August 4, 2022

Machine-learning based prediction of Alpine Foehn events using GNSS troposphere products: First results for Altdorf, Switzerland

Aichinger-Rosenberger et al. *amt-2022-33rev1*

A. General Comments

The incorporation of reviewers' comments into the revised manuscript has improved it considerably. The title change already signals the intention of the paper to prove a concept – how to use previously untapped measurements for the diagnosis and nowcasting of foehn.

A few sticky issues remain or have surfaced in the course of the major rewrite of the article. They are listed in the next section.

Some serious mistakes in the computations had been found in the original manuscript. They seem to not have been completely eliminated and top of it, some of the figures that would have helped in cross-checking the results have been dropped in the revision. The first item of the "Specific Comments" gives details.

B. Specific Comments

B.1 **Correctness of some computations in doubt**: The review of the original manuscript version had unearthed large numbers of missed events in the observations of the foehn index FI (even though they are not missing at the MeteoSwiss database as I could confirm; original Fig. 5) and the events diagnosed from GNSS measurements (original Figs. 5–7 and Figs. 9-10). How the authors "have improved the corresponding code" remains vague in the response. I would have expected to see at least one figure showing observed and diagnosed events in the revised version but these figures have been completely dropped, which eliminates a chance to at least visually inspect the appropriateness of the results. That raises nagging doubts .... What is left for a cross-check are the performance metrics. If coding mistakes had been made only in the way missing GNSS data are handled (as the response indicates), differences should be seen in the performance metrics of the test data *and* the cross-validated training data. However, the numbers for the cross-validated training data metrics in the original and the revised version (Table 2 in both) are identical whereas the metrics for the test data have changed. As far as I see, some computations still have to be wrong. Consequently, the manuscript should not be published unless the correctness of the computations can be convincingly shown.

B.2 **Selection and application of machine-learning methods**: It is important to point out in the paper that no universal truth exists for the diagnosis of foehn and even diagnoses by human experts vary considerably (Mayr et al., 2018). Therefore the classification problem belongs to the category of unsupervised learning (e.g. Hastie et al., 2009). The paper, on the other hand, uses supervised learning methods, a choice that needs justification. A consequence of the use of supervised methods is raised in the following issue.

B.3 **Dependence on foehn classified with traditional meteorological measurements**: Despite the claim of the paper (line 445 and lines 521–522) the foehn classification with GNSS data is completely dependent on meteorological data, since these were used to compute the

foehn index FI that the supervised learning methods in the paper use as truth (response variable). It will therefore not be possible to compute an independent foehn climatology at Altdorf (lines 521-522) and the quality of the classification will by design always be poorer than the one of the foehn index FI.

B.4 **Streamlining**: The manuscript can be shortened by eliminating redundancies and combining parts.

a) Section 4.1 merely repeats methods already mentioned in the introduction without adding any further information. Section 4 could start with the data and then move on to methods used. Section 5 can be combined with 4 as it also addresses the methodology.

b) "Case studies" in the results section is a misnomer since they do not refer to a select event. "Feature sets" would be more appropriate. These sets can then be introduced in the data section by combining current subsections 4.2, 4.3 and 4.4. The results can then also be presented in one single confusion matrix, making it much easier for the reader to spot the difference in performance (and simultaneously shortening the paper). However, a fourth feature set needs to be introduced to properly fulfill the purpose of feature set 3 ("case study" in the current version) – see next issue.

B.5 **Feature set 3 inadequately specified**: The stated purpose of having the feature set ("case study") 3 is to evaluate whether adding further measurement stations can compensate for having a shorter training data set. The test period with data that have not been seen by the models in the training phase must be the same for a proper comparison. Especially for a rare event such as foehn with a large interannual variability selecting different and relatively short test periods can lead to considerably different results. Put succinctly: the test period for the feature set with a shorter training period must also be set to 2019–2020 as for the other feature sets. To disentangle the effects of a shorter training period and more stations, respectively, a fourth feature set needs to be introduced that has the shorter training period 2015–2018 and no additional stations.

B.6 **Large fraction of missing data**: The discussion on the limitations of the method (around line 485) correctly mentions a significant amount of periods without data. It would be good to be quantitative – also already when describing the data in section 4 – since this amounts to approximately 1/3 and 1/2 (!) of the time (cf. Tables 5 and 7). The current set up would therefore be unsuitable for any operational use. However, since most of the machine-learning algorithms used rely on aggregating weak learners, setting the methods up in a way that alternate features are used if a particular feature is not available is possible. This is especially easy to achieve for random forests.

B.7 **Section 4.6 Performance metrics**: Comment B.2 was fulfilled to a large part. What is still missing is the statement in the beginning paragraph that all following performance measures are derived from the confusion matrix. It is incorrect to state (line 250) that the performance measures were introduced by Barnes et al. (2007). They have been around many decades before.

B.8 **Percent vs percentage points**: "Percent" is sometimes incorrectly used instead of "percentage points", e.g. line 324. The difference between the observed frequency of foehn in the foehn index – 4.7 % – does not lie within "one percent" (as stated) of the results from the two algorithms. The difference is actually 21 % and 15 %, respectively, which is substantial. What the authors meant is "within one percentage point". However, what is of interest in judging the performance is the *relative* difference of the rare event "foehn", i.e. (correct) percentages

B.9 **Misleading response**: The response to the two points raised in the section "Technical corrections" of my first review simply reads "These have been included in the revised version." This is simply not the case and a lie. Figs. 5, 9, and 10 mentioned in comment C.2 do not appear in any form in the revision any more. Topography (comment C.1) was added but the lines connecting the stations contributing to the top features were omitted.

C. Less crucial and technical issues

C.1 A less crucial item to be changed is: A confusion matrix containing 4 cells does not need a figure. A simple table will do. And if all three (or four, see comments above) feature sets from Figs. 3, 5 and 7 are combined into one table, they can be directly and easily compared.

C.2 And one technical correction: Fig. 2 needs a colorbar.

References

Barnes, Lindsey R. et al. (2007). "False Alarms and Close Calls: a Conceptual Model of Warning Accuracy". In: *Weather and Forecasting* 22.5, pp. 1140–1147. DOI: 10.1175/waf1031.1.

Hastie, T, R Tibshirani, and J Friedman (2009). *The Elements of Statistical Learning*. Second Ed. Springer, New York, p. 524. DOI: 10.1007/978-0-387-84858-7.

Mayr, Georg J. et al. (2018). "The Community Foehn Classification Experiment". In: *Bulletin of the American Meteorological Society* 99.11, pp. 2229–2235. DOI: 10.1175/bams-d-17-0200.1.