# Minor Revision: Answers to Referees

Stefan Niebler et al.

November 2021

## 1 General

1. We adjusted Figures 6,8,9, S2 and S3 to be more accessible regarding color blindness.

## 2 Remarks from previous File validation

1. Please provide a source of fig 2, 3 if you are the originator, please inform us.

   We added a reference in the corresponding figure captions, as the trademark in the images is rather small.

## 3 Reviewer 1

1. The authors have made a thorough revision of the manuscript based on the comments of the reviewers (many thanks!). The quality of the paper has improved significantly and I think it is now ready for publication.

   Thank you!

## 4 Reviewer 2

### 4.1 General

1. The authors present a revised version of their ML-based surface front detection. One certainly cannot dispute the usefulness of feature-based detection methods, but my reservation about whether manual surface analysis should guide the development of a next-generation automated feature-based method remains intact.

   We still think that manual surface analysis charts provide a sound ground truth for this method. Since we can show that the fronts derived from surface analysis charts on average have the thermodynamic properties, which would be found from a TFP-method, we use the surface analysis charts as ground truth.

2. The discrepancy between the traditional TFP-based methods, which arguably have their own shortcomings, and the presented method for emulating DWD and WCP fronts does not, in my opinion, indicate a particular weakness of the TFP-based methods, but must be considered in light of the weakness of the manual surface maps, which either fail to account for or erroneously indicate or displace relevant surface features otherwise correctly detected by the automated TFP-based method. The question remains as to what should be accepted as ground truth, and as stated earlier, I cannot recommend relying fully on manual analysis. This position is in contrast to several statements by the authors who continue to hold on to manual analysis as a ground truth and consequently continue to argue throughout the manuscript that traditional TFP-based methods are outperformed. The answer might simply be that the manual analysis is erroneous in many cases, and the ML-method has learned the bias while the TFP-based method – in fact – outperforms both. It is simply a fundamentally different viewpoint.

   Actually, we show in a comparison that the fronts determined by the weather services (and also by our network method) exhibit the characteristics in terms of temperature gradients etc., which one would expect from a front identified using the TFP-method. Thus, it is not a different viewpoint, since fronts determined from both (and different) methods should usually agree very well.

3. Even though the authors have improved their comparison between a TFP method and their new method, I still think the comparison is incorrect. A proper choice for a baseline method must always be seen relative to the ground truth. Here, manual surface charts are used as ground truth, which are drawn based on several variables at several heights, and as baseline a method is chosen which uses one variable at one height and was never developed with the specific goal to reproduce surface charts. As already mentioned, I recommended removing this comparison, since it is not necessary for the publication. Only the comparison with an earlier ML method would make sense. However, since even the authors of this study argue in their reply that they are not able to handle the code provided by of one of the earlier ML methods, I am concerned about the reproducibility of these studies. At the end of this document, I recommend another ML method that uses the same ground truth – maybe this code is more user friendly and can serve as the baseline method the author wish to have.

   We do not think that a comparison with the mentioned ML method is really meaningful. We did not find the provided code base, and the results do not seem to be very robust.

4. Nevertheless, in their revised introduction, the authors have addressed aspects of this discussion, but the need for labeled training data is so central to their method, there is basically no other option for the training

of the ML method but to accept the author decision and evaluate the manuscript considering their viewpoint.

As mentioned above, we still think that manual surface charts are a sound ground truth, therefore we use them in our study.

5. To me then, the automated method gives us gridded front data that might be useful for meteorological research related to phenomena associated with the passage of surface fronts. The presented example, a confirmation of an earlier studies that addressed the question of front-related extreme precipitation events, leaves unfortunately the question open of what exactly can be learned using ML-based methods given that the authors basically show that the method reproduces exactly what was previously found using a TFP-based method. Recommendation: It would be helpful to at least give some indication at the end of this section of what exact new physical insight can now be generated with the new method that could not be generated before.

   It is stated in the text that this evaluation was also carried out using much smaller radii on the high resolution ERA5 data. Some new results are presented in the supplement. Such an evaluation can hardly be carried out using a TFP method, since these methods have problems with high resolution data as we could see in our investigations using the baseline method.

6. The climatological application is otherwise a very nice example that would motivate a section on the issue of explainability of data-driven methods. While the presented method produces climatological patterns in agreement with previous findings, it is beyond that capable of splitting different front types in a clear manner. Of particular interest would be to understand what variables are key for the learning process and if the climatological patterns would look different if only trained on a single variable. A particular strength of the ML method could be to use a low number of input features to reproduce manual analysis. Again, to me, however the results seem to result from the combination of various input channels, while traditional methods often rely on a single variable which seem to be not sufficient to separate different front types. Layerwise backward propagation might be a simple way of showing what variables allow the network to develop this ability. Recommendation: It would be useful to give some indications in this direction at the end of the corresponding section.

   We have carried out such an investigation, but in our case this was not successful. Actually, it is well known that such an attribution method does not always work.

7. In the summary it is argued that the method can also be applied to higher-resolution data. I think this is not the case. To make the method mesh independent, the input training data would need to be converted to continues space and training would need to be performed in continues space as

is done in random feature methods or eventually also in FFT-space. There is something to be said here about mapping between Banach spaces.

It is stated in the manuscript that there are no principal obstacles to apply this method for high resolution data. We did not state that the method is mesh independent. It might be that the network (pre-) trained on ERA5 data must be further trained for the use on high resolution data.

### 4.2 Introduction

1. The authors are encouraged to add more reference to their statements relating fronts to, for example, wind gusts or extreme weather.

   We added references Catto and Dowdy (2021), Catto et al. (2015) and Martius et al. (2016) to the manuscript at Line 20

### 4.3 Some ML related questions:

1. Is the Batch normalization really needed? Usually, it accelerates the training process and additionally improves the skill. However, from a theoretical viewpoint, it is unclear why this is case and thus it might not be needed in this particular application.

   While it is certainly possible to construct and train networks without batch normalization (BN), it has a number of favorable properties that improve results significantly: Aside from faster training, the main benefit of batch normalization is an increase in generalization performance, i.e., deep networks trained with BN are less prone to overfitting. This effect has been verified empirically many times, see for example:

   Johan Bjorck, Carla Gomes, Bart Selman, Kilian Q. Weinberger Understanding Batch Normalization NeurIPS 2018

   https://proceedings.neurips.cc/paper/2018/file/36072923bfc3cf47745d704feb489480-Paper.pdf

   From a more theoretical point of view, there are connections to increasing the margin of the classifier. See for example:

   Jure Sokolic, Raja Giryes, Guillermo Sapiro, Miguel R. D. Rodrigues Robust Large Margin Deep Neural Networks https://arxiv.org/pdf/1605.08254.pdf IEEE Transactions on Signal Processing, 2017

   - Why BN in U-Nets?

   The architecture used in our paper is a U-Net, which is a type of network that uses ”skip-connections” in order to reduce training problems. One might think that the introduction of skip connections means that batch normalization is not very useful any longer, as these architectures are less prone to vanishing gradients and related problems. However, there is empirical and theoretical evidence against this, see for example:

A. Labatie: Characterizing Well-Behaved vs. Pathological Deep Neural Networks. ICLM 2018. (nicely summarized here: https://towardsdatascience.com/its-necessary-to-combine-batch-norm-and-skip-connections-e92210ca04da )

The author shows that deeper architectures are prone to confinement of the data signal to low-dimensional subspaces (the singular values of the Jacobian drop quickly; one could call this vanishing dimensionality), and show that only a combination of BN and skip-connections can reliably counteract the problem.

In summary, BN is usually included in most modern architectures as folklore-based ”best-practice”; literature gives us, nonetheless, rather strong evidence that this is highly useful from both an empirical and theoretical perspective. Aside from improving numerical conditions (gradients, singular value spectrum of the Jacobian) it also improves generalization performance / reduces overfitting tendencies.

2. Why is the drop-out chance set to 0.2? Is there any over-fitting without it? How does this relate to the problem of choosing arbitrary thresholds? I recommend a brief discussion of the sensitivity.

   Lagerquist et al. found that a high dropout worked well to avoid overfitting in their case (0.25 and 0.5). It does not appear that our network is overfitting at this point, which is why we did not use a higher dropout. Potentially a lower dropout may work as well.

3. Why did you choose 3 drop-out layers and avg. pooling steps in your U-Net architecture and not less or more?

   Using less layers performed worse in our first tests. Using 4 encoding and decoding blocks would exceed the memory of the used GPUs. This study is however not about finding the optimal network architecture, but to apply the network to a meteorological problem.

4. Why are the number of channels changing from 330 to 64 after the first encoding block, but for all further encoding it increases by a factor of two?

   With this first encoding block we tried to learn some kind of embedding of the variables. It slightly improved results but it might not be necessary for the whole cause.

5. Reference for U-Net should also be given to Shelhamer et al. 2016 (doi: 10.1109/TPAMI.2016.2572683) We added this. See Line 98

6. L. 345, how did you determine the deformation factor of k=3? Shouldn't the choice be tested against randomness in some way? How, as before, does this choice relate to the problem of choosing arbitrary thresholds? A common weakness of traditional methods.

   Basically we were looking at the width of the results from k=0. if we choose k = 1, we obtain double lines, as it cannot fully cover the deviation. k=3 works quite well as it is appears to reliably cover the width of the bias.

We did not test $k > 3$, as $k = 3$ already works. Yes this value is chosen arbitrarily, but this is not critical for the detection of fronts rather than the placement/width of the fronts related to the label bias.

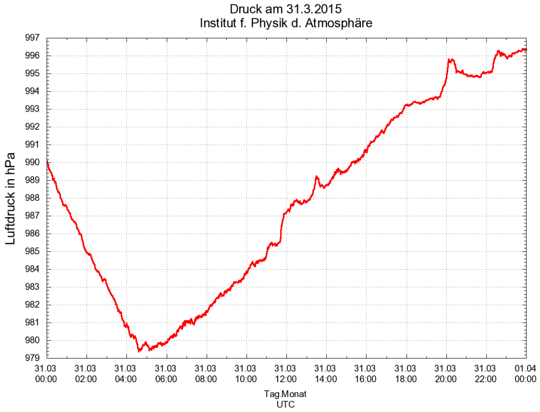
## 4.4 Section 2.2.4

1. The authors are encouraged to add more reference to their statements relating fronts to, for example, wind gusts or extreme weather.

   We do not understand where in this section this should be included, as it is not the topic of the section. As stated before we added such references into the introduction.

2. Several previous studies have questioned the usefulness of front lines in general and for the use in next-generation front detection methods. These studies rather recommend using frontal regions or frontal volumes.

   Evaluating the temporal evolution of a front at a surface station, it can be often seen that the surface front consists of a very narrow line, as e.g. can be seen in the surface pressure. We added an image from the passage of storm Niklas at the weather station in Mainz in 2015, which shows clearly the small extent of the surface front (pers. comm. P. Reutter, `https://www.ipa.uni-mainz.de/wetter-alt/wetterbesonderheiten/`)



3. Is all of what is done in this section needed simply to obtain front lines?

   Yes this section was mainly done to create lines. Furthermore this is a problem of available data as there is no labeled datasets of wide lines. And in the case of simply expanding the given frontal lines, we propose that one can simply widen the results of the network as well.

## 4.5 Section 2.3

1. I recommend removing this section and the corresponding comparison in Section 3.1.1. Also, it is noted that only midlatitude fronts are included

for the TFP method, but in Section 3.2.2. the opposite is done.

We do not agree that this section should be removed. We did not quantify the TFP method outside the midlatitudes, which we agree would not be a fair comparison. However we do not agree that not showcasing the shortcomings of a method and comparing how our method fairs in those regions should not be done.

In 3.2.2 the TFP method was applied by Catto and Pfahl. We only used our network in this section.

In 3.1.2 (which might be what you mean) the fronts outside the midlatitudes are not used in the quantification. They are used in the discussion part of the section, but we still believe that it is fair to also highlight the cases where a certain method does not work correctly and show if a proposed method can perform better there.

## 4.6 Section 2.4.3

1. Even though POD and SR are intuitive measures, I recommend to better explain the meaning of nmws and nws. The latter is “the count of all provided fronts” the former “all fronts that could be matched”. To what does provided refer to (provided by whom)? What is a front that is provided but cannot be matched?

   Line: 412 ”We define nMWS as the count of fronts provided by a weather service,”

   What is a front that is provided but cannot be matched?: Such a front is irrelevant to the evaluation. However, such a front is a front that is provided by the weather service, that did not fulfill the matching criterion (The criterion mentioned in the section starting at line: 396)

2. Fig. 6 is missing a color bar for the gray shading.

   We added the color bar to the figure

3. Fig. 6 The yellow class is labelled as “no class” but there seems to be no yellow label in the figure.

   There is at the center image bottom row. Albeit it is only a very small part.

## 4.7 Section 3.2

1. Overall, I am afraid I do not understand the purpose of this section. Is it about showing that DWD and WCP fronts have gradients?

   Yes. It further shows that both the results of the network as well as the DWD and WCP fronts exhibit the same characteristics as those that would have been found by a TFP method. This shows that the surface fronts identified by the weather services are a reasonable ground truth for learning to detect fronts.

2. Fig.9: What is the variance for the shown averaged values for each line and are the differences between the methods within or outside, for example, the range given by -/+ two times the standard deviation of the sample that went into the averaging for each method?

   We absent from a statistical test as we are averaging on rather large temperature differences. The standard deviation itself for the equivalent potential temperature for example lies at approximately $10K$ to $15K$. This section is not intended to statistically determine the difference between weather service provided and network detected fronts, but rather to show that both generate fronts that are in line with the expected behaviour (e.g. TFP criterion, temperature gradients, ...)

3. The lines all look very similar to me and may not significantly be different from each other.

   This section is intended to show that the results of our trained network correspond to the expected behaviour of the different types of fronts. It is to be expected that these results correlate with the cross sections of the weather service fronts, as long as both results truly show fronts.

4. In all honesty, this does section does not add much to the paper. This section should be removed as the paper can be published without this information.

   We do not agree. We believe that this is a very important section, as it shows that the surface fronts of the weather services are a good label, as they appear to fulfill the criterion used by the TFP methods rather well. Additionally the detected fronts also agree very well with this criterion.

### 4.8 Section 3.2.2

1. I am afraid I do not support the usage of an attribution measured that uses an attribution radius defined in terms of degrees. I would assume that 2.5 degrees correspond to a different area/distance at different latitudes so you will attribute less precipitation to fronts at higher latitudes, don't you?

   Yes, this is true. However, we tried to stay as close to the other paper as possible. In fact a $km$ based attribution may be more accurate however as you stated it would most like result in a higher matching rate, as the attribution radius in the higher latitudes would increase. This is also an advantage of our method, as it is applicable to ERA5 we can in future work use a $km$ based attribution ratio to research connection of fronts and other events more accurately.

2. Not sure if the difference between fr and a(fr) is fully clear. Is the first the number of fronts at a grid point and the second a probability? What do you mean by "grid point p is associated with a front" other than "a front occurs at p"?

We added some text to make this more clear: A "grid point p is associated with a front "a(fr)"", if it resides within a 2.5° distance of any front. On the other hand " a front occurs at p" means that a front is located at this exact grid point. The same distinction for extreme precipitation. We extended the text in the listing at Lines 676 and 679

3. Fig. 10: Maybe I missed it but why are the polar regions not shown?

   Because the other paper did not do this as well. Plus the resolution of ERA5 becomes very inaccurate the more pole ward we evaluate.

4. Fig. 10-12: Some words in the title of the figures are capitalized others not.

   we made them all lower case

## 4.9 Literature

1. The authors may consider the following paper, which appears to target the same ground truth but uses a random forest method. I guess that this is the baseline method the authors are looking for. Bochenek, B.; Ustrnul, Z.; Wypych, A.; Kubacka, D. Machine Learning-Based Front Detection in Central Europe. Atmosphere 2021, 12, 1312. https://doi.org/10.3390/atmos12101312

   Thank you for this link. We will add it to our literature. However, we do not think that at this stage we should compare against this method, as we did not find provided code base nor do the results appear very robust, regarding Table 3. So we do not believe that it provides a good baseline to compare against. We might do so in a future work. We added it with some text at Line 85.