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

Matthias Aichinger-Rosenberger[1], Elmar Brockmann[2], Laura Crocetti[1], Benedikt Soja[1], and Gregor Moeller[1]

[1]Institute of Geodesy and Photogrammetry, ETH Zürich, 8093 Zurich, Switzerland
[2]Swiss Federal Office of Topography (swisstopo), 3084 Wabern bei Bern, Switzerland

Correspondence: Matthias Aichinger-Rosenberger, maichinger@ethz.ch

Abstract. Remote sensing of water vapor using the Global Navigation Satellite System (GNSS) is a well-established technique and reliable data source for Numerical Weather Prediction (NWP). However, one of the phenomena rarely studied using GNSS are foehn winds. Since foehn winds are associated with significant humidity gradients between two sides of a mountain range, tropospheric estimates from GNSS are also affected by their occurrence. Time series reveal characteristic features like distinctive minima/maxima and significant decrease in correlation between the stations. However, detecting such signals becomes increasingly difficult for large data sets. Therefore, we suggest the application of machine learning algorithms for detection and prediction of foehn events from GNSS troposphere products. This initial study develops a new, machine learning-based method for detection and prediction of foehn events at the Swiss station Altdorf, by utilizing long-term time series of high-quality GNSS troposphere products. Data from the Automated GNSS Network Switzerland (AGNES) and various GNSS sites from neighbouring countries, as well as records of an operational foehn index are used to investigate the performance of several different classification algorithms based on appropriate statistical metrics. The two best-performing algorithms are fine-tuned and tested in three dedicated case studies. The results are promising, especially when reprocessed GNSS products are utilized and the most dense station setup is used. Detection- and alarm-based measures reach levels between 66-80% for both tested algorithms and thus are comparable to those from studies using data from meteorological stations and NWP. For operational prediction, some limitations due to the availability and quality of GNSS products in near-real time (NRT) exist. However, they might be mitigated to a significant extend by provision of additional NRT products and improved data processing in the future. Results also outline the importance of including geographically relevant stations (e.g. high-altitude stations) in the utilized data sets.

1 Introduction

Global Navigation Satellite Systems (GNSS) are used extensively for positioning and navigation applications worldwide. Additionally, they enable users to retrieve information about the state of the Earth's atmosphere, in particular the distribution of water vapor. This technique, commonly referred to as GNSS Meteorology, was first proposed three decades ago (Bevis et al., 1992) and is still gaining increasing interest from the scientific community as well as meteorological institutions. The retrieval

of atmospheric information from GNSS is based on the fact that electromagnetic signals (such as GNSS signals) are delayed when traveling through specific layers of the atmosphere. The delay experienced by a GNSS signal in the lowest part of the atmosphere (troposphere) is proportional to the water vapor content along the signal path. This fact is typically exploited in GNSS Meteorology by introducing GNSS-derived atmospheric parameters like the Zenith Wet Delay (ZWD) or the Zenith Total Delay (ZTD) in data assimilation schemes. In numerous studies, a positive impact has been demonstrated, especially on precipitation forecasts (see e.g. de Haan (2008), Brenot et al. (2013), Bennitt and Jupp (2012), Yan et al. (2009)). However, while mostly precipitation-related studies represent the current focus of research (see Guerova et al. (2016) for a comprehensive summary), other meteorological phenomena can also be investigated by means of GNSS. The number of studies on other meteorological processes is relatively small, covering phenomena such as thunderstorm activity (de Haan (2013)) or fog formation (Stoycheva and Guerova (2015), Aichinger-Rosenberger (2018)). Stoev and Guerova (2018) represent the only investigation on foehn winds using GNSS products to our knowledge, in an initial study for Bulgaria based on observations of Integrated Water Vapor (IWV).

Foehn winds are a characteristic weather phenomenon in mountainous regions all over the world, especially in the vicinity of prominent mountain ranges like the Alps (where it is typically referred to as Alpine foehn). In general, foehn can be characterized as 'a wind (which is) warmed and dried by descent, in general on the lee side of a mountain' (WMO, 1992). This definition already includes its major characteristics observed in affected areas: strong gusty winds, increasing temperatures and decreasing humidity. While there are many other effects of foehn winds (from social to economic impacts), large wind speeds and gusts are the most critical features from the perspective of operational forecasting and warning systems. In typical foehn valleys like the Reuss Valley (Switzerland) or the Wipp Valley (Austria) wind speeds up to 100 km/h are common, and even up to 200 km/h gusts can be observed at high altitude stations.

Foehn research denotes one of the major topics of (alpine) mountain meteorology (Steinacker, 2006). Despite the fact that the underlying physical processes of foehn have been studied for over a century, still some gaps in knowledge, especially concerning small-scale features, exist. As the classical thermodynamic foehn theory is not able to sufficiently explain all observed foehn events (especially those lacking precipitation), a number of different theories and extensions have been proposed. Furthermore, large observation campaigns like the Mesoscale Alpine Program (MAP) have been conducted and combined with NWP results in order to assess small-scale effects (Gohm and Mayr (2004), Lothon et al. (2006), Drobinski et al. (2007), Mayr et al. (2007)).

Despite these substantial efforts in research, classification and forecasting of foehn both are still challenging tasks. Classification by human expertise still provides the most accurate results, as dedicated experiments, comparing subjective and objective methods, reveal (Mayr et al., 2018). The ability of NWP models to predict foehn is limited by the fact that small-scale features still cannot be modelled with sufficient accuracy due to coarse representation of real-world topography (Wilhelm, 2012).

Machine learning (ML) techniques have been a major research topic in atmospheric sciences over the last decade. ML-related approaches of post-processing NWP output, known as model output statistics (MOS), have been shown to significantly enhance operational weather forecasts, see e.g. Glahn and Lowry (1972), Wilks and Hamill (2007) or Hess (2020). ML has also been used to assign uncertainty estimates to forecasts based on deep learning methods applied to previous forecasts (Scher

and Messori, 2018). Furthermore, the classification and detection of different weather types has been advanced and automated for different kinds of weather phenomena such as thunderstorms (Perler and Marchand (2009), Manzato (2005)), temperature forecasts (Yalavarthi and Shashi (2009)), wind systems (Kretzschmar et al. (2004), Otero and Araneo (2021)) or large-scale weather regimes in general (Deloncle et al., 2007). Common ML methods for such classification problems investigated in former studies are e.g.:

- Random forests: Deloncle et al. (2007)
- Adaptive boosting (AdaBoost): Perler and Marchand (2009), Sprenger et al. (2017)
- Support vector machines : Yalavarthi and Shashi (2009)
- Neural networks: Manzato (2005), Kretzschmar et al. (2004), Otero and Araneo (2021)

Only few authors have used ML approaches for detection and prediction of Alpine foehn yet. Initial studies were carried out by Sprenger et al. (2017), who applied the AdaBoost algorithm to a data set combining weather station observations with NWP output fields from the Consortium for Small-scale Modeling (COSMO) model. They found good performance of the algorithm, obtaining high values for probability of detecting foehn events (88%) and ratio for correct alarms of the algorithm (66%). The most recent study by Mony et al. (2021) showed the feasibility of using ERA5 reanalysis and climate model output instead of NWP output in a similar way as Sprenger et al. (2017). In addition, statistical mixture models were also applied for foehn diagnosis, e.g. by Plavcan et al. (2014).

The presented study represents an initial investigation on the usability of GNSS troposphere product time series for the detection and prediction of foehn events at the meteorological observation site Altdorf, Switzerland. Therefore we make use of state-of-the-art ML-based classification algorithms trained and tested on a data set spanning eleven years (2010-2020), derived from GNSS observations at sites all over Switzerland and neighbouring countries as well as a long-term record of foehn observations at Altdorf. The average performance of different ML algorithms is assessed via a cross-validation procedure. The two best-performing algorithms are trained over an eight-year and tested over a two-year period for different study setups (case studies). These case studies cover the usage of reprocessed troposphere products as well as NRT products, which could be used for operational prediction and detection of foehn events. Furthermore, we explore the performance of an extended station network, with the disadvantage of a shorter period of availability for algorithm training. Finally, we analyse the performance of our newly-developed method in detail over a week-long period covering two major foehn events at Altdorf.

2 GNSS Meteorology

As already outlined in the introduction, the concept of GNSS Meteorology is based on the fact that electromagnetic signals are delayed by the presence of the Earth's atmosphere. The signal delay is directly proportional to the refractive index n of the atmosphere. In the neutral atmosphere, the refractive index or refractivity N is composed of a dry (N_d) and wet part (N_w),

which depend on temperature T (K), as well as the dry P_d (hPa) and water vapor partial pressure e (hPa) respectively (Rüeger, 2002):

$$N = (n-1) \times 10^6 = N_d + N_w = \frac{77.6890 \cdot P_d}{T} + [\frac{71.2952 \cdot e}{T} + \frac{3.75463 \times 10^5 \cdot e}{T^2}]. \quad (1)$$

The total tropospheric delay experienced by a GNSS signal observed at an elevation el and azimuth direction a is referred to as the Slant Total Delay (STD)

$$\text{STD}(a, el) = \text{ZHD} \cdot mf_h(el) + \text{ZWD} \cdot mf_w(el) + mf_g(el) \cdot [\text{GN} \cdot cos(a) + \text{GE} \cdot sin(a)], \quad (2)$$

where ZHD (Zenith Hydrostatic Delay) represents the hydrostatic part, and ZWD the wet part of the signal delay in the zenith direction. In addition, horizontal gradients GN (north-south direction) and GE (east-west direction), accounting for the asymmetry of the atmospheric layers passed by the signal, can be estimated in GNSS processing. In order to map the delays and gradients estimated for the zenith direction to the correct elevation, mapping functions for both parts of the delay ($mf_h(e), mf_w(e)$) and the gradients ($mf_g(e)$) are used.
The total delay in the zenith direction, i.e. the Zenith Total Delay (ZTD), is the sum of the hydrostatic and wet part

$$\text{ZTD} = \text{ZHD} + \text{ZWD}. \quad (3)$$

ZHD accounts for the major part of the total delay and is largely determined by the atmospheric pressure. It can be modelled with sufficient accuracy from surface pressure observations using, e.g., the formula of Saastamoinen (Saastamoinen, 1972):

$$\text{ZHD} = \frac{0.0022767 \cdot p_s}{1 - 0.00266 \cdot \cos(2\theta) - 0.00028 \cdot H} \quad (4)$$

where p_s is the surface pressure, θ the station latitude, and H is the station height above the geoid.
ZWD represents the main signal of interest for meteorological purposes, as it is directly related to the water vapor content in the air column above the GNSS receiver, and therefore to IWV, via

$$IWV = \kappa(T_m) \cdot ZWD, \quad (5)$$

where κ denotes a semi-empirical function depending on the integrated mean temperature T_m. Thus, it shows the same high temporal and spatial variability as water vapor, making precise modelling from meteorological surface observations practically impossible. As a consequence, ZWD is commonly estimated as an unknown in GNSS parameter estimation alongside of station coordinates and the receiver clock error.

2.1 Influence of hydrometeors on GNSS signal delays

Since foehn events can occur with and without simultaneous precipitation, the influence of hydrometeor formation on GNSS signal delays should also be described in the following. In an initial investigation over two decades ago, Solheim et al. (1999)

quantified propagation delays induced in GPS signals by different types of molecular constituents such as dry air, water vapor, hydrometeors, and sand particles. They were able to show that the influence of water in both solid and liquid form on GNSS signals is significantly smaller than for its gaseous form (water vapor). However, in cases of extreme amounts of precipitation (especially in liquid form, i.e. extreme amounts of rain in very intense thunderstorms) a considerable influence for high-precision applications (as troposphere estimation) can be expected. In the framework of this study, we do not expect such severe events occurring with foehn events. Nevertheless, these events might lead to problems for our classification method (misclassification through degraded GNSS products due to high hydrometeor influence), especially in the summer months.

3 Data

3.1 GNSS station network and tropospheric products

All investigations presented in this study are based on GNSS troposphere products from the Automated GNSS Network Switzerland (AGNES). The AGNES network, which currently consists of 31 GNSS stations, was established in 2001 and is maintained by the Swiss Federal Office of Topography (swisstopo) (Brockmann et al., 2002). The capabilities of the network were extended to multi-GNSS in 2015 (Brockmann, 2016).

Reprocessed, long-term time series of hourly tropospheric delays and gradients are available for the period 1999-2020. A description of the data set as well as details on the reprocessing of GNSS data can be found in e.g. Brockmann (2015). Parts of this reprocessing were carried out in the framework of the second EUREF (International Association of Geodesy Reference Frame Sub-Commission for Europe) Permanent Network (EPN) reprocessing campaign in 2014, where GNSS data from a large number of European stations were reprocessed (Pacione et al., 2017). Therefore, also some interesting stations (from neighbouring countries such as Italy and Austria) are available and incorporated. More details on the actual selection of GNSS stations utilized for different case studies/experiments is given in Section 4.2. For this study, we use hourly GNSS troposphere products originating from reprocessing campaigns as well as operational NRT processing. The delay products are estimated every hour from 30-second measurements together with station coordinates (in this case using least-squares adjustment). Gradient products are typically estimated only every 12 hours and hourly values in between results from linear interpolation/extrapolation.

3.2 Foehn observations at Altdorf

In order to train a specific ML algorithm and evaluate its performance, a reference data set of foehn observations is needed as the target variable. This study uses time series of 10-minute estimates of foehn index (FI) calculated at the station Altdorf, following the approach presented by Dürr (2008). Altdorf is located at the exit of the Reuss Valley, between the Gotthard pass and the Lake Lucerne (see Figure 2), at a height of 449 m above mean sea level (amsl). The station has the longest time series of foehn observations in the Alps (spanning over 150 years of total observations) and FI data is provided back to 1981, for 10-minute intervals. It is also part of the National Meteorological Ground-level Monitoring Network (SwissMetNet, SMN) operated by MeteoSwiss. Currently, not only data from Altdorf but of about ten sites, frequently experiencing foehn winds,

are available on an operational level. The FI introduced by Dürr (2008) is designed for operational nowcasting and relies on typical foehn predictors such as wind speed and direction, pressure and temperature gradients, and humidity observations at the respective measurement site and surrounding stations. It returns three different integer values: 0 (no foehn), 1 (foehn-mixed air) and 2 (foehn), which are distinguished based on the predictors mentioned above. In an extensive validation against classifications by human experts, the index showed good performance for indices re-calculated back to 1981 (Gutermann et al., 2012). For a detailed description of the calculation algorithm we refer to Dürr (2008) and Gutermann et al. (2012). As we aim for a binary classification (no foehn or foehn), the cases of FI = 1 are treated as nonfoehn events and therefore mapped to value 0. Furthermore, we map the cases of foehn (FI = 2) to the value 1 for the sake of simplicity in all results shown in the following. Then, each hour in the whole data set where at least one 10-minute value indicates foehn is treated as an hour of foehn appearance, and thus a foehn event.

4 Methodology

4.1 Machine Learning Algorithms

In the course of this study, several different ML algorithms are tested in order to investigate their usability for this specific problem and to compare their performance relative to each other. The following algorithms, which have already been applied for classification of foehn or other meteorological phenomena (as listed in the introduction), are tested:

- Adaptive Boosting (AdaBoost) (Freund and Schapire, 1997)
- Gradient Boosting (GB) (Friedman, 2001)
- Multilayer Perceptron (MLP) (LeCun et al., 2012)
- Random Forest Classifier (RF) (Breiman, 2001)
- Support Vector Classifier (SVC) (Platt, 1999)
- K-Nearest Neighbor (KNN) (Cover and Hart, 1967)

As a detailed discussion of these algorithms would go beyond the scope of this study, we refer to Hsieh (2009) for a comprehensive overview.

4.2 GNSS station selection criteria

The final selection of GNSS stations, whose data we use as input features, is a difficult task for a number of reasons. The following list gives an overview of the main problems to keep in mind:

- Not every GNSS station provides a continuous data set of troposphere products. In fact, almost all of the available stations have data gaps over the chosen study period (2010-2020) and a large number of stations was only established after 2010.

– The ML-based detection/prediction can only be applied for foehn events where troposphere products from all selected GNSS stations (i.e. all features/predictors) are available. Since not all GNSS stations have data gaps at the same periods, the actual amount of possibly missed events can add up.

In order to document and cope with these challenges, we calculated detailed statistics for data availability for each station/feature setup used within this study, which are given in the respective result sections. Using these statistics, we try to balance the need to incorporate important stations (e.g. because of their geographic location) and still cover a sufficient amount of foehn events (for both training purposes and performance assessment). In the following, we therefore introduce the rules for the selection of GNSS stations applied in this study:

– Selected stations need to be relevant for the foehn classification, in terms of their geographical location. Based on previous studies we define an area of interest with latitude and longitude borders of [44.9°, 48.2°] and [5.9°, 10.75°] respectively, which is also shown in Figure 2.

– Time series from a selected station (of all troposphere products used) should cover at least 95% of all foehn events (hours of FI = 1) in the study period (2010-2020).

– The overall feature setup (GNSS stations and available products) should be available for at least 50% of all foehn events in the study period.

4.3 Feature selection from GNSS time series

The selection of features from GNSS troposphere time series is based on previous investigations on visual detection from time series of different parameters as well as on obvious choices which are expected to be impacted most by foehn conditions. Two obvious choices are visualized for December 2019 in Figure 1, namely ZWD (12-hour moving-average) at stations north (KALT, shown in blue) and south (LOMO, shown in red) of the Alpine ridge and its difference (bottom section, shown in black). In addition, foehn events at Altdorf are shown as color-coded periods (orange). Strong correlation between the contrary trends in ZWD at the two stations and the onset of foehn in Altdorf can be observed. Furthermore, the difference in ZWD between the stations reaches a (negative) maximum in the two extended foehn periods observed ($\sim$ 15.-18. and 19.-21.12.2019). These time series give a first impression how (and from which parameters) foehn events can be detected using GNSS data sets. As this becomes a very demanding task for longer periods (both visually and analytically), ML techniques are a promising tool to extend and automate such a detection process, with the additional benefit of possibly providing the ability to also predict upcoming events.

4.4 Default study setup

In the following the default feature setup for the algorithm comparison (cross-validation) and the first (reference) case study is introduced. In general, the definition of a specific setup concerns the following points:

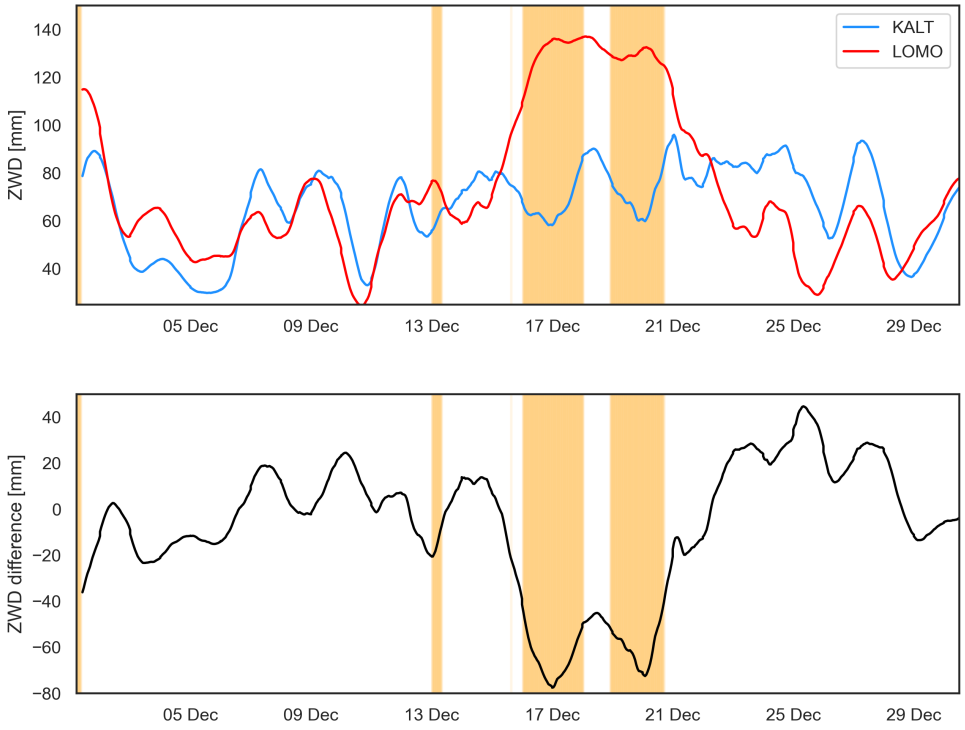

Figure 1. Time series of promising foehn predictors for December 2019. Top: ZWD (12-hour moving-averaged for visualization) from stations KALT (north of Alpine ridge, blue) and LOMO (south of Alpine ridge, red), Bottom: ZWD difference between KALT and LOMO (black). Observed foehn events at Altdorf are visualized as orange areas.

- GNSS station network: The selection of GNSS stations, from which data is utilized for training and testing the ML algorithms used, should be compliant with the criteria outlined in Section 4.2.
- Study period: In combination with the chosen station selection, a sufficient time period must be chosen to match the criteria outlined in Section 4.2.
- Features: This point defines which troposphere products should (or could) be used for this specific setup.

The detailed setup and statistics of data availability for the cross-validation and the first case study (CS1) is given in Table 1. It includes ZWD (absolute values and all possible differences between stations) and gradient products (GN and GE) from the chosen station network (Figure 2) as well as a selection of four ZHD differences which are representative differences between north-south stations in the network. Tests have also been conducted using all possible differences in ZHD (as for ZWD), but

no improvement was found using this setup. This might be explained by the fact that ZHD is largely depending on pressure, which typically does not show such small-scale variations as water vapor (and thus ZWD). Therefore, a small number of ZHD differences (i.e. pressure differences) across the Alpine ridge might be sufficient. Figure 2 provides a visualization of the full

Table 1. Default setup used for the cross-validation and CS1. For features not outlined in a specific row, by default all stations and combinations of stations are used.

Training period	2010-2018
Test period	2019-2020
Station setup	Post-processed: red triangles in Figure 2
Feature setup	ZWD, GN, GE, ZWD_diff, ZHD_diff
ZHD_diff combinations	KALT-STA2, LUZE-STA2, BOU2-STA2, SIGM-TORI, ETHZ-TORI
Total number of features	564
Foehn events (FI == 1)	5642 hours
Foehn events without GNSS data	2049 hours

station network utilized in this study. The different colored triangles represent different station setups for the case studies introduced later (red triangles represent stations used in the default setup, blue triangles represent stations which are added for the last case study). In addition, a complete list of utilized stations (including geographical coordinates) can be found in the Appendix A.

4.5 Data preparation

One of the main challenges for ML-based classification algorithms are imbalanced data sets. This imbalance is also strongly present in data sets of foehn observations, since foehn is a rather rare meteorological phenomenon. For the utilized FI data set, the average foehn probability over the 11-year period (2010-2020) amounts to only $\sim 4\%$. Thus, the ratio of under-representation of the minority class (foehn event) compared to the majority class (no foehn event) is as large as 1:25.

4.5.1 Oversampling

A common approach to overcome problems originating from high imbalance in a data set is to oversample the minority class for the training data set. One possible approach to achieve this is the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002), which we use in this study. The technique creates new (synthetic) instances of the minority class within the training data. For this study, an oversampling of observed foehn hours in the training data set by 25% was conducted using SMOTE, which improves the performance of the applied algorithms by about 20%. The value of 25% oversampling was chosen to achieve a reasonable balance between the advantage of having more usable training events (larger percentage of oversampling) and the fact that foehn is still a rather rare phenomenon (therefore also rare in possible test data sets). All results shown in the following sections are based on pre-processing using this approach.

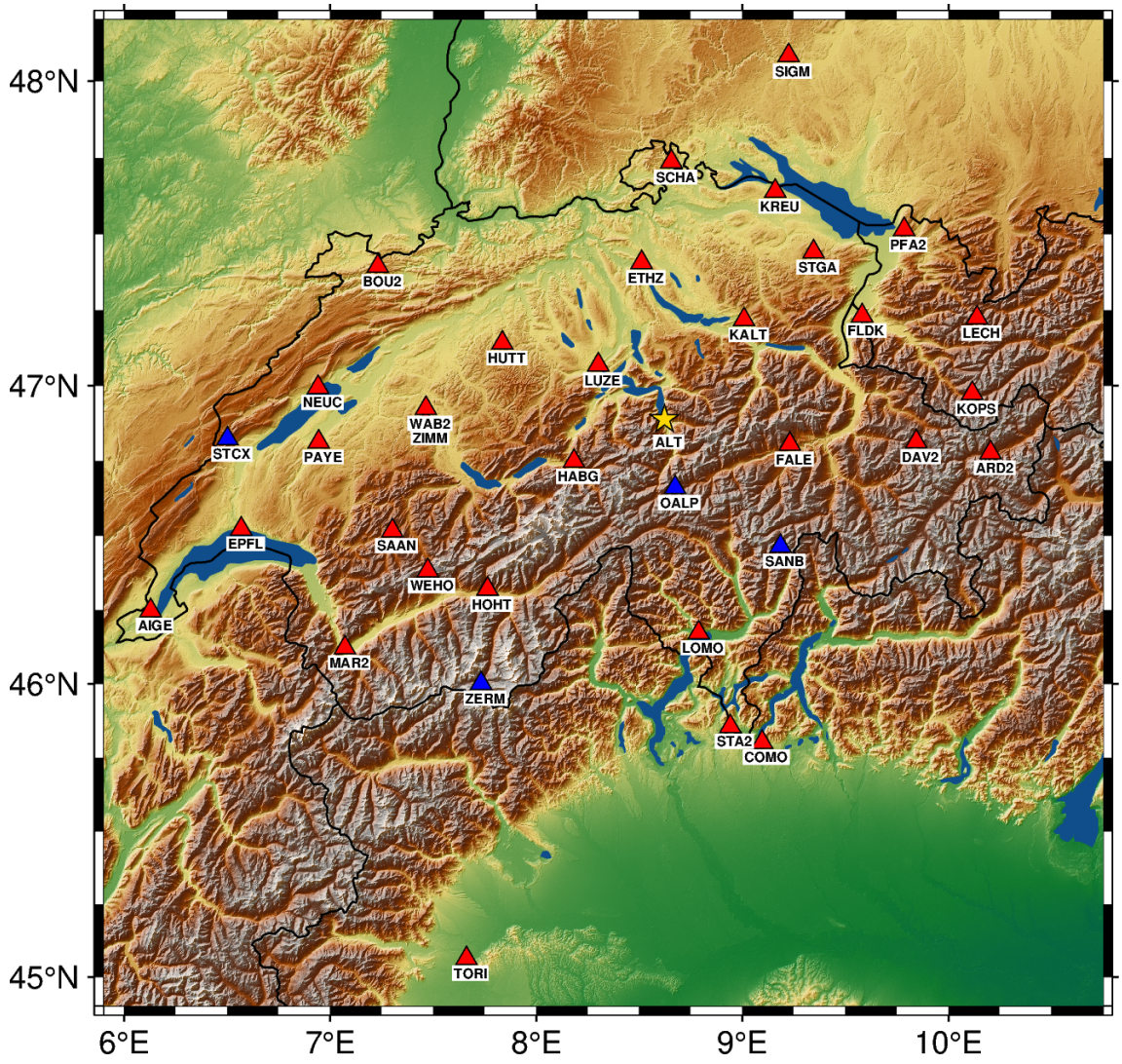

Figure 2. GNSS station network utilized for this study. Stations included in the default setup (cross-validation and all case studies) are marked as red triangles. Stations explicitly used in case study 3 (CS3) are marked as blue triangles. In addition, the SMN-station Altdorf is marked with a yellow star.

4.5.2 Shifting of FI time series

In order to assess the suitability of the GNSS troposphere products for operational prediction, a time shift of one hour is applied to the target vector (i.e. FI time series at Altdorf). As operational usage is considered a future goal of the proposed method, the shift is applied for all test cases investigated in this study (also those using post-processed GNSS products). Therefore, each prediction of a foehn event is based on GNSS observations collected one hour before a possible onset of foehn at Altdorf.

4.6 Performance metrics

As already outlined in the last section, the imbalance in data sets of foehn observations is a major obstacle for the application of ML algorithms and the assessment of their performance. For highly imbalanced data, performance metrics typically used in ML studies might not be representative and therefore other options have to be explored. In the case of the present data set, a typical performance measure such as precision alone would not be suitable as it simply compares detected/predicted foehn hours with the observed data for all time steps. Thus, it might happen that an algorithm with optimal precision predicts (almost) no foehn events at all, since this will still result in an optimal performance with regards to precision. In order to overcome these issues we adapt the following performance metrics, as introduced by Barnes et al. (2007) and used in Sprenger et al. (2017). These can be formulated as conditional probabilities P(|) and separated into detection-based:

- Probability of Detection (POD) = P(predicted | observed)
- Probability of False Detection (POFD) = P(predicted | not observed)
- Missing Rate (MR) = P(not predicted | observed)

and alarm-based metrics:

- Correct Alarm Ratio (CAR) = P(observed | predicted)
- False Alarm Ratio (FAR) = P(not observed | predicted)
- Missing Alarm Rate (MAR) = P(observed | not predicted).

As already visible from the formulations above, POD and CAR are directly connected to each other via the Bayes Theorem. This also implies that there is always a trade-off between those two parameters and therefore only one of them can be optimized, while decreasing the respective other one. Which metric should be optimized strongly depends on the actual application, as already outlined by Sprenger et al. (2017) who argued that alarm-based measures might be more relevant from a forecaster's perspective.

In addition, we adopt two measures which represent a kind of mean performance in terms of both CAR and POD for describing the results presented in the next sections. The first one is just the simple average of those two parameters combined, therefore referred to as COMB in the following:

$$\mathrm{COMB} = \frac{\mathrm{POD} + \mathrm{CAR}}{2}. \tag{6}$$

The second adopted metric is based on the F-beta score F_β (Baeza-Yates and Ribeiro-Neto, 1999), which can be formulated using the confusion matrix. The matrix reports the number of false negatives (FNs), false positives (FPs), true negatives (TNs), and true positives (TPs) and thus allows for the calculation of common performance measures in ML such as precision, recall. Plots of the confusion matrix will also be provided for all case study results in Section 6. Using precision and recall measures,

the F_β score can be computed for varying β:

$$F_\beta = (1+\beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}} \tag{7}$$

The classical F-beta score (F_1, using β = 1) represents the weighted harmonic mean of precision and recall, with a range between 0 (worst case) and 1 (optimal value). As already discussed above, a precision measure might not be representative for results of this study and therefore we use the F_2 score instead of F_1, putting more emphasis on the recall, i.e. on the detection of all foehn events:

$$F_2 = 5 \cdot \frac{\text{precision} \cdot \text{recall}}{(4 \cdot \text{precision}) + \text{recall}}. \tag{8}$$

5 Algorithm selection and tuning

This section gives an overview on the process of algorithm selection using a cross-validation approach applied for the whole training data set. Details on this approach are given in the following Section 5.1.

5.1 Choosing optimal ML algorithms

Before carrying out case studies using a specific ML algorithm, the most promising one(s) have to be identified from the list of algorithms given in Section 4.1. Therefore, a cross-validation over the training data set (2010 - 2018) was performed and evaluated using the performance metrics introduced in Section 4.6. For the cross-validation, single years of data are iteratively taken out of the training data set, serving as validation data in order to assess the performance of the outlined algorithms. This is repeated until every year serves once as validation data set. The actual implementation is carried out using the Python package scikit-learn (version 1.1.1) (Pedregosa et al., 2011). For these cross-validation runs, the default settings from the algorithm routines are used in order to get an objective, initial picture of their performance for this problem. More sophisticated approaches like including more (complex) algorithms or running grid searches for each algorithm separately before the comparison were not followed due to the "proof-of-concept" focus of the study. However, these approaches might be used in future studies on this topic, in order to optimize the overall performance (see also the future ideas described in the outlook (Section 8)).

Resulting statistics of the cross-validation are summarized in Table 2. These results indicate best performance for the SVC algorithm in terms of combined measures (COMB and F_2 score). For detection-based measures (POD), the GB algorithm achieves the highest value on average as well as for most single years. The same holds for the RF algorithm in terms of alarm-based measures (CAR), but its detection-based performance is significantly degraded compared to e.g. GB and SVC. Thus we decided to use the GB and SVC algorithms for evaluation in the test case studies, as those are the only ones providing average combined measures of over 70% (see Table 2).

Table 2. Averaged performance metrics of the cross-validation for all tested algorithms over the nine year training period 2010-2018.

Algorithm	POD	CAR	COMB	F_2	POFD	MAR
AdaBoost	0.538	0.516	0.527	0.530	0.031	0.056
GB	**0.790**	0.635	0.713	0.751	0.045	0.066
MLP	0.642	0.682	0.662	0.645	0.036	0.051
RF	0.570	**0.814**	0.692	0.605	0.032	**0.038**
SVC	0.780	0.685	**0.733**	**0.758**	0.044	0.061
KNN	0.722	0.523	0.623	0.669	0.042	0.074

5.2 Hyperparameter tuning

Based on the results of the cross-validation, the GB and SVC algorithm are chosen for in-depth tuning of their hyperparameters. Therefore, a (small-scale) grid search procedure is conducted, which is an exhaustive search over a subset of manually selected values. The performance of all hyperparameter value combinations is evaluated based on a three-fold cross-validation. Therefore, the training data set (2010-2018) is randomly divided into three folds, where two thirds are used for training while the last third serves for validation. This procedure is repeated three times until each third is used once for validation. All tested hyperparameter values, as well as the best performing value combinations, are summarized in Table 3. We waive to do a more intensive grid search procedure for reasons already mentioned in the cross-validation section ("proof-of-concept" study), although we plan to optimize the performance of the method through this in future studies.

Table 3. Tuned hyperparameters of both algorithms with their best, tested and default values. For GB, n_estimators stands for the number of boosting stages to perform and max_depth is the maximum depth of the individual regression estimators that limits the number of nodes in the tree. The contribution of each tree is limited by the learning_rate. For SVC, C is the regularization parameter and gamma represents the kernel coefficient.

Algorithm	Hyperparameter	Best Value	Tested Values	Default Value
GB	n_estimators	300	[100, 300, 500]	100
	max_depth	5	[3, 5, 8]	3
	learning_rate	0.1	[0.05, 0.1, 0.2]	0.1
SVC	C	0.1	[0.1, 1, 10, 100, 1000]	1
	gamma	'scale'	[1, 0.1, 0.01, 0.001, 0.0001,'scale']	'scale'

6 Results: Case studies

As the major performance test of the proposed method, three case studies are performed. Within these case studies, different setups regarding utilized GNSS stations and tropospheric parameters in the feature matrix are investigated.

6.1 Case Study 1: Reprocessed products

Case study 1 (CS1) investigates the performance of the chosen algorithms for an optimal combination of the largest station network and longest study period, still compliant to the selection criteria from Section 4.2. For this setup, the reprocessed time series of GNSS troposphere products are used, thus also tropospheric gradients can be utilized.

Resulting statistics for CS1 from both algorithms are given in Table 4. Both algorithms show a promising performance, which, on average, is comparable to the results reported by Sprenger et al. (2017). For the GB algorithm, POD values significantly (13%) lower than for the reference study, are balanced by an improvement in CAR by $\sim$ 10%. The SVC algorithm tends to predict more events (higher POD) and therefore also produces more false-alarms (higher POFD, lower CAR), somewhat similar to the AdaBoost algorithm of Sprenger et al. (2017). Combined measures (COMB and F_2) indicate very similar overall performance for both tested algorithms, with GB having the advantage of more balanced (almost equal) values for detection- and alarm-based measures. Furthermore, the actual probability of foehn events predicted by both algorithms over the test period (3.7/4%), lies within one percent of the actual observed probability (4.7%) Figure 3 shows the confusion matrix of GB (left)

Table 4. Performance metrics for the proposed models using post-processed troposphere products and the full feature setup (as shown in Table 1).

Algorithm	POD	CAR	COMB	F_2	POFD	MAR	P_predicted	P_observed
GB	0.753	0.764	0.758	0.7555	0.04	0.046	0.037	0.047
SVC	0.804	0.663	0.733	0.771	0.04	0.057	0.040	0.047

and SVC (right) algorithm for CS1. The actual values of TN (top right corner), FP (top left corner), FN (bottom left corner) and TP (bottom right corner) for this case study. First of all, the results show the large imbalance of the classification problem, with TN values being approximately 30 times larger than TP values. That is also the reason why we chose a logarithmic color scale for the visualization. Furthermore it can be seen, that SVC produces a significantly larger amount of FP values (false alarms), which is also reflected in the statistics given in Table 4. On the other hand, it also misses less events compared to the GB algorithm, resulting in a lower number of FNs and thus higher MAR. GB shows an approximately equal number of FP and FN values, but providing more TN values than SVC.

Another major advantage of the GB algorithm is the fact that it gives one the opportunity to assess the importance of the used features for the prediction result. In Figure 4, we show the 20 most important predictors (features) for the classification carried out for CS1. By far the best predictor is the ZWD difference between the stations FLDK and FALE, which is surprising due to the fact that both stations are not as close to Altdorf as others in the utilized network. Interestingly, also features from stations

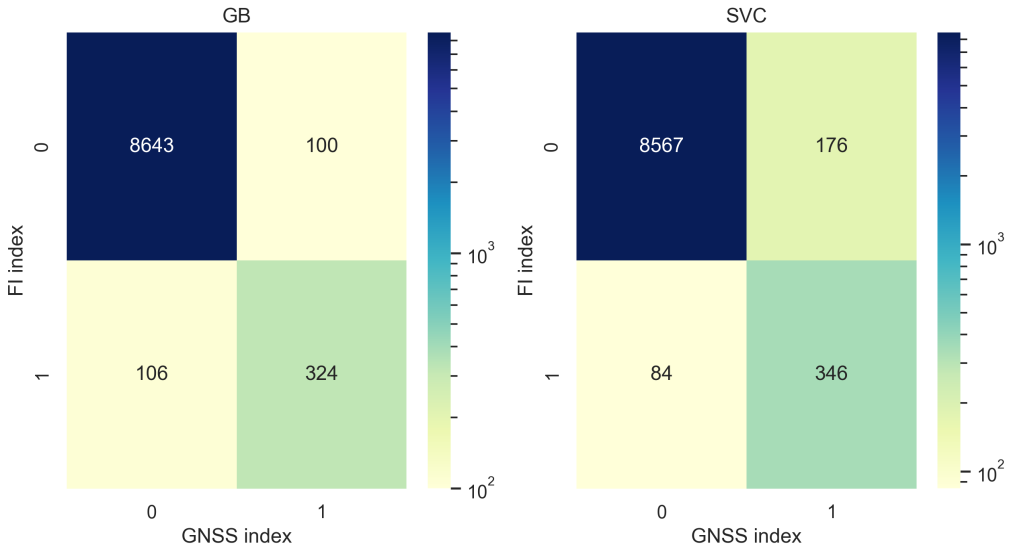

Figure 3. Confusion matrices for GB (left) and SVC (right) for the CS1 setup. Values of TN (top right corner), FP (top left corner), FN (bottom left corner) and TP (bottom right corner) are given. Please note that a logarithmic color scale is used for better visualization.

even further away from Altdorf have significant impact, most prominently ZWD and also gradient products (even for east-west direction), e.g. in the Valais area (WEHO and HOH2 stations). This is reasonable due to the fact that typical wind trajectory in the Rhone valley is east-west oriented. ZHD differences, representing larger-scale pressure gradients (such as LUZE-STA2 or ETHZ-TORI) are also found in Figure 4. This is consistent with the study from Sprenger et al. (2017), where the pressure difference between Zürich and Locarno was the single most important predictor.

6.2 Case study 2: NRT products

Case study 2 (CS2) addresses the major question to what degree the proposed method can be used for (or incorporated into) operational forecasting of foehn events. Therefore, the investigations presented for CS1 are extended by using NRT troposphere products for the two-year test period. This way, we investigate the suitability of the proposed ML algorithms for operational prediction. NRT troposphere products are currently available in form of tropospheric delays (ZHD, ZWD and ZTD), which are typically provided with a latency of approximately 30-40 minutes after each full hour. Unfortunately, no atmospheric gradients are currently delivered in NRT mode, but an extension is possible and aimed for in the near future. The missing gradient information makes it necessary to train the GB and SVC algorithms again for the dedicated period (2010-2018 of reprocessed products), but this time only using features related to tropospheric delays (ZWD, ZTD, ZWD differences, ZHD differences). The detailed setup used for CS2 is again aligned with the criteria outlined before and given in Table 5. Resulting statistics of the prediction using NRT products can be found in Table 6. In comparison to CS1, a general performance decrease in all

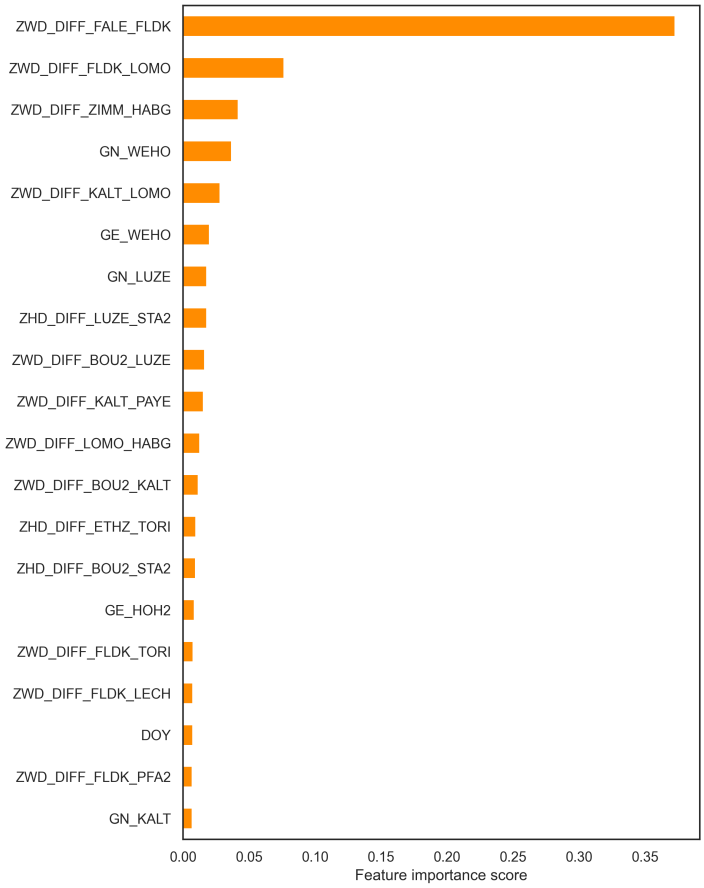

Figure 4. Feature importance score of the 20 top predictors for the GB algorithm for the setup used in CS1.

Table 5. Default setup used for CS2. For features not outlined in a specific row, by default all stations and combinations of stations are used.

Training period	2010-2018
Test period	2019-2020
Station setup	Default network: Red triangles in Figure 2
Feature setup	ZWD, ZTD, ZWD_diff, ZHD_diff
ZHD_diff combinations	KALT-STA2, LUZE-STA2, BOU2-STA2, SIGM-TORI, ETHZ-TORI
Total number of features	501
Foehn events (FI == 1)	5642 hours
Foehn events without GNSS data	2004 hours

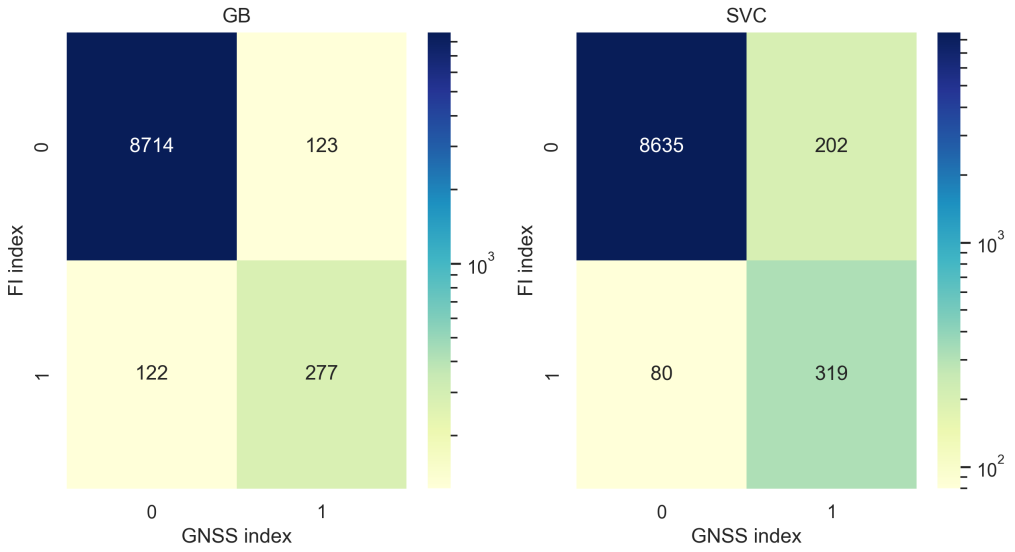

Figure 5. Confusion matrices for GB (left) and SVC (right) for the CS2 setup. Values of TN (top right corner), FP (top left corner), FN (bottom left corner) and TP (bottom right corner) are given. Please note that a logarithmic color scale is used for better visualization.

measures is apparent for both algorithms. For the SVC algorithm, the degradation is more pronounced for the alarm-based ($\sim$ -6% in CAR) than for detection-based measures (almost equal performance). The GB algorithm shows equal degradation for both types. Figure 5 shows the confusion matrix for CS2, in a similar way as Figure 3 for CS1. Results for the different

Table 6. Performance metrics for CS2 using NRT GNSS products.

Algorithm	POD	CAR	COMB	F_2	POFD	MAR	P_predicted	P_observed
GB	0.699	0.705	0.702	0.700	0.04	0.03	0.0428	0.0432
SVC	0.799	0.612	0.706	0.753	0.04	0.03	0.0564	0.0432

algorithms are almost identical to those of CS1. As for CS1, SVC produces a larger amount of FP values, but misses less events compared to the GB algorithm. GB again shows an approximately equal number of FP and FN values and provides more TP values.

This indicates the importance of gradient parameters for the proposed method, as already indicated in feature importances of CS1 (Figure 4). Furthermore, lower quality of ZWD estimates must be taken into consideration as well for the NRT solution, since lower-quality orbit and clock products have to be used for GNSS processing. Nevertheless, combined measures (F_2 and COMB) are still reaching values around 70%, which might already qualify the method as a potentially beneficial additional tool for operational forecasting. In the absence of gradient parameters, ZWD differences dominate the top 20 of most important

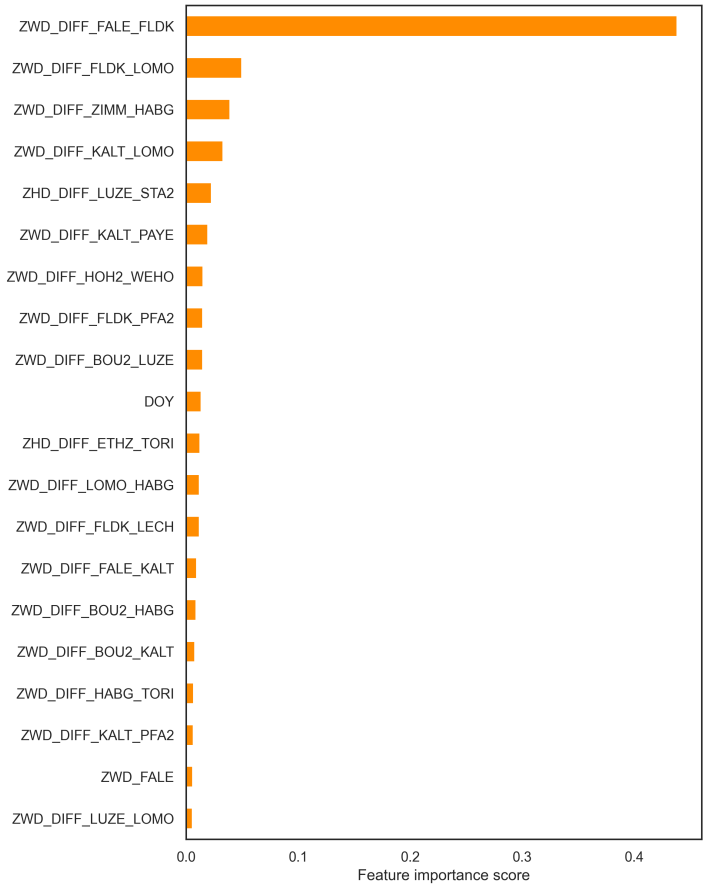

Figure 6. Feature importance score of the 20 top predictors for the GB algorithm for the setup used in CS2.

features of the GB algorithm as to be expected (see Figure 6). The dominance of the top predictor (again ZWD difference between FALE and FLDK) is even more pronounced than for CS1. Nevertheless also ZHD differences and DOY are again present and absolute ZWD for certain stations also shows up.

6.3 Case study 3: Shorter period/more stations

Case study 3 (CS3) investigates the question if the time period of training data needed can be reduced when at the same time new stations (and therefore features) are introduced. This is especially important since it allows for the introduction of geographically interesting stations (e.g. located at higher altitudes such as SANB or OALP, see Appendix A for more details), which might be beneficial for the algorithm performance. The detailed setup of this case study is again given in Table 7. The feature setup remains unchanged from the default setup, but in addition four stations (marked in blue in Figure 2) are added to the station setup. The study period is adjusted to 2015-2019 in order to again be aligned with the selection criteria and

this period is split into training and test period by 80/20% proportion. Resulting statistics are given in Table 8. The results

Table 7. Default setup used for CS3. For features not outlined in a specific row, by default all stations and combinations of stations are used.

Training period	2015-2018
Test period	2019
Station setup	Post-processed: Red and blue triangles in Figure 2
Feature setup	ZWD, GN, GE, ZWD_diff, ZHD_diff
ZHD_diff combinations	KALT-STA2, LUZE-STA2, BOU2-STA2, SIGM-TORI, ETHZ-TORI
Total number of features	738
Foehn events (FI == 1)	2650 hours
Foehn events without GNSS data	1244 hours

Table 8. Performance metrics for CS3.

Algorithm	POD	CAR	COMB	F_2	POFD	MAR	P_predicted	P_observed
GB	0.761	0.823	0.792	0.773	0.065	0.058	0.065	0.071
SVC	0.853	0.772	0.812	0.836	0.078	0.065	0.078	0.071

show large improvements for both algorithms, especially for alarm-based measures, compared to the former case studies. This outlines the importance of including geographically relevant stations (especially stations at the crest, i.e. higher altitudes) for the performance of the methods. The confusion matrix for CS3 is again visualized in Figure 7. Overall, its values show the same behaviour as already observed for the other case studies. However it should be noted that e.g. the relative amount of TPs is much higher for this case study, which is also reflected in the (improved) resulting statistics. The benefit of the newly introduced stations is also reflected in the feature importances of the GB algorithm, visualized again in Figure 8. In this case, the top predictor of CS1 and CS2 (ZWD difference between FLDK and FALE) is superseded by the ZWD difference between FLDK and SANB (station at San Bernadino pass, 1702 masl). The station SANB, as well as other high-altitude stations (such as OALP), is present a few times in the top 20 for CS3. ZHD differences are actually no longer among top predictors for CS3, which might be explained by the fact that ZWD observations (and having them at relevant location) still provides significantly more valuable information.

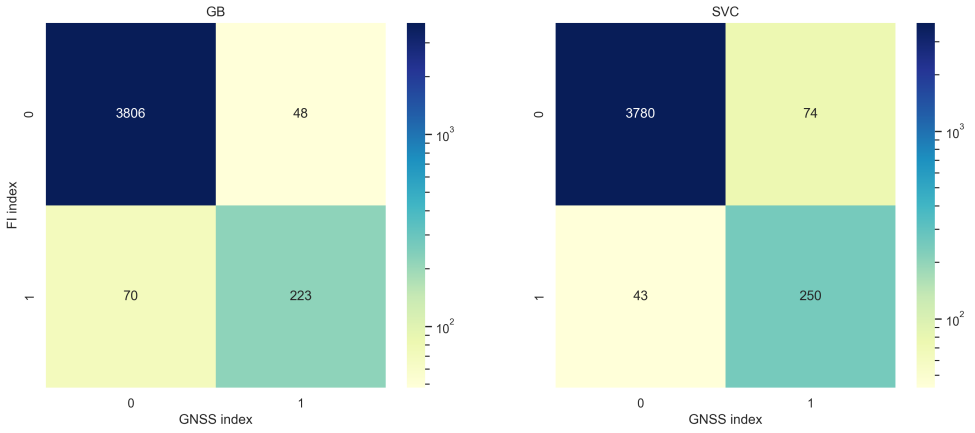

Figure 7. Confusion matrices for GB (left) and SVC (right) for the CS3 setup.Please note that a logarithmic color scale is used for better visualization.

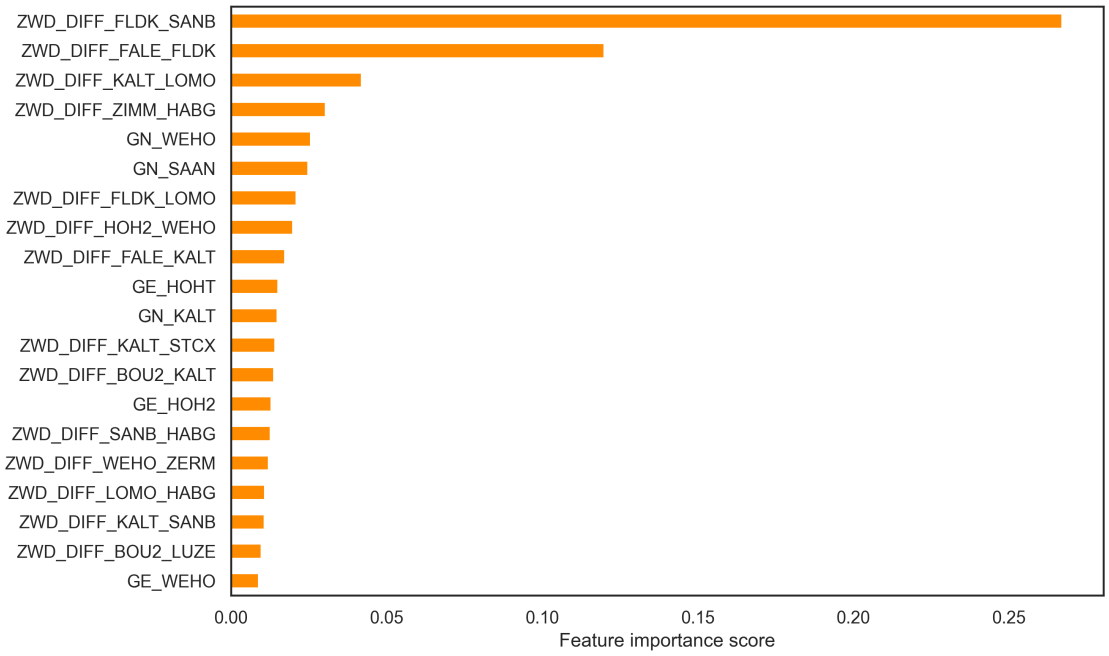

Figure 8. Feature importance score of the 20 strongest predictors for the GB algorithm, when using the CS3 setup.

7 Results: Performance analysis for distinctive foehn events

The performance statistics shown in the last section give an overview of the average performance over the test period. In order to gain more insight on the performance of our method for specific events, we look at a week-long period (including two foehn events) at Altdorf in the year 2019 in more detail. Therefore we make use of results from both algorithms for CS3, and evaluate their performance for this specific week against the operational FI.

Between 15.-21.12.2019, two major (south) foehn events were observed at Altdorf, for which we analyse the performance of our method in the following. Relevant meteorological parameters describing the situation are visualized in the upper two panels of Figure 9. The upper part shows observations of temperature (red) and relative humidity (green) as well as wind speed (black) and direction (grey) for the time period 15.-21.12.2019 (DOY 349-355). The first major event started around midnight at DOY 350 (16.12.2019), as visible in significant increase/decrease in temperature/relative humidity as well as the onset of a strong southerly flow (up to 60 km/h). This situation persisted over 48 hours, until the early morning hours of DOY 352 (18.12.2019). After a short break during daytime, the second major event started in the late evening of that same day and again lasted close to 48 hours until the evening of DOY 354 (20.12.2019), again accompanied by similar conditions observed at the SMN station.

The bottom part of Figure 9 shows the classification results of our method for GB (orange) and SVC (blue) algorithm as well as the reference FI index (red) from MeteoSwiss. First of all, it must be noted that for both the initial hours of the first major event and the last hours of the second event, no full feature matrix was available and therefore no classification was possible. This highlights the major drawback/limitation of the introduced method. However, as soon as all necessary data is available, the event is captured by both algorithms.

An overall assessment of the results shows a good average performance of the GNSS-based classification, especially for the GB algorithm. However, the transition from foehn to non-foehn and vice versa is not as accurately captured as in the operational FI index. The reason therefore might be the slightly longer response time of the GNSS-based parameters like ZWD to a change in synoptic conditions, compared to classical meteorological parameters. Thus, it might be beneficial to introduce time lags on some of the predictors depending on the actual physical parameter and the geographical location of the contributing stations. This aspect (among other ideas for improvement) is discussed further in the Outlook Section 8.

The SVC algorithm tends to over-predict foehn during the break period and therefore issues a large number of false alarms. This fact is also reflected in the statistics of all case studies shown in Section 6 (higher POD but also higher POFD/lower CAR).

In order to analyse the performance of the methods in more detail, and increase understanding of the physical processes captured in the GNSS products, we additionally analyse the time series of the most important predictors (features) of the introduced method. Figure 10 shows the six top predictors of the GB classification for CS3, as shown in Figure 8 as well as the observed foehn periods (color-coded in red). Most of them show distinctive patterns for foehn as well as non-foehn periods, which are consistent with the physical relationships between GNSS observations and meteorological conditions (as introduced in Sections 2 and 4.3) experienced for the respective period. ZWD differences show local minima or maxima for the major foehn periods, as already suspected in Section 4.3. Whether minima or maxima are observed depends on the actual

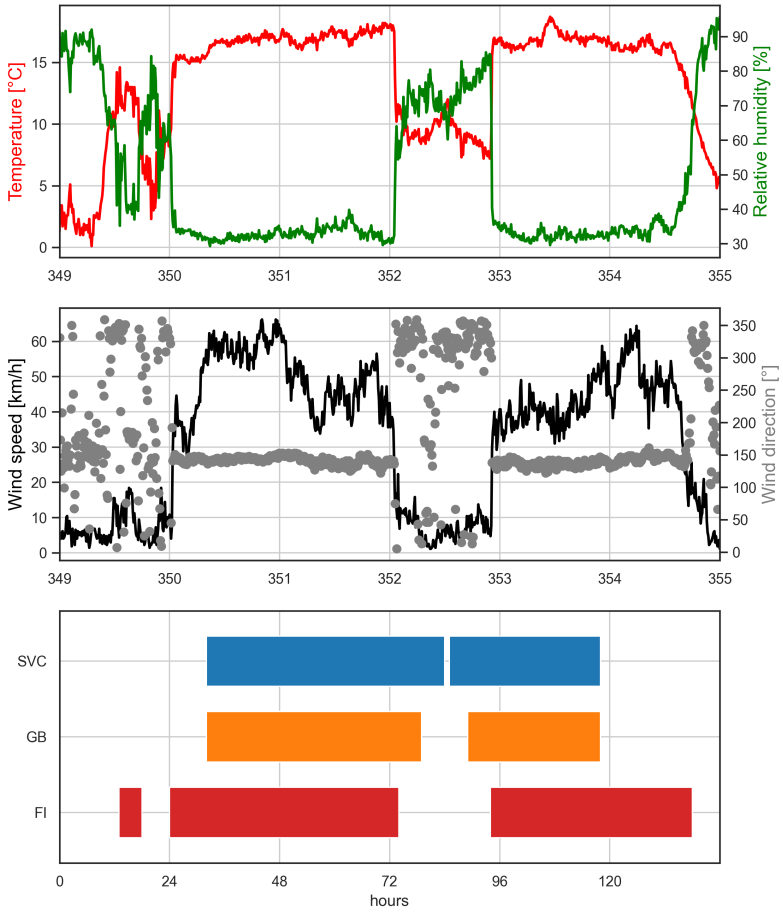

Figure 9. Time series of meteorological observations (top: temperature (red) and relative humidity (green), middle: wind speed (black) and direction (grey)). Bottom: classification results of SVC (blue) and GB (orange) and the reference FI index (red)

location of the stations and on how the difference is built. Most of the (ZWD difference) features shown in Figure 10 are built as a north-south station difference, which was expected to be a good predictor for foehn. Typically low ZWD values are observed at stations north of the Alpine ridge (such as FLDK here) and high ZWD values at stations south of the Alpine ridge (e.g. LOMO). The north gradient observed at station WEHO (GN_WEHO) is also consistent with the physical understanding, showing a southward (negative) trend for the observed foehn periods. The only feature which might not be intuitive is the ZWD difference between the stations ZIMM (near Bern) and HABG (located a bit north of the main Alpine rigde). Although the

extreme values are not as pronounced as for other predictors, they are still visible in the shown time series. This suggests that in the majority of cases, humid air reaches also the station HABG, which is located further north of the main Alpine ridge.

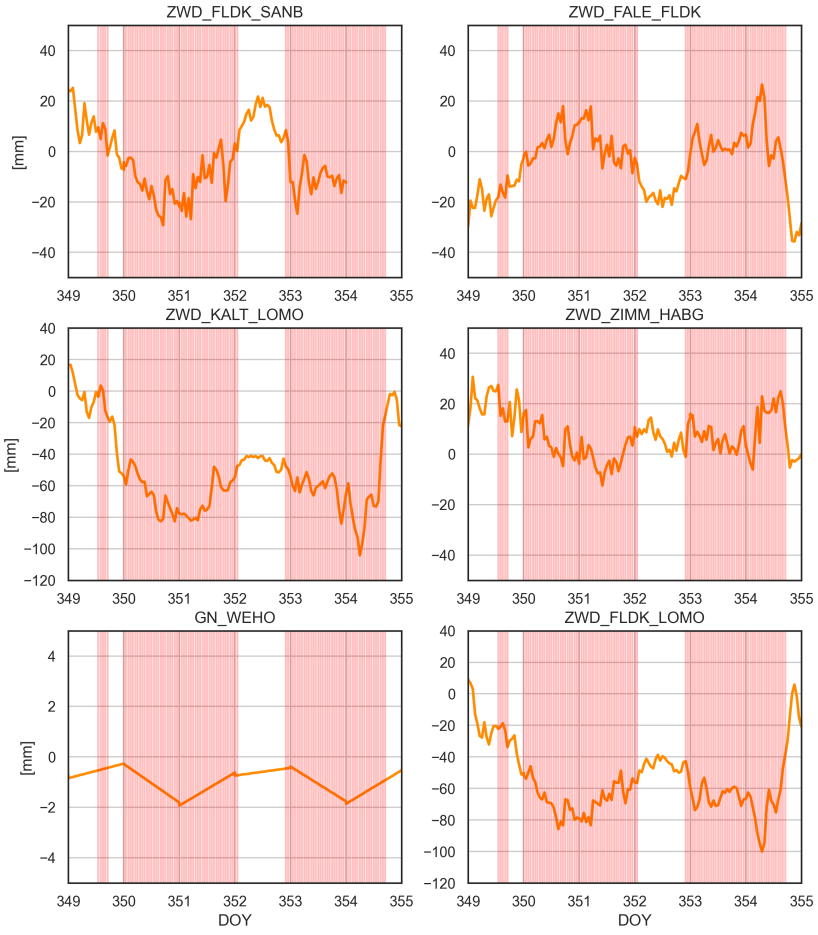

Figure 10. Time series of the six top predictors of the GB algorithm (most important features as shown in Figure 8) for the analysed time period. Observed foehn events (FI = 1) are color-coded in red.

425

8 Conclusions and outlook

In the present study, we introduced a new method for the detection and prediction of foehn events at the Swiss station Altdorf based on GNSS troposphere products and ML-based classification. We showed the performance of the introduced method by making use of an eleven-year-long data set of GNSS tropospheric parameters from a dedicated station network, namely the Swiss AGNES GNSS network as well as additional sites in neighbouring countries. Furthermore, we made use of FI observations at the SMN-station Altdorf for the chosen time period, provided by MeteoSwiss. In the course of an extensive cross-validation over the training data set (2010-2018), seven different classification algorithms were tested. The two best-performing algorithms were the GB and SVC algorithms, which were then used in the following case studies. In a first case study (CS1), we evaluated results of foehn classifications/predictions from those two algorithms over a two-year test period (2019-2020) at Altdorf. The second case study (CS2) investigated the usability of NRT GNSS products for foehn prediction, in order to assess the feasibility of these low-latency ($\sim$ 30-40 minutes) data for operational forecasting. In the third and last case study (CS3), we showed the importance of single, geographically relevant, GNSS stations such as high-altitude sites. Additionally, the performance of the method was tested in a detailed investigation of a weekly period including two strong foehn events in December 2019.

The following main conclusions can be drawn from the presented results:

- The introduced ML-based method using GNSS troposphere products can provide encouraging results. It achieves equal performance in terms of both detection-based (POD = 75-80%, POFD = 4%) and alarm-based (CAR = 66-76%, MAR = 5-6%) metrics. On average, the results of both utilized algorithms are comparable to those obtained by Sprenger et al. (2017), who used NWP data as well as observations from the validating measurement site (Altdorf), while our method does not make use of any meteorological data, at least not directly (see later point on independence).

- The most promising results can be obtained if the full station network (shown in Figure 2) can be utilized. This incorporates also stations from neighbouring countries (Austria, Italy, Germany).

- When using NRT troposphere products instead of reprocessed data, e.g. for operational prediction, some degradation of the results has to be accepted. This shows differently for the two used algorithms. For SVC, the performance loss mainly concerns alarm-based measures whereas for the GB algorithm both performance metric types are affected similarly. There are two apparent reasons for the experienced degradation related to the GNSS products. First, the NRT products do not include tropospheric gradients which show up in the top predictors for the other case studies, at least for dedicated stations. Second, it can be expected that the quality of the prediction results also is influenced by the quality of the troposphere products. For NRT products, this quality is lower due to lower-quality (but also lower latency) products used in GNSS processing (such as satellite orbits and clocks). In order to confirm this assumption, we plan to compare results using the NRT setup (using only delay-products) on both NRT and reprocessed products in a possible future study.

- The final decision on which algorithm (GB or SVC) to use is left to the actual users. Both algorithms show equal performance for combined measures (COMB and F_2), with GB providing a more balanced performance in terms of detection-

and alarm-based statistics. Another major advantage of the GB algorithm is the availability of feature importance information, which can provide valuable insight in the physical relations and meaningfulness of the results. Furthermore, its performance for the detailed event analysis presented in Section 7 was clearly better (significantly lower number of false alarms).

- As expected from the physical background of GNSS troposphere products, ZWD differences from geographically important stations serve as the most important predictors. Still, also other chosen parameter types such as north and east gradients as well as ZHD differences (representing pressure gradients) show up in the list of top features for all test cases. Furthermore, it is worth to note that not only stations in the vicinity of Altdorf, but also further away (e.g. the Valais area or the Rhine valley) play an important role. Therefore also stations from neighbouring countries can be crucial for a good performance (such as station Feldkirch (FLDK), Austria). Especially with the special setup of CS3, the importance of having additional data from high-altitude stations (SANB, OALP, ZERM) is outlined. Having the positive impacts of these stations on the result also allows for shortening the amount of data for training significantly (from nine to four years in this case).

- Choosing the optimal performance metrics and appropriate pre-processing is a key task in ML-based classification algorithms, especially when working with such a highly imbalanced data set as in this study. The actual choice for the most important metric(s) strongly depends on the actual application of the prediction method, deciding whether detection- or alarm-based measures should be preferred. Within this study we tried to tune the algorithms for an optimal balance between both metric-types and leave a possible decision to the potential users. However, as already outlined before, there exists a trade-off between POD and CAR and therefore an optimization towards one metric will always result in a shortcoming towards the other.

Overall, these initial results are promising and therefore the developed method might already aid the meteorological community as an additional tool for foehn detection and/or prediction. The biggest benefit of the method might be its (almost complete) independence of meteorological data input. With the exception of the ZHD calculation, for which surface pressure predictions from global NWP models are used, GNSS troposphere products are derived independent of meteorological data input. On the other hand, there is also a number of limitations for the introduced method:

- The most apparent limitation is that the availability of results from our method directly relies on the availability of GNSS data from all incorporated stations. As soon as data from only one station of the training network is missing, no results can be provided. This leads to a significant amount of periods for which no results can be produced, as seen e.g. in Section 7.

- Station specificity: This study only shows the applicability of the method for the FI station Altdorf. The performance achieved at this location cannot be generalized for other stations or a whole region. Tests for other meteorological stations that regularly experience foehn, both in Switzerland and neighbouring countries, are planned for future studies (see possible improvements below).

- Detection-based measures are not yet at the level of studies with meteorological data. However, alarm-based measures were higher for our study, so a combination of data sets is expected to be beneficial for the overall performance.

- Looking into specific foehn events, the introduced method shows weaknesses in terms of capturing the exact periods of foehn onset and decay, compared to the standard FI algorithm. As only a one-week period (including two major foehn events) was analysed here, this result cannot be generalized yet. Therefore, further studies are needed to look into more events in detail and, if needed, develop ways to increase the accuracy of the methods in terms of event start and end prediction.

Some of these limitations might be overcome by enhancements which can still be achieved through more detailed investigations in future studies. Possible improvements of the method we aim to investigate in the future would be:

- As stations from neighbouring countries are found to be important for the performance of the methods, extending the utilized station network for several more of them is expected to benefit the results. Especially having a denser station distribution on the southern side of the Alps (Italy) should have a positive impact on the performance of the methods.

- Introduce time lags of specific predictors, accounting for the response time of the actual parameter at a specific station to changes in the synoptic conditions indicating a foehn event or not. Pressure-related parameters (ZHD) might be applied a greater time lag (response in advance), compared to humidity/temperature-related parameters (such as ZWD). As same relation holds for the geographic location of contributing stations (south stations should response in advance). Such approaches might help to overcome the limitations in terms of exact onset/decay prediction.

- Use the developed method at other foehn locations in Switzerland and neighbouring countries. For some of those locations, the currently used setup might even provide better results than for Altdorf, based on a denser station network in those areas (e.g. Valais area).

- Enhance the nowcasting capabilities of the proposed method by including GNSS atmosphere gradients in the NRT products. If possible, gradient parameters should also be estimated at the same rate as delay products (every hour). Currently, this is done only every 6-12 hours and hourly estimates in between are interpolated linearly. This adaption might make the detection of smaller foehn events (lasting only a few hours) easier or even possible.

- Managing data gaps in GNSS time series: Interpolation techniques for missing data in GNSS troposphere time series (at least for shorter gaps) should be explored. Therefore one will need to find a balance between interpolation errors and the benefits of having a larger amount of training data.

- Optimization of the methods' performance by carrying out a more extensive grid search for hyperparameter tuning of the used algorithms or try new (possibly more sophisticated deep learning) algorithms.

- In the long term, the method might be used to generate an independent foehn climatology at Altdorf or other foehn locations in the future.

– Test the incorporation of GNSS products (especially from stations showing large impact in this study already) into algorithms based on meteorological data (as Sprenger et al. (2017)). As mentioned before, this might allow for a performance increase, especially for alarm-based statistics.

Code and data availability. Source code, trained algorithms and test data sets for running the case studies are provided with the manuscript.

Appendix A: List of GNSS stations

This section provides the full GNSS station network which is utilized for this study.

Table A1. Full list of GNSS stations utilized for this study. Indicated are Name, geographical location (Longitude, Latitude and Height), Network to which a station belongs to as well as the case studies in which it is used.

Name	Longitude [°]	Latitude [°]	Height [masl]	Network	Case studies
AIGE	6.128259	46.247775	473.8768	AGNES	All
BOU2	7.230437	47.394055	941.9955	AGNES	All
DAV2	9.843516	46.812917	1645.5747	AGNES	All
EPFL	6.567896	46.521467	460.4702	AGNES	All
ETHZ	8.510532	47.407070	594.8398	AGNES	All
FALE	9.230295	46.804491	1344.1583	AGNES	All
FLDK	9.580601	47.231347	570.3447	BEV	All
HABG	8.182777	46.747459	1147.8459	AGNES	All
HOH2	7.762704	46.319408	985.7388	AGNES	All
HOHT	7.762704	46.319408	985.7388	AGNES	All
HUTT	7.834883	47.141075	779.1001	AGNES	All
KALT	9.008414	47.217961	477.0371	AGNES	All
KOPS	10.115432	46.973951	1906.5541	BEV	All
KREU	9.160039	47.641294	529.9748	AGNES	All
LECH	10.139078	47.224056	1822.8005	BEV	All
LOMO	8.787428	46.172565	437.9931	AGNES	All
LUZE	8.300642	47.068204	542.2217	AGNES	All
MAR2	7.070694	46.122154	644.1085	AGNES	All
NEUC	6.940483	46.993829	504.6724	AGNES	All
OALP	8.673724	46.660062	2139.4693	AGNES	CS3
PAYE	6.943941	46.812141	548.7010	AGNES	All
PFA2	9.784663	47.515328	1090.0990	EUREF	All
SAAN	7.301290	46.515572	1419.5502	AGNES	All
SANB	9.184548	46.463831	1702.2351	AGNES	CS3
SCHA	8.655846	47.737566	638.1986	AGNES	All
SIGM	9.223912	48.083589	645.2889	SAPOS	All
STA2	8.941636	45.855855	417.2265	AGNES	All
STCX	6.501172	46.822386	1155.4289	AGNES	CS3
STGA	9.345949	47.441769	753.7296	AGNES	All
TORI	7.661280	45.063367	310.7408	EUREF	All
WEHO	7.472834	46.382054	2966.9314	COGEAR	All
ZERM	7.731996	46.001444	1931.1722	AGNES	CS3
ZIMM	7.465275	46.877097	956.3256	AGNES/EUREF/IGS	All

Author contributions. MAR developed the research idea, software implementation and prepared the manuscript (including formal analysis, visualization, and writing). EB produced and provided the long-term time series of GNSS troposphere products used for this research. LC prepared selected visualizations for the manuscript. All co-authors contributed to the discussion of results as well as reviewing and editing the manuscript.

Competing interests. The authors declare that no competing interests are present.

Acknowledgements. The authors would like to thank swisstopo for providing GNSS troposphere products from the AGNES network and MeteoSwiss for providing foehn index and other meteorological observations for the station Altdorf. Furthermore we thank the two anonymous reviewers for their valuable comments and suggestions which greatly helped to improve the manuscript.

References

Aichinger-Rosenberger, M.: Usability of high-resolution GNSS-ZTD data in the AROME model, Master's thesis, University of Innsbruck, https://diglib.uibk.ac.at/urn:nbn:at:at-ubi:1-27392, 2018.

Baeza-Yates, R. A. and Ribeiro-Neto, B.: Modern Information Retrieval, Addison-Wesley Longman Publishing Co., Inc., USA, 1999.

Barnes, L. R., Gruntfest, E. C., Hayden, M. H., Schultz, D. M., and Benight, C.: False Alarms and Close Calls: A Conceptual Model of Warning Accuracy, Weather and Forecasting, 22, 1140 – 1147, https://doi.org/10.1175/WAF1031.1, 2007.

Bennitt, G. V. and Jupp, A.: Operational Assimilation of GPS Zenith Total Delay Observations into the Met Office Numerical Weather Prediction Models, Monthly Weather Review, 140, 2706–2719, https://doi.org/10.1175/MWR-D-11-00156.1, 2012.

Bevis, M., Businger, S., Herring, T. A., Anthes, R. A., and Ware, R. H.: GPS Meteorology: Remote Sensing of Atmospheric Water Vapor Using the Global Positioning System, Geophys. Mag., 34, 359–425, 1992.

Breiman, L.: Random Forests, Machine Learning, 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Brenot, H., Neméghaire, J., Delobbe, L., Clerbaux, N., De Meutter, P., Deckmyn, A., Delcloo, A., Frappez, L., and Van Roozendael, M.: Preliminary signs of the initiation of deep convection by GNSS, Atmos. Chem. Phys., 13, 5425—-5449, 2013.

Brockmann, E.: Reprocessed GNSS tropospheric products at swisstopo, in: GNSS4SWEC workshop, Thessaloniki, May 11-14, 2015.

Brockmann, E.: National Report of Switzerland, in: EUREF-Symposium, San Sebastian, Spain, May 25-27, 2016.

Brockmann, E., Grünig, S., Schneider, D., Wiget, A., and Wild, U.: National report of Switzerland: Introduction and first applications of a Real-Time Precise Positioning Service using the Swiss Permanent Network AGNES, In Torres, J.A. and H. Hornik (Eds): Subcommission for the European Reference Frame (EUREF). EUREF Publication No. 10. Mitteilungen des Bundesamtes für Kartographie und Geodäsie, 23, 272–276, 2002.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P.: SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research, 16, 321–357, https://doi.org/10.1613/jair.953, 2002.

Cover, T. and Hart, P.: Nearest neighbor pattern classification, IEEE Transactions on Information Theory, 13, 21–27, https://doi.org/10.1109/TIT.1967.1053964, 1967.

de Haan, S.: Meteorological applications of a surface network of Global Positioning System receivers, Ph.D. thesis, Wageningen University, 2008.

de Haan, S.: Assimilation of GNSS ZTD and radar radial velocity for the benefit of very-short-range regional weather forecasts, Q. J. R. Meteorol. Soc., 139, 2097–2107, 2013.

Deloncle, A., Berk, R., D'Andrea, F., and Ghil, M.: Weather Regime Prediction Using Statistical Learning, Journal of The Atmospheric Sciences - J ATMOS SCI, 64, 1619–1635, https://doi.org/10.1175/JAS3918.1, 2007.

Drobinski, P., Steinacker, R., Richner, H., Baumann-stanzer, K., Beffrey, G., Bénech, B., Berger, H., Chimani, B., Dabas, A., Dorninger, M., Dürr, B., Flamant, C., Frioud, M., Furger, M., Gröhn, I., Gubser, S., Gutermann, T., Häberli, C., Häller-Scharnhost, E., Jaubert, G., Lothon, M., Mitev, V., Pechinger, U., Piringer, M., Ratheiser, M., Ruffieux, D., Seiz, G., Spatzierer, M., Tschannett, S., Vogt, S., Werner, R., and Zängl, G.: Föhn in the Rhine Valley during MAP: A review of its multiscale dynamics in complex valley geometry, Quarterly Journal of the Royal Meteorological Society, 133, 897–916, 2007.

Dürr, B.: Automatisiertes Verfahren zur Bestimmung von Föhn in Alpentälern, Arbeitsberichte der MeteoSchweiz, 223, 22 pp, 2008.

Freund, Y. and Schapire, R. E.: A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting, Journal of Computer and System Sciences, 55, 119–139, https://doi.org/https://doi.org/10.1006/jcss.1997.1504, 1997.

Friedman, J. H.: Greedy Function Approximation: A Gradient Boosting Machine, The Annals of Statistics, 29, 1189–1232, http://www.jstor.org/stable/2699986, 2001.

Glahn, H. R. and Lowry, D. A.: The Use of Model Output Statistics (MOS) in Objective Weather Forecasting, Journal of Applied Meteorology and Climatology, 11, 1203 – 1211, https://doi.org/10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2, 1972.

Gohm, A. and Mayr, G. J.: Hydraulic aspects of föhn winds in an Alpine valley, Quarterly Journal of the Royal Meteorological Society, 130, 449–480, https://doi.org/https://doi.org/10.1256/qj.03.28, 2004.

Guerova, G., Jones, J., Dousa, J., Dick, G., Haan, S., Pottiaux, E., Bock, O., Pacione, R., Elgered, G., Vedel, H., and Bender, M.: Review of the state-of-the-art and future prospects of the ground-based GNSS meteorology in Europe, Atmospheric Measurement Techniques Discussions, 9, https://doi.org/10.5194/amt-9-5385-2016, 2016.

Gutermann, T., Dürr, B., Richner, H., and Bader, S.: Föhnklimatologie Altdorf: die lange Reihe (1864-2008) und ihre Weiterführung, Vergleich mit anderen Stationen, https://doi.org/10.3929/ethz-a-007583529, 2012.

Hess, R.: Statistical postprocessing of ensemble forecasts for severe weather at Deutscher Wetterdienst, Nonlinear Processes in Geophysics, 27, 473–487, https://doi.org/10.5194/npg-27-473-2020, 2020.

Hsieh, W. W.: Machine Learning Methods in the Environmental Sciences: Neural Networks and Kernels, Cambridge University Press, https://doi.org/10.1017/CBO9780511627217, 2009.

Kretzschmar, R., Eckert, P., Cattani, D., and Eggimann, F.: Neural Network Classifiers for Local Wind Prediction, Journal of Applied Meteorology - J APPL METEOROL, 43, 727–738, https://doi.org/10.1175/2057.1, 2004.

LeCun, Y. A., Bottou, L., Orr, G. B., and Müller, K.-R.: Efficient backprop, in: Neural networks: Tricks of the trade, pp. 9–48, Springer, 2012.

Lothon, M., Druilhet, A., Bénech, B., Campistron, B., Bernard, S., and Said, F.: Experimental study of five föhn events during the Mesoscale Alpine Programme: From synoptic scale to turbulence, Quarterly Journal of the Royal Meteorological Society, 129, 2171 – 2193, https://doi.org/10.1256/qj.02.30, 2006.

Manzato, A.: The Use of Sounding-Derived Indices for a Neural Network Short-Term Thunderstorm Forecast, Weather and Forecasting, 20, 896 – 917, https://doi.org/10.1175/WAF898.1, 2005.

Mayr, G., Plavcan, D., Armi, L., Elvidge, A., Grisogono, B., Horvath, K., Jackson, P., Neururer, A., Seibert, P., Steenburgh, J. W., Stiperski, I., Sturman, A., Željko Večenaj, Vergeiner, J., Vosper, S., and Zängl, G.: The Community Foehn Classification Experiment, Bulletin of the American Meteorological Society, 99, 2229 – 2235, https://doi.org/10.1175/BAMS-D-17-0200.1, 2018.

Mayr, G. J., Armi, L., Gohm, A., Zängl, G., Durran, D. R., Flamant, C., Gabersek, S., Mobbs, S. D., Ross, A. B., and Weissmann, M.: Gap flows: Results from the Mesoscale Alpine Programme, Quarterly Journal of the Royal Meteorological Society, 133, 881–896, 2007.

Mony, C., Jansing, L., and Sprenger, M.: Evaluating Foehn Occurrence in a Changing Climate Based on Reanalysis and Climate Model Data Using Machine Learning, Weather and Forecasting, 36, 2039 – 2055, https://doi.org/10.1175/WAF-D-21-0036.1, 2021.

Otero, F. and Araneo, D.: Zonda wind classification using machine learning algorithms, International Journal of Climatology, 41, E342–E353, https://doi.org/https://doi.org/10.1002/joc.6688, 2021.

Pacione, R., Araszkiewicz, A., Brockmann, E., and Dousa, J.: EPN-Repro2: A reference GNSS tropospheric data set over Europe, Atmospheric Measurement Techniques, 10, 1689–1705, https://doi.org/10.5194/amt-10-1689-2017, 2017.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825–2830, 2011.

Perler, D. and Marchand, O.: A Study in Weather Model Output Postprocessing: Using the Boosting Method for Thunderstorm Detection, Weather and Forecasting, 24, https://doi.org/10.1175/2008WAF2007047.1, 2009.

Platt, J. C.: Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods, in: Advances In Large Margin Classifiers, pp. 61–74, MIT Press, 1999.

Plavcan, D., Mayr, G. J., and Zeileis, A.: Automatic and Probabilistic Foehn Diagnosis with a Statistical Mixture Model, Journal of Applied Meteorology and Climatology, 53, 652 – 659, https://doi.org/10.1175/JAMC-D-13-0267.1, 2014.

Rüeger, J.: Refractive Index Formulae for Radio Waves, Proc. FIG XXII International Congress, Washington, D. C., 2002.

Saastamoinen, J.: Contributions to the theory of atmospheric refraction., Bull. Geodesique, 105, 13–34, 1972.

Scher, S. and Messori, G.: Predicting weather forecast uncertainty with machine learning, Quarterly Journal of the Royal Meteorological Society, 144, 2830–2841, https://doi.org/https://doi.org/10.1002/qj.3410, 2018.

Solheim, F. S., Vivekanandan, J., Ware, R. H., and Rocken, C.: Propagation delays induced in GPS signals by dry air, water vapor, hydrometeors, and other particulates, Journal of Geophysical Research: Atmospheres, 104, 9663–9670, https://doi.org/https://doi.org/10.1029/1999JD900095, 1999.

Sprenger, M., Schemm, S., Oechslin, R., and Jenkner, J.: Nowcasting Foehn Wind Events Using the AdaBoost Machine Learning Algorithm, Weather and Forecasting, 32, 1079 – 1099, https://doi.org/10.1175/WAF-D-16-0208.1, 2017.

Steinacker, R.: Alpiner Foehn - a new verse to an old song, Promet, 32, 3–10, 2006.

Stoev, K. and Guerova, G.: Use of GNSS tropospheric products to study the foehn in Sofia, Talk: EMS Annual Meeting: European Conference for Applied Meteorology and Climatology 2018 | 3–7 September 2018 | Budapest, Hungary, 2018.

Stoycheva, A. and Guerova, G.: Study of fog in Bulgaria by using the GNSS tropospheric products and large scale dynamic analysis, Journal of Atmospheric and Solar-Terrestrial Physics, 133, 87–97, https://doi.org/https://doi.org/10.1016/j.jastp.2015.08.004, 2015.

Wilhelm, M.: COSMO-2 Model Performance in Forecasting Foehn: a Systematic Process-oriented Verification, Veröffentlichungen der MeteoSchweiz, 89, 64pp, 2012.

Wilks, D. S. and Hamill, T. M.: Comparison of Ensemble-MOS Methods Using GFS Reforecasts, Monthly Weather Review, 135, 2379 – 2390, https://doi.org/10.1175/MWR3402.1, 2007.

WMO: International Meteorological Vocabulary. WMO No. 182, 1992.

Yalavarthi, R. and Shashi, M.: Atmospheric Temperature Prediction using Support Vector Machines, International Journal of Computer Theory and Engineering, 1, 55–58, https://doi.org/10.7763/IJCTE.2009.V1.9, 2009.

Yan, X., Ducrocq, V., Jaubert, G., Brousseau, P., Poli, P., Champollion, C., Flamant, C., and Boniface, K.: The benefit of GPS zenith delay assimilation to high-resolution quantitative precipitation forecast. A case-study from COPS IOP 9., Q. J. R. Meteorol. Soc., 135, 1788—-1800, 2009.