

- Great! My forecasts are way better than the benchmarks considered (in terms of RMSE)

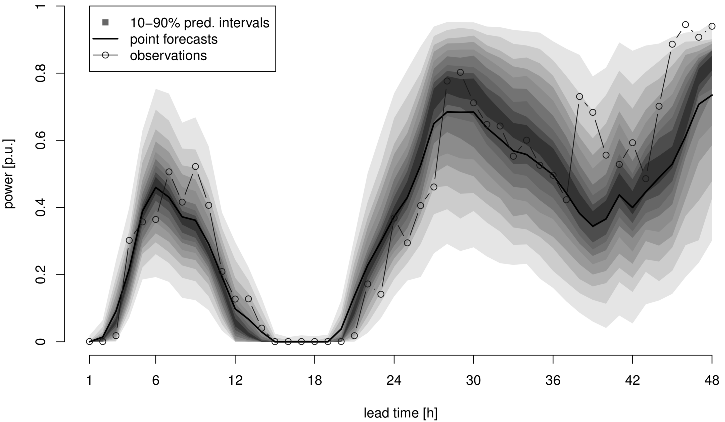


- Additional comments:
  - *persistence* is difficult to outperform for short lead times
  - the opposite holds for *climatology*

4. **Verification of probabilistic forecasts?**

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?*



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 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:

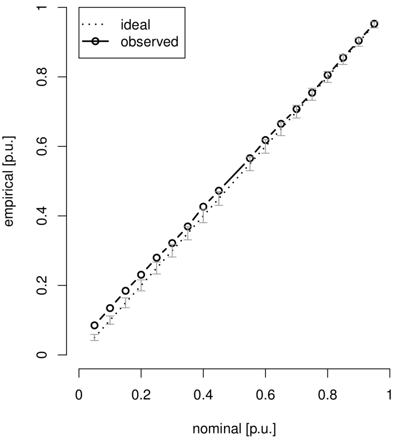
$$n^{(\alpha)}_k = \#\{\xi^{(\alpha)}_{t,k} = 1\} = \sum_{t=1}^{T} \xi^{(\alpha)}_{t,k}$$

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

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

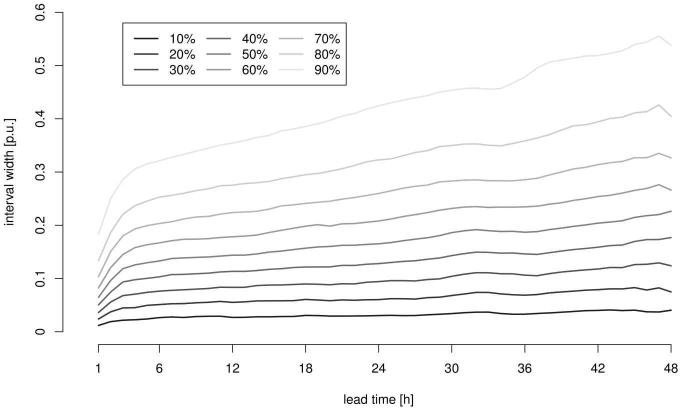
The *sharpness* of these interval forecasts is obtained by calculating their average width over the evaluation period:

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

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 intervals width increase with the lead time, reflecting higher forecast uncertainty

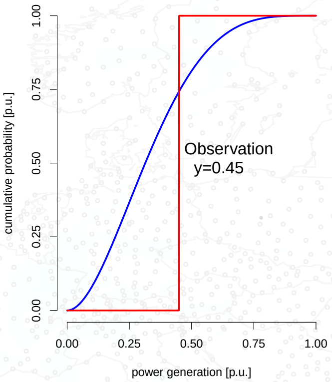
Overall skill assessment

- 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},

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

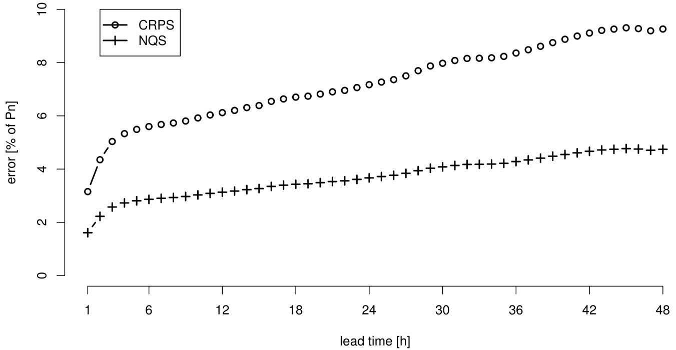


The *CRPS score value* is then given by taking its average for each of the predictive densities and corresponding observation over the evaluation period:

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

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

Now you should be ready to evaluate/handle forecasts in the "real world"!

[Extra reading: Jolliffe IT, Stephenson DB (2011). *Forecast Verification: A Practitioner's Guide in Atmospheric Science* (2nd Ed.). Wiley (link to pdf cannot be provided - available through DTU Findit)]

Thanks for your attention! - Contact: ppin@dtu.dk - web: pierrepinson.com