SCIENTIFIC COMMUNITY COMMENT: Philipp Körner

General comments:

The article introduces the tool PARAFOG v2.0, which is used to predict the occurrence of fog at certain locations. It distinguishes between radiation fog and deep, lowered stratus clouds. The authors claim to be able to predict fog with a hit rate of almost 100%, with a false alarm ration of 10 and 30% respectively.

The methodology is a further development of PARAFOG, with a significantly better hit rate. Overall, the methodology is worth publishing. However, in my opinion, there are some major weaknesses, which I would like to discuss in more detail below. I will limit myself to the things worthy of criticism. Language, stile, title, credit to other authors etc. are well done. My main criticism relates to performance analysis. After reading the abstract, the reader is curious how the authors managed to achieve such a good hit rate. However, when reading the article, disillusion sets in.

Specific Comments:

- The weights of the fuzzy logic algorithm were derived from events at the SIRTA station. However, the same fog events were also used for performance analysis. There is no differentiation between calibration and validation period.

Fuzzy logic scheme such as pre-fog alert algorithm is part of IA but is different from machine learning technics. Here, we define explicit parameters based on our observations and analysis (i.e. more than 100 STL fog events).

The best validation is however the encouraging results we obtained at Paris-Roissy, Zurich, Munich and Vienna airports by using the SIRTA' weights (Figure 9b).

- Only the results for "high" fog alert are shown. What happens after a low or medium fog alert, is not mentioned. Does every low fog alert lead to a medium and this to a high fog alert? Are there very often false alarms for low or medium fog alerts? We do not know. Nevertheless, it would be interesting for the overall assessment.

We thank you for the suggestion. HIT/MISS definition are the parameters used to evaluate performance. Alerts are tools to provide an intuitive output for the user. Since only high alerts are defined based on optimization criteria, it is logical to only evaluate their performance. Medium/low alerts are not optimized, but they are included because they can indicate the user that conditions are not 100% risk free with respect to clear weather, in a qualitative way.

- In the abstract and also in the article itself, it is written about minute resolution. However, in the corresponding place it says that the minute resolution is used to make aggregations to 45 minutes. A fog forecast is therefore often only possible in the aftermath (reanalysis), as the authors also write themselves. The operational mode, is mentioned, but results are not shown. A prediction in a reanalysis, when both the event itself and events from the future have entered the algorithm, has only little value as a quality analysis.

We think there is a misunderstanding between alert and alarm concepts. Please refer to section 5a. Alerts are used for real time, while alarms are only used for performance assessment (Figure 8). There is no conflict between near real-time analysis and reanalysis.

What we evaluated is the quality of the parameters used to determine a HIGH alert using the reanalysis. It is true that this is not exactly the same as evaluating the predictive potential, but optimizing the variables based on this result should optimize the predictions. With this regard, Figure 9 shows the statistics of what a user should expect when observing HIGH alerts.

- The exact decision-making process for assigning an alert level to the 45-minute windows is described in the paper. All assessment steps described in 5.a) sound plausible for themselves. However, especially in 5.a)-3, it is not described how exactly the numbers are derived. Why do you choose exactly 45 minutes, exactly 10 alerts and form the gradient over exactly 15 minutes? It can be assumed that the values were determined on the basis of the fog events themselves. Is that the case? If so, on which ones? One can assume, that this further weakens the statement of the performance analysis.

The assessment methodology that we propose is entirely novel and specifically designed to the PFG2 tool. Like any method, it is however not perfect. We agree that the choice of some values used for assigning an alert level to the 45-minutes alarm period may be arbitrary. Here, 45 minutes provide a minimum time to react in case of eventual fog formation. Note that PFG2 performance farther than 45 minutes before fog formation decrease and this may be linked with fog nature and the method principles (use of observations only). Hence, we are satisfied to present a hit rate of about 100% in this last 45 minutes when the signal is clear. This also implies that observable in-situ formation signals almost always appear in the last 45 minutes, but may or may not appear before (at least with this instrumental setup). Regarding the use of 10 alerts or 15 minutes gradients, it is because these parameters are optimized to provide a good compromise between HIT/FA/MISS with the present evaluation scheme.

Despite all this, it constitutes today the first complete evaluation of PFG1 and PFG2 which represent an important achievement.

- line 382++ The removal of certain events from the performance analysis manipulates the same. If the model cannot handle the situations described, then this is a weakness of the model that can either be named as such or be improved. Filtering out these events afterwards is not a solution.

In reality we do not remove certain events. At this stage it is not possible to decide if fog will re-form or dissipate, based on the instrumental setup. To tackle this issue during this period, we may set alerts to NaN in the future version of PFG.

In summary: The method and also the performance analysis have numerous degrees of freedom. The number of fog events is very small. This potentially leads to an overfitting of the method. In addition, an unknown number of events were filtered out of the performance analysis. Don't get me wrong: the tool is probably good or even very good. It's just that the validation method used is not suitable to prove this quality. The paper may be published after major revision.

We thank you for the thorough comments and changes suggested in your review of our manuscript. We agree that the assessment methodology is specifically designed to the PFG2

tool. With this regard, statements about PFG2 performance have been rewritten in the revised manuscript to consider your overall remarks.

Please the revised text now reads in section 5.b:

"*Note that this evaluation methodology has certain limitations. Arbitrary choices to consider only a 45 minutes alarm sub-period, or to have a minimum number of 10 alerts to trigger an alarm, may affect overall final performance. These parameters are optimized to provide a good compromise between hit/false alarm/miss with the present evaluation scheme, however, a sensitivity study may optimize the results. In addition, this method only evaluates the performance of PFG2 when fog events occur. Outside of these evaluation periods (3h for RAD and 24h for STL), PFG2 may deliver high alerts/alarms in pre-fog conditions such as during a stratus lowering having a cloud base height "stuck" a few tens of meters above the ground without leading necessarily to a subsequent fog event. This does not affect the PFG2 hit rate (number of hits or misses), but tend to underestimate the number of false alarms presented in this study.*"

Minor remarks:

- L143: how are 35 fog events a year fog- prone? Is that compared to other regions in France, or is there a general definition? In my ears, 35 fog events a year does not sound much.

Corrected as suggested.

- L366: "All other alerts occurring outside this period are not considered." Can you give an example? Erasing/not considering alerts in the aftermath is not suitable for a forecasting tool.

Please refer to the overall comment.