Climatology, characteristics and forcing mechanisms of warm-season cold-frontal convective cells in Germany

A Dissertation

Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor rerum naturalium (Dr. rer. nat.)

to the Department of Earth Sciences of Freie Universität Berlin

by

George Pacey

Berlin, 2024

Supervisor: Prof. Dr. Stephan Pfahl

Second examiner: Prof. Dr. Michael Kunz

Date of defense: 10.07.2024

Contents

Abstract

Cold fronts provide an environment particularly favourable for convective initiation in the mid-latitudes and can also be associated with convective hazards such as flooding, wind, hail and lightning. In this thesis, the climatology, characteristics and forcing mechanisms of cold-frontal convection are analysed by combining a radar-based cell detection and tracking dataset and an automatic front detection method applied to reanalysis data. The climatology indicates that around twice as many cells develop on cold-frontal cell days on average compared to non-cold-frontal cell days. The maximum cell frequency is found to be marginally ahead of the typical surface front location, whereas the 700 hPa front location marks the minimum cell frequency and a clear shift in regime between cells with a weakened diurnal cycle on the warm side of the 700 hPa cold front and strongly diurnally driven cells on the cold side of the 700 hPa front. Lightning strike data are analysed in Western and Central Europe and show broadly consistent results. Pre-surface-frontal cells have the largest fraction associated with mesocyclones, intense convective cores and lightning in Germany. Namely, mesocyclones were detected in around 5.0% of pre-surface-frontal cells compared to only 1.5% of non-cold-frontal cells.

To unravel the results found in the climatology, the thermodynamic environments and lifting mechanisms of convective cells are analysed by considering different variables from reanalysis data. Pre-surface-frontal cells have the highest dewpoints and convective available potential energy, which is likely a contributing factor towards the larger fraction of cells with mesocyclones, intense convective cores and lightning. Higher mid-level relative humidity than climatology was found to be particularly important to allow convective cell development on the cold-side of the 700 hPa front. While large-scale lifting is typically thought to be most relevant in the warm-sector, here its relevance is also highlighted on the cold-side of the 700 hPa front. Most cells have relatively low sunshine duration prior to development, which indicates that solar heating may not be directly relevant for the majority of cell initiation.

A more complex statistical modeling approach, incorporating one moisture, one instability, and three lifting terms, reveals a lower level of proficiency near the surface front, consistent across four distinct statistical models. Feature (predictor) importance techniques show differences in the importance of each feature depending on the front relative region. Furthermore, certain features add more skill depending on the statistical model used. Finally, a 12 feature gradient boosted model is trained to predict the probability of convective cell occurrence. The model is shown to generally represent the spatial and temporal frequency of cells around the front on new, unseen testing data.

The results in this thesis are an important step towards a deeper understanding of the drivers of cold-frontal convection depending on the region relative to the front. Furthermore, light is shed on the applications of statistical models to predict the convective cell occurrence.

Zusammenfassung

Kaltfronten bieten eine Umgebung, die für die Entwicklung konvektiver Zellen in den mittleren Breiten besonders geeignet ist und können auch mit konvektiven Gefahren wie Überschwemmungen, Wind, Hagel und Blitzen verbunden sein. In dieser Dissertation werden die Klimatologie, Charakteristika und Mechanismen solcher Konvektion analysiert, indem ein radarbasierter Erkennungs- und Verfolgungsdatensatz konvektiver Zellen in Kombination mit einer automatischen Fronterkennungsmethode auf Reanalysedaten angewandt wird. Die Klimatologie zeigt, dass sich im Durchschnitt etwa doppelt so viele Zellen an Tagen mit Kaltfronten entwickeln als an Tagen mit konvektiven Zellen ohne Kaltfront. Die maximale Zellenfrequenz liegt knapp vor der typischen Bodenfront, während die Position der 700-hPa-Kaltfront die minimale Zellenfrequenz markiert. Dabei ist ein deutlicher Regimewechsel zwischen Zellen mit einem abgeschwächten Tagesgang auf der warmen Seite und stark durch den Tagesgang getriebene Zellen auf der kalten Seite der 700-hPa-Kaltfront zu beobachten. Blitzdaten über West- und Mitteleuropa zeigen allgemein konsistente Ergebnisse. Vorläufig zur Bodenfront befindliche Zellen weisen den größten Anteil an Mesozyklonen, intensiven konvektiven Kernen sowie Blitzen in Deutschland auf. Mesozyklone wurden in etwa 5,0% dieser Zellen detektiert, während es nur 1,5% bei Zellen ohne Kaltfront waren.

Um die Ergebnisse der Klimatologie zu entwirren, werden die thermodynamischen Bedingungen und Hebungsmechanismen konvektiver Zellen unter Berücksichtigung verschiedener Variablen aus Reanalysedaten analysiert. Vorläufig zur Bodenfront befindliche Zellen weisen die höchsten Taupunkte sowie die höchste konvektiv verfügbare potentielle Energie auf, was wahrscheinlich ein Beitrag zum größeren Anteil an Zellen mit Mesozyklonen, intensiven konvektiven Kernen und Blitzen ist. Eine höhere relative Feuchtigkeit in der mittleren Troposphäre als die Klimatologie erwies sich als besonders wichtig, um die Entwicklung konvektiver Zellen auf der kalten Seite der 700-hPa-Front zu ermöglichen. Während großskalige Hebung typischerweise im warmen Sektor als relevantester Prozess angesehen wird, wird hier auch dessen Relevanz auf der kalten Seite der 700-hPa-Front hervorgehoben. Die meisten Zellen weisen eine relativ kurze Sonnenscheindauer vor ihrer Entwicklung auf, was darauf hindeutet, dass die solare Erwärmung möglicherweise für die Mehrheit der Zellbildung nicht direkt relevant ist.

Ein komplexerer statistischer Modellierungsansatz, der einen Feuchtigkeits-, Instabilitäts- und drei Hebungsterme einbezieht, zeigt eine geringere Leistungsfähigkeit in der Nähe der Bodenfront. Dieses Ergebnis ist konsistent für vier verschiedene statistische Modelle. Techniken zur Ermittlung der Bedeutung von Prädiktoren zeigen Unterschiede in der Bedeutung der Prädiktoren je nach relativer Region zur Front. Darüber hinaus tragen bestimmte Prädiktoren je nach verwendetem statistischen Modell unterschiedlich zur Genauigkeit bei. Schließlich wird ein 12-Prädiktor-Gradient-Boosting-Modell trainiert, um die Wahrscheinlichkeit des Auftretens konvektiver Zellen vorherzusagen. Das Modell repräsentiert im Allgemeinen die räumliche und zeitliche Häufigkeit von Zellen um die Front herum bei neuen, ungesehenen Testdaten.

Die Ergebnisse in dieser Dissertation sind ein wichtiger Schritt hin zu einem tieferen Verständnis der Treiber von Konvektion bei Kaltfronten je nach der relativen Region zur Front. Darüber hinaus wird ein Licht auf die Anwendungen statistischer Modelle zur Vorhersage des Auftretens konvektiver Zellen geworfen.

Preface

Convective Cells

According to the AMS Glossary of Meteorology "In the case of precipitating moist convection, [a convection cell] refers to a distinct unit of convection, often having its own closed contours of radar reflectivity and a lifetime of roughly 20–30 minutes". An example cross section through a convective cell during its three stage lifecycle is shown in Figure P1. During the mature stage, a convective cell has two primary components: an updraft and a downdraft. The updraft carries warm and moist rising air into the cell. The condensation and latent heat release further strengthens the updraft. When hydrometeors become sufficiently large, they will descend towards the surface creating the downdraft. Evaporative cooling of precipitation is an important process for accelerating the downdraft towards the surface since the air then becomes more dense and has stronger negative buoyancy. The end of the cell is due to the downdraft cutting off the updraft and hence the supply of warm and moist rising air. Therefore, separation of the updraft and downdraft is crucial for long-living convective systems, which can occur in environments with sufficient vertical wind shear (Markowski and Richardson, 2010). Convective cells are commonly associated with lightning and therefore by definition are commonly thunderstorms. Nevertheless, convective rain rates have also been observed in the absence of lightning (e.g. van den Broeke et al., 2005). For this reason, the terms convective cell and thunderstorm should not always be used interchangeably.

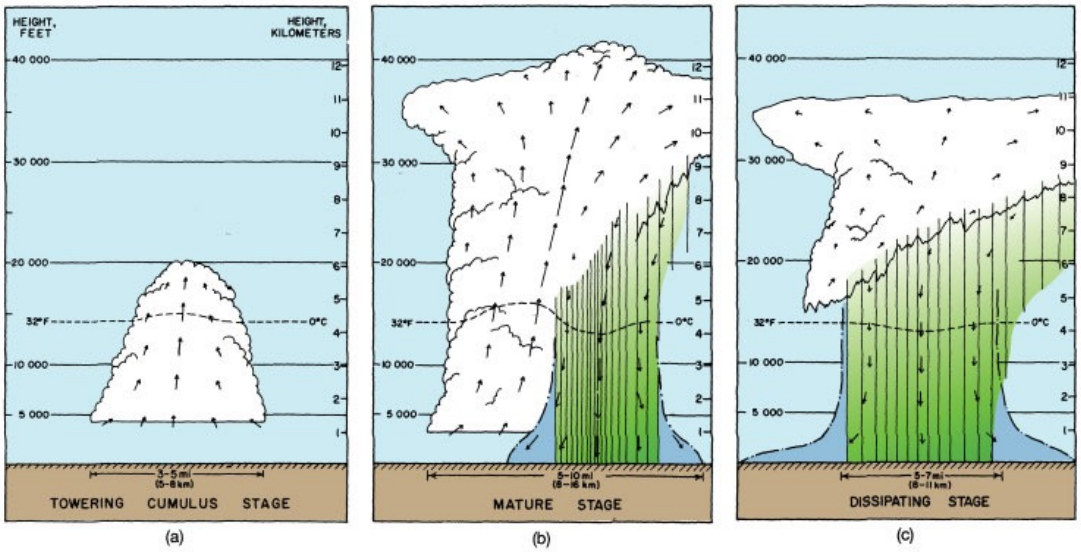


Figure P1: The three stages of a single convective cell: (a) towering cumulus stage, (b) mature stage, and (c) dissipating stage (Markowski and Richardson, 2010). Upward and downward pointing arrows represent the updraft and downdraft, respectively. Green regions represent precipitation.

The onset of deep precipitating cumulonimbus clouds (Figure P1b) is usually referred to as deep convective initiation or the onset of deep moist convection (DMC). Identifying exactly where and when this will occur, is one of the largest problems currently facing the meteorological community. This is despite the advent of increasingly high-resolution numerical models which can explicitly resolve convective initiation without the need for parameterization (e.g. Hagelin et al., 2017; Klasa et al., 2018; Reinert et al., 2023). For the last few decades, it has been widely accepted that DMC can only occur if three conditions (ingredients) are met: moisture, lift and instability (Doswell et al., 1996). That is to say, if one of the ingredients is missing then DMC will not and cannot occur. The availability of moisture and instability can be measured in a vertical profile such as a radiosonde using parameters such as specific humidity, dewpoints, lifted index and CAPE. However, determining whether parcels will be lifted to their level of free convection (LFC) is more complex and cannot be directly measured in a vertical profile. Furthermore, the lifting mechanisms that ultimately trigger convection occur on very small spatial scales and therefore are difficult to represent in numerical models (Yano et al., 2018). The initiation problem is further complicated by the fact that parcels reaching their LFC is not a guarantee of DMC due to the possibility of entrainment in the mid-levels, which in turn can dilute and weaken updrafts (e.g. Morrison et al., 2022). In such an event, only the towering cumulus stage may be reached (Figure P1a)

Convective cells have the potential to be associated with hazards such as hail, flooding and strong winds. Over the past few decades, convective loss events to the insurance industry have been steadily rising (e.g. Hoeppe, 2016), primarily due to increased exposure. The year of 2023 was yet another record in terms of losses due to SCSs (Munich Re, 2024). Germany, located in Central Europe, is particularly prone to such convective loss events with some events leading to insurance losses of over 1 billion euros (e.g. Kunz et al., 2018; Wilhelm et al., 2021). Foreseeing where convective cells will develop in time and space is crucial for mitigating the aforementioned hazards. This is because if parcels are not lifted to their LFC and thus no CAPE is released then there is no convective cell and no possibility of severe convective storms (SCSs). This is regardless of whether there is a seemingly favourable environment for SCSs (e.g. high CAPE and high vertical wind shear). If it were possible to forecast on a more local scale where and when convection will initiate, warnings could also be issued on a more local-scale. Forecasting convection (including non-severe) is also relevant for public planning and events particularly during the summer season when people often plan to be outside. The temporal scale on which convective showers and thunderstorms develop often gives the general public little time to react.

Cold fronts

Synoptic-scale cold fronts provide an environment particularly favourable for the development of convection and sometimes SCSs. The traditional definition of a front, which is often still used today, is a boundary between air masses (Bjerknes, 1919). These air masses have distinctly different properties such as temperature and humidity. Fronts are usually boundaries between warmer (often moister) and colder (often drier) air. Nevertheless, fronts characterised by strong humidity gradients but relatively small temperature gradients are also possible (e.g. split fronts; Browning and Monk, 1982). The exact definition of a front in terms of the thermodynamic field to use (e.g. potential temperature or equivalent potential temperature) and the strength of gradient is still a topic of discussion in the literature (Thomas and Schultz, 2019). Synoptic-scale fronts appear as elongated bands of high temperature and moisture gradients often extending from the centre of extratropical cyclones. The separation into warm, cold and stationary fronts depends on which air mass

is advancing. In the case of cold fronts, cold air is advancing and undercutting warmer air. The warm air is forced to rise and often leads to condensation and precipitation near the cold-frontal boundary. This ascending airstream is often referred to as the warm conveyor belt (WCB). WCBs start in the boundary layer rising through the mid-troposphere and end up in the upper troposphere (Browning, 1986). The WCB is depicted by the red arrow in Figure P2. In a classical Ana cold front situation, since the WCB has a rearward component relative to the frontal motion, the primary frontal precipitation band is after the surface cold-frontal boundary has passed (Figure P2). The primary precipitation band refers to mostly stratiform precipitation that results from condensation in the ascending WCB. On the other hand, convective precipitation has been observed to occur simultaneously ahead of the surface front (typically called pre-frontal), near the front (frontal) and behind the front (post-frontal) (e.g. Ferretti et al., 2014). The climatology, characteristics and forcing mechanisms of convection at different regions relative to the front will be a key focus of this thesis.

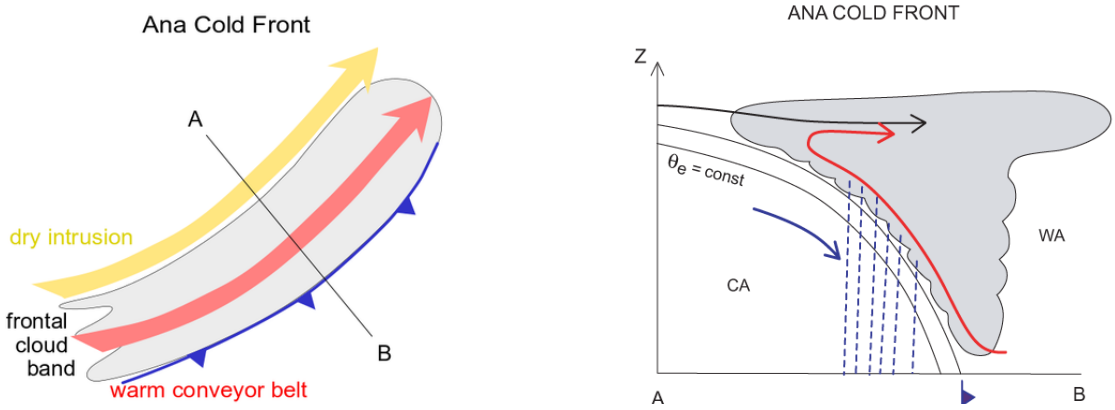


Figure P2: Conceptual model of air streams around an Ana cold front (left) and horizontal cross section through a cold front (right) (EUMeTrain, 2012). The red and yellow arrows refer to the warm conveyor belt and dry intrusion, respectively.

Forecasters are primarily interested in synoptic-scale fronts as they are the main driver of precipitation in the mid-latitudes (Catto and Pfahl, 2013), but cold fronts do not only exist on the synoptic-scale. The downdraft in a convective cell carries cooler air towards the surface. Upon reaching the surface, it spreads out laterally and forms a boundary with the warmer environmental air. The leading edge of the cold pool and the outflow is the gust front (Trapp, 2013). Similar to synoptic-scale cold fronts, gust fronts are also often associated with a wind shift and a drop in temperature. Collisions of outflow boundaries with pre-existing boundaries can provide additional convergence and lift thus favouring further convection in the vicinity of the original convective cell (e.g. Wilson and Schreiber, 1986). Therefore, it is possible to have smaller-scale cold-frontal boundaries embedded in the synoptic-scale cold-frontal environment.

Cold-frontal convection

A key characteristic of synoptic-scale cold fronts is the lifting due to the ageostrophic frontal circulation. The large-scale lifting in proximity to fronts is typically of the order of cm s^{-1} (Trapp, 2013; his chapter 5.2). With such an ascent rate, a surface based parcel would reach its LFC in the period of around a day.

Therefore, large-scale lifting alone is unlikely sufficient to allow air parcels to reach their LFC and trigger DMC. However, higher resolution simulations of fronts have indicated that the lift associated with the cross frontal circulation in localised regions along the front may be 1 order of magnitude higher (e.g. Sanders, 1955; Koch, 1984), thus may contribute to overcoming CIN. Additional smaller-scale sources of lift such as localised solar heating and orographic lift combined with the frontal lift may ultimately result in convection initiation (Doswell, 2001). Understanding the links between the different spatial scales is crucial in gaining a better understanding of the spatial and temporal onset of convective initiation. While large-scale lifting alone is not likely sufficient to lift parcels to their LFC, it can steepen lapse rates and thus increase CAPE due to adiabatic cooling (Trapp, 2013; his Figure 5.2).

Arguably, a disproportional fraction of all convective hazards worldwide occur in proximity to synoptic-scale cold fronts. The perhaps most famous example is the meeting of the dryline and a synoptic-scale cold front during spring in the Great Plains of the United States. The dryline often lies north-south separating dry and moist air masses. The meeting of the dryline and a synoptic-scale cold front is a classic setup for SCSs (Schultz et al., 2014), though additional conditions must be met thus not all cases are associated with SCSs. The importance of the front on convective initiation and the potential for SCSs is not limited to the large-scale lifting. The advection ahead of the front carries moisture rich air from the Gulf of Mexico and the upper-level flow also advects hot and dry air layers with steep lapse rates from the Rocky Mountains. Such layers are often referred to as Elevated Mixed Layers or EMLs (Carlson et al., 1983). In Europe, the dryline is a less common feature however warm-season cold-frontal convection is also sometimes associated with EMLs (Dahl and Fischer, 2016). In Europe, the source of the EMLs is typically the Spanish plateau and/or northern Africa. Such events are sometimes referred to as the "Spanish Plume" (Lewis and Gray, 2010). Aside from the potential for large instability to be released due to the steep upper level lapse rates, a pre-surface-frontal convergence line is often found below the EML (Dahl and Fischer, 2016), which is relevant for convective initiation. The southerly flow ahead of the front also carries moist rich air from the Mediterranean. Cold fronts also provide vertical wind shear, a prerequisite for organised convective storms (Markowski and Richardson, 2010). According to thermal wind balance, when horizontal temperature gradients are present, the geostrophic wind must have vertical wind shear (Holton, 2004). As the temperature gradient increases, the vertical shear of the zonal wind must also increase to maintain geostrophic balance.

Cold-frontal convection has attracted several previous studies in Europe, a large fraction of which has focused on line convection (e.g. James and Browning, 1979; Fink et al., 2009; Gatzen, 2011; Clark, 2013). Line convection is also referred to as narrow cold-frontal rainbands (NCFRs). These narrow convective lines observed on radar data often occur along the surface frontal boundary and are primarily cool-season phenomena. The aforementioned studies alluded to the importance of stronger forcing on the development of the convective line. Gatzen (2011) noted that NCFRs in Germany are often associated with severe weather. An example cross section through a cold front associated with a NCFR is shown in Figure P3. The NCFR was observed directly at the surface frontal boundary with additional convection embedded in a cold-frontal rainband.

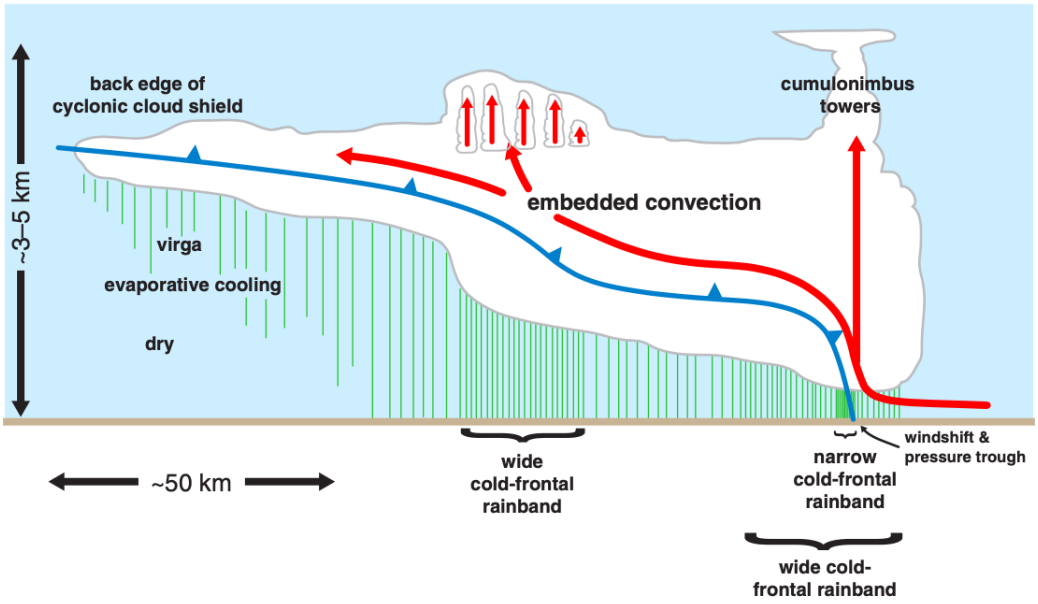


Figure P3: Schematic of the clouds, precipitation, and relative air motions associated with a cold frontal case study in the United States (motion is from left to right). The density of the green hatching is proportional to the rain rate (Markowski and Richardson, 2010).

Warm-season cold-frontal convection in Europe has been the subject of several case studies. For example, severe thunderstorms occurring in Germany during July 2013 were triggered by pre-surface-frontal convergence zones (Kunz et al., 2018). During the Convective and Orographically-induced Precipitation Study (COPS) (Wulfmeyer et al., 2011), interactions between the large-scale (e.g. synoptic-scale cold fronts) and smaller-scale influences such as orography were investigated. On the other hand, climatological studies of warm-season cold-frontal convection have largely been focused on severe convective storms. Kunz et al. (2020) found cold-frontal convective storms associated with hail produce larger hailstones and have longer tracks on average. Schemm et al. (2016) showed up to 45% of all detected hail events in north-eastern and southern Switzerland form in pre-frontal zones. Dahl and Fischer (2016) showed that around 50% of cold fronts in Western Europe are associated with pre-surface-frontal convergence lines. They also showed that the convergence lines are associated to the base of an EML. There is potential for further climatological analysis on SCSs in cold-frontal environments since previous studies have not explicitly considered specific regions relative to the front.

The term post-frontal convection has been primarily used in two ways in previous literature. First, referring to the scattered convective showers or thunderstorms which are thought to be due to the destabilisation of the air masses associated with the transport of colder air over a potentially warmer surface (e.g. Weusthoff and Hauf, 2008 and Bott, 2016a). This convection is thought to exhibit a strong diurnal cycle (e.g. Theusner and Hauf, 2004). Secondly, convection that occurs just after the surface front has passed, such as the embedded convection shown in Figure P3, may also be referred to as post-frontal convection (e.g. Schumacher et al., 2022). In this case, convection is expected to be elevated in nature since colder and drier air is present at the surface.

Primary aims of the thesis

Given the current literature and gaps in our understanding the primary aims of this thesis can be summarised in two research questions.

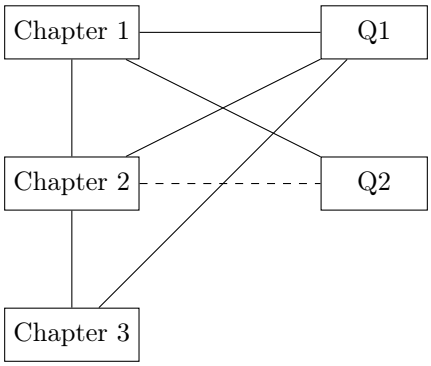
The primary research question of this thesis is:

Q1) **What are the forcing mechanisms of convective cells depending on the region relative to the front?** In other words, what processes need to come together in time and space to facilitate convective cell development depending on if convection is ahead of the front, near the front or behind the front.

The secondary research question is:

Q2) **What are the typical characteristics of convection depending on the region relative to the front and the forcing mechanisms driving those differences?** The characteristics refer to the typical organisation, intensity and risk of convective hazards.

The secondary research question will mainly be investigated in the climatological sense in Chapter 1 and partially in Chapter 2. The primary research question will be addressed using different approaches in all three chapters. A flow chart showing the links between the chapters and research questions is shown below.



Whilst these questions are the underlying focus of the thesis, several other aspects will be investigated. The motivation of each individual chapter is briefly summarised below.

Chapter 1 - The climatology and nature of warm-season convective cells in cold-frontal environments over Germany

As already discussed in the preface, some previous studies have shown cold-season convection is most common at the surface front. One should not assume this is the case for warm-season convection, especially given the possibility of pre-surface-frontal convergence lines. The term "post-frontal convection" has been

used to describe convection with distinctly different nature, both elevated convection just after the surface front has passed (see embedded convection in Figure P3) and diurnally scattered convection (Theusner and Hauf, 2004). One objective of Chapter 1 is to quantify the frequency of convection around the front as well as the diurnal cycle. Additionally, since previous studies have highlighted convection may be more likely to be severe in cold-frontal environments (e.g. Gatzen, 2011; Kunz et al., 2020), the typical nature of convection is assessed by using mesocyclone detection data, cell intensity thresholds, lightning data and hail reports. To assess whether cold-frontal convection is distinctly different from non-cold-frontal convection, a comparison to non-cold-frontal convection is made. Knowledge on cold-frontal climatology and the typical of nature of cells relative to the front could provide insightful information for forecasters and could be used as a reference for future cold-frontal convection studies in other parts of the world. The research questions of Chapter 1 are:

Q1) How do the spatial and temporal frequency of convective cells differ depending on the cell's distance from the front?
Q2) How does the nature of convective cells differ depending on the cell's distance from the front?
Q3) How do Q1 and Q2 compare to non-cold-frontal convective cells?

Chapter 2 - Thermodynamic environments and lifting mechanisms of cold-frontal convective cells during the warm-season in Germany

The environments in which convective events occur have already been studied in previous literature using variables such as CAPE, wind shear and relative humidity (Púčik et al., 2015; Taszarek et al., 2017 Taszarek et al., 2020b; Poręba et al., 2022). In Chapter 2, a similar approach is adopted with the focus again on the differences depending on the region relative to the front. As well as the thermodynamic environments, the quasi-geostrophic forcing for ascent and solar heating are considered. The primary focus is on understanding the development of convective cells and putting this in the context of Chapter 1. There will be less focus on the characteristics of convection in Chapter 2, but the environmental analysis will unravel some of the findings in the climatology from Chapter 1 regarding cell characteristics. The research questions of chapter 2 are:

Q1) How do the thermodynamic environments of convective cells vary across the front?
Q2) What is the relevance of quasi-geostrophic forcing for ascent (descent) and solar heating at different regions relative to the front?
Q3) By analysing the thermodynamic and lifting mechanisms can we explain the results of the climatology in Chapter 1?

Chapter 3 - A statistical modelling approach to better understand the forcing mechanisms of cold-frontal convective cells in Germany

Several ingredients must come together to allow the development of convective cells and links between variables are likely to be complex and non-linear. Therefore, a statistical modelling approach is used to further understand the forcing mechanisms of convective cells depending on the region relative to the front. Feature (predictor) importance techniques are utilised by training models with and without certain variables and comparing the model skill. Furthermore, a statistical model generalised on cold-frontal convective cells (at all regions relative to the front) is developed to see how well such a model could learn to reproduce

the climatology of convective cells presented in Chapter 1. The effectiveness and limitations of predictive modelling approaches are also discussed. The research questions of chapter 3 are:

Q1) Which predictors add the most skill depending on the cell-front distance?
Q2) Can statistical models be used to robustly assess feature importance and advance physical understanding?
Q3) How well can the spatial and temporal climatology of cold-frontal cells from Chapter 1 be reproduced using statistical models?

Excluding section 1.3.6, chapter 1 of this thesis is published in the *Natural Hazards and Earth System Sciences* (NHESS) journal under the Creative Commons Attribution 4.0 Licence:

Pacey, G., Pfahl, S., Schielicke, L., and Wapler, K.: The climatology and nature of warm-season convective cells in cold-frontal environments over Germany, Nat. Hazards Earth Syst. Sci., 23, 3703–3721, https://doi.org/10.5194/nhess-23-3703-2023, 2023.

I wrote all parts of the paper above, carried out the data analysis and produced all figures. The methods and conclusions in Chapter 1 slightly differ from the published article to incorporate the datasets used and the additional conclusions drawn based on section 1.3.6.

Chapter 1

The climatology and nature of warm-season convective cells in cold-frontal environments over Germany

1.1 Introduction

Atmospheric convection is a key process in the formation of certain atmospheric hazards such as flooding, wind, hail and lightning. In Europe, convective loss events to the insurance industry have risen by around a factor of 4 from about 30 events per year in 1980 to about 120 in 2014 (Hoeppe, 2016). Germany is particularly prone to such convective loss events with some events leading to insurance losses of over 1 billion euros (Kunz et al., 2018; Wilhelm et al., 2021). Taszarek et al., 2019 showed based on European Severe Weather Database reports (ESWD; Dotzek et al., 2009) that much of Germany reports 8 or more days per year with heavy convective precipitation (their Figure 7). Recent events such as the exceptional floods in Central Europe in July 2021 (Mohr et al., 2023), which was exacerbated by embedded convection, highlight the need for further information on the underlying processes that lead to heavy convective precipitation.

Deep moist convection (DMC) frequently initiates in proximity to air mass boundaries such as synoptic-scale fronts (length scale of ~1000 km), drylines and outflow boundaries (Markowski and Richardson, 2010). While cold-frontal convection is frequent during the warm-season, convection rarely initiates along the entirety of the boundary. This indicates that the precise location of convective initiation is likely due to underlying mesoscale processes (Markowski and Richardson, 2010). Such processes are not currently well-understood, thus forecasting of cold-frontal convection is challenging. Furthermore, the importance of different processes likely vary across the front, e.g, stronger forcing near the front but stronger solar heating away from the front. Previous studies including field experiments (e.g., Wulfmeyer et al., 2011; Bott, 2016a; Lee et al., 2016) have demonstrated that fronts can favour the occurrence of convection due to different processes depending on the convective initiation location relative to the front.

The region ahead of a cold front is typically referred to as pre-frontal, which typically corresponds to the warm sector of extratropical cyclones. Convection can be triggered by the convergence of warm and humid air masses near the surface. For example, near the surface front or at pre-surface-frontal convergence lines

(e.g., Wulfmeyer et al., 2011; Dahl and Fischer, 2016). The ageostrophic frontal circulation leads to lifting (Bluestein, 1993) which can help air parcels overcome convective inhibition (CIN) as well as increase convective available potential energy (CAPE) through adiabatic cooling (Trapp, 2013; his Figure 5.2). The high temperatures and moisture availability pre-frontal also increase CAPE. Convection may also be embedded in stratiform rain regions, sometimes in the form of narrow cold-frontal rainbands (Gatzen, 2011; Clark, 2013). These mostly occur in the cool-season and thus will not be explicitly addressed in this study. Convection may also initiate behind a cold front in the post-frontal environment mainly due to the destabilisation of the air masses associated with the transport of colder air over a potentially warmer surface (Weusthoff and Hauf, 2008; Bott, 2016a). Convection can occur simultaneously pre-frontal, frontal and post-frontal depending on environmental conditions (Ferretti et al., 2014).

Since cold fronts typically slope backwards with height, there is no concrete and uniform definition of post-frontal, frontal and pre-frontal convection in the literature. Some studies consider any convection after the passage of the surface cold front to be post-frontal (e.g., Schumacher et al., 2022), including convection embedded in the frontal stratiform rainband. As a surface observer this holds true, but the saturated cloud region and thus any embedded convection is likely located on the warm side of the cold front so could be considered frontal convection. Other studies consider post-frontal convection to be the diurnally driven mostly surface-based convection that is typically seen in the form of isolated showers or thunderstorms (e.g., Weusthoff and Hauf, 2008; Bott, 2016a). Where the transition from post-frontal to frontal and frontal to pre-frontal should be is also arbitrary. Therefore, we produce a cold-frontal convective cell climatology largely in a cell-front distance framework.

Previous studies on cold-frontal convection in Europe have primarily focused on narrow cold-frontal rainbands (e.g., Gatzen, 2011; Clark, 2013) or on cold-frontal hailstorms (Schemm et al., 2016; Kunz et al., 2020). Schemm et al. (2016) showed up to 45% of all detected hail events in north-eastern and southern Switzerland form in pre-frontal zones arguing that vertical wind shear and along frontal moisture transport were among the mechanisms favouring hail formation in pre-frontal environments. Kunz et al., 2020 found frontal convective storms associated with hail produce larger hailstones and have longer tracks on average and that a front was needed to trigger 50% of severe convective storms over flat-terrain in Germany. Large hail ($\geq$2 cm diameter), especially very large hail ($\geq$5 cm diameter), is often associated with supercells (e.g., Smith et al., 2012; Wapler, 2017). The defining characteristic of a supercell is a mesocyclone, that is, a deep and persistent rotating updraft. The frequency of mesocyclones and the associated hazards depending on the cell-front distance remain unexplored in any detail.

In this study we focus on warm-season cold-frontal convection (April–September) and use a broader definition of convection which is based on radar reflectivity and area thresholds thus is not solely based on convective lines (as in Gatzen, 2011 and Clark, 2013). We seek to shed light on the climatology and nature of convective cells in cold-frontal environments over Germany in a comprehensive framework, that is to say, considering the climatology and nature of cells depending on the distance from the front. By also considering lightning, mesocyclones and cell intensity in the same framework, we aim to improve understanding of convective hazard climatology in cold-frontal environments. To our knowledge, a study incorporating such aspects is not present in the current literature. We also seek to highlight the differences between cold-frontal convective cells and non-cold-frontal convective cells.

The primary research questions addressed in this chapter are as follows:

Q1) How do the spatial and temporal frequency of convective cells differ depending on the cell's distance from the front?
Q2) How does the nature of convective cells differ depending on the cell's distance from the front?
Q3) How do Q1 and Q2 compare to non-cold-frontal convective cells?

This chapter is organised as follows: Section 1.2 will introduce the convective cell detection and tracking dataset as well as the automatic front detection methods applied to ERA5 reanalysis data. We also show front detection examples for cases with convective cells in proximity. Section 1.3 will highlight the key results of the study and put them in the context of the current literature. The subsections focus on climatological cold-frontal environments, convective cell count and cell days, diurnal cycle, spatial climatology and the nature of convective cells. For the nature of cells we investigate cell lifetime, propagation speed, organisation, lightning frequency, cell intensity, and mesocyclone frequency. The nature of cells section will therefore provide novel insights into how convective hazard climatology may vary depending on the distance from the front. In Section 1.4 we summarise the results and highlight the importance of this study.

1.2 Data and Methods

Two datasets are combined: ERA5 reanalysis data (Hersbach et al., 2018a, 2020) and the KONRAD Convective Cell Detection and Tracking Dataset (Wapler and James, 2015). Automatic front detection methods are applied to the ERA5 reanalysis dataset (Section 1.2.1) and KONRAD (Section 1.2.2) is used for the convective cell definitions. At the start of the study, ERA5 was available between 1959 to present whereas KONRAD was available from 2007–2016 for the months April–September, thus the KONRAD availability period was used. The spatial and temporal resolutions of ERA5 are 1 hour and 0.25 degrees and 5 minutes and 1 km for KONRAD.

1.2.1 Front Detection

A front is generally considered a boundary between air masses with different properties such as temperature and humidity. Forecasters are primarily interested in synoptic-scale fronts as they are the main driver of precipitation in the mid-latitudes (Catto and Pfahl, 2013). Forecasters often manually position such fronts based on numerical weather prediction (NWP) output and observations. They sometimes use more technical algorithms. However, Renard and Clarke (1965) noted that different forecasters produce different analyses thus the final front location can be subjective depending on the forecaster. An additional problem is that archives of synoptic charts with fronts are sparse in both space and time. Renard and Clarke (1965) was one of the first studies that recognised the need for objective methods to detect fronts, and since then, automatic front detection in reanalysis data has attracted increasing attention (e.g., Hewson, 1998; Jenkner et al., 2010; Schemm et al., 2015; 2016, 2018; Thomas and Schultz, 2019; Rüdisühli et al., 2020). All of these studies reference the Thermal Front Parameter equation (Equation 1.1), which was first introduced by Renard and Clarke (1965).

$$\mathrm{TFP} = -\nabla|\nabla\tau| \cdot \frac{\nabla\tau}{|\nabla\tau|} \tag{1.1}$$

The term represents the rate of change of τ projected in the direction of the thermal gradient, where τ is

a thermodynamic variable (e.g., potential temperature or equivalent potential temperature). The projection takes the curvature of fronts into account. Since the term is a double derivative an inflexion point is found where the parameter is equal to zero. The maximum thermal gradient is thus where TFP=0. The horizontal wind ($\mathbf{v}$) can be projected onto the frontal line using Equation 1.2 (Hewson, 1998).

$$v_f = \mathbf{v} \cdot \frac{\nabla(\mathrm{TFP})}{|\nabla(\mathrm{TFP})|} \tag{1.2}$$

The term v_f is the horizontal wind (v) projected in the direction of the TFP gradient. It is positive at cold fronts and negative at warm fronts. The v_f threshold can be altered to consider fronts meeting a minimum advection criteria. A higher magnitude of v_f indicates stronger advection at the frontal boundary. Stationary fronts are found where $v_f \approx 0$.

Criteria

In this study, the cold-frontal regions are detected using the following criteria:

$$|\nabla\theta_e| > 6\,\mathrm{K}(100\,\mathrm{km})^{-1} \tag{A}$$

$$v_f > 1\,\mathrm{m\,s^{-1}} \tag{B}$$

$$L > 1000\,\mathrm{km} \tag{C}$$

The equivalent potential temperature is represented by θ_e and is used as the thermodynamic variable τ. The along front length is represented by L. The overlap of the θ_e gradient threshold (condition A) and velocity threshold (condition B) represents the front contour. Since we are only interested in synoptic fronts (~1000 km), fronts with an along front length (L) less than 1000 km are omitted from the analysis (condition C). The frontal line is identified at the maximum of the equivalent potential temperature gradient by applying the following condition:

$$\mathrm{TFP} = 0 \tag{D}$$

The latitude and longitude of where TFP=0 is determined using interpolation. The distance between each adjacent point is calculated and summed across the whole line to give the front length (L). The four aforementioned criteria are applied to smoothed θ_e and horizontal wind fields at 700 hPa in ERA5. The fields are smoothed 30 times using a simple smoothing function whereby the nearest four neighbours of a grid point are averaged and the process is repeated 30 times in order to remove any local-scale features. The same temperature gradient threshold and height level are used as in Schemm et al. (2016), who focused on cold fronts and hail in Switzerland. The 700 hPa pressure level is chosen to avoid interaction with orography, which is particularly an issue in central Europe (Jenkner et al., 2010, their section 4.4). Furthermore, the 700 hPa level is further from the turbulent boundary layer. We recognise the 700 hPa level is not the typical height level used by forecasters to identify fronts, but due to the complexities of automatic front detection it is important to derive smooth frontal lines for the cell-front distance calculations and limit the number of erroneously detected fronts. Equivalent potential temperature is selected over potential temperature since it is a function of both temperature and moisture thus gives the strongest frontal signal. One disadvantage to θ_e is that temperature and humidity can vary independently thus localised humidity gradients could be detected (as discussed in Thomas and Schultz, 2019). However, smaller-scale fronts like sea-breeze, land-breeze, and

gust fronts will be filtered out due to the strict 1000 km length threshold and smoothing filter. Therefore, further filtering of local and synoptic fronts, as carried out by Jenkner et al. (2010) and Rüdisühli et al. (2020) for convection-permitting datasets, is not required. The velocity threshold incorporates near-stationary to fast moving cold fronts.

The thresholds chosen (conditions A, B and C) are stricter than some previous studies. Lower thresholds would increase the number of erroneously detected fronts while higher thresholds generally reduce the number of fronts in the dataset but limit the dataset to synoptic-scale cold fronts. We value a dataset with a lower front count and higher percentage of correctly detected fronts. Cold fronts are detected in a subsection of the European domain ([40N–70N, 20W–20E], see Figure 1.1).

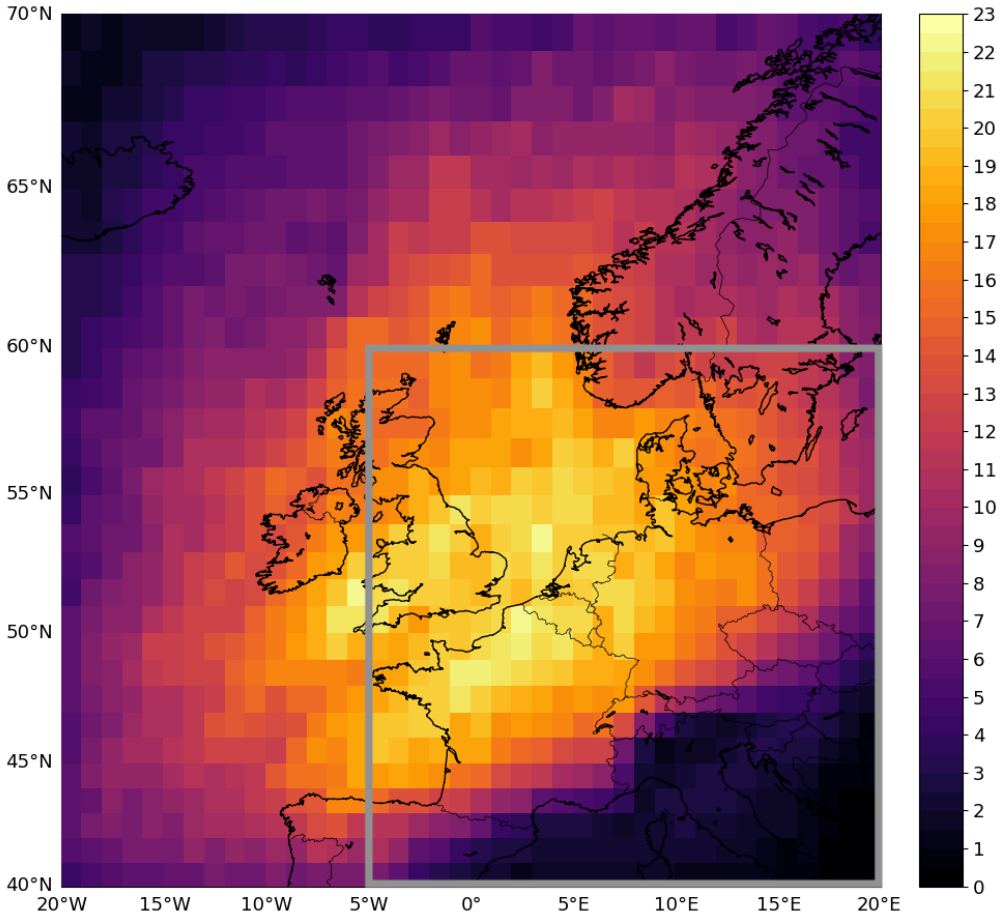


Figure 1.1: Number of cold-frontal days per warm-season between 2007–2016 for 1 x 1 degree grid boxes in the domain [40N-70N, 20W–20E]. The domain corresponds to the domain in which cold fronts were detected. If two or more fronts were detected in the grey domain ([40N-60N, 5W-20E]) at a given timestep the timestep was omitted from cell-distance calculations (see discussion is section 1.2.4).

Validation

Validation was carried out by manually viewing one year of data and several other case studies to verify that frontal features resembled synoptic-scale fronts (see examples in Figure 1.2). Figure 1.1 shows a spatial map with the number of cold-frontal days per warm season between 2007–2016 for April–September to verify plausible spatial distributions. The highest cold-frontal day frequency is over the southern half of the United Kingdom, northern France, the Low Countries and north-western Germany totalling between 20–23 front days per warm season. The minimum is found in the south-east of the domain. However, there is a sample bias for grid points at the edge of the domain since fronts extending beyond the domain may not always meet the 1000 km length criterion. Since we are focusing our analysis on Germany and the surrounding area this does not bias our results. Fronts typically decay as they experience friction from the surface over land. The cold-frontal day maximum is located close to the edge of continental Europe which could be explained by slowly propagating fronts being detected across several days. A boundary is also evident around the Alps which is likely linked to cold fronts becoming distorted and weakening when interacting with orography (Schumann, 1987). This boundary can also be identified in Kunz et al. (2020)'s cold front climatology between 2005–2014 (their Figure 3).

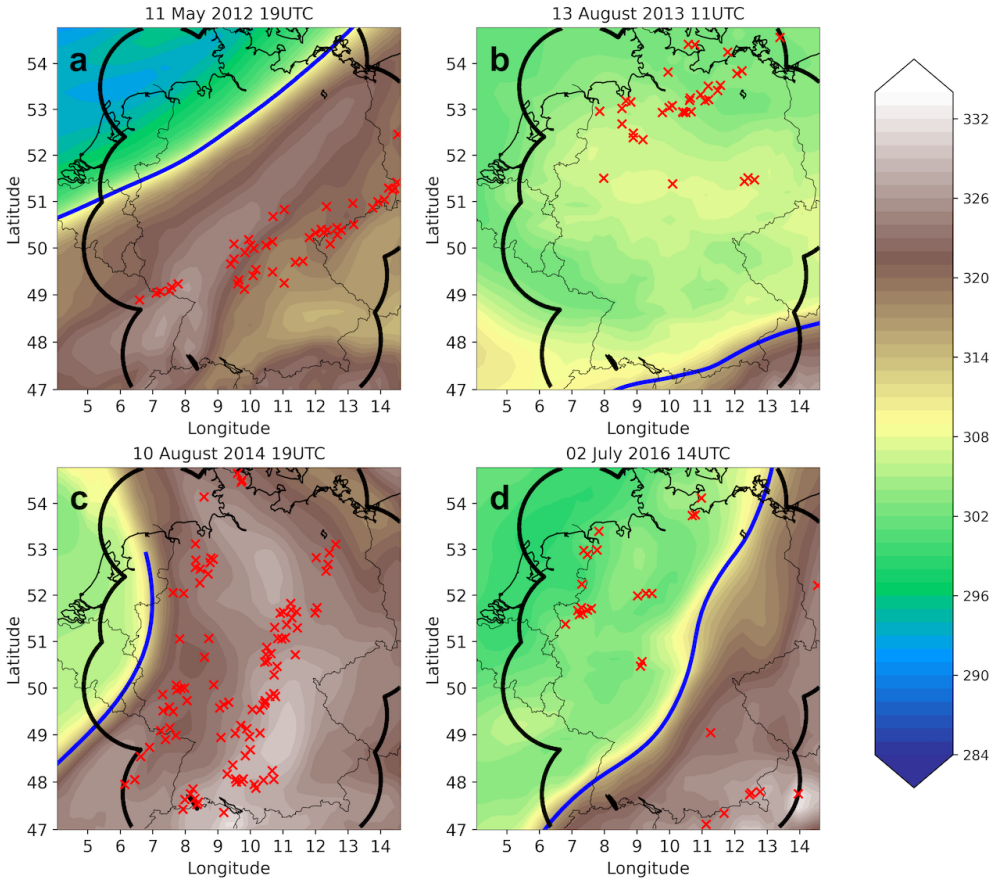


Figure 1.2: 700 hPa cold-frontal line (blue), convective cells (red crosses) and 700 hPa equivalent potential temperature in kelvin (shaded) for four cases at single timesteps. The black contour shows the KONRAD convective cell domain.

1.2.2 KONRAD Convective Cell Detection and Tracking Algorithm

KONRAD (KONvektionsentwicklung in RADarprodukten; convection evolution in radar products) is a convective cell detection and tracking algorithm (Wapler and James, 2015) originally applied to 2D radar data in the German radar domain (Figure 1.2). At the start of this study, KONRAD3D (Werner, 2017) was in the final stages of development and was not available over a long time-series so the original KONRAD based on 2D radar data was used. KONRAD is run operationally by the German Weather Service (DWD) with a spatial and time resolution of 1 km and 5 minutes, respectively. A convective cell is defined as one with 15 pixels or more exceeding 46 dBZ. Since the resolution of KONRAD is 1 km, 1 pixel ~ 1 km^2. The reflectivity is based on a 0.5 degree radar scan thus the height relative to the ground varies with distance from the radar and due to orography. Where two radar scans overlap the highest dBZ value is used. KONRAD also provides a parameter which indicates the likelihood of hail (the hail flag) ranging from 0 to 2. The parameter is based on cell intensity and area thresholds. A hail flag value of 1 is assigned if at least one pixel in the cell

exceeds 55 dBZ. A hail flag value of 2 is assigned if 12 or more pixels exceed 55 dBZ or if at least one pixel exceeds 60 dBZ. Otherwise a hail flag of 0 is assigned. Additional information in relation to the detected convective cells has been made available such as the number of lightning strikes in proximity to the cell centre and mesocyclone characteristics. The lightning strike data are based on the time of arrival method from the European VLF/LF LIghtning detection NETwork (LINET, Betz et al., 2009). A mesocyclone detection algorithm is run operationally by the DWD using doppler radar data (Hengstebeck et al., 2018). The severity of mesocyclones is given a ranking from 1 to 5. We include mesocyclones of ranking 3 or higher in order to include a sample with significant size and mesocyclone strength. To be classified as a mesocyclone with severity level 3 the following criteria must be met: 5 km equivalent diameter, 3 km depth, 10^{-3} s^{-1} azimuthal shear or 15 ms^{-1} rotational velocity. A detailed description of the algorithm and practical examples can be found in Hengstebeck et al. (2018). The convective cell, hail flag, lightning and mesocyclone definitions are summarised in Table 1.1. The number of pixels exceeding 55 dBZ, lightning and mesocyclone definitions will be utilised in section 1.3.5. However, for all other sections the convective cell definition is based on 15 pixels or more exceeding 46 dBZ. This definition includes cells of moderate to severe intensity. For mid-latitude convective rainfall the Z-R relationship $Z = 300R^{1.4}$ is commonly used (Nelson et al., 2010). A reflectivity of 46 dBZ therefore corresponds to an estimated rain rate of 33 mm hr^{-1}. Excluding section 1.3.5, we consider cells at their first detection time only to avoid repeat counting at additional timesteps in the climatology.

Table 1.1: Convective cell, hail flag 1, hail flag 2, lightning and mesocyclone definitions. Since the KONRAD dataset has a spatial resolution of 1 km, 1 pixel ~ 1 km^2.

Name	Definitions and thresholds
Convective Cell	15 pixels exceeding 46 dBZ
Hail Flag 1	1 pixel exceeding 55 dBZ
Hail Flag 2	12 pixels exceeding 55 dBZ or 1 pixel exceeding 60 dBZ
Lightning	3 or more strikes 20 km or less from the cell centre during the cell's lifetime
Mesocyclone	5 km equival ent diameter, 3 km depth, 10^{-3} s^{-1} azimuthal shear or 15 ms^{-1} rotational velocity

The KONRAD dataset has the following limitations:

- The lifetime of a cell in KONRAD does not reflect the entire lifetime of the cell, rather the time the cell exceeds the reflectivity and area thresholds. The first and last detection time refer to when the cell first and last exceeds the reflectivity and area thresholds.
- Cell splitting and merging can lead to an unrealistic lifetime in some cases (see section 2.1 of Wapler, 2021).
- During a mesoscale convective system (MCS) new cells may be detected at subsequent timesteps and assigned the first detection label due to cell recycling (see Figure A1).
- Convective lines are not trivial to identify since a large cell area could indicate a large circular cell or a long and narrow convective line.

1.2.3 Lightning strike and hail reports

Lightning strike data and hail reports are used in section 1.3.6. Lightning data were provided by the Met Office, who use an arrival time difference network (ATDnet) to detect lightning strokes (Met Office, 2020).

The data availability for lightning strike is 2008-2016 (April-September). Hail reports were provided by the ESWD (Dotzek et al., 2009) and reports with quality control level QC0+ (confirmed plausibility) or better are used. The definition for a hail report in the ESWD is hail with 2 cm or larger diameter or hail accumulations of 2 cm. The same data period as KONRAD convective cells is used for hail reports (2007–2016, April–September).

1.2.4 Cell and front detection examples

Combining the methods introduced in Section 1.2.1 and 1.2.2, four cases are shown in Figure 1.2 for single timesteps. The cold-frontal line (blue contour) is the result of applying the conditions (A, B and C). Convective cells are considered to be associated with a front in the hour following the front detection. For example, for a front detection at 13 UTC, cells that were first detected between 13:00–13:59 UTC are included. The distance between the convective cell and the 700 hPa frontal line (hereafter cell-front distance) is the distance between the convective cell centre and the nearest grid point on the 700 hPa frontal line contour (i.e., the shortest distance between the frontal line and the convective cell on the surface of an ellipsoidal model of the earth; Karney, 2013). A cell is assigned a positive distance if the average θ_e surrounding the cell location (4 grid points) is equal to or greater than the frontal line θ_e, thus indicating the cell is on the warm side of the 700 hPa front. Otherwise, the cell is assigned a negative distance indicating the cell is on the cold side of the 700 hPa front. Timesteps containing two or more cold-frontal lines in the domain [40N–60N, 5W–20E] were omitted from cell-front distance calculations due to uncertainty deriving the cell-front distance. This is a subdomain of the domain in which fronts were detected (see grey domain in Figure 1.1). This resulted in 3,506 (18.8%) of the total 18,613 frontal timesteps being omitted. We thus focus on cases where a single large-scale cold front (condition C) is present in the vicinity of Germany that meets the gradient and velocity thresholds (conditions A and B). Omitting timesteps with two or more fronts meeting the criteria does not omit cases with a primary large-scale cold front and a secondary smaller scale front, weaker front or pre-surface-frontal convergence lines. We also note that we do not explicitly consider ana and kata cold fronts separately. During a kata cold front an upper-level humidity front is observed overrunning the surface cold front (Browning and Monk, 1982).

In the examples (Figure 1.2), situations are shown with cells on the warm side of the 700 hPa front (hereafter pre-700hPa-frontal), in close proximity to the 700 hPa front (near-700hPa-frontal) and also behind the 700 hPa front (post-700hPa-frontal). In some cases they occur simultaneously in the KONRAD domain (black contour). A NE-SW oriented front is located over Germany in Figure 1.2a with pre-700hPa-frontal convective cells. A line of cells can be found between 250–400 km ahead of the 700 hPa front. The second case (Figure 1.2b) shows a cold front that has progressed across Germany and is located in proximity to the Alps. Cells are located post-700hPa-frontal 350–750 km behind the 700 hPa front. Like Figure 1.2a, Figure 1.2c also shows a pre-700hPa-frontal case but with a front oriented more towards the N-S direction and with more curvature. A larger number of cells are associated with this case with cells in a larger range of distances between 100–750 km ahead of the 700 hPa front. The final example (Figure 1.2d) shows a front oriented NE-SW through the middle of the radar domain with both pre-700hPa-frontal and post-700hPa-frontal cells.

1.3 Results

1.3.1 Climatological cold-frontal environments

To understand the importance of different processes and the typical environment across the front, we plot climatological means of four possible drivers of convection depending on the distance from the front: convergence of the wind field at different pressure levels, CAPE, surface dewpoints and the total sky direct solar radiation at the surface (hereafter solar heating; Figure 1.3). The means are calculated using all instances that an ERA5 grid point has a certain front distance and is not weighted by individual timesteps. The mean of the convergence is shown by the colorbar, whereas the means of CAPE, solar heating and surface dewpoints are normalised between one and zero and shown as black lines. The variable means are normalised to compare the typical magnitude of each variable at different distances from the front. All variables are obtained from ERA5 reanalysis data; the same dataset in which fronts were detected. The means were calculated in the domain of Germany only.

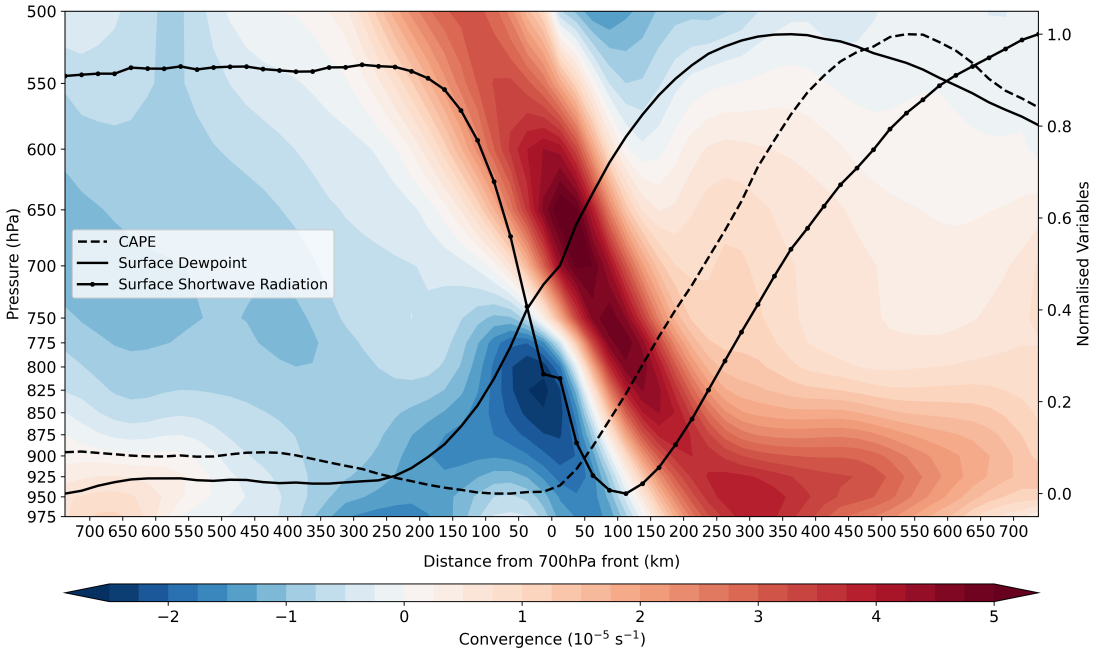


Figure 1.3: Climatological convergence at pressure levels (shaded), CAPE (dashed line), surface dewpoints (solid line), surface shortwave radiation (solid line with circular markers) in ERA5 depending on the distance from the 700 hPa front. Excluding convergence, the climatological means are normalised between 0 and 1. The CAPE in ERA5 is calculated based on the most unstable parcel below 350 hPa.

Convergence of the wind field

Convergence and divergence of the wind field is shown by the positive values (red) and negative values (blue), respectively (Figure 1.3). The strongest climatological near-surface convergence is around 300 km ahead of the 700 hPa front. The result is consistent with understanding of typical cold frontal slopes since the 700 hPa level is typically found ~3 km above the surface and cold fronts have a typical slope of ~1:100 (Bott, 2016b).

This indicates that the climatological location of the surface convergence zone, which is linked to the surface front, is 300 km ahead of the 700 hPa front. The surface convergence decreases ahead of the maximum but convergence is present up to 750 km ahead of the 700 hPa front. Near-surface convergence of the wind field is generally linked to lift; a prerequisite for convective initiation (Doswell et al., 1996). However, 200 km behind the maximum (100 km ahead of the 700 hPa front) there is divergence near the surface (blue shading in Figure 1.3). The strongest climatological divergence is found around 850 hPa under the 700 hPa front location. The implication is that if the cold front location is observed at the 700 hPa level on a synoptic chart the strongest divergence of the wind field can be expected at this location in space. Near-surface divergence of the wind field is generally linked to descending motion, which could inhibit convective initiation.

Convective Available Potential Energy

CAPE is a measure of atmospheric instability; a prerequisite for deep moist convection (Doswell et al., 1996). It is also a very commonly used parameter in severe convective storm forecasting. CAPE is climatologically maximum 550–600 km ahead of the 700 hPa front (shown by the dashed line in Figure 1.3). The highest CAPE available in the environment is therefore typically ahead of the maximum near-surface convergence. CAPE is minimum at the 700 hPa frontal line and slightly increases further behind the 700 hPa front.

2-metre dewpoint temperature

Surface dewpoints are a measure of moisture availability at the surface. Figure 1.3 shows that surface dewpoints are maximum 350–400 km ahead of the 700 hPa front, which is marginally ahead of the maximum climatological near-surface convergence. Further ahead of the 700 hPa front, surface dewpoints decrease but not as quickly as the decrease towards the 700 hPa front. The surface dewpoints reach a minimum after the passage of the 700 hPa front.

Total sky direct solar radiation at the surface

Solar heating is linked to increased surface temperatures, which contributes towards atmospheric instability and lifting of air parcels near the surface. Solar heating is minimum 100–150 km ahead of the 700 hPa front (solid line with circular markers in Figure 1.3), which indicates increased cloudiness. The maximum in large-scale precipitation in ERA5 was also found at this distance relative to the front (not shown). Further ahead of the 700 hPa front solar radiation gradually increases reaching a maximum 700–750 ahead of the 700 hPa front. The increase in the other direction towards the post-700hPa-frontal environment is sharper, which can be explained by cloudiness of the warm-sector whereas the post-700hPa-frontal environment remains largely cloud-free.

Surface front relative to the 700 hPa front

Since near-surface convergence is relevant for lifting parcels to their level of free convection (one of the ingredients for deep moist convection; Doswell et al., 1996), the strongest climatological near-surface convergence (Figure 1.3) is used as a reference point for the surface front location relative to the 700 hPa front. Furthermore, the surface front would be expected to be 300 km ahead of the 700 hPa front since the 700 hPa level is typically found ~3 km above the surface and cold fronts have a typical slope of ~1:100. While the slope of the front and corresponding surface front location relative to the 700 hPa front are likely to vary slightly

per case study, we proceed assuming the climatological location of the surface front (hereafter surface front) is 300 km ahead of the 700 hPa front for this study.

1.3.2 Convective cell count and days

During the period 2007–2016 for the months April–September, 258,218 cells were detected in KONRAD (Section 1.2.2). On cold-frontal cell days, 163,255 cells (63%) were detected (see Table 1.2). A cold-frontal cell day is defined as one where at least one convective cell was detected within 750 km of the 700 hPa frontal line. The remaining 94,963 (37%) cells were detected on non-cold-frontal days. A cell was associated with a cold front on 614 days, thus approximately a third of warm-season days are associated with cold-frontal convection in Germany. This does not imply a new frontal system every 3 days as many cases are associated with slowly propagating fronts across several days. While a slightly larger percentage of all cell days were non-cold-frontal cell days (56%), over twice as many cells develop on average on cold-frontal cell days (266 cells per day) compared to non-cold-frontal cell days (123 cells per day) (Figure 1.4). Considering all 1389 days convective cells were detected in KONRAD, the 95th percentile of the cells per day is 725 accounting for 80 days; 80% of these days occurred on cold-frontal cell days. This highlights that extreme events (in terms of cell count per day) are significantly more likely on cold-frontal cell days. The differences in the cell per day count could reflect convection being more widespread across Germany and/or more cells associated with each cell cluster. During an MCS for example a larger amount of cells are typically detected in KONRAD as a continuous line exceeding 46 dBZ may not always be present so several cells are detected within the MCS. Secondly, due to cell recycling in an MCS, KONRAD will likely detect new cells at subsequent timesteps and assign them the first detection label. Mesoscale convective systems could therefore also explain the large differences in the cell count per day between cold-frontal cell days and non-cold-frontal cell days. These results are consistent with findings from Wapler and James (2015) who found the synoptic patterns with the highest number of convective cells were Cyclonic Southerly, Anticyclonic South-Easterly, Cyclonic South-Westerly and Trough over Western Europe. Such synoptic patterns are likely to indicate the presence of a cold front excluding Anticyclonic South-Easterly. The cell statistics are summarised in Table 1.2.

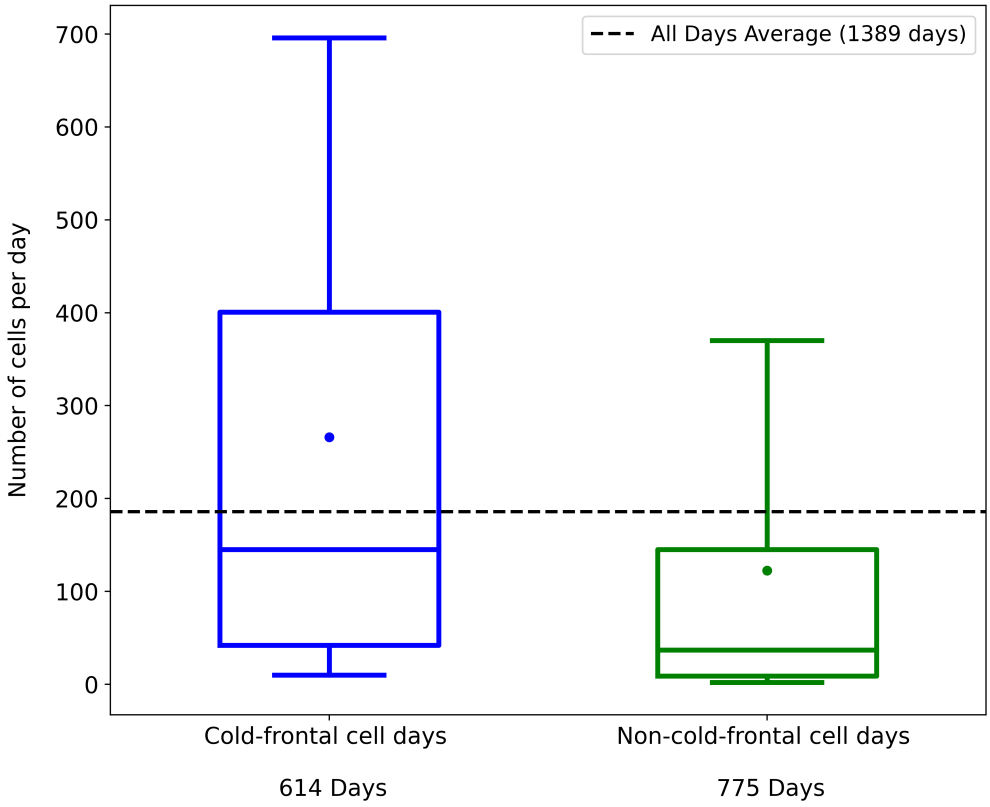


Figure 1.4: Number of cells per day box plots for cold-frontal cell days and non-cold-frontal cell days. A cold-frontal cell day is defined as one where at least one convective cell was detected within 750 km of the 700 hPa frontal line, otherwise the day is classified as a non-cold-frontal cell day. The median is represented as a horizontal line, dots represent the mean, boxes represent the 25th–75th percentile values, and whiskers represent the 10th and 90th percentile values. The all days mean (1389 days) is shown by the dashed black line.

Table 1.2: Summary of convective cell datasets. A cold-frontal cell day is defined as one where at least one convective cell was detected within 750 km of the 700 hPa cold-frontal line, otherwise the day is classified as a non-cold-frontal cell day.

Name	Cold-frontal cell days	Non-cold-frontal cell days	All cell days
Cell Count	163,255	94,963	258,218
Day Count	614	775	1,389
Cells per day	265.9	122.5	185.9

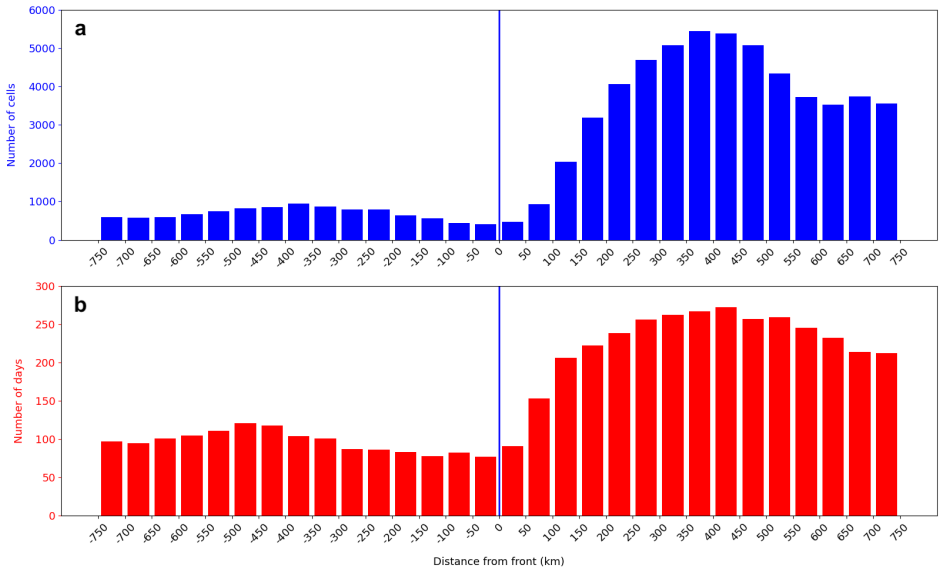


Figure 1.5: Number of convective cells (a) and number of convective cell days (b) depending on the cell-front distance. Positive and negative distances represent the pre-700hPa-frontal and post-700hPa-frontal environments, respectively.

Figure 1.5a shows the number of convective cells (blue) depending on the cell-front distance. Positive distances indicate pre-700hPa-frontal cells and negative distances post-700hPa-frontal cells. A total of 65,567 cells were found in proximity (±750 km) to the 700 hPa frontal line between 2007–2016 (April–September). The 65,567 cells are a subset of the 163,255 cells on cold-frontal cell days, the remaining cells occurred outside of the 750 km range, on timesteps with two or more fronts or on timesteps with no cold front detected. The most common region for cells is pre-700hPa-frontal accounting for 84% of all cells. The highest cell frequency is 350–400 km ahead of the 700 hPa front. This corresponds to the location marginally ahead of the mean surface front location (section 1.3.1) and corresponds to the where surface dewpoints are climatologically highest (Figure 1.3). Further ahead of the front the cell frequency decreases but remains significantly higher than post-700hPa-frontal cells. Furthermore, a slight increase in the cell frequency is observed around 650–700 km ahead of the 700 hPa front which could be linked to pre-surface-frontal convergence lines (Dahl and Fischer, 2016). Behind the surface front the cell frequency steeply declines towards the 700 hPa front. The 700 hPa frontal line corresponds to the minimum cell frequency. This also justifies our use of the 700 hPa level since it marks a clear shift in the probability of convective cells. This minimum could be linked to the descent observed behind a cold front (Davies and Wernli, 2015), as also shown by the maximum climatological divergence in the lower levels (Figure 1.3). Furthermore, the lowest climatological CAPE is at the 700 hPa front location. Synoptic-scale fronts are typically analysed by forecasters at the surface or 850 hPa, thus forecasters should not expect the lowest convective activity to be after the front has passed at 850 hPa or the surface, rather after the frontal passage at 700 hPa. Post-700hPa-frontal cells account for 16% of all cells. Similar to the pre-700hPa-frontal environment, there is a relatively normalised distribution around the maximum frequency 350–400 km behind the front. Assuming post-700hPa-frontal convection is primarily driven by solar heating, a certain period is likely required after the front passes before convective initiation is favourable. Figure 1.3 highlights that solar heating increases sharply during and after the passage of the

700 hPa front and CAPE also increases slightly.

The number of convective cell days is shown in Figure 1.5b. The number of convective cell days refers to the number of days a cell was found between each 50 km interval shown in Figure 1.5. A similar distribution around the front is observed as in Figure 1.5a but with a slight shift in the maximum further from the 700 hPa front. This is true both post-700hPa-frontal and pre-700hPa-frontal with the new maximum found -500 to -450 km and 400 to 450 km from the 700 hPa front, respectively. While approximately 5 times as many cells were found around the pre-700hPa-frontal maximum (~5000 cells) compared to the post-700hPa-frontal maximum (~1000 cells), the number of cell days differs by a factor of 2.5. Therefore, pre-700hPa-frontal cases are associated with a greater number of cells than post-700hPa-frontal cases which explains the significant difference in the magnitude of the peak when looking at the total number of cells.
Since cold fronts typically weaken as they move over continental Europe there is possibility of a sample bias with more pre-700hPa-frontal environments being sampled than post-700hPa-frontal environments. This bias was assessed by shuffling each cold-frontal timestep in the analysis period to a random timestep and then deriving the cell-front distances again. This process was repeated 100 times and a new cell count around the front was produced using the random sample (see Figure A2). The sample bias can be seen for pre-700hPa-frontal environments, however when dividing the actual cell count by the random sample cell count we see no significant differences in the distribution around the 700 hPa front compared to Figure 1.5a. Figure A2 also highlights that between 100–1000 km ahead of the 700 hPa front the cell count is higher than the random sample cell count.

1.3.3 Diurnal Cycle

Convection is known to generally exhibit a diurnal cycle with a maximum in the afternoon hours and a minimum in the early morning hours (Ban et al., 2014). Figure 1.6 shows the cell frequency as a function of the cell-front distance and hour of the day. As noted in Figure 1.5, the maximum cell count is found pre-700hPa-frontal with the minimum at the 700 hPa front. Figure 1.6 highlights that most post-700hPa-frontal cells develop during the daytime hours (especially between 10–18 UTC) with low cell frequency outside of this period. In contrast, pre-700hPa-frontal cells do develop in the night hours and the cell maximum is shifted later in the day. The exception to this is further from the front in the region 750–1000 km where cells have a stronger diurnal cycle. We speculate that the lifting associated with the front is weaker this far from the front thus cells are largely diurnally driven. This result also justifies our distance criterion at which we consider a convective cell to be associated with a front, that is, within 750 km of the 700 hPa frontal line.

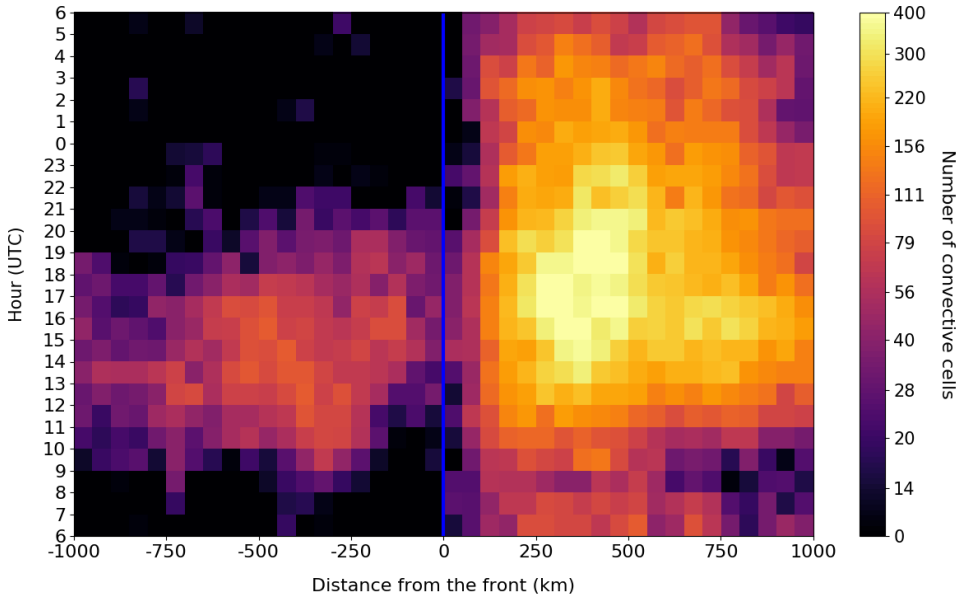


Figure 1.6: Number of convective cells (non-linear colorbar) depending on distance from the 700 hPa front (horizontal axis) and hour of the day (vertical axis). Negative distances indicate post-700hPa-frontal cells while positive distances indicate pre-700hPa-frontal cells. Note the non-linear colorbar.

Figure 1.7 shows the same data as Figure 1.6 but with an average taken across the pre-700hPa-frontal (0 to 750 km) and post-700hPa-frontal (-750 to 0 km) environments. The diurnal cycle of non-cold-frontal cells (green) is shown for reference. A strong diurnal cycle for post-700hPa-frontal cells (blue line) is evident with a maximum between 15–16 UTC. The distribution for non-cold-frontal cells is similar with a maximum also between 15–16 UTC. The pre-700hPa-frontal maximum is found between 16–17 UTC, a later peak than in the non-cold-frontal and post-700hPa-frontal categories. The weakened cycle is evident by the larger percentage of cells occurring during the night-time hours. Between 21–06 UTC, 31% of pre-700hPa-frontal cells develop compared to only 9% and 13% of post-700hPa-frontal and non-cold-frontal cells, respectively. A similar weakening of the diurnal cycle was observed for mesocyclones compared to lightning in Germany (Wapler et al., 2016). This indicates that supercells, where a deep and persistent mesocyclone is present, may also be linked to the observed weakened diurnal cycle. Mesocyclone frequency depending on the cell-front distance will be investigated in section 1.3.5. Furthermore, Morel and Senesi (2002) found that MCSs dissipate most commonly at 21 Local Solar Time (LST) on average in Europe. Surowiecki and Taszarek (2020) found a similar result showing the majority of MCSs dissipate around 19–20 UTC in Poland, though squall line/bow echo MCSs most frequently dissipate at midnight. MCSs typically live longer than individual cells, and new cells may be triggered throughout their lifecycle, e.g., through interaction with the cold pool which is less dependant on solar heating. We speculate therefore that the observed pre-700hPa-frontal weakened diurnal cycle could be linked to mesoscale convective systems. We also investigated the monthly cycle of convective cells depending on the cell-front distance but no significant differences were found with the peak season being mid-July for cold-frontal cells. The peak season for non-cold-frontal cells was early July.

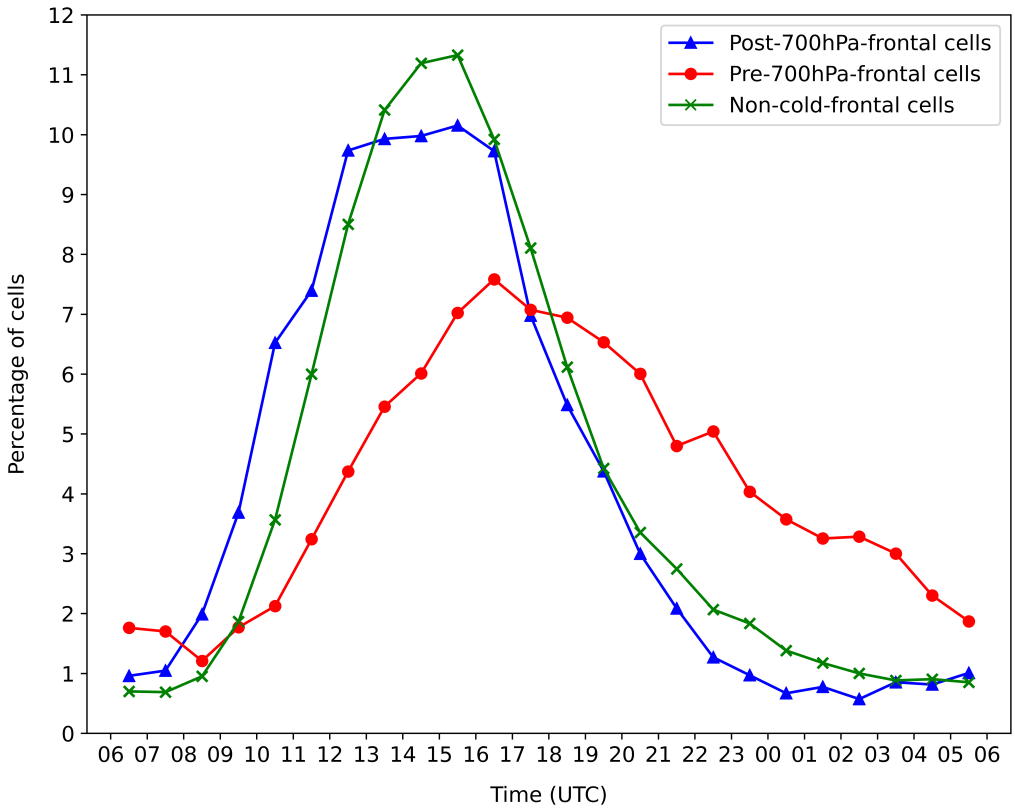


Figure 1.7: Diurnal cycle of convective cells for post-700hPa-frontal (blue), pre-700hPa-frontal (red) and non-cold-frontal cells (green).

1.3.4 Spatial cell climatology

The frequency of convective events varies in different parts of Europe (Groenemeijer et al., 2017; Taszarek et al., 2019 and 2020a). Schemm et al. (2016) also showed the fraction of hail events associated with a synoptic-scale cold front varied spatially across Switzerland and the vicinity. In this section we produce spatial maps of convective cell days per warm season in the KONRAD radar domain (grey contour in Figure 1.8). As in Figure 1.4, a cell day is defined as one where at least one convective cell was detected, in this case within a 1 x 1 degree grid box.

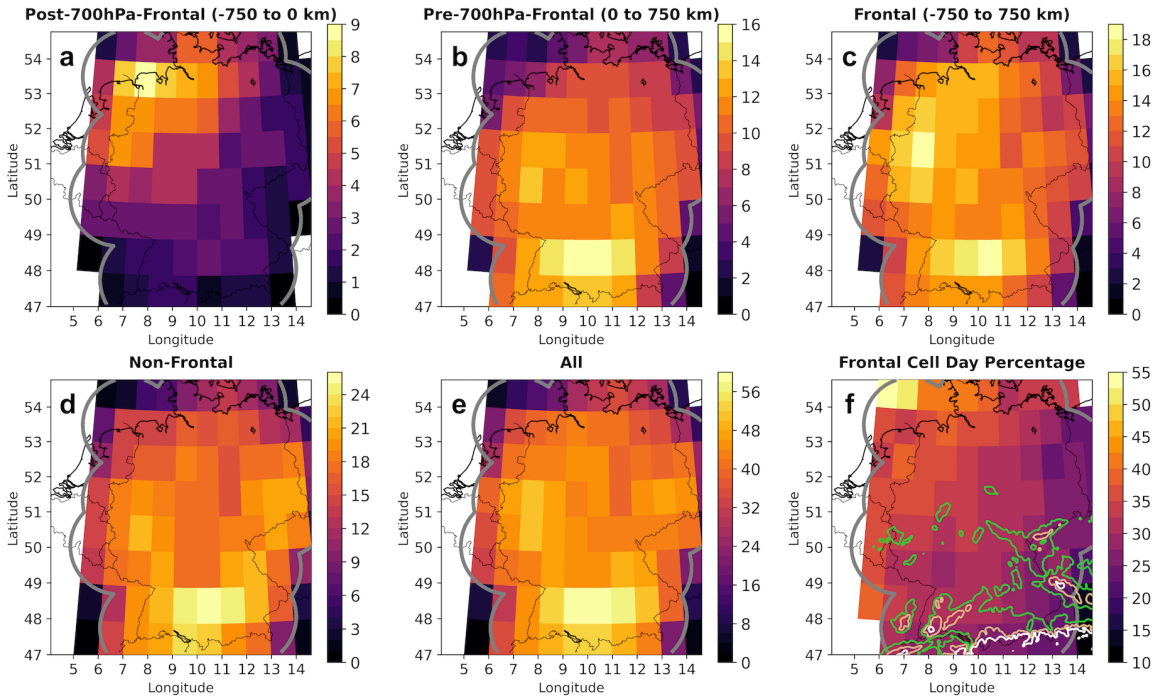


Figure 1.8: Number of convective cell days per warm season for 1 x 1 degree grid boxes for post-700hPa-frontal cells (a), pre-700hPa-frontal cells (b), cold-frontal cells (c), non-cold-frontal cells (d), all cells (e). Panel f shows the percentage of convective cell days associated with cold fronts (panel c divided by panel e) and the green, brown and white contours represent 500, 1000 and 1500 metres elevation, respectively. Note the different colorbar levels for each subplot.

Convective cell days

Distinct spatial distributions were found for pre-700hPa-frontal and post-700hPa-frontal cells (Figure 1.8). Post-700hPa-frontal cells (Figure 1.8a) are most frequent in north-west Germany developing between 5–9 days per warm season. The highest frequency is near the coastline, which could be linked to sea-breezes, higher moisture availability and increased lift with cooler air being advected over a warmer surface. Moving south and east the number of cells per warm season reduces to 1–3. Pre-700hPa-frontal cells (Figure 1.8b) are most frequent in the southern half of Germany, particularly in the far south near the borders with France, Switzerland and Austria. Cells develop between 8–16 days per warm season in the majority of the domain in the pre-700hPa-frontal environment. The exceptions are grid boxes outside of Germany which are likely under-sampled as they are located further from radars. The combined pre-700hPa-frontal and post-700hPa-frontal convective cell spatial distribution shows one maximum in southern Germany and another in north-west Germany (Figure 1.8c). North-east Germany receives the lowest number of cold-frontal cell days per warm season. The non-cold-frontal (Figure 1.8d) and all cells spatial distributions (Figure 1.8e) are very similar to pre-700hPa-frontal cells (Figure 1.8b) with a maximum cell frequency in southern Germany. The high cell frequency in southern Germany is largely linked to the orography in this region which can modify the mesoscale circulation and environment (see elevation contours on Figure 1.8f). Several studies have shown that thunderstorms are more frequent near mountainous regions (e.g., Piper and Kunz, 2017; Taszarek et al., 2019).

Cold-frontal cell day fraction

Figure 1.8f shows the percentage of all convective cell days associated with cold fronts for 1 x 1 degree grid boxes. Around 35–40% of convective cell days were associated with cold fronts for much of north-western Germany. This could be linked to the high frequency of cold fronts in this region (Figure 1.1). The percentage decreases moving further south and east. In south-eastern and eastern Germany, western Czechia and western Poland only 20–30% of cell days were cold-frontal. These regions are mostly in proximity to elevated terrain (see contours on Figure 1.8f). Kunz et al. (2020) found a similar result showing that over complex terrains, the proportion of frontal severe convective storms is around 10–15%, while over flat terrain 50% require a frontal trigger.

Cold front types

The orientation and geographical location of cold fronts could potentially be used as an indicator for the likelihood of cells. K-means clustering, an unsupervised machine learning algorithm (section 14.3.1 in Wilks, 2006), is used to cluster frontal timesteps into clusters such that each cluster contains fronts with similar characteristics (i.e., spatially located and orientated in a similar way). A cropped domain ([40N–60N,0–20E]) of the original front detection domain (Figure 1.1) was used. A total of 4,230 frontal timesteps with convective cells were available to be clustered. Since the input into the algorithm is binary in our case (i.e., 1 for a cold-frontal grid point and 0 for a non-cold-frontal grid point) no normalisation is required.

K-means clustering requires the cluster number to be predefined. The elbow method and silhouette score were first tested between cluster numbers 2 and 50 (see Figure A3) to determine the optimal cluster number. However, no clear elbow could be identified, which is often the case for many real world problems. The highest silhouette score was for 2 clusters which consisted of a cluster with fronts close to the Alps and another cluster with fronts to the north-west of Germany. However, high variance was found within both clusters and no clear orientation of the front could be identified indicating several different fronts were contained in each cluster. Since the elbow method and silhouette score proved ineffective, plots of cluster numbers 10 through 50 were produced to find the optimal cluster number in terms of not too many clusters resulting in large similarities between clusters (and the front type) but not too little resulting in fronts with different characteristics being grouped in the same cluster. Thirty was selected subjectively as the number of clusters based on viewing the aforementioned plots. Six clusters (1,133 timesteps) were removed due to a lack of continuity in the clusters. We draw frontal contours (Figure 1.9) where 50% of timesteps in the respective cluster had the front located at that grid point. Plots for cluster numbers 15, 35 and 50 are shown in the supplement (Figures A4, A5 and A6).

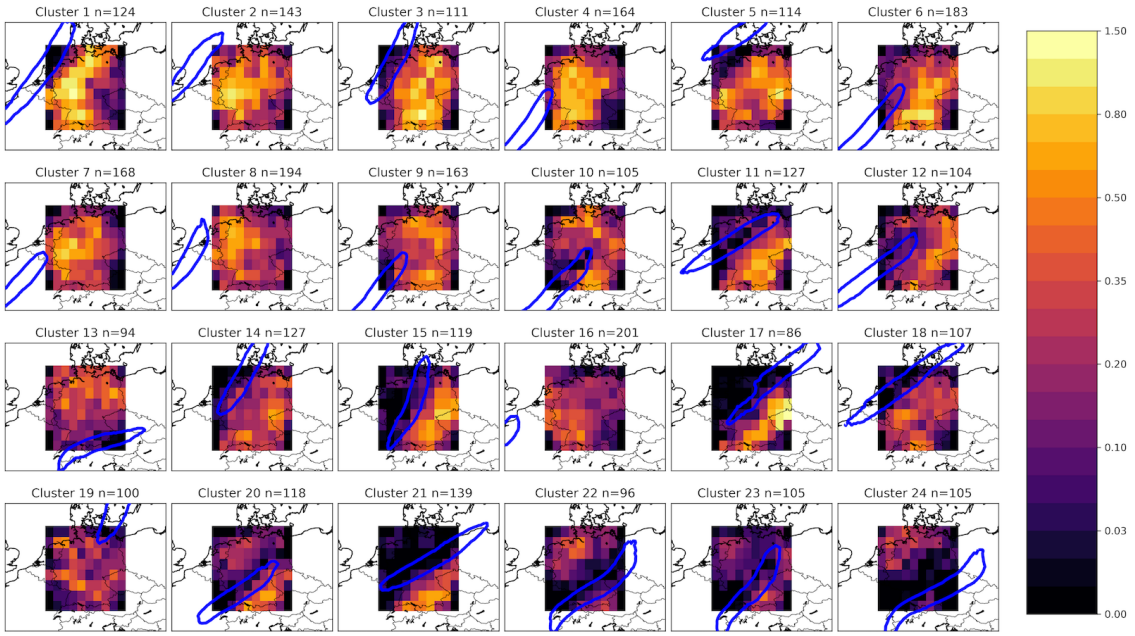


Figure 1.9: Clustered fronts using the k-means clustering algorithm in the domain [40N–60N,0–20E]. Subplot header represents the number of timesteps (n) contained in each cluster. The colorbar represents the number of convective cells per 1 x 1 degree grid box per timestep. The blue contours indicate where 50% of timesteps in the respective cluster have the front located at that grid point.

The resulting 24 front clusters are shown in Figure 1.9 sorted by the overall largest number of convective cells per timestep. The colorbar represents the number of cells per timestep per 1 x 1 degree grid box. Cluster 1 has the highest cell frequency, which is associated with a NE-SW oriented front extending from Scandinavia to France. The highest cell frequency is in the western half of Germany ahead of the 700 hPa front, consistent with our results shown in Figure 1.5. The convective cell minimum around the 700 hPa frontal location can also be located in several clusters (e.g., Clusters 11, 15 and 20). Cluster 2 and 3 show similar front types to Cluster 1 but are positioned slightly further west (Cluster 2) or east (Cluster 3). Post-700hPa-frontal cells are mostly attributed to clusters 13, 20, 22, 23 and 24 when the front has progressed further inland across Germany. A period of solar heating is likely required after the front passes before convective initiation is favourable. Clusters 22, 23 and 24 also highlight how cold fronts become deformed on their interaction with the Alps (Schumann, 1987). Cluster 17 is worthy of mentioning due to the high cell frequency per timestep near the German-Czech border. Piper and Kunz (2017) found increased thunderstorm activity in this region based on lightning data between 2001–2014. These plots provide useful insight for forecasters highlighting where in respect to the front cells are most likely for certain front types and their corresponding geographical locations.

1.3.5 The nature of convective cells

As mentioned in Section 1.2.2, KONRAD defines a convective cell based on reflectivity and area thresholds. Namely, a convective cell is defined as a cell with 15 pixels (~15 km^2) or more exceeding 46 dBZ (see Table 1.1). Furthermore, a hail flag value of 1 is assigned if at least one pixel in the cell exceeds 55 dBZ. A hail

flag value of 2 is assigned if 12 or more pixels exceed 55 dBZ or if at least one pixel exceeds 60 dBZ. Using the same definitions of cold-frontal and non-cold-frontal cell days as in Figure 1.4, Figure 1.10 highlights that mostly all cell days have at least one cell with the hail flag value 1; 96% of cold-frontal cell days and 87% of non-cold-frontal cell days. However, with increasing number of pixels a discrepancy is found. For example, 69% of cold-frontal cell days have a cell with 12 pixels exceeding 55 dBZ (similar criterion for hail flag 2), opposed to only 46% of non-cold-frontal cell days. The discrepancy becomes larger for increasing number of pixels. Around 30% of cold-frontal cell days have a cell with 50 pixels exceeding 55 dBZ, opposed to only 10% of non-cold-frontal days. These results highlight that given cells initiate in proximity to a cold front, cells with larger intense convective cores are more likely to develop. These larger cores could be linked to an increased number of strong convective lines present on cold-frontal cell days. The 55 dBZ threshold has also been proposed as a reflectivity threshold for hail (Mason, 1971). Several studies have evaluated the threshold (e.g., Hohl et al., 2002; Wapler et al., 2012; Kunz and Kugel, 2015). More recently, Wapler (2017) validated the hail flag against ESWD reports showing that 72% of 2 cm or larger hail reports and 84% of 5 cm or larger hail reports were labelled hail flag 2. Wapler (2017) also found that the synoptic pattern most commonly associated with hail in Germany is a cyclonic south-westerly flow. Such a synoptic pattern is often associated with a cold front and sometimes the advection of an elevated mixed layer originating from North Africa or the Spanish plateau (Lewis and Gray, 2010; Dahl and Fischer, 2016). The steep lapse rates in the midlevels can contribute to atmospheric instability which is one factor relevant for large hail (Púčik et al., 2015; Taszarek et al., 2020b). Schemm et al. (2016) found up to 45% of hail events in north-eastern and southern Switzerland formed in pre-frontal zones. They argued that vertical wind shear and along frontal moisture transport were among the mechanisms favouring hail formation in pre-frontal environments. Additionally, Kunz et al. (2020) showed frontal hailstorms produce larger hail and have a longer track on average. Future studies should utilise hail reports such as those collected in the ESWD and automatic front detection on a pan-European scale to better understand the link between cold fronts and hail in Europe.

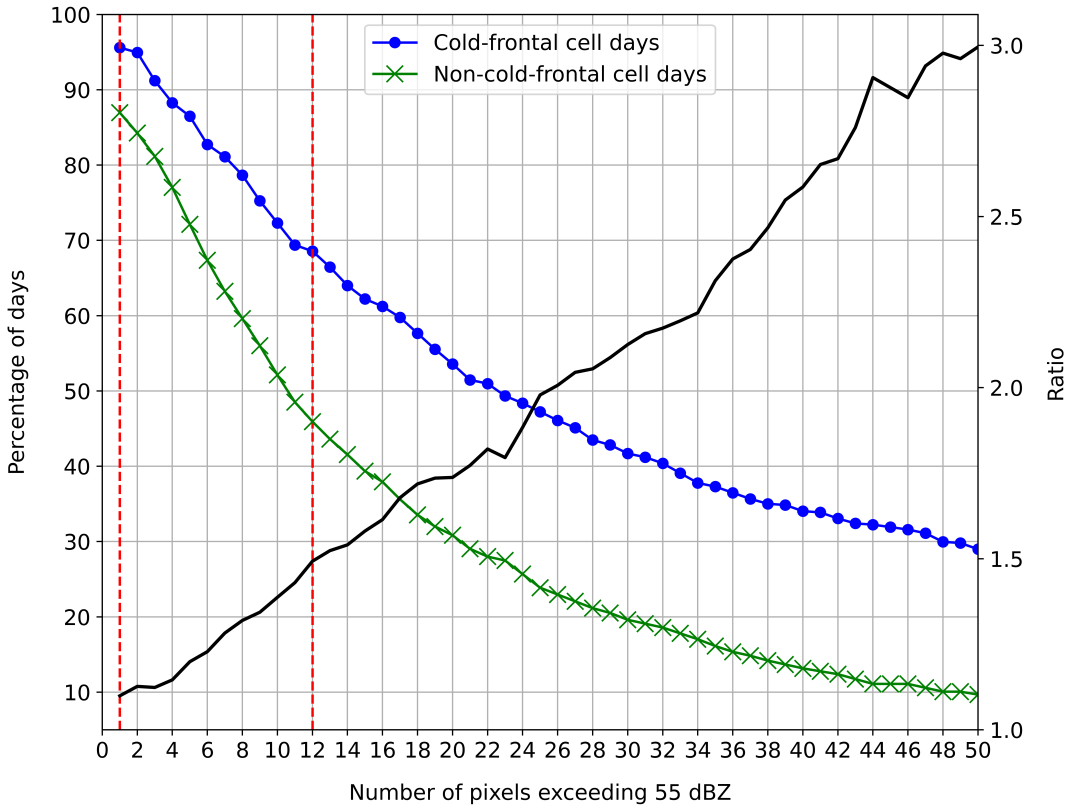


Figure 1.10: Percentage of days with at least one cell with n pixels exceeding 55 dBZ for cold-frontal cell days (blue) and non-cold-frontal cell days (green). The ratio (black) of frontal to non-cold-frontal cell days is shown on the secondary axis. The first dashed vertical red line represents the hail flag 1 and second dashed vertical line represents the hail flag 2 (excluding the 1 pixel exceeding 60 dBZ criterion).

Figure 1.11 shows several characteristics of convective cells depending on the cell-front distance. The average for non-cold-frontal cells (dashed green line) is shown for reference. The longest mean cell lifetime (Figure 1.11a) of around 20 minutes is found 450–750 km ahead of the 700 hPa front, thus pre-surface-frontal. For the first cell detection the lifetime assigned is 5 minutes and for subsequent timesteps 5 minutes is added. Pre-700hPa-frontal cells further than 150 km from the 700 hPa front have a longer lifetime than the non-cold-frontal 18 minute average. Post-700hPa-frontal and near-700hPa-frontal cells have the shortest lifetime. Due to the convective cell definition of 46 dBZ (15 pixels), the cell lifetime refers to the time a cell exceeds these thresholds rather than the overall lifetime of the cell, thus the actual lifetime of the convection transitioning from the development, mature and decay phase would be longer. Cell splitting and merging can also lead to an unrealistic lifetime in some cases (see section 2.1 of Wapler, 2021).

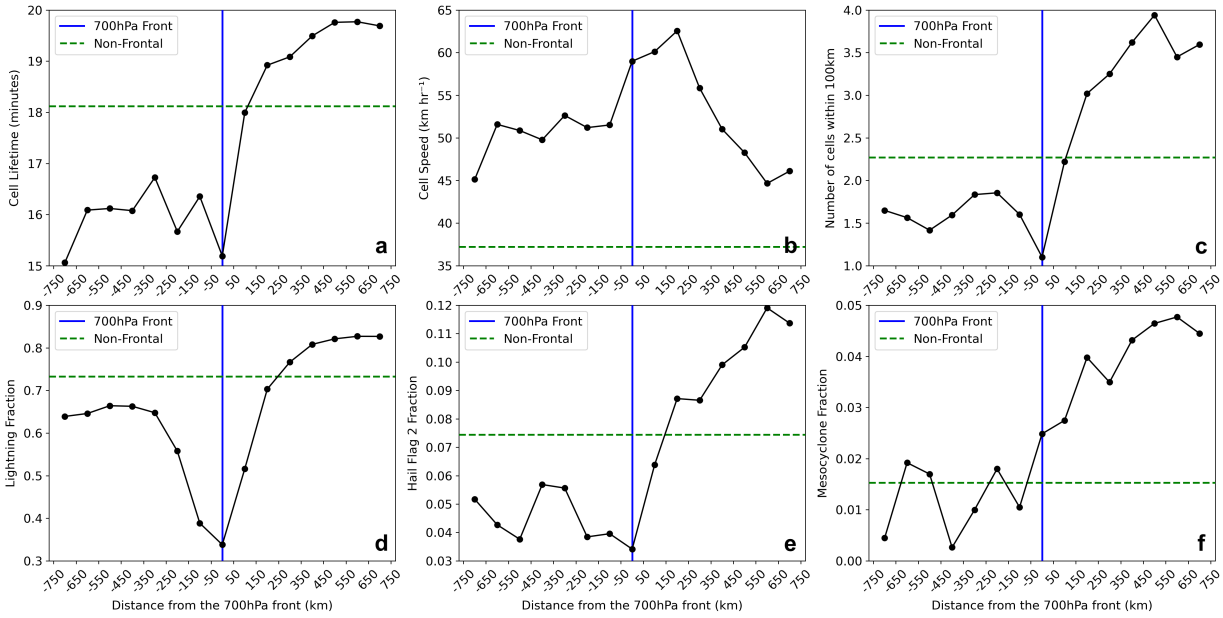


Figure 1.11: Mean cell lifetime (a), mean cell speed (b), mean number of cells within 100 km of cell centre (c), fraction of cells with lightning (d), fraction with hail flag 2 (e) and fraction with mesocyclones (f) depending on the cell-front distance (km). The dashed green line represents the non-cold-frontal cells fraction for reference.

Wapler and James (2015) showed that convective cells in Germany propagate in different directions and with different speeds depending on the synoptic pattern. The cell speed is important as it can favour the likelihood of certain convective hazards. Slowly propagating convective systems typically lead to an increased chance of flooding. For example, a series of slowly moving thunderstorms under weak flow led to severe flooding in Germany in 2006 (Piper et al., 2016). On the other hand, faster propagating systems are more likely to be associated with severe convective winds. For example, Gatzen et al. (2020) showed that both warm and cold season bow echoes in Germany occurred under strong 500 hPa flow. In cold-frontal environments we find cells propagate faster than non-cold-frontal cells at all locations relative to the front (Figure 1.11b). The maximum mean cell speed is found 150–250 km ahead of the 700 hPa front. Two minimums are found, one 550–650 km ahead of the 700 hPa front (pre-surface-frontal) and 650–750 km behind the 700 hPa front.

Once convection has initiated, secondary convection is often found in proximity due to interactions between cell outflows and other boundaries (Wulfmeyer et al., 2011). Radar and satellite observations highlight that post-frontal convective cells are typically scattered in nature (Theusner and Hauf, 2004; Weusthoff and Hauf, 2008). Theusner and Hauf (2004) showed that 72% of post-700hPa-frontal cells were single cells based on 39 warm-season days in Germany. We calculate the number of cells present within a 100 km radius at the detection time to provide an estimate of the aforementioned interactions (Figure 1.11c). We find that pre-surface-frontal cells (350–750 km) have the largest number of cells present in a 100 km radius (around 4 other cells). Cell outflow boundaries interacting with other surface convergence lines such as the surface front and pre-surface-frontal convergence lines could result in increased convergence and the initiation of new cells. The minimum is found near-700hPa-frontal. where only 1 additional cell is present on average. Post-700hPa-frontal cells are also associated with a smaller amount of cells in proximity (1–1.5 cells) compared to pre-700hPa-frontal cells and non-frontal cells (2.25 cells).

Figure 1.11d indicates the fraction of cells associated with lightning depending on the cell-front distance. A cell is considered to be associated with lightning if 3 or more lightning strokes are detected 20 km or less from the cell centre during the cell's lifetime. We require at least 3 strokes to filter out any erroneous detected strokes, as was carried by Púčik et al. (2015) to confirm presence of a thunderstorm. The lightning fraction maximum is found 550–650 km ahead of the 700 hPa front. Other pre-surface-frontal cells are also commonly associated with lightning (around 80% of cells). The fraction of cells associated with lightning sharply declines with increasing proximity to the 700 hPa cold front. The fraction of cells with lightning increases again in the post-700hPa-frontal environment but the overall fraction with lightning is lower than most pre-700hPa-frontal cells and the non-frontal fraction (0.73). Observational studies have indicated that low instability in the lower mixed-phase region can discriminate between convection associated with lightning and non-lightning convection (e.g., van den Broeke et al., 2005). Differences in the typical vertical profile could vary across the front explaining the observed differences in lightning probability. Figure 1.3 shows that across the front CAPE is climatologically lowest near the 700 hPa front corresponding to the minimum cell lightning fraction and highest 550–600 km ahead of the 700 hPa front corresponding to the highest cell lightning fraction.

Figure 1.11e is complementary to Figure 1.10 highlighting it is generally pre-surface-frontal cells that have an increased likelihood of becoming intense compared to non-cold-frontal cells. Like the cell lifetime, cells further than 150 km ahead of the 700 hPa front surpass the non-cold-frontal average. The largest fraction with the hail flag (around 12%) is cells 550–650 km ahead of the 700 hPa front, around 200 km ahead of the cell frequency maximum (Figure 1.5), and where CAPE is climatologically highest (Figure 1.3). As with the lightning fraction, near-700hPa-frontal cells have the lowest fraction associated with the hail flag 2. Following the result that near-700hPa-frontal cells are typically less intense one may suspect the lower lightning fraction is simply related to cell intensity. However, upon considering cells that only meet the minimum cell threshold (15 pixels at 46 dBZ and no pixels at 55 dBZ), the same distribution was found with the lowest lightning fraction near-700hPa-frontal and the highest pre-surface-frontal. The largest post-700-hPa-frontal hail flag 2 fraction is found 250–450 km behind the 700 hPa front, closely corresponding to the maximum frequency of post-700 hPa frontal cells (Figure 1.5).

Mesoyclones, where a rotating updraft is present, are known to be linked to severe convective hazards such as hail and tornadoes. For example, Wapler (2017), showed 75% of hail events in Germany were associated with mesocyclones. Based on the mesocyclone criteria introduced in section 1.2.2 (see Table 1.1), around 5% and 1.5% of pre-surface-frontal cells and non-cold-frontal cells had mesocyclones, respectively. Near-700hPa-frontal cells are also more likely to be associated with a mesocyclone than non-cold-frontal or post-700hPa-frontal cells. A key result from this section is that while convective cells are most frequent near the surface front (Figure 1.5a), when they do develop further ahead of the surface front they are as likely or more likely to be associated with lightning, 55 dBZ cores and mesocyclones, especially in comparison to non-cold-frontal cells.

1.3.6 Lightning strike and hail report frequency

In order to understand if the results of the climatology are refined to Germany or could also apply on a larger European scale, lightning strike data are used in the same 700 hPa front framework. ESWD hail reports are also investigated in the same framework. The datasets are explained in more detail in section 1.2.3. Lightning frequency exhibits a similar distribution around the front in a sub-European domain (Figure 1.12) to the cell frequency in Germany (Figure 1.5a), but with a slight shift in the maximum to the region 400–450 km ahead

of the 700 hPa front. The low lightning frequency after the passage of the 700 hPa front is particularly clear, with an increase in the frequency at further distances behind the 700 hPa front. A slight increase in lightning frequency is observed around 650–700 km ahead of the 700 hPa front, possibly linked to pre-surface-frontal convergence lines (Dahl and Fischer, 2016). The similarity between Figures 1.5 and Figure 1.12 indicates that the frequency around the front is not confined to Germany nor to the KONRAD convective cell definitions. Nevertheless, the overall fraction of convective cells with lightning across the front must also be considered (Figure 1.11d).

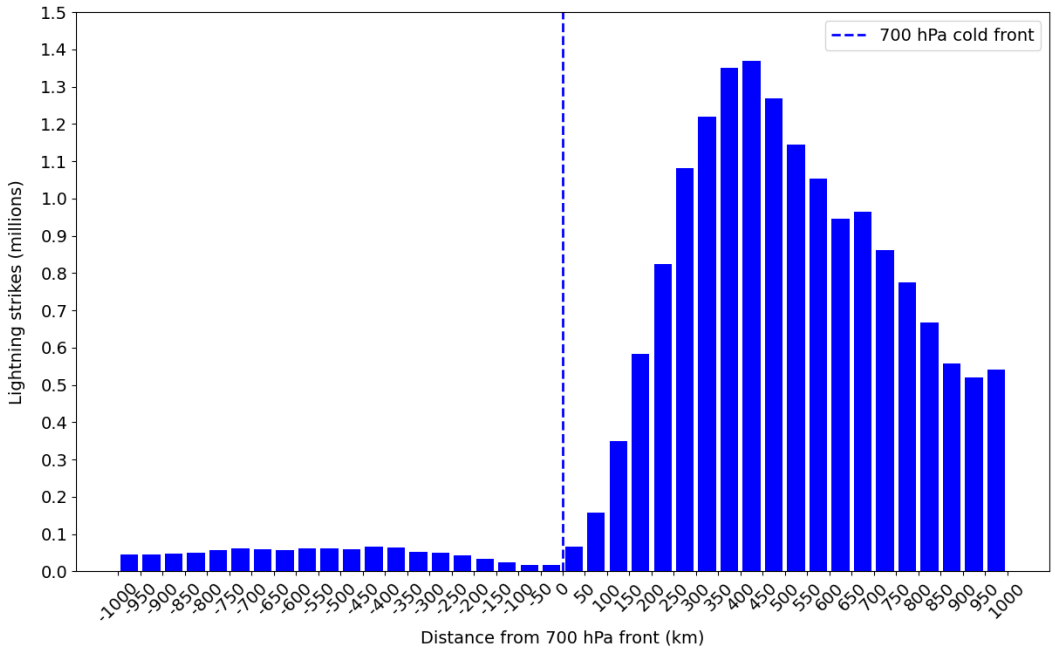


Figure 1.12: Lightning strike frequency between 2008–2016 in a sub-European domain (see grey domain in Figure 1.1) depending on the distance from the 700 hPa front. Note that a 1000 km range is used rather than 750 km used in Figure 1.5.

The convective cell diurnal cycle heatmap (Figure 1.6) as well as the spatial map frequency for post-700hPa-frontal (Figure 1.8a) and pre-700hPa-frontal (Figure 1.8b) were created with the lightning strike data and showed broadly consistent results (Figures A7, A8 and A9).

Out of the 2736 ESWD cold-frontal hail reports available between 2007–2016 (April-September), 96% were reported in the pre-700hPa-frontal region (Figure 1.13). Hail was rarely reported behind the 700 hPa front. The third most common region for hail is in the 200–300 km front relative region, which would typically be behind the surface front. This is a reminder that the risk of convective hazards does not cease to exist once the surface front has passed, consistent with some previous literature. For example, Schumacher et al. (2022) documented a severe convective wind event in the United States that occurred after the surface front had passed. Since there is a relatively low hail report frequency near the 700 hPa (and behind), the level could be used as a reference point in operational forecasting for where hail is less likely to occur. On the other hand, the highest hail frequency occurs marginally ahead of the surface front, similar to where convective cells and lightning are most frequent. As seen for convective cells (Figure 1.5) and lightning strikes (Figure 1.12), the frequency of hail is not symmetric around the mean surface front location. There is increased likelihood at

distances ahead of the surface front compared to the same distance behind the surface front.
The available hail reports likely only represent a subset of all hail reports due to underreporting with a bias towards central Europe (e.g. Groenemeijer et al., 2017 and Taszarek et al.; 2020a). Therefore, it was not practical to assign each hail report to a convective cell in Germany and then calculate the fraction of cells with hail and enable an analysis similar to that in Figure 1.11. As a consequence, it is not known with certainty whether hail is more likely pre-surface-frontal given convective cell detection. Figure 1.11e does support this hypothesis, but there are limitations of using radar based proxies for hail. KONRAD is based on a 0.5 degree radar scan thus the height relative to the ground varies with distance from the radar. At larger distances from the radar the reflectivity being sampled is above the ground so a 55 dBZ region with hail may not actually reach the ground due to melting.

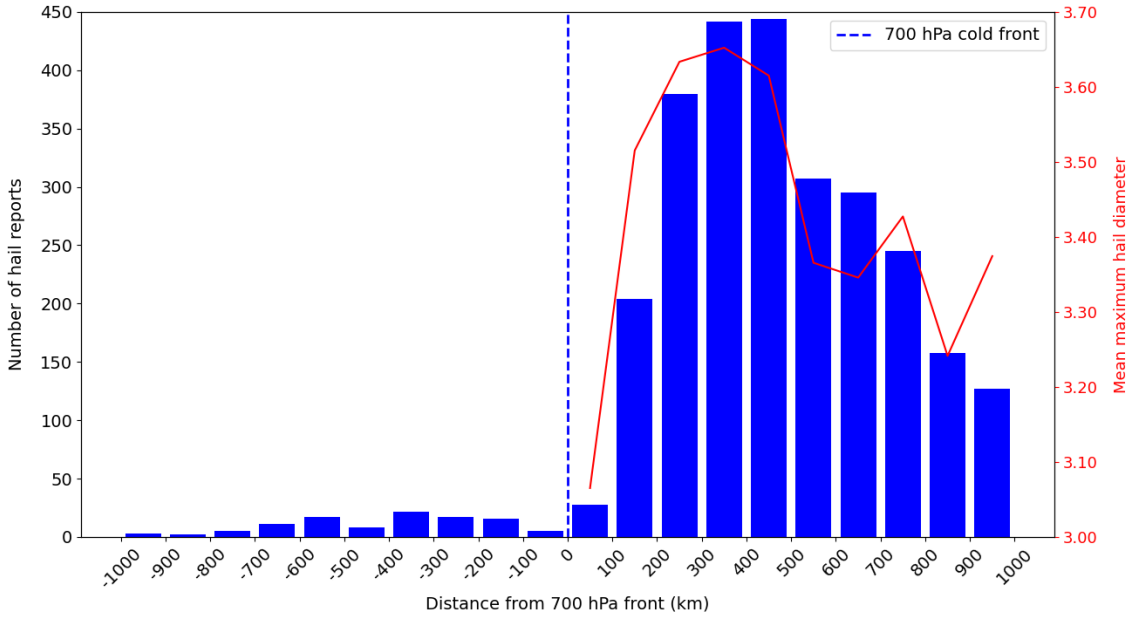


Figure 1.13: ESWD hail report frequency between 2007–2016 in a sub-European domain (see grey domain in Figure 1.1) depending on the distance from the 700 hPa front.

The secondary axis (red) on Figure 1.13 shows the mean maximum hail diameter. The mean refers to the mean across each 100 km front relative region. Since there are few reports post-700hPa-frontal and not all reports have the maximum hail diameter available, only pre-700hPa-frontal hail diameters are analysed. There is some variation in the mean maximum hail diameter across the front, however the trend broadly follows the frequency of hail reports. The maximum and minimum mean diameter is at 300 to 400 km and 0 to 100 km, respectively. The mean hail diameter on non-cold-frontal hail days was 3.4 cm, slightly lower than the 300–400 km front relative region. Kunz et al. (2020) also found cold-frontal hailstones to generally have larger diameters than their non-cold-frontal counterparts.

1.4 Conclusions

A novel climatology of cold-frontal convective cells in the German radar domain (Figure 1.2) has been produced using automatic front detection methods (section 1.2.1) and a convective cell detection and tracking dataset (section 1.2.2) between 2007–2016 for April–September. Furthermore, the nature of cells such as the intensity, lightning strike frequency and mesocyclone frequency has been analysed. Previous cold-frontal convection studies in Europe had primarily focused on narrow cold-frontal rainbands (Gatzen, 2011; Clark, 2013) or on cold-frontal hailstorms (Schemm et al., 2016; Kunz et al., 2020). In this study, we focused on the nature and climatological differences of convective cells depending on their distance from the front as well as making a comparison with non-cold-frontal cells.

The primary findings of the study are outlined below:

- Cold-frontal cell days are associated with around twice as many cells on average compared to non-cold-frontal cell days (Figure 1.4). Cold-frontal cell days are more likely to be associated with cells with large 55 dBZ cores. Around 30% of cold-frontal convective cell days were associated with at least one cell with a 50 km^2 area exceeding 55 dBZ opposed to only 10% of non-cold-frontal cell days (Figure 1.10).

- Convective cells are most frequent 350–400 km ahead of the 700 hPa front (Figure 1.5a), which is marginally ahead of the typical surface front location (section 1.3.1; Figure 1.3). Surface dewpoints are highest in the same region (Figure 1.3).

- The minimum cell frequency is directly at the 700 hPa front with over 5 times less cells developing compared to the 350–400 km maximum (Figure 1.5a). The minimum CAPE is found at the 700 hPa front and the strongest divergence in the lower levels (Figure 1.3).

- Post-700hPa-frontal cells most commonly develop 350–400 km behind the 700 hPa front (Figure 1.5a), where solar heating increases after the passing of the 700 hPa front and CAPE also increases slightly (Figure 1.3).

- Post-700hPa-frontal and non-cold-frontal cells exhibit a typical convective diurnal cycle with a maximum between 15–16 UTC. Pre-700hPa-frontal cells exhibit a weakened diurnal cycle with around a third of all cells developing between 21–06 UTC (Figures 1.6 and 1.7). The weakened cycle is consistent with observations of mesoscale convective systems and mesocyclones in Europe.

- Post-700hPa-frontal cells are most frequent in north-west Germany, particularly close to the coast (Figure 1.8a). Pre-700hPa-frontal and non-cold-frontal cells show similar spatial distributions with cells most frequent in the southern half of Germany (Figures 1.8b and 1.8d). Cells in south-east and eastern Germany are less associated with cold fronts than other regions (Figure 1.8f). Cold-frontal cell frequency is highest in Germany when the 700 hPa front is orientated NE-SW to the west or north-west of Germany (Figure 1.9).

- Pre-700hPa-frontal cells (particularly pre-surface-frontal cells) are longer-living and more often associated with 55 dBZ cores, lightning and mesocyclones than cells in other categories. The opposite is true for post-700hPa-frontal cells compared to non-cold-frontal cells excluding mesocyclones where less significant differences were found between the non-cold-frontal and post-700hPa-frontal fractions.

CAPE is climatologically highest ahead of the surface front (Figure 1.3) providing some indication as to why a higher fraction of pre-surface-frontal cells have 55 dBZ cores, lightning and mesocyclones (Figure 1.11). Cells at all distances relative to the front propagate faster on average than non-cold-frontal cells (Figure 1.11).

- A preliminary climatology of lightning and hail reports in a sub-European domain showed that lightning strikes and hail reports are most frequent 400–450 km ahead of the 700 hPa front and least frequent near-700hPa-frontal (Figures 1.12 and 1.13). This is broadly consistent with the convective cell distribution around the front in Germany (Figure 1.5a).

These results act as an important reference for future studies on cold-frontal convection. Since our results have shown the climatology and nature of convective cells vary substantially in Germany depending on their cell-front distance and in comparison to non-cold-frontal cells, this research should prompt further studies on cold-frontal convection especially regarding links to convective hazards. Future studies should also be performed on a pan-European scale to highlight any potential spatial differences in cold-frontal convection climatology.

Chapter 2

Thermodynamic environments and lifting mechanisms of cold-frontal convective cells during the warm-season in Germany

2.1 Introduction

What exactly drives convective cells to develop in space and time is not currently well-understood. There has been improvement in recent times of the representation of convective initiation in numerical models; primarily due to convective permitting models (CPMs) at increased resolution. However, biases regarding the positioning, timing and intensity of convection still remain (e.g. Kain et al., 2008; Klasa et al., 2018). What is clear is the necessity of three primary ingredients to facilitate deep moist convection: moisture, lift and instability (Doswell et al., 1996). Vertical wind shear is also required to allow convective storms to become organised. Moisture, instability and wind shear can be directly measured in the storm's environment using vertical profiles. Traditionally, this was done using observational proximity soundings (e.g. Brooks et al., 1994; Púčik et al., 2015; Kolendowicz et al., 2017; Taszarek et al., 2017; Pacey et al., 2021). However, the advent of ERA5 reanalysis data on a global scale with a spatial and temporal resolution of 0.25 degrees and 1 hour has enabled more localised studies where observational soundings are lacking (e.g. Taszarek et al., 2020b); Calvo-Sancho et al., 2022: Poręba et al., 2022; León-Cruz et al., 2023). The precursor ERA-Interim also allowed analysis of convective storm environments (e.g. Taszarek et al., 2018), albeit at a coarser spatial and time resolution.

Most previous studies have sought to better understand the environments in which convective storms form by analysing quantities such as mid- and low-level moisture, instability, wind shear etc, but few studies have analysed the availability of individual lifting mechanisms as well as moisture and instability. Determining whether air parcels will be lifted to their level of free convection (LFC) is essential for convective initiation and is often difficult to forecast. It is worth mentioning that moist air parcels reaching their LFC is not a guarantee of deep moist convection due to the possibility of entrainment (Morrison et al., 2022). Locating where and when convective cells develop and the responsible mechanisms is essential for accurate prediction

of convective hazards such as hail, strong winds and tornadoes. This is because if parcels are not lifted to their LFC and thus no CAPE is released then there is no convective cell and no possibility of SCSs. This is regardless of whether there is a seemingly favourable environment for SCSs (e.g. high CAPE and high vertical wind shear). Indeed, warm season forecast failures for severe convective storms are often related to difficulties anticipating convective initiation (Markowski and Richardson, 2010). Despite this, the environments in which severe convective storms form have arguably attracted more attention in previous literature compared to convective initiation.

Individual lifting (triggering) mechanisms such as sea-breezes, outflow boundaries and quasi-geostrophic forcing for ascent cannot be directly sampled in environmental vertical profiles. The latter is thought to be particularly relevant in proximity to cold fronts. Due to the weaker horizontal gradients frontal lifting is typically weaker during the warm-season. The literature would benefit from studies quantifying the relevance of frontal lifting at different regions relative to the front, especially during the warm-season. Scholars have alluded to the role of solar heating combined with frontal lift in ultimately determining where and when convection occurs in proximity to cold fronts (e.g. Doswell, 2001). Behind the front, where there is generally large-scale descending motion (Figure 1.3), solar heating seems to carry more importance (Figures 1.6 and 1.7), as discussed in detail in section 1.3.3. However, when convection occurs the large-scale descent may be weaker or localised areas of ascending motion may aid the development of convective cells.

In this chapter, individual lifting mechanisms such as solar heating and quasi-geostrophic forcing for ascent as well as the thermodynamic environments of cold-frontal convective cells will be analysed. Vertical wind shear will also be analysed as it may both positively and negatively affect convective initiation (Peters et al., 2022). Moist updrafts in sheared environments have been shown in simulations to have lower terminus heights than their non-sheared counterparts (e.g. Peters et al., 2019). On the other hand, updrafts are typically wider in sheared environments so may be less susceptible to entrainment. As in Chapter 1, automatic front detection methods and a convective cell detection and tracking dataset are combined and a special focus is placed on the variation of the environment and lifting mechanisms depending on the distance from the front. Rather than only focusing on ERA5 grid points associated with convective cells, grid points not associated with cells while convection is ongoing elsewhere as well as grid points with no cells away from cell regions are also considered. This allows analysis of how well certain variables can distinguish between grid points where convective cells *did* and *did not* develop. The separation into the three categories will be explained in section 2.2.2. The primary research questions addressed in this chapter are as follows:

Q1) How do the thermodynamic environments of convective cells vary across the front?
Q2) What is the relevance of quasi-geostrophic forcing for ascent (descent) and solar heating at different regions relative to the front?
Q3) By analysing the thermodynamic and lifting mechanisms can we explain the results of the climatology in Chapter 1?

This chapter is organised as follows:

Section 2.2 introduces the methods used in this chapter, which are similar to those used in Chapter 1. Station data from the German Weather Service (DWD) are also used alongside ERA5 and KONRAD. Section 2.3 performs a statistical comparison of the environments and lifting mechanisms of convective cells

depending on the distance from the front. Section 2.4 puts the results in the context of Chapter 1.

2.2 Methodology

As in Chapter 1, a convective cell detection and tracking dataset (KONRAD; Wapler and James, 2015) and ERA5 data (Hersbach et al., 2020) are combined between 2007–2016 for the months April–September. Cold fronts and convective cells are detected in the same way as introduced in section 1.2 so will not be reintroduced in this chapter. Rather than deriving the distance from the cell centre to the 700 hPa front, the distance between ERA5 grid points and the front is calculated. Convective cells are then assigned to ERA5 grid points in space and time. The KONRAD dataset has a spatial and temporal resolution of 1 km and 5 minutes respectively. ERA5 has a 0.25 degree spatial resolution and a 1 hourly temporal resolution. Spatially, the ERA5 grid points within the maximum north, south, west and eastern extent of the cell area are labelled as convective cell grid points. Since some cells have lower area than the grid size, the bounds are increased by 0.125 degrees (half a grid point) to ensure every cell is associated to at least one grid point. Applying this approach ensures that the area where the convection is occurring is labelled as a convective cell grid point. Temporally, cells are associated to the timestep before the first cell detection time. For example, a cell first detected between 14:00–14:59 UTC is assigned to the 14 UTC timestep. This is to avoid sampling the post-convective environment, which is particularly important for thermodynamic variables such as convective available potential energy (CAPE).

2.2.1 Variables

The variables analysed in this chapter broadly follow the ingredients based methodology proposed by Doswell et al. (1996); moisture, lift and instability. Wind shear is also considered as it related to the organisation of convective cells (e.g. Markowski and Richardson, 2010) and may both positively and negatively affect convective initiation (e.g. Peters et al., 2022). Large-scale and convective precipitation are analysed to see how ERA5 represents precipitation across the front. A full list of the variables analysed in this study are shown in Table 2.1.

Table 2.1: A list of variables analysed in this chapter and the associated level and dataset.

Variable	**Level**	**Dataset**
Dewpoint temperature	2-metres above ground level	ERA5
Relative Humidity	700 hPa, 850-500 hPa average	ERA5
CAPE[1]	single level	ERA5
Convective Inhibition	different departure levels	ERA5[2]
Q-vector convergence	850 hPa, 700 hPa, 500 hPa	ERA5
Vertical Velocity	850 hPa, 700 hPa, 500 hPa	ERA5
Total incoming solar radiation	surface	ERA5
Sunshine hours	surface	DWD Station Data
Vertical Wind Shear	surface-500 hPa, 0-3 km, 0-6 km AGL[3]	ERA5
Precipitation	surface	ERA5

[1]The CAPE used in this thesis is the ERA5 CAPE parameter downloaded from the Copernicus Climate Data Store (Hersbach et al., 2018b). CAPE is derived considering parcels departing from different model levels below the 350 hPa level and the departure level with the highest CAPE is retained. In essence, the ERA5 CAPE parameter is the most unstable CAPE (MUCAPE). The following is assumed:(i) the parcel of

air does not mix with surrounding air; (ii) ascent is pseudo-adiabatic (all condensed water falls out) and (iii) other simplifications related to the mixed-phase condensational heating.

[2]Methods to calculate CIN are shown under the heading **Convective Inhibition Dataset**

[3]Above ground level (AGL)

Q-vectors

Q-vectors are derived using the Python package MetPy (May et al., 2022), which derives Q-vectors in the following way (Equation 2.1).

$$\begin{aligned} Q_i &= -\frac{R}{\sigma p}\left[\frac{\partial u_g}{\partial x}\frac{\partial T}{\partial x} + \frac{\partial v_g}{\partial x}\frac{\partial T}{\partial y}\right] \\ Q_j &= -\frac{R}{\sigma p}\left[\frac{\partial u_g}{\partial y}\frac{\partial T}{\partial x} + \frac{\partial v_g}{\partial y}\frac{\partial T}{\partial y}\right] \end{aligned} \tag{2.1}$$

Where:

Q_i : x component of the Q-vector
Q_j : y component of the Q-vector
R : Gas constant for dry air
σ : Static stability parameter
p : Pressure
u_g : u component of geostrophic wind
v_g : v component of geostrophic wind
T : air temperature

The u and v components of the geostrophic wind are derived from the geopotential height field. The geopotential height fields are first smoothed using a simple smoothing function whereby the nearest four neighbours of a grid point are averaged. The process is repeated 50 times. The air temperature field (T) is also smoothed 50 times. The smoothing is required to filter out local scale features to be left with the large-scale circulation. Smoothing values between 10 and 100 were tested and 50 was selected as it showed a realistic and smooth frontal circulation with Q-vector convergence ahead of the front and divergence behind the front.

DWD Sunshine Hours Station Data

In section 2.3.6 solar heating is analysed using total incoming solar radiation from ERA5. Observational sunshine duration data from the German Weather Service (DWD) is also analysed to see if the two datasets are in agreement. The data was downloaded from the Open Data Server (Kaspar et al., 2019). The 10-minute station observations of sunshine duration (Deutscher Wetterdienst, 2024) are used for the years 2010–2016 between 09-18 UTC. Only convective cells that were detected within 25 km of a station are considered as

solar radiation can vary on small spatial scales. Since the cell detection in KONRAD (section 1.2.2) refers to the time where the threshold of 46 dBZ is reached, clouds would be expected just before cell detection as cells would be in their developing phase (e.g. Figure P1a). Therefore, the sunshine duration in the period 90 minutes to 30 minutes before cell detection is analysed to ensure the pre-convective environment is sampled. For example, if a cell is detected at 18:10, the sunshine duration between 16:40 and 17:40 is considered.

Convective Inhibition Dataset

The Convective Inhibition (CIN) parameter available from the Climate Data Store assigns a missing value if CIN exceeds 1000 J kg^{-1} or if there is no cloud-base present (Hersbach et al., 2018b). So that a CIN value is present for all grid points, CIN is obtained from an alternative data source (thundeR; Taszarek et al., 2023). CIN is derived in thundeR using ERA5 model level data for three model parcel departure levels: most unstable CIN (MUCIN), mixed-layer CIN (MLCIN) and surface-based CIN (SBCIN). The CIN parameters are calculated by integrating negative parcel buoyancy between the parcel initialisation height and the level of free convection (LFC). The most unstable parcel refers to the parcel with the highest equivalent potential temperature between the surface and 3 km above ground level (AGL). The mixed-layer parcel is calculated by averaging the potential temperature and mixing ratio between the surface and 500 metres AGL and initialising from surface. The surface based parcel is the parcel nearest to the surface.

2.2.2 Definitions of non-cell regions, cell regions and cell grid points

Rather than solely focusing on where convection occurred (cell grid points), the surrounding regions are also assessed (cell regions), as well as regions where no convection occurred (non-cell regions). This comprehensive approach allows us to understand how well certain variables can distinguish between grid points where convective cells *did* and *did not* develop. One hundred kilometre front relative bins are first created. Following the approach in Chapter 1, we only consider grid points within 750 km of the 700 hPa front. Using 100 km intervals there are a total of 15 bins (Table 2.2). Each ERA5 grid point within 750 km of a 700 hPa front detected between 2007–2016 and April–September is associated to one of these bins. If the grid point is associated with at least one convective cell then this is labelled a *convective cell grid point*. If a convective cell has been detected within a certain bin at a certain timestep but *not* at the given grid point then this grid point is labelled as a *cell region*. If no convective cell has been detected within a certain bin at a certain timestep then this grid point is labelled *non-cell region*. The number of grid points associated to each category at different distances from the front is shown in Table 2.2 assigning of a grid points to each category is visualised in Figure 2.1 for 4 independent timesteps. Using the top-left panel in Figure 2.1 as an example, most grid points are labelled non-cell regions (green). Cells were only detected in the bins 250–350 km and 450–550 km, therefore most grid points in this region are labelled cell regions (purple). Grid points where convection occurred (cell grid points) are in yellow.

Table 2.2: Number of grid points in each category (non-cell regions, cell regions and cell grid points) depending on the distance from the 700 hPa front.

Distance from front	Non-cell regions	Cell regions	Cell grid points	Total
-750 to -650 km	375,897	33,165	1,230	410,292
-650 to -550 km	403,822	38,920	1,571	444,313
-550 to -450 km	428,715	49,085	1,942	479,742
-450 to -350 km	458,274	52,914	2,120	513,308
-350 to -250 km	483,334	49,239	1,975	534,548
-250 to -150 km	512,047	42,658	1,753	556,458
-150 to -50 km	539,207	39,002	1,128	579,337
-50 to 50 km	522,995	38,111	1,035	562,141
50 to 150 km	530,638	88,049	3,301	621,988
150 to 250 km	516,253	134,347	7,863	658,463
250 to 350 km	503,009	166,937	10,348	680,294
350 to 450 km	505,580	167,999	11,275	684,854
450 to 550 km	525,694	147,972	9,524	683,190
550 to 650 km	555,236	123,031	7,571	685,838
650 to 750 km	556,448	108,553	7,072	672,073

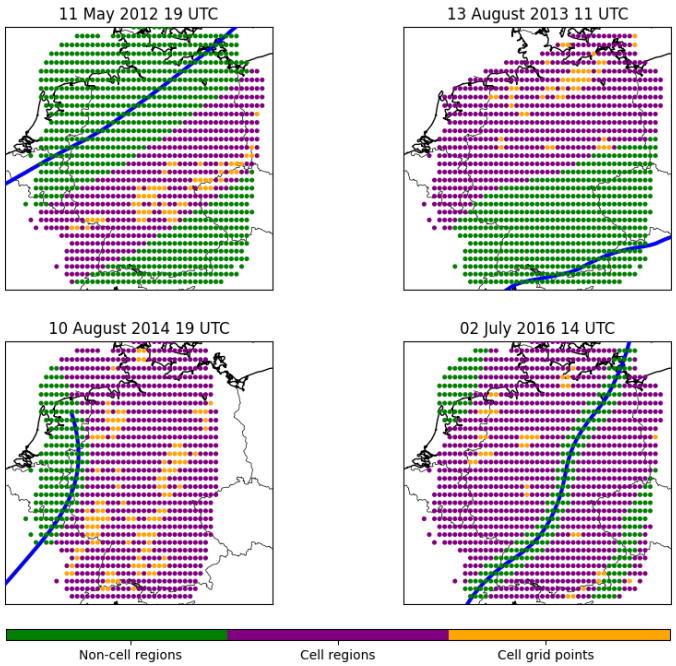


Figure 2.1: Visualisation of how cell grid points (yellow), cell regions (purple) and non-cell regions (green) are defined for four timesteps on different days. The 700 hPa frontal line is shown by the blue line and was detected using the methods defined in section 1.2.1.

2.3 Results

2.3.1 Moisture

Surface dewpoints (2-metres above ground level in this case) are a measure of moisture availability at the surface. While sufficient moisture directly at the surface is not essential for convective initiation since convection can be elevated (e.g. Corfidi et al., 2008), surface dewpoints are a commonly used tool by forecasters. Lower dewpoint air generally requires more lifting to reach the lifting condensation level (LCL). Figure 2.2 shows that pre-700hPa-frontal cells develop in environments with a surface dewpoint of around 15–16 °C on average. Cells 650–750 km ahead of the 700 hPa front have the highest mean dewpoints; the furthest distance ahead of the front. Post-700hPa-frontal cells formed in environments with lower surface dewpoints ranging between 12-14 °C. Dewpoints at cell grid points are around 1 °C higher than at cell region grid points on average. The difference in the mean between cell grid points and non-cell region grid points is significant at all distances relative to the front. The difference between the cell grid point mean and non-cell regions mean ranges between 3-4 °C.

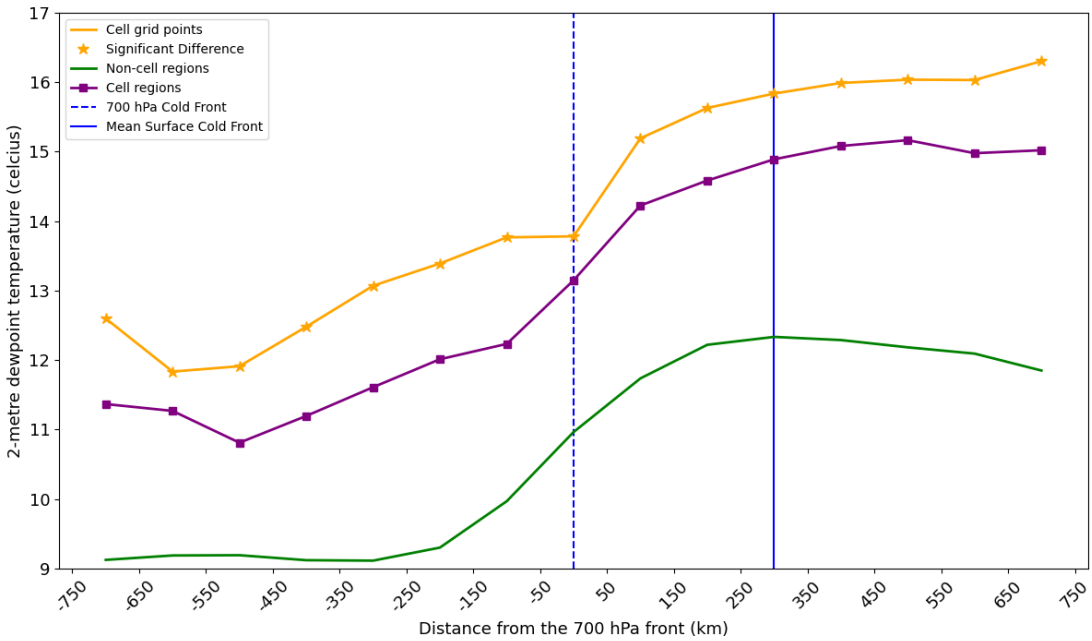


Figure 2.2: 2-metre dewpoint temperature (celcius) depending on distance from the 700 hPa front (km) for cell grid points (orange), cell region grid points (purple) and non-cell region grid points (green). Stars indicate that the convective cell grid point mean is significantly different from the non-cell grid point mean at the 95% confidence level based on a Welch's t-test, which does not assume equal population variance.

Mid-level moisture is also relevant for the initiation of deep moist convection since entrainment of dry environmental air can lead to updraft dilution, hence influencing updraft buoyancy. Recent work has shown that how susceptible the updraft is to dry environmental air is dependant on the updraft width below the level of free convection (LFC) (Morrison et al., 2022). However, the updraft, which occurs on the storm-scale,

will not be looked at in this study. The relative humidity (RH) is analysed at 700 hPa (hereafter RH700) as parcels have typically already passed the LFC at this level so entrainment is important in determining whether updrafts can reach upper levels of the atmosphere.

Cells form in environments with a mean RH700 of between 60–70% post-700hPa-frontal and 70-85% pre-700hPa-frontal (Figure 2.3). Cells have the largest RH between 50-250 km ahead of the 700 hPa front, which is also the case for cell regions and non-cell regions. The RH700 at cell grid points is significantly higher than non-cell regions at all distances from the front except in the region 50–150 km ahead of the 700 hPa front. A larger difference exists between cell grid points and non-cell region grid points post-700hPa-frontal (up to 30% difference) compared to pre-700hPa-frontal (up to 10% difference). This indicates that a larger enhancement of RH700 humidity is required to facilitate post-700hPa-frontal cell development. On the other hand, the warm-sector typically already has high upper-level moisture content. Excluding the region 50-250 km, cell regions have slightly lower RH than cell grid points but are above non-cell region grid points. These results are consistent with environment studies of lightning in Europe which found that lightning is less favourable when mid-level relative humidity is low (Westermayer et al., 2017) and that 75% of lightning cases in Poland had RH700 of 65% or higher (Poręba et al., 2022; their Figure 11). The average RH between 850-500hPa (Figure 2.3) shows a very similar result as RH700.

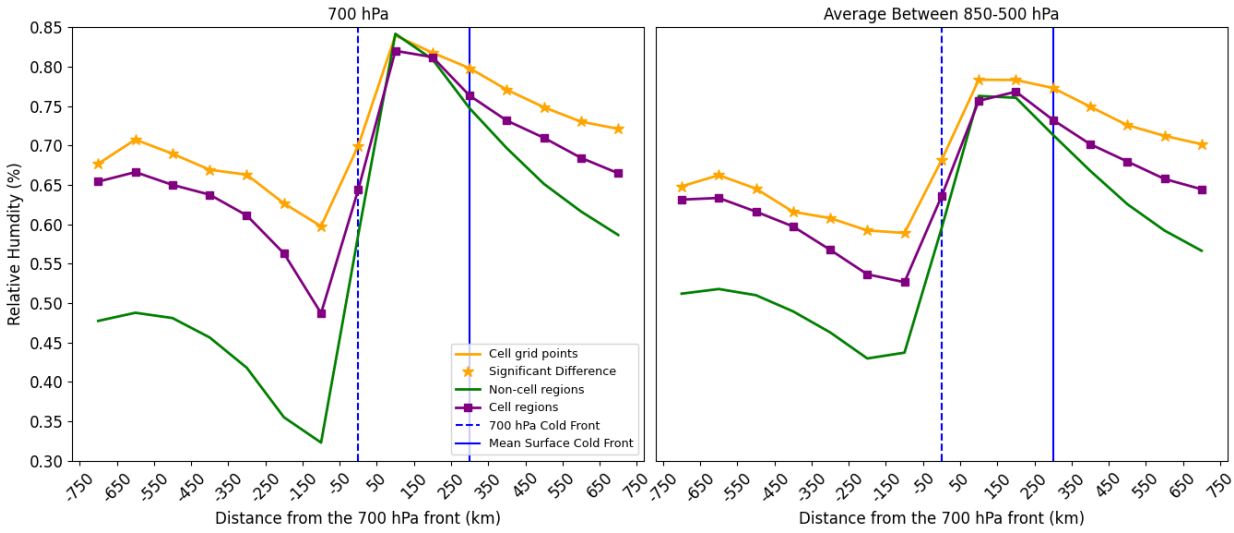


Figure 2.3: As Figure 2.2 but for relative humidity at 700 hPa (%) and mean relative humidity between 850–500 hPa.

2.3.2 Instability

CAPE is a measure of atmospheric instability; a prerequisite for deep moist convection (Doswell et al., 1996). It is also a very commonly used parameter in severe convective storm forecasting. Figure 2.4 shows that pre-surface-frontal cells have the highest mean CAPE; up to around 650 J kg^{-1} 700-750 km ahead of the front. In comparison, post-700hPa-frontal cells occur in environments of lower CAPE; between 150–250 J kg^{-1} on average. The mean CAPE at convective cell grid points is between 5–8 times higher than non-cell regions depending on the front relative region and is significant at the 95% confidence level. The difference between cell grid points and non-cell regions is particularly large pre-surface frontal. Like the moisture variables the cell region means are between the non-cell region and cell grid point means.

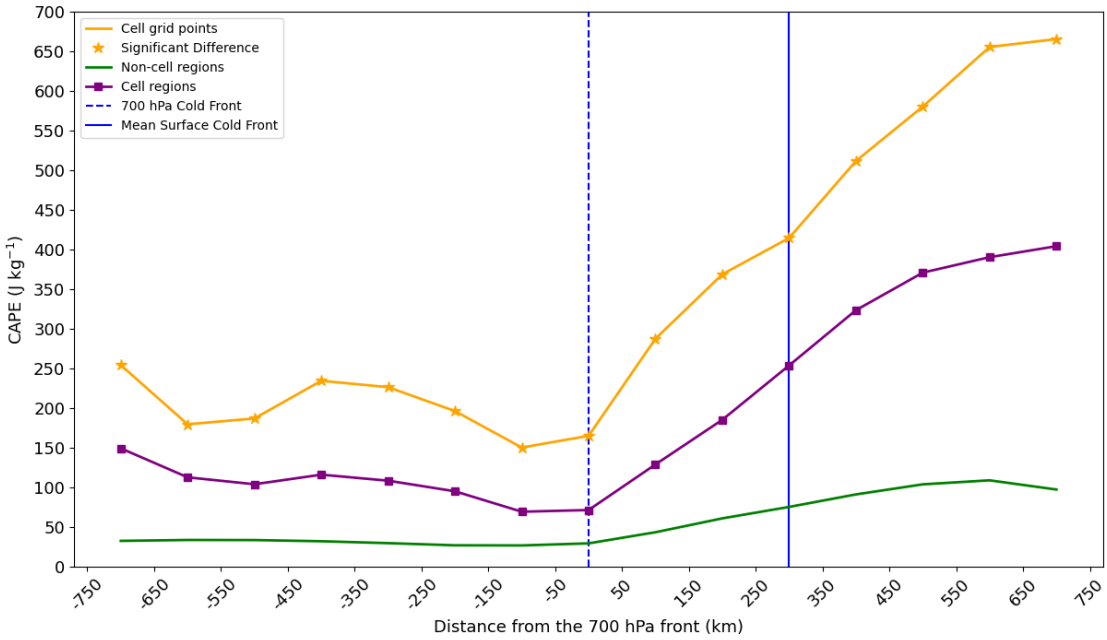


Figure 2.4: As Figure 2.2 but for CAPE. The ERA5 CAPE variable uses the parcel with the highest CAPE considering different departure levels below 350 hPa.

2.3.3 Convective Inhibition

Environments with high convective inhibition (CIN) require stronger lifting so that parcels can reach their LFC. The Convective Inhibition parameter available from the Climate Data Store assigns a missing value if CIN exceeds 1000 J kg^{-1} or if there is no cloud-base present (Hersbach et al., 2018b). So that a CIN value is present for all grid points, CIN is obtained from an alternative data source (thundeR; Taszarek et al., 2023). CIN is derived from ERA5 in thundeR but using a different method to the parameter available on the Climate Data Store. CIN is considered using parcels with different departure levels: most unstable parcel (MUCIN), mixed-layer parcel (MLCIN) and surface based parcel (SBCIN). Due to the large quantity of data that would have needed to be requested only CIN at convective cell grid points is available for analysis. The reader is referred to section 2.2.1 for further details.

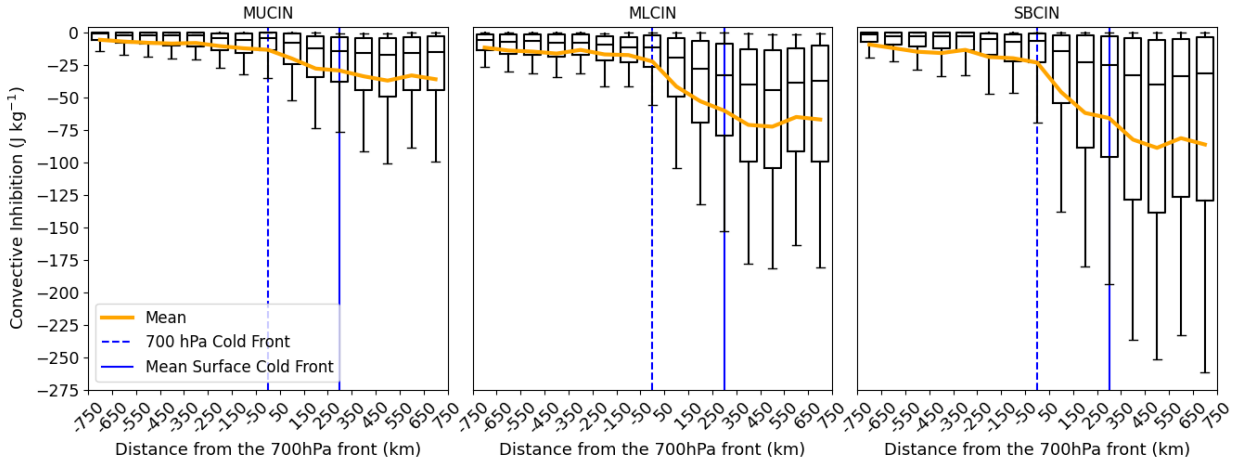


Figure 2.5: MUCIN, MLCIN and SBCIN (J kg^{-1}) for convective cells. The whiskers represent the 10th and 90th percentiles, the median is represented by the horizontal black line and the mean by the orange line. Negative CIN indicates stronger convective inhibition.

Pre-surface-frontal cells formed in environments with the strongest MUCIN, MLCIN and SBCIN with a mean of around -82 J kg^{-1} of SBCIN (Figure 2.5). Post-700hPa-frontal cell environments had relatively weaker SBCIN in comparison (-15 J kg^{-1}). Therefore, more CIN needed to be overcome to initiate convective cells pre-surface-frontal compared to post-700hPa-frontal. However, the stronger pre-surface-frontal CIN may be advantageous for large CAPE build up (see Figure 2.4) since convection will not be initiated prematurely (Ludlam, 1980). MUCIN is lower than MLCIN and SBCIN for pre-700hPa-frontal cells which could be explained by surface inversions (mostly at night and early morning) leading to higher CIN for parcels departing closer to the surface. Figure 2.5 was reproduced using daytime cells only (09–18 UTC) (Figure B1) showing that pre-surface-frontal and post-700hPa-frontal cells have a mean of -8 J kg^{-1} and -28 J kg^{-1} of SBCIN, respectively. A similar result was found for MLCIN and MUCIN. While the overall CIN is weaker during the daytime, pre-surface-frontal CIN is still stronger than post-700hPa-frontal during the daytime. Therefore, the CIN differences depending on the front relative region are not only because a higher fraction of pre-surface-frontal cells occur during the night (Figure 1.6).

2.3.4 Quasi-geostrophic forcing for ascent

Q-vector convergence is a commonly used diagnostic by forecasters to highlight areas of geostrophic forcing for ascent or descent. Q-vectors are derived from the quasi-geostrophic equations (Hoskins et al., 1978). The reader is referred to section 2.2.1 for further details on how Q-vectors are calculated. The 850 hPa level is the typical height of the boundary layer during the daytime and the location of the strongest capping inversion. Rising air parcels must overcome this inversion in order to reach the level of free convection (LFC). The 700 hPa and 500 hPa levels are also analysed to understand the importance of mid-level geostrophic forcing for ascent. Since large-scale lift is typically of the order of cm s^{-1}, it is unlikely sufficient alone to allow air parcels to reach their LFC (Trapp, 2013; his chapter 5.2). However, frontal simulations have indicated the lift associated with the cross frontal circulation may be 1 order of magnitude higher (Koch, 1984; his Figure 7), thus may contribute to overcoming CIN. Synoptic-scale ascent can also steepen lapse rates due to adiabatic

cooling (Trapp, 2013; his Figure 5.2).

Figure 2.6 shows that cells marginally ahead of the surface front have the strongest convergence of the Q-vector at 850 hPa (hereafter QVEC850), which coincides with where convection is most frequent (Figure 1.5a). The strongest QVEC850 convergence of non-cell region grid points is at the mean surface front location. The strongest QVEC850 divergence, which is linked to descending motion, is near-700hPa-frontal for all categories. The strongest QVEC850 divergence closely correlates with where convection is least frequent (Figure 1.5a). However, near-700hPa-frontal the QVEC850 divergence is weaker at cell grid points compared to cell region grid points and non-cell regions. From 750 km behind the 700 hPa front up to near the surface front there is no significant difference between the QVEC850 for non-cell regions and cell regions indicating the importance of QVEC850 specifically where the convective cells are detected. QVEC700 convergence shows a similar result to QVEC850 with slightly weaker mean convergence and divergence.

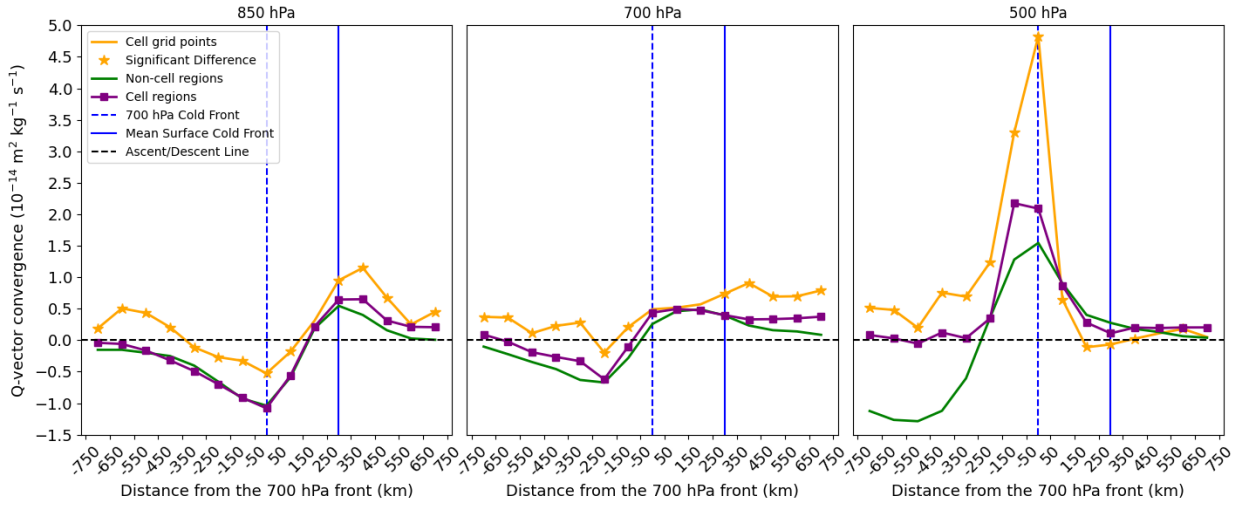


Figure 2.6: As Figure 2.2 but for Q-vector convergence at 850 hPa, 700 hPa and 500 hPa (left to right). Positive and negative values indicate convergence (ascending motion) and divergence (descending motion) of the Q-vector, respectively. Q-vectors are derived using the methodology described in section 2.2.1.

Q-vector convergence at 500 hPa (hereafter QVEC500 convergence) is strongest near-700hPa-frontal across all three categories (Figure 2.6). Pre-700hPa-frontal cells formed in environments with weaker QVEC500 convergence than non-cell regions and cell regions, which is in contrast to QVEC700 and QVEC850 convergence. QVEC500 convergence at cell grid points near-700hPa-frontal (-50 to 50 km) is over 3 times stronger compared to non-cell regions and around 5 times stronger compared to the maximum cell grid point means at 700 hPa and 850 hPa. Non-cell regions post-700hPa-frontal generally have QVEC500 divergence but cell regions and cell grid points have convergence of QVEC500. This result highlights the importance of upper-level forcing particularly on the development of convective cells particularly at the 700 hPa front and also post-700hPa-frontal. The forcing for ascent could be linked to a post-frontal trough or may also act to destabilise upper-layers and hence increase CAPE.

2.3.5 Vertical Velocity

The vertical velocity at 850, 700 and 500 hPa ($w_{850\ \mathrm{hPa}}$, $w_{700\ \mathrm{hPa}}$ and $w_{500\ \mathrm{hPa}}$) is shown from left to right in Figure 2.7. Like Q-vector convergence the variable can be used to highlight areas of ascending and descending motion. However, the vertical velocity is not solely linked to vertical motion due to geostrophic forcing for ascent or descent. This means that additional sources of ascending motion such as areas of convection may also show a signal in the vertical velocity variable in ERA5. At all levels the cell region and non-cell grid point means are similar to the Q-vector in terms of where ascending and descending motion are present (Figure 2.6). However, there is a difference at cell grid points as there is mean ascending motion at all locations relative to the front at all three vertical levels. The strong anomaly at the 700 hPa front which was seen for QVEC500 is not seen for $w_{500\ \mathrm{hPa}}$. The highest vertical velocity is at 500 hPa at cell grid points near the surface front (around 0.06–0.07 m s^{-1}). Ahead of the 700 hPa front vertical motion is maximised at 500 hPa; consistent with observations of quasi-geostrophic vertical motion being maximised around 500 hPa (Holton and Hakim, 2013). It is not clear whether the vertical velocity anomalies at cell grid points are related to convection being partially resolved in ERA5 or relate to stronger large-scale lifting than climatology where convective cells were detected. The non-cell grid points mean suggest that the lifting from the front at 850 hPa, 700 hPa and 500 hPa is maximised between the surface front and the 700 hPa front.

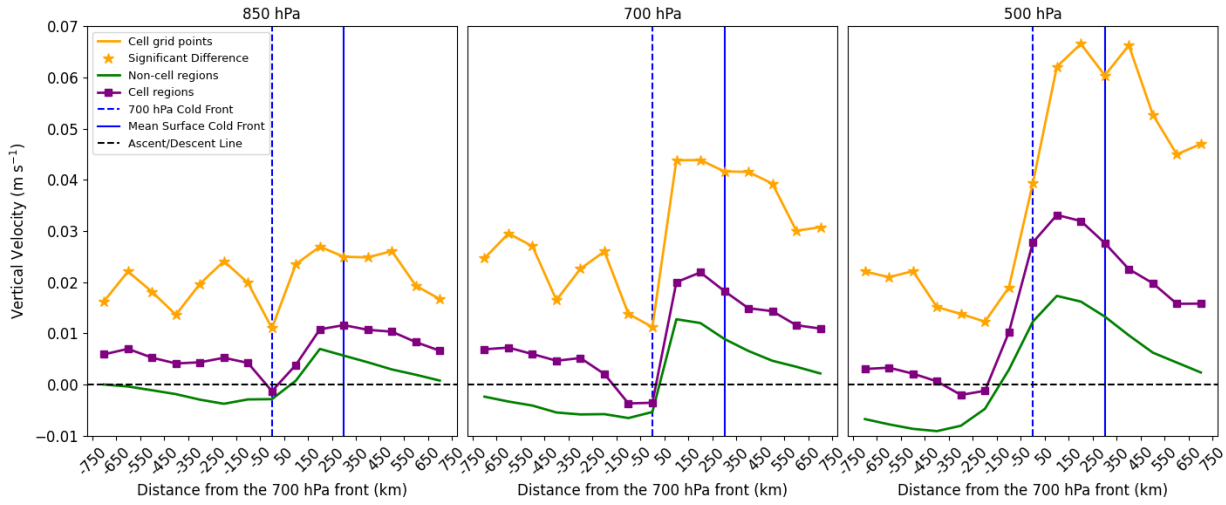


Figure 2.7: As Figure 2.2 but for vertical velocity at 850, 700 and 500 hPa. The signs of vertical velocity are reversed so that positive and negative values indicate ascending and descending motion, respectively.

2.3.6 Solar heating

Solar heating (insolation) is linked to increased surface temperatures, which contributes towards atmospheric instability (Markowski and Richardson, 2010; their Figure 7.9c). Solar heating also gives parcels positive buoyancy near the surface, which can help parcels to be lifted to their LFC. The total incoming solar radiation (hereafter solar radiation) is shown in Figure 2.8 for timesteps between 09–18 UTC only since solar radiation is weaker during the early hours of the morning, the late evening and not possible during the night. The variable refers to the radiation accumulated during the hour prior to the ERA5 timestep, thus prior to convective cell detection. Post-700hPa-frontal cells develop with the largest solar incoming radiation (around

250 W m^{-2}). The lowest solar radiation is ahead of the 700 hPa front but behind the surface front (50 to 150 km region). Solar radiation at cell grid points between -250 to 50 km and pre-surface-frontal is lower than at non-cell regions. The result may seem counterintuitive as solar radiation is thought to be a driver of convective initiation so higher solar radiation would be expected prior to convective initiation. The result may indicate that ERA5 struggles with the timing of convective initiation thus produces convective clouds before the actual convective initiation.

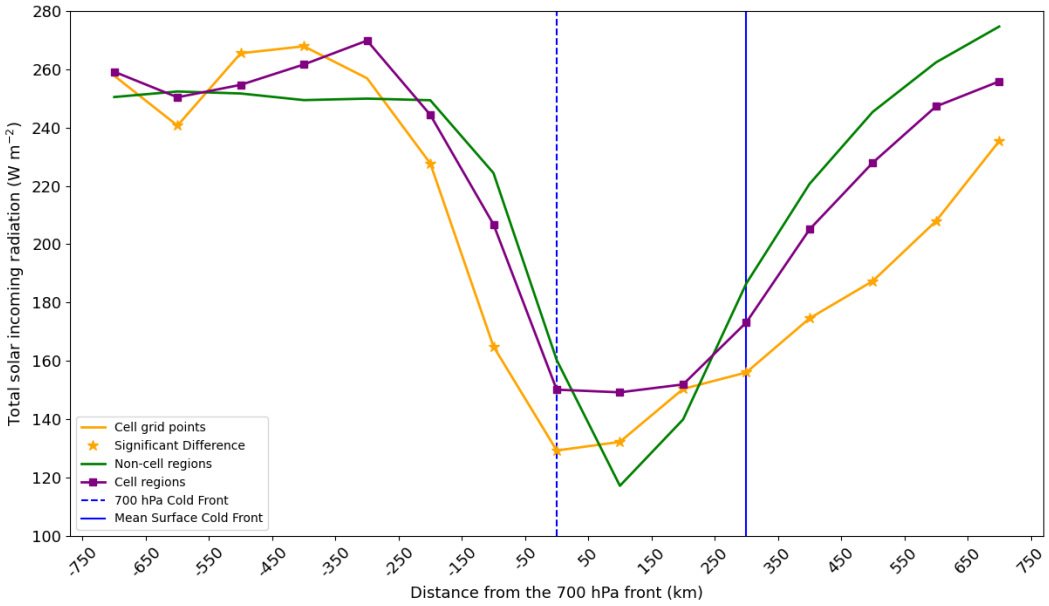


Figure 2.8: As Figure 2.2 but for total incoming solar radiation only using timesteps between 09–18 UTC (W m^{-2}).

To further understand the relevance of solar heating on the development of convective cells surface station data between 2010–2016 from the German Weather Service are analysed. The advantage of also using an observational dataset is that it does not have the potential aforementioned model biases. This in turn allows a more robust assessment of the amount of sunshine (clouds) prior to convective cell detection. Since the cell detection in KONRAD (section 1.2.2) refers to the time where the threshold of 46 dBZ is reached clouds would be expected just before cell detection as cells would be in their developing phase. Therefore, the sunshine duration in the period 90 minutes to 30 minutes before cell detection is analysed to ensure the pre-convective environment is sampled. For example, if a cell is detected at 18:10, the sunshine duration between 16:40 and 17:40 is considered. As in Figure 2.8, only cells detected between 09–18 UTC are considered. The reader is referred to section 2.2.1 for further details.

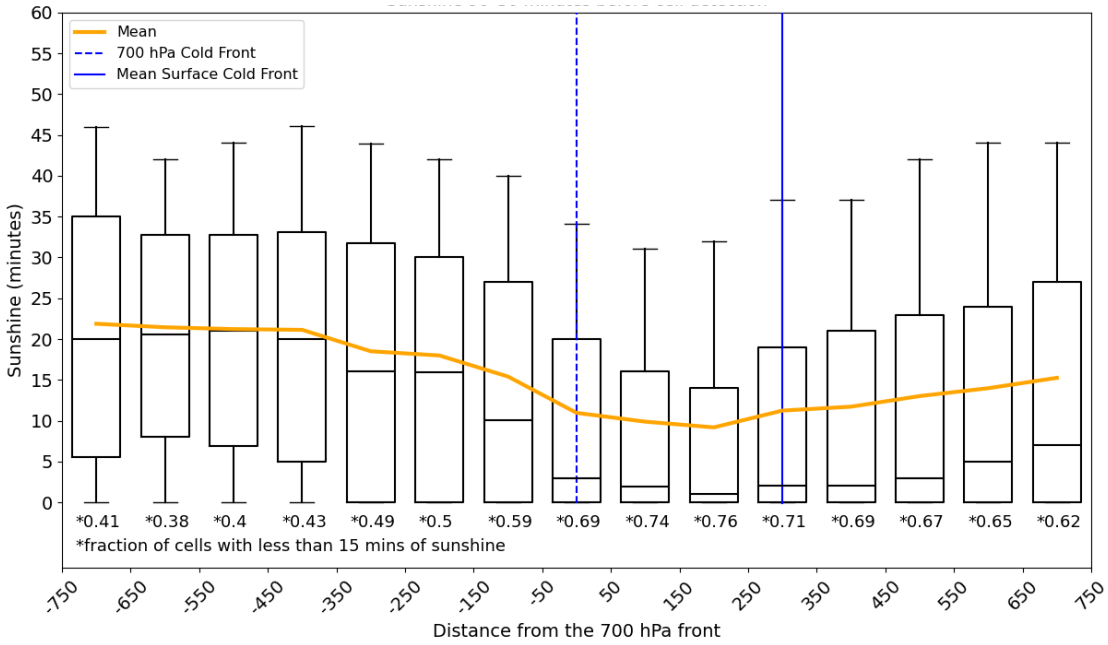


Figure 2.9: Sunshine minutes between 90 and 30 minutes before cell detection. Only cells detected between 09–18 UTC are considered. The whiskers represent the 10th and 90th percentiles, the median is represented by the horizontal black line and the mean by the orange line. The fraction of cells with less than 15 minutes of sunshine is shown under the box plots.

Figure 2.9 shows that most convective cells develop with limited sunshine prior to detection. Post-700hPa-frontal cells and pre-700hPa-frontal cells develop with an average of 20 and 12 minutes of sunshine prior to initiation, respectively. Furthermore, 67% and 44% of pre-700hPa-frontal and post-700hPa-frontal cells had less than 15 minutes of sunshine prior to initiation, respectively. The fraction of cells with less than 15 minutes of sunshine for each 100 km front relative region is shown underneath the box plots in Figure 2.9. Very few cells had a full 60 minutes of sunshine in the 90–30 minute period before detection. The lowest mean sunshine is for cells 150 to 250 km from the front, where cells perhaps typically initiate in the frontal cloud band where solar radiation is lowest (Figure 2.8). Another reason why the sunshine duration is relatively low prior to cell detection could be because cells develop in proximity to pre-existing convective systems and cloud regions. The high percentage of post-700hPa-frontal cells with limited sunshine before initiation is particularly interesting as the post-frontal environment is known to be relatively cloud-free usually with just scattered cumulus. The limited sunshine for most cells before detection (Figure 2.9) and cells having lower total solar radiation than non-cell regions (Figure 2.8) could be partly attributed to cloud cover from pre-existing convective cells. The importance of cold pools on convective initiation has been already highlighted in previous literature. For example, Hirt et al. (2020) showed using high-resolution model simulations for case studies that up to 50% of convective initiation is at the edge of cold pools. At the edge of cold pools it is likely to be cloudy and convection may not be directly triggered by solar heating. The reader is reminded that only cells at the first detection time are considered. In the case of an MCS, several new cells may be detected at subsequent timesteps and assigned the first detection label due to cell recycling (Figure A1). This could also happen for smaller multicell clusters where new cells triggered by neighbouring cells are assigned the first detection label.

2.3.7 Vertical Wind Shear

While vertical wind shear is typically associated with the organisation of convection, cells reaching 46 dBZ may require some wind shear in the environment. When wind speed significantly increases with height (high wind shear) updrafts become titled thus giving more separation of the updraft and downdraft. The downdraft is then less likely to hinder the ascent of warm and moist air (updraft) allowing the possibility of longer-living convective systems. Vertical wind shear can both positively and negatively affect convective initiation (Peters et al., 2022). Numerical simulations have shown that parcel buoyancy may be reduced in high-shear environments due to entrainment (e.g. Markowski and Richardson, 2010 and Peters et al., 2019). More recent work has shown that given updrafts meet an initial width and shear threshold they can widen further, which in turn reduces their susceptibility to entrainment (Morrison et al., 2022 and Peters et al., 2022).

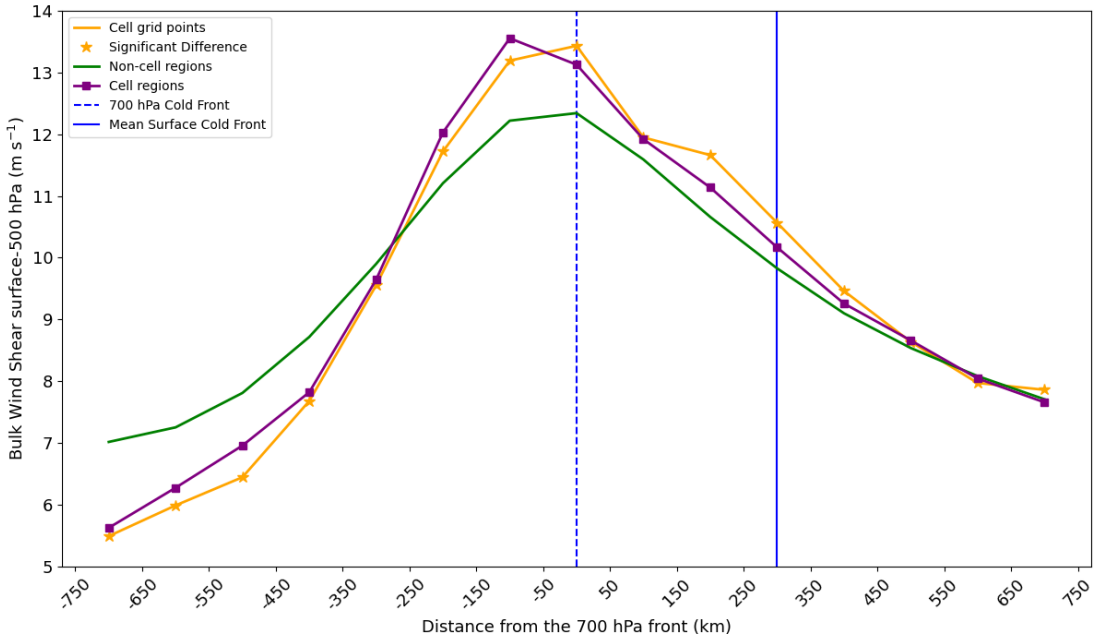


Figure 2.10: As Figure 2.2 but for wind shear between the surface and 500 hPa.

The bulk wind shear between the surface and 500 hPa is highest for convective cell grid points near-700hPa-frontal with around 13 m s^{-1} on average (Figure 2.10). The wind shear near-700hPa-frontal is also higher than non-cell grid points by around 1 m s^{-1}. There is less difference between cell regions and cell grid points near-700hPa-frontal. The wind shear decreases at increasing distance from the 700 hPa front. At a certain distance behind the 700 hPa front wind shear is lower at cell regions and cell grid points compared to non-cell region grid points. While there is some asymmetry around the maximum, post-700hPa-frontal cells generally form in environments with comparable wind shear compared to pre-surface-frontal cells. Therefore, the result from Chapter 1 showing that pre-surface-frontal cells have a larger fraction of 55 dBZ cores and mesocyclones (Figure 1.11) cannot be explained by a lack of wind shear post-700hPa-frontal. Rather thermodynamics likely play a more important role.

Since bulk wind shear between the surface and 500 hPa could potentially lead to a lower vertical distance

being sampled for grid points at higher elevation, the bulk wind shear between the surface and 6 km above ground level (AGL) as well as between the surface and 3 km AGL are also analysed at cell grid points using ERA5 model level data (Taszarek et al., 2023). Even though there are differences in the magnitude of the wind shear between the bulk wind shear surface-500 hPa (Figure 2.10) and 0-6 km AGL (Figure 2.11), the regions of maximum and minimum wind shear around the front are very comparable. Cells near the 700 hPa front have a particularly high 0-6 km AGL mean wind shear of around 23 m s^{-1}. Supercells have been shown in several studies to form in environments with around 20 m s^{-1} of 0-6 km AGL of shear (e.g Doswell and Evans, 2003). Therefore, thermodynamics likely explain why a higher fraction of pre-surface-frontal cells are associated with mesocyclones than near-700hPa-frontal (Figure 1.11e) due to the more frequent overlap of high shear and high CAPE environments pre-surface-frontal.

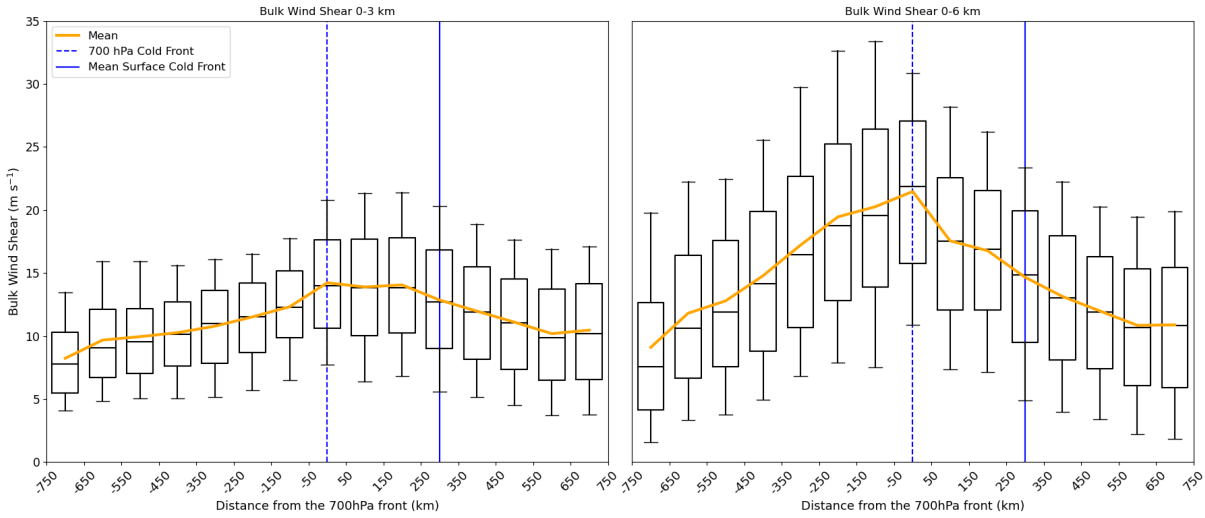


Figure 2.11: As Figure 2.5, but for bulk wind shear between the surface and 3 km and between the surface and 6 km AGL.

2.3.8 Precipitation

To see how ERA5 precipitation is represented across the front, large-scale and convective precipitation are shown in Figure 2.12. The large-scale precipitation is highest for all three categories between the 700 hPa and surface front. This is in agreement with the classical conceptual model of an Ana cold front where the primary precipitation region associated to a cold front is behind the surface front (Figure P2). In this case, the primary precipitation refers to mostly stratiform precipitation that results from condensation in the ascending warm conveyor belt. Figure 2.8 also shows that the lowest solar radiation is at this location relative to the front. Total cloud cover and high cloud cover for non-cell regions are also highest at this location (Figure B2) as well as vertical velocity at non-cell grid points (Figure 2.7). At most regions relative to the front the large-scale precipitation is higher at cell grid points compared to non-cell regions.

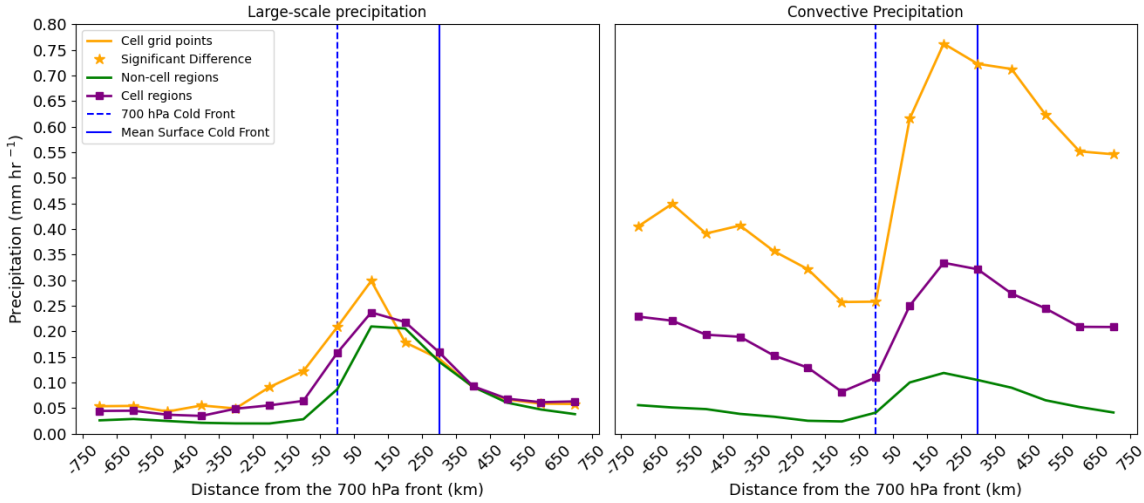


Figure 2.12: As Figure 2.2 but for large-scale precipitation and convective precipitation.

Convective precipitation is maximum 150–250 km ahead of the 700 hPa front for all three categories. This maximum is 150 km behind the convective cell maximum shown in Figure 1.5a. The differences could relate to weaker precipitation than the 46 dBZ being more frequent closer to the front or ERA5 struggling to resolve the location of convection relative to the front. The minimum convective precipitation is 50–150 km behind the 700 hPa front (again for all three categories) before increasing slightly further behind the front. The increase behind the 700 hPa front is consistent with the slight increase seen in the convective cell climatology (Figure 1.5a). Convective precipitation at cell grid points is significantly higher compared to non-cell grid points and shows a similar trend to the vertical velocity fields (Figure 2.7). This corroborates the hypothesis that the vertical velocity signal comes from convection being partially resolved in some cases.

2.4 Unravelling the findings in the climatology

A statistical comparison between cell grid points, cell regions and non-cell regions (section 2.2.2) depending on the region relative to the front was made in section 2.3. How the results in this study relate to Chapter 1 is discussed in this section with a focus on three front relative regions. The regions are considered are pre-surface-frontal cells (300 to 750 km ahead of the 700 hPa front), near-700hPa-frontal (-50 to 50 km from the 700 hPa front) and post-700hPa-frontal cells (-700 to -50 km from the 700 hPa front).

2.4.1 Pre-surface-frontal cells

Convection is much more frequent pre-surface-frontal than post-700hPa-frontal (Figure 1.5). A key result from this chapter is that while the environment quickly becomes unfavourable behind the surface front, this is not the case in the other direction ahead of the front. For example, 750 km ahead of the 700 hPa front (around 450 km ahead of the surface front) CAPE, surface dewpoints, solar radiation and Q-vector convergence are much higher than at the 700 hPa front (only 300 km behind the surface front). These results may explain why cell frequency remains high ahead of the front but sharply decreases towards the 700 hPa front. This is

despite the fact that CIN is also higher for pre-surface-frontal cells. The larger CAPE at pre-surface-frontal cells may contribute towards the larger fraction of 55 dBZ cores, mesocyclones and lightning (Figure 1.11).

2.4.2 Near-700hPa-frontal cells

Convection is least frequent surrounding the 700 hPa frontal line location (Figure 1.5). Divergence of the Q-vector at 850 hPa (and hence descending motion) is typically found at this region relative to the front, which would act to hinder the development of convective cells. This is supported by vertical velocity at 850 hPa which shows descending motion. Furthermore, the mean solar radiation and CAPE is lower than at other regions relative to the front. The higher wind shear (Figure 2.10) may also contribute towards the low cell frequency since deep moist convection may struggle to initiate when initial updraft width is low and shear is high (e.g. Morrison et al., 2022 and Peters et al., 2022).

2.4.3 Post-700hPa-frontal cells

Post-700hPa-frontal cells are associated with lower cell frequency that pre-surface-frontal cells (Figure 1.5) and almost always occur during the daytime (Figure 1.6). The strong diurnal cycle highlights the importance of solar heating on cell development. Generally Q-vectors at 850 hPa, 700 hPa and 500 hPa are divergent at this location relative to the front which would act to hinder the development of convective cells. However, at post-700hPa-frontal cell grid points weaker divergence or even convergence of the Q-vectors is observed. This may indicate that additional large-scale lift may be required as well as solar heating to initiate cells. The large-scale forcing for ascent could be linked to a post-frontal trough or may also act to destabilise upper-layers and hence increase CAPE. Another interesting finding is that vertical wind shear is comparable for post-700hPa-frontal and pre-surface-frontal cells on average. This means that a lack of wind shear is likely not the reason for the large differences in the fraction of cells with 55 dBZ cores and mesocyclones (Figure 1.11), but is rather due to thermodynamics. Post-700hPa-frontal cells form in environments with around 3 °C lower dewpoints (Figure 2.2) and 500 J kg $^{-1}$ of CAPE on average compared to pre-surface-frontal cells. Figure 2.3 showed that mid-level post-700hPa-frontal relative humidity at 700 hPa is generally low (35–50%). In theory, it should be easier for parcels to reach their LFC due to the lower CIN, though they may be more susceptible to entrainment due to the dry mid-troposphere and thus not be able to reach the threshold of a convective cell (46 dBZ). The mid-troposphere must therefore be moister than usual in order to allow convective cell development at this region relative to the front.

2.5 Conclusions

In this chapter, the thermodynamic environments and lifting mechanisms associated with warm-season cold-frontal convective cells are analysed by combining automatic front detection and cell detection methods from Chapter 1 and incorporating several ERA5 variables. Surface observation data from the DWD are also analysed (section 2.2). As in Chapter 1, a strong focus is placed on differences depending on the distance from the front. The primary findings of this chapter are highlighted below:

- Pre-700hPa-frontal cells form in environments with around 3 °C higher dewpoints (Figure 2.2) and 500 J kg $^{-1}$ higher CAPE (Figure 2.4) compared to post-700hPa-frontal cells. Both CAPE and dewpoints at convective grid points are significantly higher than non-cell regions at all distances relative to the front.

- Cell grid points between the 700 hPa front and surface front form in environments with the highest mid-tropospheric RH (Figure 2.3), which is linked to the frontal cloud band (Figures 2.8 and B2). Post-700hPa-frontal cell grid points had RH700 around 20–30% higher than non-cell regions, highlighting the importance of higher RH700 than usual to facilitate convective cell development and reduce the likelihood of entrainment.

- Pre-surface-frontal cells form in environments with the strongest Q-vector convergence (synoptic-scale lift) at 850 hPa and 700 hPa (Figure 2.6) and also have the strongest CIN (Figure 2.5). The stronger CIN may however contribute towards the build up of CAPE since convection will not be initiated prematurely (Figure 2.4).

- Near-700hPa-frontal cells form in the environments with the strongest Q-vector convergence at 500 hPa (Figure 2.6); around 3 times stronger than at non-cell regions, highlighting the importance of upper-level forcing. However, the role of this upper-level forcing on cell development is not entirely clear, but may be linked to a post-frontal trough or may also act to destabilise upper-layers and hence increase CAPE.

- Post-700hPa-frontal cell regions have mean divergence of the Q-vector at 850 and 700, whereas weaker divergence or convergence of the Q-vectors is found at cell grid points (Figure 2.6), indicating the importance of large-scale forcing for cell development. At 500 hPa, Q-vectors are mostly divergent at post-700hPa-frontal non-cell regions, but convergent at cell grid points.

- Excluding cells between the 700 hPa and surface front, solar radiation is lower at convective cell grid points compared to cell regions and non-cell regions (Figure 2.8). Station data shows that most cells develop with limited sunshine duration prior to development, particularly pre-700hPa-frontal cells (Figure 2.9). The cloudiness before cell detection indicates that solar heating may not be directly relevant for the majority of cell initiation. The cloud could be from pre-existing convection and therefore would highlight the importance of outflow boundaries on cell development.

These results advance understanding of the thermodynamic environments in which cold-frontal convective cells form and the importance of different lifting mechanisms on cell development depending on the region relative to the front. These results also help to explain several of the findings in the climatological analysis in Chapter 1.

Chapter 3

A statistical modelling approach to better understand the forcing mechanisms of cold-frontal convective cells in Germany

3.1 Introduction

The development of statistical models and the use of data driven approaches have become increasingly popular in recent times. In the field of meteorology and climate, the interest has extended beyond meteorologists, also to technology companies such as NVIDIA, Google (Deep Mind) and Huawei, who have developed global data-driven weather forecasting models (Pathak et al., 2022; Lam et al., 2023; Bi et al., 2023). The models can produce realistic thermodynamic and dynamic atmospheric fields forward in time, on par with those produced by numerical weather prediction (NWP) (Bouallègue et al., 2024). ECWMF have also developed an artificial intelligence weather prediction model (AIFS; ECMWF, 2024). Data driven approaches rely on large amounts of historical data for training. During the training, the goal is to establish relationships between the predictors and the target variable(s). The primary advantage compared to conventional NWP is that once the model is trained it is relatively inexpensive computationally to produce predictions from the model. Conventional NWP models often take on the order of 60 minutes (or more) to run, opposed to one order of magnitude quicker for data driven models. However, the models cited above were trained based on ERA5 reanalysis data, thus conventional numerical models are still required.

The aforementioned data driven weather prediction models leveraged neural networks, which are among the most computational expensive machine learning algorithms. They also have very limited interpretability when several hidden layers are used. This means it is difficult to determine how exactly the network arrives at its predictions. There are, however, less expensive and more simplistic statistical modelling techniques that have also produced reasonable predictions. For example, Rädler et al. (2019) and Battaglioli et al.

(2023) developed Generalised Additive Models (GAMs) to estimate the probability of lightning and hail occurrence based on ERA-Interim (Rädler et al., 2019) and ERA5 (Battaglioli et al., 2023) predictors. Based on the outputted modelled probabilities, historical trends of hail and lightning were reconstructed from past to present. In the United States, random forest, which uses an ensemble of decision trees to make predictions, have been used to produce probabilistic forecasts of severe convective storms (e.g. Mecikalski et al., 2021; Schumacher et al., 2021; Hill et al., 2023). Leinonen et al. (2022) used gradient tree boosting for nowcasting of thunderstorm hazards in the north-eastern United States. Aside from prediction, these models can also potentially be used to understand the importance of different mechanisms (features) on determining an outcome (e.g. likelihood of convection).

The aim of this chapter is to expand on the simple univariate statistical comparison made in Chapter 2 by training statistical models with various predictors (features) and assessing the model's ability to predict convective cell occurrence. The statistical models can unravel the complex relationships that may exist between different predictors (e.g. CAPE, dewpoints, solar heating, vertical velocity etc) and the target variable (convective cell occurrence). Such models can also assign weights to each variable depending on its importance to the model. Since the links between different convective ingredients, especially the lifting mechanisms, are likely to be complex and non-linear, statistical modelling is a viable approach to the problem.

The lifting mechanisms considered in Chapter 2 (quasi-geostrophic forcing for ascent and solar heating) will be extended to include orographic influences. Orographic influences are considered as climatologies of convective storms often show higher frequency in proximity to mountainous regions or areas of higher elevation (e.g., Piper and Kunz, 2017; Taszarek et al., 2019). In Chapter 1, pre-700hPa-frontal cells and non-cold-frontal cells were shown to be most frequent in southern Germany (Figure 1.8). The ways in which orography favours the development of convection can vary (Kirshbaum et al., 2018). Flow approaching slopes is forced to rise thus provides lift. This is usually referred to as forced ascent. On the other hand, flow around terrain with valleys may split and converge at the end of the valley providing lift. Convergence zones due to flow splitting and channelling were observed during the Convective and Orographically-induced Precipitation Study (COPS) in south-western Germany/eastern France (Wulfmeyer et al., 2011).

The analysis in this chapter will be split into two parts. First, models with a low number of variables (4–5 variables) will be developed where variables have limited multicollinearity (section 3.3). The lack of multicollinearity between variables ensures a robust assessment of feature importance. The feature importance will be assessed by training and testing a model with and without features at different regions relative to the front. For this analysis, 4 different predictive modelling algorithms will be tested to see if the importance of features and model skill are consistent across each algorithm. Second, a larger number of variables (60) will be considered to produce a best possible prediction model for cold-frontal convective cells (section 3.4). The model's ability to generalise to new unseen testing data will be assessed by reproducing some of the results from the climatology in Chapter 1. The strengths and limitations of the models will also be discussed. Section 3.2 will introduce the data, predictive models and validation techniques used in this chapter. The two approaches (section 3.3 and section 3.4) will begin with the specific model setup to that section and explain how features are selected.

The primary research questions addressed in this chapter are as follows:

Q1) Which predictors add the most skill depending on the cell-front distance?
Q2) Can statistical models be used to robustly assess feature importance and advance physical understanding?
Q3) How well can the spatial and temporal climatology of cold-frontal cells from Chapter 1 be reproduced using predictive modelling techniques?

3.2 Data, models and validation

Probabilistic models are developed using the dataset constructed in Chapter 2 with ERA5 grid points and cell-front distances. The list of variables is also extended from those used in Chapter 2 (Table 2.1) to include additional ERA5 variables and also non-ERA5 variables. Convective cell grid points are labelled 1 indicating a true event (convective cell occurrence). Cell regions and non-cell regions are labelled 0 indicating a false event (no convective cell occurrence). Unlike Chapter 2, where the full German radar domain was used (Figure 1.2), only the domain of Germany was used (Figure 3.1). The German radar domain extends partly into the North Sea and Baltic Sea where the dynamics of convective cell development may differ compared to over land. Regions outside of Germany are also towards the edge of the radar domain, thus subsetting Germany reduces the number of potential missed detections. This reduces the number of potentially erroneously labelled non-convective cell grid points when there was indeed a convective cell. Since 78% of convective cells during April–September occur during the summer months (KONRAD; section 1.2.2), only June, July and August are used. Selecting only the summer months also reduces the class imbalance significantly, that is, the ratio of true and false events. In this case, this is the percentage of all ERA5 grid points labelled as convective cells compared to non-convective cells. Class imbalance is a common issue in machine learning since a model could end up being biased towards the majority class (e.g. Krawczyk, 2016).

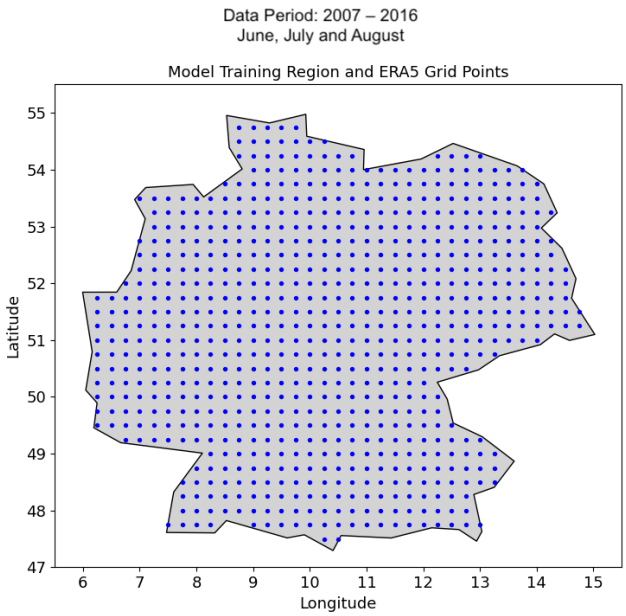


Figure 3.1: Data time period and ERA5 grid points used to train and test models. The ERA5 grid points and data time period are a subset of what was used in Chapters 1 and 2.

3.2.1 Predictive Models

A total of 4 predictive models are used in this chapter. The models were selected in a way that they use distinct methods to produce predictions. They are trained and tested based on predictor variables with the target variable being convective cell occurrence at a given grid point. The models are introduced below.

Generalised Linear Models

Generalised Linear Models (GLMs) are a class of linear models that are flexible in modelling various types of response variables, including regression or classification tasks (Nelder and Wedderburn, 1972). Unlike traditional linear regression, GLMs allow for non-normal distributions of the response variable. The model assumes a linear relationship between the predictors $(X_1, X_2, \ldots, X_p)$ and the linear predictor.

$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p \tag{3.1}$$

where η is the linear predictor and $\beta_0, \beta_1, \ldots, \beta_p$ are coefficients estimated during model fitting.

The linear predictor is transformed via a link function:

$$g(\mu) = \eta \tag{3.2}$$

In this case, the logit link function is used:

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right) \tag{3.3}$$

where $g(\cdot)$ is the link function and μ represents the expected value of the response variable.

The logit function can be inverted to map real numbers $(-\infty,\infty)$ to probabilities (0,1).

$$\mu = \frac{1}{1 + e^{-\eta}} \tag{3.4}$$

Generalised Additive Models

Generalised Additive Models (GAMs) are a class of statistical models used for regression or classification tasks (Hastie and Tibshirani, 1986). They are an extension of linear models, allowing for non-linear relationships between predictors and the target variable. GAMs model the relationship between predictors $(X_1, X_2, \ldots, X_p)$ and the response variable (Y) as a sum of smooth functions (f_i) of the predictors, allowing complex and non-linear relationships to be established:

$$Y = \beta_0 + f_1(X_1) + f_2(X_2) + \ldots + f_p(X_p) + \epsilon \tag{3.5}$$
$$\text{where } \beta_0 \text{ is the intercept and } \epsilon \text{ is the error term}$$

Such non-linear relationships cannot be established with a GLM.

K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a supervised learning algorithm used for classification and regression tasks (Cover and Hart, 1967). For classification tasks, the prediction for a new data point is based on the majority class in the feature space. The value of k must be predefined and refers to the number of nearest neighbours to consider. The proximity between data points is measured using the Euclidean distance. For example, assuming we have a dataset (Table 3.1) with three predictors: X_1, X_2, and X_3, and we want to predict the class of a new data point (Y) .

Table 3.1: Sample dataset with 3 predictor variables (X_1, X_2, and X_3) and a binary target variable (Y).

Row	X_1	X_2	X_3	Y
1	2	3	1	0
2	1	5	2	1
3	4	2	3	0
4	3	4	4	1

If want to predict the class of a new data point with $X_1 = 2$, $X_2 = 4$, and $X_3 = 3$, we calculate the distances between the new data point and each existing point in the dataset. The Euclidean distance is used:

$$\text{Row 1 Distance} = \sqrt{(2-2)^2 + (3-4)^2 + (1-3)^2} = \sqrt{5}$$
$$\text{Row 2 Distance} = \sqrt{(1-2)^2 + (5-4)^2 + (2-3)^2} = \sqrt{3}$$
$$\text{Row 3 Distance} = \sqrt{(4-2)^2 + (2-4)^2 + (3-3)^2} = \sqrt{8}$$
$$\text{Row 4 Distance} = \sqrt{(3-2)^2 + (4-4)^2 + (4-3)^2} = \sqrt{2}$$

If $k = 3$, the three closest points are chosen, which are rows 4, 2, and 1. The class of the new data point is based on the majority class of the nearest neighbours or can be expressed as a probability. Rows 4 and 2 have the class label 1 and row 1 has the class label 0. Therefore, the prediction for the new data point would be class 1, or this could be expressed as a probability of $\frac{2}{3}$ since two thirds of the k-nearest neighbours had the class label 1.

XGBoost

XGBoost (Extreme Gradient Boosting) is an optimised implementation of gradient boosting (Friedman, 2001), a machine learning technique that builds a predictive model by combining the predictions of multiple decision trees (see example in Figure 3.2). Each decision tree can be built with a different number of predictors rather than using all predictors. XGBoost iteratively improves the model's overall performance by correcting the errors made by the previous trees. The number of iterations and trees constructed is usually referred to as "boosting rounds". The strength of gradient boosting lies in the ensemble of all trees rather than any individual tree.

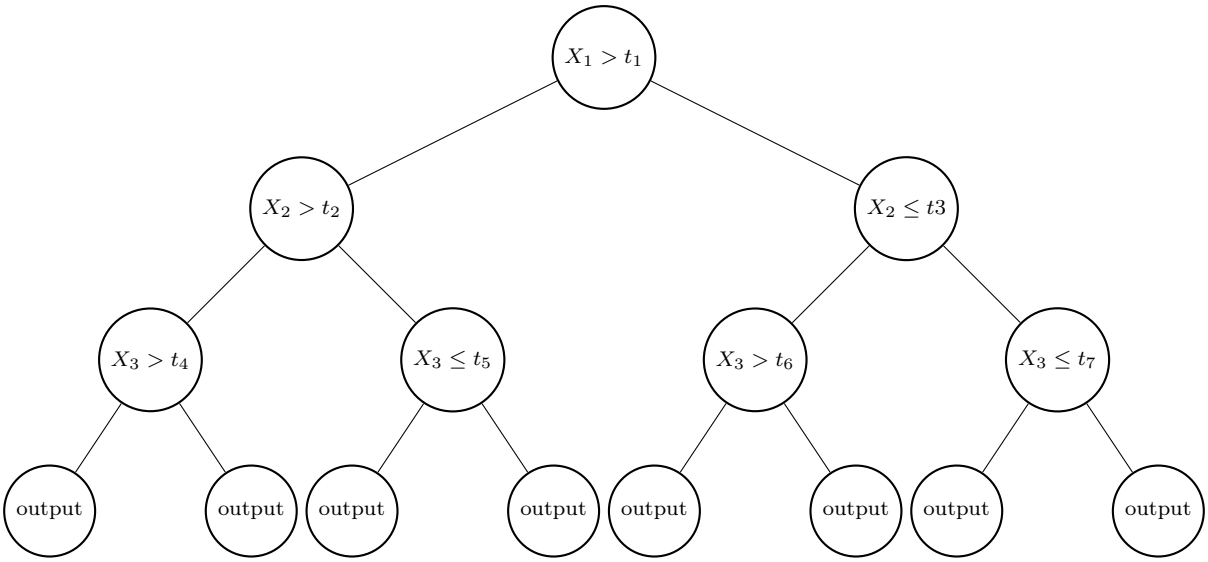


Figure 3.2: An example decision tree with conditions based on features X_1, X_2, and X_3, where t_n are threshold values. The final prediction is obtained by summing the output scores from all trees and then applying a link function to get the raw probability between 0 and 1.

3.2.2 Validation

The models are validated using the ROC AUC (receiver operating characteristic area under the curve; Wilks, 2019). The metric is commonly used for classification problems. The ROC AUC is calculated by deriving the true positive rate (Equation 3.6) and false positive rate (Equation 3.7) at different classification thresholds. The confusion matrix is shown in Table 3.2. To derive the ROC AUC, the true positive rate is plotted against the false positive rate at different classification thresholds and then the area under the curve is calculated. An advantage of the ROC AUC is that the decision threshold does not need to be subjectively defined.

Table 3.2: Confusion matrix showing true positive (TP), false positive (FP), false negative (FN), and true negative (TN) labels.

	Actual Label True	Actual Label False
Predicted Label True	TP	FP
Predicted Label False	FN	TN

The True Positive Rate (TPR), also known as Sensitivity or Recall, measures the proportion of actual positives that are correctly identified:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.6}$$

The False Positive Rate (FPR) measures the proportion of actual negatives that are incorrectly identified as positives:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{3.7}$$

The ROC AUC shows overall how well a model can discriminate between positive or negative classes across different classification thresholds. A ROC AUC value of 1.0 indicates a perfect model and a value of 0.5 indicates a model no better than chance. The threshold at which the model is no better chance is usually called the baseline. The baseline for the ROC AUC remains fixed at 0.5 regardless of the class imbalance.

3.3 Feature importance in the cell-front relative framework

The goal of this section is to further understand the importance of different features depending on the region relative to the front by using statistical models. The features chosen are linked to the three primary ingredients needed for deep moist convection: moisture, lift and instability.

3.3.1 Model Setup

XGBoost, KNN, GAM and GLM (see section 3.2.1) are trained and tested separately at 200 km front relative regions: -700 to -500 km, -500 to -300 km, -300 to -100 km, -100 to 100 km, 100 to 300 km, 300 to 500 km and 500 to 700 km. A 200 km interval is chosen to allow a sufficient number of grid points in each sample. The range is reduced from 750 km to 700 km since 1500 km is not divisible by 200. The number of grid points and percentage of true events for each front relative region are summarised in Table C1. In this section, 80% of grid points are used for training and 20% for testing, which is a typical split used in machine learning. The feature importance of the different predictors is assessed by training models with and without features and comparing the ROC AUC values (see section 3.2.2). As a robustness measure of the results, the skill is tested with all data points being used as test data at least once using k-folds cross validation (i.e. k=1 through to k=5). The reader is referred to Appendix C for further details on k-folds cross validation. The domain and time period are as described in Figure 3.1.

Hyperparameters are settings that can be adjusted to influence how a model learns and makes its predictions. The number of estimators is a key hyperparameter for KNN and XGBoost since it can have a large effect

on model performance. For KNN, the number of estimators represents the number of nearest neighbours to consider. The final prediction is based on the number of neighbours predicting a certain class. Initial testing showed the default value of 5 neighbours gave poor performance for KNN. For XGBoost, the number of estimators represents the number of boosting rounds or iterations. Trees are built sequentially, where each subsequent tree corrects the errors made by the previous ones. Adding additional trees may improve model performance up to a certain point, but also increases the computational cost. Therefore, a compromise must be made between model skill and computational cost. The model skill was first tested for the 300 to 500 km front relative region depending on the number of estimators (Figure C2). Considering the highest ROC AUC and computational cost, 150 estimators (boosting rounds) were chosen for XGBoost and 150 estimators (nearest neighbours) were chosen for KNN. Figure C2 was reproduced using the -500 to -300 km front relative region and no significant differences were observed (not shown).

3.3.2 Feature Selection

To ensure a robust assessment of feature importance depending on the cell-front relative region, it was essential to select features with limited multicollinearity. The features were selected using an ingredients based approach, that is moisture, lift and instability. Only a single moisture and instability term were selected as using multiple moisture and instability terms were likely to have to strong correlations. CAPE is used as the instability term as it is the most commonly used measure of instability. A transformation of CAPE to the power of one quarter ($\text{CAPE}^{0.25}$) is used to damp CAPE for higher values. Indeed, Westermayer et al. (2017) showed that for CAPE in the range 800–2800 J kg^{-1} the relative frequency of lightning is largely constant. Total Column Water Vapour (TCWV) is used as the moisture variable since it is a measure of the total moisture availability throughout the entire column, rather than on a specific level. For example, using a surface-based moisture term (e.g. surface dewpoints) may not be effective as convection can be elevated (e.g. Corfidi et al., 2008). The lifting features considered are proxies for solar heating, orographic influences and large-scale lifting. The solar heating is considered using a cosine transformation of the hour of the day (Equation 3.8). Such a transformation is useful since if time is represented using just a numerical value like hours of the day (0 to 23), the model may interpret the time as a continuous, linear sequence. This can lead to false assumptions e.g. 23:00 and 00:00 UTC being far apart when in reality they are close together.

$$\cos_{\text{time}}(h) = 0.5 \cdot \cos\left(\frac{\pi}{12} \cdot (h - 16)\right) + 0.5, \quad h \in [0, 23] \tag{3.8}$$

where h represents the hour (UTC).
$\cos_{time}$ as a function of h is shown in Figure C3.

Orographic influences are considered by using the elevation above mean sea level (ELEV) as a feature. In principle, the model can also learn to associate lower elevation with higher cell probability, which may be a relevant factor for convective initiation due to sea-breezes. The large-scale lifting feature is the Q-vector convergence maximum between 900–500 hPa (QVEC_{max}). Initial testing with univariate GLMs showed that this parameter added more skill than Q-vector convergence on individual pressure levels. The correlations between the five features are shown in Figure 3.3 for each front relative region. Some correlation between the instability and moisture terms was unavoidable due to the links between moisture and CAPE (Markowski and Richardson, 2010; their Figure 7.9b). A negative correlation between TCWV and ELEV is expected

since most moisture is concentrated in the lower troposphere near mean sea level. The individual lifting mechanisms have limited correlation, which allows a robust assessment of the importance of different lifting mechanisms on convective cell development.

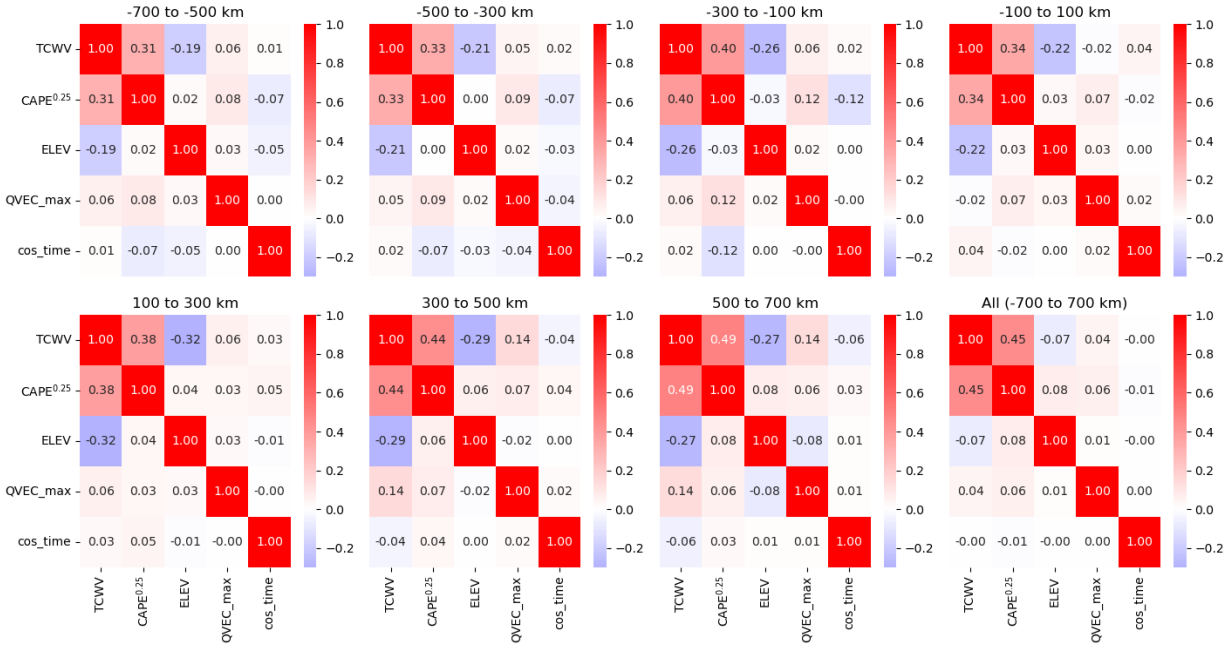


Figure 3.3: Correlation matrix for the 5 features used in section 3.3 depending on the cell-front relative region and all data (bottom right).

3.3.3 5 feature model

The skill of XGBoost, KNN, GAM and GLM is assessed depending on the cell-front relative region first using all five variables (Figure 3.4). The mean ROC AUC of all 5 folds is shown in the top-left panel. All four models have the highest predictive power in the region -300 to -100 km and then show decreasing predictive skill with the minimum skill in the region 100 to 500 km ahead of the 700 hPa front. The minimum skill corresponds to the region surrounding the surface front (300 km ahead of the 700 hPa front; section 1.3.1). The predictive skill then increases again for all models ahead of the surface front. XGBoost generally shows the highest predict skill across the front, however near the surface front KNN has slightly higher skill. GLM shows the lowest skill at all cell-front relative regions, which is not the most surprising result since GLM uses the most simplistic algorithm and cannot learn complex non-linear relationships. The results are broadly consistent across all folds, indicating the result is not sensitive to a certain subsample of testing data.

The lower model skill (consistent for all predictive model types) near the surface front could be linked to organised convection where outflow boundaries favour the development of cells in the vicinity. Such interactions between cells and organised convection are not explicitly considered in the model. Figure 1.11 showed that pre-surface-frontal cells are associated with the largest number of cells within 100 km, this included the region 500–700 km. The region 500–700 km shows higher skill than 300–500 km. Since cells closer to the surface front and ahead of the surface front have a comparable number of cells in proximity,

the difference in skill across the front cannot be only explained by convective organisation. A multitude of factors are likely at play, possibly including differences in how well ERA5 variables are resolved depending on the cell-front distance. The lowest skill near the surface front and highest skill in the region -300 to -100 km is not unique to the 5 variables used here. The result was also identified during initial tests with GLMs using 9 variables (Figure C4).

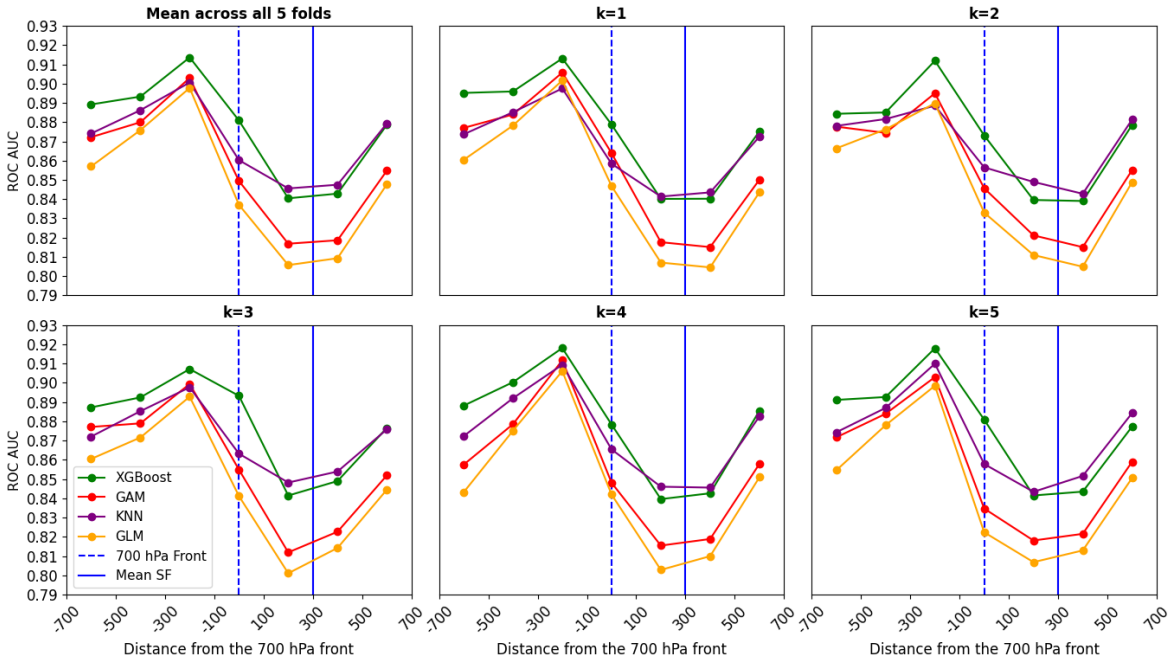


Figure 3.4: ROC AUC with variables TCWV, CAPE$^{0.25}$, cos$_{time}$, QVEC$_{max}$ and ELEV depending on the distance from the front for k-folds 1–5 and mean of all folds (top-left). The solid and dashed vertical blue lines represent 700 hPa front and mean surface front (SF) location, respectively. The model type is shown in the legend.

3.3.4 CAPE and TCWV feature importance

To understand the importance of each feature on the models' predictive skill, the models are retrained and retested with each feature removed. A comparison is made between the ROC AUC with and without each feature. While there are slight differences in the magnitude of the change depending on the k-fold, there is a general trend for the total column water vapour to carry increasing feature importance going from further distances behind the 700 hPa front to further distances ahead of the 700 hPa front (Figure 3.5).

On the other hand, CAPE shows a generally opposite trend with higher importance behind the front compared to ahead of the front (Figure 3.6). However, there is a secondary peak between 100 to 300 km according to the mean across all folds. CAPE has the least importance in the region 500 to 700 km. CAPE may carry less importance ahead of the surface front since this is where CAPE is climatologically highest (Figure 1.3). On the other hand, since there is a 0.45 correlation between CAPE$^{0.25}$ and TCWV in the 500 to 700 km region (Figure 3.3), it could be that the model is simply relying more on the TCWV predictor. When the predictor CAPE was used without transformation to the power of one quarter lower feature importance was found especially for GLM.

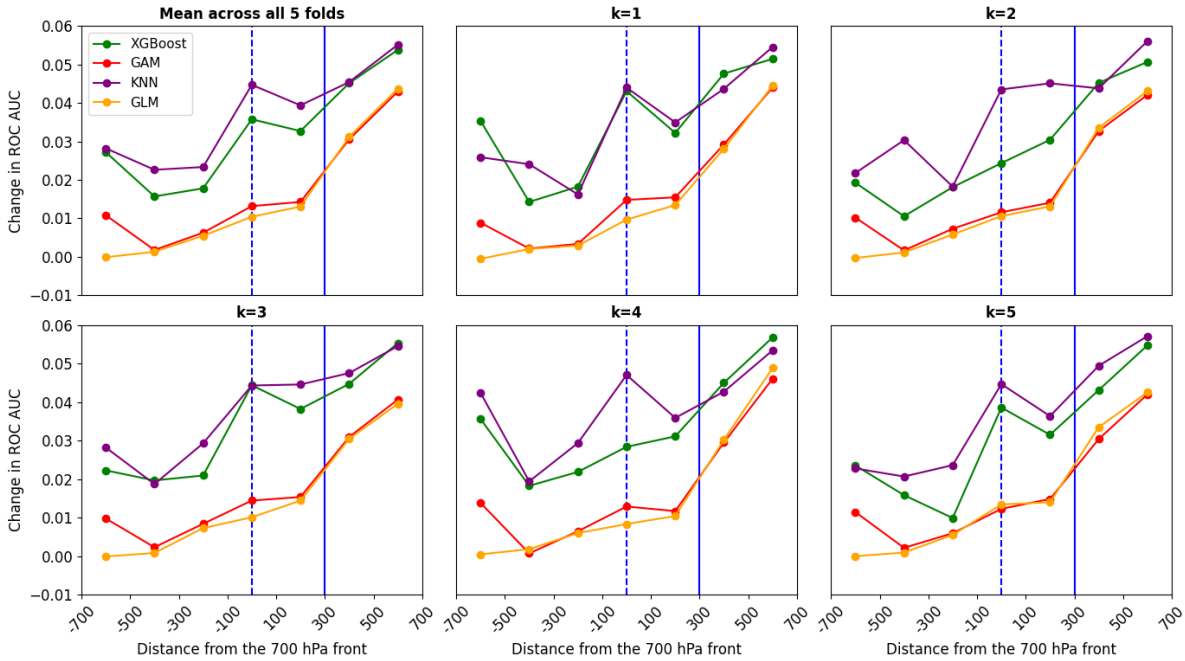


Figure 3.5: Feature importance of TCWV depending on the distance from the front for all 5 k-folds and mean of all k-folds (top-left). The dashed and solid vertical blue lines represent 700 hPa front and mean surface front (SF) location, respectively. Note the differing y-axis ranges for each feature importance figure (Figures 3.5–3.9).

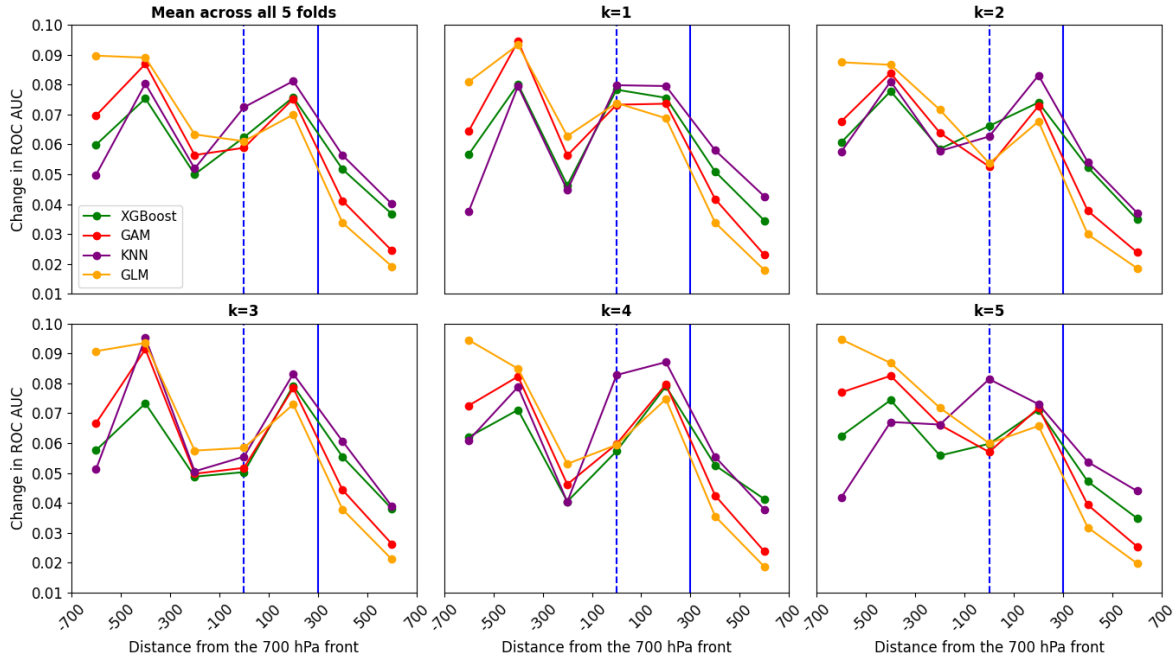


Figure 3.6: As Figure 3.5 but for CAPE$^{0.25}$. Note the differing y-axis ranges for each feature importance figure (Figures 3.5–3.9).

3.3.5 Solar heating feature importance

The $\cos_{time}$ feature is a proxy for solar heating (Equation 3.8 and Figure C3). The feature shows decreasing feature importance going from further distances behind the 700 hPa to further distances ahead of the 700 hPa front. This is very consistent with Figure 1.6 which shows post-700hPa-frontal cells generally occur during the daytime whereas pre-700hPa-frontal cells have a weakened diurnal cycle. In the region -700 to -100 km, the $\cos_{time}$ feature adds comparable skill regardless of which predictive model is used. On the other hand, the feature adds more skill in the region -100 to 700 km for XGBoost and KNN.

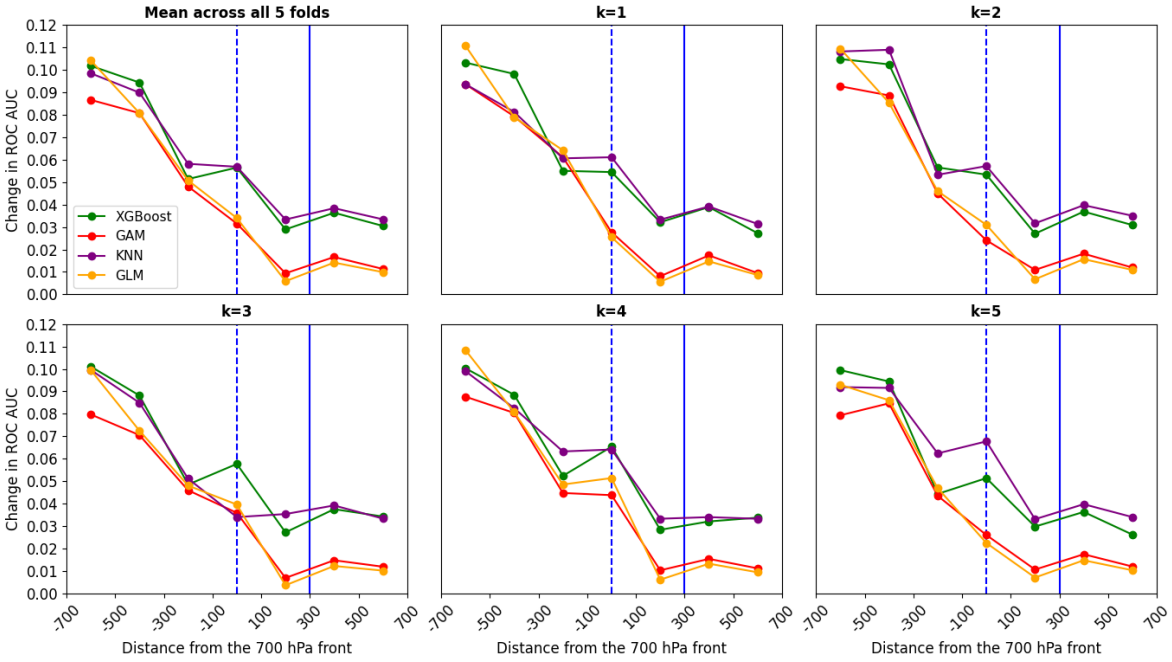


Figure 3.7: As Figure 3.5 but for the $\cos_{time}$ feature. Note the differing y-axis ranges for each feature importance figure (Figures 3.5–3.9).

3.3.6 Large-scale lifting feature importance

The large-scale lifting feature is most important for all four model types in the cell-front relative region -100 to 100 km with decreasing importance away from this region. The highest importance at -100 to 100 km is consistent with the result in Figure 2.6. A relatively symmetrical relationship is observed around the maximum. The large-scale lifting feature adds more skill for XGBoost and KNN models compared to GAM and GLM. Whether the large-scale lift is relevant for destabilisation in the upper-level levels (Trapp, 2013, his Figure 5.2) or triggering convection remains unclear.

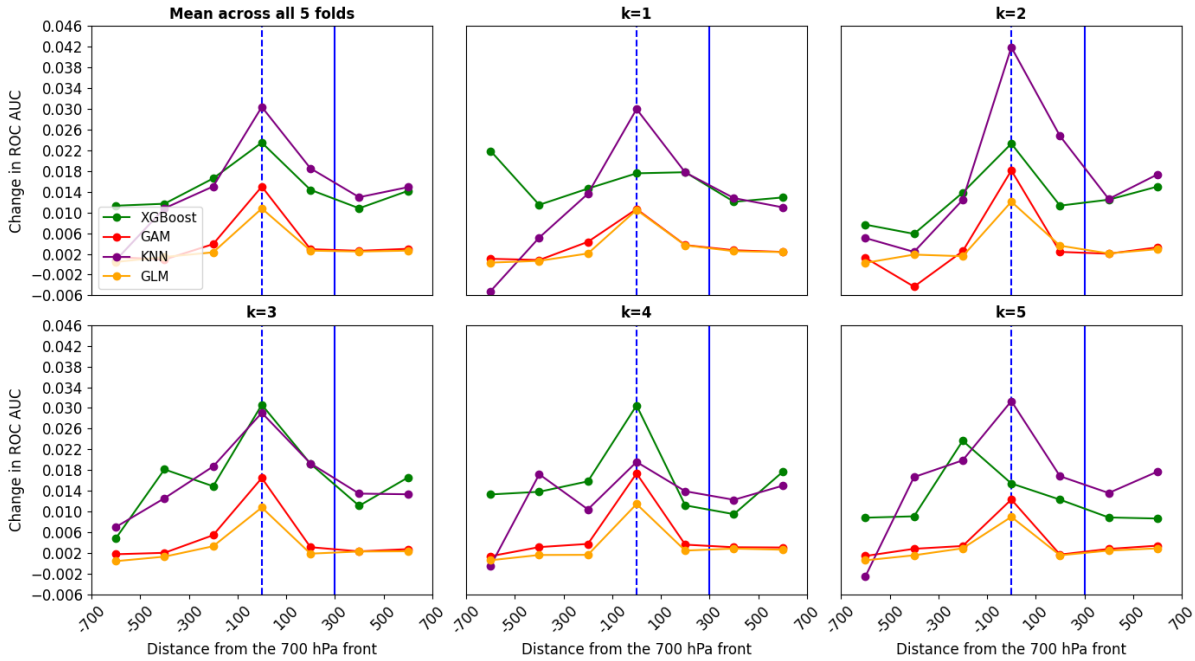


Figure 3.8: As Figure 3.5 but for the large-scale lifting feature. Note the differing y-axis ranges for each feature importance figure (Figures 3.5–3.9).

3.3.7 Elevation feature importance

The elevation feature is the height above mean sea level. In principal, the predictive models can learn to associate lower or higher elevation with higher probability of convective cells. As shown in Figure 1.8, pre-700hPa-frontal cells have the highest frequency in southern Germany where the elevation is higher and post-700hPa-frontal cells in north-western Germany (particularly on the coast), where the elevation is low. In contrast to other features, the elevation feature has the most importance at different cell-front relative regions depending on the predictive model used. GLM and GAM show the highest importance in the region furthest ahead of the front (500 to 700 km), albeit with a change in AUC under 0.01. XGBoost and KNN show that the elevation feature has the most importance at the 700 hPa front (-100 to 100 km). Considering the nature of the algorithms, the result may highlight that elevation alone does not increase the likelihood of convective cells and other conditions are required (e.g large-scale lifting and solar heating). An increase in the elevation feature importance is observed for KNN and XGBoost from -300 km to -700 km, which may be related to coastal areas (low elevation) being more favourable for convective cell development (again conditional on solar heating). The large difference in the skill added (particularly near-700hPa-frontal) also emphasises the importance of testing different predictive modelling algorithms since one approach may be able to better capture the patterns between a feature (and its interactions with other features) and the target variable.

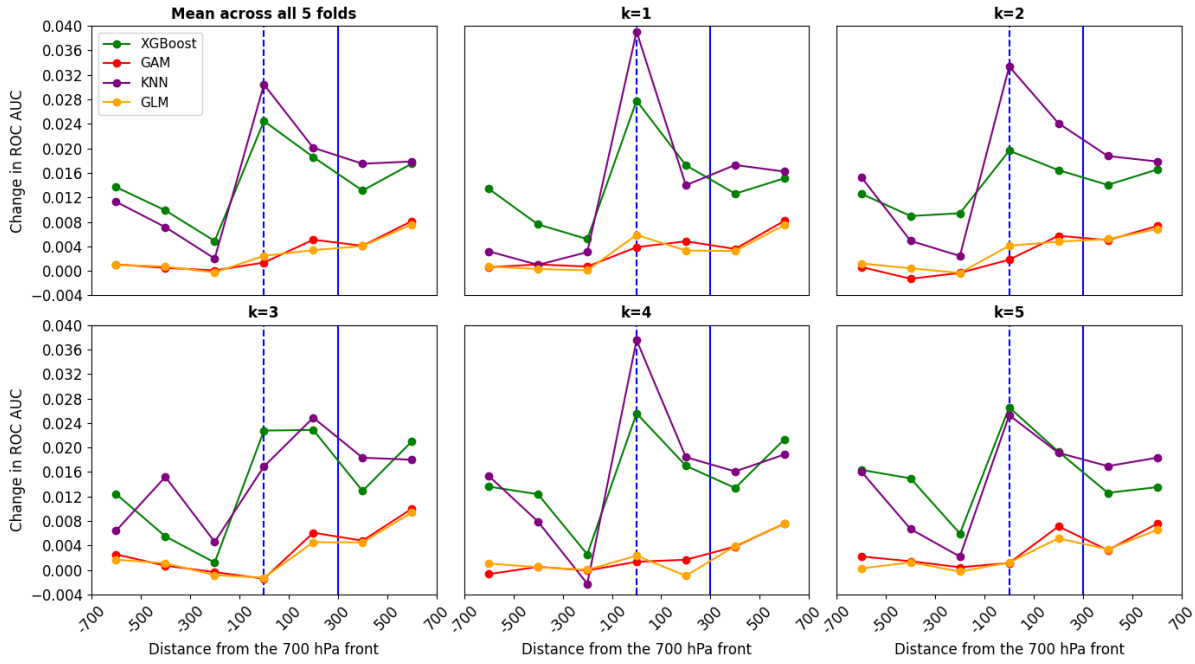


Figure 3.9: As Figure 3.5 but for the elevation feature. Note the differing y-axis ranges for each feature importance figure (Figures 3.5–3.9).

The importance of 5 features depending on the front relative region and statistical model used has been evaluated. While thus far the discussion in this section has been based on the importance of each feature relative to the front, it is worth noting that some features overall carry more importance than other features. Using XGBoost and CAPE as an example, the feature adds 0.04–0.10 AUC depending the distance from the front. The skill added is lowest in the region 500–700 km (0.04 change in AUC), but this still exceeds the skill added by the large-scale lifting term and elevation term at all front relative regions. However, this result should not necessarily be interpreted as CAPE being more important for convective cell development, since it has been well established that lift and moisture are also required for convective initiation (Doswell et al., 1996). The higher importance of CAPE compared to the large-scale lifting and elevation could be explained by the fact that in the absence of large-scale lifting and topography convection can still occur due to other possible sources of lift (e.g. outflow boundaries and localised solar heating). Furthermore, the large-scale lifting term does not account for mesoscale lifting. On the other hand, without instability (CAPE) a necessary ingredient for convection is missing. The convective cell definition of 46 dBZ may also contribute towards the higher feature importance of CAPE, since moderately strong updrafts are likely required to reach such a cell intensity.

3.4 A unified cold-frontal cell model

The goal of this section is to develop a best possible predictive model and test its ability to reproduce parts of the climatology presented in Chapter 1. The focus is not specifically on feature importance so predictors with high correlations are also permitted. Nevertheless, which predictors the model relies on the most to make predictions will be discussed.

3.4.1 Model Setup

A larger number of predictors are considered in this section (Table C2). The model's ability to generalise to new unseen testing data will be assessed again using the ROC AUC but also in more detail by reproducing some of the results from the climatology in Chapter 1. The algorithm XGBoost is selected due to a trade-off between computational cost (Figure C2) and skill (Figure 3.4). In machine learning, a 70/30 or 80/20 training/test split is typically used (as applied in section 3.3). However, if 7 years were selected as training this would only leave 3 years of training data. Since we want to evaluate the model's ability to reproduce the climatology, 5 years are used for training and testing. Odd years are used for training: 2007, 2009, 2011, 2013, 2015 and even years are used for testing: 2008, 2010, 2012, 2014 and 2016. It is worth noting though there is not a universal approach when selecting the training and testing split since it depends on several factors such as the dataset size, nature of the data and the application. The model in this section is trained using ERA5 grid points within 750 km of the 700 hPa front. A single model is trained in this section, unlike section 3.3 where a new model was trained at each 200 km front relative region. This is because the focus is on developing a generalised cold-frontal convective cell model. The domain and time period are as described in Figure 3.1.

3.4.2 Feature Selection

A total of 60 predictors (featuresare considered in this section. These include most variables used in Chapter 2 as well as several additional variables. Additional variables linked to atmospheric instability are included such as lapse rates and maximum theoretical updraft velocity ($\sqrt{2 \times \text{CAPE}}$). The u and v components of wind and wind speed at different pressure levels are used since wind direction and speed influence moisture advection. The elevation above mean sea level is used as a proxy for topographic influences on convection. Since convection may occur at lower elevations next to higher elevations (e.g. downslope) the distance from the nearest elevation exceeding 200, 500 and 800 metres were considered (ELEV200, ELEV500, ELEV800). The full list of predictors is shown in the appendix (Table C2).

Training XGBoost with 60 features may add noise to the model and using a larger number of features comes with increasing computational cost. The optimal number of features was selected using Recursive Feature Extraction with Cross-Validation (RFECV; Kuhn and Johnson, 2013). The model is initially trained with all features (60 in this case) and the least important features are removed until the minimum number of features is reached. RFECV selects the number of features that maximise the model's performance based on the cross-validation result, that is to say, using different training and testing datasets. The minimum number of features is selected as 6 due to previous domain knowledge that multiple factors are relevant for the development of convective cells. Indeed, the model skill reduces steeply after 10 or lower features (Figure C5) with the optimal number of features as 12. The best 12 features and their feature importance are shown in Table 3.3 and Figure 3.10. The feature importance is derived using the built-in functionality of XGBoost which is based on how much each feature contributes to reducing the model error during the splitting of tree nodes. The ROC AUC depending on the selected number of features is shown in Figure C5.

Table 3.3: Summary of the best 12 features categorised into thermodynamic, kinematic and other. Predictor full names are shown in Table C2.

Thermodynamic	**Kinematic**	**Other**
$T_{2\text{m}}$ $T_{500\text{ hPa}}$	$w_{850\text{ hPa}}$	$\cos_{\text{time}}$
$LR_{700-500\text{ hPa}}$ CAPE	$w_{700\text{ hPa}}$	HIHC
$RH_{850-500\text{ hPa}}$ $RH_{700\text{ hPa}}$	$w_{500\text{ hPa}}$	
TCWV		

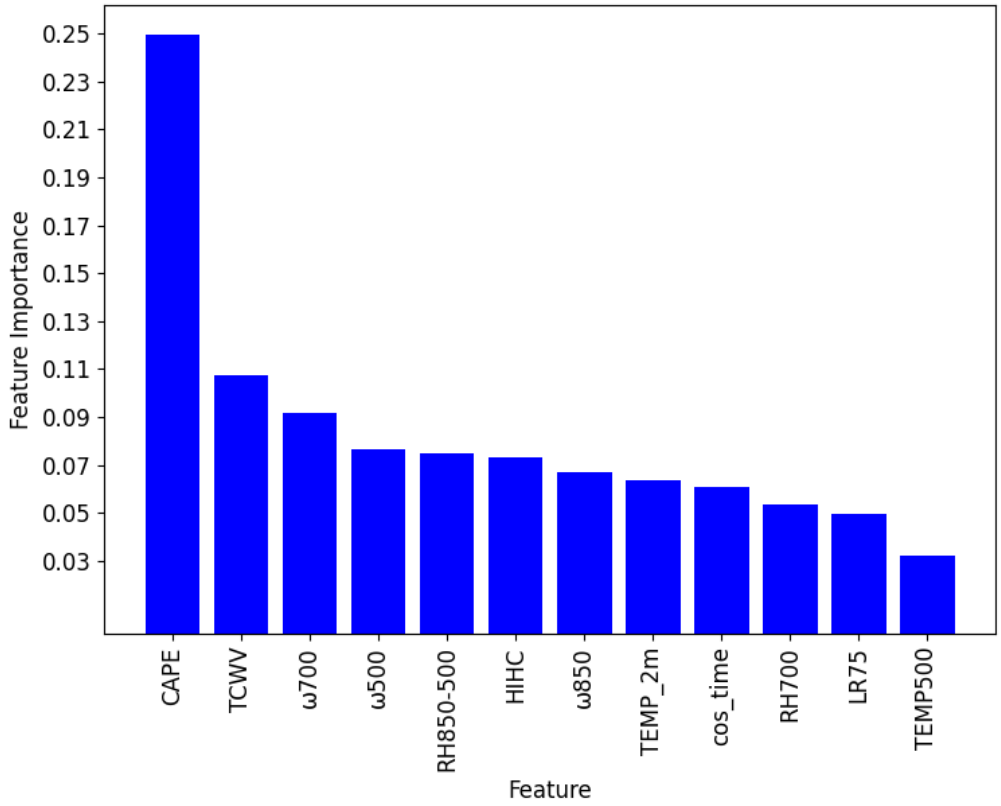


Figure 3.10: Summary of the best 12 features and feature importance to the model. Higher numbers indicate higher feature importance. Predictors are ranked from most important to least important from left to right. The feature importance values sum to one.

The best 12 predictors include the three ingredients for deep moist convection. The instability is represented by CAPE and also by LR75. Steeper lapse rates between 700 to 500 hPa generally allow stronger buoyancy in the mid-levels. Colder air aloft also increases instability (TEMP500). LR75 and TEMP500 were however the least important out of the 12 predictors (Figure 3.10), possibly owing to their partial collinearity with CAPE. Total column water vapour (TCWV) is a measure of the integrated water vapour content of the atmosphere. While specifically lower level moisture is generally thought to be most relevant for convective initiation, the majority of moisture is concentrated in the lower levels of the atmosphere. Given the result that TCWV is in the best 12 predictors and moisture at specific levels are absent, correlation coefficients between TCWV and Td_{2m}, QHUM900, QHUM850 are derived. The correlations are 0.79, 0.77 and 0.79, respectively. The correlation between TCWV and the mean of QHUM900, QHUM850, QHUM700 is 0.89. The high correlation indicates that the TCWV feature does contain information about the low-level moisture

availability. Thus, using a specific level such as QHUM900 or surface dewpoints compared to a combined parameter across a column may be less effective as parcels can depart from different levels depending on the situation. Mid-level relative humidity (RH700 and RH850–500) is also in the best 12 variable combination. As discussed in detail in section 2.3.1, mid-level relative humidity is relevant due to the possibility of entrainment. Drier environmental air may entrain into the updraft thus hindering the development of convective cells. Vertical velocity at all three levels are present in the best 12 variable combination. A clear distinction between vertical velocity at non-cell regions and cell grid points was observed in Figure 2.7. However, it is still not clear whether a signal for vertical velocity would be present in ERA5 when convective precipitation is resolved. Therefore, one cannot assume the result indicates the importance of large-scale lifting, although it may play some role. Higher surface temperatures ($TEMP_{2m}$) reduce CIN and increase CAPE (Trapp, 2013). If the higher temperatures are a result of stronger solar heating this also favours lifting of air parcels near the surface. High cloud cover is also found to be among the best predictors. The presence of high cirrus clouds may indicate existing deep moist convection and enhanced high cloud cover. Convection often initiates in the vicinity of pre-existing convection due to gust fronts (e.g. Lima and Wilson, 2008; Hirt et al., 2020). Asides from being linked to stronger solar heating, the time of day ($\cos_{time}$) feature may carry importance since different predictors may be more relevant depending on the time of day. The decision tree and ensemble nature of XGBoost allows the development of trees with different conditions. For example, one tree may be constructed with high probability of convection being dependant on $\cos_{time}$ being in the nighttime range (i.e. $\cos_{time}$=0) and other features being above a certain threshold. Another tree may be developed with a high probability of cells when $\cos_{time}$ is in the daytime range ($\cos_{time}$=1) and other features being above a certain threshold. Interpretation of individual trees are not explored in this thesis. Such a task is also not trivial since the XGBoost model is initially developed with 100 trees and with a maximum tree depth of 6.

An important caveat to note is that the absence of a parameter in the 12 feature model does not indicate unimportance of that parameter for convective cell development. For example, some wind shear is generally present across cold fronts so may carry less importance than if the model were trained on all cases including non-cold-frontal cells. Furthermore, since the model is trained on all grid points within 750 km of the 700 hPa front rather than a cell-front relative region (as done in section 3.3), the focus is on generalising to cold-frontal cells at all cell-front relative regions. An additional note to consider is that correlated features may appear as less important if other strong predictors are present. This is highlighted in the feature ranking for the RFECV (shown in Figure C6). Among the worst predictors is w_{max}, however, the best performing predictor in the best 12 predictor model is CAPE (Figure 3.10). CAPE and w_{max} have a Pearson correlation of 0.92. Therefore, the predictor w_{max} does not add any unique information since it is just a transformation of CAPE. If CAPE were removed from the model, w_{max} would likely become one of the most important predictors. This also highlights the importance of using features with limited correlation to robustly assess feature importance (as done in section 3.3). Figure C6 also shows that the elevation terms were among the worst predictors and removed first during the recursive feature extraction process. Since the importance of elevation seems to vary in importance across the front (Figure 3.9), it may not work as well on a generalised cold-frontal cell model. CAPE did also vary in importance across the front, with the least importance to the XGBoost model at 500–700 km showing a change in AUC of 0.04 (Figure 3.6). Nevertheless, this is still a larger change than where the elevation was most important in the front relative framework showing a change in AUC of around 0.024 at -100 to 100 km (Figure 3.9).

3.4.3 Hyperparameter Tuning

Many models have important input parameters which cannot be directly estimated from the data. Therefore, the model must be trained and tested with different combinations of these parameters. Hyperparameters influence how a model learns and makes its predictions. Tuning hyperparameters for XGBoost and other machine learning algorithms is crucial for achieving optimal performance and improving model accuracy. Since XGBoost has such a large number of hyperparameters, not all hyperparameters are tuned. The selected hyperparameters to be tuned are the number of boosting rounds, the learning rate, the maximum depth of trees, the sample size used in training and the fraction of predictors used for each tree. Tuning the number of boosting rounds is important as too few trees might result in underfitting, while too many trees can lead to overfitting and increased training time. A lower learning rate allows for more conservative updates, which can lead to better generalisation but may require more boosting rounds. Tuning the maximum depth helps prevent overfitting by limiting the complexity of individual trees. Computational cost increases with increasing maximum tree depth. A smaller subsample size can reduce overfitting by introducing more randomness, especially in the presence of highly correlated features or imbalanced datasets. By randomly selecting a subset of features for each tree, the model will not rely too heavily on any particular subset of features, thereby improving generalisation.

The optimal hyperparameters were chosen by iterating through different combinations of hyperparameters and deriving the ROC AUC. Several combinations resulted in an ROC AUC of around 0.90. Therefore, the ability of the model to predict the monthly and yearly cell count was also considered. The selected hyperparameters are shown below:

number of boosting rounds : 558
maximum tree depth : 3
learning rate : 0.05
fraction of features used for each tree : 0.5
subsample size : 0.5

3.4.4 Reproducing the climatology

The probabilities outputted by the model for the testing years 2008, 2010, 2012, 2014 and 2016 are summed to give a cell count estimate.For example, if a probability of 0.25 occurs at four consecutive timesteps at the same grid point then the model would be indicating that one cell at that grid occurred during that time. The cell count refers to a cell being detected at a given grid point, rather than the raw number of cells.
For the years 2010, 2012 and 2014 the model accurately predicts the yearly cell count (Figure 3.11a). The cell count is slightly overestimated in 2008 and underestimated in 2016. Overall, the model can distinguish between a low and high cold-frontal convective cell season, however further validation years would be required for a more robust assessment. The root-mean-square error (RMSE) for the yearly cell counts 372.1. The

overestimation of the cell count in 2008 is linked to July and August 2008 (Figure 3.11b) with an overestimation of around 300 cells. Whereas, in 2016 the model underestimated the cell count during June and July. The lowest observed cell month was May 2010; accurately reproduced by the model. As with the yearly count, the model generally represents the monthly cycle of convective cells well distinguishing between high and low cell count months. The RMSE for the monthly cell counts is 221.4. When the original 60 variable model without hyperparameter tuning was used to reproduce the yearly and monthly cell counts the RMSE was 769.5 and 348.8, respectively. This highlights the importance of feature selection and hyperparameter tuning during model development.

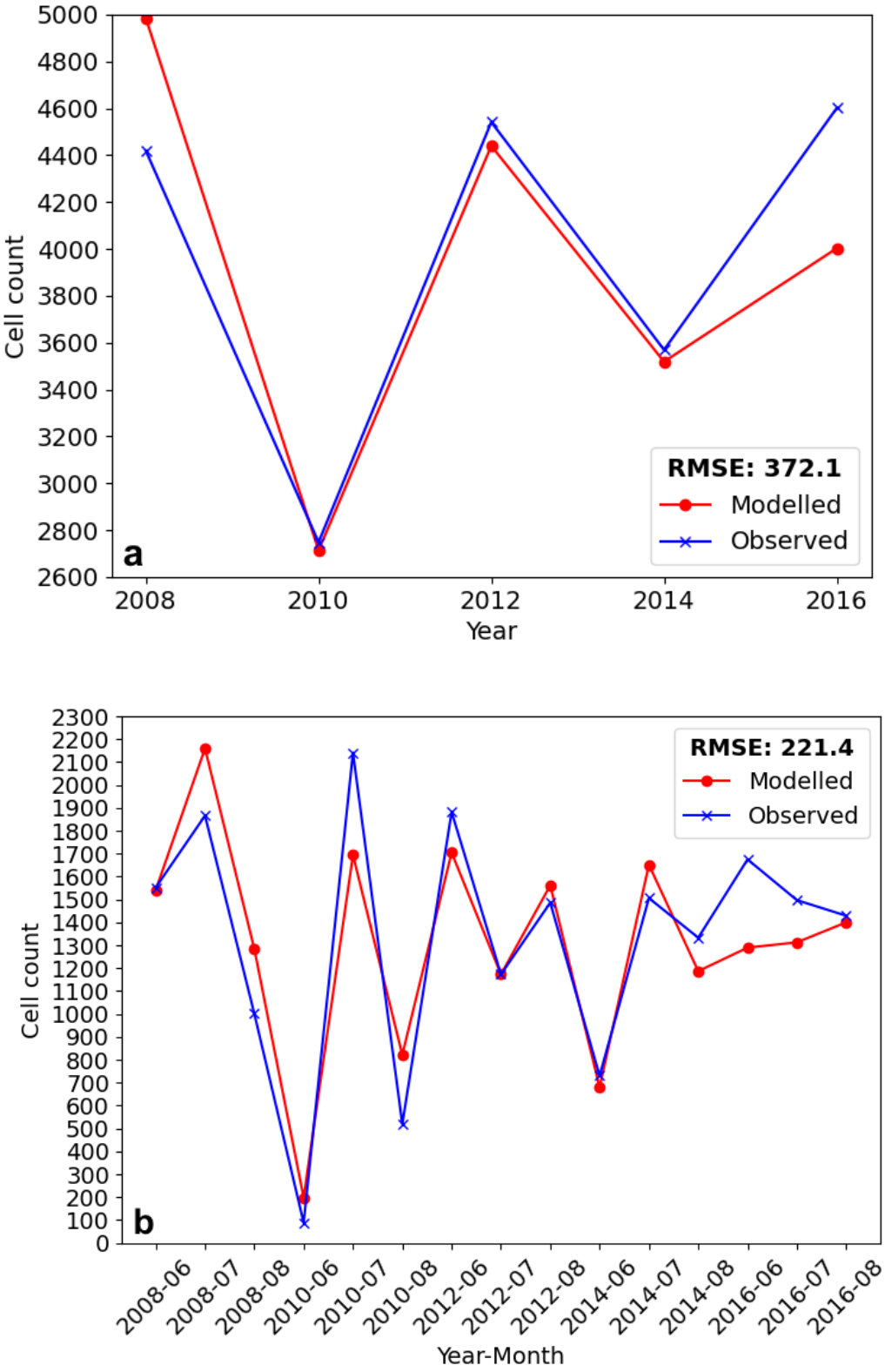


Figure 3.11: Yearly cell count (a) and monthly cell count (b) for the model (red) and observations (blue).

Figure 1.5a from Chapter 1 is recreated showing the modelled and observed cell count depending on the distance from the 700 hPa front. The model is able to accurately reproduce the cell frequency around the 700 hPa front with the maximum cell count in the region 350–450 km and lowest frequency surrounding the 700 hPa front (Figure 3.12). The slight increase in convective cells in the region -450 to -350 km is also captured by the model. However, there are some cell-front relative regions with underestimation and overestimation of the cell count. The largest overestimation is in the region 50–150 km by around 700 cells. The region 350–450 km (the maximum cell count region) has the largest underestimation by around 600 cells. The largest underestimation could be linked to organised convection near the surface front where lifting due to outflow boundaries is not explicitly considered in the predictors. The overestimation in the region 50–150 km could be linked to the vertical velocity, as vertical velocity is climatologically highest in the 50–150 km region (Figure 2.7), though convective cells are not most frequent here.

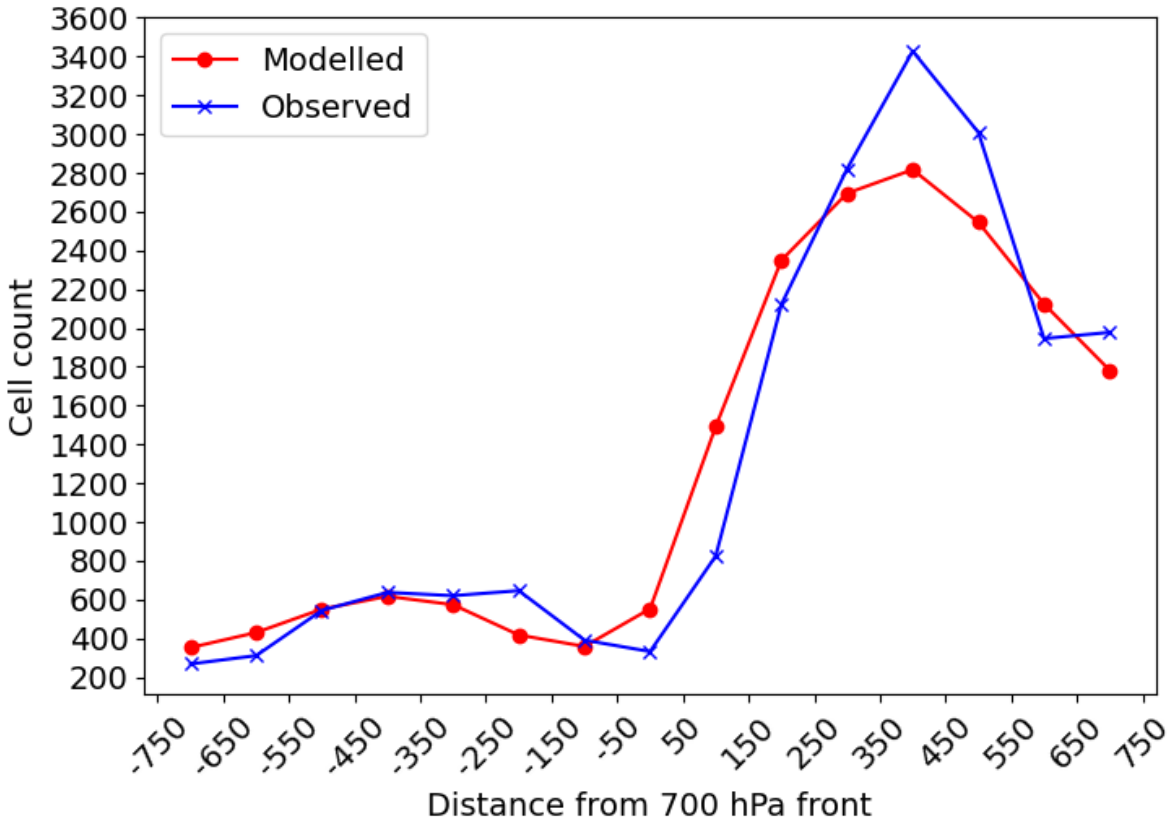


Figure 3.12: Cell count depending on the distance from the 700 hPa front for the model (red) and observations (blue). Testing years: 2008, 2010, 2012, 2014 and 2016.

The heatmap shown in Figure 1.6 is recreated with the modelled and observed cell count for the training years. Unlike Figure 1.6, the cell-front distance range is limited to 750 km as this is the range on which the model was trained. The model correctly shows that post-700hPa-frontal cells rarely occur during the night-time and are mostly confined to the daytime (Figure 3.13). When comparing individual boxes of the heatmap, larger differences are observed indicating the model struggles to identify specifically where and

when in time and space convective cells are likely to occur. The time of day where pre-700hPa-frontal cells typically occur is also captured by the model. The lower cell frequency during the morning hours 550–750 km ahead of the 700 hPa front is captured too. Similar to post-700hPa-frontal, there are larger differences when comparing individual grid boxes and a large underestimation is observed in the cell count, primarily surrounding the region with the maximum cell count. However, the fact that the model produces the highest and lowest cell frequencies in the correct cell-front relative regions and time of day shows the model has learned to some degree what drives convective cell development across the front.

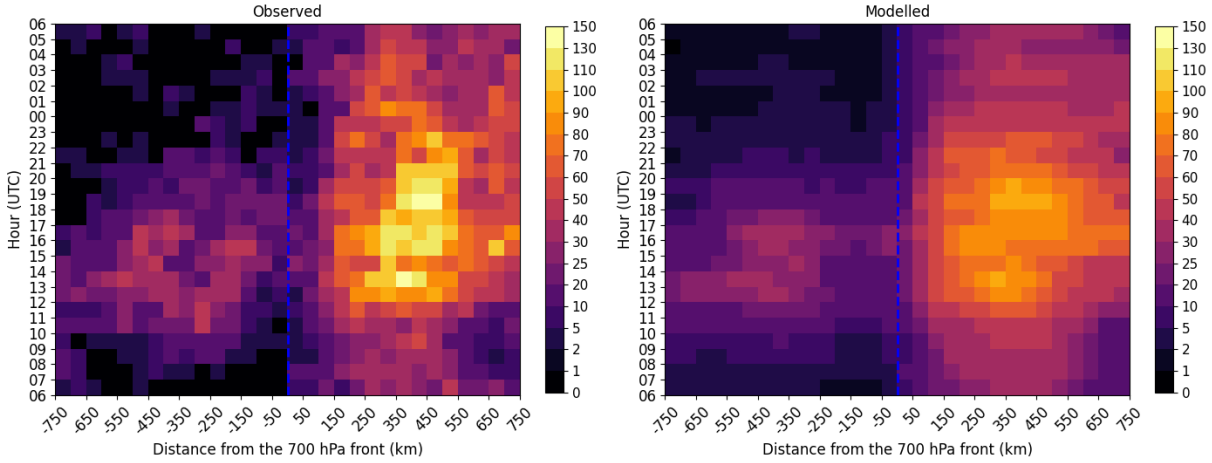


Figure 3.13: Cell count (non-linear colorbar) depending on the cell-front distance (horizontal axis) and hour of the day (vertical axis) for the observations (left) and model (right). The blue horizontal dashed line represents the 700 hPa front location. Testing years: 2008, 2010, 2012, 2014 and 2016.

3.5 Conclusions

Probabilistic models of cold-frontal convective cells in Germany were developed during the summer season between 2007–2016. The predictors primarily came from ERA5 but additional predictors such as the elevation above mean sea level and the time of day were also considered. Two approaches were taken: first, 4 predictive modelling types were trained and tested depending on the region relative to the front. A low number of predictors with limited multicollinearity were used in order to robustly assess the importance of different mechanisms (features) depending on the distance from the front. Second, a larger number of predictors were considered to produce a best possible prediction model. Feature elimination techniques and hyperparameter tuning were utilised to produce an optimal model generalised on cold-frontal convective cells. The resulting model was applied to new and unseen testing data to see how well the model could reproduce the climatology from Chapter 1.

The primary results of this chapter are summarised below:

- Using one moisture, one instability and 3 lifting features, XGBoost was the most skillful model according to the ROC AUC validation metric, except near the surface front where KNN slightly outperformed XGBoost. GAM outperformed GLM at all cell-front relative regions but underperformed XGBoost and KNN (Figure 3.4).
- All the predictive modelling algorithms exhibit the highest skill in the region -300 to -100 km and the lowest skill surrounding the surface front (Figure 3.4). Convective organisation near the surface front as well as representation of variables in ERA5 are likely contributors towards this result.
- The time of day feature shows decreasing feature importance going from further distances behind the 700 hPa to further distances ahead of the 700 hPa front (Figure 3.7), consistent with results in Figures 1.6 and 1.7. The large-scale lifting feature shows the most importance near the 700 hPa front (Figure 3.8), consistent with results in Figure 2.6.
- The elevation feature showed the most importance near the 700 hPa front for KNN and XGBoost but showed little importance for GLM and GAM (Figure 3.9). Considering the nature of the algorithms, the result may highlight that orographic influences alone do not increase the likelihood of convective cells near-700hPa-frontal and other conditions must be met (e.g. solar heating and large-scale lifting).
- A gradient boosted model (XGBoost) with 60 predictors was streamlined to a lower number of predictors using a feature elimination technique. The optimal number of features was 12 which contained several thermodynamic variables such as CAPE and total column water vapour and mid-level relative humidity (Table 3.3). Vertical velocity at 3 different pressure levels, the time of day and high cloud cover fraction were also among the strongest predictors. The 12 features are in agreement with our physical understanding of the ingredients required for convective initiation.
- CAPE was the most important predictor followed by total column water vapour and vertical velocity at 700 hPa (Figure 3.10). During the feature elimination process, the maximum theoretical updraft speed ($\sqrt{2 \times \text{CAPE}}$) was among the least important predictors (Figure C6). The low importance is primarily due to its high correlation with CAPE, which highlights the importance of using features with limited correlation to robustly assess feature importance in statistic models.
- Accumulating the probabilities outputted from the model across a 5 year period unseen by the model during training showed that the model could generally reproduce the monthly and yearly cold-frontal cell frequency well (Figure 3.11). The region where convective cells are most and least frequent relative to the front was also represented well (Figure 3.12).
- The model also learned that post-700hPa-frontal cells have a strong diurnal cycle and pre-700hPa-frontal cells have a weakened diurnal cycle (Figure 3.13), but had difficulty representing the exact cell count for specific cell-front relative regions.

These results are encouraging in light of using statistical models and data driven techniques to predict the likelihood of convection. However, they do come with their limitations, especially if they are to be used to further understand physical processes using feature importance techniques. Since the use of such approaches is likely to increase, the meteorology community should be aware of the limitations regarding the interpretability.

Concluding remarks

The climatology, characteristics and forcing mechanisms of warm-season cold-frontal convective cells have been analysed by considering the region relative to the front using automatic front detection methods applied to ERA5 data (e.g. Jenkner et al., 2010; Thomas and Schultz, 2019) and a cell detection and tracking dataset (KONRAD; Wapler and James, 2015). The primary focus has been on Germany, but preliminary work considering lightning strike data on a larger European domain has shown broadly consistent results for the climatology presented in Chapter 1. The principal objective of this thesis has been to further understand the forcing mechanisms which drive cold-frontal convection, with a special focus on the region relative to the front. This was motivated by the fact that there are gaps in our understanding of where convection will develop in time and space in proximity to cold fronts during the warm-season and the underlying drivers. Since some previous studies have indicated links between cold fronts and hazards (e.g. Gatzen, 2011; Schemm et al., 2016; Kunz et al., 2020), the typical characteristics of convection depending on the region relative to the front were also considered. In this thesis, cold fronts were detected at the 700 hPa level and this level has been used as a reference point throughout. While analysing cold fronts at the 700 hPa level is not conventional in operational forecasting, the selection of the level as a reference point led to several interesting findings which are highlighted in the discussion below. Given that the typical slope of cold fronts is 1:100 and the climatological near-surface convergence is 300 km ahead of the 700 hPa front, this has been used as a reference point for the surface front (see discussion in 1.3.1).

The first approach to achieve the goals of thesis was to understand the spatial and temporal frequency as well as the characteristics of cold-frontal convection depending on the region relative to the front in a climatological sense. A key result from this section was that the location of the 700 hPa front in space marked the lowest convective activity (Figures 1.5a and 1.12). The convective activity then increases slightly at distances further behind the 700 hPa front, but remains much lower than on the warm side of the 700 hPa front. The 700 hPa front also marks a clear shift in regime between convection with a weakened diurnal cycle on the warm side of the 700 hPa cold front and strongly diurnally driven convection on the cold side of the 700 hPa front (Figures 1.6 and A7). Previous studies had alluded to the fact that post-frontal convection is diurnally driven (e.g. Theusner and Hauf, 2004) but due to the backward sloping nature of cold fronts it was not clear exactly where this would be relative to the surface front or upper-level fronts. Convective activity is most frequent marginally ahead of the surface front (Figures 1.5a and 1.12). Cold-season convection studies had often noted that convection is directly at the surface front (e.g. Gatzen, 2011). While at larger distances ahead of the surface front the cell frequency decreases, the fraction of cells with mesocyclones, intense 55 dBZ cores and lightning is equal to or higher than closer to the surface front (Figure 1.11). Therefore, even though cells are less likely ahead of the surface front compared to closer to the surface front, given cell detection

pre-surface-frontal cells are as likely or more likely to be associated with intense cores, mesocyclones and lightning. On a more general note, around 30% of cold-frontal convective cell days were associated with at least one cell with a 50 km^2 area exceeding 55 dBZ opposed to only 10% of non-cold-frontal cell days (Figure 1.10). In addition, cold-frontal cell days were associated with twice as many cells on average compared to non-cold-frontal cell days (Figure 1.4).
The results from the climatology have an important implication on operational weather forecasting. For example, a forecaster may choose to locate fronts at the 700 hPa level and use this as a reference point for lower risk of convective activity. Once the 700 hPa front has passed a location, convection is only likely to occur during the daytime. The surface front is likely already a focus for convective activity for most forecasters, but these results indicate that pre-surface-frontal convection need not be ruled as less severe than convection occurring closer to the front. These results from the climatology may also form a basis for future climatological studies of cold-frontal convection in other parts of Europe and worldwide.

Chapter 2 of this thesis focused on the thermodynamic and lifting mechanisms associated with cold-frontal convective cells in Germany, again considering the region relative to the front. Pre-700hPa-frontal cells form in environments with around 3 °C higher dewpoints and 500 J kg^{-1} higher CAPE compared to post-700hPa-frontal cells (Figures 2.2 and 2.4). The dewpoints and CAPE are highest for pre-surface-frontal cells. The moisture advection ahead of the surface front and the higher pre-surface-frontal CIN (Figure 2.5) likely allows instability to build up avoiding the pre-mature release of convection. Pre-surface-frontal cells have the highest CAPE is likely a contributing factor towards the higher fraction of pre-surface-frontal cells with lightning, mesocyclones and 55 dBZ cores. Higher mid-level relative humidity than climatology was found to be particularly important to allow post-700hPa-frontal convective cell development (Figure 2.3). The higher mid-level relative humidity reduces the likelihood of entrainment and thus updraft dilution. The most striking finding from the large-scale lifting analysis using Q-vector diagnostics was a strong anomaly in the upper-level forcing when cold-frontal cells occur near the 700 hPa frontal line (Figure 2.6). Whether the upper level forcing is important to destabilise upper-layers to increase CAPE and/or relevant for lifting air parcels is not entirely clear. A strong anomaly for ascending motion in the vertical velocity field was found at convective cell grid points (Figure 2.7), raising questions as to whether a signal in ERA5 would be present due to ERA5 partially resolving convection. An analysis of solar heating showed that solar radiation at most front relative regions is lower at convective cell grid points compared to cell regions and non-cell regions (Figure 2.8). Station data showed that most cells develop with limited sunshine duration prior to development, particularly pre-700hPa-frontal cells (Figure 2.9). Therefore, solar heating may not be directly relevant for the majority of cell initiation. The cloudiness before cell detection could be linked to cloudiness from pre-existing convective cells and thus would stress the importance of outflow boundaries and cell interactions on convective cell development. Since cells do have a diurnal cycle (albeit weaker pre-700hPa-frontal), solar heating may be relevant for the first cell initiation and then further cells may be triggered by outflow boundaries. This could explain why post-700hPa-frontal convective cells almost always occur during the daytime but have limited sunshine duration before initiation.
The results in chapter 2 leave several interesting open questions for future work. For example, is the large-scale lifting more important for increasing instability or also relevant for lifting parcels to their LFC and maintaining positive buoyancy? Given the large fraction of cells with low sunshine duration prior to cell detection, to what degree do outflow boundaries play a role on convective initiation depending on the distance

from the front?

Chapter 3 of this thesis used a statistical modelling approach to expand on the simple univariate statistical comparison made in Chapter 2. Statistical models were trained with various predictors (features) and the model's ability to predict convective cell occurrence was assessed. Using 4 predictive modelling algorithms (section 3.2.1), the highest model skill was behind the 700 hPa front and lowest skill was near the surface front for all four predictive models (Figure 3.4). Convective organisation near the surface front as well as representation of variables in ERA5 are likely contributors towards this result. Feature importance methods showed that a solar heating feature based on the time of day was most important post-700hPa frontal (Figure 3.7) and a large-scale lifting feature was most important near the 700 hPa front (Figure 3.8), consistent with results in previous chapters. The elevation feature showed the most importance near the 700 hPa front for two predictive models, but showed little importance for the other two models (Figure 3.9). Considering how each predictive model makes predictions, the differences may be linked to the fact that orographic influences alone do not increase the likelihood of convective cells near-700hPa-frontal and that other additional sources of lift are required (e.g. solar heating and large-scale lifting). The result also emphasises the importance of testing different predictive models as one model may better capture the patterns between a certain feature (and interactions with other features) and the target variable.

To develop a best possible prediction model generalised on cold-frontal convective cells, a gradient boosted model was developed. The most skillful predictors were several thermodynamic variables such as CAPE and total column water vapour and mid-level relative humidity (Figure 3.10). The importance of upper-level humidity across the front and CAPE was also found in Chapter 2 (Figures 2.3 and 2.4). Vertical velocity at 3 different pressure levels, the time of day and high cloud cover fraction were also among the strongest predictors. On the contrary, the maximum theoretical updraft speed was among the least important highlighting the importance of using features with limited correlation to robustly assess feature importance. The probabilities outputted from the model were accumulated across a 5 year period unseen by the model during training. This showed that the model could generally reproduce the monthly and yearly cold-frontal cell frequency well (Figure 3.11). The region where convective cells are most (and least) frequent relative to the front was also represented well (Figure 3.12), albeit with an underestimation and overestimation in some front relative regions. The model also learned that post-700hPa-frontal cells have a strong diurnal cycle and pre-700hPa-frontal cells have a weakened diurnal cycle (Figure 3.13), but had difficulty representing the exact cell count for specific cell-front relative regions.

These results are encouraging in light of using statistical models and data driven techniques to predict convective events. However, they do come with their limitations, especially if they are to be used to further understand physical processes using feature importance techniques.

Key acronyms and definitions

AGL Above Ground Level

CAPE Convective Available Potential Energy*

CIN Convective Inhibition

cos$_{time}$ A cosine transformation of the time of day

CPM Convective Permitting Model

DWD Deutscher Wetter Dienst (German Weather Service)

ECWMF European Centre for Medium-Range Weather Forecasts

ESWD European Severe Weather Database

ERA5 ECMWF Reanalysis version 5

GAM Generalised Additive Model

GLM Generalised Linear Model

HIHC High cloud cover

KONRAD KONvektionsentwicklung in RADarprodukten; convection evolution in radar products) - a convective cell detection and tracking algorithm

K-fold cross-validation A technique used to assess the robustness of model performance by dividing the dataset into k subsets (folds) and evaluating the model for each subset

KNN K-nearest neighbors algorithm

TOTC Total cloud cover

LCL Lifting condensation level

LFC Level of free convection

LR Lapse Rate

LST Local Solar Time

NWP Numerical Weather Prediction

Post-700hPa-frontal The cold side of the 700 hPa front

Pre-700hPa-frontal The warm side of the 700 hPa front

Pre-surface-frontal Ahead of the surface front

RFECV Recursive Feature Elimination with Cross-Validation

RH Relative Humidity

ROC AUC Receiver operating characteristic area under the curve

TCWV Total Column Water Vapour

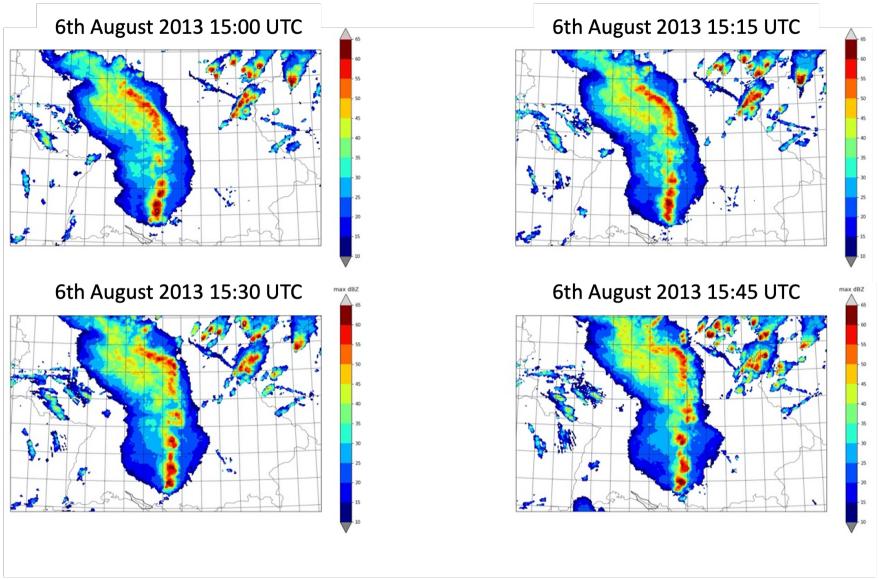
QVEC Q-vector convergence

UTC Coordinated Universal Time

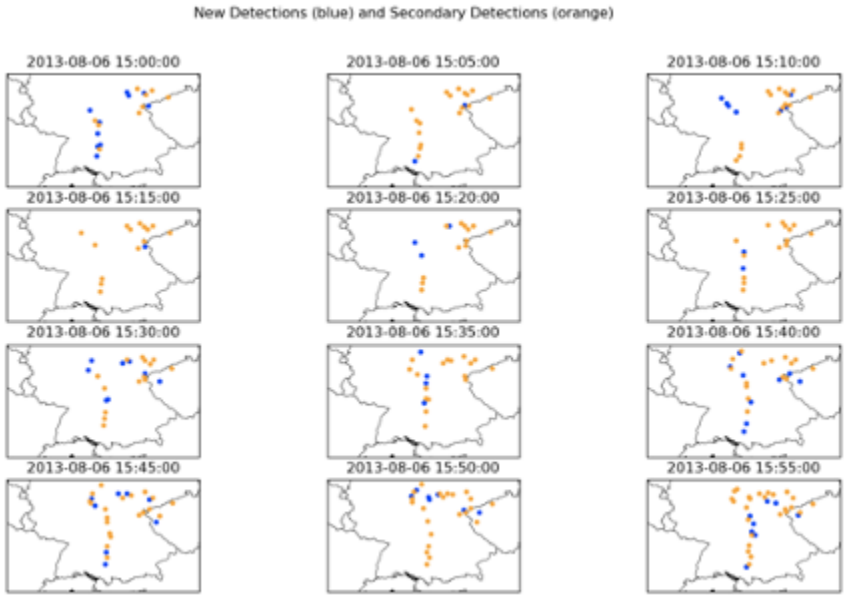
XGBoost Extreme Gradient Boosting (Machine Learning Algorithm)

***The CAPE used in this thesis is the ERA5 CAPE parameter downloaded from the Copernicus Climate Data Store (Hersbach et al., 2018b). CAPE is derived considering parcels departing from different model levels below the 350 hPa level and the departure level with the highest CAPE is retained. In essence, the ERA5 CAPE parameter refers to the most unstable CAPE (MUCAPE). The following is assumed: (i) the parcel of air does not mix with surrounding air; (ii) ascent is pseudo-adiabatic (all condensed water falls out) and (iii) other simplifications related to the mixed-phase condensational heating.

Appendix A - Chapter 1



(a) OPERA Rada Data 06 August 2013 15:00 – 15:45 UTC



(b) KONRAD Cell Detections 06 August 2013 15:00 – 15:55 UTC

Figure A1: KONRAD cell detections and comparison to radar data during an MCS case study on 6 August 2013. OPERA Radar Data (Huuskonen et al., 2014) shows the max dBZ (a) and KONRAD cell detections (b). First cell detections in blue and secondary or later detections in orange.

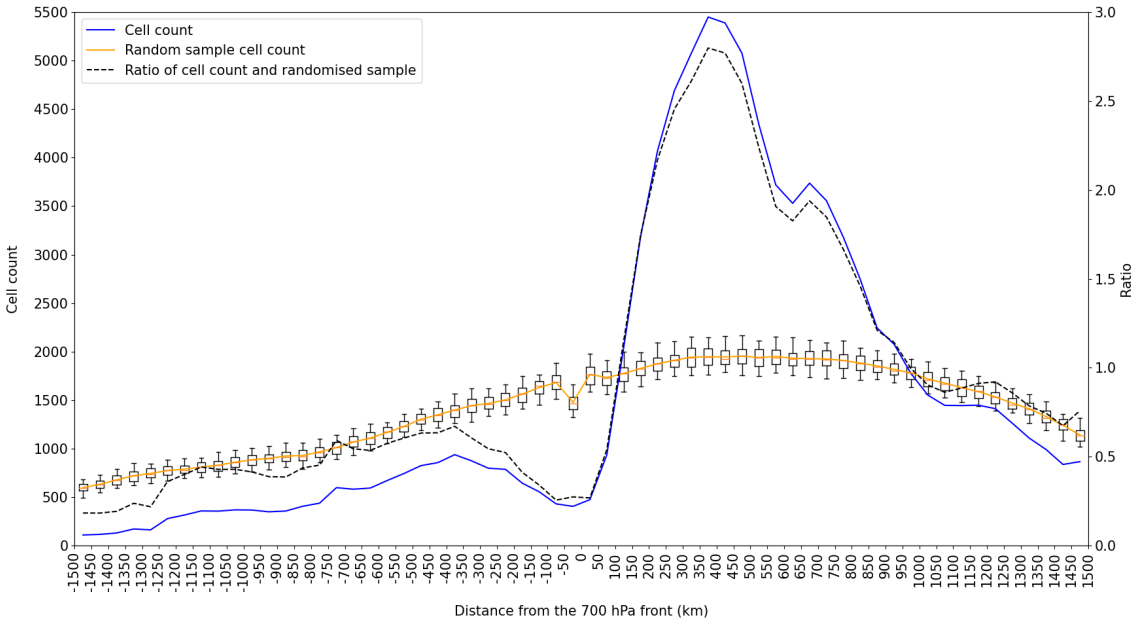


Figure A2: Cell count (blue), random sample cell count mean (orange) and ratio of cell count and randomised sample (black) on the secondary axis. The random sample was performed 100 times and the 5th and 95th percentiles are shown by the whiskers of the box plots.

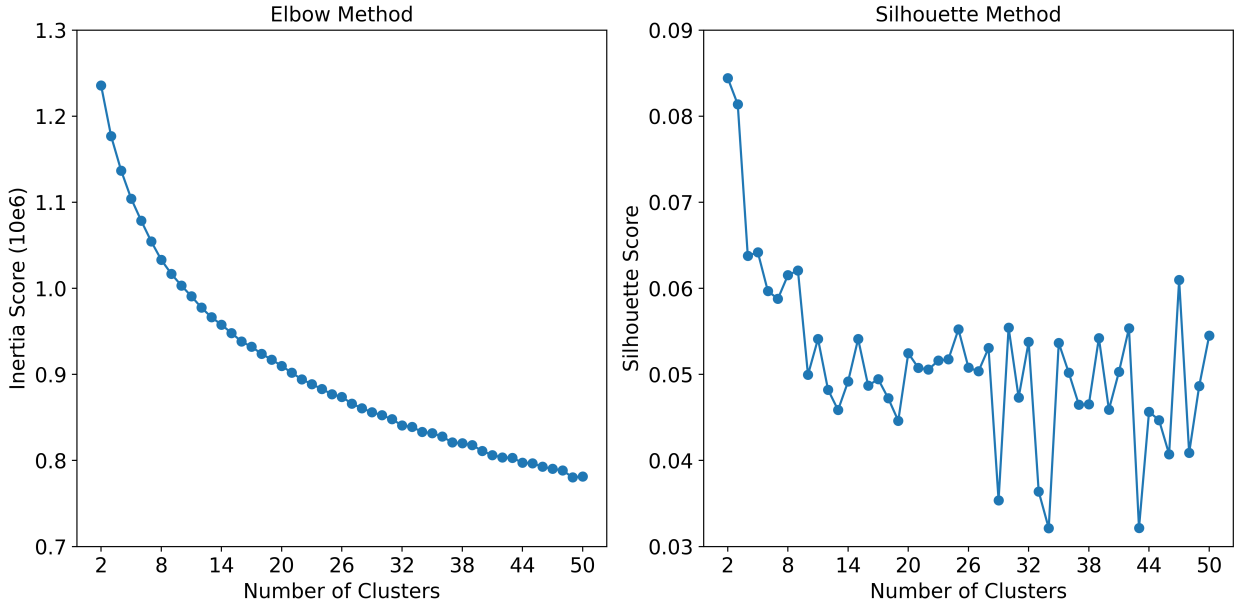


Figure A3: Elbow method (left) and silhouette score method (right) applied for cluster numbers between 2 and 50.

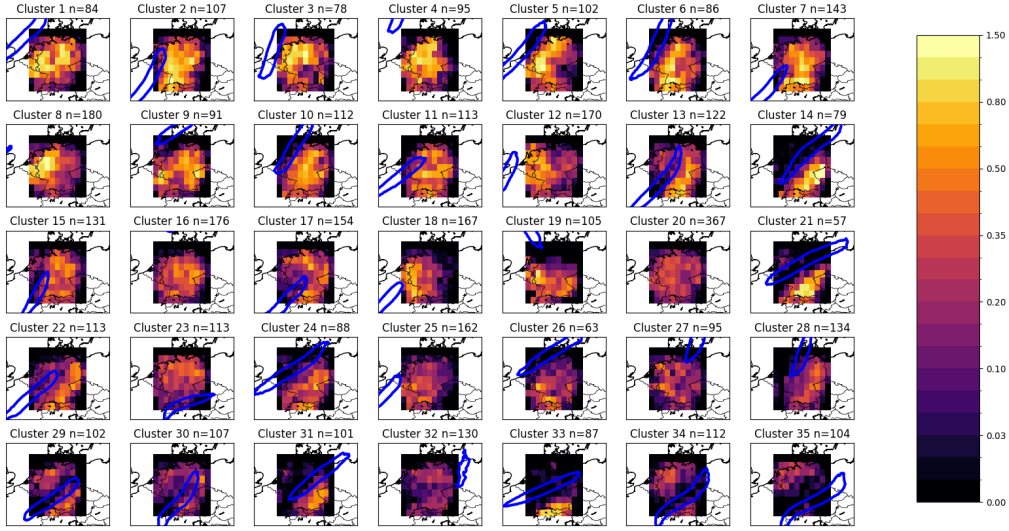


Figure A4: As Figure 1.9 but for 15 clusters. Absence of a blue contour indicates high within-cluster variance and no common front type associated to that cluster.

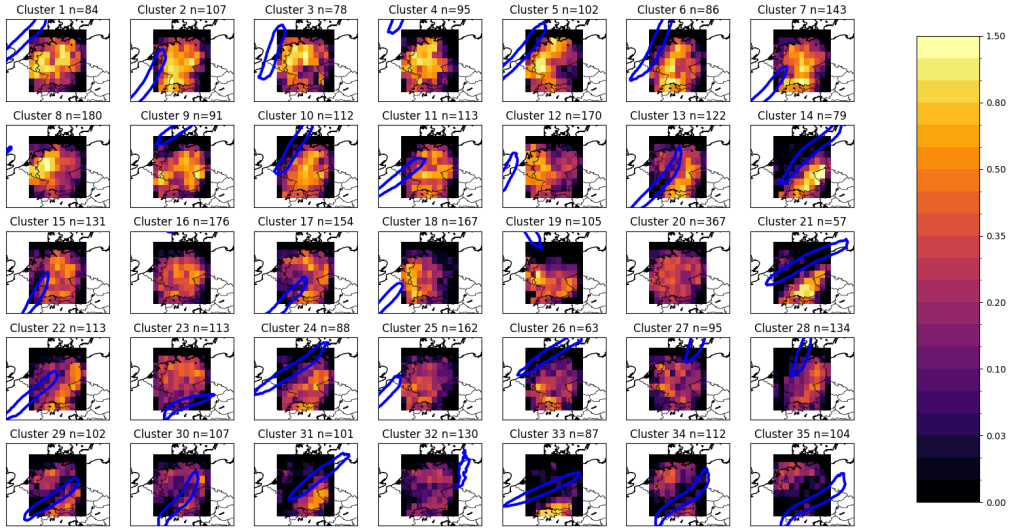


Figure A5: As Figure 1.9 but for 35 clusters. Absence of a blue contour indicates high within-cluster variance and no common front type associated to that cluster.

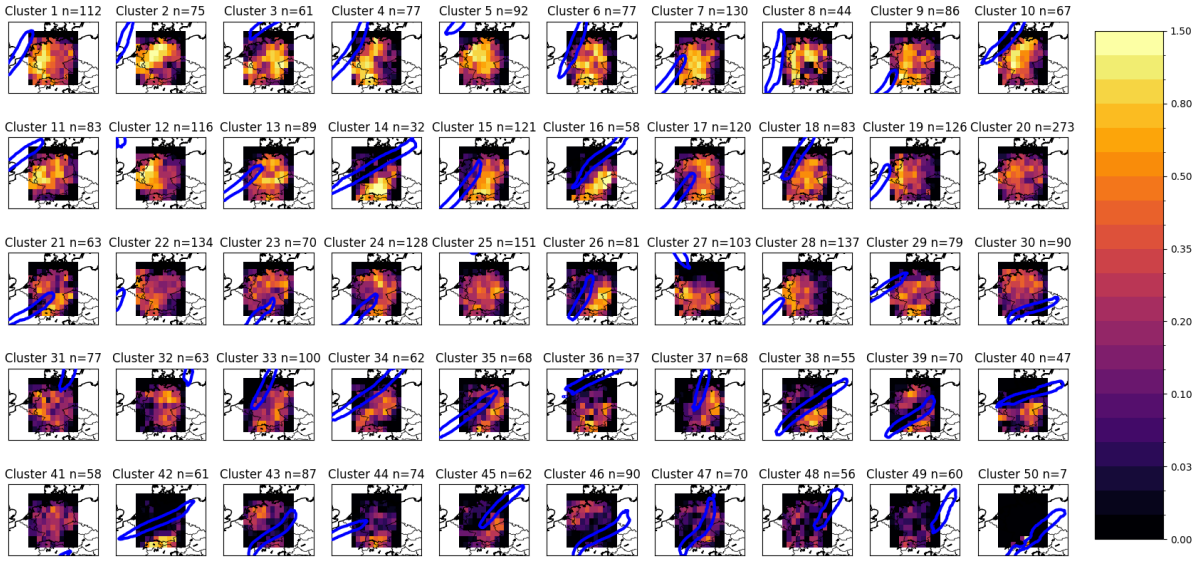


Figure A6: As Figure 1.9 but for 50 clusters. Absence of a blue contour indicates high within-cluster variance and no common front type associated to that cluster.

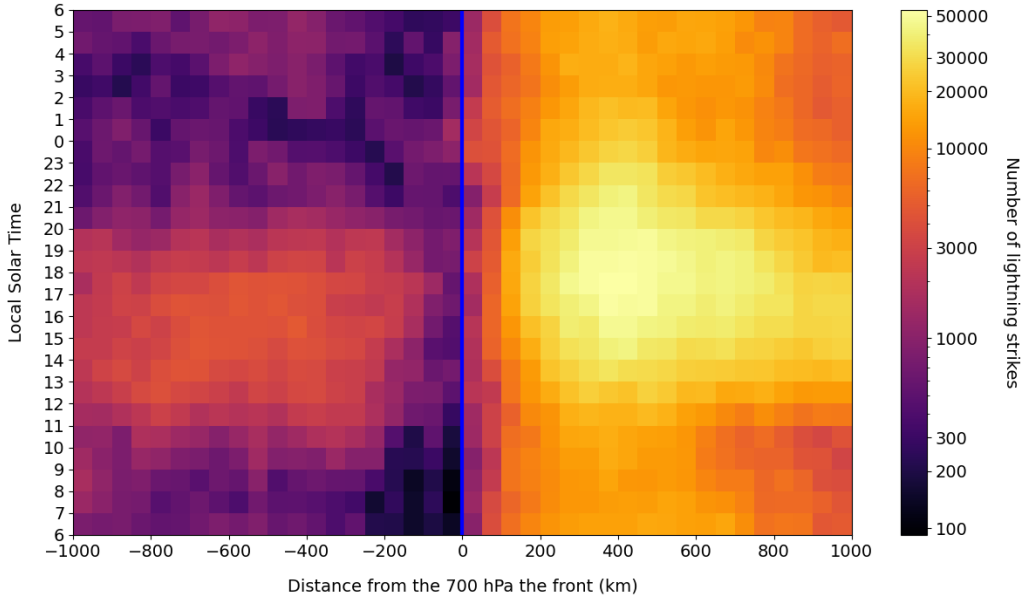


Figure A7: As Figure 1.6 but for lightning strike frequency in a sub-European domain (see grey domain in Figure 1.1) depending on the distance from the 700 hPa front. Note the non-linear colorbar.

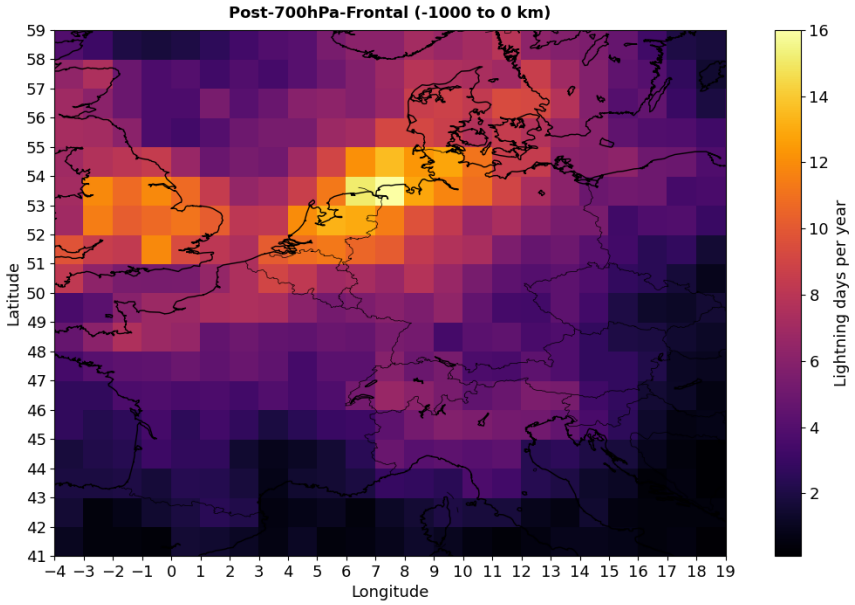


Figure A8: As Figure 1.8 but for post-700hPa-frontal lightning strike frequency. A lightning day refers to a day with at least 1 strike in a given grid box. Note that a 1000 km range is used rater than 750 km used in Figure 1.8a.

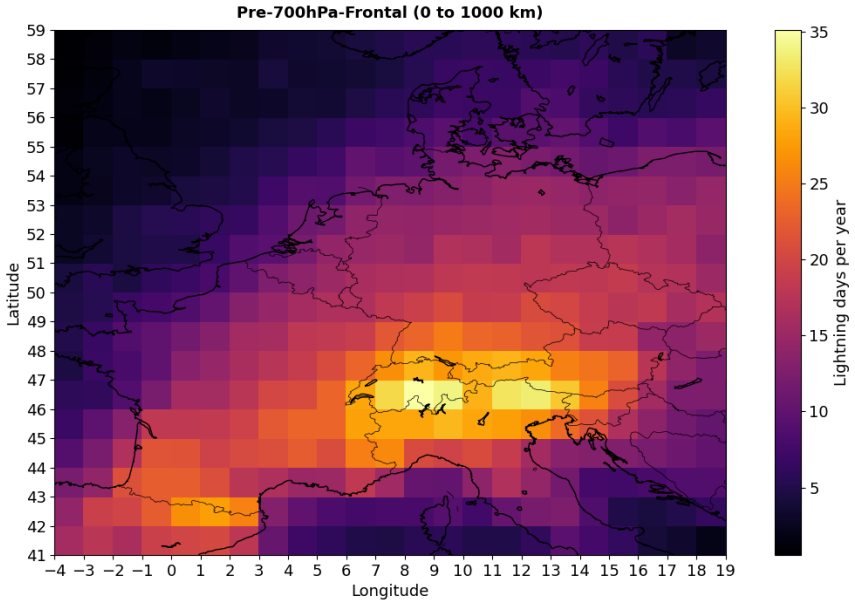


Figure A9: As Figure A8 but for pre-700hPa-frontal. A lightning day refers to a day with at least 1 strike in a given grid box. Note that a 1000 km range is used rater than 750 km used in Figure 1.8b.

Appendix B - Chapter 2

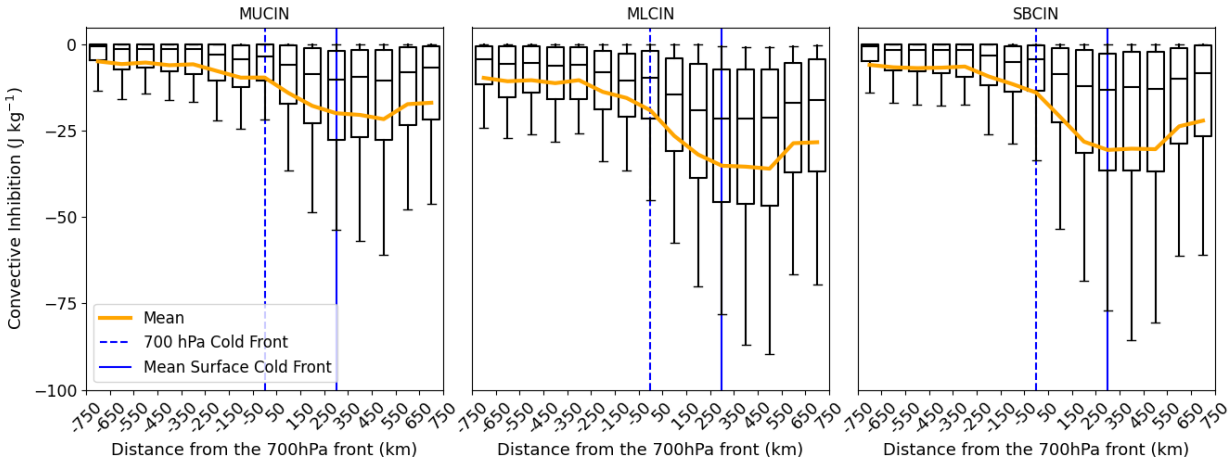


Figure B1: As Figure 2.5 but for cells between 09–18 UTC only. Note the smaller y-axis range compared to Figure 2.5.

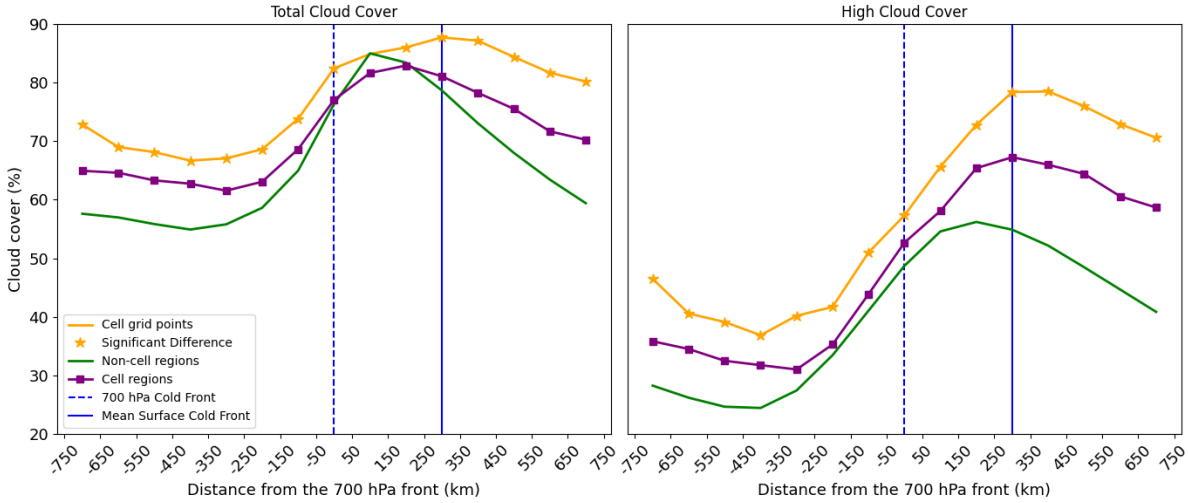


Figure B2: As Figure 2.2 but total cloud cover (left) and high-cloud cover (right).

Appendix C - Chapter 3

Table C1: Number of data (grid) points between 2007–2016 (June–August) and percentage of true events for each front relative model used in section 3.3.

Distance from the 700 hPa front	Number of grid points	Percentage of true events
-700 to -500 km	277,294	0.655%
-500 to -300 km	365,829	0.759%
-300 to -100 km	423,204	0.594%
-100 to 100 km	435,604	0.391%
100 to 300 km	460,329	1.847%
300 to 500 km	477,710	2.634%
500 to 700 km	472,753	1.900%
Total	2,912,723	1.300%

Table C2: Full predictor list considered in section 3.4 with short name, units and levels.

Predictor	**Short Name**	**Units**	**Level**
Dewpoint	Td_{2m}	°C	2-metres AGL
Specific humidity	QHUM	kg kg^{-1}	900, 850, 700, 500
Total column water vapour	TCWV	kg m^{-2}	integrated through entire atmosphere
Relative Humidity	RH	%	850, 700, 600, 500, 850-500hPaav
Convective Available Potential Energy	CAPE	J kg^{-1}	integrated between LFC and EL
Maximum theoretical updraft velocity	w_{max}	m s^{-1}	integrated between LFC and EL
Lapse Rates	LR75, LRSL	°C	700–500 hPa, surface-850 hPa
Air temperature	TEMP	°C	surface, 900, 850, 700, 500 hPa
Total Incoming Solar Radition	TISR	W m^{-2}	surface
Q-vector convergence	QVEC	m^2 kg^{-1} s^{-1}	900,850,700,500,900-500max,900-500mean
Convergence	CONV	s^{-1}	900, 850, 700, 500 hPa
Vertical Velocity	w	m s^{-1}	850, 700, 500 hPa
Wind shear	WSHR	m s^{-1}	surface-500 hPa
U component of wind	UWND	m s^{-1}	925, 850, 700, 500 hPa
V component of wind	VWND	m s^{-1}	925,850,700,500 hPa
Wind speed	WSPD	m s^{-1}	925, 850, 700, 500, 250 hPa
Total Cloud Cover	TOTC	%	single level
High Cloud Cover	HIHC	%	single level
Elevation	ELEV	km	single level
Distance from elevation	ELEV(height)m	km	200m, 500m, 800m
Hour of the day	HOUR	hours	single level
Cosine time function	cos_{time}	dimensionless	single level
Distance from the 700 hPa front	FrontCellDist	km	single level
Longitude	LON	degrees	single level
Latitude	LAT	degrees	single level

K-fold cross-validation

K-fold cross-validation is a technique used to assess model performance and the model's ability to generalise to new unseen data. The original dataset is divided into K subsets of approximately equal size, which are called "folds". The model is trained K times, where one of the folds is used as the testing set and the remaining folds are used as the training set.

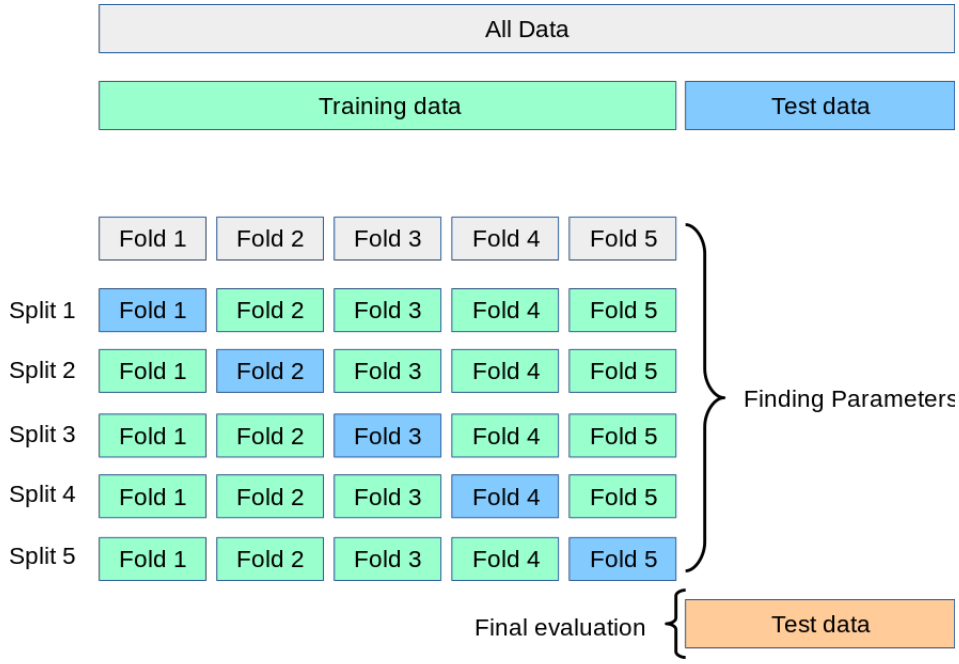


Figure C1: Visualisation of k-folds cross-validation where k=5 (Pedregosa et al., 2011)

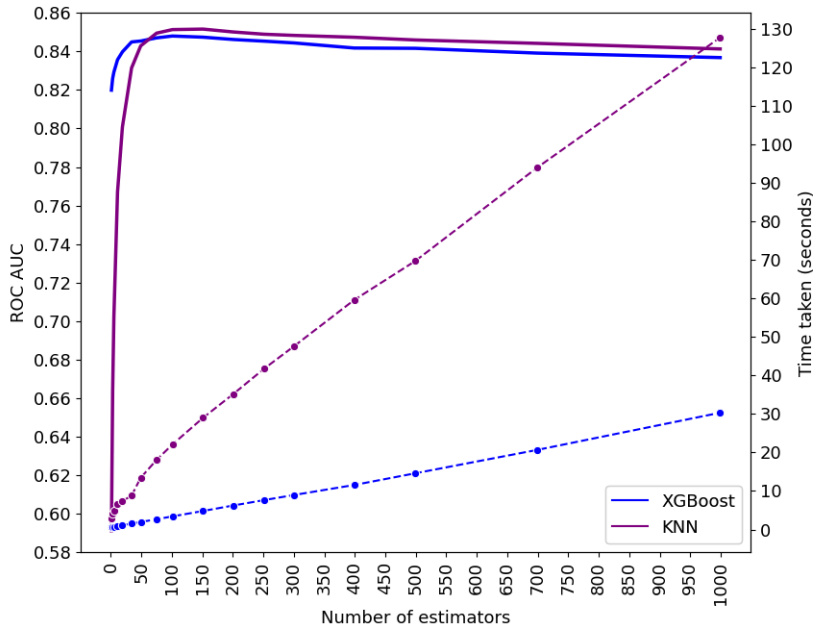


Figure C2: ROC AUC depending on the number of estimators for XGBoost (blue) and KNN (purple). The dashed lines represent the time taken to train and test the models (right y-axis).

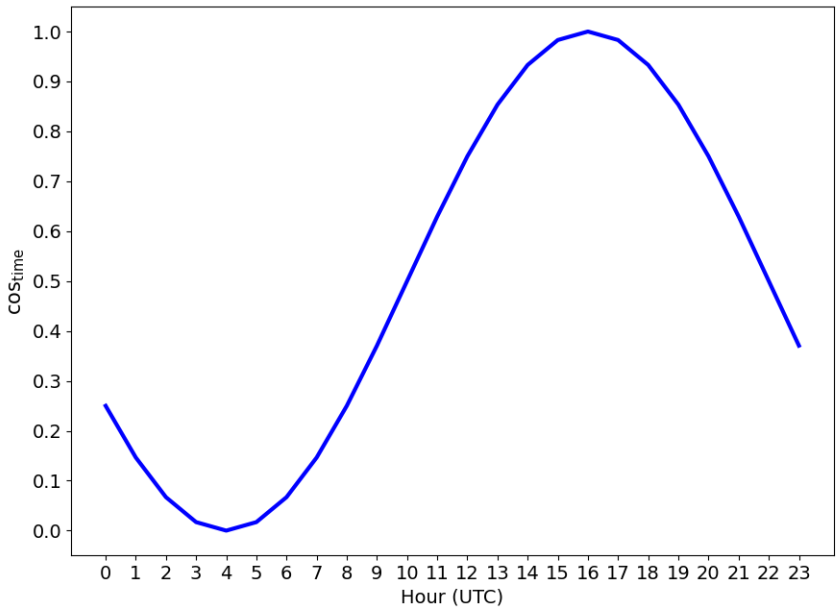


Figure C3: $\cos_{time}$ as a function of h

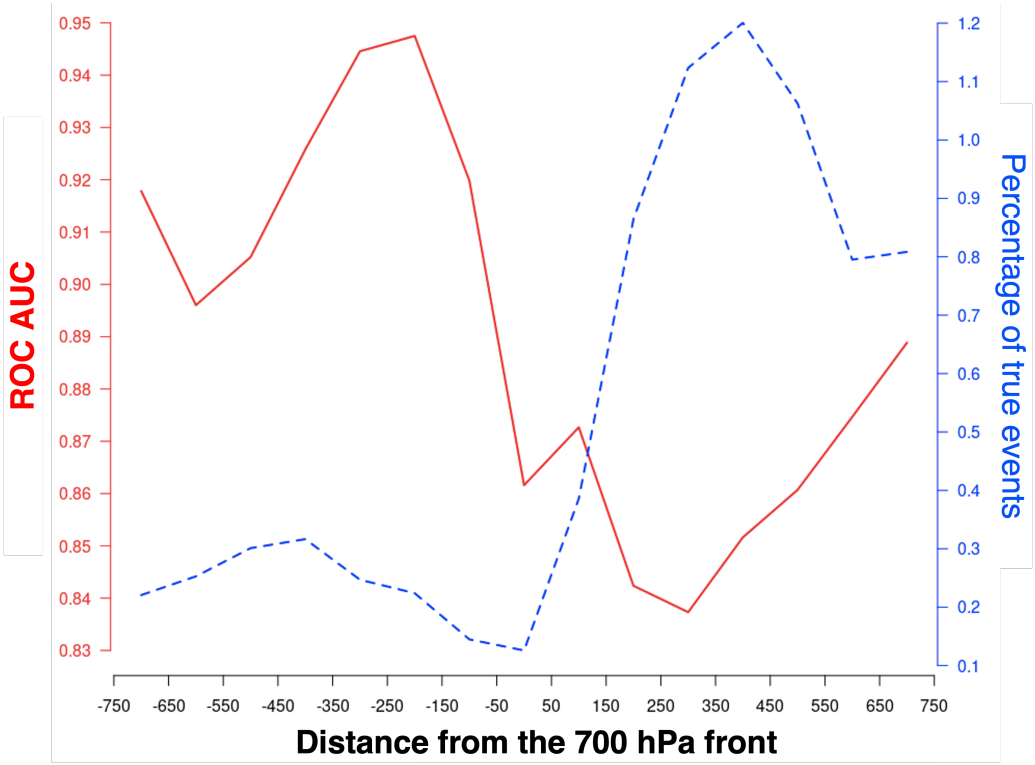


Figure C4: GLM trained with TEMP_{2m}, TEMP900, TEMP700, TEMP500, Td_{2m}, QHUM900, QHUM700, QHUM500 and $w_{500\ \mathrm{hPa}}$. Model was trained and tested using 2007–2009 (April–September) and 2010–2016 (April–September), respectively. Variable short names are shown in Table C2.

Recursive feature elimination with cross-validation (RFECV)

Recursive feature elimination (RFE) is a machine learning method for feature selection. A model is first trained with all available features. Features are ranked based on their importance or contribution to the model's performance. RFECV eliminates the least important feature(s) from the dataset. The process is repeated iteratively until the predefined number of features is reached or until a specific performance criterion is met. In section 3.4, the performance criterion is the ROC AUC. At each iteration of feature elimination, RFECV (RFE with cross validation) uses cross-validation to evaluate the model's performance. This means the feature importance is assessed using different training and testing splits and averaged over all folds. The RFECV process was carried out using the Python scikit-learn package (Pedregosa et al., 2011)

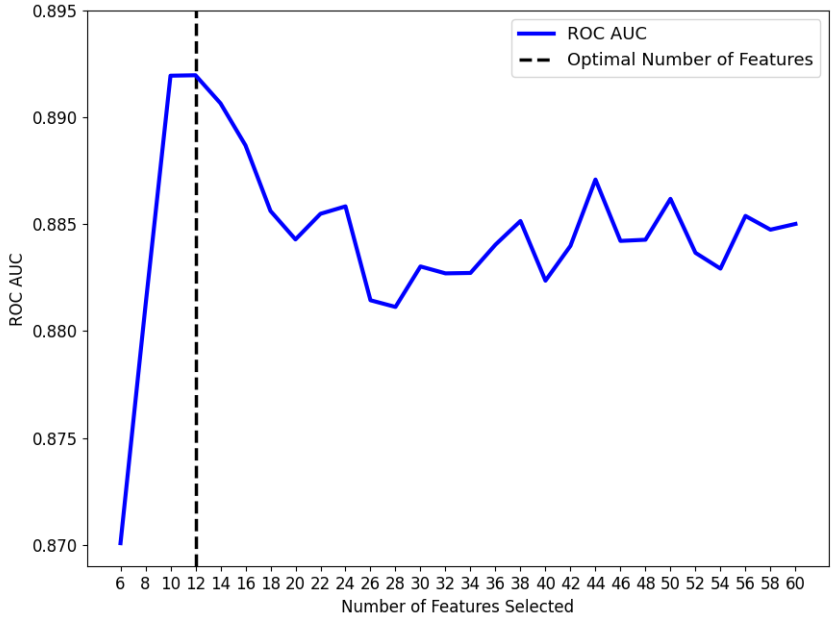


Figure C5: ROC AUC depending on the number of features selected during the RFECV performed in section 3.4.

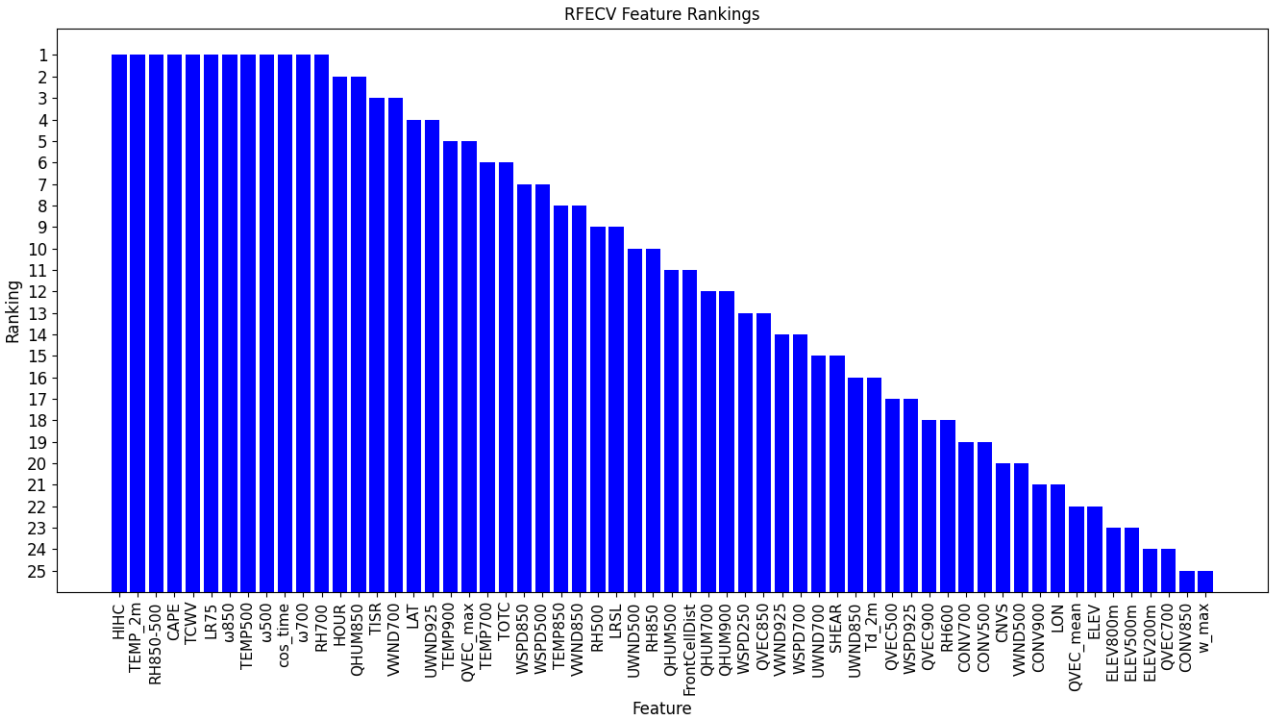


Figure C6: Feature ranking for RFECV used in section 3.4. Lower numbers indicate higher ranking. Features with the same ranking are not ordered on the x-axis in terms of importance.

Bibliography

N. Ban, J. Schmidli, and C. Schär. Evaluation of the convection-resolving regional climate modeling approach in decade-long simulations. *Journal of Geophysical Research: Atmospheres*, 119(13):7889–7907, 2014. doi: 10.1002/2014JD021478. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014JD021478.

F. Battaglioli, P. Groenemeijer, T. Púčik, M. Taszarek, U. Ulbrich, and H. Rust. Modeled multidecadal trends of lightning and (very) large hail in europe and north america (1950–2021). *Journal of Applied Meteorology and Climatology*, 62(11):1627 – 1653, 2023. doi: https://doi.org/10.1175/JAMC-D-22-0195.1. URL https://journals.ametsoc.org/view/journals/apme/62/11/JAMC-D-22-0195.1.xml.

H. D. Betz, K. Schmidt, P. Laroche, P. Blanchet, W. P. Oettinger, E. Defer, Z. Dziewit, and J. Konarski. Linet—an international lightning detection network in europe. *Atmospheric Research*, 91(2-4):564–573, 2009. doi: 10.1016/j.atmosres.2008.06.012.

K. Bi, L. Xie, H. Zhang, and et al. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619:533–538, 2023. doi: 10.1038/s41586-023-06185-3. URL https://doi.org/10.1038/s41586-023-06185-3.

J. Bjerknes. On the structure of moving cyclones. *Monthly Weather Review*, 47(2):95 – 99, 1919. doi: 10.1175/1520-0493(1919)47<95:OTSOMC>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/47/2/1520-0493_1919_47_95_otsomc_2_0_co_2.xml.

H. B. Bluestein. *Synoptic-dynamic meteorology in midlatitudes. Volume II. Observations and theory of weather systems*. Oxford University Press, 1 1993. URL https://www.osti.gov/biblio/6227041.

A. Bott. *Synoptische Meteorologie: Methoden der Wetteranalyse und -prognose, Mesoskalige meteorologische Prozesse*, page 475. Springer Berlin Heidelberg, Berlin, Heidelberg, 2016a. doi: 10.1007/978-3-662-48195-0_12. URL https://doi.org/10.1007/978-3-662-48195-0_12.

A. Bott. *Synoptische Meteorologie: Methoden der Wetteranalyse und -prognose, Fronten und Frontalzonen*, page 397. Springer Berlin Heidelberg, Berlin, Heidelberg, 2016b. doi: 10.1007/978-3-662-48195-0_11. URL https://doi.org/10.1007/978-3-662-48195-0_11.

Z. B. Bouallègue, M. C. A. Clare, L. Magnusson, E. Gascón, M. Maier-Gerber, M. Janoušek, M. Rodwell, F. Pinault, J. S. Dramsch, S. T. K. Lang, B. Raoult, F. Rabier, M. Chevallier, I. Sandu, P. Dueben, M. Chantry, and F. Pappenberger. The rise of data-driven weather forecasting: A first statistical assessment of machine learning-based weather forecasts in an operational-like context. *Bulletin of the American*

Meteorological Society, 2024. doi: 10.1175/BAMS-D-23-0162.1. URL https://journals.ametsoc.org/view/journals/bams/aop/BAMS-D-23-0162.1/BAMS-D-23-0162.1.xml.

H. E. Brooks, C. A. Doswell, and J. Cooper. On the environments of tornadic and nontornadic mesocyclones. *Weather and Forecasting*, 9(4):606 – 618, 1994. doi: 10.1175/1520-0434(1994)009<0606:OTEOTA>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/wefo/9/4/1520-0434_1994_009_0606_oteota_2_0_co_2.xml.

K. A. Browning. Conceptual models of precipitation systems. *Weather and Forecasting*, 1(1):23 – 41, 1986. doi: 10.1175/1520-0434(1986)001<0023:CMOPS>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/wefo/1/1/1520-0434_1986_001_0023_cmops_2_0_co_2.xml.

K. A. Browning and G. A. Monk. A simple model for the synoptic analysis of cold fronts. *Quarterly Journal of the Royal Meteorological Society*, 108(456):435–452, 1982. doi: https://doi.org/10.1002/qj.49710845609. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.49710845609.

C. Calvo-Sancho, J. Díaz-Fernández, Y. Martín, P. Bolgiani, M. Sastre, J. J. González-Alemán, D. Santos-Muñoz, J. I. Farrán, and M. L. Martín. Supercell convective environments in spain based on era5: hail and non-hail differences. *Weather and Climate Dynamics*, 3(3):1021–1036, 2022. doi: 10.5194/wcd-3-1021-2022. URL https://wcd.copernicus.org/articles/3/1021/2022/.

T. N. Carlson, S. G. Benjamin, G. S. Forbes, and Y.-F. Li. Elevated mixed layers in the regional severe storm environment: Conceptual model and case studies. *Monthly Weather Review*, 111(7):1453 – 1474, 1983. doi: 10.1175/1520-0493(1983)111<1453:EMLITR>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/111/7/1520-0493_1983_111_1453_emlitr_2_0_co_2.xml.

J. L. Catto and S. Pfahl. The importance of fronts for extreme precipitation. *Journal of Geophysical Research: Atmospheres*, 118(19):10,791–10,801, 2013. doi: 10.1002/jgrd.50852. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/jgrd.50852.

M. Clark. A provisional climatology of cool-season convective lines in the uk. *Atmospheric Research*, 123: 180–196, 2013. ISSN 0169-8095. doi: 10.1016/j.atmosres.2012.09.018. URL https://www.sciencedirect.com/science/article/pii/S0169809512003201. 6th European Conference on Severe Storms 2011. Palma de Mallorca, Spain.

S. F. Corfidi, S. J. Corfidi, and D. M. Schultz. Elevated convection and castellanus: Ambiguities, significance, and questions. *Weather and Forecasting*, 23(6):1280 – 1303, 2008. doi: 10.1175/2008WAF2222118.1. URL https://journals.ametsoc.org/view/journals/wefo/23/6/2008waf2222118_1.xml.

T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967. doi: 10.1109/TIT.1967.1053964.

J. M. L. Dahl and J. Fischer. The origin of western european warm-season prefrontal convergence lines. *Weather and Forecasting*, 31(5):1417 – 1431, 2016. doi: 10.1175/WAF-D-15-0161.1. URL https://journals.ametsoc.org/view/journals/wefo/31/5/waf-d-15-0161_1.xml.

H. Davies and H. Wernli. Dynamical meteorology | quasigeostrophic theory. In G. R. North, J. Pyle, and F. Zhang, editors, *Encyclopedia of Atmospheric Sciences (Second Edition)*, pages 393–403. Academic Press,

Oxford, second edition edition, 2015. ISBN 978-0-12-382225-3. doi: 10.1016/B978-0-12-382225-3.00326-1. URL https://www.sciencedirect.com/science/article/pii/B9780123822253003261.

Deutscher Wetterdienst. Description of solar radiation data from german weather service. https://opendata.dwd.de/climate_environment/CDC/observations_germany/climate/10_minutes/solar/DESCRIPTION_obsgermany_climate_10min_solar_en.pdf, 2024. Last Accessed on 15th February 2024.

C. Doswell. *Severe Convective Storms—An Overview*, volume 28. 11 2001. doi: 10.1175/0065-9401-28.50.1.

C. A. Doswell and J. S. Evans. Proximity sounding analysis for derechos and supercells: an assessment of similarities and differences. *Atmospheric Research*, 67-68:117–133, 2003. ISSN 0169-8095. doi: https://doi.org/10.1016/S0169-8095(03)00047-4. URL https://www.sciencedirect.com/science/article/pii/S0169809503000474. European Conference on Severe Storms 2002.

C. A. Doswell, H. E. Brooks, and R. A. Maddox. Flash flood forecasting: An ingredients-based methodology. *Weather and Forecasting*, 11(4):560 – 581, 1996. doi: 10.1175/1520-0434(1996)011<0560:FFFAIB>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/wefo/11/4/1520-0434_1996_011_0560_fffaib_2_0_co_2.xml.

N. Dotzek, P. Groenemeijer, B. Feuerstein, and A. Holzer. Overview of essl's severe convective storms research using the european severe weather database eswd. *Atmos. Res.*, 93:575–586, 2009.

ECMWF. First update on aifs. https://www.ecmwf.int/en/about/media-centre/aifs-blog/2024/first-update-aifs, 2024. Last Accessed: March, 2024.

EUMeTrain. Eumetrain - satellite meteorology: Cloud formation, 2012. URL https://resources.eumetrain.org/satmanu/CMs/Cf/navmenu.php?page=2.0.0. Accessed in January 2024.

R. Ferretti, E. Pichelli, S. Gentile, I. Maiello, D. Cimini, S. Davolio, M. M. Miglietta, G. Panegrossi, L. Baldini, F. Pasi, F. S. Marzano, A. Zinzi, S. Mariani, M. Casaioli, G. Bartolini, N. Loglisci, A. Montani, C. Marsigli, A. Manzato, A. Pucillo, M. E. Ferrario, V. Colaiuda, and R. Rotunno. Overview of the first hymex special observation period over italy: observations and model results. *Hydrology and Earth System Sciences*, 18 (5):1953–1977, 2014. doi: 10.5194/hess-18-1953-2014. URL https://hess.copernicus.org/articles/18/1953/2014/.

A. Fink, T. Brücher, V. Ermert, K. A, and J. Pinto. The european storm kyrill in january 2007: Synoptic evolution, meteorological impacts and some considerations with respect to climate change. *Natural Hazards and Earth System Sciences*, 9, 03 2009. doi: 10.5194/nhess-9-405-2009.

J. H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29 (5):1189–1232, 2001. ISSN 00905364. URL http://www.jstor.org/stable/2699986.

C. Gatzen. A 10-year climatology of cold-season narrow cold-frontal rainbands in germany. *Atmospheric Research*, 100(4):366–370, 2011. ISSN 0169-8095. doi: 10.1016/j.atmosres.2010.09.018. URL https://www.sciencedirect.com/science/article/pii/S0169809510002565. 5th European Conference on Severe Storms.

C. P. Gatzen, A. H. Fink, D. M. Schultz, and J. G. Pinto. An 18-year climatology of derechos in germany. *Natural Hazards and Earth System Sciences*, 20(5):1335–1351, 2020. doi: 10.5194/nhess-20-1335-2020. URL https://nhess.copernicus.org/articles/20/1335/2020/.

P. Groenemeijer, T. Púčik, A. M. Holzer, B. Antonescu, K. Riemann-Campe, D. M. Schultz, T. Kühne, B. Feuerstein, H. E. Brooks, C. A. Doswell, H.-J. Koppert, and R. Sausen. Severe convective storms in europe: Ten years of research and education at the european severe storms laboratory. *Bulletin of the American Meteorological Society*, 98(12):2641 – 2651, 2017. doi: 10.1175/BAMS-D-16-0067.1. URL https://journals.ametsoc.org/view/journals/bams/98/12/bams-d-16-0067.1.xml.

S. Hagelin, J. Son, R. Swinbank, A. McCabe, N. Roberts, and W. Tennant. The met office convective-scale ensemble, mogreps-uk. *Quarterly Journal of the Royal Meteorological Society*, 143(708):2846–2861, 2017. doi: https://doi.org/10.1002/qj.3135. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3135.

T. Hastie and R. Tibshirani. Generalized additive models. *Statistical Science*, 1(3):297–310, 1986. ISSN 08834237. URL http://www.jstor.org/stable/2245459.

T. Hengstebeck, K. Wapler, D. Heizenreder, and P. Joe. Radar network–based detection of mesocyclones at the german weather service. *Journal of Atmospheric and Oceanic Technology*, 35(2):299 – 321, 2018. doi: 10.1175/JTECH-D-16-0230.1. URL https://journals.ametsoc.org/view/journals/atot/35/2/jtech-d-16-0230.1.xml.

H. Hersbach, B. Bell, P. Berrisford, G. Biavati, A. Horányi, J. Muñoz Sabater, J. Nicolas, C. Peubey, R. Radu, I. Rozum, D. Schepers, A. Simmons, C. Soci, D. Dee, and J.-N. Thépaut. Era5 hourly data on pressure levels from 1959 to present. *Copernicus Climate Change Service (C3S) Climate Data Store (CDS). (Accessed on 08–JAN-2021)*, 2018a. doi: 10.24381/cds.bd0915c6.

H. Hersbach, B. Bell, P. Berrisford, G. Biavati, A. Horányi, J. Muñoz Sabater, J. Nicolas, C. Peubey, R. Radu, I. Rozum, D. Schepers, A. Simmons, C. Soci, D. Dee, and J.-N. Thépaut. Era5 hourly data on single levels from 1959 to present. *Copernicus Climate Change Service (C3S) Climate Data Store (CDS). (Accessed on 08–JAN-2021)*, 2018b. doi: 10.24381/cds.bd0915c6.

H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, G. De Chiara, P. Dahlgren, D. Dee, M. Diamantakis, R. Dragani, J. Flemming, R. Forbes, M. Fuentes, A. Geer, L. Haimberger, S. Healy, R. J. Hogan, E. Hólm, M. Janisková, S. Keeley, P. Laloyaux, P. Lopez, C. Lupu, G. Radnoti, P. de Rosnay, I. Rozum, F. Vamborg, S. Villaume, and J.-N. Thépaut. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730): 1999–2049, 2020. doi: 10.1002/qj.3803. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803.

T. D. Hewson. Objective fronts. *Meteorological Applications*, 5(1):37–65, 1998. doi: 10.1017/S1350482798000553. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1017/S1350482798000553.

A. J. Hill, R. S. Schumacher, and I. L. Jirak. A new paradigm for medium-range severe weather forecasts: Probabilistic random forest–based predictions. *Weather and Forecasting*, 38(2):251 – 272, 2023. doi: 10.1175/WAF-D-22-0143.1. URL https://journals.ametsoc.org/view/journals/wefo/38/2/WAF-D-22-0143.1.xml.

M. Hirt, G. C. Craig, S. A. K. Schäfer, J. Savre, and R. Heinze. Cold-pool-driven convective initiation: using causal graph analysis to determine what convection-permitting models are missing. *Quarterly Journal of the Royal Meteorological Society*, 146(730):2205–2227, 2020. doi: https://doi.org/10.1002/qj.3788. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3788.

P. Hoeppe. Trends in weather related disasters – consequences for insurers and society. *Weather and Climate Extremes*, 11:70–79, 2016. ISSN 2212-0947. doi: 10.1016/j.wace.2015.10.002. URL https://www.sciencedirect.com/science/article/pii/S2212094715300347. Observed and Projected (Longer-term) Changes in Weather and Climate Extremes.

R. Hohl, H.-H. Schiesser, and D. Aller. Hailfall: the relationship between radar-derived hail kinetic energy and hail damage to buildings. *Atmospheric Research*, 63(3):177–207, 2002. ISSN 0169-8095. doi: 10.1016/S0169-8095(02)00059-5. URL https://www.sciencedirect.com/science/article/pii/S0169809502000595.

J. R. Holton. An introduction to dynamic meteorology: Chapter 3 elementary applications of the basic equations. volume 88 of *International Geophysics*, pages 57–85. Academic Press, 2004. doi: https://doi.org/10.1016/S0074-6142(04)80037-8. URL https://www.sciencedirect.com/science/article/pii/S0074614204800378.

J. R. Holton and G. J. Hakim. Chapter 6 - quasi-geostrophic analysis. In J. R. Holton and G. J. Hakim, editors, *An Introduction to Dynamic Meteorology (Fifth Edition)*, pages 171–211. Academic Press, Boston, fifth edition edition, 2013. ISBN 978-0-12-384866-6. doi: https://doi.org/10.1016/B978-0-12-384866-6.00006-4. URL https://www.sciencedirect.com/science/article/pii/B9780123848666000064.

B. J. Hoskins, I. Draghici, and H. C. Davies. A new look at the -equation. *Quarterly Journal of the Royal Meteorological Society*, 104(439):31–38, 1978. doi: https://doi.org/10.1002/qj.49710443903. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.49710443903.

A. Huuskonen, E. Saltikoff, and I. Holleman. The operational weather radar network in europe. *Bulletin of the American Meteorological Society*, 95(6):897 – 907, 2014. doi: 10.1175/BAMS-D-12-00216.1. URL https://journals.ametsoc.org/view/journals/bams/95/6/bams-d-12-00216.1.xml.

P. K. James and K. A. Browning. Mesoscale structure of line convection at surface cold fronts. *Quarterly Journal of the Royal Meteorological Society*, 105(444):371–382, 1979. doi: https://doi.org/10.1002/qj.49710544404. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.49710544404.

J. Jenkner, M. Sprenger, I. Schwenk, C. Schwierz, S. Dierer, and D. Leuenberger. Detection and climatology of fronts in a high-resolution model reanalysis over the alps. *Meteorological Applications*, 17(1):1–18, 2010. doi: 10.1002/met.142. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/met.142.

J. S. Kain, S. J. Weiss, D. R. Bright, M. E. Baldwin, J. J. Levit, G. W. Carbin, C. S. Schwartz, M. L. Weisman, K. K. Droegemeier, D. B. Weber, and K. W. Thomas. Some practical considerations regarding horizontal

resolution in the first generation of operational convection-allowing nwp. *Weather and Forecasting*, 23(5): 931 – 952, 2008. doi: 10.1175/WAF2007106.1. URL https://journals.ametsoc.org/view/journals/wefo/23/5/waf2007106_1.xml.

C. F. F. Karney. Algorithms for geodesics. *Journal of Geodesy*, 87(1):43–55, 2013. doi: 10.1007/s00190-012-0578-z.

F. Kaspar, F. Kratzenstein, and A. K. Kaiser-Weiss. Interactive open access to climate observations from germany. *Advances in Science and Research*, 16:75–83, 2019. doi: 10.5194/asr-16-75-2019. URL https://asr.copernicus.org/articles/16/75/2019/.

D. J. Kirshbaum, B. Adler, N. Kalthoff, C. Barthlott, and S. Serafin. Moist orographic convection: Physical mechanisms and links to surface-exchange processes. *Atmosphere*, 9(3), 2018. ISSN 2073-4433. doi: 10.3390/atmos9030080. URL https://www.mdpi.com/2073-4433/9/3/80.

C. Klasa, M. Arpagaus, A. Walser, and H. Wernli. An evaluation of the convection-permitting ensemble cosmo-e for three contrasting precipitation events in switzerland. *Quarterly Journal of the Royal Meteorological Society*, 144(712):744–764, 2018. doi: https://doi.org/10.1002/qj.3245. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3245.

S. E. Koch. The role of an apparent mesoscale frontogenetic circulation in squall line initiation. *Monthly Weather Review*, 112(10):2090 – 2111, 1984. doi: https://doi.org/10.1175/1520-0493(1984)112<2090:TROAAM>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/112/10/1520-0493_1984_112_2090_troaam_2_0_co_2.xml.

L. Kolendowicz, M. Taszarek, and B. Czernecki. Atmospheric circulation and sounding-derived parameters associated with thunderstorm occurrence in central europe. *Atmospheric Research*, 191:101–114, 2017. ISSN 0169-8095. doi: https://doi.org/10.1016/j.atmosres.2017.03.009. URL https://www.sciencedirect.com/science/article/pii/S0169809517302806.

B. Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, Nov 2016. ISSN 2192-6360. doi: 10.1007/s13748-016-0094-0. URL https://doi.org/10.1007/s13748-016-0094-0.

M. Kuhn and K. Johnson. *An Introduction to Feature Selection*, pages 487–519. Springer New York, New York, NY, 2013. ISBN 978-1-4614-6849-3. doi: 10.1007/978-1-4614-6849-3_19. URL https://doi.org/10.1007/978-1-4614-6849-3_19.

M. Kunz and P. I. Kugel. Detection of hail signatures from single-polarization c-band radar reflectivity. *Atmospheric Research*, 153:565–577, 2015. ISSN 0169-8095. doi: 10.1016/j.atmosres.2014.09.010. URL https://www.sciencedirect.com/science/article/pii/S0169809514003676.

M. Kunz, U. Blahak, J. Handwerker, M. Schmidberger, H. J. Punge, S. Mohr, E. Fluck, and K. M. Bedka. The severe hailstorm in southwest germany on 28 july 2013: characteristics, impacts and meteorological conditions. *Quarterly Journal of the Royal Meteorological Society*, 144(710):231–250, 2018. doi: 10.1002/qj.3197. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3197.

M. Kunz, J. Wandel, E. Fluck, S. Baumstark, S. Mohr, and S. Schemm. Ambient conditions prevailing during hail events in central europe. *Natural Hazards and Earth System Sciences*, 20(6):1867–1887, 2020. doi: 10.5194/nhess-20-1867-2020. URL https://nhess.copernicus.org/articles/20/1867/2020/.

R. Lam, A. Sanchez-Gonzalez, M. Willson, P. Wirnsberger, M. Fortunato, F. Alet, S. Ravuri, T. Ewalds, Z. Eaton-Rosen, W. Hu, A. Merose, S. Hoyer, G. Holland, O. Vinyals, J. Stott, A. Pritzel, S. Mohamed, and P. Battaglia. Graphcast: Learning skillful medium-range global weather forecasting, 2023.

K.-O. Lee, C. Flamant, V. Ducrocq, F. Duffourg, N. Fourrié, and S. Davolio. Convective initiation and maintenance processes of two back-building mesoscale convective systems leading to heavy precipitation events in southern italy during hymex iop 13. *Quarterly Journal of the Royal Meteorological Society*, 142 (700):2623–2635, 2016. doi: 10.1002/qj.2851. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.2851.

J. Leinonen, U. Hamann, U. Germann, and J. R. Mecikalski. Nowcasting thunderstorm hazards using machine learning: the impact of data sources on performance. *Natural Hazards and Earth System Sciences*, 22(2): 577–597, 2022. doi: 10.5194/nhess-22-577-2022. URL https://nhess.copernicus.org/articles/22/577/2022/.

M. W. Lewis and S. L. Gray. Categorisation of synoptic environments associated with mesoscale convective systems over the uk. *Atmospheric Research*, 97(1):194–213, 2010. ISSN 0169-8095. doi: https://doi.org/10.1016/j.atmosres.2010.04.001. URL https://www.sciencedirect.com/science/article/pii/S0169809510000888.

J. F. León-Cruz, E. Caetano, J. Cortés-Ramos, C. Dominguez, and J. M. Méndez-Pérez. Thunderstorm and hailstorm environments in mexico. *International Journal of Climatology*, 43(10):4379–4395, 2023. doi: https://doi.org/10.1002/joc.8093. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.8093.

M. A. Lima and J. W. Wilson. Convective storm initiation in a moist tropical environment. *Monthly Weather Review*, 136(6):1847 – 1864, 2008. doi: 10.1175/2007MWR2279.1. URL https://journals.ametsoc.org/view/journals/mwre/136/6/2007mwr2279.1.xml.

F. H. Ludlam. Clouds and storms: The behavior and effect of water in the atmosphere., 1980.

P. Markowski and Y. Richardson. *Mesoscale Meteorology in Midlatitudes*, volume 2 of *Advancing weather and climate science*. Wiley, Somerset, 1. aufl. edition, 2010. ISBN 0470742135.

B. J. Mason. *THE PHYSICS OF CLOUDS. 2ND ED.* CLARENDON PR, 1971.

R. M. May, K. H. Goebbert, J. E. Thielen, J. R. Leeman, M. D. Camron, Z. Bruick, E. C. Bruning, R. P. Manser, S. C. Arms, and P. T. Marsh. Metpy: A meteorological python library for data analysis and visualization. *Bulletin of the American Meteorological Society*, 103(10):E2273 – E2284, 2022. doi: 10.1175/BAMS-D-21-0125.1. URL https://journals.ametsoc.org/view/journals/bams/103/10/BAMS-D-21-0125.1.xml.

J. R. Mecikalski, T. N. Sandmæl, E. M. Murillo, C. R. Homeyer, K. M. Bedka, J. M. Apke, and C. P. Jewett. A random-forest model to assess predictor importance and nowcast severe storms using high-resolution radar–goes satellite–lightning observations. *Monthly Weather Review*, 149(6):1725 – 1746, 2021.

doi: 10.1175/MWR-D-19-0274.1. URL https://journals.ametsoc.org/view/journals/mwre/149/6/MWR-D-19-0274.1.xml.

Met Office. Lightning strike location data. https://www.metoffice.gov.uk/binaries/content/assets/metofficegovuk/pdf/data/adtnet_data_sheet.pdf, 2020. Met Office, 3 pp.

S. Mohr, U. Ehret, M. Kunz, P. Ludwig, A. Caldas-Alvarez, J. E. Daniell, F. Ehmele, H. Feldmann, M. J. Franca, C. Gattke, M. Hundhausen, P. Knippertz, K. Küpfer, B. Mühr, J. G. Pinto, J. Quinting, A. M. Schäfer, M. Scheibel, F. Seidel, and C. Wisotzky. A multi-disciplinary analysis of the exceptional flood event of july 2021 in central europe – part 1: Event description and analysis. *Natural Hazards and Earth System Sciences*, 23(2):525–551, 2023. doi: 10.5194/nhess-23-525-2023. URL https://nhess.copernicus.org/articles/23/525/2023/.

C. Morel and S. Senesi. A climatology of mesoscale convective systems over europe using satellite infrared imagery. ii: Characteristics of european mesoscale convective systems. *Quarterly Journal of the Royal Meteorological Society*, 128(584):1973–1995, 2002. doi: 10.1256/003590002320603494. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1256/003590002320603494.

H. Morrison, J. M. Peters, K. K. Chandrakar, and S. C. Sherwood. Influences of environmental relative humidity and horizontal scale of subcloud ascent on deep convective initiation. *Journal of the Atmospheric Sciences*, 79(2):337 – 359, 2022. doi: 10.1175/JAS-D-21-0056.1. URL https://journals.ametsoc.org/view/journals/atsc/79/2/JAS-D-21-0056.1.xml.

Munich Re. Record thunderstorm losses and deadly earthquakes: the natural disasters of 2023, 2024. URL https://www.munichre.com/en/company/media-relations/media-information-and-corporate-news/media-information/2024/natural-disaster-figures-2023.html.

J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972. ISSN 00359238. URL http://www.jstor.org/stable/2344614.

B. R. Nelson, D.-J. Seo, and D. Kim. Multisensor precipitation reanalysis. *Journal of Hydrometeorology*, 11(3):666 – 682, 2010. doi: https://doi.org/10.1175/2010JHM1210.1. URL https://journals.ametsoc.org/view/journals/hydr/11/3/2010jhm1210_1.xml.

G. P. Pacey, D. M. Schultz, and L. Garcia-Carreras. Severe convective windstorms in europe: Climatology, preconvective environments, and convective mode. *Weather and Forecasting*, 36(1):237 – 252, 2021. doi: 10.1175/WAF-D-20-0075.1. URL https://journals.ametsoc.org/view/journals/wefo/36/1/WAF-D-20-0075.1.xml.

J. Pathak, S. Subramanian, P. Harrington, S. Raja, A. Chattopadhyay, M. Mardani, T. Kurth, D. Hall, Z. Li, K. Azizzadenesheli, P. Hassanzadeh, K. Kashinath, and A. Anandkumar. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators, 2022.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

J. M. Peters, W. Hannah, and H. Morrison. The influence of vertical wind shear on moist thermals. *Journal of the Atmospheric Sciences*, 76(6):1645 – 1659, 2019. doi: https://doi.org/10.1175/JAS-D-18-0296.1. URL https://journals.ametsoc.org/view/journals/atsc/76/6/jas-d-18-0296.1.xml.

J. M. Peters, H. Morrison, T. C. Nelson, J. N. Marquis, J. P. Mulholland, and C. J. Nowotarski. The influence of shear on deep convection initiation. part i: Theory. *Journal of the Atmospheric Sciences*, 79(6):1669 – 1690, 2022. doi: https://doi.org/10.1175/JAS-D-21-0145.1. URL https://journals.ametsoc.org/view/journals/atsc/79/6/JAS-D-21-0145.1.xml.

D. Piper and M. Kunz. Spatiotemporal variability of lightning activity in europe and the relation to the north atlantic oscillation teleconnection pattern. *Natural Hazards and Earth System Sciences*, 17(8): 1319–1336, 2017. doi: 10.5194/nhess-17-1319-2017. URL https://nhess.copernicus.org/articles/17/1319/2017/.

D. Piper, M. Kunz, F. Ehmele, S. Mohr, B. Mühr, A. Kron, and J. Daniell. Exceptional sequence of severe thunderstorms and related flash floods in may and june 2016 in germany – part 1: meteorological background. *Natural Hazards and Earth System Sciences*, 16(12):2835–2850, 2016. doi: 10.5194/nhess-16-2835-2016. URL https://nhess.copernicus.org/articles/16/2835/2016/.

S. Poręba, M. Taszarek, and Z. Ustrnul. Diurnal and seasonal variability of era5 convective parameters in relation to lightning flash rates in poland. *Weather and Forecasting*, 37(8):1447 – 1470, 2022. doi: https://doi.org/10.1175/WAF-D-21-0099.1. URL https://journals.ametsoc.org/view/journals/wefo/37/8/WAF-D-21-0099.1.xml.

T. Púčik, P. Groenemeijer, D. Rýva, and M. Kolář. Proximity soundings of severe and non-severe thunderstorms in central europe. *Monthly Weather Review*, 143(12):4805 – 4821, 2015. doi: 10.1175/MWR-D-15-0104.1. URL https://journals.ametsoc.org/view/journals/mwre/143/12/mwr-d-15-0104.1.xml.

D. Reinert, F. Prill, H. Frank, M. Denhard, M. Baldauf, C. Schraff, C. Gebhardt, C. Marsigli, and G. Zängl. Dwd database reference for the global and regional icon and icon-eps forecasting system. *https://www.dwd. de/DWD/forschung/nwv/fepub/icon_ database_ main. pdf (accessed March 2024)*, 2023.

R. J. Renard and L. Clarke. Experiments in numerical objective frontal analysis1. *Monthly Weather Review*, 93:547–556, 1965.

S. Rüdisühli, M. Sprenger, D. Leutwyler, C. Schär, and H. Wernli. Attribution of precipitation to cyclones and fronts over europe in a kilometer-scale regional climate simulation. *Weather and Climate Dynamics*, 1 (2):675–699, 2020.

A. T. Rädler, P. Groenemeijer, E. Faust, and R. Sausen. Detecting severe weather trends using an additive regressive convective hazard model (ar-chamo). *J. Appl. Meteor. Climatol.*, 57:569–587, 2019.

F. Sanders. An investigation of the structure and dynamics of an intense surface frontal zone. *Journal of Atmospheric Sciences*, 12(6):542 – 552, 1955. doi: 10.1175/1520-0469(1955)012<0542:AIOTSA>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/atsc/12/6/1520-0469_1955_012_0542_aiotsa_2_0_co_2.xml.

S. Schemm, I. Rudeva, and I. Simmonds. Extratropical fronts in the lower troposphere–global perspectives obtained from two automated methods. *Quarterly Journal of the Royal Meteorological Society*, 141(690): 1686–1698, 2015. doi: 10.1002/qj.2471. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.2471.

S. Schemm, L. Nisi, A. Martinov, D. Leuenberger, and O. Martius. On the link between cold fronts and hail in switzerland. *Atmospheric Science Letters*, 17(5):315–325, 2016.

S. Schemm, M. Sprenger, and H. Wernli. When during their life cycle are extratropical cyclones attended by fronts? *Bulletin of the American Meteorological Society*, 99(1):149 – 165, 2018. doi: 10.1175/BAMS-D-16-0261.1. URL https://journals.ametsoc.org/view/journals/bams/99/1/bams-d-16-0261.1.xml.

D. M. Schultz, Y. P. Richardson, P. M. Markowski, and C. A. Doswell. Tornadoes in the central united states and the. *Bulletin of the American Meteorological Society*, 95(11):1704 – 1712, 2014. doi: 10.1175/BAMS-D-13-00252.1. URL https://journals.ametsoc.org/view/journals/bams/95/11/bams-d-13-00252.1.xml.

R. S. Schumacher, A. J. Hill, M. Klein, J. A. Nelson, M. J. Erickson, S. M. Trojniak, and G. R. Herman. From random forests to flood forecasts: A research to operations success story. *Bulletin of the American Meteorological Society*, 102(9):E1742 – E1755, 2021. doi: 10.1175/BAMS-D-20-0186.1. URL https://journals.ametsoc.org/view/journals/bams/102/9/BAMS-D-20-0186.1.xml.

R. S. Schumacher, S. J. Childs, and R. D. Adams-Selin. Intense surface winds from gravity wave breaking in simulations of a destructive macroburst. *Monthly Weather Review*, 2022. doi: 10.1175/MWR-D-22-0103.1. URL https://journals.ametsoc.org/view/journals/mwre/aop/MWR-D-22-0103.1/MWR-D-22-0103.1.xml.

U. Schumann. Influence of mesoscale orography on idealized cold fronts. *Journal of Atmospheric Sciences*, 44(23):3423 – 3441, 1987. doi: 10.1175/1520-0469(1987)044<3423:IOMOOI>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/atsc/44/23/1520-0469_1987_044_3423_iomooi_2_0_co_2.xml.

B. T. Smith, R. L. Thompson, J. S. Grams, C. Broyles, and H. E. Brooks. Convective modes for significant severe thunderstorms in the contiguous united states. part i: Storm classification and climatology. *Weather and Forecasting*, 27(5):1114 – 1135, 2012. doi: https://doi.org/10.1175/WAF-D-11-00115.1. URL https://journals.ametsoc.org/view/journals/wefo/27/5/waf-d-11-00115_1.xml.

A. Surowiecki and M. Taszarek. A 10-year radar-based climatology of mesoscale convective system archetypes and derechos in poland. *Monthly Weather Review*, 148(8):3471 – 3488, 2020. doi: 10.1175/MWR-D-19-0412.1. URL https://journals.ametsoc.org/view/journals/mwre/148/8/mwrD190412.xml.

M. Taszarek, H. E. Brooks, and B. Czernecki. Sounding-derived parameters associated with convective hazards in europe. *Monthly Weather Review*, 145(4):1511 – 1528, 2017. doi: 10.1175/MWR-D-16-0384.1. URL https://journals.ametsoc.org/view/journals/mwre/145/4/mwr-d-16-0384.1.xml.

M. Taszarek, H. E. Brooks, B. Czernecki, P. Szuster, and K. Fortuniak. Climatological aspects of convective parameters over europe: A comparison of era-interim and sounding data. *Journal of Climate*, 31(11):4281 – 4308, 2018. doi: 10.1175/JCLI-D-17-0596.1. URL https://journals.ametsoc.org/view/journals/clim/31/11/jcli-d-17-0596.1.xml.

M. Taszarek, J. Allen, T. Púčik, P. Groenemeijer, B. Czernecki, L. Kolendowicz, K. Lagouvardos, V. Kotroni, and W. Schulz. A climatology of thunderstorms across europe from a synthesis of multiple data sources. *Journal of Climate*, 32(6):1813 – 1837, 2019. doi: 10.1175/JCLI-D-18-0372.1. URL https://journals.ametsoc.org/view/journals/clim/32/6/jcli-d-18-0372.1.xml.

M. Taszarek, J. T. Allen, P. Groenemeijer, R. Edwards, H. E. Brooks, V. Chmielewski, and S.-E. Enno. Severe convective storms across europe and the united states. part i: Climatology of lightning, large hail, severe wind, and tornadoes. *Journal of Climate*, 33(23):10239 – 10261, 2020a. doi: 10.1175/JCLI-D-20-0345.1. URL https://journals.ametsoc.org/view/journals/clim/33/23/jcliD200345.xml.

M. Taszarek, J. T. Allen, T. Púčik, K. A. Hoogewind, and H. E. Brooks. Severe convective storms across europe and the united states. part ii: Era5 environments associated with lightning, large hail, severe wind, and tornadoes. *Journal of Climate*, 33(23):10263 – 10286, 2020b. doi: https://doi.org/10.1175/JCLI-D-20-0346.1. URL https://journals.ametsoc.org/view/journals/clim/33/23/jcliD200346.xml.

M. Taszarek, B. Czernecki, and P. Szuster. thunder - a rawinsonde package for processing convective parameters and visualizing atmospheric profiles. 05 2023. doi: 10.5194/ecss2023-28.

M. Theusner and T. Hauf. A study on the small scale precipitation structure over germany using the radar network of the german weather service. *Meteorologische Zeitschrift*, 13(4):311–322, 2004. doi: 10.1127/0941-2948/2004/0013-0311.

C. M. Thomas and D. M. Schultz. What are the best thermodynamic quantity and function to define a front in gridded model output? *Bulletin of the American Meteorological Society*, 100(5):873 – 895, 2019. doi: 10.1175/BAMS-D-18-0137.1. URL https://journals.ametsoc.org/view/journals/bams/100/5/bams-d-18-0137.1.xml.

R. J. Trapp. *The Initiation of Deep Convective Clouds*, page 121–161. Cambridge University Press, 2013. doi: 10.1017/CBO9781139047241.006.

M. S. van den Broeke, D. M. Schultz, R. H. Johns, J. S. Evans, and J. E. Hales. Cloud-to-ground lightning production in strongly forced, low-instability convective lines associated with damaging wind. *Weather and Forecasting*, 20(4):517 – 530, 2005. doi: 10.1175/WAF876.1. URL https://journals.ametsoc.org/view/journals/wefo/20/4/waf876_1.xml.

K. Wapler. The life-cycle of hailstorms: Lightning, radar reflectivity and rotation characteristics. *Atmospheric Research*, 193:60–72, 2017. ISSN 0169-8095. doi: 10.1016/j.atmosres.2017.04.009. URL https://www.sciencedirect.com/science/article/pii/S0169809516306020.

K. Wapler. Mesocyclonic and non-mesocyclonic convective storms in germany: Storm characteristics and life-cycle. *Atmospheric Research*, 248:105186, 2021. ISSN 0169-8095. doi: 10.1016/j.atmosres.2020.105186. URL https://www.sciencedirect.com/science/article/pii/S0169809520311224.

K. Wapler and P. James. Thunderstorm occurrence and characteristics in central europe under different synoptic conditions. *Atmospheric Research*, 158-159:231–244, 2015. ISSN 0169-8095. doi: https://doi.org/10.1016/j.atmosres.2014.07.011. URL https://www.sciencedirect.com/science/article/pii/S0169809514002749.

K. Wapler, M. Goeber, and S. Trepte. Comparative verification of different nowcasting systems to support optimisation of thunderstorm warnings. *Advances in Science and Research*, 8(1):121–127, 2012. doi: 10.5194/asr-8-121-2012. URL https://asr.copernicus.org/articles/8/121/2012/.

K. Wapler, T. Hengstebeck, and P. Groenemeijer. Mesocyclones in central europe as seen by radar. *Atmospheric Research*, 168:112–120, 2016. ISSN 0169-8095. doi: 10.1016/j.atmosres.2015.08.023. URL https://www.sciencedirect.com/science/article/pii/S0169809515002719.

M. Werner. Konrad3d: A new tool for detection and nowcasting of convective cells at dwd. *Second European Nowcasting Conference, Offenbach, Germany*, 2017. URL http://eumetnet.eu/wp-content/uploads/2017/07/enc_book_of_abstract.pdf.

A. T. Westermayer, P. Groenemeijer, G. Pistotnik, R. Sausen, and E. Faust. Identification of favorable environments for thunderstorms in reanalysis data. *Meteorologische Zeitschrift*, 26(1):59–70, 02 2017. doi: 10.1127/metz/2016/0754. URL http://dx.doi.org/10.1127/metz/2016/0754.

T. Weusthoff and T. Hauf. The life cycle of convective-shower cells under post-frontal condition. *Quarterly Journal of the Royal Meteorological Society*, 134(633):841–857, 2008. doi: 10.1002/qj.260.

J. Wilhelm, S. Mohr, H. J. Punge, B. Mühr, M. Schmidberger, J. E. Daniell, K. M. Bedka, and M. Kunz. Severe thunderstorms with large hail across germany in june 2019. *Weather*, 76(7):228–237, 2021. doi: 10.1002/wea.3886. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/wea.3886.

D. S. Wilks. *Statistical methods in the atmospheric sciences*. International Geophysics Series, 2nd edition, 2006.

D. S. Wilks. Chapter 9 - forecast verification. In D. S. Wilks, editor, *Statistical Methods in the Atmospheric Sciences (Fourth Edition)*, pages 369–483. Elsevier, fourth edition edition, 2019. ISBN 978-0-12-815823-4. doi: https://doi.org/10.1016/B978-0-12-815823-4.00009-2. URL https://www.sciencedirect.com/science/article/pii/B9780128158234000092.

J. W. Wilson and W. E. Schreiber. Initiation of convective storms at radar-observed boundary-layer convergence lines. *Monthly Weather Review*, 114(12):2516 – 2536, 1986. doi: 10.1175/1520-0493(1986)114<2516:IOCSAR>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/114/12/1520-0493_1986_114_2516_iocsar_2_0_co_2.xml.

V. Wulfmeyer, A. Behrendt, C. Kottmeier, U. Corsmeier, C. Barthlott, G. C. Craig, M. Hagen, D. Althausen, F. Aoshima, M. Arpagaus, H.-S. Bauer, L. Bennett, A. Blyth, C. Brandau, C. Champollion, S. Crewell, G. Dick, P. Di Girolamo, M. Dorninger, Y. Dufournet, R. Eigenmann, R. Engelmann, C. Flamant, T. Foken, T. Gorgas, M. Grzeschik, J. Handwerker, C. Hauck, H. Höller, W. Junkermann, N. Kalthoff, C. Kiemle, S. Klink, M. König, L. Krauss, C. N. Long, F. Madonna, S. Mobbs, B. Neininger, S. Pal, G. Peters, G. Pigeon, E. Richard, M. W. Rotach, H. Russchenberg, T. Schwitalla, V. Smith, R. Steinacker, J. Trentmann,

D. D. Turner, J. van Baelen, S. Vogt, H. Volkert, T. Weckwerth, H. Wernli, A. Wieser, and M. Wirth. The convective and orographically-induced precipitation study (cops): the scientific strategy, the field phase, and research highlights. *Quarterly Journal of the Royal Meteorological Society*, 137(S1):3–30, 2011. doi: 10.1002/qj.752. URL `https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.752`.

J.-I. Yano, M. Z. Ziemiański, M. Cullen, P. Termonia, J. Onvlee, L. Bengtsson, A. Carrassi, R. Davy, A. Deluca, S. L. Gray, V. Homar, M. Köhler, S. Krichak, S. Michaelides, V. T. J. Phillips, P. M. M. Soares, and A. A. Wyszogrodzki. Scientific challenges of convective-scale numerical weather prediction. *Bulletin of the American Meteorological Society*, 99(4):699 – 710, 2018. doi: 10.1175/BAMS-D-17-0125.1. URL `https://journals.ametsoc.org/view/journals/bams/99/4/bams-d-17-0125.1.xml`.

Acknowledgements

First and foremost I would like to thank Prof. Dr. Stephan Pfahl and Vertr.-Prof. Dr. Lisa Schielicke as my supervisors during this PhD project. I have been impressed by their regular availability for meetings and consistent support throughout my PhD journey. I would also like to thank all members of the Weather and Climate Processes group at FU Berlin for their useful comments and insight on my research as well as feedback on posters and presentations for upcoming conferences. Thanks again to Prof. Dr. Stephan Pfahl as well as Prof. Dr. Michael Kunz for being the examiners of this thesis.

This research would not have been possible without funding from Deutsche Forschungsgemeinschaft through grant CRC 1114 "Scaling Cascades in Complex Systems, Project Number 235221301" under the project C06 "Multi-scale structure of atmospheric vortices". Being part of a CRC also allowed me to connect with several other PhD students from different fields and attend several personal development training sessions. The funding also allowed me to attend several international conferences and workshops, including one in the United States. The discussions and knowledge gained during these conferences and workshops enabled me to push the boundaries of my research further.

There are several other people to thank, all of which I cannot name. Thanks to Dr. Kathrin Wapler for providing support with the KONRAD cell tracking and detection dataset. Thanks to Dr. Stefan Rüdisühli for the discussions and insight regarding automatic front detection methods. Thanks to Assoc. Prof. Dr. Johannes Dahl for useful discussions and interpretations of certain results. Thanks to Prof. Dr. Henning Rust and Dr. Nico Becker for helpful discussions regarding statistical modelling. Thank you to all secretaries and administration staff who have assisted with processing travel applications and other administration tasks. Last but not least, thank you to all my family and friends who have supported me through my academic journey, especially those outside of Germany who supported my decision to live abroad and pursue a PhD.

Declaration of originality

I hereby declare that the report submitted is my own unaided work. All direct or indirect sources used are acknowledged as references. I am aware that the thesis in digital form can be examined for the use of unauthorized aid and in order to determine whether the report as a whole or parts incorporated in it may be deemed as plagiarism. For the comparison of my work with existing sources I agree that it shall be entered in a database where it shall also remain after examination, to enable comparison with future theses submitted. Further rights of reproduction and usage, however, are not granted here. This paper was not previously presented to another examination board and has not been published.