Answers to Comments Referee 1

General response

First, we want to thank referee one for this extensive review and a lot of valuable comments and suggestions on how to improve the manuscript. In the following, we will go through all specific comments and give some details how we implemented most of the referees' points to improve the manuscript.

Specific comments

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

Thanks for cross-checking the foehn event data, this is a very valuable and the most obvious point for improvement, which we have addressed. We have improved the corresponding code and included detailed statistics on the overall number of foehn events as well as events which had to be excluded due to lack of GNSS data. Essentially, we have now defined criteria for the selection of stations and the study period to make sure we do not lose the majority of foehn events due to lack of GNSS data (one station missing is enough to not produce any result). The missing response dates are properly excluded.

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

Code and test data set will be made available with the revised manuscript in form of an online repository, in order to make the results reproducible. We have also tried to give more details on the algorithms used in the revised version of the manuscript, although plenty thereof was already given in section about hyperparamter tuning in the old version. The software package is the 1.0.1 version of sklearn. The number of covariates was added to the study setup tables.

However, we would also like to mention that comparable studies (such as Sprenger et. al (2017)) also did not publish any additional material (neither code nor data), therefore we still see the upload of additional material as optional.

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

The original idea was to test several algorithms of different types and find out which ones work best for this task. Those algorithms (which are tested) were also described in the introduction, indicating that these have already been used (with success) in similar studies. Therefore, we chose them for the cross-validation. For sure there are a lot more possibilities to choose from but at some point, a decision for one/several algorithms had to be made. However, it is reasonable to exclude decision trees here, which we did for the revised manuscript.

Concerning Support Vector Classifiers/Machines: The implementation of SVC in sklearn can also handle non-linear boundaries. Therefore, it takes the keyword “kernel” which can be set to e.g. “polynomial” but also “linear” in case one wants linear boundaries. The default value would be RBF (Radial Basis Function), which was set in our analyses.

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

The main goal of the study is not finding the ultimate best performing algorithm for usage of GNSS data to detect/predict foehn events but rather showing that this is possible at all using machine-learning based classification (proof-of-concept). Although we already do quite a bit of fine tuning for the two selected algorithms, applying an extensive grid-search for one specific algorithm would be an interesting investigation of a follow-up study. We realized that therefore we need to adjust some formulations in the manuscript (such as the title even), which might be too promising for potential readers.

As mentioned in the last point, we wanted to start out this initial study with a cross-comparison of several algorithms to find potential candidates for further optimization for dedicated case studies. Therefore, we chose the default settings of all algorithms (i.e. their sklearn implementation) in the cross-comparison as this seems to secure the most independent assessment possible. We are still convinced that this a reasonable way to approach this problem, since no prior experience on which algorithm (and which settings) will work best is available (when using GNSS data for training). Thus, using default settings (which work best for a majority of classification problems) should be the most objective choice in our opinion. For sure some (not chosen) algorithms might outperform the ones proposed here after extensive tuning. However, at some point one will also face time constraints as this potentially becomes a (possibly) never-ending optimization of settings.

For these reasons, we have sticked to our initial approach but putting much more emphasis on the performance of the two chosen algorithms in the case studies.

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

This is certainly a very valuable point, which also was investigated further. However, we feel that we already have addressed the basic physical understanding through showing the response of ZWD values (and differences between them for station north/south of the Alpine ridge) in Figure 2 of the initial manuscript. That is where the basic idea for this study comes from. Furthermore, the new figures added which show the feature importance of the GB algorithm already show the importance of certain stations or differences between them (such as FLDK and SANB or KALT and LOMO) and indicates that these differences have the highest impact on the classification. We have included additional plots of the top predictors for selected foehn events which show the physical connection and enhance physical understanding.

ZWD maps: Although we have some means to do so at hand, ZWD maps are not trivial to produce. Furthermore, the accuracy of such maps is limited by the density of the station network because of interpolation errors (especially with height). Switzerland might has one of the densest networks in Europe, but the horizontal resolution is still not comparable to a NWP grid or the SwissMetNet station network. Therefore, we did not include ZWD maps.

6. ***Ultimate reason for the method: Why should foehn be diagnosed from GNSS-derived information? Weather station data give a more reliable answer for specific locations and such information was actually used as truth to approximate with GNSS data and machine learning algorithms. The method***

described in the paper cannot be used to diagnose foehn in locations without weather stations either, since it was trained on only one station and the transferability to other locations is not shown in the paper. The paper uses the approach for nowcasting 1 hour into the future and mentions that NWP models fare poorly with foehn quoting a paper from 2012. NWP models and their spatial resolution have dramatically improved in the decade since then. I would guess that MeteoSwiss has a current performance evaluation of COSMO1 available for Altdorf, against which the results of the paper could be measured. Results should also be compared to a simple persistence model, i.e. nowcasting the same no/foehn state as in the current hour.

As already mentioned before, the aim of this study was not to give an ultimate, stand-alone method for foehn detection/prediction but rather showing that the method introduced denotes an additional tool that might be helpful to achieve those tasks. The main aim is still a proof-of-concept for a completely new approach. Therefore, we scaled back some of the promises given in the manuscript (e.g. in the title). Also, this study (just as Sprenger et. al (2017)) focuses on detecting/predicting foehn events at the specific location of Altdorf. Therefore, we have added this information to the title. In a further study we plan to extend the investigations to other foehn locations (in Switzerland or also Austria), which for sure will require dedicated training of the algorithms at the respective stations. Therefore, the method will always be somehow specific for a certain location. Concerning NWP performance and evaluation, we have not got any update on the performance and not found any new studies on it. Therefore, it seems reasonable to us to quote the latest one available, even when already from 2012, just as our reference study did.

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

It was the original plan to use all available data (1999-2020) for this study. However, several stations which are important for the method (as seen in the feature importance plot) are not available since 1999. Therefore, one will lose those products/stations as features/predictors for the entire classification as all used features have to be available for the whole training and test period. We have now defined dedicated criteria for the selection of GNSS stations and study periods, based on statistics about covered foehn events.

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

This comparison is already done/included when computing the performance measures but including overall values as a separate statistic is also possible. Specific stratification by season is not shown since for the two-year test period these statistics might not be robust, as the number of foehn events covered is already quite low in general. For a possible follow-up study (with a larger test data set) we might investigate the seasonal performance of the method. As suggested, we have also added a section showing such a week-long time series plot, also including meteorological data. Most of the top 6 predictors show expected behavior for the observed foehn periods. We hope this increases the physical understanding of the approach.

Less crucial points

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

We incorporated a section on this topic.

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

Performance metrics: The section was shortened, and the confusion matrix is given for the case studies.

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

A few stations from Italy are part of the available data set (as well as some Austrian and German stations), those are now part of the training and test data. For most other stations in Northern Italy, we don’t have data available (yet). The inclusion of more stations from different regions is also planned for a follow-up study.

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

This point has been addressed in the revised manuscript. *Plavcan et al., 2014* was added.

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

Moving-average filtering is applied to reduce noise present in the ZWD estimates. The moving average is centered and there is no particular reason for choosing a 12-hour window beside optimized performance in noise reduction (for visualization). The ZWD estimates used were originally also filtered, which also explains “lost” foehn events at the very end of the test period (last hours of 31.12.2020). However, tests reveal comparable performance using unfiltered estimates. Therefore, we present results using unfiltered ZWD data in the revised manuscript, in order to not lose additional foehn events.

The “hourly” GNSS products are estimated every hour from 30-second measurements together with station coordinates (in this case using least-squares adjustment). We added this information to the manuscript.

Technical corrections

These have been included in the revised version.

Answers to Comments Referee 2

General response

The paper „Prediction of Alpine Foehn from time series of GNSS troposphere products using machine learning" shows first the selection of the ML methods and then usage of two of them on a GNSS tropospheric data set (tropospheric delays and gradients) to detect the foehn occurrences. It is a very new field of study as most of the GNSS meteorology research focuses on the precipitation/humidity parameters rather as foehn. Also the usage of the machine learning algorithms is interesting. I found the paper very well written. The only drawbacks of the paper are: 1. Sometimes a more extended discussion on the results is lacking, 2. The figures (especially Fig.5-10) could be made more interesting.

We want to thank referee two for the positive feedback on the manuscript and valuable comments how to further improve it. We have provided a more detailed discussion of certain aspects and removed/changed several plots. All specific comments are addressed below.

Specific comments

1. ***Title: since you always work on the past data (even with the NRT approach), it is rather a 'detection' than a 'prediction', so maybe the title could be changed accordingly***

We have changed the title (e.g. specifying the location Altdorf in the title) as the original title might be too promising. However, as we introduce the time shift on the FI time series, we actually do a prediction (for the next hour).

2. ***Line 3: 'lee/luv' – a specific terminology, maybe worth explaining (at least in the Introduction, however 'luv' doesn't appear anywhere else than the abstract***

Lee/luv terminology was left out of the entire manuscript

3. ***Line 68 'COSMO (Consortium for Small-scale Modeling).' -> 'Consortium for Small-scale Modeling COSMO)'; the full name should go before abbreviation***

Has been changed

4. ***Line 90: This is not the exact formula from Rueger and I think there is a mistake there: However, I would recommend sticking to the original formulation as then you have a clear distinction between the dry and water vapor parts.***

Thanks for this hint. Although we did not find a particularly different version of the formula, we have included a changed version (which shows the split in dry and wet part more clearly) in the revised manuscript. Maybe you could give the exact version you mean and we could still implement it in the manuscript.

5. ***Figure 1: Would be nice to see the topography in this Figure to better visualize foehn***

Has been updated

6. ***Section 4.1: I would recommend giving here at least very brief overview of the selected methods***

As the manuscript is quite long already, we actually would like to leave the reader here with the references given describing the methods.

7. ***Line 163: '(negative) maximum' - > why not use 'minimum' here?***

We would interpret 'minimum' as close to zero but one can for sure argue to use minimum here as well.

8. ***Fig.3 and Table 2 show exactly the same information, so I would recommend removing one of them, especially that Fig. 3 is not even addressed in the main text.***

Figure 3 (along with some others) has been removed.

9. ***Figure 4: Make the foehn line more pronounced***

Figure 4 was removed from the manuscript, as the cross-validation/algorithm selection was shortened

10. ***Line 259: Would be good to comment here what the chosen parameters mean***

Actually, this is done in the caption of (what is now) Table 3.

11. ***Figure 5: Maybe you could add vertical lines so the reader can more easily compare the data for particular dates; also you do not comment this plot in the text***

Figure 5 has been removed

12. ***Figure 6 and 7: Maybe there is a way to plot them together for better comparisons of the two methods?***

These plots have been removed, a direct comparison over one week can now be found in the last results section.

13. ***Line 284: A more detailed discussion about the features would be advantageous***

Is now included

14. ***Figure 9: Why not add here a line also of the match with GB (not only with the adjusted one); also it seems like the event of Oct 2020 was caught by the algorithm but in a different epoch – maybe it is something to look into***

Figure 9 has also been removed

15. ***Line 312: Would be nice to see here more discussion on why you change the threshold and how it is done***

The entire section on the threshold optimization has been removed from the manuscript, as the updated results suggest that it is not needed anymore (at least not as much for the old results)

Author's changes in manuscript

Main changes:

- **Title of the manuscript**:
  Title has been changed as the old one might have been to promising, reference to the station Altdorf was included
- **Detailed statistics of data availability**:
  Due to the extensive comments of Reviewer 1 on missing foehn events in the plots provided in the original manuscript, we now introduced detailed statistics on the data availability (for GNSS data mainly) for each case study. These missing periods originated in the fact that the GNSS time series contain data gaps for a number of stations. Since our method needs data from the full station network selected to be able to predict/detect at a certain time period, we cannot produce output if only one GNSS station is missing. This leads to a rather large number of foehn events for which we cannot produce predictions/detections. Detailed statistics on how many foehn events we are able to produce output per case study are now included in the manuscript. Furthermore, the newly chosen study setup is based on criteria influenced by the number of covered foehn events, see next point.
- **Selection criteria for station/feature setup**:
  We newly defined three selection criteria for the feature setup (mainly concerns the station selection) which are outlined in the new section …. Aim was to find a balance between using an appropriately dense GNSS network while still catching the majority of foehn events, in order to have a good amount of training data and more robust statistics.
- **Case studies:**
  The number of case studies was changed (from two to three), with the new one investigating the situation of having less data (only 5 instead of 10 years) but a better station distribution (additional high-altitude stations).
- **Section showing a detailed performance analysis for one week:**
  As requested by referee 1, we added this last section which shows two major foehn events in one week of December 2019, comparing our index against the operational one from MeteoSwiss. In addition, meteorological data (temperature, humidity and wind) as well as the six top predictors of our method are shown.
- **Plots in general:**
  The majority of the old plots were removed (except for feature importances) and only the tables with statistics were left in. As requested, confusion matrices are now plotted in addition.

Some of the minor changes:

- References on machine-learning based classification were shorten to exclusively atmospheric science studies (request from Referee 1)
- Mixture-models (Plavcan et. al, 2014) were added in the introduction
- Section on the influence of hydrometeors on GNSS troposphere estimates was added

- Equation for refractivity (Rueger paper) was changed (comment by Referee 2)
- Performance metrics section was shortened
- New station map with the newest station setup (and setups distinguished between case studies) as well as topography was added