Answers to Comments Referee 1

General response

We want to thank referee one for the second extensive review and advice on some points where the manuscript still could/should be improved/corrected. Comments B1 and B5 were very valuable for the last corrections and extensions. We tried to incorporate most of the other comments as well.

We also want to clearly state here that missing some corrections was not done intentionally and the majority of comments on the first version has been implemented.

In the latest version, one figure combining the dropped plots (now Figure 6) was re-introduced to allow for a cross-check. This one includes all foehn events recorded by MeteoSwiss and our predictions (only events where all GNSS data was available) for all four experiments.

In the following, we will go through all specific comments brought up.

Specific comments

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 Consequently, the manuscript should not be published unless the correctness of the computations can be convincingly shown.***

   Unfortunately, the scores of the cross-validation have not been updated in the last version: We thank you for cross-checking, this needed to be corrected. The new scores can now be found in the latest version of the manuscript. However, the general outcome has not changed, and the Gradient-Boosting and Support-Vector-Classifier algorithms are still the best performing in terms of combined scores (COMB and F2) and therefore chosen for the feature setup experiments as we now call them (see point 4). Furthermore, these results did never have any impact on the general conclusion of the paper, which is drawn from the feature setup experiments.

Moreover, we still show observed and diagnosed events, see Figure 9 in the old version (comparison with station data from Altdorf for December 2019). However, as this only represents ~one week of data, we decided to re-introduce comparison plots for diagnosed vs observed events for the experiments. They are now combined into Figure 6 (in the newest version). This hopefully should erase last doubts about the results.

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

   We have tried to point this out even more than in the newest version (lines 81-87). How to tackle this problem adequately without supervision is very uncertain to us. Unsupervised learning can cluster/group the data in order to discover something that is not visible otherwise. Considering the highly imbalanced data set and the multitude of phenomena affecting tropospheric delays, it is doubtful that a clear cluster would emerge solely corresponding to foehn events, not to mention foehn at a specific location.

   Therefore, it was an obvious choice to apply supervised learning with a reference label (FI index from Dürr (2008)), whose quality has been proven by comparisons to human forecasts (as shown in Dürr (2008)). It is important to note that supervised machine learning is commonly applied to problems with imperfect labels.

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

3. ***Dependence on foehn classified with traditional meteorological measurements. Despite the claim of the paper (line 445 and lines 521–522) the foehn classification with GNSS data is completely dependent on meteorological data, since these were used to compute. foehn index FI that the supervised learning methods in the paper use as truth (response variable). It will therefore not be possible to compute an independent foehn climatology at Altdorf (lines 521-522) and the quality of the classification will by design always be poorer than the one of the foehn index FI.***

   While it is true the method is not fully independent of meteorological data (as not even the estimation of GNSS parameters is completely), the dependence is limited to meteorological data of the training period. Thus, our algorithm(s) is/are dependent exactly on meteorological data from 2010-2018 or 2015-2018, depending on the actual experiment. However, it is not dependent on contemporaneous meteorological data, that is what we would call “completely dependent”. Still, we relativized/excluded the benefits of “almost/near-independence” in the latest version, compared to the last one, as it should be clear to potential readers anyway which data is needed to obtain results.

Furthermore, it might not be possible to compute a fully independent climatology for Altdorf, but for a e.g. 30-year time period excluding 2010/2015-2018, the dependence will be very small. Still, as this might also not be a main use case in the future, we excluded this point from the discussion/conclusion.

In general, we think the fact that for this method (and any supervised learning method) the performance will always be weaker than the “truth” method (here FI index) in direct comparison should actually be quite clear to potential readers of the journal without further explanation. Still, we have added this fact as a point to the limitations/conclusions.

It was also never expected that GNSS measurements will perform better than measures based on actual meteorological data. It is already quite satisfying that the agreement is relatively high.

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

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

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

   a) We tried to combine some of these parts in the new version. Sections 3, 4 and 5 have been combined and rearranged. Section 3 (Methodology as a whole) now starts with datasets, then introduces the default feature setup and finally gives information on the ML algorithms tested and how the two, which were further used, are chosen.

   b) “Case studies” was changed “feature setups”, we hope that this suits better.

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

Thank you for bringing this point up. It is true that the results of FS1/2 and FS3 (or CS in old version) are not comparable 1:1. We therefore added the requested fourth setup (FS4), which contains the shorter period (2015-2019) but no additional stations (so the station setup of FS1 and FS2). The new results confirm the old ones, i.e. the additional stations bring benefits for the method (see feature importances). However, the decreased number of covered events when new stations are missing must be considered. We also note that 2019 had a higher foehn probability than e.g. 2020, which also explains (some of) the relatively better performance of FS4 compared to FS1.

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

   The large fraction of missing data is for sure the biggest problem of the current version of our method. We have stated this several times and provided detailed statistics on the missing data in the manuscript. As mentioned by you, it is also stated in the discussion of disadvantages/future improvements. We have now added quantitative measures such as percent points of missing events for each FS (in tables describing the setups).

   As the aim of this study is a proof-of-concept, it was not the intention to present a method which is ready for operational usage. The large fraction of missing data is for sure the most urgent problem to be dealt with. Nevertheless, if this can be solved, one could even guess (we leave it at this word) that without these missing periods, results might even be better as also more foehn events would show up in the training data.

   The issue will anyhow be investigated in further studies, where different solutions for the problem can be tested. Thank you for your proposal of using alternate features, this might be a good way to deal with it. Another easy (but computationally more expensive) possibility is to train a set of different algorithms, iteratively excluding GNSS data from one station in the feature setup each time. This gives (at least) the opportunity to provide a continuous solution as long as only data from one station is missing. We have added both strategies to the outlook section.

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

   We have added this statement in the new version (lines 246 and 247).

   The sentence (line 250) on the performance measures has been corrected to “see e.g. Barnes et. al (2007)”.

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

   We changed "percent" to "percent points". The relative difference of 21% and 15% might be more substantial, but the result is still quite solid in our opinion.

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

   We are very sorry to have missed this issue which was due to the fact that the plots have been dropped and the response was not properly updated. However, we also want to clearly state here that all this was definitely not done intentionally.

   We have re-introduced the content of Figures 5, 9 and 10 and combined them for the new setups in one figure (Figure 6).

   We have also added the lines to the map in the newest version.

Comments C:

1.) The confusion matrices of all experiments can now be found in Table 8.
2.) The colorbar (legend) was added to the station map.

Answers to Comments Referee 2

General response

Thank you for the revised manuscript, I think it is greatly improved, also regarding the plots. I just have some few final remarks. Let me start with addressing some of the older comments.

We want to thank the referee for his/her kind words and comments which have significantly contributed to improve the manuscript. The remaining comments and remarks are answered in the following:

1. ***Eq.1: I asked the authors in the previous review to change the formula from Rüger and now, it is correct (but the authors asked me what was wrong before):The first two terms were 77.68P/T + 6.3938 e/T When it should be 77.68P/T - 6.3938 e/T (minus instead of plus), because if we substitute P=Pd+e, then we get: 77.68(Pd+e)/T - 6.3938 e/T=77.68Pd/T + 71.2952e/T (the original formulation from Rüger).***
   Thank you very much for the clarification.

2. ***"6. Section 4.1: I would recommend giving here at least very brief overview of the selected methods Answer: As the manuscript is quite long already, we actually would like to leave the reader here with the references given describing the methods" -> In my opinion the manuscript is not dramatically long and a brief overview would really help the readers that are not familiar with all the ML methods to get a grasp of the methods the authors are using. I would suggest to at least write something about your two chosen methods, GB and SVC.***

   We have now tried to combine this issue with point number 4, providing a short description of the two used algorithms (GB and SVC) as well as the tuned hyperparameters of them.

3. ***Line 163: '(negative) maximum' - > why not use 'minimum' here? Answer: We would interpret 'minimum' as close to zero but one can for sure argue to use minimum here as well" -> just a clarification, in mathematics the local and global minima (and as well a maxima) of a function do not have anything to do with being close to zero but they are simply the largest and smallest values of a function. In here, (for the differences) you have two obvious local minima (out of which one is also the global minimum)***
   Thanks for the clarification, we have changed this.

4. ***Line 259: Would be good to comment here what the chosen parameters mean. Answer: Actually, this is done in the caption of (what is now) Table 3" -> I still think it is a good idea to put also an explanation in the main text.***

   Was done in combination with point 2, see above.

Some new comments:

1. ***Line 130: '1999-2020' – later in the text you state that you only use the data from 2010 or even 2015, so what is a point of mentioning the data from 1999-2010? I***

understand that it is to point out you have these longer time series but it is confusing for the reader.
We have changed it to 2010-2020, thanks for pointing this out.

2. ***Figure 2: I would make the coordinate font smaller and the stations font larger.***
   The station font was made larger and the coordinate font smaller to improve visibility.

3. ***Line 323: Please put dot before 'Figure'. In general, I would suggest to check the punctuation, there are some commas missing and some too many (e.g. Line 327: 'Furthermore it can be seen, that' -> 'Furthermore, it can be seen that' etc.)***
   We have gone through the manuscript once again and updated/corrected everything we found.

4. ***Line 375, 378: remove 'again' – you do not show the same again, these are new plots***

   Requested by referee one, we decided to drop the plots of the confusion matrices and instead give their entries summarized in Table 10. Therefore, also these lines have been dropped/modified.

Answers to Comments Referee Report 3

General response

We want to thank referee three for the comments on how to further improve it. All specific comments are addressed below.

Specific comments

1. ***$mfh(e)$、 $mfw(e)$、 $mfg(e)$ should be el in the formula 2, e stands for water vapor***

Thanks for having a close look here. We have updated the formula.

2. ***Please describe clearly why beta is set to 2 in formula 8.***

Beta is a real number factor which determines the weighting of recall vs. precision (or vice versa) in the F-score. As already mentioned in the manuscript, for our (highly imbalanced) classification problem recall is a better representation of accuracy than precision. This is because optimal performance in terms of precision would/could result in the best performing algorithm never predicting foehn, as it is such a rare phenomenon. Thus, we use the F2 score (beta = 2) which weights the recall parameter two times larger than precision. We added some more description to the manuscript.

3. ***Please label the abscissa of the upper two graphs in Figure 9***

We have added the appropriate label DOY to the figure (now Figure 7).

4. ***CS1 in label of table 1 should be full named***

The label has been updated as requested, but keep in mind that (due to the request of referee 1) it now states FS1, for "Feature setup".

5. ***The training period and test period of CS3 are different from CS 1 and CS 2 does this affect the evaluation of the model?***

The short answer to this is simply yes, it does. Choosing e.g. 2020 as the test period will yield different results to a certain extent. However, our main intention for this setup was to still apply the 80/20 training/test data rule-of-thumb as also done for the other case studies/experiments. We added a fourth feature set (station setup from CS1 and 2 for the reduced period) to achieve a fair comparison with the extended station setup. This was requested by referee one.

6. ***From line 404 to 405 please cite some papers or explanations that illustrate that GNSS parameters have slightly longer response times to o a change in synoptic conditions***

What was meant here was actually the fact that different GNSS stations (and their absolute values) might experience a change in synoptic conditions later (or actually sooner depending on their location). Therefore, this was not formulated correctly and updated in the newest version. We are sorry for confusing here.

However, also the temporal resolution, which the tropospheric parameters estimated, might also play a role in cases of rapid changes in synoptic conditions. As correlation of the tropospheric parameters with other station parameters requires longer observation periods, e.g. to separate the zenith total delay from the height component or the east-west gradient from longitude, a stacking period of 15 min is recommended for tropospheric parameter estimation [Dach et. al, 2015]. In recent works, the temporal resolution is further reduced to 1 min [e.g., Hadas and Hobinger, 2021] using a Kalman filter approach. Therefore, additional relative constrains are introduced leading to slightly longer response times in case of rapid changes in synoptic conditions.

Still, as we have hourly values of both troposphere estimates and foehn index, this might not be a huge problem for our results.

Dach, R., S. Lutz, P. Walser, P. Fridez (Eds); 2015: Bernese GNSS Software Version 5.2. User manual, Astronomical Institute, University of Bern, Bern Open Publishing. DOI: 10.7892/boris.72297; ISBN: 978-3-906813-05-9.

Hadas T., Hobiger T.: Benefits of Using Galileo for Real-Time GNSS Meteorology. IEEE Geoscience and Remote Sensing Letters, Vol. No. , Piscataway, NJ, USA 2020, pp. 1-5. DOI: 10.1109/LGRS.2020.3007138

7. ***In line 515, whether to consider using other observations to supplement missing tests instead of linear interpolation?***

This depends on what is meant with “other observations”.

Using additional meteorological measurements would most likely help to improve the performance of the method. However, this is not our intention as the goal is to rely only on GNSS-based observations.

Using observations from different GNSS stations (which are currently not part of the feature setup) for filling gaps should be possible but takes a more sophisticated setup of the ML algorithm and a more detailed analysis of their impact. This might be a topic addressed in a follow-up study. Another possibility which we might also investigate in the future (and also added to the conclusions/outlook section) is to build a larger set of algorithms, each one excluding one GNSS station from the feature setup. This could provide a continuous solution as long as not more than one station is missing.

Author's changes in manuscript (from revised version)

Main changes:

- **Updated scores of cross-validation/algorithm selection**:
  We changed the scores of the cross-validation to those achieved by the new study setup. These changes were unfortunately missed to update in the revised version. Although the changes are minor and not influencing the following results and conclusion, we thank referee one for pointing towards this issue.
  Furthermore, there are small changes in some of the statistics of the three FS experiments due to:
    1.) A bug that was encountered in the calculation of POFD and MAR, which has now been corrected for all experiments. Therefore, these values (slightly) changed for all results shown.
    2.) The fact that we re-ran all the experiments using the newest sklearn-version (1.1.2), in order to cross-check if there are any changes when using this version. Slight changes in some cases occur, which are now updated in the manuscript and in the model code/software provided. Thus, the 1.1.2 version of sklearn should be used when using the provided code.

  However, these changes have not influenced significantly the conclusions of the study.

- **Addition of a fourth experiment**:
  As requested by referee one, we added a fourth experiment (or feature setup) to enable a better comparison with FS3, which uses a different training/test setup (2015-2018/2019) compared to the first two experiments (2010-2018/2019-2020).
  Feature setup 4 (FS4) now contains the station setup of FS1 and FS2 but the training/test split of FS3. Therefore, the comparison between FS3 and FS4 now directly addresses the impact of additional stations on the results.
- **Re-introduction of prediction time series plots for comparison**:
  As requested by referee one, we re-introduced plots to compare results from the GB/SVC algorithms with observed events (FI from MeteoSwiss) over the whole test period for the different experiments. All four of those lots have been combined into what now is Figure 6. As stated in the caption of Figure 6, please note that this now includes all observed events, so also those were no GNSS data is available and thus no prediction could be made.
- **“Case study” has been changed to “Feature setup”**:
  As requested by referee one, we changed “case study” to “feature setup”.
- **Sections combined**:
  As requested by referee one, we combined former sections 3, 4 and 5 under “Methodology”. These parts have also been slightly rearranged. The section now starts with data sets, then introduces the default feature/station setup and finally presents the ML algorithm cross-validation and descriptions of the two chosen algorithms.

Some of the minor changes:

- Short descriptions of the Gradient-Boosting and Support-Vector-Classifier algorithms were added (as requested by referee two)
- Lines connecting the five top predictors (ZWD differences) were added to Figure 2 as requested by referee one. We are sorry to have missed this issue in the first revision. Furthermore, the station font was increased as suggested by referee two and a colorbar (what we would interpret as a legend here) was added.