JGR Atmospheres

RESEARCH ARTICLE
10.1029/2023JD040126

U-Net Segmentation for the Detection of Convective Cold Pools From Cloud and Rainfall Fields

Jannik Hoeller[1,2], Romain Fiévet[2], Edward Engelbrecht[1,3], and Jan O. Haerter[1,2,3,4]

[1]Integrated Modeling, Leibniz Centre for Tropical Marine Research, Bremen, Germany, [2]Niels Bohr Institute, Copenhagen University, Copenhagen, Denmark, [3]Physics and Earth Sciences, Constructor University Bremen, Bremen, Germany, [4]Department of Physics and Astronomy, University of Potsdam, Potsdam, Germany

Key Points:

- A neural network is trained to detect cold pools from cloud top temperature and precipitation
- The method works reliably and a pseudo-3D approach, which includes dynamic cold pool evolution, performs best
- The method offers new perspectives for the cold pool detection from geostationary satellite observations over tropical land

Correspondence to:
J. Hoeller,
jannik.hoeller@leibniz-zmt.de

Citation:
Hoeller, J., Fiévet, R., Engelbrecht, E., & Haerter, J. O. (2024). U-Net segmentation for the detection of convective cold pools from cloud and rainfall fields. *Journal of Geophysical Research: Atmospheres*, *129*, e2023JD040126. https://doi.org/10.1029/2023JD040126

Received 3 OCT 2023
Accepted 8 DEC 2023

Abstract Convective cold pools (CPs) mediate interactions between convective rain cells and help organize thunderstorm clusters, in particular mesoscale convective systems and extreme rainfall events. Unfortunately, the observational detection of CPs on a large scale has been hampered by the lack of relevant near-surface data. Unlike numerical studies, where fields, such as virtual temperature or wind, are available at high resolution and frequently used to detect CPs, observational studies mainly identify CPs based on surface time series, for example, from weather stations or research vessels—thus limiting studies to a regional scope. To expand to a global scope, we here develop and evaluate a methodology for CP detection that relies exclusively on data with (a) global availability and (b) high spatiotemporal resolution. We trained convolutional neural networks to segment CPs in high-resolution cloud-resolving simulation output by deliberately restricting ourselves to only cloud top temperature and rainfall fields. Apart from simulations, such data are readily available from geostationary satellites that fulfill both (a) and (b). The networks employ a U-Net architecture, popular with image segmentation, where spatial correlations at various scales must be learned. Despite the restriction imposed, the trained networks systematically identify CP pixels. Looking ahead, our methodology aims to reliably detect CPs over tropical land from space-borne sensors on a global scale. As it also provides information on the spatial extent and the relative positioning of CPs over time, our method may unveil the role of CPs in convective organization.

Plain Language Summary Cold pools come about when rain evaporates underneath thunderstorm rain cells. Such cold pools are colder and thus denser than the surrounding air, which makes them flow over the surface. The associated gust fronts can often be felt when observing thunderstorm clouds in nature, as strong, but relatively short-lasting, winds occurring near the thunderstorm downpour. Such cold pools can bring about clumps of thunderstorms, which can deliver large quantities of rainfall within areas of approximately 100 km across. Thus, detecting cold pools on these and larger scales is important, but so far difficult due to the challenge of observing the air currents underneath clouds from satellite. We here present a method that may be able to do just that, namely detecting cold pools from above, by identifying signatures left behind in cloud patterns, for example, arc-like shapes along the gust front. We demonstrate the algorithms capabilities using high-resolution simulation data, where we are able to "know" the true result. We use an artificial intelligence framework to carry out this statistical task. We suggest that our method should be applicable to satellite data and thus give new insight into cloud organization at large scales.

1. Introduction

Cold pools (CPs) are volumes of atmospheric air that are cooled by evaporation of precipitation. The resultant denser air experiences negative buoyancy (Markowski & Richardson, 2011), resulting from convective downdrafts, or microbursts (Lundgren et al., 1992). At Earth's surface, CPs spread horizontally as density currents (Drager & van den Heever, 2017; Droegemeier & Wilhelmson, 1985; Zuidema et al., 2017). While expanding radially along the surface, CPs can be characterized as consisting of (a) a head, which can measure hundreds to several thousand meters vertically and (b) a shallower interior, which is separated from the head by a wake region (Benjamin, 1968; Cafaro & Rooney, 2018; Droegemeier & Wilhelmson, 1987; Fiévet et al., 2023; Kneller et al., 1999; Kruse et al., 2022; Markowski & Richardson, 2011; Meyer & Haerter, 2020; Simpson, 1980).

Substantial mechanistic significance has been attributed to the thin surface of horizontal convergence between the CP head and the ambient atmosphere, often referred to as CP gust front. On the one hand, this CP gust front

features pronounced vertical wind speed. On the other, the initial negative-buoyancy anomaly near the CP's gust front is gradually reduced as the CP spreads, a possible consequence of enhanced surface latent and sensible heat fluxes (Drager et al., 2020; Tompkins, 2001; Torri & Kuang, 2016). Thus, in the course of the lateral expansion, buoyant ambient air can be lifted (Drager & van den Heever, 2017), resulting in further condensation and convection. "Communication" between cells (Simpson, 1980) can thus be encoded by such mechanical and thermodynamic effects.

Although the relative contributions of the thermodynamic versus the dynamic effect are still under debate (Fuglestvedt & Haerter, 2020; Romps & Jeevanjee, 2016; Torri et al., 2015) and can depend on the environmental and boundary conditions (Feng et al., 2015), there is substantial consensus that CPs can trigger new clouds in their vicinity and modify the subcloud moisture distribution (Böing et al., 2012; Drager & van den Heever, 2017; Haerter et al., 2020; Jensen et al., 2022; Schlemmer & Hohenegger, 2016).

Due to their ability to influence the organization (Böing, 2016; Haerter et al., 2019; Schlemmer & Hohenegger, 2016), the maintenance of squall line structures (Rotunno et al., 1988; Schumacher & Rasmussen, 2020), and potential intensification of thunderstorms (Houze & Betts, 1981; Zipser et al., 1981), CP characteristics have been investigated in past decades, often within numerical studies: virtual temperature (T_v) anomalies have been used to track CPs in cloud-resolving simulations by patch detection (Schlemmer & Hohenegger, 2016) or by unsupervised image segmentation (Gentine et al., 2016). Drager and van den Heever (2017) compared the utility of different variables for CP identification in simulations. Torri and Kuang (2019) used a Lagrangian tracking to investigate CP collisions. Focusing on the dynamical gust front, Fournier and Haerter (2019) and Henneberg et al. (2020) introduced tracking algorithms targeting the thin (dynamic) convergence rings surrounding each CP.

Over continental regions in the tropical climate and during midlatitude summer, CP dynamics has been implicated in the evolution of so-called mesoscale convective systems (MCS; Jensen et al., 2022), which are extended clusters, $\mathcal{O}$(100 km), of thunderstorms cells (Houze, 2004; Schumacher & Johnson, 2008; Schumacher & Rasmussen, 2020) and have been found to contribute to the majority of tropical rainfall and dominate potential increases in extremes (Tan et al., 2015). Despite their climatic relevance, tropical MCS are still nearly impossible to forecast by simulations (Fritsch & Carbone, 2004; Sukovich et al., 2014).

At smaller spatiotemporal scales, idealized simulations allow for detailed analysis of specific mechanisms as they provide near-continuous output data for many variables—making CP detection and tracking feasible. Yet, numerical studies still depend on the model and resolution chosen, so that findings do not immediately carry over to the real world. Since traditional general circulation models are too coarse to resolve CP processes (Feng et al., 2015; Fiévet et al., 2022), CP mechanisms are mostly studied in high-resolution simulations within limited domain sizes or, less commonly, by including specific parameterizations (Grandpeix & Lafore, 2010; Rio et al., 2009). In both cases, the validity of the outcome is limited by the underlying modeling assumptions.

Observational CP studies are usually hampered by the lack of spatially resolved near-surface data. Using point-like near-surface station observations at subhourly temporal resolution, the spread of a given CP can however be detected as a rapid temperature drop in the time series (de Szoeke et al., 2017; Kirsch et al., 2021; Kruse et al., 2022; Vogel et al., 2021; Zuidema et al., 2017). At times, perhaps as a result of strong surface sensible heat fluxes (Knippertz et al., 2007), a dew point temperature increase, rather than a temperature drop, is found to be a robust characteristic of CP gust fronts, for example, in arid regions (Emmel et al., 2010; Redl et al., 2015). Also combinations with dynamic features, such as changes in wind speed have been employed (Emmel et al., 2010; Redl et al., 2015). While Emmel et al. (2010) ultimately validated each event subjectively with infrared satellite imagery, Redl et al. (2015) implemented a criterion based on satellite microwave data. Making use of the dust signature in brightness temperature differences, Caton Harrison et al. (2021) devised an algorithm to automatically identify and track Saharan CP outflows based on gradients of the corresponding brightness temperature difference fields. In a validation based on 35 manually identified CP outflows, the approach achieved a detection rate of 74.2% (26/35). Also "thin line" echoes associated with the CP leading edge in radar imagery have been used in manual CP detections (Borque et al., 2020; Brandes, 1977; Engerer et al., 2008; Wakimoto, 1982).

CP detection over the ocean has occasionally benefited from the imprint that CPs leave on the sea surface (Atlas, 1994). During the Variability of American Monsoon Systems (VAMOS) Ocean–Cloud–Atmosphere–Land Survey Regional Experiment (VOCALS-REx) in the Southeast Pacific, Wilbanks et al. (2015) applied an air density increase criterion to detect stratocumulus-topped boundary layer CPs from in situ ship measurements.

Garg et al. (2020) introduced a method based on wind gradients from space-borne scatterometers to detect CP regions over tropical oceans. To validate their method, they employed in situ air temperature measurements from buoys and a simulation. As part of the Elucidating the Role of Cloud–Circulation Coupling in Climate (EUREC4A) field campaign (Stevens et al., 2021), Touzé-Peiffer et al. (2022) developed a method for detecting CPs over tropical oceans based on the mixed-layer height in atmospheric soundings.

Detecting CPs over continents at a spatially extended scale can be more cumbersome. Reaching real-world CP detection results for the continental areas is however the only realistic benchmark for CP representation in numerical models, and thus the proper dynamics of MCS. To approach automatized and global-scale CP detection over land surfaces, we developed an algorithm using convolutional neural networks (CNNs) which relies only on quantities that are observable from geostationary satellite imagery. The algorithm was trained and tested using data from idealized cloud-resolving simulations, where all field variables are available over the entire domain. The chosen configurations correspond to an atmosphere over tropical land surfaces. To our knowledge, this is the first approach for detecting CPs over land on a global scale based on satellite data. Our algorithm may allow for new insight into the role of CPs in convective organization and the formation of weather extremes.

2. Methods

CNNs, which gradually coarsen a field of input data through filtering operations, are widely used for classification and segmentation problems. Upon each filtering step, spatial correlations at larger and larger scales are distinguished. Whereas CNNs for classification problems group the entire input data field into a set of classifiers, CNNs for segmentation problems return to the resolution of the input to mark each pixel as being of one of several categories. For the problem at hand, we wish to mark each pixel in the 2D plane as either belonging to a CP or not—thus the segmentation technique is appropriate.

2.1. Simulation Data

In order to simplify the generation of labeled data sets, the network training and testing is conducted using data from numerical simulations. To this end, the cloud-resolving three-dimensional atmosphere simulator System for Atmospheric Modeling (SAM; Khairoutdinov & Randall, 2003), version 6.11, is used. It resolves the Euler equations in the anelastic approximation on a staggered mesh. Convective fluxes are evaluated using a fifth-order finite difference scheme from Yamaguchi et al. (2011) and turbulent dissipation is modeled by an eddy-viscosity-based closure. Moist thermodynamics is resolved by transporting liquid and ice water static energy, total precipitating and nonprecipitating water mass fractions, and uses a bulk single-moment microphysics closure scheme.

The configuration chosen for this study corresponds to an atmosphere over an idealized moist tropical land surface. It is similar to the configuration studied by Jensen et al. (2022) which exhibited strong and complex CP activity and is therefore suited to design and test our detection method. Parts of the underlying simulation output were used in Hoeller, Fiévet, and Haerter (2023) to devise a detection and tracking method for CPs in cloud-resolving simulation data. The computational domain has a size of $L_x = L_y = 240$ km in the horizontal directions and extends vertically to a maximum altitude of $L_z = 26$ km. It is discretized by an orthogonal mesh of horizontal resolution $\Delta x = \Delta y = 200$ m and vertical resolution Δz increasing from $\Delta z(z = 25\text{ m}) = 50$ m to $\Delta z(z = 25\text{ km}) = 1{,}000$ m over 100 levels. In the following, we use $n_x, n_y \in [0, N]$, as integers labeling the indices of the horizontal model grid, with $N = 1{,}200$ the total number of grid points in either horizontal dimension. The lateral boundary conditions are set to be periodic. Relevant two-dimensional simulated fields are sampled instantaneously every $\Delta t = 10$ min. We often denote the discrete time $t = t_n \equiv n\Delta t$, measured from the beginning of the simulation, by the integer time step n.

Surface heat fluxes are evaluated using Monin-Obukhov similarity theory with a saturated humidity (moist surface) condition and a prescribed diurnally varying temperature T, with an average of $T_0 = 298$ K. The temperature amplitude ΔT, defined as half the diurnal temperature range between maximum and minimum, is chosen to represent plausible ranges measured for tropical land (Sharifnezhadazizi et al., 2019). The effect of the surface forcing is to trigger idealized diurnally varying convective activity typical of tropical land surfaces: moist convection tends to develop and self-organize during the afternoon hours—giving rise to a complex organizational pattern. The nocturnal cooling reduces convective activity and precipitation rates typically reach a domain-wide minimum during the early morning hours of the subsequent model day. In order to work with a diverse set of

atmospheric conditions, four different configurations are considered, where $\Delta T \in \{2, 4\}$ K and wind shear is either switched off or set to a temporally and spatially averaged vertical profile over tropical Africa (LAT: 5.5°–16°N, LON: −20°–10°E) obtained from ERA5 reanalysis data for July 2016. The vertical profile consists in a piecewise linear profile with zero velocity below $z = 1$ km, linearly increasing speeds from 0 to 16 m s^{-1} up to 19 km altitude and 16 m s^{-1} beyond. Depending on their ΔT, we term the configurations either "diu2K" or "diu4K." Configurations with wind shear further obtain the addition "wind," that is, "diu2K wind" denotes the setup with $\Delta T = 2$ K and wind shear.

2.1.1. Ground Truth Labeling

Labeled data sets are derived from simulation output based on a CP detection and tracking algorithm (CoolDeTA) devised by Hoeller, Fiévet, and Haerter (2023). CoolDeTA applies a *k*-means algorithm to the sum of the normalized lowest domain level fields of virtual temperature and horizontal wind speed to classify each pixel in the two-dimensional field as either "potential CP" or "no CP" without defining a fixed threshold. We note that, in the following, only such potential CP areas can contain actual CPs. Individual CP instances are differentiated and labeled based on a watershed algorithm. The starting points for the flooding of the watershed algorithm are the downdraft centers within spatially contiguous rain patches with a surface rain intensity, r_{int}, exceeding a threshold of $r_{\mathrm{int}} \geq r_0$. Like Hoeller, Fiévet, and Haerter (2023), we apply $r_0 = 2$ mm h^{-1}. Additionally, the flooding starts in the centers of tracked CP instances from the previous time step, if there are any. To detect only robust CP instances, we keep rain patches and potential CP areas only when their area $A \geq A_0$ with $A_0 = 3$ km^2, as opposed to the 2 km^2 threshold used by Hoeller, Fiévet, and Haerter (2023). Valid rain patches with a downdraft center, which is located in an area classified as "no CP" by the *k*-means algorithm, will not be labeled as CPs. The same is true for potential CP areas which do not coincide with the center of a downdraft or a tracked CP instance from the previous time step.

Providing the fields of virtual temperature, T_{v}, and both horizontal and vertical wind speed in the lowest domain level, as well as r_{int}, CoolDeTA can identify, label, and track each CP instance individually while storing additional information. To use the simplest possible case in the present study, we here keep the labels derived based on CoolDeTA binary, comprising the two classes "CP" and "no CP."

2.1.2. Input Variables

Regarding the potential input for the neural network, SAM outputs several variables which are accessible from space-borne data. The present study focuses on the cloud top temperature, T_{CT}, and r_{int}. For T_{CT}, we employ the standard output of SAM where T_{CT} equals the cloud top temperature for cloudy domain columns and the surface temperature otherwise. T_{CT} and r_{int} are readily available from infrared emissions and an increasing number of precipitation products. Depending on the region of interest and problem-specific requirements in terms of spatial and temporal resolution, as well as accuracy, these precipitation products can be multisatellite products with global coverage such as IMERG (Huffman et al., 2015) or even products based on ground-based weather radars.

2.1.3. Training and Test Sets

For each of the four simulation setups, output is available for 7.5 simulation days in total. The first 3 days are considered as spin-up phase (Hoeller, Fiévet, & Haerter, 2023). After these spin-up days, we employ the simulation output of day 4 for network training and validation. While we train networks based on the training set, the validation set is used to monitor the progress of the training on separate data which has not been trained on. As is common in supervised learning, we randomly split the data by assigning each instance with a probability of 75% to the training set and with 25% to the validation set. Throughout the entire training, including different network trainings, we keep the obtained allocation fixed to facilitate the comparison between networks. Yet, the order of the training data is randomly shuffled every training epoch, that is, every time the training set is passed through a neural network.

When all network trainings are completed, the final performance is evaluated based on a test set (Willemink et al., 2020) consisting of simulation output of day 6, that is, the test set is not considered at any earlier stage. Although the observed CPs, and the complex pattern formed by their interaction, are unique for each simulation day, the offset of 1 day between the test set and the training and validation set guarantees fully independent sets w.r.t. the distribution of relevant quantities such as humidity.

To ensure sufficient variation between consecutive time steps of the data sets, we consider only every second time step of the corresponding simulation output for the training and validation set, and every fourth time step

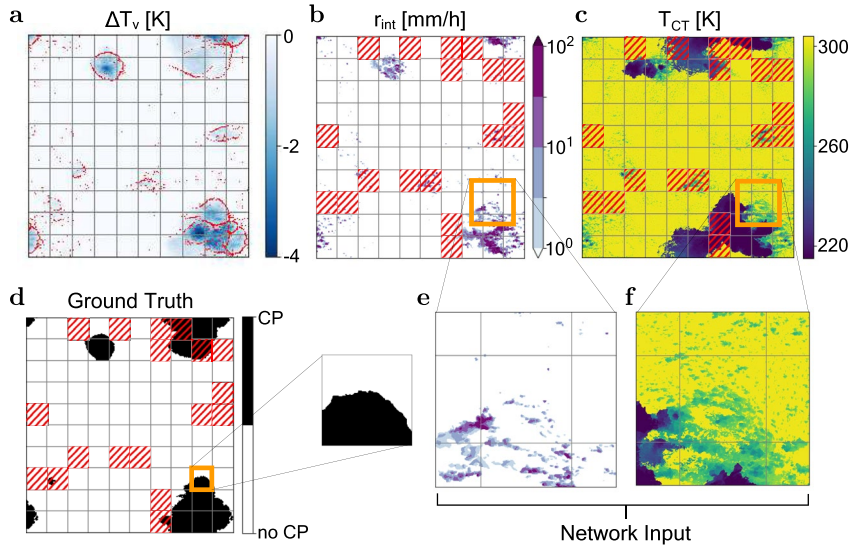

Figure 1. Defining patches for neural network input and ground truth. (a) Time step 497, that is, 80 min before T_{max} on simulation day 4, of "diu4K wind," showing near-surface virtual temperature anomaly, ΔT_v, with superimposed dynamical gust front, that is, $w > \overline{w} + 2\sigma_w$ (red scatter). The superimposed grid represents the individual $n_p \times n_p$ pixel patches, processed by the neural network. (b) Analogous to (a) but for surface rain intensity, r_{int}. Patches that were omitted from the data set are hatched. (c) Analogous to (a) but for cloud top temperature, T_{CT}. (d) Ground truth labeling showing cold pool (CP) areas as black regions; a single patch is enlarged for clarity. (e) Highlighted patch, including padding, for r_{int}. (f) Analogous to (e) but for T_{CT}.

for the test set. This is particularly important for the test set to prevent any distortion of the final results due to correlated data.

In order to reduce the computational cost and accelerate network training, we subdivided every $N \times N$ pixel output field, termed "image," into 100 subregions, which we refer to as "patches," of $n_p \times n_p$ pixels each. As our downsampling requires the integer n_p to be a power of 2 (see Section 2.2) and to compromise between computational effort and prediction skill, we chose $n_p = 128$ as the linear dimension of each two-dimensional patch. To accomplish this, each original output image is padded from $n_x = 1{,}200$ to $n_{x,pad} \equiv 1{,}280$ in compliance with the periodic lateral boundary conditions.

Eventually, each network prediction requires an input and a corresponding ground truth to optimize and/or evaluate the network performance. While the ground truth corresponds to an $n_p \times n_p$ pixel patch of the output images with derived labels, the input consists of stacked patches corresponding to T_{CT} and r_{int}. To compensate for lacking context information at patch boundaries, the input patch is $n_p/2$ pixels larger than the underlying ground truth patch on either side, resulting in an input patch size of $2n_p \times 2n_p$ pixels (Figure 1). Although the additional $n_p/2$ pixels on each side are thus ranging into the adjacent ground truth patch, this overlap does not distort the results as the final network prediction only comprises the underlying central $n_p \times n_p$ pixel patch.

To ensure a robust training process and reliable results, we manually checked the ground truth labeling of every patch in the data set. We omitted patches if they (a) contained at least 1, but less than 25 pixels (i.e., 1 km^2) of class "CP," (b) were in the center of a larger convective system with a gust front significantly beyond the boundaries of the input patch, (c) were poorly labeled by the CP detection algorithm, or (d) featured ambiguous scenes where an unequivocal verification of the labeling is not possible. For the evaluation of both (c) and (d), the dynamical gust front, that is, $w > \overline{w} + 2\sigma_w$ served as main indicator: clear offsets between gust front and boundaries of ground truth CPs were interpreted as poor labeling, discontinuous, and thus dissipating gust fronts as ambiguous cases. Omitted patches were excluded from the data set. We did not modify the individual pixels which belong to a certain CP instance. Only if a certain CP instance in an otherwise accurately labeled patch was a complete artifact, we set all pixels of that CP instance in the patch to "no CP" and kept the corrected patch in the data set.

As simulation setups affect the cloud and rainfall patterns associated with CPs, we considered patches from simulations with different environmental conditions. Yet, both simulations with imposed wind profiles feature

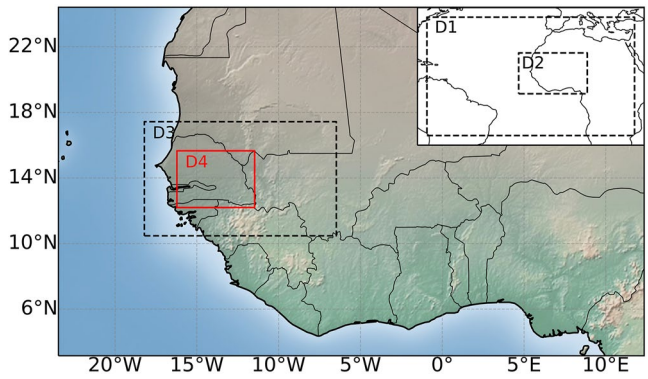

Figure 2. Weather Research and Forecasting (WRF)-nested domains for Senegal case study. Map of Africa showing the regional/outer domain (dashed), denoted as D1 at 9 km horizontal resolution. D2, D3, and D4 denote the nested domain boundaries at horizontal resolutions 3, 1, and 0.333 km, respectively. The red domain is the primary region used in this case study.

prevailing easterly winds. To allow the network to capture underlying patterns independent of the wind direction, we rotate each patch of the two simulations with wind by 90°, 180°, and 270° and add the resulting patches to the data sets. Extending data sets with slightly modified copies of the data based on operations such as rotations or translations is a common approach to increase the amount and diversity of data and is called data augmentation (Chlap et al., 2021; Shorten & Khoshgoftaar, 2019).

Data imbalances due to the underrepresentation of classes or features in the training set are a common issue of learning algorithms (He & Garcia, 2009). Taking the reduced convective activity due to nocturnal cooling into account, the majority of patches does not contain any CP pixels in the ground truth data and features only the class "no CP." We compensated for this by randomly removing a certain number of these patches (Shi et al., 2021). By experiment, we selected the number of patches with only class "no CP" to be 4% of the training and validation set. The other extreme are patches with class "CP" only. Surface temperature oscillations promote the sudden organization of CPs into convective systems (Haerter et al., 2020; Jensen et al., 2022). Since the surface areas of these convective systems often exceed the patch size, a great number of patches have class "CP" only. However, omitting patches in the center of larger convective systems according to (b) already lowered the number of patches with class "CP" only to ≈5.5% resulting in a sufficiently balanced training and validation set distribution (Figure 3a) with mean class "CP" fractions of 0.26 (diu2K), 0.31 (diu2K wind), 0.38 (diu4K), and 0.52 (diu4K wind). We chose not to balance the distribution of the test set in order to not affect the results in any way.

The resulting data set for training and validation comprises 4,208 patches with 984 different CP instances. Each CP instance represents a certain CP event as labeled by CoolDeTA. Depending on both their lifetime and their spatial extent, individual CP instances can be present in multiple patches in the data set. Thus, the number of CP instances is higher for diu2K (420) and diu4K (389) compared to the two setups with wind shear, diu2K wind (107) and diu4K wind (68), which feature less, but often larger and more persistent CP events. The number of CP instances provides an indication of the CP variety within the data set. It should be noted though that CPs and their patterns are often affected by complex interaction processes. Although a specific CP event might undergo significant changes during its lifetime and even merge with other CP events, it would still be counted as one CP instance by CoolDeTA. Accordingly, we consider the provided numbers of CP instances a rather conservative estimate of the CP variety in the data set. The test set comprises 7,226 patches with 183 CP instances. With 44 CP instances in diu2K, 31 in diu2K wind, 77 in diu4K, and 31 in diu4K wind, they are more uniformly distributed among the different simulations. We attribute this also to the fact that we considered only every fourth time step for the test set, thus excluding some short-lived CP instances. Since the test set was not balanced, 6,684 of the 7,226 patches contain only class "no CP."

2.1.4. Senegal Case Study

In addition to the network training and testing, we also conducted a numerical simulation over West Africa to validate the method under more realistic conditions. For this case study, we used the nonhydrostatic Advanced Research Weather Research and Forecasting (WRF) model version 4.3 (Skamarock et al., 2021) for a 24-hr numerical simulation on 4 August 2022 over Senegal in West Africa (Figure 2). The date was chosen primarily for the high frequency of convective systems during this period, as well as the occurrence of smaller-scale convection in other parts of the study domain. To obtain numerical simulation data at a relatively high grid resolution, in this case 333 m, a nesting approach was necessary. We therefore created four domains with grid spacing of 9, 3, 1, and 0.333 km with a one-way nesting strategy. Only data from the 0.333 km domain simulation are used in this study.

National Oceanic and Atmospheric Administration/National Centers for Environmental Prediction Global Forecasting Model (GFS) forecasts, with a 0.25° horizontal grid spacing, were employed for the initial and boundary conditions in the 9 km outer domain, provided at 3 hourly intervals. Fifty-five vertical levels were chosen, such that their vertical spacing decreases closer to the surface. To represent surface fluxes, we used the Noah Land Surface Model scheme with soil temperature and moisture at four layers. The 3–0.333 km domains were run with explicit convection. In the 9 km domain, we used the Kain-Fritsch convection scheme with a mass-flux approach.

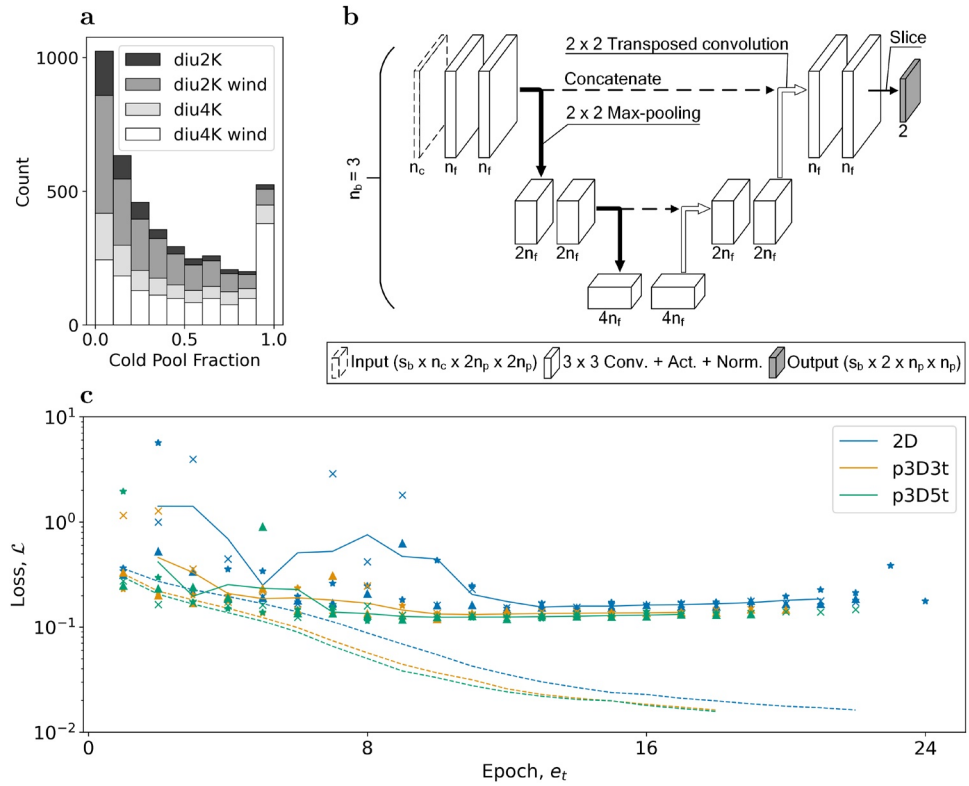

Figure 3. Training statistics and U-Net architecture. (a) Distribution of the data set used for network training and validation w.r.t. the fraction of class "CP" in each patch. (b) U-Net architecture for cold pool segmentation, here for the case with three filtering blocks ($n_b = 3$). The number of input channels n_c represents the number of different variables provided to the network as input. In the case of pseudo-3D models, the number of input channels, n_c = number of variables × number of utilized time steps, n_t. The number of output channels comprises the two classes "CP" and "no CP." (c) Loss, $\mathcal{L}$, as a function of the epoch, e_t, for the 2D, p3D3t, and p3D5t neural networks. Dashed lines represent running averages of training loss for all training runs of a respective neural network type. Thin colored are running averages of validation loss for all training runs of a respective neural network type additionally averaged over a centered window of three e_t; different symbols correspond to the validation loss of the different training runs. Note: As the mean variance of the training loss for the three neural network types is only between 2.5×10^{-6} (2D) and 6.1×10^{-6} (p3D3t), markers for the training loss of different training runs are not visualized.

Finally, we applied the large-eddy simulation approach to represent the planetary boundary layer (PBL) in the 333 m domain, while the Yonsei University PBL scheme was employed for the other domains.

2.2. Network Architecture

As mentioned, instead of predicting one specific label per provided input image (classification), the detection of CPs requires an output, such as "CP" or "no CP," for every pixel of the image (segmentation). A common architecture used for segmentation is the U-Net (Ronneberger et al., 2015), a CNN that consists of an encoder path and a decoder path. In the encoder path, input images are downsampled after every block, allowing the network to learn features at larger scales. A common downsampling method, where the output is generated from the input by considering only the maximum value of a moving window of size $s \times s$ and which we also apply in the present study with $s = 2$, is max pooling. By reducing the resolution of the image in each downsampling step, typically by a factor of 2 as we do here, the network can learn features at different scales. To be able to capture the underlying correlations, the number of filter layers is doubled with every downsampling step. In the decoder path, on the other hand, the images are upsampled again via transposed convolution or interpolation to finally enable pixel-wise predictions. After each upsampling step, concatenated filter layers of the same depth encoder block provide additional information. The employed U-Net architecture for the simplified case with three vertical blocks ($n_b = 3$) is depicted in Figure 3b.

Apart from n_b and the starting number of filter layers n_f, neural networks and U-Nets in particular offer a variety of modeling choices, termed hyperparameters, to tune. After an exploration phase, in which we identified hyperparameters significant for our network along with promising orders of magnitude based on training and validation performances, we investigated the following seven hyperparameters in more detail: n_b, ultimately chosen as $n_b = 6$; n_f, ultimately chosen as $n_f = 64$; the activation function, ultimately chosen as LeakyReLU; the normalization strategy, ultimately chosen as batch normalization; the loss function, ultimately chosen as combination of cross entropy loss and dice loss; the learning rate l_r, ultimately chosen as exponentially decaying function $l_r = 10^{-5} \times \gamma^{e_t}$ with e_t as the training epoch and $\gamma = 0.9$; and the batch size s_b, ultimately chosen as $s_b = 8$. Activation functions are nonlinear functions and a fundamental part of CNNs. Following convolutional layers in the convolution block (cf., Figure 3b), activation functions enable the network to capture complex patterns. Typically, convolution blocks are completed by normalization steps, which can support an efficient learning process (Ioffe & Szegedy, 2015). While the loss function is the function to be minimized during training, l_r controls the corresponding optimization step size. The number of instances considered per optimization step is the batch size. Typically, training batch sizes are greater one to reduce the risk of getting stuck in local minima.

In order to determine the most promising network configuration w.r.t. the seven hyperparameters, we conducted a number of experiments based on the training and validation set. Instead of analyzing all possible combinations of configurations, we limited the number of experiments by structuring them in two stages. Starting from a first guess reference configuration for which all seven hyperparameters were defined pragmatically, the first stage consists of multiple levels, each containing experiments for a group of hyperparameters with all their combinations. After each level, the reference configuration is updated based on the best candidates of those hyperparameters. Due to their close relation, we grouped l_r with s_b (Group 1), activation function with normalization strategy and loss function (Group 2), and n_b with n_f (Group 3). Whereas the hyperparameters in Group 1 are essential for robust learning and thus investigated first, the hyperparameters in Group 3 are examined last as larger numbers of n_b and n_f, which were expected to be advantageous, would slow down the remaining experiments significantly.

Since some hyperparameters could have candidates with similarly good performance so that the best candidate might thus change for other configurations, we performed a second stage of experiments with all combinations of these candidates plus some fine-tuned ones.

Depending on the convolution kernel, CNNs can be categorized into 2D and 3D CNNs. Conventional end-to-end 2D CNNs receive 2D input, which may consist of multiple channels, for example, 2D fields of different variables, apply 2D convolutions, that is, convolutions with 2D kernel matrix, and generate a corresponding 2D output, whereas 3D CNNs analogously process 3D data. At the expense of significantly higher computational cost, 3D CNNs are thus able to learn correlations in a third dimension based on the 3D convolution kernel. As we are interested in 2D segmentations and the simplest model possible, we selected the 2D version. However, since CPs are density currents and exhibit gust fronts typically emanating radially from a precipitation cell center, expansion over time constitutes one of the main CP features (Benjamin, 1968). In order to include this time-dependent component and potentially enable the network to learn the correlations between consecutive time steps, we also implement the so-called pseudo-3D approach. The term "pseudo-3D," as introduced by Vu et al. (2020), represents a model class that is intermediate between conventional 2D CNNs and 3D CNNs. In pseudo-3D models, the information of the third dimension (here time) is inserted as additional input channels to the network, therefore without modifying the network's 2D architecture. As a consequence, the total number of input channels of pseudo-3D models depends not only on the number of input variables provided, but on the product of the numbers of input variables and utilized time steps. Thus, pseudo-3D models might potentially benefit from time-dependent information without being as computationally expensive as end-to-end 3D models (Vu et al., 2020). In the present study, we investigate the pseudo-3D model with three (p3D3t) and five time steps (p3D5t). Time steps are thereby centered about the time step for which a prediction is to be made.

2.3. Loss and Evaluation Metrics

The selection of an appropriate loss function depends on the specific problem at hand. All loss functions use the pixel-wise network prediction $U = [U^{(0)}, U^{(1)}]$, consisting of the two output channels $U^{(0)}, U^{(1)} \in \mathbb{R}^{n_p \times n_p}$, where indexes "0" and "1" indicate the "no CP" and "CP" channels, respectively, and compare U with the corresponding ground truth derived by CoolDeTA, denoted $V \in \mathbb{N}^{n_p \times n_p}$, where $V_{kl} \in \{0, 1\}$, indicating "no CP" and "CP," respectively.

We examined several loss functions during the experiments. For this purpose, we rescaled each pixel $U_{kl}^{(j)}$ in U to the range [0,1] so that the "probabilities" of both the "no CP" and "CP" channel sum up to one. We term the result of this so-called "softmax" function u. The corresponding function is written as

$$u_{kl}^{(j)} \equiv \frac{e^{U_{kl}^{(j)}}}{e^{U_{kl}^{(0)}} + e^{U_{kl}^{(1)}}}, \text{for } j \in \{0, 1\}. \tag{1}$$

In order to compare u to the ground truth, we split V analogously to the prediction via one-hot encoding into two slices of binary data $v = [v^{(0)}, v^{(1)}]$, that is, $v_{kl}^{(0)} = 1 - V_{kl}$ and $v_{kl}^{(1)} = V_{kl}$. As loss functions, we employed a cross entropy loss which is often used as default in image segmentation and defined as

$$\mathcal{L}_{\text{CE}}(u, v) = \sum_{j,k,l} \frac{-v_{kl}^{(j)} \log\left(u_{kl}^{(j)}\right)}{\sum_{m,n,q} v_{nq}^{(m)}}, \tag{2}$$

a soft Dice coefficient loss, defined as

$$\mathcal{L}_{\text{Dice}}(u, v) = 1 - \frac{2 \sum_{j,k,l} u_{kl}^{(j)} v_{kl}^{(j)} + \epsilon}{\sum_{j,k,l} u_{kl}^{(j)} + \sum_{j,k,l} v_{kl}^{(j)} + \epsilon}, \tag{3}$$

where $\epsilon = 1$ is a constant preventing divisions by zero (Jadon, 2020), and a combination of both

$$\mathcal{L}(u, v) = \alpha \mathcal{L}_{\text{Dice}}(u, v) + (1 - \alpha)\mathcal{L}_{\text{CE}}(u, v), \tag{4}$$

with $\alpha = 0.5$. Whereas $\mathcal{L}_{\text{Dice}}$ can deal with imbalanced data sets (Milletari et al., 2016) and focuses on how good the predicted CPs overlap the ground truth CPs, $\mathcal{L}_{\text{CE}}$ evaluates the difference between the probability distributions of u and v. For our problem, we chose $\mathcal{L}$ as loss function as it combines the strengths of both $\mathcal{L}_{\text{Dice}}$ and $\mathcal{L}_{\text{CE}}$ and outperformed both these functions during the experiments.

For the evaluation of the trained networks, we distinguish between patches containing only one of the two classes for the corresponding ground truth data and patches with at least one pixel of both classes. In the former case, the only evaluation metric will be pixel accuracy, PA, which evaluates the fraction of predictions that are correct, defined as

$$\text{PA} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{5}$$

In Equation 5, TP and TN indicate true positive and true negative predictions, respectively, whereas FP and FN denote false positive and false negative predictions, respectively.

In case the ground truth patch contains at least one pixel of both classes, we additionally calculate the intersection over union, IOU,

$$\text{IOU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}. \tag{6}$$

The IOU score is a measure of how well the specific objects of prediction and ground truth overlap one another, ranging from zero, where no overlap is found, to unity, for perfect overlap. Furthermore, we consider Precision and Recall, defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{7}$$

and

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{8}$$

As IOU both Precision and Recall range from zero, where no "CP" pixel was correctly identified, to unity, for a perfect prediction. However, shedding light on different components of the prediction, they help to understand potential sources of good and bad performances.

Table 1
Mean Test Performances of Different Models

Patches	Model	PA (%)	IOU	Precision	Recall
Both classes	2D	93.8	0.71	0.84	0.83
	p3D3t	94.8	0.75	0.83	0.88
	p3D5t	94.5	0.74	0.84	0.87
Only no CP	2D	99.8			
	p3D3t	99.9			
	p3D5t	99.9			
Only CP	2D	92.0			
	p3D3t	94.1			
	p3D5t	85.9			

Note. Presented are mean performances for pixel accuracy (PA), intersection over union (IOU) score, Precision, and Recall for patches with at least one pixel of both classes "CP" and "no CP" in the ground truth (both classes) and PA for patches with only pixel of class "no CP" (only no CP) or "CP" (only CP) in the ground truth.

To enable a more application-oriented perspective on the performance of the three models, we define cold pool objects (CPOs) as spatially four-connected regions of ≥25 "CP" pixels (≥1 km^2) and evaluate both the probability of detection, POD, and the false alarm ratio, FAR, defined as

$$\mathrm{POD} = \frac{D}{D + M}, \tag{9}$$

and

$$\mathrm{FAR} = \frac{\mathrm{FA}}{\mathrm{FA} + D}, \tag{10}$$

with the numbers of successfully detected CPOs, D, missed CPOs, M, and false alarm, FA. The minimum CPO size of 25 "CP" pixels ensures that only robust predictions are considered. Ground truth CPOs are considered detected if (a) predicted CPOs overlap at least 50% of their area and (b) at least 50% of the area of the predicted CPOs falls inside ground truth CPOs. Condition (b) makes sure that only skilled predictions with CPO areas in the correct order of magnitude are considered successful detections. As the smallest ground truth CPO in the test set comprises 59 pixels, the defined minimum size does not affect the CPO detection. Undetected ground truth CPOs are considered missed CPOs. Predicted CPOs which do not coincide with any "CP" pixel of the ground truth are considered false alarms.

2.4. Network Validation

We plot the training and validation losses for the 2D and both pseudo-3D models as a function of the epoch, e_t (Figure 3c). e_t describes how many times the entire training set has been passed through the neural network. The loss measures the quality of the prediction, where a value of zero means perfect prediction. Instead of defining a fixed e_t, we stop the training if the validation loss has not improved for 10 consecutive e_t. Taking into account the stochasticity involved in the training process, we conducted three runs for each model. As might be expected, the *training loss* decreases monotonically with the data employed for learning, that is, e_t, and reaches a value close to zero for our maximum e_t of 22–24. Notably, for intermediate e_t, both pseudo-3D neural networks perform better than the 2D counterpart, whereas for the final e_t, the three are essentially indistinguishable.

However, a good value of training loss does not necessarily imply optimal *validation loss*, a measure of prediction quality for a previously unseen data set. Indeed, we find that intermediate e_t (≈10) yield lowest validation loss for all three cases, such that a global minimum occurs. This type of optimum at intermediate e_t is typical of neural networks and is often interpreted as large e_t constituting a form of overfitting w.r.t. the training data—yielding less than optimal behavior for the unknown validation data. Yet, the minimum is characterized by an asymmetric increase of validation loss, where somewhat larger e_t lead to only small increases in validation loss. Further, we again find quantitative improvements in validation loss for the pseudo-3D cases, which systematically reach lower values of loss than 2D.

3. Results

For the final evaluation of the trained neural networks, we now employ the test set, that is, the data for day 6 of each simulation. We ensure that the results obtained are on the conservative side, by considering only the worst run with the greatest final validation loss for each model. We conclude this chapter with a case study, where we shed light on potential sources of misclassifications through realistic examples from a simulation over Senegal.

3.1. Test Set Performance

We quantify the utility of our segmentation method by applying typical performance metrics (Table 1). A key measure is pixel accuracy (PA), which is generally high (mean PA $\gtrsim$ 94%) for all models, with the pseudo-3D models performing slightly better than the 2D model. The intersection over union (IOU) score denotes the fidelity

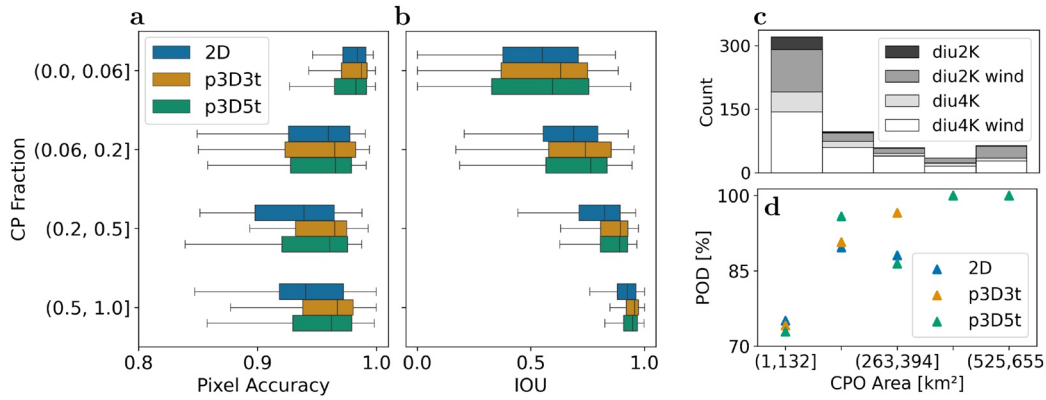

Figure 4. Selected test performances of different models. (a) Distributions of pixel accuracy for each neural network, grouped into quartiles of cold pool (CP) fraction with ranges, as indicated along the vertical axis. Colored bars represent the interquartile range IQR = $Q3 - Q1$ of the three tested models, with the first quartile $Q1$ and the third quartile $Q3$, along with the corresponding median (vertical dash). Whiskers range from $Q1 - 1.5 \times$ IQR (minimum) to $Q3 + 1.5 \times$ IQR (maximum). Outliers w.r.t. this range are not visualized. (b) Analogous to (a) but for the intersection over union (IOU) score. Note that for both metrics a value of unity reflects a perfect prediction. (c) Distribution of spatially contiguous test set cold pool objects (CPOs) w.r.t. their CPO area. (d) Percentage of successfully detected CPOs from (c), POD, for varying CPO area. Note the shared horizontal axis between (c, d) and the overlapping markers for the largest two CPO area intervals in (d).

of spatial overlap of ground truth CPs and neural network-predicted CPs and is thereby sensitive to the underlying CP areas, yielding lower values than PA for all models. Again the pseudo-3D models achieve higher mean IOU scores of 0.75 (p3D3t) and 0.74 (p3D5t) compared with 0.71 for the 2D model. As mean Precision is almost equally high for all models (Table 1), the difference in IOU is mainly driven by the higher mean Recall of the pseudo-3D models, that is, they miss less "CP" pixels than the 2D model.

In order to investigate the sensitivity of the network performances w.r.t. the CP fraction in the patch, we group PA and IOU score into quartiles of CP fraction. For all these quartiles, PA is high (PA $\gtrsim$ 0.95) for all models (Figure 4a). Yet, systematic differences exist: Generally, PA is greatest for small CP fraction and somewhat decreases for intermediate fractions, where it then seems to saturate. This behavior is expected, since (a) the majority of the training and validation set patches contained only small fractions of class "CP," slightly biasing the neural networks toward "no CP" predictions and (b) regions without "CP" pixels often feature neither precipitation, nor clouds, simplifying the network prediction. Overall, PA is somewhat greater for the pseudo-3D models, however, this benefit is nearly lost for small CP fractions, a finding we attribute to the potential noise at the early stages of CP expansion: in p3D3t and p3D5t, where additional time steps are included, data taken before the onset of the CP might contribute to the training—thus obscuring the signal of actual CP expansion.

The IOU score (Figure 4b) can be substantially lower for the smallest CP fraction quartile, with some improvement for the pseudo-3D models. This loss for small CP fraction is however not surprising to us, as for small CP fraction there will often be only few pixels in a patch which actually qualify as CP pixels and small spatial displacements of these pixels in the predicted data can already lead to a drastic reduction of the IOU. Refined measures could be designed that still assign a score to a minimally displaced CP pixel. However, physically relevant CPs, for example, in terms of collision effects (Fiévet et al., 2023; Meyer & Haerter, 2020) and intense precipitation (Jensen et al., 2022) tend to cover larger patch fractions and the IOU score is systematically high—again with best performances for the pseudo-3D models.

We now turn to test patches which contain only "no CP" or "CP" pixels in the ground truth. For the former case, PA yields near-perfect accuracy (Table 1). Thus, the models show high fidelity in capturing cases where CPs are not present, most likely due to the absence of precipitating clouds in a majority of the patches. PA is however substantially reduced in the latter case (Table 1). The reduction in PA is especially pronounced for p3D5t, thus the model where five time steps were used. We attribute this loss of accuracy to the temporal mixing of patches with and without CP pixels, whereby the lack of CP pixels at earlier stages may skew the results.

In Figures 4c and 4d, we evaluate the percentage of successfully detected CPOs, POD (see Section 2.3), as a function of CPO area. The results are quite clear: larger CPOs are detected at quite high fidelity ($\gtrsim$90%), whereas the fidelity

Table 2
Detection Performance on the Test Set for the Different Simulations

Simulation	Total CPOs	2D		p3D3t		p3D5t	
		POD	FAR	POD	FAR	POD	FAR
diu2K	35	0.71	0.24	0.71	0.32	0.83	0.29
diu2K wind	180	0.88	0.15	0.86	0.22	0.87	0.23
diu4K	83	0.90	0.04	0.92	0.04	0.93	0.09
diu4K wind	292	0.80	0.21	0.83	0.23	0.78	0.30
All	590	0.84	0.17	0.84	0.21	0.83	0.25

Note. For each neural network, the probability of detection, POD, and the false alarm ratio, FAR, are shown. The test set patches and thus also the contained cold pool objects (CPOs) to be detected are identical for all networks. Note that only patches with at least one CPO in the ground truth were evaluated here.

for the smallest area class is lower (Figure 4d). Again, a clear improvement in detection cannot be achieved for either of the three models, even though a slight improvement is seen for pseudo-3D models for the intermediate area classes.

Table 2 shows both POD and the false alarm ratio (FAR) for each simulation. Whereas POD is similarly good for all models, the pseudo-3D networks feature higher and thus worse FAR of 0.21 (p3D3t) and 0.25 (p3D5t), compared to 0.17 of the 2D network. In case of the p3D5t model, this means on average one spuriously predicted CPO for every three successfully detected CPOs. However, as the mean validation losses of the p3D5t training runs are lowest in comparison to the other models (Figure 3c), this should not be a problematic characteristic for p3D5t, but is most likely caused by an unfavorable epoch to stop the training run. Apart from lower detection rates for CPOs from "diu2K," which are mainly attributed to a high proportion of CPOs in the smallest area class, the performance of the networks seems to be relatively independent w.r.t. the simulation setup.

As the morphology of patterns is so diverse and quantification of spatial pattern overlap always requires to make choices as to the metrics used, we also provide a qualitative discussion on typical cases now. We visualize several predictions based on the test set and present 2D fields of rainfall intensity (r_{int}), cloud top temperature (T_{CT}), and virtual temperature anomaly (ΔT_v) as well as the ground truth segmentation and predictions of the three neural network models side by side (Figure 5). The cases selected represent a range of circumstances: in some cases, cloud patterns are rather obvious and yield reasonable segmentation for all models (Figure 5a). Only in a few cases, some models miss CPOs completely (Figure 5b). As in the presented example, these CPOs are generally rather small and weak, and often associated with cloud-free gust fronts. Where different aspects overlap temporally, such as cirrus from previous convection obscuring the present scene (Figure 5c), all models may struggle with proper segmentation. In fact, cirrus clouds are a major source of false positives, as all models associate very cold T_{CT} with CPs. However, whether cirrus clouds eventually lead to false positives depends also on their pattern. For the pseudo-3D models, simultaneous advection seems to increase the probability of false positives. Although cases with advection pose additional challenges, all models perform well for large CPs with large cloud-free areas, for example, Figure 5d. Yet, for cases in which the parent convection partly dissipated (Figure 5e) or dissipates (Figure 5f) pseudo-3D models give results which are physically more accurate w.r.t. the plausibility of the gust front. The same seems to be true for scenes with advected parent convection (Figure 5g)—likely due to the fact that parts of the gust front are obscured when only using single patches, but revealed when taking a sequence of time steps into account. As a general outcome, all models perform reasonably well on the test cases described, yet, the distinction between 2D and pseudo-3D quality metrics is not as clear cut and should be assessed dependent on the scientific questions in focus.

3.2. Case Study: Detecting Cold Pools Over Senegal

As we trained the neural networks solely with data from idealized simulations, they are not intended for direct application to observational data. However, to gain a better understanding of potential challenges associated with transitioning to more realistic data, we now apply the neural networks to segment CPs in the case study data from Senegal (see Section 2.1). Due to its location in the transition zone from tropical savannah to arid steppe, sea breezes from the Atlantic Ocean, and orographic effects of the Northern Guinea Highlands, the region of the case study features complex cloud and rainfall patterns and is thus well-suited for a realistic trial. Contrary to the evaluation of the test set predictions, we here refrain from analyzing any quantitative metrics based on ground truth images. This way, we avoid omitting any patches and can thus test the network performance over the entire simulation domain, including patches with ambiguous scenes.

To obtain the predictions for the entire domain, we subdivided the simulation domain into patches of $n_p \times n_p$ pixels, which the networks can process, and recombined the network-generated segmentations. Unlike in the network test, we use the neural networks of all three training runs for the prediction of each model: Only if all three networks of a certain model segment a pixel as "CP," it is predicted as "CP." Otherwise, it is predicted as "no CP." Hereby, we can identify consistent features w.r.t. false negatives and positives, which were learned by a certain model during all training runs.

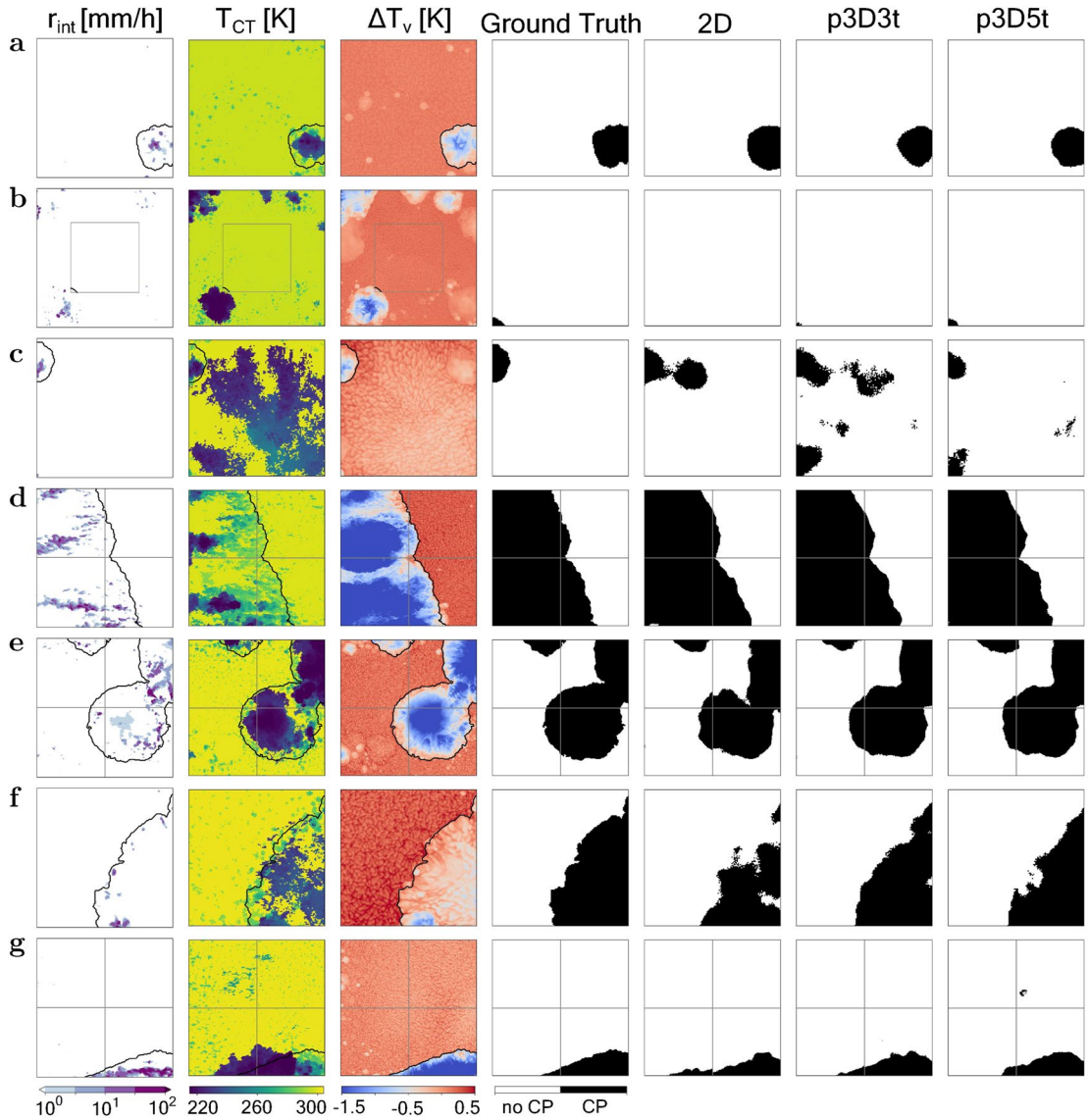

Figure 5. Examples of cold pool predictions based on the test set. Two-dimensional fields of surface rain intensity, r_{int}, cloud top temperature, T_{CT}, and near-surface virtual temperature anomaly, ΔT_v, for various examples, along with ground truth segmentations based on CoolDeTA, as well as predictions of the 2D and pseudo-3D neural networks. For comparison, black contours in r_{int}, T_{CT}, and ΔT_v indicate the boundary of the corresponding ground truth. (a) Morning cold pool (CP; time step 740) from "diu2K." (b) Analogous to (a) but for time step 744. For clarity, r_{int}, T_{CT}, and ΔT_v are plotted with their additional overlap while ground truth and predictions are only shown for the $n_p \times n_p$ pixel patch indicated by the gray frame. (c) CP from "diu2K" which developed during the afternoon (time step 780) at the boundary of a recently dissipated convective system, represented by high-altitude cirrus remnants. (d) Parts of an eastward propagating gust front of a convective system from "diu2K wind" (time step 772) with large cloud-free areas ($\gtrsim$300 km^2) and new emerging rain cells. (e) Afternoon scene (time step 772) from "diu4K" with parts of an early stage CP in the north of the upper left patch and parts of a convective system which consists of CPs at different stages. (f) Gust front of a convective system from "diu4K" (time step 780) with dissipating parent convection. (g) Northern part of a CP from "diu4K wind" (time step 780) where westward advected parent convection masks parts of its CP gust front. Note that superimposed grids represent the individual $n_p \times n_p$ pixel patches, processed by the neural networks.

In Figure 6, we present examples of the network predictions for the entire domain along with the corresponding 2D fields of r_{int}, T_{CT}, and ΔT_v. The patches processed by the networks are indicated by the superimposed grid. The selected cases represent CPs in different stages of their life cycle at different times of the day. In most of these cases, all neural networks detect the CPs and their gust fronts reasonably well. When the CP is small enough so that the networks can track its gust front (Figures 6a and 6e), the networks are even able to identify most of the associated CP region correctly, although parts of the gust front are obscured by deeper clouds (Figures 6a and 6e) or rain-free (Figure 6e). In later stages of the life cycle, when the CP is large compared to the patch size, the networks may struggle to properly detect CP regions where the gust front is already too far beyond the patch

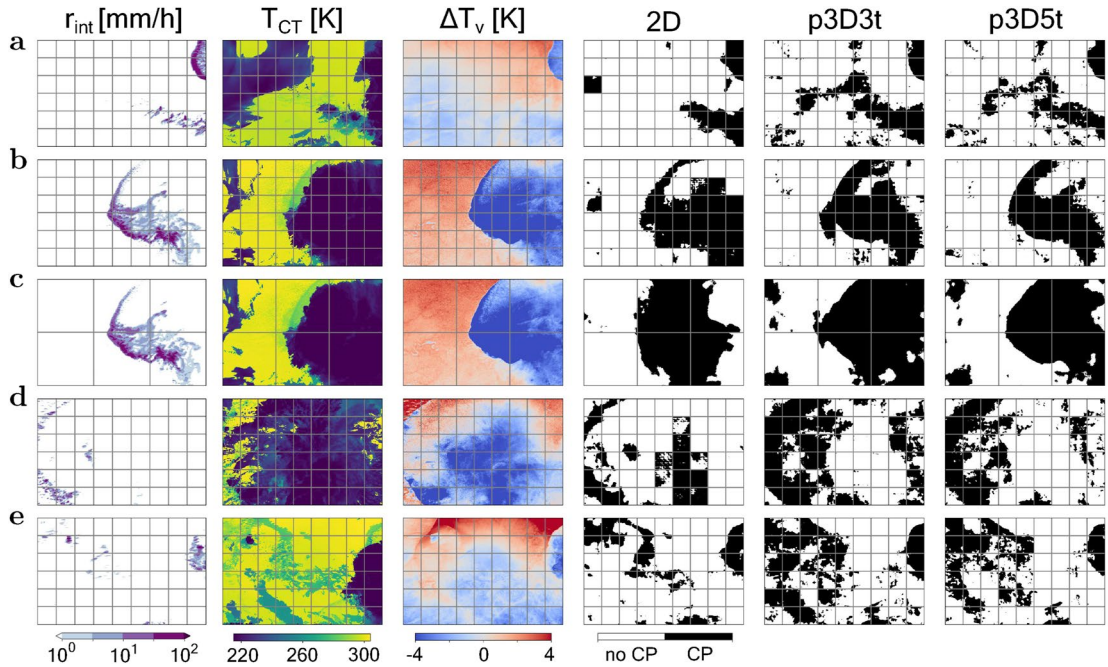

Figure 6. Examples of cold pool predictions for the case study. Two-dimensional fields of surface rain intensity, r_{int}, cloud top temperature, T_{CT}, and near-surface virtual temperature anomaly, ΔT_v, for various examples, along with predictions of the 2D and pseudo-3D neural networks. (a) Early stage cold pool (CP) entering the northeastern part of the domain at 07:40 UTC. (b) The CP from (a) but in a mature stage at 11:10 UTC. (c) Analogous to (b) but with the resolution lowered by a factor of 3 by computing the mean values. (d) The CP from (a) but in a dissipating stage at 14:40 UTC. (e) Late evening scene (22:10 UTC) with two CPs entering the domain from the east and the south, a sea breeze coming from the west and the dissipating CP from (a–c) in the center. Note that superimposed grids represent the individual $n_p \times n_p$ pixel patches, processed by the neural networks.

boundaries (Figures 6b and 6d). Yet, enlarging the networks' field of view by lowering the resolution compensates for this effect for all networks (Figure 6c).

However, the example cases also reveal challenges: The segmentations of (dissipating) CPs in later stages can be somewhat noisy (Figures 6d and 6e). This applies to both large convective systems (Figure 6d) and smaller systems (southeast in Figure 6e). For large systems, the limited field of view of the neural networks can again play a role when the gust front is located too far outside the patch boundaries. In this regard, particularly the 2D networks tend to uniformly segment patches as only CP when the (cold) cloud cover is much larger than the patch (Figures 6a, 6b, and 6d), possibly due to less contextual information compared to the pseudo-3D networks with multiple input time steps. Concerning false positive classifications, organized low-level clouds seem to constitute a potential source (Figures 6a and 6e), especially when associated with rain. As a result, particularly the pseudo-3D networks struggle to distinguish the sea breeze coming from the west in Figure 6e from CPs.

4. Conclusion and Outlook

CPs likely play a key role in organizing the atmospheric convective cloud and precipitation field (Böing, 2016; Böing et al., 2012; Haerter, 2019; Haerter et al., 2019, 2020; Muller et al., 2022; Nissen & Haerter, 2021; Schlemmer & Hohenegger, 2016). Robust detection of CP processes leading to the formation of thunderstorm clusters could enable better understanding of how convective systems organize through the interaction of CPs, such as lifting and collision processes, and how heavy precipitation events associated with MCS emerge.

The present study demonstrates that CPs can be detected in simulation data via an artificial neural network by employing variables readily available from geostationary satellite observations, namely cloud top temperature and precipitation. Using these two variables only, our networks were able to detect CPs in data from cloud-resolving simulations with an overall mean accuracy between 93.8% (2D) and 94.8% (p3D3t) for patches with at least one pixel of both classes, ≥99.8% for patches without any pixel of class "CP," and between 85.9% (p3D5t) and 94.1% (p3D3t) for patches with pixel of class "CP" only.

We conducted several experiments to identify the most promising architecture for our network. The computationally most expensive architecture, using 6 blocks and 64 starting filters, performed best, as might be expected—given the physical insight that CPs over tropical land are often linked to organized convective systems with spatial and temporal correlations at different scales. Whereas already the two-dimensional input fields gave satisfactory results, we find that taking into account three to five time steps does improve the performance further, comparable to the improvements found in Vu et al. (2020) for some of their data sets. Including several time steps within the input channels is a computationally inexpensive means of mimicking a three-dimensional input data set.

The training and test data sets contain data from different simulation setups, which correspond to an atmosphere over an idealized moist tropical land surface. The comparison between the two diurnal forcings is important as results show qualitatively different cloud organization, such as the formation of pronounced convective systems for a larger diurnal range but more scattered, smaller CPs for a smaller range. Assessing large-scale wind effects is important, as it compares the prominent model idealization of no wind shear (Bretherton et al., 2005; Manabe et al., 1965; Tompkins & Craig, 1998) with the more realistic sheared case. In the Senegal case study, we additionally applied the trained neural networks to simulation output from a realistic simulation setup. Our overall finding is that the detection works well for all these cases.

While the trained neural networks reliably detected gust fronts obscured by higher clouds in most situations, we identified several potential sources of misclassification. Concerning false positives or spuriously predicted CPs, the most common sources are (cirrus) clouds with very cold T_{CT} and organized low-level clouds, particularly in scenes with simultaneous advection and/or rain. When transitioning to actual satellite data, the former could be addressed by replacing the cloud top temperature input with satellite channels or products that also respond to cloud opacity. Focusing on false negatives, that is, missed CPs or CP pixels, one of the main sources are patches in the center of larger convective systems where the CP gust front is too far beyond the patch boundaries and thus out of the networks' sight. Yet, we showed that enlarging the networks' field of view by lowering the image resolution can compensate for this effect to some extent. Another source of false negatives are CP gust fronts without any signal in the cloud or rainfall field, that is, cloud-free gust fronts without any rain. However, unlike in our idealized simulations, CPs in more realistic simulations without a fixed surface temperature or even in satellite data would be noticeable in such situations based on their colder air compared to their environment. When transitioning to such data, this characteristic could be learned by the neural networks and used for the detection of CPs in cases with clear sky gust fronts.

Looking ahead, the obvious next step is to apply the method to actual satellite data. To make full use of the available satellite channels and address the identified sources of misclassification, this step will most likely involve new network training based on real observations. Likely, several new challenges will need to be addressed, such as the lower spatial resolution of the available data. The lower resolution may require us to focus on CPs that have already evolved into larger-scale structures, thus increasing the minimum detectable CP size. Yet, by training the neural networks with some of the multiple available satellite channels instead of the cloud top temperature input used so far, the neural network performance may benefit from additional information about the atmosphere, such as water vapor content, cloud phase, and cloud particle size and make use of potential patterns hidden so far.

To avoid inconsistencies between the neural network inputs w.r.t. their spatial and temporal resolutions, the selected satellite channels could be combined with a precipitation product based on calibrated infrared images. Such precipitation products are derived by calibrating infrared images from geostationary satellites with rain rates from low Earth orbiting passive microwave satellite sensors. While not applicable for low rain rates, a precipitation product based on calibrated infrared images might provide a sufficient estimate to detect CPs, particularly over tropical land. Considering the case-dependent accuracy of these products as well as potential errors associated with spatial and temporal interpolations when using IMERG data, it might be worth testing the neural networks without precipitation input in future studies based on satellite data.

Ultimately, being able to extract self-organization effects from observational data will enable us to improve cloud-resolving models that still struggle to capture organizational effects with high fidelity. For this purpose, the network training should additionally focus on minimizing the number of spuriously predicted CPs, for example, by adding more examples of organized and/or precipitating clouds to the training set that do not produce any CPs. By enlarging the variety of CPs in the data set, also the applicability of the method could be extended. One way forward could be to advance CP interaction parameterizations in coarser-scale models.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

Figures were made with Matplotlib version 3.5.2 (Caswell et al., 2022; Hunter, 2007) and seaborn version 0.12.2 (Waskom, 2021). The neural network was implemented using PyTorch version 1.11.0 (Paszke et al., 2019). Both the code and the data sets for the training and testing of the neural networks were used in version 1.0 and are licensed under Creative Commons Attribution 4.0 International (Hoeller, Fiévet, Engelbrecht, & Haerter, 2023). The idealized simulations used for training and testing are run with the cloud-resolving three-dimensional atmosphere simulator System for Atmospheric Modeling (SAM; Khairoutdinov & Randall, 2003), version 6.11. For the Senegal case study simulation, we used the nonhydrostatic Advanced Research Weather Research and Forecasting (WRF) model version 4.3 (Skamarock et al., 2021). Ground truth labels were derived from this simulation output based on the cold pool detection and tracking algorithm by Hoeller, Fiévet, and Haerter (2023).

Acknowledgments
The authors gratefully acknowledge funding by a grant from the VILLUM Foundation (Grant 13168) and the European Research Council (ERC) under the European Union' Horizon 2020 research and innovation program (Grant 771859), and the Novo Nordisk Foundation Interdisciplinary Synergy Program (Grant NNF19OC0057374). This work used resources of the Deutsches Klimarechenzentrum (DKRZ), granted by its Scientific Steering Committee (WLA) under project ID bb1166. Open Access funding enabled and organized by Projekt DEAL.

References

Atlas, D. (1994). Footprints of storms on the sea: A view from spaceborne synthetic aperture radar. *Journal of Geophysical Research*, *99*(C4), 7961–7969. https://doi.org/10.1029/94JC00250

Benjamin, T. B. (1968). Gravity currents and related phenomena. *Journal of Fluid Mechanics*, *31*(2), 209–248. https://doi.org/10.1017/s0022112068000133

Böing, S. J. (2016). An object-based model for convective cold pool dynamics. *Mathematics of Climate and Weather Forecasting*, *2*(1), 43–60. https://doi.org/10.1515/mcwf-2016-0003

Böing, S. J., Jonker, H. J., Siebesma, A. P., & Grabowski, W. W. (2012). Influence of the subcloud layer on the development of a deep convective ensemble. *Journal of the Atmospheric Sciences*, *69*(9), 2682–2698. https://doi.org/10.1175/JAS-D-11-0317.1

Borque, P., Nesbitt, S. W., Trapp, R. J., Lasher-Trapp, S., & Oue, M. (2020). Observational study of the thermodynamics and morphological characteristics of a midlatitude continental cold pool event. *Monthly Weather Review*, *148*(2), 719–737. https://doi.org/10.1175/mwr-d-19-0068.1

Brandes, E. A. (1977). Gust front evolution and tornado genesis as viewed by Doppler radar. *Journal of Applied Meteorology*, *16*(4), 333–338. https://doi.org/10.1175/1520-0450(1977)016<0333:GFEATG>2.0.CO;2

Bretherton, C. S., Blossey, P. N., & Khairoutdinov, M. (2005). An energy-balance analysis of deep convective self-aggregation above uniform SST. *Journal of the Atmospheric Sciences*, *62*(12), 4273–4292. https://doi.org/10.1175/jas3614.1

Cafaro, C., & Rooney, G. G. (2018). Characteristics of colliding density currents: A numerical and theoretical study. *Quarterly Journal of the Royal Meteorological Society*, *144*(715), 1761–1771. https://doi.org/10.1002/qj.3337

Caswell, T. A., Droettboom, M., Lee, A., de Andrade, E. S., Hoffmann, T., Klymak, J., et al. (2022). matplotlib/matplotlib: Release (version 3.5.2) [Software]. Zenodo. https://doi.org/10.5281/zenodo.6513224

Caton Harrison, T., Washington, R., & Engelstaedter, S. (2021). Satellite-derived characteristics of Saharan cold pool outflows during boreal summer. *Journal of Geophysical Research: Atmospheres*, *126*, e2020JD033387. https://doi.org/10.1029/2020JD033387

Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., & Haworth, A. (2021). A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, *65*(5), 545–563. https://doi.org/10.1111/1754-9485.13261

de Szoeke, S. P., Skyllingstad, E. D., Zuidema, P., & Chandra, A. S. (2017). Cold pools and their influence on the tropical marine boundary layer. *Journal of the Atmospheric Sciences*, *74*(4), 1149–1168. https://doi.org/10.1175/jas-d-16-0264.1

Drager, A. J., Grant, L. D., & van den Heever, S. C. (2020). Cold pool responses to changes in soil moisture. *Journal of Advances in Modeling Earth Systems*, *12*, e2019MS001922. https://doi.org/10.1029/2019MS001922

Drager, A. J., & van den Heever, S. C. (2017). Characterizing convective cold pools. *Journal of Advances in Modeling Earth Systems*, *9*, 1091–1115. https://doi.org/10.1002/2016MS000788

Droegemeier, K. K., & Wilhelmson, R. B. (1985). Three-dimensional numerical modeling of convection produced by interacting thunderstorm outflows. Part I: Control simulation and low-level moisture variations. *Journal of the Atmospheric Sciences*, *42*(22), 2381–2403. https://doi.org/10.1175/1520-0469(1985)042<2381:TDNMOC>2.0.CO;2

Droegemeier, K. K., & Wilhelmson, R. B. (1987). Numerical simulation of thunderstorm outflow dynamics. Part I: Outflow sensitivity experiments and turbulence dynamics. *Journal of the Atmospheric Sciences*, *44*(8), 1180–1210. https://doi.org/10.1175/1520-0469(1987)044<1180:NSOTOD>2.0.CO;2

Emmel, C., Knippertz, P., & Schulz, O. (2010). Climatology of convective density currents in the southern foothills of the Atlas Mountains. *Journal of Geophysical Research*, *115*, D11115. https://doi.org/10.1029/2009JD012863

Engerer, N. A., Stensrud, D. J., & Coniglio, M. C. (2008). Surface characteristics of observed cold pools. *Monthly Weather Review*, *136*(12), 4839–4849. https://doi.org/10.1175/2008mwr2528.1

Feng, Z., Hagos, S., Rowe, A. K., Burleyson, C. D., Martini, M. N., & Szoeke, S. P. (2015). Mechanisms of convective cloud organization by cold pools over tropical warm ocean during the AMIE/DYNAMO field campaign. *Journal of Advances in Modeling Earth Systems*, *7*, 357–381. https://doi.org/10.1002/2014MS000384

Fiévet, R., Meyer, B., & Haerter, J. O. (2022). On the sensitivity of convective cold pools to mesh resolution. *Earth and Space Science Open Archive*, *24*. https://doi.org/10.1002/essoar.10512297.1

Fiévet, R., Meyer, B., & Haerter, J. O. (2023). On the sensitivity of convective cold pools to mesh resolution. *Journal of Advances in Modeling Earth Systems*, *15*, e2022MS003382. https://doi.org/10.1029/2022MS003382

Fournier, M. B., & Haerter, J. O. (2019). Tracking the gust fronts of convective cold pools. *Journal of Geophysical Research: Atmospheres*, *124*, 11103–11117. https://doi.org/10.1029/2019JD030980

Fritsch, J. M., & Carbone, R. (2004). Improving quantitative precipitation forecasts in the warm season: A USWRP research and development strategy. *Bulletin of the American Meteorological Society*, *85*(7), 955–966. https://doi.org/10.1175/bams-85-7-955

Fuglestvedt, H. F., & Haerter, J. O. (2020). Cold pools as conveyor belts of moisture. *Geophysical Research Letters*, *47*, e2020GL087319. https://doi.org/10.1029/2020GL087319

Garg, P., Nesbitt, S. W., Lang, T. J., Priftis, G., Chronis, T., Thayer, J. D., & Hence, D. A. (2020). Identifying and characterizing tropical oceanic mesoscale cold pools using spaceborne scatterometer winds. *Journal of Geophysical Research: Atmospheres*, *125*, e2019JD031812. https://doi.org/10.1029/2019JD031812

Gentine, P., Garelli, A., Park, S., Nie, J., Torri, G., & Kuang, Z. (2016). Role of surface heat fluxes underneath cold pools. *Geophysical Research Letters*, *43*, 874–883. https://doi.org/10.1002/2015GL067262

Grandpeix, J.-Y., & Lafore, J.-P. (2010). A density current parameterization coupled with Emanuel's convection scheme. Part I: The models. *Journal of the Atmospheric Sciences*, *67*(4), 881–897. https://doi.org/10.1175/2009jas3044.1

Haerter, J. O. (2019). Convective self-aggregation as a cold pool-driven critical phenomenon. *Geophysical Research Letters*, *46*, 4017–4028. https://doi.org/10.1029/2018GL081817

Haerter, J. O., Böing, S. J., Henneberg, O., & Nissen, S. B. (2019). Circling in on convective organization. *Geophysical Research Letters*, *46*, 7024–7034. https://doi.org/10.1029/2019GL082092

Haerter, J. O., Meyer, B., & Nissen, S. B. (2020). Diurnal self-aggregation. *npj Climate and Atmospheric Science*, *3*(1), 30. https://doi.org/10.1038/s41612-020-00132-z

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/tkde.2008.239

Henneberg, O., Meyer, B., & Haerter, J. O. (2020). Particle-based tracking of cold pool gust fronts. *Journal of Advances in Modeling Earth Systems*, *12*, e2019MS001910. https://doi.org/10.1029/2019MS001910

Hoeller, J., Fiévet, R., Engelbrecht, E., & Haerter, J. (2023). Cold pool detection U-Net release (version 1.0) [Software and Dataset]. Zenodo. https://doi.org/10.5281/zenodo.8376599

Hoeller, J., Fiévet, R., & Haerter, J. O. (2023). Detecting cold pool family trees in convection resolving simulations [Software]. *Authorea Preprints*. https://doi.org/10.22541/essoar.167768111.14938799/v1

Houze, R. A. Jr. (2004). Mesoscale convective systems. *Reviews of Geophysics*, *42*, RG4003. https://doi.org/10.1029/2004RG000150

Houze, R. A. Jr., & Betts, A. K. (1981). Convection in gate. *Reviews of Geophysics*, *19*(4), 541–576. https://doi.org/10.1029/RG019i004p00541

Huffman, G. J., Bolvin, D. T., Braithwaite, D., Hsu, K., Joyce, R., Xie, P., & Yoo, S.-H. (2015). NASA Global Precipitation Measurement (GPM) Integrated Multi-satellitE Retrievals for GPM (IMERG). *Algorithm Theoretical Basis Document (ATBD) Version*, *4*(26).

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning* (pp. 448–456).

Jadon, S. (2020). A survey of loss functions for semantic segmentation. In *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)* (pp. 1–7).

Jensen, G. G., Fiévet, R., & Haerter, J. O. (2022). The diurnal path to persistent convective self-aggregation. *Journal of Advances in Modeling Earth Systems*, *14*, e2021MS002923. https://doi.org/10.1029/2021MS002923

Khairoutdinov, M. F., & Randall, D. A. (2003). Cloud resolving modeling of the ARM summer 1997 IOP: Model formulation, results, uncertainties, and sensitivities. *Journal of the Atmospheric Sciences*, *60*(4), 607–625. https://doi.org/10.1175/1520-0469(2003)060<0607:CRMOTA>2.0.CO;2

Kirsch, B., Ament, F., & Hohenegger, C. (2021). Convective cold pools in long-term boundary layer mast observations. *Monthly Weather Review*, *149*(3), 811–820. https://doi.org/10.1175/mwr-d-20-0197.1

Kneller, B. C., Bennett, S. J., & McCaffrey, W. D. (1999). Velocity structure, turbulence and fluid stresses in experimental gravity currents. *Journal of Geophysical Research*, *104*(C3), 5381–5391. https://doi.org/10.1029/1998JC900077

Knippertz, P., Deutscher, C., Kandler, K., Müller, T., Schulz, O., & Schütz, L. (2007). Dust mobilization due to density currents in the Atlas region: Observations from the Saharan Mineral Dust Experiment 2006 field campaign. *Journal of Geophysical Research*, *112*, D21109. https://doi.org/10.1029/2007JD008774

Kruse, I. L., Haerter, J. O., & Meyer, B. (2022). Cold pools over the Netherlands: A statistical study from tower and radar observations. *Quarterly Journal of the Royal Meteorological Society*, *148*(743), 711–726. https://doi.org/10.1002/qj.4223

Lundgren, T., Yao, J., & Mansour, N. (1992). Microburst modelling and scaling. *Journal of Fluid Mechanics*, *239*, 461–488. https://doi.org/10.1017/s002211209200449x

Manabe, S., Smagorinsky, J., & Strickler, R. F. (1965). Simulated climatology of a general circulation model with a hydrologic cycle. *Monthly Weather Review*, *93*(12), 769–798. https://doi.org/10.1175/1520-0493(1965)093<0769:SCOAGC>2.3.CO;2

Markowski, P., & Richardson, Y. (2011). *Mesoscale meteorology in midlatitudes* (Vol. 2). John Wiley & Sons.

Meyer, B., & Haerter, J. O. (2020). Mechanical forcing of convection by cold pools: Collisions and energy scaling. *Journal of Advances in Modeling Earth Systems*, *12*, e2020MS002281. https://doi.org/10.1029/2020MS002281

Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)* (pp. 565–571).

Muller, C., Yang, D., Craig, G., Cronin, T., Fildier, B., Haerter, J. O., et al. (2022). Spontaneous aggregation of convective storms. *Annual Review of Fluid Mechanics*, *54*(1), 133–157. https://doi.org/10.1146/annurev-fluid-022421-011319

Nissen, S. B., & Haerter, J. O. (2021). Circling in on convective self-aggregation. *Journal of Geophysical Research: Atmospheres*, *126*, e2021JD035331. https://doi.org/10.1029/2021JD035331

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* (Vol. 32, pp. 8024–8035). Curran Associates, Inc. Retrieved from http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Redl, R., Fink, A. H., & Knippertz, P. (2015). An objective detection method for convective cold pool events and its application to northern Africa. *Monthly Weather Review*, *143*(12), 5055–5072. https://doi.org/10.1175/mwr-d-15-0223.1

Rio, C., Hourdin, F., Grandpeix, J.-Y., & Lafore, J.-P. (2009). Shifting the diurnal cycle of parameterized deep convection over land. *Geophysical Research Letters*, *36*, L07809. https://doi.org/10.1029/2008GL036779

Romps, D. M., & Jeevanjee, N. (2016). On the sizes and lifetimes of cold pools. *Quarterly Journal of the Royal Meteorological Society*, *142*(696), 1517–1527. https://doi.org/10.1002/qj.2754

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (Vol. 9351, pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28

Rotunno, R., Klemp, J. B., & Weisman, M. L. (1988). A theory for strong, long-lived squall lines. *Journal of the Atmospheric Sciences*, *45*(3), 463–485. https://doi.org/10.1175/1520-0469(1988)045<0463:ATFSLL>2.0.CO;2

Schlemmer, L., & Hohenegger, C. (2016). Modifications of the atmospheric moisture field as a result of cold-pool dynamics. *Quarterly Journal of the Royal Meteorological Society*, *142*(694), 30–42. https://doi.org/10.1002/qj.2625

Schumacher, R. S., & Johnson, R. H. (2008). Mesoscale processes contributing to extreme rainfall in a midlatitude warm-season flash flood. *Monthly Weather Review*, *136*(10), 3964–3986. https://doi.org/10.1175/2008mwr2471.1

Schumacher, R. S., & Rasmussen, K. L. (2020). The formation, character and changing nature of mesoscale convective systems. *Nature Reviews Earth & Environment*, *1*(6), 300–314. https://doi.org/10.1038/s43017-020-0057-7

Sharifnezhadazizi, Z., Norouzi, H., Prakash, S., Beale, C., & Khanbilvardi, R. (2019). A global analysis of land surface temperature diurnal cycle using MODIS observations. *Journal of Applied Meteorology and Climatology*, *58*(6), 1279–1291. https://doi.org/10.1175/jamc-d-18-0256.1

Shi, J., Dang, J., Cui, M., Zuo, R., Shimizu, K., Tsunoda, A., & Suzuki, Y. (2021). Improvement of damage segmentation based on pixel-level data balance using VGG-Unet. *Applied Sciences*, *11*(2), 518. https://doi.org/10.3390/app11020518

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, *6*(1), 60. https://doi.org/10.1186/s40537-019-0197-0

Simpson, J. (1980). Downdrafts as linkages in dynamic cumulus seeding effects. *Journal of Applied Meteorology*, *19*(4), 477–487. https://doi.org/10.1175/1520-0450(1980)019<0477:DALIDC>2.0.CO;2

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Liu, Z., Berner, J., et al. (2021). A description of the advanced research WRF model release (version 4.3) (No. NCAR/TN-556+STR). https://doi.org/10.5065/1dfh-6p97

Stevens, B., Bony, S., Farrell, D., Ament, F., Blyth, A., Fairall, C., et al. (2021). EUREC^{4}A. *Earth System Science Data Discussions*, *2021*, 1–78.

Sukovich, E. M., Ralph, F. M., Barthold, F. E., Reynolds, D. W., & Novak, D. R. (2014). Extreme quantitative precipitation forecast performance at the weather prediction center from 2001 to 2011. *Weather and Forecasting*, *29*(4), 894–911. https://doi.org/10.1175/waf-d-13-00061.1

Tan, J., Jakob, C., Rossow, W. B., & Tselioudis, G. (2015). Increases in tropical rainfall driven by changes in frequency of organized deep convection. *Nature*, *519*(7544), 451–454. https://doi.org/10.1038/nature14339

Tompkins, A. M. (2001). Organization of tropical convection in low vertical wind shears: The role of cold pools. *Journal of the Atmospheric Sciences*, *58*(13), 1650–1672. https://doi.org/10.1175/1520-0469(2001)058⟨1650:OOTCIL⟩2.0.CO;2

Tompkins, A. M., & Craig, G. C. (1998). Radiative–convective equilibrium in a three-dimensional cloud-ensemble model. *Quarterly Journal of the Royal Meteorological Society*, *124*(550), 2073–2097. https://doi.org/10.1002/qj.49712455013

Torri, G., & Kuang, Z. (2016). Rain evaporation and moist patches in tropical boundary layers. *Geophysical Research Letters*, *43*, 9895–9902. https://doi.org/10.1002/2016GL070893

Torri, G., & Kuang, Z. (2019). On cold pool collisions in tropical boundary layers. *Geophysical Research Letters*, *46*, 399–407. https://doi.org/10.1029/2018GL080501

Torri, G., Kuang, Z., & Tian, Y. (2015). Mechanisms for convection triggering by cold pools. *Geophysical Research Letters*, *42*, 1943–1950. https://doi.org/10.1002/2015GL063227

Touzé-Peiffer, L., Vogel, R., & Rochetin, N. (2022). Cold pools observed during EUREC^{4}A: Detection and characterization from atmospheric soundings. *Journal of Applied Meteorology and Climatology*, *61*(5), 593–610. https://doi.org/10.1175/jamc-d-21-0048.1

Vogel, R., Konow, H., Schulz, H., & Zuidema, P. (2021). A climatology of trade-wind cumulus cold pools and their link to mesoscale cloud organization. *Atmospheric Chemistry and Physics*, *21*(21), 16609–16630. https://doi.org/10.5194/acp-21-16609-2021

Vu, M. H., Grimbergen, G., Nyholm, T., & Löfstedt, T. (2020). Evaluation of multislice inputs to convolutional neural networks for medical image segmentation. *Medical Physics*, *47*(12), 6216–6231. https://doi.org/10.1002/mp.14391

Wakimoto, R. M. (1982). The life cycle of thunderstorm gust fronts as viewed with Doppler radar and rawinsonde data. *Monthly Weather Review*, *110*(8), 1060–1082. https://doi.org/10.1175/1520-0493(1982)110<1060:TLCOTG>2.0.CO;2

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Wilbanks, M. C., Yuter, S. E., De Szoeke, S. P., Brewer, W. A., Miller, M. A., Hall, A. M., & Burleyson, C. D. (2015). Near-surface density currents observed in the southeast pacific stratocumulus-topped marine boundary layer. *Monthly Weather Review*, *143*(9), 3532–3555. https://doi.org/10.1175/mwr-d-14-00359.1

Willemink, M. J., Koszek, W. A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., et al. (2020). Preparing medical imaging data for machine learning. *Radiology*, *295*(1), 4–15. https://doi.org/10.1148/radiol.2020192224

Yamaguchi, T., Randall, D. A., & Khairoutdinov, M. F. (2011). Cloud modeling tests of the ultimate–macho scalar advection scheme. *Monthly Weather Review*, *139*(10), 3248–3264. https://doi.org/10.1175/mwr-d-10-05044.1

Zipser, E. J., Meitín, R. J., & LeMone, M. A. (1981). Mesoscale motion fields associated with a slowly moving gate convective band. *Journal of the Atmospheric Sciences*, *38*(8), 1725–1750. https://doi.org/10.1175/1520-0469(1981)038<1725:MMFAWA>2.0.CO;2

Zuidema, P., Torri, G., Muller, C., & Chandra, A. (2017). A survey of precipitation-induced atmospheric cold pools over oceans and their interactions with the larger-scale environment. *Surveys in Geophysics*, *38*(6), 1283–1305. https://doi.org/10.1007/s10712-017-9447-x