New scoring methods for weather forecasts

Isabelle Sanchez, Fabien Stoop, Michaël Zamo, Francis Pouponneau, Nicole Girardot
1 - Verification of sensible weather
   - Motivations
   - Calculation
   - Application

2 - Modified CRPS calculation
   - Context
   - Calculation
   - First results

Summary
1 - Verification of sensible weather:
   - Motivations
   - Calculation
   - Application

2 - Modified CRPS calculation
   - Context
   - Calculation
   - First results

Conclusion
Objectives:

- Score for final public production (illustration below)
- Synthetic score
- Values easy to understand
- Linked with user feeling
- Same score from short to medium range (up to day 7)
The goals define the choices:

- Score for final public production → parameters: sensible weather (sun, rain, storm, snow…), total cloud cover, temperature, wind speed
- Synthetic score → aggregation of scores for these parameters
- Easy to understand → value from 0 (bad forecast) to 1 (good forecast)
- Linked with user feeling → put more weight where the error is more sensible (see next slides)
Verification of sensible weather

Calculation for temperature:
(the score goes from 0 for bad forecast to 1 for good forecast)

Value of the score according to the forecast value

<table>
<thead>
<tr>
<th>Score</th>
<th>Observed Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>obs - 3°C</td>
</tr>
<tr>
<td></td>
<td>obs - 1.5°C</td>
</tr>
<tr>
<td></td>
<td>obs</td>
</tr>
<tr>
<td>1</td>
<td>obs + 1.5°C</td>
</tr>
<tr>
<td></td>
<td>obs + 3°C</td>
</tr>
</tbody>
</table>
Calculation for windspeed:

The principle is the same as for temperature, except that the width of the tolerance interval enlarges when the windspeed becomes stronger.

Figure 3: Note du taux de réussite pour les FF.
Verification of sensible weather

Calculation for weather type:

Problem: the forecast classes of weather change with the time range.
In order to have a score valid for all time ranges, 6 weather types are considered:
- Clear sky
- Cloudy sky
- Rain
- Snow
- Thunderstorm
- Fog / mist

For each type, the forecast is expressed as a probability: Pf
the observation, Po, is 0 or 1.

The score for each type is: \[|1 - (Pf + Po)|\]

The final score for sensible weather is the worst of the 6 values.
Aggregation of temperature, windspeed and weather scores:

In order to get a synthetic value, the final score is computed as a weighted average of the three scores.

The weights are:
- 50% for weather type
- 30% for temperature
- 20% for windspeed

These weights aim to represent the importance of weather factors for the final users.
Application:

The scores are computed every day. Different kind of plots exist:

Score and tendency for last 10 days

Score and tendency for last 3 months

<table>
<thead>
<tr>
<th>Région</th>
<th>Indices CDPH</th>
<th>Indices CDPME</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>J1</td>
<td>J3</td>
</tr>
<tr>
<td>France</td>
<td>0.80</td>
<td>0.74</td>
</tr>
<tr>
<td>Nord</td>
<td>0.81 ~</td>
<td>0.75 ~</td>
</tr>
<tr>
<td>Ile-de-France-Centre</td>
<td>0.77</td>
<td>0.72</td>
</tr>
<tr>
<td>Nord-Est</td>
<td>0.80</td>
<td>0.75</td>
</tr>
<tr>
<td>Centre-Est</td>
<td>0.79</td>
<td>0.73</td>
</tr>
<tr>
<td>Sud-Est</td>
<td>0.79</td>
<td>0.74</td>
</tr>
<tr>
<td>Sud-Ouest</td>
<td>0.81</td>
<td>0.76</td>
</tr>
<tr>
<td>Ouest</td>
<td>0.80</td>
<td>0.75</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Région</th>
<th>Indices CDPH</th>
<th>Indices CDPME</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>J1</td>
<td>J3</td>
</tr>
<tr>
<td>France</td>
<td>0.82 ~</td>
<td>0.76 ~</td>
</tr>
<tr>
<td>Nord</td>
<td>0.84 ~</td>
<td>0.78 ~</td>
</tr>
<tr>
<td>Ile-de-France-Centre</td>
<td>0.84 ~</td>
<td>0.79 ~</td>
</tr>
<tr>
<td>Nord-Est</td>
<td>0.81</td>
<td>0.75</td>
</tr>
<tr>
<td>Centre-Est</td>
<td>0.80 ~</td>
<td>0.74 ~</td>
</tr>
<tr>
<td>Sud-Est</td>
<td>0.79</td>
<td>0.74</td>
</tr>
<tr>
<td>Sud-Ouest</td>
<td>0.82 ~</td>
<td>0.77</td>
</tr>
<tr>
<td>Ouest</td>
<td>0.84 ~</td>
<td>0.79 ~</td>
</tr>
</tbody>
</table>
Application:
Chronology of the scores over 3 months

Indice de qualité global du 19/02/2016 au 18/05/2016 (moy. sur 3 jours)
Indice de qualité global, rès. 12 HUTC, réf. ANA_CLIM, échant. France

Global score

Score for temperature

Score for weather type

Z500 RMSE
3 models
Summary:

- The score gives satisfaction according to the objectives.

- The visualisation gives a real time information about final forecast quality.

- The temporal evolution of the score shows bad forecasts situations.

- Final validation is in progress.
1 - Verification of sensible weather:
   ▪ Motivations
   ▪ Calculation
   ▪ Application

2 - Modified CRPS calculation
   ▪ Context
   ▪ Calculation
   ▪ First results

Conclusion
Modified CRPS calculation

Context:

- TAC sub-group on extreme events verification
- Search for proper scores
- Have the possibility to give more attention to certain ranges of values

References:
- « Forecaster’s Dilemma: Extreme Events and Forecast Evaluation », Lerch et al., 2015
Usual CRPS calculation:

\[
\text{CRPS(\text{forecast})} = \frac{1}{\text{ncases}} \sum_{i=1}^{\text{ncases}} \int_{x=-\infty}^{x=-\infty} \left( F_i^f(x) - F_i^o(x) \right)^2 dx
\]
Modified CRPS calculation:

consider a given range of values → apply a weight function $w(x)$

$$\text{CRPS(forecast)} = \frac{1}{n\text{cases}} \sum_{i=1}^{n\text{cases}} \int_{x=-\infty}^{x=\infty} w(x) \left( F_i^f(x) - F_i^o(x) \right)^2 dx$$

Different weight functions are tested:

- Indicator left
- Indicator right
- Normal CDF left
- Normal CDF right
Data used:

- Parameters: 10m-windspeed, 6h-precipitations, 24h-precipitations
- PEARP ensemble system (based on ARPEGE), starting at 18h UTC.
- Time ranges from 0/6/24 to 108h
- Period: mai 2015 to april 2016
- Area: France
- Verification against synoptic observations
First results:
Comparison uniform/left/right weight for windspeed – threshold 5 m/s

→ Daily cycle is more pronounced with weighted CRPS (clear difference for valid time 12h)
→ no tendency with time range
First results:
Comparison uniform/left/right for 6h-precipitations – threshold 4 mm

→ Light deterioration with time range
→ Impact of threshold / frequency of the event
First results:
Comparison uniform/left/right for 24h-precipitations – threshold 10 mm

→ Light deterioration with time range
→ Impact of threshold / frequency of the event
Modified CRPS calculation

First results:
Comparison indicator and normal CDF weights for windspeed – threshold 15 m/s

→ indicator and normal CDF weights are very similar
First results:
Evolution with threshold

→ This behavior is well known for the Brier score, and is named « degeneracy » of the score: this is linked with the frequency of the assessed event.
So what?

- The weight method works, first results are encouraging and coherent with the Brier score behavior.

- It can be used to compare different ensembles for a given range of event.

- Problem of degeneracy will be further investigated:
  - Test the skill version of weighted CRPS.
  - Use weight functions based on the quantiles, not the absolute forecast values.
→ two current ways of investigation have been presented

→ first one tries to consider final user perception of the forecast error. It is nearly in operations.

→ second one aims to focus on extreme events. First results are encouraging.
Thank you for your attention!