Automated detection and classification of synoptic scale fronts from atmospheric data grids
Final Author Comments

September 28, 2021

1 General response

First of all, we thank both reviewers for their helpful comments and suggestions which lead to a significant improved manuscript. This document contains our responses to all reviewer comments.

2 Referee 1

This is a good, well-written paper that should be of interest to the readership. However, I have a couple of minor comments that could be addressed in a revised version of the paper.

Comments:

1. The paper is longer than it needs to be, and some information is spread over the paper which makes it difficult to extract the relevant pieces. E.g. when introducing the vertical levels in l.117 and then mentioning that you only use 9 pressure levels in l.196. Why not describe the data-set augmentation with the data in section 2.1?

   We have rewritten some parts of the paper and tried to condense some redundant parts of the manuscript. We reordered some sections, put the discussion of the results into section 3, deleted the discussion section, and added some more text to the conclusion.

2. Section 3.1 and 3.2: I have tried for a while to understand why you present results for the validation AND the test dataset and gave up. Why do you need section 3.1? You are writing in l.300: "We validated our model during training using 1460 samples of data from 2017. We evaluated our trained models on 1 year of data from 2016 using an object based evaluation described as described later in this section." This does not really explain why you need the two sections. Also, section 3.1 starts with "The trained models were evaluated on test sets..." which generates ultimate confusion. Do you loose any information when removing section 3.1? Maybe I am missing something.

   We agree that the information from Section 3.1 does not add any important information to the overall manuscript. In combination with the length of the Manuscript and the fact that we plan to add an application of the method, we decide to remove Section 3.1. Additionally we will rearrange some information regarding the evaluation from this section into the methodology section to remove the spreading.

3. l.193: I do not understand this. You say you "ignore the outer 20 pixel". But then you are saying that the brighter areas can be used as input in the caption of Figure 4. Are they used as input but not predicted? But then the output domain should be smaller than the input domain in Figure 3...? And why do you crop to 128x256 pixel (l.199)? And then there is again a confusing mentioning of the 5 degree border in the caption of Table 2...

   Yes, the bright+dark shade describe possible input regions, the dark shades are the regions, where a valid output can be located. The network essentially provides an output of the same size as the input, however as outer pixel may have far less information, we decide to ignore the outer 20 pixel of the output, such effectively reducing the size of the output region. So yes for a given input domain, the output domain is smaller, due to this. The 128×256 crop is performed to ensure that our networks input is divisible by 8 as well as to create some more samples to draw from during training. We rewrote the respective section to make it more clear, why and how this is done.

4. l.8: I would not call the baseline model "ETH". ETH is a very large institution.

   We now refer to the method as "baseline"

5. l.21: Maybe add a reference to the Mei-Yu front?

   We added a suitable reference

6. l.22: "Determining the position and propagation of surface fronts plays an important role for weather forecasting". Well, the prediction of the position, yes. But is the same true for the automatic detection? Fronts can easily be identified in field maps by the trained eye. Why do we need the ability to detect them automatically with ML? I do understand why, but it would be good if this would be made more explicit in the intro, otherwise it seems that you have a hammer and are searching for nails.

   The detection of fronts in operational weather forecasts is of course an important task. However, even for postprocessing of model data this would be interesting. Beside the task of operational weather centers, an automatic classification of fronts in meteorological data set is of interest for research purposes. As we show in the newly added application, we can use the automatic detection for statistical evaluations, e.g. for the connection of fronts with extreme precipitation events. Other examples are obvious, as e.g. the connection of clouds and convection to fronts of different types. We added some text in the manuscript to highlight this purpose.

7. l.24: What are empirical guidelines?

   Many weather services as the DWD have some guidelines how to determine fronts. Of course, the physical variables play a role. However, also some empirical connections and features might be helpful for certain regions to determine fronts and other features manually. This was personal communication with the DWD. Unfortunately, we cannot provide detailed examples. We deleted the word empirical, since it is misleading.

8. Section 2.1: Maybe I missed it, but do you actually state the resolution of the NWS and DWD datasets somewhere (or the resolution equivalent of the PNG image)?

   NWS labels are given as coordinate pairs with a $0.1°$ resolution, DWD images come at a 4389×3114 pixel resolution, from which we extract coordinate pairs. We added this information to the paper

9. Figure 3: I do not understand the encode and decode blocks. Can you add some info here? Also, what are the white boxes the "copy" arrows end in?

   The code and decode blocks are just several convolution, ReLU and BatchNorm operators in sequence. We decided to put references to those at the corner of the image, to not unnecessarily convolute the network architecture image. The white Boxes are the results of the copy operators. Basically: The box at the beginning of the gray arrow is copied along the arrow and the destination of this copy is described by the white box. This white box is concatenated to the result of the upsample operation coming from below. We tried to make this more clear in the manuscript, by adding some text, also in the figure caption

10. l.198: "If both labels are available". What does this mean? At a certain point in time? Why should this matter?

    Yes, we refer to a certain point in time. As our data sets from the NWS and DWD do only overlap for a subset of our data set, we will have input data, where only a label from one weather services is available. In the case that we have a label for both weather services for a certain point in time, we randomly select which one to use. The result of this selection can vary between epochs. On the other hand if only a single label is available, we do not need to randomly select, which label to use, as there only is one. We updated the respective section in the manuscript

11. Table 2: The whole caption should be reformulated. "For the global region this border is included within the mentioned range." ?

    We made a new table, where we explicitly state the input and output regions used during training.

12. l.242: This paragraph is important but very difficult to understand. It should be rewritten.

    We have rewritten the paragraph and added a figure to visualize this approach. We also added more details about the method.

13. l.279: I would not use "t" for the index of the channels as "t" is often used for time.

    We have changed the index

14. l.280-282: I do not understand this. “individually for each batch”? “more emphasize onto classification”? Either equation (2) holds, or not.

    The loss weighs how many samples in a batch contain certain labels, which is why the weights may differ between batches. We have rewritten this part and updated the loss to clearer represent that the weights used for the loss are dependent on the available labels in the batch.

15. l.289: Why did you not evaluate the baseline at 0.25 degree? I guess there are good reasons, but please state them.

    The baseline natively runs at 1 degree. An upscaling of the method to 0.25 degree was unfeasible, because of the additional small scale features, disturbing the gradients of the thermodynamic variables. Simply upscaling the results would introduce additional ambiguities in the placement of the labels. A downscaling seemed to be the more accurate choice at this point. We added some text about this into the manuscript.

16. Table 3: You can as well remove the “Stationary” line.

    We removed it.

17. Table 5: “The suffix “all”. . . ” I do not understand this sentence.

    We show two slightly different evaluation methods in the manuscript, which were also applied to this evaluation. In the updated version we only use one of these when comparing against the baseline method, which we explain in the previous section. This ultimately means that we do not need the distinction in this table. We therefore removed the sentence.

18. l.488: I find this a bit confusing. You would not leave out a certain region in a real-world application, so why here?

    As Referee 2 noted, the baseline method is not suited for application outside the midlatitudes. As Greenland is located outside the midlatitudes, it shouldn’t have been in the quantitative evaluation for a fair comparison. In our previous evaluation it was apparent that Greenland caused a lot of false positive fronts detected by the baseline method which drastically reduces the correlation coefficient. Removing Greenland from the evaluation was to show that the low score was mainly caused by this comparatively small region, while the rest is pretty accurate. Nonetheless, as stated we decided to restrict this evaluation to the midlatitudes, which in return already excludes Greenland making this distinction obsolete.

Typos etc:

1. l.51: typically → done
2. l.253: predicted fronts → rewritten
3. l.301: remove “described” → done
4. l.346: “be be” → removed
5. l.349: “slight edge”? → rewritten
6. l.351: “fact that training” → done
7. l.403: “most likely” → done
8. l.445: “and the European data” → wrote ”than the European data”
9. Caption Figure 7: “on the for the”→ done
10. l.514: “for is the lack” → added a this
11. l.439: “However,” → done

3 Referee 2

Automated feature recognition has proven useful in gaining scientific knowledge of the dynamics and relationships between various atmospheric flow features such as cyclones, jets, and surface fronts. However, there are a variety of automated methods to identify the relevant feature of interest because even trained experts do not agree on how to define a feature. This applies to surface fronts for which no general definition exists. Therefore, improved methods that help to gain additional insight into the nature of fronts are important, however I am have some concerns on whether ML-based method trained on surface analysis is the next step.

General comment

1. There is no single accepted front definition and different weather centers use their own definitions based partly on physical considerations, partly on training and experience, partly on the specific local meteorology, and sometimes simply artistic. It is therefore questionable whether a front identification should be guided by manual surface maps or physical arguments. This dilemma is nicely summarized in Uccellini et al (1992) and Sanders (1999) and was lately reviewed in Schemm et al. (2018) and Thomas and Schultz (2019a). I recommend that the authors review these earlier studies; their introduction comes in its first paragraph without a single reference (and there are numerous studies that link fronts to extreme weather that could be referenced). Also, there is little historic background provided.

   The front dilemma can be summarized with the following example: The UK MetOffice automated surface analysis regularly displays double fronts, while the DWD chart never shows these fronts – see Fig. 2 in Thomas and Schultz (2019a). Instead, DWD-front are Norwegian-like and hemispheric spanning, which is more art than science. The missing double fronts are however real and important to detect. They will be missed if trained on DWD charts.

   Related to the definition of fronts, there is one stream of front definitions that is based on baroclinic instability and there is also a second front definition based on air-mass boundaries – see Thomas and Schultz (2019b). The two are mixed up in this study, for example, when the authors speak about fronts that are associated with the propagation of extratropical cyclones but thereafter describe fronts as air-mass boundaries. These air-mass boundaries, which provide little baroclinicity, are very interesting for research. If one wants to detect these, one cannot train a ML method on DWD charts, although it seems as if DWD uses many meteorological variables to draw their front, which is common for the air-mass boundary definition of fronts but not that based on baroclinicity. Maybe DWD excludes mesoscale fronts in general?

   Overall, I therefore reject manual surface charts as ground truth, baseline method or "gold standard" for verification. The surface maps are biased, inhomogeneous, only partly based on physical reasoning and cannot be transferred between different regions. I find a tool that learns these biases, here the DWD bias, difficult to use for research purposes, though they might be useful in an automated DWD workplace environment. Even though the authors try to alleviate some of these issues with the blurring of the front position shown in their Fig. 5, I hesitate to conlcude that ML-based fronts trained on surface charts is the way forward.

   We respectfully disagree that manual surface charts are not suited as a baseline. In our (updated) evaluation we show cross sections of both our detected fronts and the provided weather service fronts at 850hPa, especially for equivalent potential temperature, as used e.g. for the baseline method. Both fronts seem to show the characteristics that more traditional TFP methods are based on, e.g. TFP = 0. This alone seems to refute the believe that the surface fronts do not follow any physical reasoning. Further you state that the weather service label contain personal bias of the executing meteorologists and acknowledge that we try to alleviate some of these biases. However you misunderstood that we are trying to inflate the label. We do use a label deformation procedure adjusting the labels prior to evaluation. The idea behind this was that it would help to relocate biased labels, assuming that they are at least somewhat correctly placed. As mentioned before the cross sections seem to indicate that this is indeed the case, which is why we believe that those are suited training data and the output of the trained network is a suitably good front detection. At last we also use **two** different weather services (NWS and DWD) for training, which should reduce the individual bias of each of those sets by simply introducing more meteorologists in our training data. Actually, we can show that the performance of the method increases if we use training data from more than one weather service. Finally, we want to note here that even if the training data stemming from DWD does not include double fronts, a closer inspection of the movie in the SI shows that indeed the network can provide double fronts, e.g. fronts in the North Atlantic at 0:39 and 0:45, as well as in the South Atlantic at 0:30 east of North and South America. Obviously, the network is able to generalize the features even if these fronts are not explicitly included in one training data set.

2. The manuscript has a strong technical nature with only little insight into meteorology or front dynamics. I would recommend considering a transfer of this manuscript to GMD.

   While there are other papers within WCD that are of similar technical nature, we did add an application of our trained network researching the connection of extreme precipitation and our detected fronts using the ERA5 data and 1 hourly aggregated precipitation, similar to a study by Catto and Pfahl. Therefore, we think that WCD is still an appropriate choice for the manuscript.

3. The presented comparison against a second front detection method, which is based on the thermal front parameter (TFP), is odd. First, I recommend not to call it "the ETH method", because this is not known to the community and ETH is a large institution. Maybe TFP method? Second, I recommend providing more background about the TFP, which goes back to Renard and Clarke (1965). The TFP implementation by Jenkner et al (2010), which is used here as reference method, is unique, because it places the front where TFP=0, which is at the center of the frontal zone. However, this is not where most meteorologist place the front. Most center, including DWD and ECMWF, place it where MAX(TFP)=0, which is at the leading edge of the frontal zone. So, there is a mismatch. This important difference is not explained in the current manuscript and because the width of a frontal zone can easily encompass a couple of hundred kilometers, the here used "baseline method" will be due to its design in most situations do not agree with the DWD charts. Basically, a method that was trained to reproduce DWD fronts, which it does very well, is compared against a method that was intentionally designed not to agree with DWD fronts because the front line is placed in a different location. It is thus not a meaningful comparison (and this explains much of what is found in Lines 376-4079) and I recommend that the comparison is removed.

   We changed the name of the method into "baseline method". We added some information that the placement of the baseline TFP method may be offset compared to the weather services, to make it clear to the reader. However, in our CSI evaluation we do not calculate the exact matching of fronts but rather use a more soft criterion. To match fronts we use a search radius of $D = 250\,\mathrm{km}$. To be considered a match only the median distance of a detected front from the closest pixel of a provided labeled front hast to be below D. We thought that this is an already loose enough criterion to match fronts, however we additionally added another evaluation where we evaluate the baseline TFP method using $D = 500\,\mathrm{km}$ essentially doubling the search radius by adding another 250 km to the radius. Also we would like to remind you, that we did not solely use the DWD data for evaluation but also that of the NWS, i.e. from two weather services. Actually, the method is very flexible, such that additional training data sets can easily implemented.

4. More meaningful would be a comparison against another ML-based method, for example, that of Lagerquist et al. (2019), who pioneered ML-based front detection. This would be more insightful because it is not clear at this point which neural-network architecture is most suitable for front detection and why this is the case. I find Fig. 10 in Lagerquist et al. (2019) very helpful. A similar figure plus a direct comparison of these two ML-based methods would thus be of high merit.

   We tried using the method, however we could not get it to run in a feasible time. Additionally the method is executed and evaluated on a different grid and spacing which makes a direct comparison inaccurate at least.

5. It is not advisable to transfer an automated front detection from a low-resolution grid to a high-resolution grid without intensive retuning and testing. How was this retuning done? By how much was the detection thresholds for the front gradient increased? By how much was the minimum length criterion changed? Did the authors increase the minimum advection speed to separate stationary from non-stationary fronts? A method developed for a ERA-Interim 1x1 degree grid (or for a 2-km grid as in Jenkner et al. 2010) should not be transferred to another grid spacing. Further, does DWD use a minimum front length and are the authors using the same threshold? At the same time, while the front threshold presumably was increased when going from a 1x1 degree grid to a 0.5x0.5-degree grid, the number of fronts ideally should not change (see Fig. 2 in Thomas and Schultz (2019b) on the dependency to the threshold). More details on the retuning that was done when preparing the comparison is needed in this manuscript.

   For the tuning of the algorithm we tested different settings for the number of parameters in the algorithm, namely the minimum advection speed ($3\,\mathrm{m\,s^{-1}}$ to $6\,\mathrm{m\,s^{-1}}$), the minimum temperature gradient ($4 \times 10^{-2}\,\mathrm{K\,km^{-1}}$ to $5 \times 10^{-2}\,\mathrm{K\,km^{-1}}$), the minimum front length (500 km to 700 km), the number of gridpoints a front object has to contain (15 to 50), the value of a smoothing parameter for frontal lines (5 to 30), the gap allowed between two segments of large THE gradient from them to be considered as one front object (5 km to 100 km), and the number of times a digital filter is applied to the equivalent potential temperature gradient field (5 to 10). For the tuning exercise we considered the three month

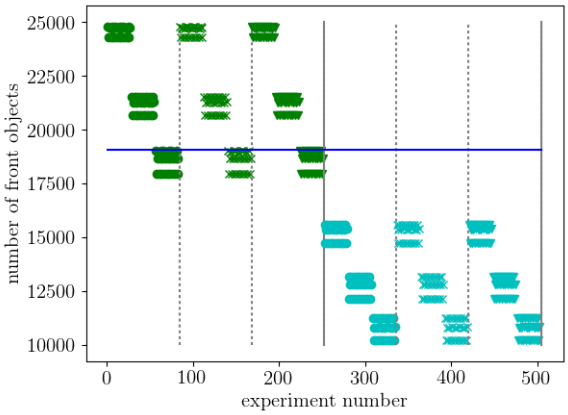

Figure 1: Number of front objects identified in the northern and southern extra-tropics from the ERA5 data-set for different choices of the free parameters. Green symbols correspond to climatologies computed with a minimum THE gradient of $4 \times 10^{-2}\,\mathrm{K\,km^{-1}}$ and cyan symbols to those with a minimum THE gradient of $5 \times 10^{-2}\,\mathrm{K\,km^{-1}}$. Within each group experiments are grouped according to minimum advection velocities of $3\,\mathrm{m\,s^{-1}}$, $4\,\mathrm{m\,s^{-1}}$ and $5\,\mathrm{m\,s^{-1}}$ (groups of points separated by dashed lines) and minimum front length of 500 km, 600 km and 700 km (groups of points between the dashed lines). The blue horizontal line shows the number of fronts detected with the original algorithm in the ERA-Interim data-set.

of December 2013, January 2014, and February 2014, for which both the ERA-Interim and the ERA5 reanalysis are available. For the three month in total 1947 different climatologies were computed and compared by considering the number of fronts detected in the extratropics ($-60°$ to $-30°$ N and $30°$ to $-60°$ N), the geographical location, the length of fronts detected, and a visual inspection of individual cases from similar performing parameter combinations.
In the original version of the algorithm those parameters, i.e. that applied to the ERA-Interim data at 1° grid spacing, are set to $3\,\mathrm{m\,s^{-1}}$, $4 \times 10^{-2}\,\mathrm{K\,km^{-1}}$, 500 km, 15, 5, 100 km, and 5. We decided to keep as much of the physical values, which are determined in units of $\mathrm{km^{-1}}$ or km identical to the original algorithm. The motivation is to retain similar physical properties of the front. Except from smoothing of gradients on a lower resolution grid, it is expected that the definition of the parameters in units of km or $\mathrm{km^{-1}}$ already reflects the change in grid resolution. However, we adjusted the number of the minimum number of grid-points in a front object (changed from 15 to 20), the number of filter applications (changed from 5 to 7), and the smoothing parameter (changed from 5 to 15) to reflect the larger spatial resolution of the ERA5 data-set (here: 0.5°). These parameters determine the smoothing of the equivalent potential temperature gradient field and the "straigthening" of frontal lines and hence we deem them more appropriate for adjustment to different grid spacings.
Fig. 1 shows the number of fronts detected in the northern and southern extra-tropics for different experiments. In all the shown experiments the filter is applied 7 times to the equivalent potential temperature gradient field, as fewer applications lead to large increase in the number of detected fronts and a shift towards short and disconnected frontal features and a more applications to a large decrease in the number of detected fronts. The number of fronts in the ERA5 data-set with the re-tuned algorithm the number of detected fronts is about 30 % larger than in the original data-set, which could be remedied by increasing the minimum length of fronts to 700 km or increasing the minimum potential temperature gradient to a value between $4 \times 10^{-2}\,\mathrm{K\,km^{-1}}$ and $5 \times 10^{-2}\,\mathrm{K\,km^{-1}}$. Considering also the distribution of front length (not shown) and the spatial distribution of front occurrence (see Fig. 2), we decided to not change the values of minimum front length and the minimum potential temperature gradient as both do not yield a benefit in terms of the spatial front distribution.
The information, which parameters of the front detection algorithm have been adjusted has been included in the revised version of the manuscript.

6. I was disappointed to see an equivalent-potential temperature based front definition purposefully applied to latitudes outside of midlatitudes. Front detection methods based on equivalent-potential temperature (called TH in the next statement) are well known to be unsafe for usage outside of the midlatitudes, for example, Schemm et al. (2015, p. 1696) noted: "... clearly indicates that the TH method is influenced by semi-permanent convergence zones and tropical convection (although a minimum advection threshold is applied). Tropical features which, from a synoptic viewpoint, would not be regarded as a 'front', are

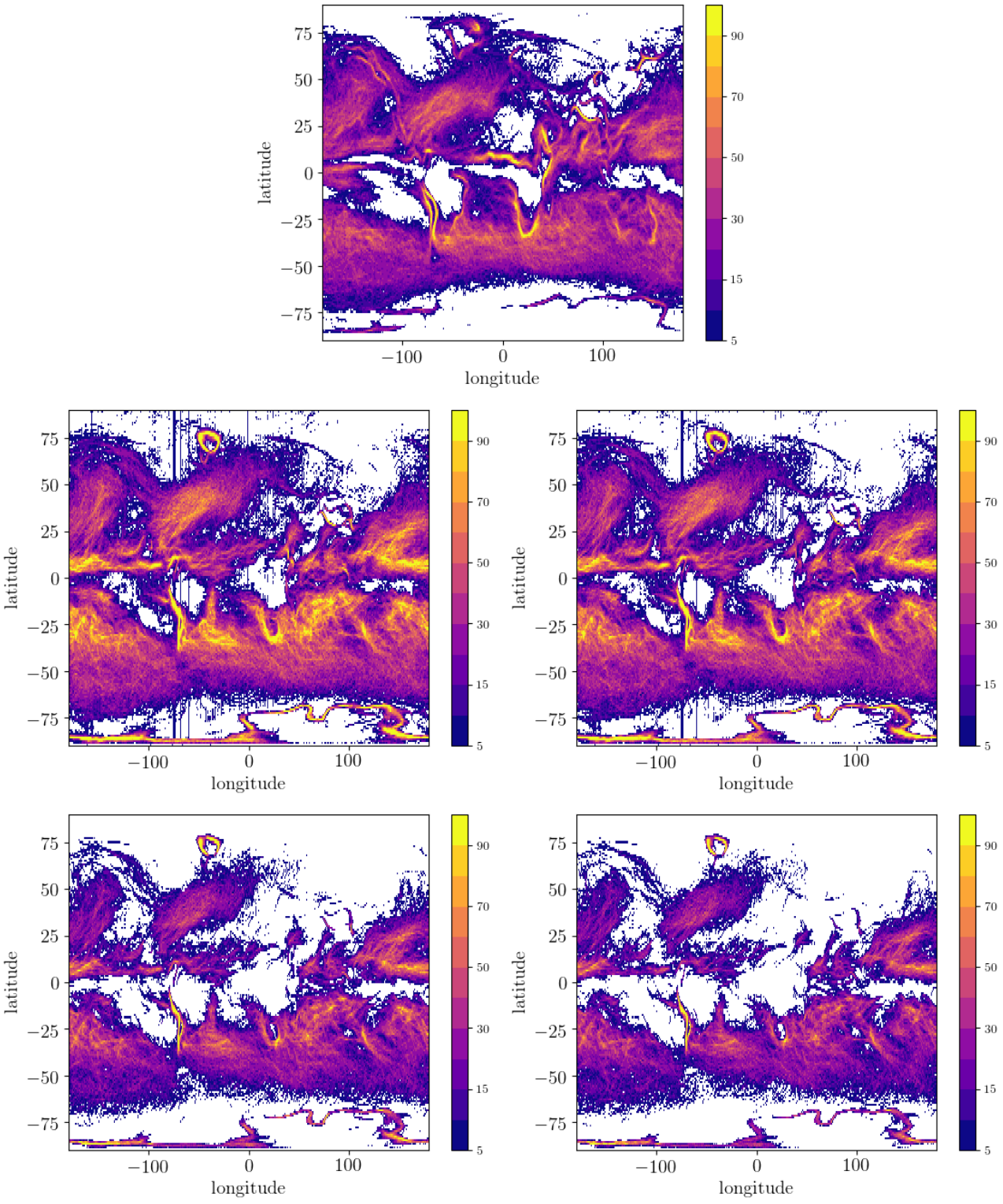

Figure 2: Number of times a front is detected at a specific grid-point in the ERA-Interim data-set (top row) and the ERA5 data-set using different minimum THE gradients (middle row: $4 \times 10^{-2}\,\mathrm{K\,km^{-1}}$, bottom row: $5 \times 10^{-2}\,\mathrm{K\,km^{-1}}$) and different minimum frontal length (left: 500 km, right: 700 km).

identified as such. Accordingly, TH methods should be used with care if applied outside midlatitudes", which is a nice way to say that it should not be done. Further they note "... as the θe gradient can be dominated solely by moisture gradients, especially in tropical latitudes, this results in the detection of several quasi-stationary fronts (which form along mountain crests, or in association with land – sea contrasts) which must be removed in a post-processing step" (Schemm et al. 2015, p.1687). Against these recommendations the authors decided to apply the θe method to subtropical and tropical latitudes and afterward, not too surprisingly, conclude that it detects numerous of non-cyclone related fronts. What was the intention behind this? The section between L.460-470 is therefore misleading.

We understand that the used baseline method is not suited for the application outside the midlatitudes. As a result we will restrict our quantitative evaluation to the midlatitudes, as a fair way to evaluate the performance. But we still believe that an a qualitative evaluation outside the midlatitudes is of interest to highlight the differences in how the network performs in the regions where the baseline method struggles / should not be used. We do agree with you that we should make it more clear that the baseline is not designed for these regions. To do this we added some text when describing the baseline as well as adding a gray overlay to the climatology images to indicate this. Finally, we want to add that although we are aware of the restriction of front detection methods for the tropics, we added the results for the tropics about the connection between extreme precipitation and fronts. This inclusion was also motivated by the results of Catto and Pfahl (2013), who also applied a TFP mehod in the tropics for their comparisons.

7. The conclusion is short, with only a technical statement and an outlook but no conclusion related to weather and climate dynamics. Maybe you could try to conclude on how and why the ML-based method is able to distinguish mobile from stationary fronts (such as those along the coastlines or mountains), which would yield additional process understanding and it is a mayor struggle for traditional TFP-based methods.

   We reordered some sections, put the discussion of the results into section 3, deleted the discussion section, and added some more text to the conclusion. We added another chapter regarding the connection between our detected fronts and extreme precipitation. A direct explanation of such a deep learning architecture is hard and it is a very current field of research, to get reliable conclusive information from a neural network. Showing the physical cross-sections indicates that the algorithm respects the wind speed close to stationary fronts to detect those, seeing how it is far lower than the other types of fronts. However it is hard to detect whether this is causation or simply correlation.

Minor comments:

1. L. 19 "much of the literature is on the larger-scale fronts" – research on mesoscale fronts is a very active field of research as well.

   We added some text into this direction

2. L. 15 "are a vital part of the communication of weather to the public and the public perception of weather in general" – Most people use Apps; fronts are no longer a major part of modern weather communication.

   We added a comment on this

3. L. 27: "The former methodology goes back to the work by Hewson (1998)" – I guess it goes back to Renard and Clarke (1965).

   We added the reference and some text

References:

• Lagerquist, Ryan, Allen, John T., and McGovern, Amy, 2020, "Climatology and Variability of Warm and Cold Fronts over North America from 1979 to 2018" Journal of Climate Vol. 33, No. 15, 1520-0442

• Lagerquist, R., A. McGovern, and D. Gagne II, 2019: Deep learning for spatially explicit prediction of synoptic-scale fronts. Wea. Forecasting, 34, 1137–1160.

• Sanders, F., 1999: A proposed method of surface map analysis. Mon. Wea. Rev., 127, 945–955

• Schemm, S., Sprenger, M., & Wernli, H. (2018). When during Their Life Cycle Are Extratropical Cyclones Attended by Fronts?, Bulletin of the American Meteorological Society, 99(1), 149-165.

• Thomas, Carl M. and Schultz, David M., 2019, "Global Climatologies of Fronts, Airmass Boundaries, and Airstream Boundaries: Why the Definition of "Front" Matters" Monthly Weather Review Vol. 147, No. 2, pp 691, 1520-0493

• Thomas, Carl M. and Schultz, David M., 2019, "What are the Best Thermodynamic Quantity and Function to Define a Front in Gridded Model Output?" Bulletin of the American Meteorological Society Vol. 100, No. 5, pp 873, 1520-0

• Uccellini, L. W., S. F. Corfidi, N. W. Junker, P. J. Kocin, and D. A. Olson, 1992: Report on the surface analysis workshop at the National Meteorological Center 25–28 March 1991. Bull. Amer. Meteor. Soc., 73, 459–471.

We included most reference into the text