https://doi.org/10.5194/wcd-2021-27
Preprint. Discussion started: 25 May 2021

Weather and Climate Dynamics
Discussions
Open Access
EGU

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

Stefan Niebler[1], Annette Miltenberger[2], Bertil Schmidt[1], and Peter Spichtinger[2]

[1]Institut für Informatik, Johannes Gutenberg Universität Mainz, Staudingerweg 7, 55128 Mainz, Germany
[2]Institut für Physik der Atmosphäre, Johannes Gutenberg Universität Mainz, Becherweg 21, 55128 Mainz, Germany

Correspondence: Stefan Niebler (stnieble@uni-mainz.de)

Abstract. Automatic determination of fronts from atmospheric data is an important task for weather prediction. In this paper we introduce a deep neural network to detect and classify fronts from multi-level ERA5 reanalysis data. Model training and prediction is evaluated using two different regions covering Europe and North America. We apply label deformation within our loss function which removes the need for skeleton operations or other complicated post processing steps as observed in other work, to create the final output. We observe good prediction scores with CSI higher than 62.9% and a Object Detection Rate of more than 73%. Frontal climatologies of our network are highly correlated (greater than 79.6%) to climatologies created from weather service data. Evaluated cross sections further show that our networks classification is physical plausible. Comparison with a well-established baseline method (ETH Zurich) shows a better performance of our network classification.

1 Introduction

Atmospheric fronts are ubiquitous structural elements of extratropical weather. The term *front* refers to a narrow transition region between airmasses of different density and/or temperature. These airmass boundaries play an important role for understanding the dynamics of midlatitude weather. Fronts are often associated with significant weather, such as intense precipitation and high gust speeds. Hence, fronts in the sense of separating polar from more subtropical airmasses are a vital part of the communication of weather to the public and the public perception of weather in general. Frontal surfaces exist also on smaller scales, e.g. in the context of sea-breeze circulation or local circulation patterns in mountainous regions. However, the focus here and in much of the literature is on the larger-scale fronts that can extend over several hundred kilometres and are often associated with extra-tropical cyclones. Quasi-stationary fronts exist that can extend over large distance and do typically not move strongly over time, e.g. the Mei-Yu front. These stationary fronts are as well foci of significant surface weather.

Determining the position and propagation of surface fronts plays an important role for weather forecasting. The traditional manual approach to front detection is based on the expertise of weather analysts at operational meteorological services, along some (even empirical) guidelines. With the advent of large, gridded reference data-sets, e.g. ERA-40 reanalysis, in the second half of the past century the drive for objective means to detect fronts automatically set in. Currently used methods are typically relying on detecting strong gradients in either temperature and humidity fields (e.g., by using equivalent potential temperature or wet-bulb temperature) or in wind fields. The former methodology goes back to the work by Hewson (1998), who suggested an automatic method to detect fronts in fairly coarse data sets based on the so-called "thermal front parameter". In his and subsequent studies this is often related to the second spatial derivative of the temperature, and one or more "masking parameters", i.e. thresholds of thermal gradients along the front or in adjacent regions. This or conceptually similar methods have been used in numerous studies to determine the global or regional climatological distribution of fronts (e.g. Berry et al., 2011; Jenkner et al., 2010).

For the investigation of fronts on the southern hemisphere Simmonds et al. (2012) suggested an alternative approach that investigates the Eulerian time rate of change of wind direction and speed in the lower troposphere at a given location. A comparison of the two methods to identify fronts on a global climatological scale by Schemm et al. (2015) revealed some agreement between the fronts detected, but also regional difference and systematic biases in the detection of certain front types by both algorithms: For example, the "thermal" method detects more reliably warm fronts than the method based on lower tropospheric wind speed and direction. In addition, the orientation of detected fronts differs in general between the two methods. In consequence Schemm et al. (2015) also find differences in the global distribution of fronts and the amplitude of seasonal variations in front occurrence frequency.

While it is well known that different front detection methods provide different outputs (e.g. Schemm et al., 2015; Hope et al., 2014), an objective ground-truth is difficult to find. Most studies developing or testing automatic detection schemes rely on manual analysis as the "gold standard" to test the accuracy and for tuning free parameters in the automatic detection schemes (e.g., Hewson, 1998; Berry et al., 2011; Bitsa et al., 2019). However, it should be noted that manual analysis is affected to a large degree by subjective decisions, and hence the focus, interest and expertise, of the person conducting the analysis. Shakina (2014) reports results from an inter-comparison study of different manual front analysis carried out independently in different divisions of the Russian Meteorological service up until the 1990s. Comparing the different archives agreement on the presence or absence of a front in any one 2.5° × 2.5° box was found in 84.8 % of cases. However, if only the presence of fronts in any one grid box is considered agreement dropped to 23 % to 30 % depending on the type of front. Shakina (2014) further suggests that disagreement mainly arises from the detection and positioning of secondary or occluded fronts which typicall are associated with less marked changes in surface weather. It is likely that the differences between manual analysis by different forecasters in the meantime have not reduced, but they may potentially be reduced by strict guidelines for forecasters on the key decision features for positioning fronts.

Despite a none negligible subjectivity of manual analysis, it still offers many advantages over automatic methods:

1. In contrast to most automatic detection methods many different aspects, including temperature, wind, and humidity fields, surface pressure, but also surface precipitation and wind, are taken into account.

2. Manual analysis does not rely strongly on the choice of (arbitrary) thresholds that are needed in most automatic front detection algorithms.

3. Experience of analysts can be taken into account, especially on regional scales (e.g. with complicated terrain as in the Alps)

In order to address the over-reliance on specific variables some recent studies have suggested methods that combine not only temperature and humidity data but also include information on the wind field (e.g. Ribeiro et al., 2016; Parfitt et al., 2017), or information on Eulerian changes in mean sea-level pressure (e.g. Foss et al., 2017). Nevertheless these extended algorithms that are so far mainly used in regional studies still rely on choosing appropriate thresholds for the magnitude of thermal gradients or changes in the wind direction and speed.

The necessity of manually designing metrics and selecting thresholds for automatic front detection can be at least partly overcome by employing statistical methods and machine learning approaches. The key idea with this approach is that based on manual analysis a complex statistical method retrieves as much consistent information on patterns, important variables, and thresholds as is available in manual analyses and coinciding state of the atmosphere, e.g. from reanalysis data-sets. Previous attempts on using machine learning approaches for front detection are discussed in more detail the following section. The overall aim of our paper is to investigate the degree to which machine learning approaches are able to replicate manual analysis on a case-study and climatological scale and the degree to which the learned features are consistent with meteorological expectations on the physical properties characterising a frontal surface.

Recently different groups have used Artificial Neural Networks (ANNs) to predict frontal lines from atmospheric data. Biard and Kunkel (2019) used the MERRA-2 data-set to predict and classify fronts over the North American continent. Their network also classifies their predicted fronts using the four types: warm, cold, stationary, and occlusions. They used labels provided by the North American weather service (NWS).

Lagerquist et al. (2019) used the North American Regional Reanalysis (NARR) data-set Mesinger et al. (2006), to predict synoptic cold and warm fronts over the North American continent also using the NWS labels. While the network of Biard and Kunkel (2019) creates an output on the input domain, the network of Lagerquist et al. (2019) predicts the probability for a single pixel and needs to be applied to each pixel consecutively. Both methods rely on postprocessing steps like morphological thinning to create their final representation of frontal data. In their evaluation they used an object based evaluation method, which we also adapt. Additionally, both methods only use a 2D mask for each input variable not making use of multiple pressure or height levels. Matsuoka et al. (2019) on the other hand used a U-Net architecture (Ronneberger et al., 2015) to predict stationary fronts located near Japan. Our provided network uses a more sophisticated U-Net approach to predict and classify all four types of fronts, without the need of morphological post processing. Additionally we evaluate our approach similar to Lagerquist et al. (2019) using an object based evaluation method. Unlike the previous methods we incorporate data from two different weather services, the NWS and the German Weather Service (Deutscher Wetterdienst, DWD) and also evaluate on both regions. We additionally compare our predicted fronts against the method developed by Schemm et al. (2015) as baseline. We use the ERA5 reanalysis data (Hersbach et al., 2020) from the European Centre for Medium-Range Weather

Forecasts (ECMWF) at a $0.25°$ grid at multiple pressure levels for each variable. This data-set exhibits a higher resolution than the NARR ($32\,\mathrm{km}$ grid) and used MERRA-2 data-set by Biard and Kunkel (2019) ($1°$ grid). Additionally, we used multiple pressure levels to refine our results.

In Section 2 we will describe our used network architecture, data and evaluation methods. In Section 3 we explain our evaluation methods and display our evaluation results on the training and test data set, before the discussing these results in section 4. Section 5 provides a short outlook for future improvements.

2 Materials and Methods

For each spatial grid point our proposed algorithm predicts a probability distribution, describing how likely it is that the point belongs to each of our possible five classes: warm, cold, occlusion, stationary, or background. Our method predicts that estimate from a 4-dimensional input consisting of multiple channels located on a 3-dimensional multilevel geospatial grid, which was flattened to a 3-dimensional input by combining the atmospheric channel and level dimension. For this task we use a convolutional neural network (CNN) architecture to automatically learn atmospheric features that correspond to the existence of a weather front at spatial grid points. We use a supervised learning approach, in which we provide ground truth data of frontal data sampled from two different weather services. We adjust hidden parameters of the CNN in order to optimize a loss function measuring the quality of our weather front prediction. CNN architecture and training will be explained in further detail in this section. Our network was implemented, trained, and tested using Pytorch 1.6 (Paszke et al., 2019). Parallel Multi-GPU training was implemented using Pytorch's DistributedParallel package. The provided code was run using Python 3.8.2.

2.1 Data

We will briefly describe which channels and gridpoints were used as training input from the ERA5 reanalysis data (Hersbach et al., 2020). Furthermore, we will describe the format of the corresponding label data and in the case of the DWD label data, how it was generated.

2.1.1 ERA5 Reanalysis Data

Our model input consists of a multichannel multilevel spatial grid provided by ECMWFs ERA5 reanalysis data-set. Each channel denotes a different atmospheric variable, while levels consist of a subset taken from the $L137$ level definition (ECMWF, 2021). Data is represented on a spatial grid with a grid-spacing of $0.25°$ in both latitudinal and longitudinal direction. Since we do not expect to obtain relevant information from high altitude level data, we decided to restrict ourselves to levels within the inclusive interval $[105, 137]$, representing pressure levels between surface pressure and about $700\,\mathrm{hPa}$. This range contains both the ground level information as well as the $850\,\mathrm{hPa}$ pressure level information, both of which are commonly used to detect fronts. As the pressure levels are defined as parameters of an affine transformation of the surface level pressure, we added the surface level pressure as an extra channel to the data, to calculate the exact pressure values of each level.

Table 1. Mean and variance of the individual variables used for normalization of input data.

variable	(unit)	mean	variance
t	K	2.75355461e+02	3.20404803e+02
q	kg kg^{-1}	5.57926815e-03	2.72627785e-05
u	m s^{-1}	1.27024432	6.74232481e+01
v	m s^{-1}	1.0213897e-01	4.36244384e+01
w	Pa s^{-1}	5.87718196e-03	4.77972548e-02
sp	hPa	8.65211548e+04	1.49460630e+08
kmPerLon	km/°	0.64	0.09

For the actual training we restrict ourselves to 5 multilevel variables: temperature (t), specific humidity (q), zonal wind velocity (u, East-West), meridional wind velocity (v, North-South), and vertical velocity (w), respectively. In addition the surface pressure (sp) and longitudinal distance per pixel in km ($kmPerLon$) are considered. The distance between two pixel at a certain degree latitude is derived by assuming a spherical shape of the globe, while surface pressure was added to our data using the merge operation of the Climate Data Operators (CDO) (Schulzweida, 2019). ERA5 data is normalized with respect to a global mean and variance sampled from data of the year 2016. The resulting mean and variance values are listed in Table 1.

2.1.2 NWS Front Label Data

For training on the North American continent we use the HiRes Coded-Surface-Bulletins (csb) of the North American National Weather Service (National Weather Service, 2019). The latter data ranges from 2003 up to 2018. It was previously used by Biard and Kunkel (2019) and Lagerquist et al. (2019). Each front in a csb file consists of an identifier, describing the type of front, followed by a series of coordinate pairs, defining a polyline of the front. We do not perform any pre-processing on this data. In accordance with our available data we restricted the use of the latter to the years 2012 - 2017 using only snapshots in a 6-hour interval to keep the amount of data balanced compared to the DWD data during training.

2.1.3 DWD Front Label Data

For training over Europe and the Northern Atlantic we use label data extracted from the surface analysis maps of the Deutscher Wetterdienst (DWD) for the years 2015 to 2019. Unlike the Coded-Surface-Bulletins, these maps are not provided as polylines, but rather as a PNG images of a region containing both the North Atlantic as well as Western Europe. For an example of such an image see Fig. 1 (left panel). To use the labels we extract each individual front, by creating coordinate pairs, which describe the front as a polyline, similar to csb.

The depicted fronts are color coded within an image, which allows us to easily separate them from the background. We do not need information about the direction of a front. Thus, we also remove the symbolic identifiers like half-circles and triangles,

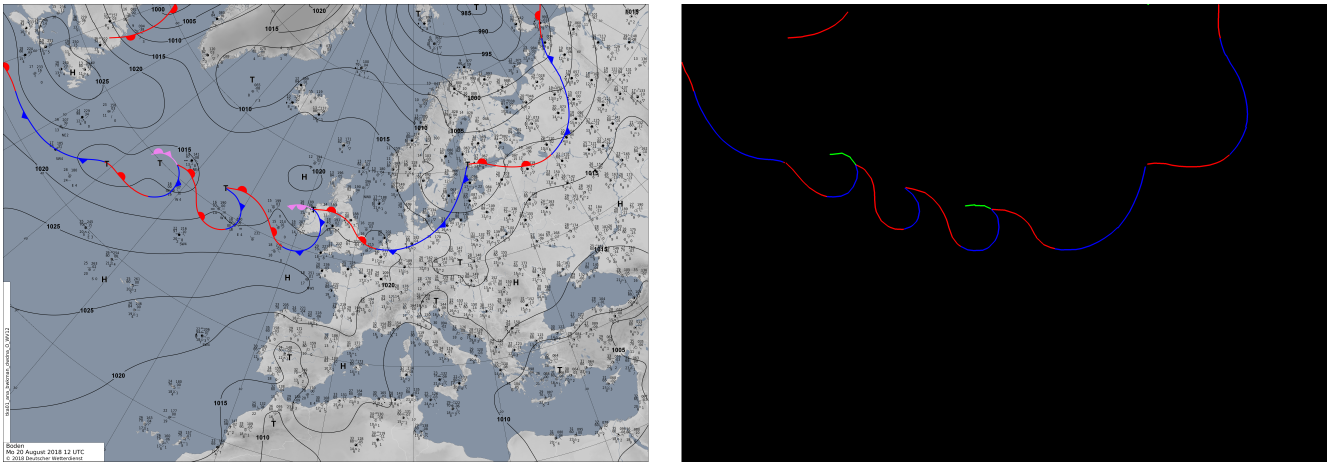

Figure 1. Example of well extracted fronts. Blue and red lines correspond to cold and warm fronts respectively as in the original images. Green lines correspond to the occlusions which are pink in the input image. Note that stationary fronts are originally depicted as alternating warm and cold fronts. For this reason we cannot distinguish those from regular cold and warm fronts.

indicating the directions of a front. Otherwise, these symbols could create false positive coordinate points in the label data. Our algorithm therefore first filters all fronts of a specific type by filtering all pixel of the corresponding color. In a second step we erase all additional symbols on each line. Subsequently, latitude and longitude coordinate pairs along each line are extracted in order to describe each front in terms of a polyline. In Fig. 1 (right panel) we show an example of a processed image file, redrawn onto the same projection as the input image. Blue and red lines in both panels correspond to cold and warm fronts respectively, while green lines correspond to occlusions, which are pink in the left panel.

In certain cases our method fails to correctly extract the frontal lines. Some of these cases, like small gaps within a front, can be ignored as the lower resolution of the ERA5 grid masks them anyways. However, there are cases with larger gaps, wrongly extracted objects or wrongly connected fronts. Gaps originate from two factors. One is that another object is drawn on top of a frontal line, effectively splitting the gap into two parts. The other cause of a gap is an odd placement of the frontal-symbols where an aggregation of multiple symbols on a short segment occurs. As our method removes sections where a symbol is placed before reconnecting the remaining points, crowded placement of these symbols may make the remaining part of the front too short to be considered relevant and as such will be omitted. Wrongly extracted objects occur mostly with storms that are depicted in the same color as a warm front. As such our extraction method wrongly extracts these objects as well. The last cause of error occurs when we try to sort the extracted coordinate pairs of a single front. In some cases the sorting method may end up stuck in a local minimum, resulting in a wrong order of points. While it may be possible to fix some of these errors by preprocessing the original images provided by the DWD we instead chose to completely remove faulty images from our data-sets. An example of such a faulty extracted image is shown in Fig. 2.

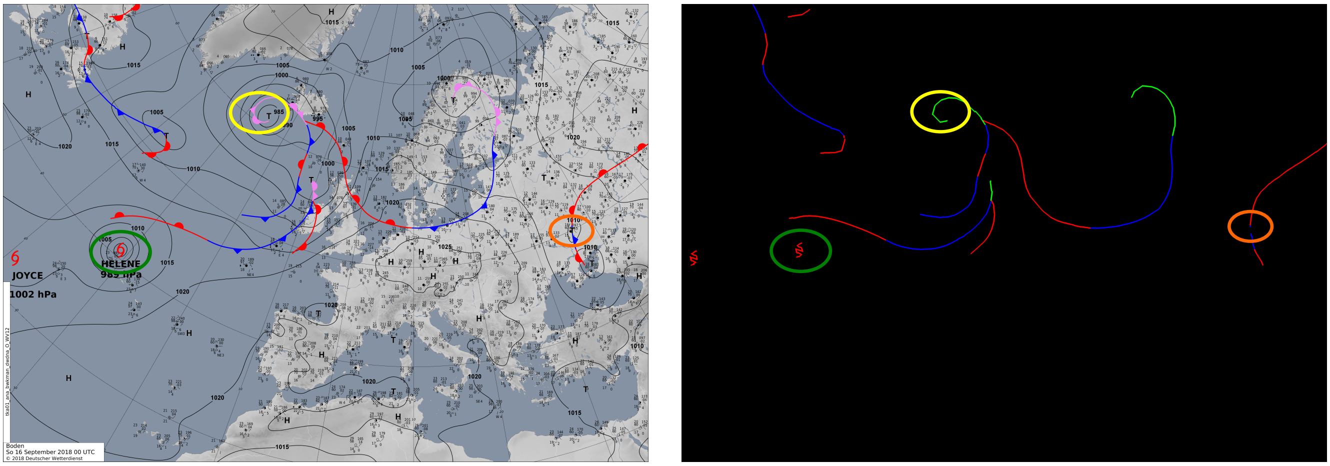

Figure 2. Example of badly extracted fronts. **Green Circle:** Object that is not a front, but has the same color coding is wrongly extracted as a front. **Orange Circle:** Unrelated symbol is drawn over the front. The front could not be extracted completely. **Yellow Circle:** Frontal symbol is placed in an area with high curvature. The curvature is not extracted exactly, as the symbol is removed during the procedure and the lose ends are connected with a straight line.

2.2 Network Design and Training

2.2.1 Network Architecture

Neural networks are a machine learning technique where a network consisting of several layers is used to extract feature representations of an input at different levels. Each layer transforms its input into an output map, the layers feature map. These feature maps can then be used as an input for consecutive layers which enables the network to learn more detailed features within the data. In a Convolutional Neural Network (CNN) the most common transformation function is a convolution of the input image with a convolution mask where each entry is a trainable, latent parameter of the network. During training these parameters are adjusted to optimize a loss function, which measures the quality of the output of the network. In our case we use a U-Net Architecture originally introduced by Ronneberger et al. (2015) for biomedical segmentation. The proposed architecture consists of several consecutive blocks that gradually extract features from the data and reduce the spatial dimension of the input data to extract features on multiple scales. These blocks are followed by a number of expansive blocks which gradually increase the resolution up to the original scale. Additionally at each resolution scale a so called skip connection allows the final feature map of a encoding block to directly serve as additional input to the corresponding decoding block. These skips improve the networks ability to localize the features, as the upsampled features only hold coarse localization information. In our network we use convolutional layers as explained before. Additionally we use Rectified Linear Unit (ReLU), Batch Normalization, Pooling and upsampling layers, whose functionality we will briefly explain.

- ReLU layers are used to introduce non linearity into the network. They transform each input x as $ReLU(x) = max(0,x)$

https://doi.org/10.5194/wcd-2021-27
Preprint. Discussion started: 25 May 2021

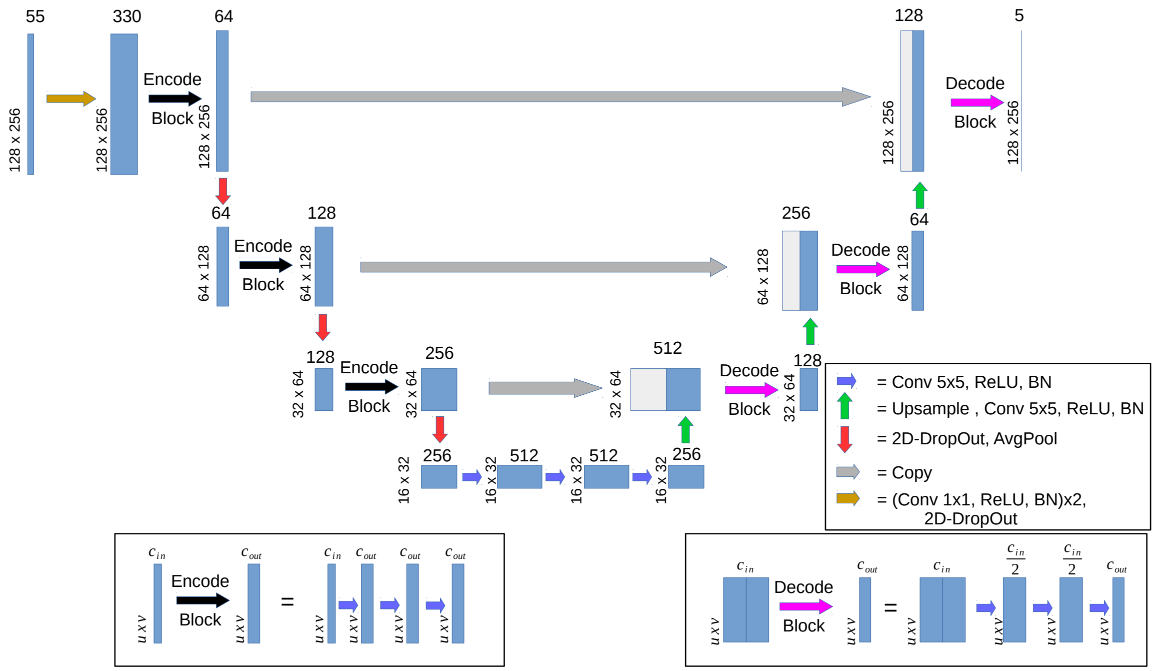

Figure 3. U-Net architecture used for this paper. The first convolution of the input data uses a 1 × 1 sized kernel instead of 5 × 5.

- Batch Normalization layers normalize the batched input to a mean of 0 and variance of 1. They can have additional learnable affine parameter.
- Pooling layer transform several input grid points to a single output gridpoint. Common operations are averagePooling or maxPooling where the grid points are combined calculating the average or maximum of the input respectively. This operation is used to reduce the resolution of the feature map.
- Upsample layers are a simple upsampling of a grid point to increase the resolution of the feature map.

A sketch of the used architecture is shown in Fig. 3

We use Pytorch's DistributedParallel package to enable training on multiple GPUs in parallel. Training is performed on a single node, with each GPU acting on a fixed shard of the available data.

2.2.2 Data-Set Augmentation

ERA5 reanalysis data is available for the whole globe. Our labeled ground truth data however resides in the analysis regions of the corresponding weather services. Therefore, we restrict ourselves to these subset regions of the ERA5 data-set for training.

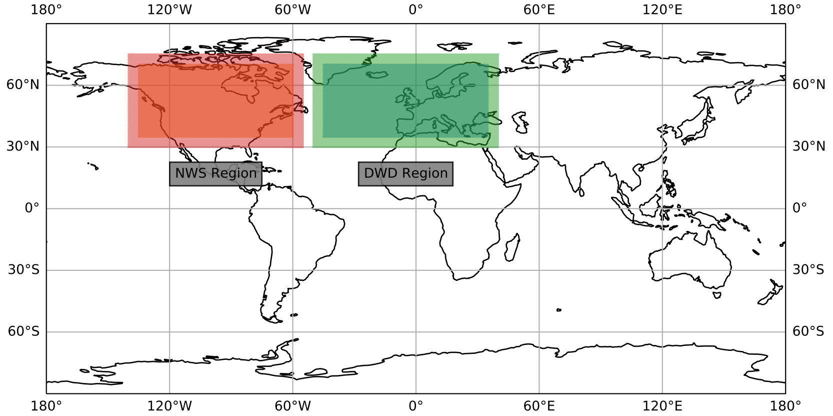

Figure 4. Bounding Boxes for the two regions used for training. The brighter area can be used as input, but is never evaluated

We further chose to restrict ourselves to a rectangular subgrid which is completely within the analysis region of the weather service. For the DWD data-set we restrict ourselves to the region ranging from 35° to 70° north and $-40°$ to 35° east. For the NWS data-set we use the area between 35° to 70° degree north and $-135°$ to $-60°$ degree east. These regions are depicted in Fig. 4 with a darker shade. During training our network ignores the outer 20 pixel (5°) of the input at each border to prevent the case where some fronts may have insufficient data due to image cropping. With respect to this, we can additionally use the depicted brighter shaded area to generate our input, as the network output is still only evaluated within the darker shaded area. Due to hardware constraints on our available GPU hardware we further restrict ourselves to only 9 pressure levels of the data-set. The resulting dimensions for each data and the file information are listed in Table 2.

During training we randomly select from which weather service we pick our labels, if both labels are available. If only a single label is available it is always chosen for that data. We crop a 128×256 pixel sized sub-grid residing within the corresponding weather services analysis region (including the 5° border) as described above from the ERA5 data. The same crop is applied to the label data by removing each vertex, where neither the vertex itself nor a neighbor vertex is located within the crop region. We applied a random horizontal and vertical flip as data augmentation to further increase sample count for our data set. It is important to note that, whenever data is horizontally flipped the sign of the input variable v has to be flipped as well.

For vertical flips the same has to be applied to the u input variable, as these variables describe a vector field rather than a stationary value. Flipping of the data might also lead to a better representation of fronts in the Southern Hemisphere, which seem to be mirrored at the equator. We added random dropouts into each of our encoding layers with a 0.2 dropout rate in order to reduce overfitting.

Table 2. The used evaluation region for each weather service region and the global region. For each weather service region an additional 5° border is added to not reduce the size of the evaluation region. For the global region this border is included within the mentioned range. The used data files contain the global region for all used variables except surface pressure and latitude where the global data is calculated from a single level slice (sp) or broadcast from the extracted latitudes (latitude). The row file shows the dimensions of the datafile where we extract our data from during training and evaluation. The files contain a higher resolution of levels than we use in this work.

Weather Service	Latitudes	Longitudes	Levels	#Voxel
DWD	$[35^\circ N, 70^\circ N]$	$[-40^\circ E, 35^\circ E]$	$[105, 137, 4]$	$140 \times 300 \times 9$
NA	$[35^\circ N, 70^\circ N]$	$[-135^\circ E, -60^\circ E]$	$[105, 137, 4]$	$140 \times 300 \times 9$
Global	$]-90^\circ N, 90^\circ N]$	$[-180^\circ E, 180^\circ E[$	$[105, 137, 4]$	$720 \times 1440 \times 9$

2.2.3 Training

Our model is trained using stochastic gradient descent with Nesterov momentum of 0.9 to minimize the loss function. The initial learning rate is set to $0.005 \cdot \#Ranks$, where $\#Ranks$ corresponds to the number of processes used for the parallel training. We train the network for several epochs. Within each epoch the algorithms randomly trains on a permutation of the complete training data set. Every 10 epochs we measure the training loss. If the test loss does not improve for 10 test phases we divide the learning rate by 10 up to a minimum of $1e-7$ and reset the count, if the learning rate was changed. If the test loss does not improve for 20 test phases (200 epochs) and we cannot reduce the learning rate anymore we stop training. Additionally we set a maximum of 10000 training epochs or 3 days time as stopping criteria. At each test step, we save a snapshot of the network if the test loss is better than the currently best test loss. Our final network is the resulting network which yielded the lowest test error.

2.2.4 Loss and Evaluation

As described by Lagerquist et al. (2019) the frontal polylines are subject to two non-negligible causes of bias: inter- and intra-meteorologist. The first bias describes the effect that two meteorologists may consider the exact location of a front at different pixels, the occurrence of a front at all, or which exact shape the frontal curve follows. The second bias describes the effect that the same meteorologist may have a bias on the placement of frontal data coming from previously placed fronts by the same person. This is due to the fact that at subsequent forecast analyses different persons carry out the analysis. The transformation of these curves into poly-lines and the application onto a different resolution is subject to creating additional label displacements. While these problems are present in most human labeled data it is more peculiar in this specific case because the ideal poly-line shows a width of only *a single pixel*. As a result each ever so slight displacement introduces a large per pixel disparity between two fronts, as the intersection of the sets of pixels that describe these fronts ends up being close to non existent. As an example consider the frontal line depicted in Fig. 5 a. Predicting the same front with a one pixel displacement to the right, would lead to zero intersecting pixels. The same result would be achieved by simply not predicting any front at all. However, qualitatively one would clearly consider the first case a better prediction than the latter. This especially holds true if we consider that the provided label itself might be displaced due to one of the biases mentioned before. This has at least two negative effects. First, the gradient information is really sparse, as a close prediction will be considered false positive just as a far off prediction. Additionally, the label offset may lead to the case that a labeled front is not located at the physical frontal position, essentially creating a false label with wrong underlying atmospheric properties. One way to handle this might be to enlarge the extracted front label, such that they are more likely to cover the correct location. A possible approach for this is shown in Fig. 5 b. While it still introduces false positive labels the penalty for the prediction of misplaced labels is less pronounced. From our studies and the results of previous studies (e.g., Matsuoka et al., 2019; Lagerquist et al., 2019; Biard and Kunkel, 2019) it seems apparent that a deep learning architecture learns this bias in label placement and as a results predicts enlarged lines, exhibiting a larger width than the provided label. When using enlarged labels this effect is further enhanced, creating even larger predictions. To regain positional accuracy previous work used a post-processing step to extract thin lines from wider network predictions.

We decided to take another approach, where instead of inflating the provided labels we allow the labels to be slightly deformed before evaluation. This approach inherently covers the uncertainty of the label placement, as it allows for changing the label placement if necessary. At the same time the labels remain polylines, preventing inflation of the predicted lines. For our implementation we used this approach. For each front consisting of N vertices, we deform the front by extracting each vertex v_n, $0 \leq n < N$. For each v_n we calculate the pixel position $(u,v)_n$ within the image domain and then extract a $k \times k$ grid g_n centered at $(u,v)_n$. This grid describes all considered possible locations for the n^{th} vertex. A possible front is considered a sequence of N points $p_0, p_1, ..., p_{N-1}$ where $p_n \in g_n$ for each $0 \leq n < N$. For each front in the label we now have k^{2N} possible deformations. Consider a front prediction image Im. For each front we choose the deformation that scores best according to a matching function comparing each possible deformation with the predicted fronts of Im. As the deformation is restricted to a $k \times k$ grid we ensure that such a matching is only applied locally, as we only strive to counteract the label position bias, which we generally expect to be small. The matching procedure does not take classification into account, but rather tries to match a front against the whole set of predicted front. We do not change the labels class during extraction. Thus each front is extracted as the class provided by the weather service. We implemented the matching procedure using C++ and Pybind11 Jakob et al. (2017). This method comes at the risk, that instead of predicting the position of the front the network may end up detecting a systematic displacement of the front within the range of the $k \times k$ grid. We believe this could happen for two possible reasons.

- The label bias exhibits a systematic displacement itself.
- k is chosen too large.

The first case is actually a problem within the labels themselves and it is generally questionable whether these labels are suitable for training at all. The second case may lead to a problem, where each prediction is within vicinity of an actual front, even though the prediction is far off. As a result we chose $k = 7$, allowing each vertex to displace itself up to 3 pixel in each direction, limiting the scope of movement to a sensible range.

https://doi.org/10.5194/wcd-2021-27
Preprint. Discussion started: 25 May 2021

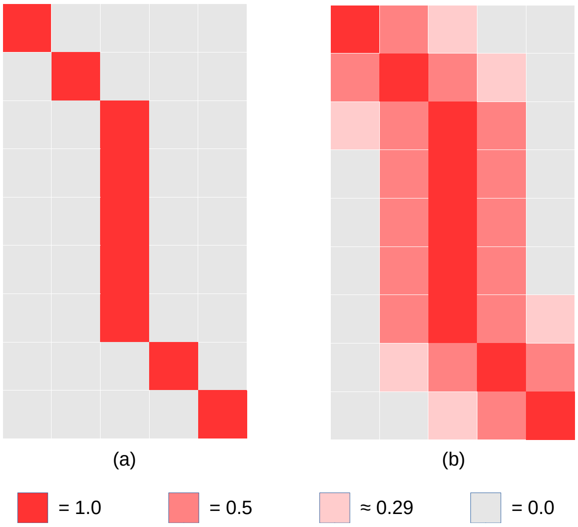

Figure 5. Comparison of line extraction methods: (a) only extracting the provided polylines, (b) extracting a 3-pixel wide polyline. The difference in pixel weight is optional.

2.3 Loss Functions

During training we extract the label lines as described in Section 2.2.4. As a loss function we decided to use a loss based on *Intersection over Union* (IoU), which we evaluate for each output channel individually, before combining them by a weighted average. This loss inherently circumvents the problem that in each channel most of our output, belongs to the background as it does not contain a front. While the original formulation of IoU is used for sets and therefore a strictly binary labeling, we used an adjusted version that works with floating point probabilities. This loss function is also used by Matsuoka et al. (2019). However, they only evaluate it on a single output channel. The definition of loss for a single output channel is shown in Equation 1:

$$L(p,x) = \frac{\sum_i p_i \cdot x_i}{\sum_i p_i \cdot p_i + \sum_i x_i \cdot x_i - \sum_i p_i \cdot x_i} \qquad (1)$$

Here L denotes the loss function, x is the extracted label image and p the prediction of our network. p_i and x_i are respectively the i^{th} pixel of either p or x.

As our networks generates a multichannel output we calculate our adjusted loss E (see Eq. (2)) as follows. We use a softmax activation function to turn our network output into probabilities for each channel.

$$E(p,x) = \frac{\sum_t w_t L_t(p,x)}{\sum_t w_t} \qquad (2)$$

Our first output channel corresponds to the background label, which corresponds to the absence of fronts. We invert this output, by subtracting it from 1, to get a value describing the presence of fronts. We then calculate the single channel loss L_t

for each of our 5 output channels (front, warm, cold, occlusion, stationary) denoted as $t \in 0,1,2,3,4$ individually and combine these three losses using a weighted average of each of these 5 losses, with the corresponding weights w_t. We set w_0 to 0.2 to put more emphasize onto classification. The remaining weights for the classes $t \in 1,2,3,4$ are calculated individually for each batch. For each batch nz_t denotes the amount of samples in this batch, that contain a label of class t. We calculate an intermediate weight b_t for $t > 0$ as $b_t = \frac{batchsize}{nz_t}$, where $batchsize$ is the number of samples in a batch. From this intermediate weight we obtain w_t as $w_t = (1 - w_0)\frac{b_t}{\sum_{k=1}^{4} b_k}$

2.4 Baseline Method

We compare our results against a baseline method provided by ETH Zurich. The method introduced by Jenkner et al. (2010) and later modified by Schemm et al. (2015) uses thermal gradients and other information to predict fronts. While the method was originally designed to work on a $1°$ resolution grid, we adjusted the hyper parameters of the method to allow it to run on a $0.5°$ grid. Our network works on a $0.25°$ resolution grid and outputs on the same domain. Therefore, when comparing against the baseline method we resample the network output to a $0.5°$ resolution using a 2D maximum pooling operation.

3 Results

We trained and evaluated multiple models. Each model is trained using 6 GPUs on a single node of the Mogon II cluster of the Johannes-Gutenberg-University. Each node contains 6 Nvidia GTX1080 GPUs and an Intel Xeon CPU E5-2650 v4 with 24 cores and hyperthreading. Data was staged in prior to training to enable reading from a local SSD rather than the parallel file system.

- A model using 8526 samples of 6 years from 2012-2014, 2015/03-2015/12 and 2018-2019 using labels from both NWS and DWD.
- A model using 5608 samples of 4 years from 2012-2014 and 2015/03-2015/12 using only labels from NWS.
- A model using 4142 samples of 3 years from 2015/03-2015/12, 2018 and 2019 using only labels from DWD.

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. A softmax activation function is applied to the raw network output before any evaluation or post processing steps. We performed evaluation on the DWD data-set and the NWS data-set separately and provide the same evaluations for the networks that were only trained using DWD or NWS Data respectively. This results in a total of 6 evaluations, which are listed in Tables 3 and 4.

3.1 Evaluation on validation data

The trained models were evaluated on test sets from 2017 for both the NWS and the DWD label sets. For evaluation we calculated the CSI similar to Lagerquist et al. (2019). As the domain of our predictor is a latitude and longitude grid we need to use the great-circle distance between two points to estimate distance in kilometers. We evaluated the distance by modeling the earth as a perfect sphere with a radius of 6371 kilometers. Network output is transformed into front-objects in three steps.

- Set all predictions with a value lower than 0.45 to 0, all others to 1
- Use one iteration of 8-connected binary dilation and calculate all different connected components. Each connected component is considered a front.
- Filter the labeled image with the binary mask from step 1 to remove the dilation effect.
- remove all fronts that consist of less than 2 pixel

. The same transformation is applied to the provided weather service fronts. As the label data is binary, the first step has no effect in that case. Note that some provided weather service fronts are separate lines in the label file, but end up as a single longer front due to being connected due to the coarser grid used in the analysis, e.g. $0.25°$. The last step of object conversion is performed to remove short frontal fragments that may have been caused by cropping of the region. We do not perform any of these steps for the baseline method as it already contains a filtering step within the algorithm itself.

A predicted front F_p is considered to be matched to the weather service label if the median distance of each pixel of F_p to the nearest labeled pixel of the same class in the weather services label image is less than D km. The same is applied vice versa for the weather service fronts compared against the network output. Each class of front can only be matched to pixel of the same class, however each frontal object is matched against the whole set of objects of the same class, rather than just a single other object.

We define n_{MWS} as the count of fronts provided by a weather service, that could be matched against the prediction, while n_{WS} is the count of all provided fronts. Similarly, n_{MP} describes the count of all predicted fronts, that could be matched against the weather service fronts, while n_P describes the total count of predicted fronts. With these values we can then calculate the *Critical Success Index* (CSI), *Probability of Object Detection* (POD), and *Success Rate* (SR) as described in Eq. 3, 4, and 5, respectively. As mentioned by Lagerquist et al. (2019) these measurements are also applied in other scenarios, like the verification of tornado warnings by the NWS (Brooks, 2004). The success rate describes the probability that a predicted front corresponds to an actual front from the labeled data-set, while the POD describes the probability that an actual front is detected by the network. SR and POD could easily be maximized at the cost of the other, by either not predicting anything or classifying each pixel as a front instead. The CSI serves as a measurement that penalizes such degenerate optimizations as it maximizes only when both values yield good results. Generally speaking a high CSI score is preferable. Whether it is more important to have a high POD or SR depends on the task at hand and whether it is more important that the detection is more sensitive or more accurate.

$$POD = \frac{n_{MWS}}{n_{WS}} \tag{3}$$

$$SR = \frac{n_{MP}}{n_P} \tag{4}$$

$$CSI = \frac{1}{POD^{-1} + SR^{-1} - 1} \tag{5}$$

Table 3. CSI, POD and SR values for $D = 250\ \mathrm{km}$ evaluated on DWD data for 2017. Warm fronts tend to be detected worse than the other classes while cold fronts are generally well detected. Stationary fronts are not available for DWD labels and are therefore not listed.

Training region	NWS			DWD			Both		
	CSI	POD	SR	CSI	POD	SR	CSI	POD	SR
Binary	51.2%	67.2%	68.4%	67.6%	79.4%	82.1%	66.9%	77.6%	82.8%
Warm	21.8%	24.6%	65.4%	51.9%	61.8%	76.3%	51.4%	62.0%	75.0 %
Cold	40.0%	50.0%	66.5%	58.2%	70.3%	77.2%	57.0%	68.8%	76.8 %
Occlusion	35.7%	43.8%	65.6%	53.2%	70.2%	68.7%	52.3%	67.3%	70.2%
Stationary		–			–			–	

The resulting CSI, POD and SR for $D = 250\ \mathrm{km}$ are displayed in Table 3 and 4 for the binary task which only considers the classes front and no-front, as well as the individual scores for each of the four frontal classes. Tables S1 and S2 in the SI additionally display the case where each front can only be matched against a single other front object rather than the whole set of fronts of a class. We can see that using this metric harshly reduces the object detection rate, while keeping the Success Rate at a similar level than the other metric. This indicates that our network tends to detect larger fronts as multiple short segments, which each by itself does not fulfil the matching criterion. This would unnecessarily punish the provided output. For this reason we added the secondary evaluation metric where each individual front can be be matched against the complete set of fronts of a single type. The provided results show that the network excels at the pure front detection task with CSI scores of 66.9% (DWD) or 62.9% (NWS). At the same time the network evaluates with a POD and SR exceeding 76.6%, with a slight edge on detecting the DWD labels. The classification scores are comparably lower with a class CSI ranging between 35.8% and 57.0%. Across all tests warm and stationary fronts appear to be harder to classify for the network than cold fronts or occlusions. Another interesting observation is the fact training on a single region does not provide a good generalization onto the other region, which is expressed by low CSI scores when training on only the DWD (NWS) data and evaluating on the NWS (DWD) data. At the same time training on both regions yields comparable scores as the single region trained networks, which implies that using as many regions as possible is desirable. Generally, this might be originating in different synoptic structures of cyclones and their associated fronts over the North American continent and over the North Atlantic.

3.1.1 Comparison Against Baseline

We additionally evaluated the CSI score on a coarser 0.5° resolution grid and compare the results against our baseline algorithm, evaluated on the same grid. Our baseline does not classify its results which is why we only display and compare the task of front detection and forgo any classification results. We evaluated both evaluation metrics. As shown in Table 5 our network (NET) outperforms the baseline algorithm (ETH) in all evaluated scenarios and metrics with a more than twice as high of a CSI score.

Table 4. CSI, POD and SR values for $D = 250\text{km}$ evaluated on the NWS data 2017. Warm fronts tend to be detected worse than the other classes while cold fronts are generally well detected. The network trained purely on DWD data, could not learn stationary fronts, as they are not included in the training data, which is why these are not listed.

Training region	NWS			DWD			Both		
	CSI	POD	SR	CSI	POD	SR	CSI	POD	SR
Binary	63.9 %	77.4 %	78.5 %	43.7 %	49.5 %	78.9 %	62.9 %	76.6 %	77.8 %
Warm	37.7 %	55.7 %	53.9 %	20.7 %	42.0 %	28.9 %	35.8 %	54.5 %	51.1 %
Cold	54.3 %	68.9 %	71.9 %	38.1 %	46.6 %	67.6 %	54.2 %	69.9 %	70.7 %
Occlusion	48.2 %	68.9 %	61.6 %	35.4 %	53.2 %	51.4 %	47.8 %	68.8 %	61.0 %
Stationary	42.5 %	55.2 %	64.8 %		–		40.9 %	52.7 %	64.4 %

Table 5. Comparison of the CSI, POD and SR of the ETH algorithm against our network (NET) for the data of 2017. As the algorithm provided by the ETH does not classify fronts we use the binary-classification evaluation for our network. (quasi-)stationary fronts were removed from the network output as well as the NWS label, as the ETH algorithm should not predict those. For the DWD Label these could not be reliably removed, due to the labels ambiguity. The suffix "all" describes the case where a front can be matched against the whole set of fronts at once, while "single" describes the case where a front can only be matched against a single front. The Network clearly outperforms the ETH algorithm in all cases. We can see that the ETH algorithm is better in predicting fronts in the DWD regions rather than the NWS region.

Method	Evaluation on DWD Region			Evaluation on NWS Region		
	CSI	POD	SR	CSI	POD	SR
ETH-all	29.5 %	43.0 %	48.5 %	20.8 %	41.5 %	29.4 %
ETH-single	22.2 %	29.4 %	47.6 %	18.9 %	37.6 %	27.5 %
NET-all	66.7 %	76.3 %	84.1 %	54.9 %	72.0 %	69.8 %
NET-single	63.3 %	72.3 %	83.5 %	52.8 %	69.7 %	68.6 %

3.2 Evaluation and Comparison on Test Data

We further evaluate our data on an independent test data set, which consists of 1463 samples for the DWD region and 1464 samples for the NWS region. We first evaluated the CSI scores as we did for the validation set. However, we only report the score where each front is compared against the whole set of fronts. We also re-evaluate the CSI scores on the coarser 0.5° grid and compared our results against our baseline on this data-set. These results are shown in Tables 6 and 7.

The CSI, POD and SR scores for the test data-set are comparable to the validation set and suggest that our network generalizes well onto other data-sets. The comparison to the ETH algorithm also shows similar to the test data-set that our network strongly outperforms the baseline in all measured scores.

Table 6. CSI, POD and SR as in Tables 3 and 4 but for data from our test set from 2016

Training region	Both					
Testing region	DWD			NWS		
	CSI	POD	SR	CSI	POD	SR
Binary	64.2 %	73.5 %	83.5 %	64.8 %	79.2 %	78.0 %
Warm	47.6 %	56.3 %	75.4 %	35.8 %	55.8 %	49.9 %
Cold	54.3 %	64.4 %	77.6 %	55.6 %	71.7 %	71.3 %
Occlusion	50.8 %	65.2 %	69.7 %	48.7 %	71.5 %	60.3 %
Stationary		–		43.0 %	56.7 %	64.0 %

Table 7. Comparison of the CSI, POD and SR of the ETH algorithm against our network for the data of 2016. As the algorithm provided by the ETH does not classify fronts we use the binary-classification evaluation for our network. (quasi-)stationary fronts were removed from the network output as well as the NWS label, as the ETH Algorithm should not predict those. For the DWD Label these could not be reliably removed, due to the labels ambiguity. The suffix "all" describes the case where a front can be matched against the whole set of fronts at once, while "single" describes the case where a front can only be matched against a single front. The Network clearly outperforms the ETH algorithm in all cases. We can see that the ETH algorithm is better in predicting fronts in the DWD regions rather than the NWS region.

Method	Evaluation on DWD Region			Evaluation on NWS Region		
	CSI	POD	SR	CSI	POD	SR
ETH-all	28.8 %	41.7 %	48.3 %	21.1 %	41.3 %	30.2 %
ETH-single	22.6 %	30.0 %	47.6 %	19.3 %	37.3 %	28.6 %
NET-all	64.5 %	72.6 %	85.2 %	56.5 %	73.6 %	70.9 %
NET-single	61.0 %	68.6 %	84.7 %	54.5 %	71.1 %	70.0 %

3.3 Variation of Physical Variables across Frontal Surfaces

The network occasionally disagrees with the weather services labels. To examine this in more detail we compare physical quantities on a line across the frontal border. We create such a cross section for each pixel that corresponds to a front in 3 steps.

- Estimate the direction (45° interval) of the frontal border at the given point
- Sample points orthogonal to the frontal border, centered at the given point
- Use the scalar product of the wind direction vector and the normal front vector to sort the sampled points along wind direction

The results are accumulated and the mean is presented in Fig. 6 for the DWD frontal data-set. The corresponding plots for the NWS front data-set are shown in the supplement (Fig. S1). In the left column we evaluated the variation in temperature,

https://doi.org/10.5194/wcd-2021-27
Preprint. Discussion started: 25 May 2021

specific humidity, and the temperature gradient across the frontal surface based on fronts locations (i) identified by the machine learning algorithm (dashed lines) and (ii) indicated in the surface analysis from the DWD (solid lines). The meteorological data are taken from the surface pressure level of the ERA5 data-set. The temperature gradient is calculated using finite differences across the sampled temperature. For both front location data-sets the temperature is clearly increasing (decreasing) across the frontal surface for cold (warm) fronts, as would be expected from the physical definition of these features (Fig. 6 a). For cold fronts the across-frontal change in temperature is similar, but fronts identified by DWD are located on average at slightly warmer temperatures. For the identified warm fronts the across-frontal temperature variation is on average larger than for the DWD labels. This maybe explained by the assignment of some warm fronts with weak temperature gradients to the additional category of stationary fronts by our machine learning algorithm. Note that this category does not exist in the DWD data-set. For occluded fronts there is only a small across-frontal temperature variation as could be expected and again this is consistent across both data-sets.

For most automatic front detection algorithms the across-frontal temperature gradient is of importance, which is shown in Fig. 6 b. Again we see very similar patterns for both the DWD and our front data-set. In both data-sets the frontal surface is located at the onset of a region with strong change in the horizontal temperature gradient. This is consistent with the physical definition of frontal zones and agrees with the manually designed automatic front detection algorithms. Again the only notable difference occurs for warm fronts, where the location of the frontal surface seems to be better or more consistently placed relative to the region of strongest change in temperature gradient by the machine learning algorithm. The relative positioning of the frontal surface in the horizontal temperature gradient field is very similar for occluded and stationary fronts compared to cold and warm fronts, which is encouraging in terms of a unique and systematic placement of frontal surfaces.

Finally, we also investigate the change of specific humidity across the frontal surface (Fig. 6 c). As expected these changes are strongly correlated with the across-frontal temperature pattern: For occluded and stationary fronts we find a maximum in specific humidity at the location of the frontal surface, while for warm (cold) fronts specific humidity decreases (increases) across the frontal surface in the propagation direction. Also consistent with the temperature patterns, specific humidity values are generally higher across occluded fronts in our front data-set compared to the DWD data-set and the specific humidity is larger on the warm side of cold and warm fronts. This is mostly likely related to the longer and more continuous front segments identified in the machine learning algorithm compared to the DWD data-set, i.e. in general also identify more southerly frontal points. The specific humidity gradient across warm and cold fronts is more pronounced for fronts identified with our algorithm. At least for warm-fronts that is consistent with the stronger temperature variation. For cold fronts again the extension to warmer, more southern points likely explains the difference.

When comparing the frontal zone structure over North America according to NWS labels and our generated labels, generally also consistent structures are found (see SI) with deviations mirroring broadly those identified for the DWD data.

Overall, from the good agreement in physical structures across the identified frontal surfaces from our algorithm and from the manual weather service analysis we can conclude that our algorithm detects physically meaningful positions. The positioning of the frontal surfaces is further consistent with physical intuition and interpretation prevalent in literature.

Finally, we also investigate the physical structure of fronts from

https://doi.org/10.5194/wcd-2021-27
Preprint. Discussion started: 25 May 2021

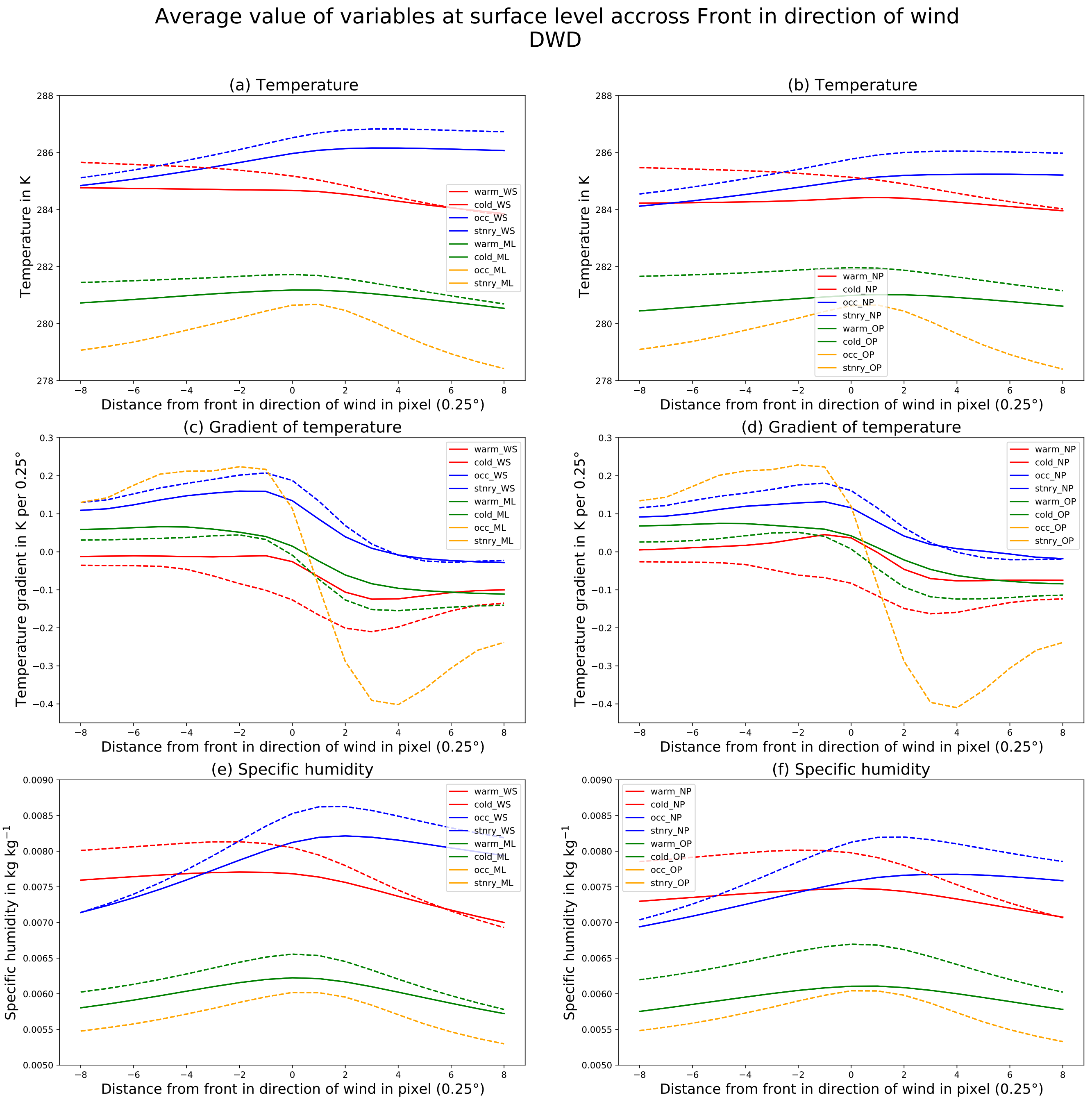

Figure 6. (a) Mean of the temperature, (c) temperature gradient and (e) specific humidity of provided (solid, WS) and network generated (dashed, ML) front labels for the DWD Data. (b),(d),(f): The same for not-predicted (solid, NP) or over-predicted (dashed, OP) front labels.

- our machine learning method that are not within 3 pixel to any weather service label (OP), and
- the weather service labels that are not within 3 pixel to any machine learning predicted front (NP).

These can be loosely interpreted as *false positive predictions*, e.g. predictions of our network that were not labeled by the weather service (OP), and *false negative predictions*, e.g. labels of the weather service that where not predicted by our network (NP). The range of 3 pixel corresponds to approximately $0.75°$. Note that we do not have a label for stationary front in the DWD labels, which is why no curve is shown in the plots. The corresponding composites of temperature, temperature gradient and specific humidity are shown in the right column in Fig. 6 for the European region and in the right column in the Fig. S1 (SI) for the North American region. First lets have a closer look at the NP cases: For warm fronts there is an obvious signal that the temperature change is very small across the fronts that are not detected by our algorithm. This suggests that many weak warm fronts exist in the manually generated front data set, which represents either stationary fronts or warm fronts perpetuated from earlier analysis times. For cold fronts and occluded fronts the picture is not that clear. For both cases the network seems to miss the more northern fronts (front parts) as indicated by the lower mean temperature compared to the detected fronts. However, there is a hint towards less structure in the temperature gradient for these cases. Secondly, for the OP cases the frontal structure seems to be very similar to the structure across frontal surface identified in the DWD data. This suggests that the additional fronts (or frontal segments) identified by the network a physically valid examples of fronts and are consistent with the overall physical idea of frontal surface structures.

The same comparison for the NWS data-set suggests that again weaker frontal structures are missed by the network, while additionally identified fronts are in their structure physically consistent with the fronts labeled in the manual analysis. This analysis additionally confirms the physical consistency and meaningfulness of the network generated front labels.

3.4 Comparison of Frontal Climatology

To further investigate the soundness of our predictions we created frontal climatologies for the year 2016 for both the provided weather service labels as well as our network and the baseline method. While the respective weather services only provide labels within their analysis region, both the network and the ETH algorithm can be executed on the global grid. The resulting climatologies are shown in Fig. 7.

First we compare the climatology for the North Atlantic / European region from the manually labeled data-set with the climatology of network generated fronts. In the DWD climatology the North Atlantic storm track is clearly visible as a band of heightened front occurrence stretching from the East coast of North America to the channel (Fig. 7 c). Frontal activity is tampering off inwards of the European west coast. The climatology of the network generated fronts has a very similar overall structure with a strongly enhanced frontal frequency in the storm track region (Fig. 7 a). Frontal frequency is somewhat larger at the beginning of the storm track. This may be related to the training with North American manual analysis, which naturally has a stronger focus on the early cyclone lifecycle and the European data. Over the Channel and North Sea Coast of Europe frontal frequency in the network generated data-set is somewhat lower than in the DWD data-set, which maybe related to the

https://doi.org/10.5194/wcd-2021-27
Preprint. Discussion started: 25 May 2021

inclusion of stationary fronts in the latter but not the former. We have seen also in the previous section that very weak warm fronts, as may exist further into the European continent are often not detected by the network. In both data-sets a slightly enhanced frontal frequency around Iceland is evident.

Next we compare the climatology for the North American region from the manually labeled data-set with the climatology of network generated fronts. The manual labels indicate the onset of the storm track with enhanced frontal frequencies just off the North American East Coast and secondary peaks in frontal frequencies in the lee of the Rocky Mountains and along the West Coast (Fig. 7 d). The climatology of network generated fronts captures all three maxima in the frontal frequency in roughly the same location (Fig. 7 a). However, frontal frequency in the lee of the Rocky mountains and along the West Coast are more pronounced in the network generated climatology. We are under the impression that the network tends to assign labeled warm fronts as stationary and vice versa. These shifts may explain the different frontal frequency.

Finally, we compare the global climatology of network generated front labels to those generated by ETH automatic front detection algorithm (compare Fig. 7 a and b). The striking first difference between the two climatologies is the much larger spatial extend of regions with high frontal frequency in the second data-set. This is evident both in the storm track regions on both hemispheres but also the subtropical regions. In the subtropics regions of large gradients in equivalent potential temperature exist and these are picked up by the automatic front detection algorithm. However, their structure and origin differs from fronts in the extratropics. It appears that the network is able to detect this difference in the structure, while focusing solely on equivalent potential temperature and frontal propagation speed does not. In absence of any manual data-set that can serve as ground truth it is difficult to judge the physical meaningfulness of the climatological patterns emerging from either algorithm. And indeed in the case of the subtropics may strongly depend on the purpose and definition of what is considered a frontal structure. In the storm track regions on both hemispheres both data-sets show consistently enhanced frontal frequencies over similar geographical regions. They only differ in the zonal extend of the regions with enhanced activity and the absolute values of frontal frequencies. In the only region, where we have an independent, manually generated data-set often considered as the "ground truth", the climatology of network generated fronts is in closer agreement with the former than the climatology from the ETH automatic front detection algorithm. For the southern hemisphere or the North Pacific we currently do not have any such data-set available. The second striking differnce is the high frontal frequency along orographic barriers in the climatology from the ETH automatic front detection algorithm, i.e. along the Andes, Greenland, Himalaya and Antarctic coast line. These maxima in frontal activity are largely absent from the climatology of network generated fronts consistent with the manually labeled data-sets. It appears that the network correctly discriminates between temperature and humidity gradients arising only because of the presence of significant topography from those caused by dynamically generated air mass boundaries. In contrast, focussing solely on the advection speeds in regions of large equivalent potential temperature gradients seem not to suffice. Overall, the global picture emerging from the extrapolation of the network trained on the North American, North Atlantic and European domain performs well also on a global scale and correctly identifies regions of high frontal activity expected from previous investigations and the known general circulation patterns. While physically plausible, this is of course no vigorous evaluation of the performance of the extrapolation to different regions of the globe. In future work should investigate this aspect in a more quantitative manner with manually labeled data-sets from other parts of the globe. To quantify the former qualitative

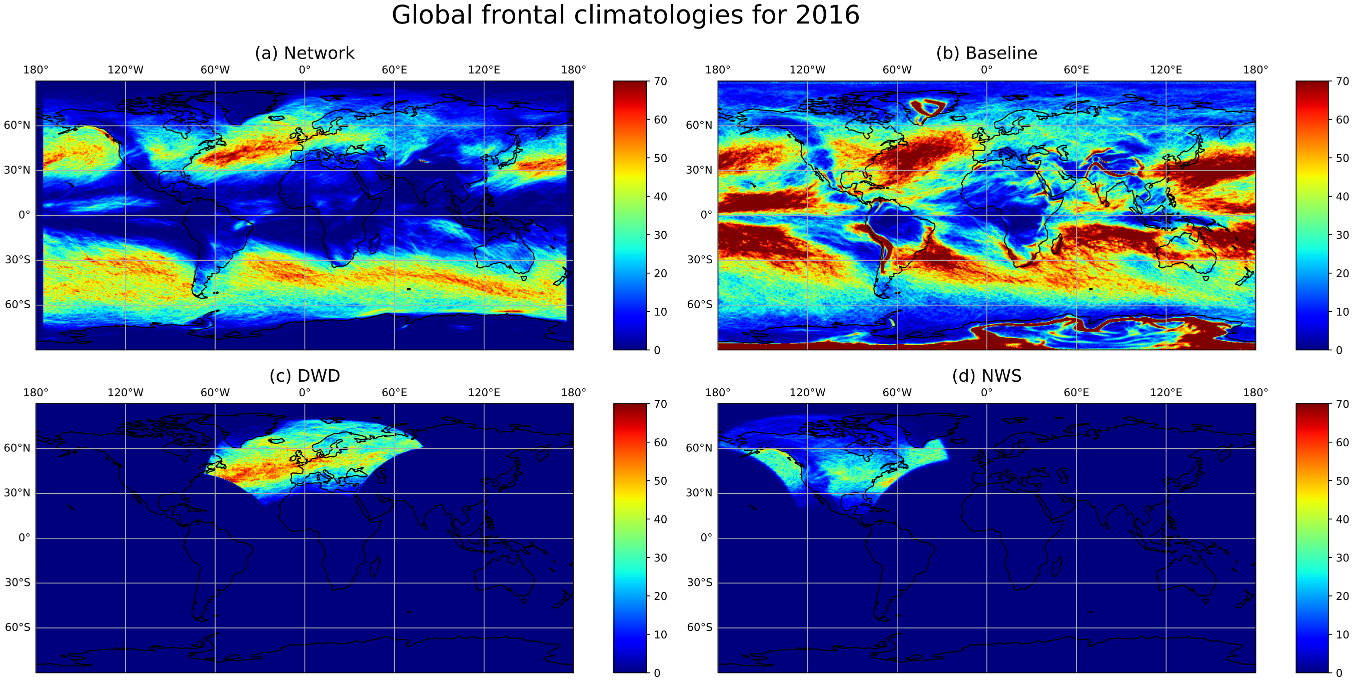

Figure 7. (a) A global frontal climatology of the Network executed on the 0.25° resolution grid and resampled to 0.5° resolution for the year 2016. The outer 5° border denotes the region where the network does not provide a valid prediction. (b) A global frontal climatology of the ETH algorithm executed on the for the year 2016. (c) Climatology of the provided DWD Frontal labels for 2016. (d) Climatology of the provided NWS Front labels for 2016. Red denotes more than 70 fronts. The front count was clipped at 70 for visual representation. Stationary Fronts are explicitly excluded from the climatology of the network generated and NWS labeled data. The global climatology from the ETH algorithm does not include fronts propagating at less than $3\,\mathrm{m\,s^{-1}}$. The DWD data-set may include stationary fronts, as we were unable to reliably separate them from warm or cold fronts.

discussion of the climatologies we evaluated the Pearson correlation coefficient of the created climatologies within the regions described in Table 2. As the ETH algorithm does not provide stationary fronts, we excluded stationary fronts when creating the climatologies. However, due to the ambiguity in the label for stationary fronts in the DWD label data, we were unable to remove those from the data-set. As a result stationary fronts are likely to still be present in the DWD data. The correlation coefficients are provided in Table 8. Our Network outperforms the baseline algorithm for both regions, with correlation coefficients greater than 79.0%. The ETH algorithm performs badly at the DWD region, as it detects many false positive fronts at the Greenlandian coast. If we remove the section spanning $[60^{\circ}N, 70^{\circ}N]$, $[-45^{\circ}E, -30^{\circ}E]$ from the evaluation its score increases from 34.8% to 59.7%, while our network keeps a high correlation of 79.6%. For the NWS data-set the scores are more similar. We however found that the south eastern corner ($[30^{\circ}N, 35^{\circ}N]$,$[-65^{\circ}E, -60^{\circ}E]$) of our used NWS region is not covered by the NWS. If we remove this area we can increase correlation score from 82.4% to 85.0%, which is even higher than the score obtained at

https://doi.org/10.5194/wcd-2021-27
Preprint. Discussion started: 25 May 2021

Table 8. Pearson correlation coefficient of the predicted fronts of the ETH Algorithm (ETH) and our trained Network (NET) against the provided labels of the weather services for 2016. The second row denotes the regions, where the results were evaluated. Training Region respectively corresponds to the regions described in Table 2. Without Greenland corresponds to the DWD Training Region without a region containing Greenland ($[60°, 70°]$N, $[-45°, -30°]$E). Without South-east corresponds to the NWS Training Region with the south eastern corner ($[30°, 35°]$N,$[-65°, -60°]$E) cropped from the region as it does not belong to the NWS analysis region. Stationary fronts were excluded for from all climatologies except the DWD labels.

Method	Correlation against DWD		Correlation against NWS	
	Training Region	Without Greenland	Training Region	Without South-east
ETH	34.8%	59.7%	73.9%	75.4%
NET	82.1%	79.6%	82.4%	85.0%

the DWD region.

Overall, the investigation of the front climatology agrees well with physically expected patterns and climatologies from manually generated frontal data-sets. This lends additional physical credibility to the network generated frontal labels. A physically plausible global climatological pattern further suggests that the learned frontal identification can be extrapolated training region. We found that for this is necessary to including data from two sufficiently different geographic regions, i.e. North America and North Atlantic / Europe, as well as to augment the data-set by including also zonally mirrored examples of the frontal cases (not shown). The latter was found to be particular important for a good performance in the southern hemisphere.

4 Discussion

From our results we can see that there are non-negligible differences in the expression of fronts between north America and western Europe. Detection results of the networks that were trained using only a single label show that one cannot simply transfer a trained detection network to any other region, without a drastic loss in CSI scores. The presented Cross Sections further show that while the different frontal types have common characteristics across the continents (e.g. direction of temperature gradient) there are still differences in the intensity of these characteristics on the neighboring regions (e.g. different temperatures). Additionally we cannot neglect the fact that the labeling of the DWD-Regions is ambiguous regarding stationary, cold and warm fronts. This implies that the inclusion of further data-sets - for example the data-sets used by Matsuoka et al. (2019) or data of the southern hemisphere - may create even better results. The latter is especially interesting as it allows for a better quantitative global evaluation of our method. Further research on how to handle the label bias may also be beneficial, considering that the rules for classifying a front may be different between data-sets as well.

Table 6 (test data-set) and Tables 3 and 4 (verification data-set) show our network excels at the detection of fronts, resulting in high CSI scores of more than 66.0% or 62.9% for both regions. Classification quality of warm and Stationary fronts is worse than cold and occlusions. A possible explanation for is the lack of a clear distinction of these two front classes from the DWD data, which in return leads to more false classifications due to the ambiguity. When changing the matching criterion from matching against all fronts to matching being only possible against a single front, we can observe a high drop in Object detection rate from 77.6% / 76.6% to 72.6% / 71.1% for the DWD / NWS Region as shown in the SI (Tables S1 and S2). At the same time the success rate barely changes. This indicates that our network tends to not fully cover large frontal regions with a single front but rather multiple smaller, disjointed fronts. However we do not observe the same drop in the classification scores which indicates that this mostly affects fronts where the network is unsure about the correct label. Another reason may be the fact that the binary detection case merges fronts that consist of alternating classes into a single long front. If some of these alternating classes are not predicted by the Network, shorter fragments are created.

In future work separating the detection from the classification task may be beneficial, seeing the good detection rates of the presented network in the binary case. We would also like to further explore the application and effect of other methods to handle the label bias, such as the method described by Acuna et al. (2019).

The provided climatologies for the network agree very well with the labels of the weather services. In combination with the provided Cross sections this further enhances our belief that our proposed network can be a useful tool for the detection of fronts. A possible application may be to use the network as a supportive tool for the weather services to propose location and classifications of frontal data to the respective meteorologists. We further provide a video supplement visualizing the network outputs at a 1 hour time resolution of January 2016 for an almost global region spanning $[85°N, -175°E]$ to $[-85°N, 175°E]$ Niebler (2021b). The background consists of the normalized specific humidity input at surface level. See section $S3$ in the SI for further details on how the video was created. Our network was trained solely on data from our two regions, located at the northern hemisphere. The displayed climatologies as well as the video supplement however appear to show physically plausible results for the southern hemisphere ranging from $[-30°N$ to $-70°N]$. The storm track is clearly expressed in our climatology and the general shape, composition and motion of fronts appear plausible in the video supplement. While this is only a qualitative observation, it seems to contradict our claim that training on one region is insufficient for extrapolation onto other regions. However we believe that this is due to the fact that this region is mostly covered by sea. As a result there is far less topographic influence, which causes the extrapolation from the northern Atlantic onto the southern hemispherical data to be less erroneous. However any of this needs to explicitly be evaluated in future work. Finally comparison of our networks output with the provided weather service labels in Figure 8 shows the effect of our proposed loss functions. Our network tends to predict smoother shaped fronts, which are not always located on top of the label provided by the corresponding weather services. However our networks nonetheless outputs thin lines with reasonable transitions between fronts, while not requiring the application of morphological post processing operations.

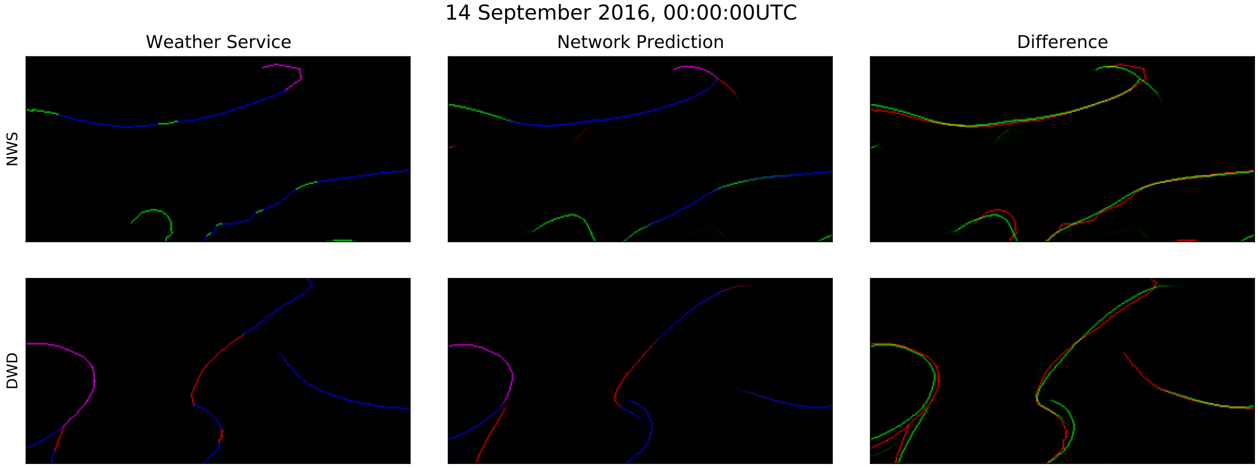

Figure 8. Comparison of Fronts from provided labels and network generated output. Colors are chosen to correspond to color choices in literature, except for stationary fronts, which are colored pink. Difference image shows a direct comparison of frontal placement by the weather service (red) and the network (green), ignoring classification. All displayed examples are at 14 September 2016, 00:00:00 UTC.

5 Conclusions

We trained a neural network on a loss function, that allows to classify and predict fronts across the input regions. Our applied loss function leads the network to predict clearly localized fronts without the need of morphological post processing thinning operations. Our network clearly outperforms the compared baseline method which is a widely applied method for frontal detection. We showed that we cannot simply transfer a locally trained network onto any other region but rather need to train on several data-set to obtain a reliable general front detection. Climatology results indicate that a transfer on oceanic regions may be feasible, however this has to be evaluated in future research. It is also desirable to further investigate up to which degree extrapolation onto different regions is possible and to investigate whether or not generalization onto global data is possible from just a few subregions.

Code and data availability. The code and a trained model are available at Niebler (2021a). ERA5 Reanalysis data can be accessed via the ECMWF climate data center. Used NWS Frontal label is available with doi: 10.5281/zenodo.2642801 (National Weather Service, 2019). We are not allowed to make the used DWD Label data publicly available, access to the data may however be granted by the DWD.

Video supplement. A video supplement showing predicted fronts for January 2016 is available at https://doi.org/10.5446/53399 (Niebler, 2021b)

Author contributions. Stefan Niebler implemented, and trained the network. He also evaluated the baseline method as well as the network. Bertil Schmidt, Annette Miltenberger, Peter Spichtinger, and Stefan Niebler wrote the draft of the manuscript. Bertil Schmidt, Annette Miltenberger, and Peter Spichtinger proposed and supervised the project. All authors edited the manuscript and analyzed the results.

Competing interests. The authors declare that they have no conflict of interest

Acknowledgements. The study is supported by the project “Big Data in Atmospheric Physics (BINARY)”, funded by the Carl Zeiss Foundation. We acknowledge the ECMWF for providing access to the ERA5 Reanalysis data. We further acknowledge the ETH Zurich and especially Michael Sprenger for providing the code for the used baseline method. Label data for the European continent and Northern Atlantic was provided by the Deutscher Wetterdienst. Label data for the North American continent was provided and made publicly available by the North American Weather Service. We further acknowledge the ZDV of the Johannes Gutenberg University and the Mogon II Super Cluster for providing the necessary hardware to execute our experiments.

References

Acuna, D., Kar, A., and Fidler, S.: Devil is in the Edges: Learning Semantic Boundaries from Noisy Annotations, 2019.

Berry, G., Reeder, M. J., and Jakob, C.: A global climatology of atmospheric fronts, Geophysical Research Letters, 38, https://doi.org/10.1029/2010GL046451, 2011.

Biard, J. and Kunkel, K.: Automated detection of weather fronts using a deep learning neural network, Advances in Statistical Climatology, Meteorology and Oceanography, 5, 147–160, https://doi.org/10.5194/ascmo-5-147-2019, 2019.

Bitsa, E., Flocas, H., Kouroutzoglou, J., Hatzaki, M., Rudeva, I., and Simmonds, I.: Development of a Front Identification Scheme for Compiling a Cold Front Climatology of the Mediterranean, Climate, 7, https://doi.org/10.3390/cli7110130, 2019.

Brooks, H. E.: TORNADO-WARNING PERFORMANCE IN THE PAST AND FUTURE: A Perspective from Signal Detection Theory, Bulletin of the American Meteorological Society, 85, 837 – 844, https://doi.org/10.1175/BAMS-85-6-837, 2004.

ECMWF: L137 model level definitions, https://www.ecmwf.int/en/forecasts/documentation-and-support/137-model-levels, access date: 2021-05-18, 2021.

Foss, M., Chou, S. C., and Seluchi, M. E.: Interaction of cold fronts with the Brazilian Plateau: a climatological analysis, International Journal of Climatology, 37, 3644–3659, https://doi.org/10.1002/joc.4945, 2017.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/https://doi.org/10.1002/qj.3803, 2020.

Hewson, T. D.: Objective fronts, Meteorological Applications, 5, 37–65, https://doi.org/10.1017/S1350482798000553, 1998.

Hope, P., Keay, K., Pook, M., Catto, J., Simmonds, I., Mills, G., McIntosh, P., Risbey, J., and Berry, G.: A Comparison of Automated Methods of Front Recognition for Climate Studies: A Case Study in Southwest Western Australia, Monthly Weather Review, 142, 343–363, https://doi.org/10.1175/MWR-D-12-00252.1, 2014.

Jakob, W., Rhinelander, J., and Moldovan, D.: pybind11 – Seamless operability between C++11 and Python, https://github.com/pybind/pybind11, 2017.

Jenkner, J., Sprenger, M., Schwenk, I., Schwierz, C., Dierer, S., and Leuenberger, D.: Detection and climatology of fronts in a high-resolution model reanalysis over the Alps, Meteorological Applications, 17, 1–18, https://doi.org/10.1002/met.142, 2010.

Lagerquist, R., McGovern, A., and II, D. J. G.: Deep Learning for Spatially Explicit Prediction of Synoptic-Scale Fronts, Weather and Forecasting, 34, 1137 – 1160, https://doi.org/10.1175/WAF-D-18-0183.1, 2019.

Matsuoka, D., Sugimoto, S., Nakagawa, Y., Kawahara, S., Araki, F., Onoue, Y., Iiyama, M., and Koyamada, K.: Automatic Detection of Stationary Fronts around Japan Using a Deep Convolutional Neural Network, SOLA, 15, 154–159, https://doi.org/10.2151/sola.2019-028, 2019.

Mesinger, F., DiMego, G., Kalnay, E., Mitchell, K., Shafran, P. C., Ebisuzaki, W., Jović, D., Woollen, J., Rogers, E., Berbery, E. H., Ek, M. B., Fan, Y., Grumbine, R., Higgins, W., Li, H., Lin, Y., Manikin, G., Parrish, D., and Shi, W.: North American Regional Reanalysis, Bulletin of the American Meteorological Society, 87, 343 – 360, https://doi.org/10.1175/BAMS-87-3-343, 2006.

National Weather Service: National Weather Service Coded Surface Bulletins, 2003-, https://doi.org/10.5281/zenodo.2642801, 2019.

Niebler, S.: stnie/FrontDetection: Front Detection Network Release 0.0.1, https://doi.org/10.5281/zenodo.4770096, 2021a.

Niebler, S.: Detected Fronts January 2016, Copernicus Publications, https://doi.org/10.5446/53399 $Lastaccessed : 18May$2021, 2021b.

Parfitt, R., Czaja, A., and Seo, H.: A simple diagnostic for the detection of atmospheric fronts, Geophysical Research Letters, 44, 4351–4358, https://doi.org/10.1002/2017GL073662, 2017.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library, in: Advances in Neural Information Processing Systems 32, edited by Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., pp. 8024–8035, Curran Associates, Inc., http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf, 2019.

Ribeiro, B. Z., Seluchi, M. E., and Chou, S. C.: Synoptic climatology of warm fronts in Southeastern South America, International Journal of Climatology, 36, 644–655, https://doi.org/10.1002/joc.4373, 2016.

Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, 2015.

Schemm, S., Rudeva, I., and Simmonds, I.: Extratropical fronts in the lower troposphere–global perspectives obtained from two automated methods, Quarterly Journal of the Royal Meteorological Society, 141, 1686–1698, https://doi.org/10.1002/qj.2471, 2015.

Schulzweida, U.: CDO User Guide, https://doi.org/10.5281/zenodo.3539275, 2019.

Shakina, N. P.: Identification of zones of atmospheric fronts as a problem of postprocessing the results of numerical prediction, Russian Meteorology and Hydrology, 39, 1–10, https://doi.org/10.3103/S1068373914010014, 2014.

Simmonds, I., Keay, K., and Bye, J. A. T.: Identification and Climatology of Southern Hemisphere Mobile Fronts in a Modern Reanalysis, Journal of Climate, 25, 1945–1962, https://doi.org/10.1175/JCLI-D-11-00100.1, 2012.