Authors‘ Response to Referee Comment #1

Sedlak et al., 2023: Analysis of 2D airglow imager data with respect to dynamics using machine learning

We would like to thank Anonymous Referee #1 for his valuable comments.

Referring to the specific comments:

1. Explain the rationale behind the decision to use 70% of the available dataset for training; 20% for validation and 10% for testing.

   These are typical sizes, the exact one is heuristic. But in general, the training dataset should be by far the largest, because the weights are actually adjusted based on the samples of the training dataset, and the validation dataset is larger than the test dataset, because it is used to evaluate all possible adjustments during and after training. Thus, it is used quite often and should contain a large variety of examples, and the test data set is used only as a final check to prevent overfitting on the validation data and thus confirm that the performance on the validation data set can be generalized.

2. Explain the reason for 100 epochs (line 157) (see also lines 267-268). What determines the number? For example, could it be 50 or 200?

   During training, the performance on the validation dataset is monitored, and for both models, most of the improvement happens within the first 50 epochs, and there is barely any improvement after that. So, we could have chosen fewer epochs, but it doesn't matter because the final model is chosen on the best configuration (model of the epoch with the best performance on the validation dataset) during training. More than 100 epochs wouldn't be useful because, as mentioned before, there is hardly any improvement after the first 50 epochs.

3. One can readily appreciate the difficulty of manual classification as discussed in line 409 and following. However, the number of time steps in the test dataset that were deemed to be misclassified (1110 and 1040) in Table 4 is a cause of concern. If I understand this correctly, these were manual classifications. The manuscript describes the negative impact of these manual misclassifications on the statistical measures of the neural network classifier (lines 425-427; lines 430-435 and lines 503-504). Why do the authors not repeat the calculation of the statistical measures using the “correct” classification to establish the “true” value of the statistical measures?

   In Table 4, we only looked at the confused calm and dynamic sequences, but in order to correct the statistical measures, we would have to analyze:

   - all the other dynamic sequences that are not predicted to be dynamic
   - all the other sequences predicted to be dynamic but not classified as dynamic
   - at least all correct calm and dynamic predictions, if they are really correct, since both the manual classifications and the model predictions tend to confuse calm and dynamic episodes.

   If we had adjusted the statistical measures without doing this, we would have biased the statistical measures in a particular direction.

Analyzing this by hand is again very time consuming (much more time consuming because it affects a large portion of the test data) and does not improve the model, and we wanted to invest that time in improving the model.

4. Lines 430-435 state that the NN-classifier is superior to manual classification at distinguishing between “calm” and “dynamic” episodes, which is indeed good news for the method, but leaves the reader wondering if the statistical measures have a great deal of validity.

   That's true, and we tried to emphasize that in our discussion. But the statistical measures still give an indication of the strengths and weaknesses of the model.

5. A second even more disturbing issue arises with the large number of incorrect manual classifications. Since the test data is only 10% of the total (70% training; 20% validation in line 152), perhaps a large proportion of the training and validation were manually classified incorrectly to start with, thereby having a negative effect on both the training and the validation.

   That's also something that we tried to emphasize in our discussion, because it's actually a really good thing. Because it shows that you can train a model that works very well even if you don't have perfectly classified datasets.

Referring to the minor corrections:

Title: The suggested alternative title would be another good choice. We decided to leave out title unchanged.

Line 26: 2017b changed to 2017

Line 29/30: Changed.

Line 38: Changed.

Line 42: ‚exist‘ replaced by ‚propagate‘

Line 54: Changed.

Line 69: Changed.

Line 71: Changed.

Line 73: Changed.

Line 81: changed to ‚has already been described in‘

Line 83: There is a half-sized space between 640 and pixels which we use as a separator between numbers and units throughout the article. No changes made.

Line 88: Usage of half-sized space: see comment to Line 83. We changed ‚pixels‘ to ‚pixel‘.

Line 89: See Line 83.

Line 90: Corrected. We replaced the date representation as suggested.

Line 91: Following the suggestion of Referee #2 we changed the wording to „during".

Line 92: Changed.

Line 93: Inserted ‚of'.

Line 94: Changed.

Line 97: Changed.

Line 107: Changed.

Line 117: Changed.

Line 122: Changed.

Line 125: Changed.

Line 133: See Line 83.

Line 135: Changed.

Line 140: Changed.

Line 157. See answer to Specific Comment #2 above.

Line 162: Changed.

Line 168: Changed.

Line 196: Omitted.

Line 237: Omitted.

Line 291: There is no space in our original file hence this must have been an artifact due to pdf export.

Line 328: We replaced all lower " by upper ones.

Line 332: Changed.

Line 335: Corrected.

Line 337: We corrected this as suggested.

Line 365: Corrected.

Line 368: Omitted.

Line 382: Changed.

Line 393: Corrected.

Table 5: Thank you for finding this inconsitency! We updated the values in Table 6 (former Table 5).

Line 428: Changed.

Line 429: Changed.

Line 456: Changed.

Line 483f: Corrected.

Line 492: Changed.

Line 493: Changed.

Line 509f: Changed.

Line 541: Changed.

Line 552: Changed.

Line 555: Changed.

Line 607: Changed.