The authors have significantly improved the clarity and readability of the manuscript by incorporating the comments from both previous reviews. As the authors emphasized in their reply to the last comment, this manuscript is supposed to focus on the method and a detailed analysis of the TSI long-term dataset is envisaged for future publications. To give due consideration to this purpose of the manuscript, it is necessary to further relate the presented method to the existing literature. For example:

- *Gnanadesikan, R. (1977) Methods for Statistical Data Analysis of Multivariate Observations, Wiley. ISBN 0-471-30845-5 (p. 83–86)*
- *Alpaydin, Ethem (2010). Introduction to Machine Learning. MIT Press. ISBN 978-0-262-01243-0 (chapter 5, 10)*

Please ensure that already established techniques are described using the proper technical terms where applicable. I would argue that the presented method to detect 22deg halos on TSI images is indeed a binary classification (as opposed to the statement on P16, L7) but based on statistical (multivariate) analysis rather than machine learning. However, there is considerable overlap between both fields (statistical analysis vs. machine learning), especially considering training and testing of the algorithm as well as the related technical terms.

From my point of view, the following 3 major points have to be addressed before the manuscript can be published:
(The following comments refer to the revised manuscript and the authors' comments (AC) on the previous review, which are highlighted in blue.)

1. **Training and testing**: state-of-the-art techniques exist for "training" and "testing" a classification algorithm (see e.g. Alpaydin 2010). The most important requirement is using a new dataset for testing the algorithm which was not used for training. The algorithm presented here seems to use the same data from March 2018 for both training and testing (cf. P13, L1-3). Please revise the manuscript and, if necessary, the presented method accordingly.

2. **Linear classification**: the method of assigning a "sky type" or "ice halo score" presented in this study seems to be very similar to *Fisher's linear discriminant* (Gnanadesikan, p. 83-86) or *Linear Discriminant Analysis* (LDA) (Alpaydin, chapters 5 and 10). Both use feature vectors weighted by the Mahalanobis distance and a threshold to assign new data to one of the classes (linear classification). Please discuss and add citations where appropriate.

   Moreover, please revise the manuscript using the correct technical terms which can be found in the literature (e.g. Alpaydin 2010): e.g. "(expandable) master table" probably refers to "training dataset", which contains "feature vectors".

3. **Sky type classification**: the TSI images in this study were separated into the categories "Cirrostratus" and three levels of cloud fraction "Cloudy" (CLD), "Partly Cloudy" (PCL), and "Clear" (CLR), which were defined by their visual appearance. This method is different

compared to previous studies, which used Lidar observations and a temperature threshold to identify ice clouds (Sassen et al. 2003 and Forster et al. 2017). Using a different method, makes it very difficult to compare the results (P15, L18-21). As stated in the previous review, I see the potential of this study especially in comparing the results to previous studies (and different locations). Therefore, the same criteria should be used to define the basic population of "cirrus clouds".

Furthermore, the choice of sky types does not seem to be very suitable for the described goal of this study: "With the goal of using these long-term image records to provide supporting information [to] the presence of smooth, hexagonal ice crystals in cirrus clouds from observations of 22deg halos , we developed an algorithm that assigns sky type and halo scores to long-term series of TSI images" (P16, L15-17).
Although the majority of 22deg halos coincides with "CS", a significant amount (44% for Jan and 38% for Feb) coincides with "Partly cloudy" and "Cloudy" skies (cf. Tab. 6 "% sky type of all halo instances"). So the sky type categories used here are apparently not a good indicator of whether the present clouds are able to produce a 22deg halo and are therefore not suitable for drawing conclusions about ice crystal microphysics in halo-bearing clouds in general.
It is mentioned several times throughout the manuscript that the sky type classification of the images is used to infer information about the "presence of smooth crystalline habits among the cloud particles" (e.g. P15, L28-30). To answer this question, it would be necessary to identify ice clouds and separate them from other sky types including clear sky, as it was done in Sassen et al. 2003 and Forster et al. 2017.
Nevertheless, it is possible of course to draw conclusions from the frequency of 22deg halos in "CS" skies, but it has to be stated explicitly. In the citation above (P16, L15ff), the words "cirrus clouds" would have to be replaced by "CS", for example. Cirrostratus is only a subcategory of Cirrus, as e.g. Cirrocumulus.

Please address these concerns and revise the manuscript by accurately describing which sky type the results actually refer to when interpreting the results, drawing conclusions, and comparing them with other studies. Please explain in the manuscript the reasons for choosing these specific sky type categories and their merit for the goal of the study.

In the following, please find specific remarks to each of the four points summarized above:

Specific remarks

1. **Training and testing**:
    a. P13, L1-3: "The sections of the record in which visual and algorithm differed were inspected again, at which point either the visual assessment was adjusted, or the misclassified images were included in the Master table in order to train the algorithm toward better recognition."
    Apparently, the March dataset is used for tuning the algorithm, i.e. for finding the classification threshold. This is not equivalent to testing. For the latter, the trained and tuned algorithm is tested against completely new data and should not be

updated simultaneously. In order to avoid a bias in the assessment of the final classification quality of the algorithm, the final trained version should be applied to a dataset which was not used for training (cf. literature for state-of-the-art techniques). This implies that the training data set actually contains 80 + 44,026 images. Please address this point.

b. P16, L11: “Further training is easy to incorporate via a master table which provides means and covariance matrices to the algorithm.”
“Training” the algorithm presented in this study means finding a threshold that best separates the two classes. Adding new feature vectors to the “master table” is usually referred to as generating training data.

c. P13, L17: "Upon inspection of the numerical values for IHS, it becomes clear that `a cut-off is needed to assign an image with a label of halo/no halo. This cut-off value is arbitrary and dependent on factors such as w and C_0, as well as the quality of the calibration. **Our testing places it at around 4000 for the month of March**.”
Which values were used for the other months? To my understanding, training the algorithm should result in *one* threshold value which will be applied to the whole TSI dataset. The sentence highlighted above gives the impression that a separate threshold is determined for each month. In that case, it will require a lot of work tuning the algorithm for each month separately for this large dataset. Please clarify.

d. AC: Both, C0 and w are arbitrarily chosen, and are passed as a parameter as befits the question. The reference to w=4 images is specific for the day data in figure 5. For the evaluation in section 3, w=3.5 minutes. This limits the time resolution for halo appearances to 3.5 minutes, but smooths out false halo singals encountered in the record for that month. The equation references have been corrected in the renumbering of equations.
This should be explained in the manuscript since it affects the results for the mean duration of 22deg halos in Tab. 6. and the histogram in Fig. 8.
If a different value for w is used for each day, the first bin (0-5 min) in Fig. 8 will be mainly subjected to the this choice (should actually be 4-5 min?). Why is the choice of w changed? It should be constant throughout the analysis.

e. P15, L2: “Due to the time-broadening applied via Eqn (16), the display time cannot be resolved below 3 minutes.”
P12, L20: “The broadening w in Eqn (16) was chosen as 4 images for this example, which means the Gaussian half width corresponds to 2 minutes.”
See previous comment. Please double-check, is it 2, 3, 3.5, or 4 minutes?

2. Linear classification:

a. P10, L22 “An image IHS and STS are assigned as the average over all scoring quadrants.” How were the results calculated for each individual quadrant in Tab. 6? By a linear combination as for the Linear Discriminant Analysis?

b. P16, L7: “The algorithm presented here for TSI data […] does not characterize halos in a binary decision, but rather assigns a continuous ice halo score to an image [...]” The presented algorithm does classify halos in a binary decision after computing the score. This is true for other classification algorithms as well, e.g. for the random forest classifier. Please correct this statement.

3. **Sky type classification**:
   a. P15, L18-21: “For example, in January we found that 9 % of all cirrostratus skies were accompanied by a 22deg halo. In the data for April, this fraction increased to 22% of all cirrostratus skies. We also have registered halos for a portion of partly cloudy skies, and for cloudy skies. No halos have been registered in any of the clear skies. This is certainly consistent with the observations of Forster et al (Forster et al., 2017).”
   The reference of the last sentence “this is certainly consistent with the observations of Forster et al” is not quite clear. Does it refer to “No halos have been registered in any of the clear skies” or 22% of all cirrostratus skies show halos?
   If the comparison refers to the fraction of cirrostratus skies accompanied by a 22deg halo, this statement is not correct. Sassen et al. 2003 and Forster et al. 2017 use Lidar data (and a temperature threshold) to identify clouds dominated by ice crystals. So even though the resulting numbers are similar for April, the population is different.

   b. A quantity that could be directly compared is the overall frequency of all 22deg halos with >=1/4. Could you provide a number?

Minor comments:

- P15, L28-30: “One of the conclusion to be made from the relation between STS and IHS concerns the confidence in the presence of smooth crystalline habits among the cloud particles, as shown only in a one-fifth fraction of all cirrostratus.”

  Please clarify this sentence: what is the conclusion here? The average fraction of 22deg in CS sky types amounts to 16.25%, i.e. rather 1/6 than 1/5.

- AC: “[…]I have not been able to visually and reliably discriminate parhelia in any TSI image. An algorithm specifically for parhelia was therefore not attempted. With the separation into quadrants, any existing parhelia would form right on the boundary between top and bottom quadrant, and basically average into the radial intensity of this quadrant. […] The top quadrants, if not overexposed, may give halo signals. But again – parhelia can not be visually distinguished in those images.

  It would be worth adding a short sentence to the manuscript, describing that 22deg parhelia could, in principle be mis-classified as 22deg halo, but due to the coarse image resolution and low brightness, they have not been detected so far in the TSI images.

- P7, L17 ff: Was the image distortion accounted for in addition to the coordinate transformation in Eq. 9? If it was assumed negligible, this should be stated in the text and expected errors could be cited from Long et al. (cf. Fig. 5)
- Long et al. appears twice in the references.
- P7, L17: plain -> plane
- P12, L27: 44,026 images vs. Tab. 6 44,057 images
- Table 5: It seems that the attempt was made here to combine two tables into one. I would suggest limiting the table to the assessment of the algorithm compared to the visual inspection, which has to be assumed as "ground truth" here. That means the table should express only the second part of the sentence: "86 % of all algorithm CS skies also identify as CS if inspected visually". The first part of the sentence would focus on assessing the visual classification of the images against the algorithm, which is not the primary interest here. If the authors consider both results equally important, I would suggest separating the results into 2 tables.

  For example:

	Visual			
Algorithm	CS	PCL	CLD	CLR
CS	**86%**	3%	1%	6%
PCL	…	**91%**	…	
CLD			**98%**	
CLR				**93%**