User-oriented Assessment of Monthly Forecasts

Alice LEMANT, Arnaud MOUNIER, David PELOT

INTRODUCTION

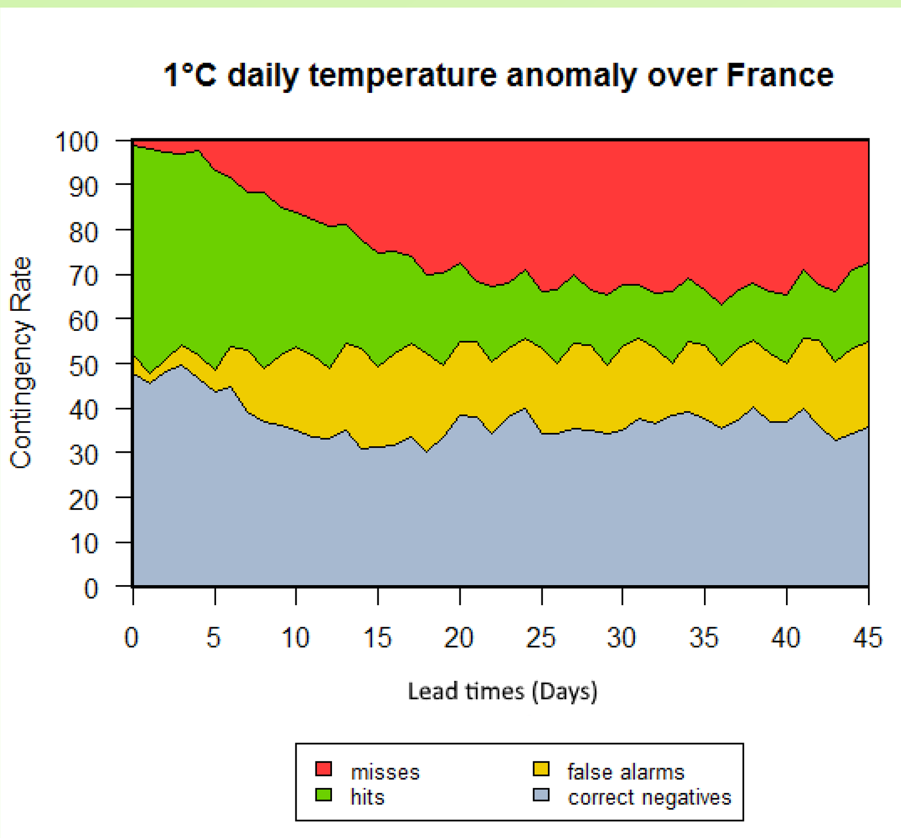
Monthly forecasts are provided to more and more users who require an estimation of its reliability. As the users are not familiar with advanced scoring methods, the goal of this study is to investigate the capacities of 2-meter temperature monthly forecasts with basic comprehensive scores.

DATA and METHODOLOGY

- post-processing : statistical adaptation on local French cities (multiple linear regression + pseudo perfect prediction method).
- period : from December 2015 to January 2018 (up to 46 days lead time), the sample size is 220 with monday and thursday runs of monthly forecast.
- temperature anomalies are calculated with comparison to 30 years climatology based on observations and ERA interim's analysis.
- the anomalies are calculated and averaged over different times and spatial scales (from town to country and day to month).
- the scores are based on contingency tables and its four components : misses (M), hits (H), false alarms (FA) and correct negatives (CN), respectively red, green, yellow and grey on the four colours plots below.

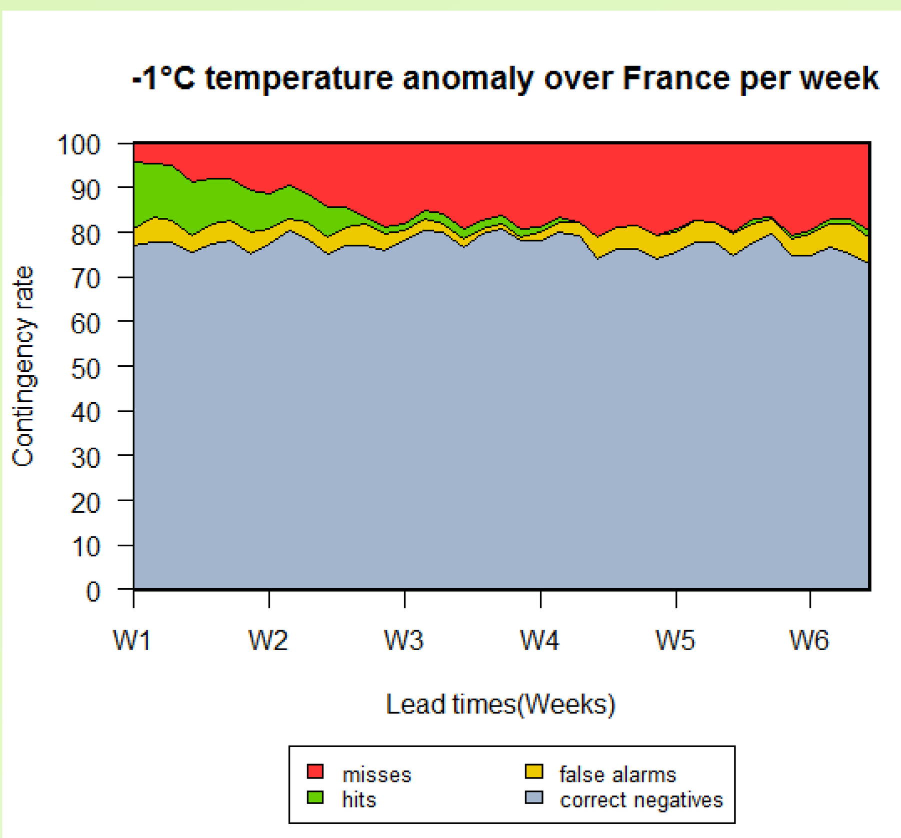
TEST OF DIFFERENT TIME AVERAGING : Day, Week, Month

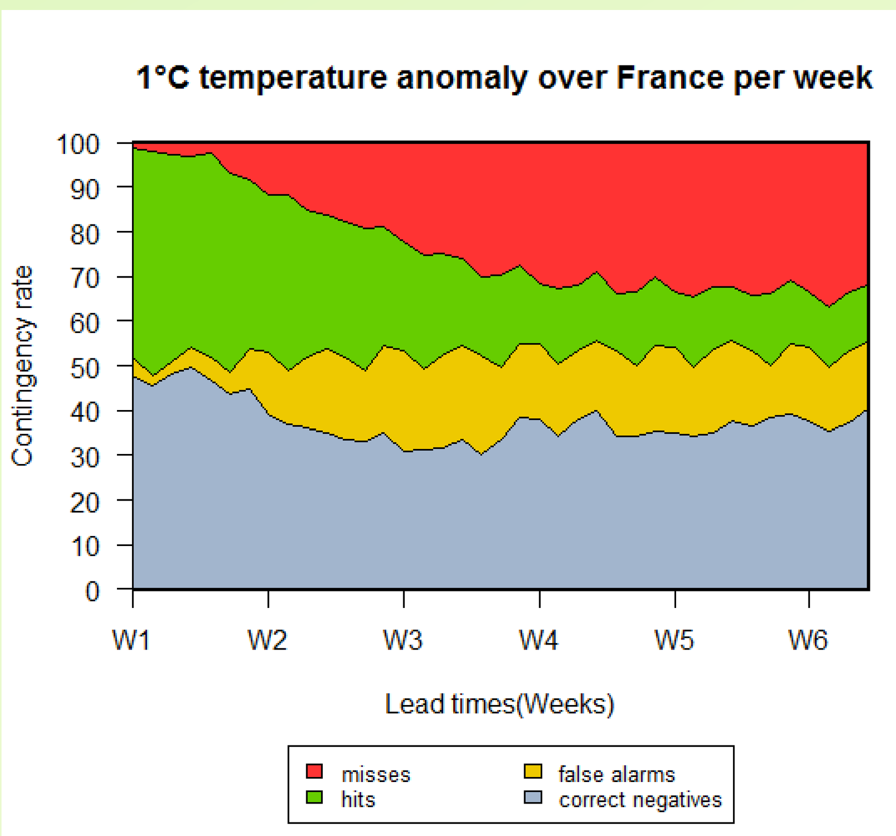
In this part, we compare forecasted and observed anomalies. We used the same threshold for the forecasted and observed anomalies.
The four plots below show the four components of the contingency tables, according to the lead time.



The graph shows a daily temperature anomaly over France >1°C. The model attests good results until day 15 : the number of hits remains superior to the number of misses until this lead time, with a reasonable number of false alarms.

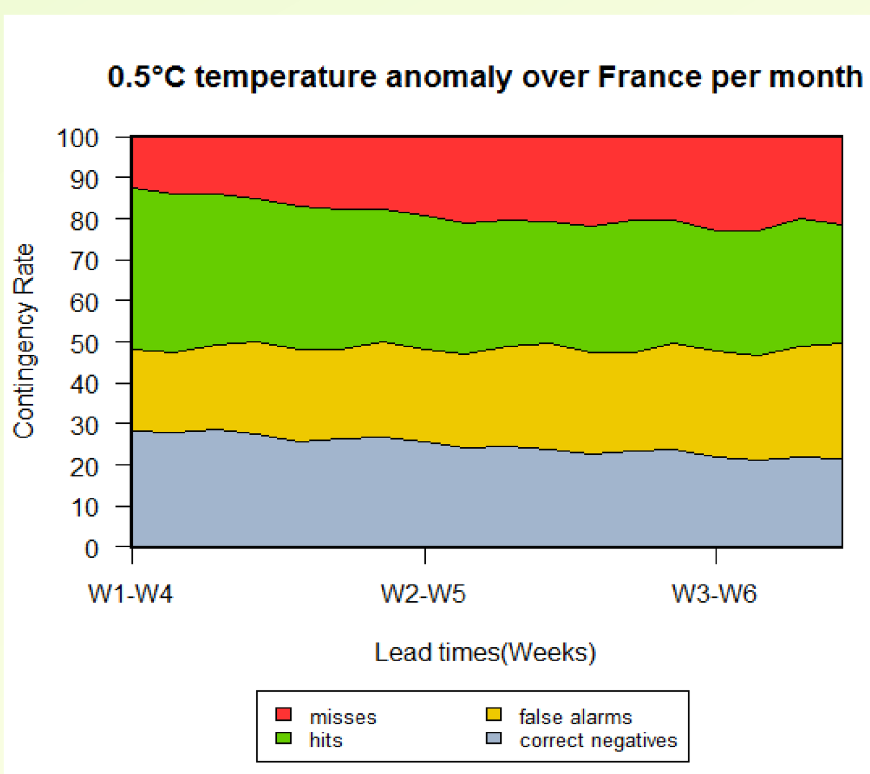
From day 25 to day 45, the curves are quite stationary with equal repartition between good forecasts (hit + correct negatives) and wrong forecasts (misses + false alarms).





On these 2 graphs, we focused on a 1°C temperature anomaly (positive and negative) averaged over a week. Note that, on x-axis, W1 means average from day 1 to day 7, W2 from day 8 to day 14 and so on.
The obvious assessment is that there is much more warm anomalies than cold ones. Furthermore the model detects warm anomalies better than cold ones, but with more false alarms. From week 4 the model is unable to detect anomalies inferior to -1°C.
If we compare one day to one week +1°C anomalies, there are more hits than misses up to week 3 with week averages, so this improves the skill of the forecast.

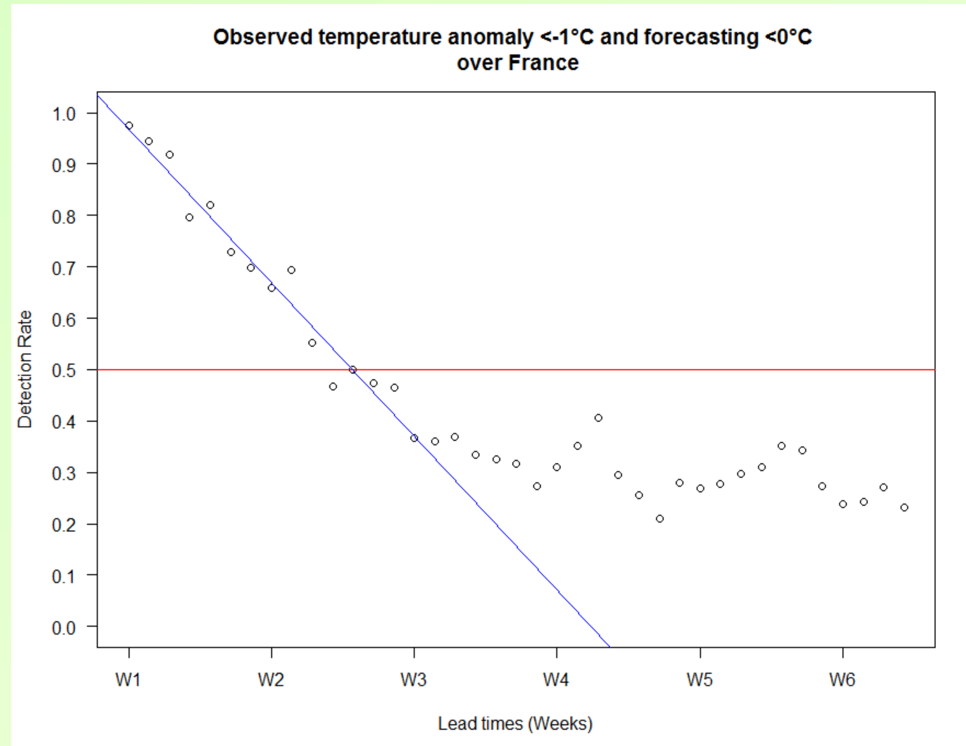
This plot shows anomalies averaged over one month. The threshold is +0.5°C in order to have enough cases.
There is a slight decrease of good forecasts (hits + correct negatives), but the proportion of good ones stays larger than bad ones up to week 3-6.
This result is interesting if we consider the use of monthly forecast as an indication of the first month of seasonal forecast.

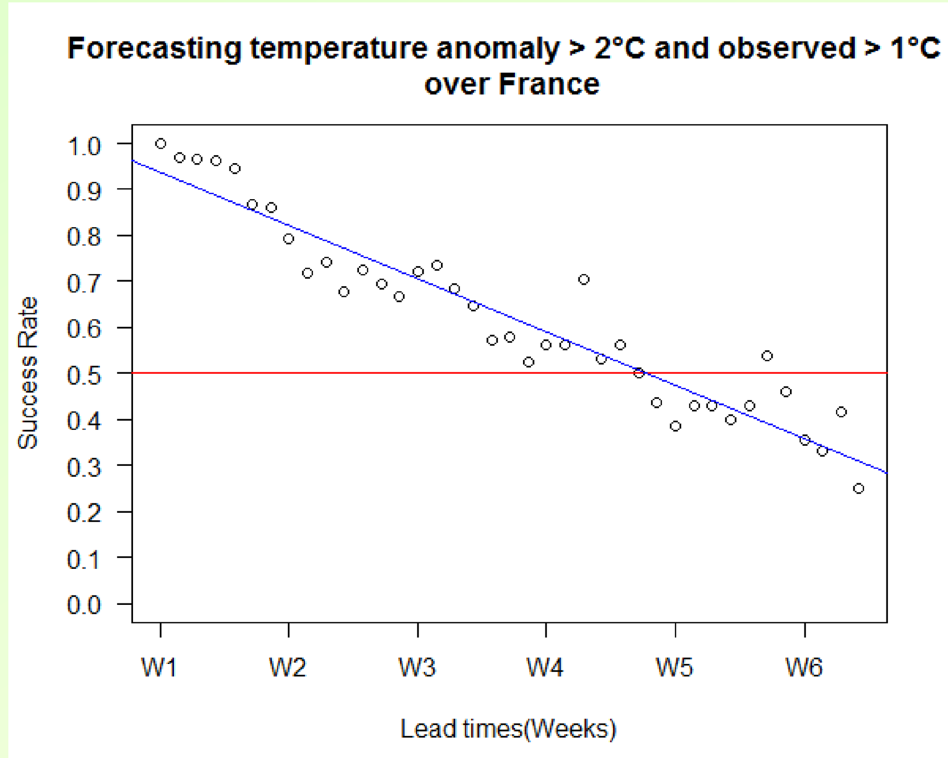


BETTER SCORES WITH DIFFERENTIATED ANOMALIES

Here we try to improve the detection of cold anomalies : we compare -1°C observed anomalies with 0°C forecasted anomalies.

Only the detection rate (hit/hit+misses) is plotted. The rate is larger than 50% up to week 2.5, and it stays above 20 % up to the last lead time (significantly non zero).





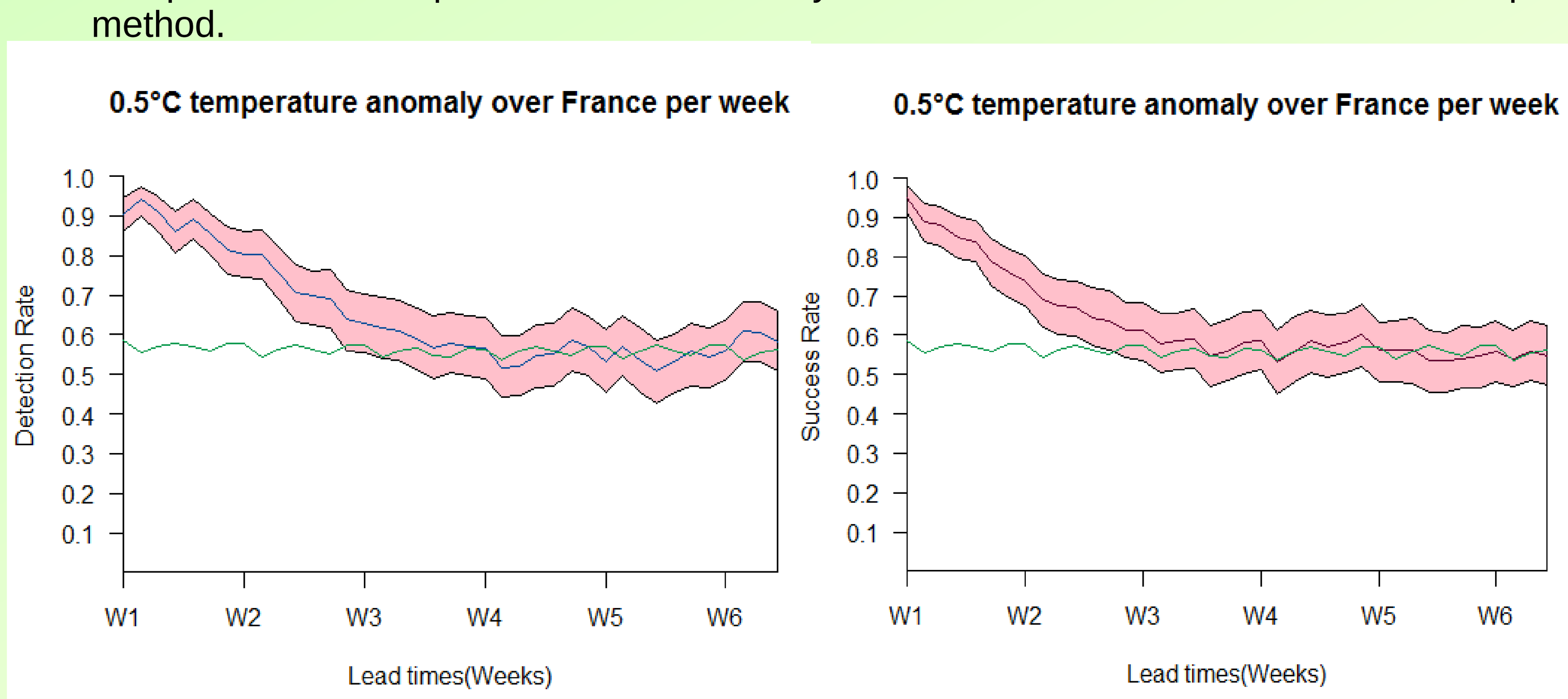
Here we compare +1°C observed anomalies with +2°C forecasted anomalies, in order to avoid false alarms.

The success rate (hit/hit+false alarms) is plotted. The rate is larger than 50% up to week 4.

Depending on user needs, this method can be used to decrease the number of false alarms.

COMPARISON TO CLIMATOLOGY

On the plots below, the green line shows the score for the climatology.
The pink area corresponds to the uncertainty of the score, calculated with the bootstrap method.



Here are the detection rate (left) and success rate (right) for +0.5°C anomalies over one week.
These plot show that the forecast outperforms the climatology up to week 3. This confirms the results obtained on the left part.

TEST OF DIFFERENT SPACE AVERAGING

Spatial averaging over one region or all the country have been tested.
Plots are not shown, but the impact of spatial averaging compared to time averaging is not significant.

CONCLUSION

- The model has more difficulties forecasting cold anomalies.
- Cold anomalies are more successfully anticipated with differentiated thresholds.
- The model is significantly better than climatology forecast until week three.
- Time averaging permits the use of monthly forecast for longer lead times.

FUTURE PLANS

- The main restriction of this study is the short period and particularly the lack of cold events. So the extension of the period will clearly improve the results reliability.
- The second restriction is the deterministic approach : assessment of ensemble benefit has to be explored, typically the link between ensemble spread and error, together with the use of probabilistic scores.