The DOI for this manuscript is doi: 10.1175/BAMS-D-17-0200.1

The final published version of this manuscript will replace the preliminary version at the above DOI once it is available.

If you would like to cite this EOR in a separate work, please use the following full

Mayr, G., D. Plavcan, L. Armi, A. Elvidge, B. Grisogono, K. Horvath, P. Jackson, A. Neururer, P. Seibert, J. Steenburgh, I. Stiperski, A. Sturman, Ž. Vecenaj, J. Vergeiner, S. Vosper, and G. Zängl, 2018: The Community Foehn Classification Experiment. Bull. Amer. Meteor. Soc. doi:10.1175/BAMS-D-17-0200.1, in press.

The Community Foehn Classification Experiment

Georg J. Mayr*

University of Innsbruck, Innsbruck, Austria

David Plavcan†

University of Innsbruck, Innsbruck, Austria

Laurence Armi

Scripps Institution of Oceanography, University of California San Diego, La Jolla, California

Andrew Elvidge

University of East Anglia, Norwich, UK

Branko Grisogono

University Zagreb, Zagreb, Croatia

Kristian Horvath

Meteorological and Hydrological Service, Zagreb, Croatia

Peter Jackson

University of Northern British Columbia, Prince George, Canada

Alfred Neururer

Central Institution for Meteorology and Geodynamics, Innsbruck, Austria

Petra Seibert

University of Natural Resources and Life Sciences, Vienna, Austria

James W Steenburgh

University of Utah, Salt Lake City, USA

Ivana Stiperski

University of Innsbruck, Innsbruck, Austria

Andrew Sturman

University of Canterbury, Christchurch, New Zealand

Željko Večenaj

University of Zagreb, Zagreb, Croatia

Johannes Vergeiner

Central Institution for Meteorology and Geodynamics, Innsbruck, Austria

Simon Vosper

Met Office, Exter, UK

Günther Zängl

Deutscher Wetterdienst, Offenbach, Germany

**Corresponding author address:* Georg J. Mayr, Department of Atmospheric and Cryospheric Sciences, University of Innsbruck, Innrain 52, 6020 Innsbruck, Austria.

E-mail: georg.mayr@uibk.ac.at

†Current affiliation: UBIMET, Vienna, Austria

ABSTRACT

Strong winds crossing elevated terrain and descending to its lee occur over mountainous areas worldwide. Winds fulfilling these two criteria are called “foehn” in this paper although different names exist depending on region, sign of temperature change at onset, and depth of overflowing layer. They affect local weather and climate and impact society. Classification is difficult because other wind systems might be superimposed on them or share some characteristics. Additionally, no unanimously agreed-upon name, definition nor indications for such winds exist. The most trusted classifications have been performed by human experts. A classification experiment for different foehn locations in the Alps and different classifier groups addressed hitherto unanswered questions about the uncertainty of these classifications, their reproducibility and dependence on the level of expertise. One group consisted of mountain meteorology experts, the other two of Masters degree students who had taken mountain meteorology courses, and a further two of objective algorithms. Sixty periods of 48 hours were classified for foehn/no foehn at five Alpine foehn locations. The intra-human-classifier detection varies by about 10 percentage points (interquartile range). Experts and students are nearly indistinguishable. The algorithms are in the range of human classifications. One difficult case appeared twice in order to examine reproducibility of classified foehn duration, which turned out to be 50% or less. The classification dataset can now serve as a testbed for automatic classification algorithms, which - if successful - eliminate the drawbacks of manual classifications: lack of scalability and reproducibility.

1. Introduction

Many processes and phenomena in the atmosphere need to be diagnosed – from low pressure systems with fronts in midlatitudes and hurricanes in the tropics, to fog or lightning. Some diagnoses are easy to make. Hearing thunder identifies lightning, and not being able to see a building less than 1 km away during daytime indicates fog. These diagnoses can even be automated with suitable instrumentation – to identify lightning from its signature in the electromagnetic waves it emits and fog from scattering of a light source. Some processes and phenomena, however, are much harder to classify, often because not enough information is available or the process itself is insufficiently understood. Lately, methods from statistics and machine learning in combination with a huge increase in computing power have been harnessed with ever increasing success to tackle more and more difficult classification tasks, earning them the label "artificial intelligence". Arguably the largest progress has been made in classifying images, from spotting a dog on a photo to identifying a particular person. The underlying neural-network algorithms, however, typically need thousands or even hundreds of thousands of pre-classified images provided by humans in order to "learn". Such "supervised" learning is much easier than "unsupervised learning" for which no "truth" exists. This is the area where classifications by human experts are still the gold standard, albeit with several drawbacks: lack of scalability and reproducibility, and unknown error rates. Because only few people have the required expertise to perform a classification, which takes a substantial amount of time, the classification task cannot be extended to an arbitrarily large number of instances, and comparison of classifications among different experts or by the same expert performed at different times are at best extremely rare.

A group of experts collaborated recently on such a task to remedy two of the classification drawbacks by providing estimates of classification uncertainty and reproducibility, and a database against which existing and future algorithms can be tested. The classification task identified periods of downslope windstorms in time series of weather station measurements.

Such windstorms result from winds that cross topographic obstacles and accelerate as they descend to their lee. They occur over mountainous locations worldwide and are known by different names, which are sometimes also used to refer to an additional characteristic. Because no all-encompassing name exists this article will use “foehn” for simplicity *without* implying a temperature increase during its onset, or a specific region. Foehn affects local weather and climate and impacts agriculture (growing conditions due to temperature and humidity changes; top soil erosion), tourism (reliable spots for wind and kite surfing), artificial snow making (change of wet-bulb temperatures), air pollution (trapping pollutants in cold pools underneath the foehn layer, or sweeping them away in case of break-through), human health (reduction of air pollution), forest fires (intensifying them to uncontrollable extents), ground traffic (toppling trucks; snow or sand drifts; blasting of vehicles with sand and small rocks), and air traffic (closure of runways when crosswinds are too high). The increasing density of automatic weather stations allows the observation of such winds at progressively more locations. Classification, however, is difficult because other wind systems such as radiatively-driven downslope/downvalley winds might be superimposed on foehn, share some of its characteristics, or because not enough information is available. The difficulty is compounded because no unanimously agreed-upon definition of foehn and its indications exist, foehn occurs in a variety of synoptic-scale and mesoscale settings, and different names are being used depending on region, the sign of temperature change at its onset, and its depth.

2. Classification task

Nevertheless, two unanimously agreed-upon characteristics are that air crosses an obstacle and that it descends and accelerates on the downwind side causing strong winds. A fairly simple conceptual model of the flow situation after the onset of foehn, corroborated by field campaigns, laboratory experiments, computer simulations, and theoretical investigations, is shown in Fig. 1. Unfortunately, no continuous measurements covering the vertical cross-section are routinely available for classification; only weather stations at the ground. Nowadays with the proliferation of automatic weather stations and mesonets in some regions, measurement(s) close to the crest of the obstacle are also available so that the first foehn characteristic of air *crossing* the topographic obstacle can be checked. The second characteristic that air *descends* leads to adiabatic warming and consequentially to a decrease in relative humidity. It can be examined through differences between crest and downwind station of variables which are approximately conserved in foehn flow, such as potential temperature or mixing ratio.

Classification is made more ambiguous by processes for which potential temperature and mixing ratio are not conserved, i.e. turbulent mixing within the foehn flow, at the surface and its upper interface; mixing air in from tributaries; phase changes of water (formation and evaporation of liquid and solid particles); daytime warming and nighttime cooling due to surface sensible heat flux. How large these diabatic effects are varies with season, time of day, location, and large-scale and mesoscale flow configurations. Information about their contribution is not readily available so that classifications become difficult and possess an unknown and variable degree of uncertainty.

3. The community foehn classification experiment

The community foehn classification experiment set out to quantify the uncertainty of human foehn classifications, to compare them to machine classifications and to provide a data set for the development of foehn classification algorithms. Three groups of human experts and two objective algorithms faced the task of identifying foehn periods. The first group (most of them are co-authors of this paper) consisted of 26 seasoned experts in mountain meteorology from different continents with operational or research backgrounds and thus a broad range of concepts of what constitutes foehn. The other two groups are students taking the advanced weather forecasting course at the University of Innsbruck in 2016 (34) and 2017 (18), respectively. The student groups had a fairly homogeneous level of expertise because they had received four hours of lectures on foehn and had to apply it in homework problems in their advanced weather forecasting course. It was explained to the students why it was crucial for the outcome of this study that they worked completely independently. In addition to human experts, two algorithms were used that also employ the concept shown in Fig. 1. One, labeled A1 henceforth, is in operational use by the Swiss weather service. It uses percentiles of the distribution of the difference of potential temperature between crest and downstream locations (small, cf. Fig. 1), wind speed (high) and relative humidity (low) as hard thresholds for the classification of three categories: no foehn, foehn air mixed with cold valley air, and foehn. The second algorithm, A2, in operational use at the University of Innsbruck, learns from the data by itself and does not use hard thresholds. It uses so-called statistical mixture models to fit two or more parametric distributions to the observed distribution of classifying variables, such as potential temperature difference between crest and downwind stations, and wind speed, to yield a probability for foehn between 0 and 1, instead of merely a binary yes/no classification. Both algorithms require that the appropriate directional sector for foehn winds be manually set.

The classification experiment was designed to strike a balance between ideal goals and practical feasibility for the human classifiers. Therefore, five topographically different locations of differing annual foehn frequency in the Swiss Alps were selected (Table 1 and Fig. 2). Twelve 48-hour periods at each station yielded a total of 60 cases, for which the experts had to classify south foehn periods lasting at least 1 h at 30-minute resolution. One of the co-authors, who did not himself manually classify (D. Plavcan), selected these cases based on results from the two automated classification algorithms, A1 and A2, to cover all permutations: phases of foehn/no-foehn for which both, only one, or none agreed. Cases contained none, one or several foehn periods, respectively. Unbeknownst to the classifiers, one difficult 48-hour period appeared twice in order to estimate reproducibility.

Each participant received a wind-speed-coded wind rose for each location, a pseudo-3D image of the location from Google Earth, exact coordinates, plots of meteorological variables for each of the 60 periods of 48 hours, and instructions that contained an annotated example of an additional case reproduced here in Fig. 3. To classify only south foehn events, air had to cross the Alpine crest from south to north as indicated by wind direction at the crest plotted in black instead of gray, which is fulfilled for the whole 48-hour-period in this case. Three periods of foehn are inferred; from 9:00—10:20, 11:10—14:30 and 31:00—45:20 (as hh:min). During these periods, similar potential temperatures at crest and the classification location imply the second foehn characteristic of lee-slope descent. Wind directions are from the appropriate sector[1] and wind speeds are higher. Temperatures increase at the onset of each period, presumably when foehn erodes an underlying shallow cold pool. Humidity also drops, reflecting the draw-down of drier air from higher altitudes. Because *relative* humidity (%) instead of specific humidity (g/kg) is plotted, the temperature increase additionally contributes to a drop in relative humidity.

4. Results

The three human groups classified foehn duration during the 12 x 48-h periods at each of the five locations broadly similarly as Fig. 4 shows. Median durations (colored horizontal lines) are

[1] deduced from wind roses and topography maps; not shown

within a few percentage points of each other. The group of mountain meteorology experts have the most diverse backgrounds and consequently concepts of what constitutes foehn. As a result, their classification variation is larger than that of the second group of students who all had the same foehn concept instilled in their course. The variation of the first group of students, on the other hand, is larger; mainly because of a few outliers at each location.

The variation and thus classification uncertainty is smallest at location 4, a station at the northern edge of the Alps. The largest uncertainty occurred for location 1, where foehn can potentially blow from several wind sectors and for which the crest station might not always be representative of the upstream conditions.

The agreement between the algorithms and human classifications varies. A1 is within a few percentage points of the medians of the human groups at locations 2 and 3; A2 at locations 1 and 4. However, they are at the margins of human classifications for locations 2 (A2), 3 (A2) and 5 (A1 and A2), and A1 is even outside at locations 1 and 4.

a. Classification example

Fig. 5 shows the classifications from the three groups of human classifiers and the two algorithms for one of the 60 cases. At about midday of the *second* day potential temperature at the valley station 1 reached a value close to that of the crest station (purple line), indicating descent of air. Wind speeds also increased. In the evening the signals in the variables reverse, indicating the cessation of foehn conditions. Human classifications agree on a core period of foehn from 11:00-14:30 (labeled “easy” in the figure) but differ in onset and end times, with end times less unanimous than onset times. The two algorithms classify similarly.

The nighttime period between day’s 1 and 2, on the other hand, is more difficult. About 60% of the experts and students classified it as foehn (labeled “difficult”), again agreeing for the core period but differing for onset and even more so for end times. On the evening of the first day the wind direction changed into the foehn sector. At the same time both average and peak wind speeds increased and potential temperature also increased. Unlike the "easy" period, however, potential temperature is five Kelvin colder than at the crest, which likely led the other 40% to classify it as a radiatively cooled nocturnal downslope/downvalley flow. Air originating from a different level than represented by the crest station (cf. Fig. 2) and mixing of foehn air with radiatively cooled air from the valley and its tributaries might have been responsible for such a large difference. The three-category algorithm A1 classifies no foehn, whereas the mixture model algorithm A2 gives a probability close to 1 that it is foehn. The decrease and fluctuations of the probability towards the end of the period stems from the decrease and fluctuations in wind speed and later on the increase in potential temperature difference.

This "difficult" period indicates that a simple "yes" or "no" might not be enough for all applications when it comes to classifying foehn flows, for example because of the superposition of foehn and a radiatively cooled downvalley wind. Algorithm A1 adds the third category of "mixed foehn/valley air" (although it does not classify it as such in this particular case). Algorithm A2 gives a continuous probability of foehn occurrence.

b. Changes in classification uncertainty

Over all 60 cases, delineating the beginning and end of a foehn event had a higher variability among all classifiers. Although the majority of classified foehn events started with a temperature increase, uncertainty was not clearly different from events which started with no change or a decrease of temperature. Classification uncertainty was also higher for nighttime than daytime for similar reasons as in the "difficult" period of Fig. 5. Classification uncertainty also varied somewhat seasonally with low uncertainty in fall (SON) and winter (DJF) months, highest uncertainty in spring (MAM) particularly among human classifiers, and medium uncertainty in summer months (JJA).

c. Reproducibility

To evaluate reproducibility, one of the more difficult cases (at location 1) occurred twice in the data set, unbeknownst to the classifiers. Fig. 6 shows the relative frequency of the absolute difference of foehn duration classified at the first occurrence and the second occurence of that case. Ideally and for perfect reproducibility, the difference in classified foehn duration among the identical cases is zero. However, fewer than half of the classifiers achieved perfect reproducibility.

This lack of reproducibility is worrisome although probably less extreme for easier cases. Nevertheless, it corroborates the first author's personal experience of classifying foehn at different locations globally.

d. Dataset

The dataset will be available at UC Irvine, which hosts a large repository of classification data sets, at `https://archive.ics.uci.edu/ml/about.html`.

5. Conclusion

Several lessons have been learned from this experiment that add on the one hand supporting evidence to what was previously at least informally known from other classification tasks ((i)-(iii), and on the other hand ((iv) - (vi)) add new knowledge: (i) Busy experts are willing to volunteer a chunk of their scarce time provided the classification task is an intellectually challenging puzzle; (ii) Human experts use implicit (and in the case of the Masters students explicitly taught) physically-based concepts to help them distinguish between the two categories of foehn/no foehn;

(iii) Expert classifications carry uncertainty and are not even necessarily reproducible, which needs to be quantified (as here) or at least considered when interpreting results using such classifications; (iv) Uncertainty is largest for onset and even more so for the ending of a foehn event and also larger during the night; (v) Combining advanced statistical and/or machine learning models with physically-based concepts for choosing their input variables yields similar results to those of human experts. In addition, they easily scale to longer time series or more locations and are reproducible, which is a fundamental scientific requirement and allows the comparison of different data sets (foehn occurrence at different locations in this case). It is thus highly recommended to develop objective classification procedures, ideally without having to resort to manually specified and/or hard limits. If the algorithms are additionally made available as packages of open-source languages, foehn classifications can easily be reproduced by other researchers; (vi) Diagnoses contain more information when they are probabilistic instead of binary yes/no – a concept that has already been implemented for a long time in (weather) forecasts.

In addition to shedding light on human and machine classification of foehn, the dataset allows the testing of existing and newly developed algorithms for unsupervised learning tasks when truth is not known, such as in the case of foehn occurrence. It can also serve a community interested in estimating the accuracy of previous human foehn classifications and climatologies.

Acknowledgments. Many thanks to Achim Zeileis for discussions how to best design the experiment. A huge thanks to the experts and the 2016 and 2017 cohorts of the Advanced Weather Forecasting class who classified these 60 periods.

References

Brinkmann, W. A. R., 1971: What is a foehn? *Weather*, **26 (6)**, 230–240, doi:10.1002/j.1477-8696.1971.tb04200.x.

Dürr, B., 2008: Automatisiertes Verfahren zur Bestimmung von Föhn in Alpentälern. Tech. Rep. 223, Arbeitsberichte der MeteoSchweiz, 22 pp. URL http://www.meteoschweiz.admin.ch/content/dam/meteoswiss/de/Ungebundene-Seiten/Publikationen/Fachberichte/doc/ab223.pdf.

Elvidge, A. D., and I. A. Renfrew, 2016: The causes of foehn warming in the lee of mountains. *Bulletin of the American Meteorological Society*, **97 (3)**, 455–466, doi:10.1175/bams-d-14-00194.1.

Jackson, P. L., G. Mayr, and S. Vosper, 2013: Dynamically-driven winds. *Mountain Weather Research and Forecasting: Recent Progress and Current Challenges*, F. K. Chow, S. F. J. de Wekker, and B. J. Snyder, Eds., Springer, chap. 3, 121–218, doi:10.1007/978-94-007-4098-3_3.

Leisch, F., 2004: FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software*, **11 (8)**, 1–18, doi:10.18637/jss.v011.i08.

Plavcan, D., G. J. Mayr, and A. Zeileis, 2014: Automatic and probabilistic foehn diagnosis with a statistical mixture model. *J. Appl. Meteor. Climatol.*, **53**, 652–659, doi:10.1175/jamc-d-13-0267.1.

Richner, H., and P. Hächler, 2013: Understanding and forecasting Alpine foehn. *Mountain Weather Research and Forecasting: Recent Progress and Current Challenges*, F. K. Chow, S. F. J. de Wekker, and B. J. Snyder, Eds., Springer, 219–260, doi:10.1007/978-94-007-4098-3_4.

Seibert, P., 2012: The riddles of foehn - introduction to the historic articles by Hann and Ficker. *Meteorologische Zeitschrift*, **21 (6)**, 607–614, doi:10.1127/0941-2948/2012/0398.

von Ficker, H., 1910: On the formation of foehn winds on the northern side of the Alps (translated and edited by Volken, E. and S. Brönnimann). *Meteorologische Zeitschrift*, **21 (6)**, 597–605, doi:10.1127/0941-2948/2012/0532.

LIST OF TABLES

TABLE 1. Weather station locations used for foehn classification with their long-term foehn frequencies determined from automatic algorithms A1 and A2.

location	latitude N	longitude E	altitude [m MSL]	frequency from A1[%]	from A2 [%]	town
1	46.30287	7.84294	639	6	10	Visp
2	46.88702	8.62181	438	5	5	Altdorf
3	47.12745	9.51753	457	4	4	Vaduz
4	47.42546	9.39847	776	2	2	St. Gallen
5	47.03643	8.30097	457	<1	<1	Luzern
c (crest)	46.65346	8.61625	2287	-	-	Guetsch

LIST OF FIGURES

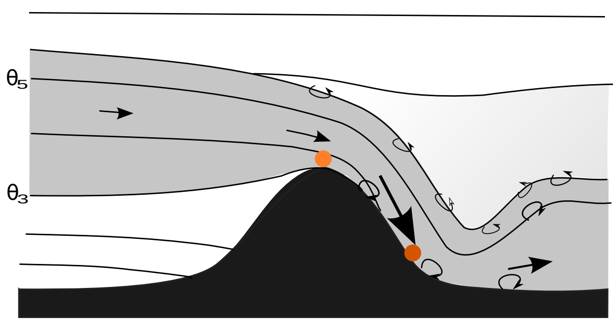

FIG. 1. Conceptual model of well-established foehn flow (dark grey shading) along a vertical cross-section exhibiting the two core characteristics of air crossing the obstacle (black), which can be ridge-like, a strait, or a pass, and descending to its lee. Flow approximately along isentropes (thin lines); straight arrows indicate wind speed, curved ones turbulence. Colored dots are exemplary weather station locations at crest and downwind (c.f. Table 1).

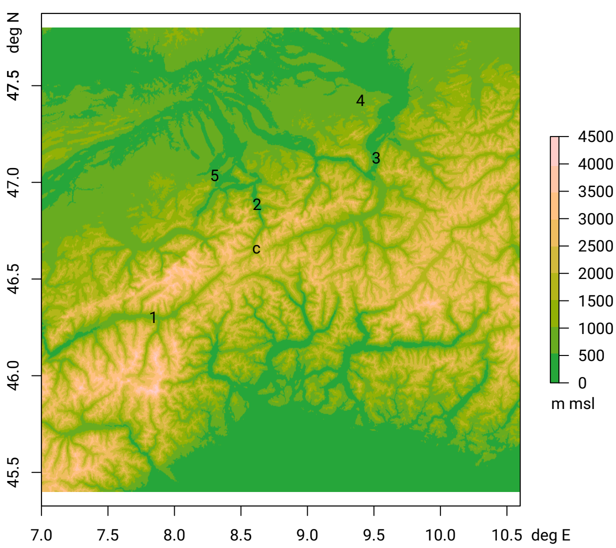

FIG. 2. Topography (m msl) and location of stations for which foehn was classified (cf. Table 1). Measurements at the crest location (C) were used to assist classification at all locations. Digital elevation model at 250 m horizontal resolution from SRTM (`http://gisweb.ciat.cgiar.org/TRMM/SRTM_Resampled_250m/`gisweb.ciat.cgiar.org/TRMM/SRTM_Resampled_250m/).

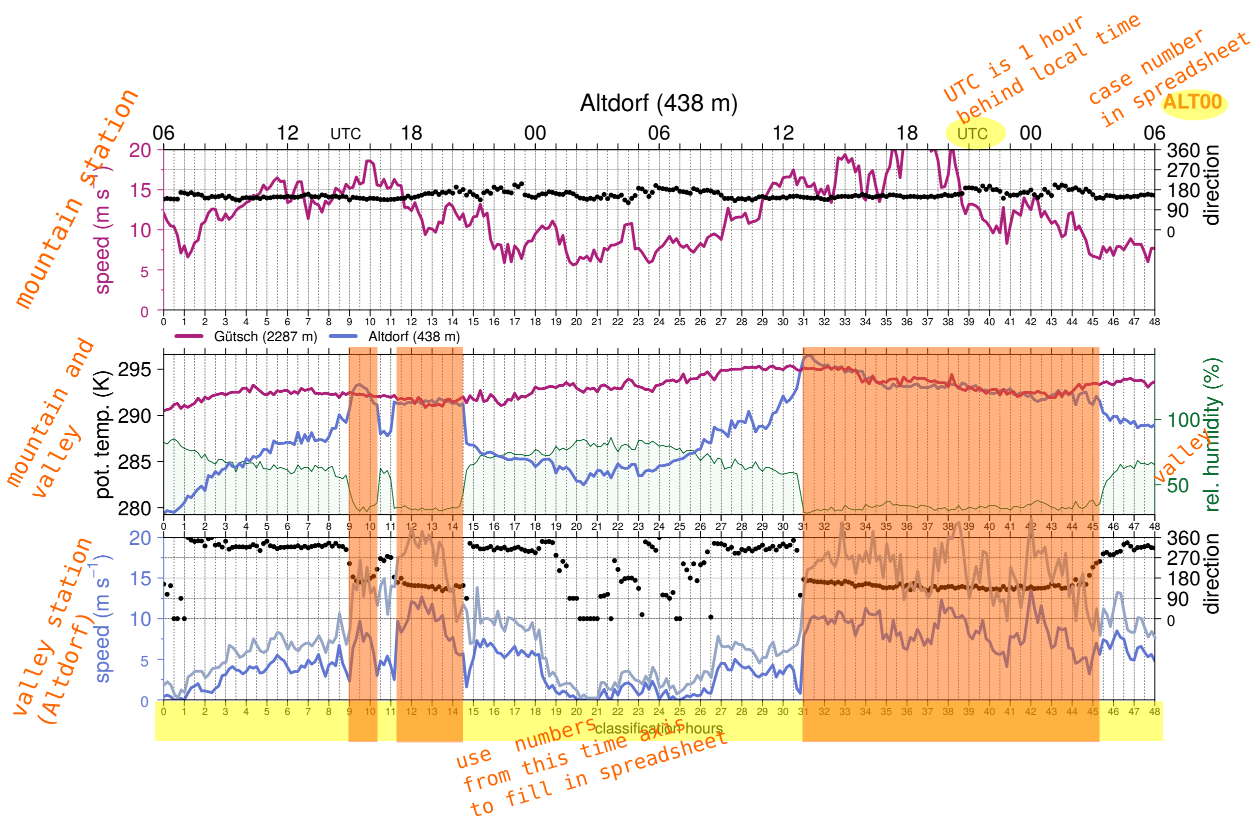

FIG. 3. Annotated time series of an additional case at location 2 (Altdorf) supplied to classifiers with the instruction package and other material. Upper panel: wind speed (magenta) and direction (degrees from N; in grey but bold black when from foehn sector) at the crest station (2287 m amsl); center panel: potential temperature at crest station (magenta) and classification location (blue) and relative humidity at classification location (green shaded); lower panel: wind speed (blue), gusts (light blue) and direction (black; degrees from N) at location 2. All values (except gusts) are averages over the previous 10 minutes. A hypothetical but not unreasonable classification of three foehn episodes at the station Altdorf is marked by orange rectangles (9:00-10:20; 11:10-14:30; 31:00-45:20). Foehn episodes had to be classified at a resolution of complete half hour segments and a minimum duration of 1 hour. In this example, foehn was classified between 9.0 – 10.0, 11.5 – 14.5, and 31.0 – 45.0.

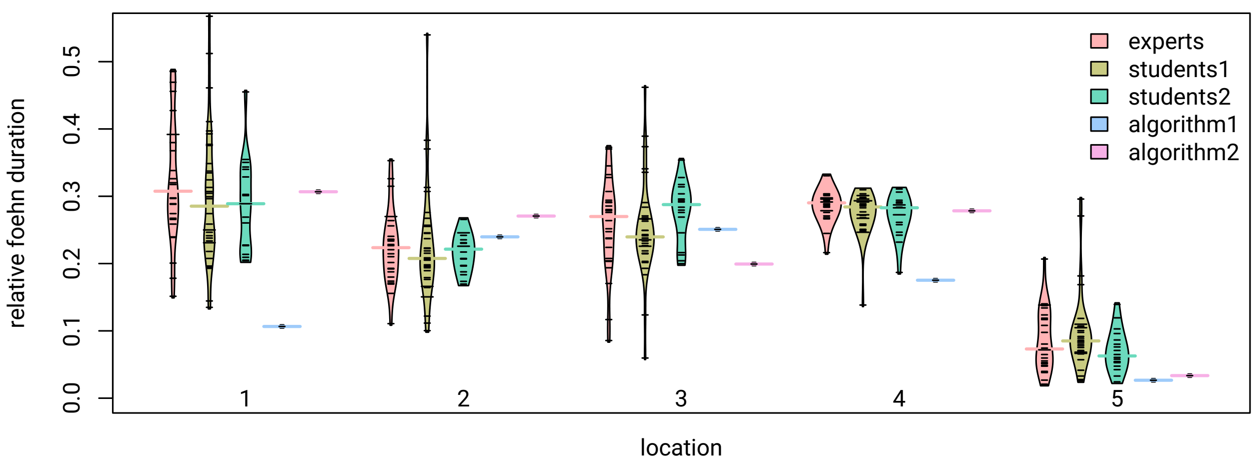

FIG. 4. Beanplots of classified foehn duration at each location relative to the total duration of the time series of 12 times 48 hours stratified by classifier groups: experts, two Masters student groups, and the two algorithms. (A1 for foehn mixed with valley air and pure foehn combined; for A2 a threshold of foehn probability of at least 50% is used). Black lines indicate individual classifications, colored lines the median of each group. Areas are the empirical densities of each group.

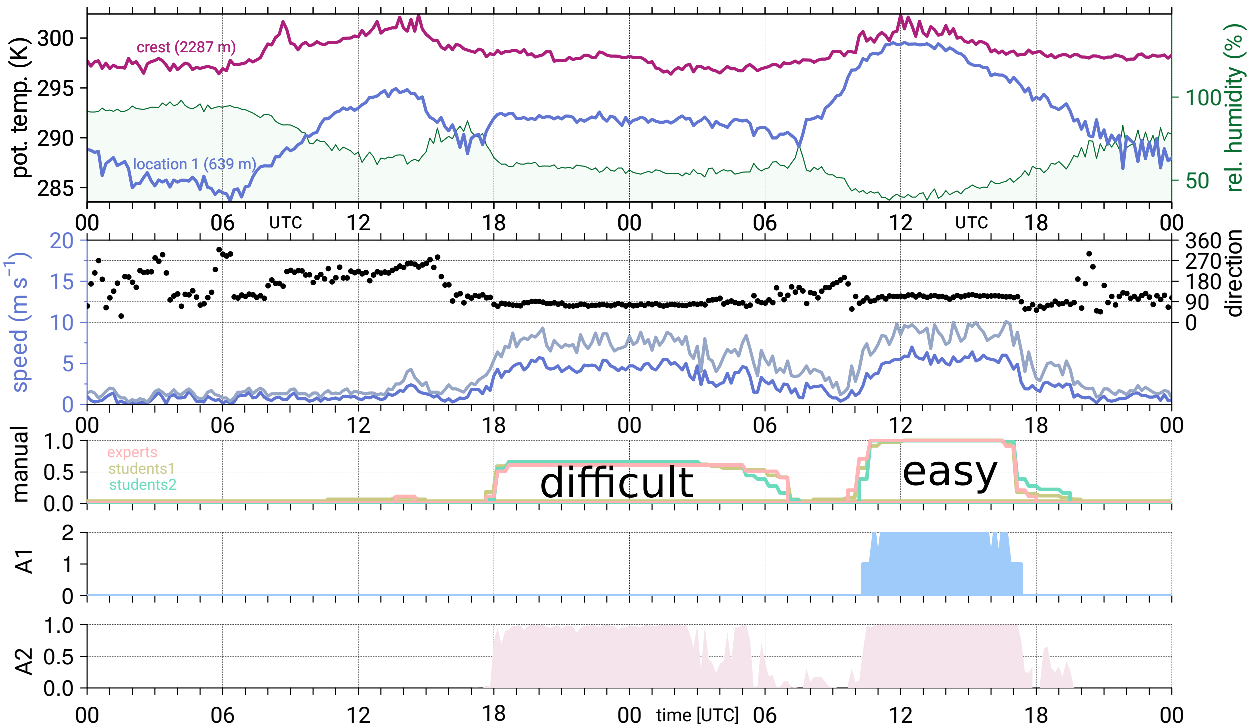

FIG. 5. Classification case at location 1. First panel: potential temperature (blue) and relative humidity (green) with added potential temperature at crest (purple). Second panel: wind direction (black), wind speed average (dark blue) and gusts (light blue) at location 1. Third panel: proportion of human classifier groups that classified foehn during the time series. Fourth panel: classifications with the three-category algorithm A1 (no foehn (0), foehn mixed with valley air (1), foehn (2)); Last panel: probability of foehn from the statistical mixture model A2.

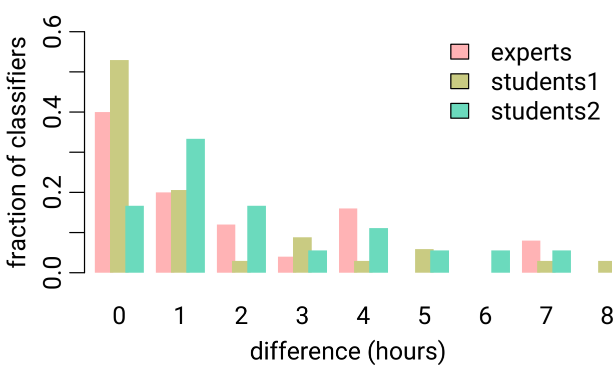

FIG. 6. Histogram of absolute difference in classified foehn duration (hours) between two identical cases. For perfect reproducibility all classifiers should have had 0 h difference. The bars are for the hour prior to and including the labeled duration difference.