Answers to Comments Referee 1

General response

First of all, 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 first suggestions on strategies how we will implement 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 already have addressed. After a first check of the corresponding code, we encountered a bug when mapping the 10-min foehn index observations to full hours. Furthermore, missing GNSS products at certain times have also caused the exclusion of certain timestamps of the training/testing periods. We have already improved the corresponding code and will include detailed statistics on the overall number of foehn events as well as events which had to be excluded due to lack of (GNSS) training data. The missing response dates are also excluded. We will provide extensive statistics of available/useable data (both GNSS and Foehn index) in the revised manuscript.

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)?*

   The code and data set (pre-processed data, i.e. feature matrix for the test cases) will be made available with the revised manuscript in form of an online repository. We will also try to give more details on the algorithms used in the revised version of the manuscript. The software package is the latest version of sklearn (1.0.2). The number of covariates will be added.
   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-validations. This reason will be added in the manuscript for clarification. 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 will do 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. We will include information on these settings in the manuscript.

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 want to stick to our initial approach, however we will still fine-tune one or two other algorithms (e.g. Random Forest) and evaluate their performance for the case studies as well.

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 will also be 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. That is where the basic idea for this study comes from. Furthermore, Figure 8 already shows the importance of certain stations or differences between them (such as HABG and SANB or KALT and SANB) and indicates that these differences have the highest impact on the classification. We will include additional plots of the differences in ZWD for these stations for selected foehn events which might show a 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. Therefore, we are not sure if we already can incorporate such maps in this study or try to address this point in a follow-up study.

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 will scale 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 will add 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 will contact MeteoSwiss for some details on the most recent performance.

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, a number of 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. This was the ultimate reason to choose the timespan 2010-2020. We will add this information in the manuscript. Extension of the approach to other (foehn index) stations is planned for a follow-up study.

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 actually already done/included when computing the performance measures but including overall values as a separate statistic is also possible. Stratification by season is a good idea which will be implemented by us. As suggested, we will also add such a week-long time series plot, also including meteorological data as this will make it easier to see the performance for dedicated events and increase physical understanding.

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*

   This is a topic where only little research has been done so far, but we will incorporate 2-3 relevant studies in the introduction.

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: we will shorten this section as suggested and give the confusion matrix.

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

   Yes, there are stations in Italy, but 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 will be addressed in the revised manuscript.

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, recent tests reveal comparable performance using unfiltered estimates. Therefore, we will 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 will add this information to the manuscript.

Technical corrections

These will be included in the revised version.