February 17, 2022

Prediction of Alpine Foehn from time series of GNSS troposphere products using machine learning

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

A. General Comments

With foehn winds, air descends behind a topographic obstacle and often lowers moisture content there. The paper attempts to exploit this characteristic to diagnose and nowcast the occurrence of foehn (the response variable) with GPS-satellite derived integral moisture content variables as covariates. The response is an independently derived foehn indicator at a single station in Switzerland and the covariates are integral moisture measurements, their spatial gradients and horizontal differences among a few dozen stations in Switzerland. This is a unique, not yet exploited data set for the diagnosis and nowcasting of foehn. It could therefore be used to gain a better understanding of the mutual effects of foehn and moisture fields and to better diagnose or nowcast the occurrence of foehn but the paper falls short in both aspects as described in section B.

The results of the manuscript in its current form cannot be reproduced. Not enough details are given of the specific settings for the various machine learning algorithms and the data sets are also not available. Worryingly, a substantial amount of observed foehn events in Figs. 5, 6, 7, 9, 10 are absent, which casts doubt also on proper data handling for the rest of the paper.

B. Specific Comments

B.1 Improper data handling: The almost complete absence of foehn events in the last quarter of 2020 (which is part of verification period in the paper) from Figs. 5, 6, 7, 9, 10 seems suspicious since fall is a main foehn season. Indeed, a retrieval of the foehn index data for Altdorf for the period Sept-Dec 2020 shows 599 10-minute intervals with foehn instead of what seems to be only a single data point in the paper. Such a mistake casts serious doubt on proper data handling in the rest of the paper; especially since no summary statistics of the data are given, e.g. percentage of missing values in both response and covariates, range of values for the covariates. The same data check for Sept-Dec 2020 revealed that 240 data points of the foehn index are missing. Are those missing response dates properly excluded from the computation of the scores?

B.2 Not reproducible: Since neither code (only upon request) nor data are available and almost no specifics about the settings of the machine learning algorithms nor the version of the software package are given the results cannot be reproduced; even a plausibility check for appropriateness of the algorithm settings is not possible. And: how many covariates are actually used (add to table 1)?

B.3 Choice of machine learning algorithms: Why are exactly the algorithms listed in subsection 4.1 chosen among many possible candidates and why are so many used (see next issue)? Since random forests as ensembles of decision trees outperform them: why are decision trees included? Support vector classifiers assume a linear boundary between two classes (foehn/no foehn in this case) whereas support vector *machines* can handle non-linear boundaries. Why are SV classifiers chosen instead of SV machines?

B.4 Lack of performance optimization: If best possible performance of foehn diagnosis/nowcasting with GPS data is a goal then the setting of all algorithms should be tuned first instead of using default settings in the particular software package to select two and only tune these two. Other methods - if properly tuned - might work better.

B.5 Lack of physical understanding: The application of integral water vapor information from GPS satellites to the diagnosis of foehn is unique. Therefore an attempt is needed to understand details of the integral water vapor fields and their relation to foehn. Since most information lies in the ZWD field (cf. Fig. 8) figures with its average spatial distribution during foehn events and non-foehn events (of similar sample size as foehn events) will be helpful – similar to Sprenger et al. (2017) for pressure. Such maps should ideally be stratified by season. Since water vapor content is highly variable, an exploration behind the reasons of success and failure of the model diagnosis should be undertaken. Deep foehn situations, for example, might have a strong humidity gradient across the Alps, whereas shallow foehn cases or the onset of foehn events might have weak gradients. Since the GNSS stations are not collocated with the foehn station, consequences for the model performance should be explored, e.g. with maps of ZWD for foehn situations. A different avenue to pursue for increasing understanding is using an individual tree from a random forest model to illustrate how that model separates foehn cases from no-foehn cases.

B.6 Ultimate reason for the method: Why should foehn be diagnosed from GNSS-derived information? Weather station data give a more reliable answer for specific locations and such information was actually used as truth to approximate with GNSS data and machine learning algorithms. The method described in the paper cannot be used to diagnose foehn in locations without weather stations either, since it was trained on only one station and the transferability to other locations is not shown in the paper. The paper uses the approach for nowcasting 1 hour into the future and mentions that NWP models fare poorly with foehn quoting a paper from 2012. NWP models and their spatial resolution have dramatically improved in the decade since then. I would guess that MeteoSwiss has a current performance evaluation of COSMO1 available for Altdorf, against which the results of the paper could be measured. Results should also be compared to a simple persistence model, i.e. nowcasting the same no/foehn state as in the current hour.

B.7 Larger data set: To become more confident about the usability of integral moisture data for foehn diagnosis, more foehn locations should be included. Several more locations exist in Switzerland, for which a foehn index is available. To get more robust error estimates and performance scores, using the longer data set 1999-2020 mentioned in line 120 would be helpful. Line 125 merely states that only 2010-2020 is used without giving a reason.

B.8 Verification: Comparing total number of foehn hours from foehn index and the algorithms - stratified by season - should give an overall impression of the performance. To get an impression of the performance, a week-long time series containing one or more foehn events should be shown that includes the foehn index and the values of the four dominant features (as given in Fig. 8); if possible together with meteorological data of wind speed and direction, relative humidity and temperature.

Less crucial items to be changed are:

B.1 Give a short summary of how hydrometeors affect ZWD and ZHD and what that means for the applicability of the data set to foehn diagnosis, since foehn can happen with and without precipitation-sized particles.

B.2 Performance metrics: Subsection 4.4. can be shortened drastically by giving a confusion matrix and listing the scores derived from it in a table. After all, these are well-known scores in literature. Numbers for the confusion matrix should be given for both the test and training period.

B.3 Performance might be improved further by having GNSS information further south. Are there no such stations in Italy?

B.4 Focus the machine learning aspects in the introduction only on classification, the task at hand in the paper. You might add a further machine learning method to foehn diagnosis, namely mixture models (Plavcan et al., 2014).

B.5 Why are 12-hour moving average values of ZWD used in Fig. 2? Is the averaging window centered or asymmetric? Are the covariates used in the algorithms also 12-hour moving averages or "hourly troposphere products" as line 265 states? Does "hourly" mean an average over the hour or an instantaneous value every hour?

C. Technical Corrections

C.1 Fig. 1: Add topography and draw the lines between the stations, which contribute the top 4 features shown in Fig. 8.

C.2 Combine figures and tables to ease comparison: table 4 with table 5; Fig. 9 with Fig. 10 and rearrange Fig. 5 to have "observed" in center and the results from both methods immediately above and below (also bring lines closer together - a small detail that helps comparison).

References

Plavcan, David, Georg J. Mayr, and Achim Zeileis (2014). "Automatic and Probabilistic Foehn Diagnosis with a Statistical Mixture Model". In: *Journal of Applied Meteorology and Climatology* 53, pp. 652–659. DOI: `10.1175/jamc-d-13-0267.1`.

Sprenger, Michael et al. (2017). "Nowcasting Foehn Wind Events Using the AdaBoost Machine Learning Algorithm". In: *Weather and Forecasting* 32.3, pp. 1079–1099. DOI: `10.1175/waf-d-16-0208.1`.