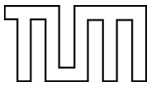

Technische Universität München
Department of Mathematics

Master's Thesis

# Development of an enhanced additive logistic regression model for European thunderstorms and their associated hazards

Emma Allwright

Supervisor: Prof. Dr. Claudia Czado
Advisors: Prof. Dr. Claudia Czado & Özge Sahin

Submission Date: 30.01.2021

I hereby declare that this thesis is my own work and that no other sources have been used except those clearly indicated and referenced.

Munich,

Emma Allwright

# Abstract

The potential cost associated with damage caused by severe hail storms is becoming an increasingly considered aspect of natural catastrophe modelling in the insurance industry. To calculate this potential cost we need to understand both the damage an event would cause, and the probability of said event. This thesis focuses on the latter and uses recently released meteorological data to develop a regression model for the probability of severe hail.

We first discuss the meteorological theory of the development of thunderstorms and how severe hail grows, followed by the necessary statistical theory to analyse and fit regression models to our data. We then preprocess and perform a statistical analysis on our data. A generalised additive model and logistic model are fit for the probability of a thunderstorm and severe hail given that there is a thunderstorm respectively. These two models can then be multiplied to calculate the probability of severe hail. We compare our two separate models to those from another study to verify our results.

# Acknowledgements

I would first and foremost like to thank Prof. Claudia Czado for her supervision throughout my thesis. Her help, feedback and guidance made this possible and I am immensely grateful for the time she has invested in this project. I am also deeply thankful for the support and feedback Özge Sahin has provided. Thirdly I would like to thank Dr. Anja Rädler for her support and advice, especially with the meteorological and coding parts of this thesis. Together I could not have asked for a better team to supervise me and I am truly grateful for this experience.
I would also like to thank Munich RE for providing this topic, the European Centre for Medium Range Weather Forecasts and the European Severe Storms Laboratory, including the European Severe Weather Database, for providing the data used in this thesis, and of course, my friends and family for their support throughout this journey.

# Contents

# 1 Introduction

Severe hail is a phenomenon which threatens both life and property [26], so it is of interest to the public, meteorologists and the insurance industry to know when and where severe hail will occur. The growth of hail is a microphysical process which occurs on a very small scale within a severe convective storm (thunderstorm). The scale on which thunderstorms occur is large enough that they can be resolved and numerically modelled in high resolution weather forecasts, as seen in [13] and [9]. However, neither severe hail nor thunderstorms develop on a scale which can be numerically modeled using most historical data sets, including the ERA5 reanalysis data set used in this thesis, as discussed in [27].
We will use a probabilistic method to model severe hail. Some current statistical models used to calculate the probability of severe hail focus on the long-term applicability of the model, for example in [23], where one of their foci is the long-term variability of the probability of hail over a time period of 60 years, or [28], which focuses on applying their model to future climate scenarios. This thesis, however, uses data only from the years 2007 to 2018. We seek to improve upon current statistical models focusing on the current climate, such as [27], using recently released, more detailed data to fit a logistic model for the probability of severe hail. This is calculated as in Equation (1.0.1) by multiplying the probability of a thunderstorm by the probability of severe hail, given that there is a thunderstorm:

$$p(hail) = p(storm) \times p(hail|storm). \tag{1.0.1}$$

To determine the probabilities of a thunderstorm and severe hail during a thunderstorm, we will fit two regression models.
This thesis is structured in the following manner: firstly, Chapter 2 will give an overview of the physics governing the initiation of thunderstorms, the growth of hail and the reasoning behind why we have chosen to investigate the considered covariates. Chapter 3 then discusses the required theory for the binary regression models applied later in the thesis, including the fitting of a generalised linear model (GLM), specifically the case of logistic regression, overdispersion and the fitting of a generalised additive model (GAM). Chapter 4 then discusses the method used to preprocess the data so that it is in a form ready for analysis. In Chapter 5 the exploratory data analysis is performed. For both the thunderstorm and hail models we propose a logistic model and a GAM, then analyse the goodness of fit of each of these to select the most appropriate models. Finally we compare predicted values from our selected models to predicted values from the models presented in [28].
This thesis has been written together in partnership with Munich RE, and it is a major goal that in the future the work from this thesis will be applied by the meteorologists with whom we worked to aid in the understanding and modelling of the damages caused by severe hail events in Europe.

# 2 Meteorological Background

This section provides a brief meteorological background to the variables used in the model, thunderstorm initiation and hail formation.

## 2.1 Meteorology Glossary

A small dictionary is provided containing frequently used technical meteorological terms and abbreviations.

**Severe hail**: Hail stones with a diameter of at least 2 cm.

**Thunderstorm**: A severe convective storm with cloud to ground lightning.

**Convection**: Refers to the transport of a quantity, typically heat, due to fluid or gas rising or sinking, for further detail see Page 1 of [5]. In the context of a storm, this is due to warm air rising in an 'updraft' and cooler, denser air sinking in a 'downdraft' [5].

**Orography**: The terrain of a region, for example, mountains, hills or valleys. The height is reported in metres above sea level [10].

**Sea breeze**: Onshore wind which develops due to the differences in heating between a landmass and body of water. Warm, inland air rises and is replaced by cooler air from over the ocean, for further detail see Page 90 of [12].

**Frontal system**: The transition zone between two air masses which have different densities is called a front, Page 298, [1]. This can be caused, for example, by different temperatures or humidities of the air masses as seen in Section 8.1.2 of [32]. It is interesting to note that increased air humidity will decrease the density of the air, because water has a smaller molecular mass than air.

**Air parcel**: A volume of air which weighs 1 kg (approximately $0.77\,\mathrm{m}^3$).

**Free convection layer** [m]: The height in the atmosphere at which a parcel of air lifted from the ground has positive buoyancy relative to the environment conditions, discussed in Section 3.1 of [20]. This means that the parcel of air, when lifted to this height, would rise in relation to its surrounding environment.

**Equilibrium layer** [m]: The height in the atmosphere at which a parcel of air lifted from the ground has 0 buoyancy, meaning that there is no longer a force causing the parcel to rise, discussed in Section 3.1 of [20].

$\mathbf{u_{sfc}}$, $\mathbf{u}_{6\,\mathrm{km}}$ $[\mathrm{m\,s^{-1}}]$: The eastward component of the windspeed at surface level and at 6 km above ground level. A negative value represents a westward direction.

$\mathbf{v_{sfc}}$, $\mathbf{v_{6km}}$ $[\mathrm{m\,s^{-1}}]$: The northward component of the windspeed at surface level and at 6 km above ground level. A negative value represents a southward direction.

**ESWD**: European Severe Weather Database.[1]

**ECMWF**: European Centre for Medium Range Weather Forecasts.[2]

[1] `www.eswd.eu`

[2] `www.ecmwf.int`

**EUCLID**: European Cooperation for Lightning Detection.[3]

**Resolution** (of a data set): The distance between measurements, measured in degrees for latitude and longitude, and hours for time.

**Reanalysis data**: A combination of weather forecasts and observations used to give a best estimate for historic weather conditions [11].

**ERA5**: Released in 2020, this is the $5^{th}$ generation of reanalysis global climate and weather data produced by the European Centre for Medium Range Weather Forecasts (ECMWF). Data is available dating back to 1950 and has a resolution in (time, latitude, longitude) of (hourly, 0.25° and 0.25°). The data also contains 37 pressure levels from 1000 hPa to 1 hPa, see [8] and [7].

**ERA-interim**: This is the $4^{th}$ generation of reanalysis climate and weather data released by the ECMWF, with resolution in (time, latitude, longitude) of (6 hourly, 0.75°, 0.75°). The data also contains 37 pressure levels from 1000 hPa to 1 hPa [6].

**Grid box**: A one hour by 0.25° by 0.25° 'box' representing the conditions at this point in time and space.

**lat, lon** : latitude, longitude.

**Case**: A case is a gridbox which has at least one report of hail, or one cloud to ground lightning detection, for hail cases and lightning cases respectively.

**Data point**: Used as a synonym for a statistical observation. The word 'observation' has distinct and different meanings in statistics and meteorology, thus 'data point' is used to avoid confusion as to the context in which the word is meant.

## 2.2 Thunderstorm Initiation

Thunderstorms are a spectacular example of severe weather phenomena. To the insurance industry however, these events represent a large potential loss due to their associated hazards. The focus of this thesis, severe hail, is associated with both crop damage and damage to buildings and motor vehicles as discussed and shown in the introduction and Table 1 of [26]. This table of hail stone size and the frequency of damage types is shown in Figure 1.
To develop, thunderstorms require moisture in the atmosphere at ground level, instability in the atmosphere and a trigger mechanism as discussed in Section 1.1.b of [5]. Instability is measured by the difference in temperature at different heights in the atmosphere. A trigger mechanism creates 'lift', where the elevation of a parcel of air is increased relative to its surrounding environment. Examples of trigger mechanisms are sea breezes, frontal systems, heating caused by the sun and orography, see Section 2.1 of [28]. The effect of orography is clearly visible in Figure 2 which shows many more thunderstorms in closer proximity to the European Alps in the lower region of the figure.
Additionally, the type and longevity of thunderstorms is affected by deep layer wind shear (DLS), as discussed in Section 3 of [24]. DLS is explained in further detail in Section 2.5. The following paragraphs describe the different types of convection in thunderstorms, following Section 2.1.6 of [28].

[3] `www.euclid.org`

TABLE 1. Number of reports containing description of damage to different types of objects and with specified hail size.

| Hail size (cm) | Farmland gardens | Trees | Greenhouses | Roofs | House windows |
|---|---|---|---|---|---|
| <2 | 18 | 2 | 0 | 0 | 0 |
| 2–2.9 | 326 | 90 | 5 | 1 | 0 |
| 3–3.9 | 584 | 277 | 40 | 59 | 0 |
| 4–4.9 | 392 | 193 | 16 | 103 | 27 |
| 5–5.9 | 158 | 47 | 8 | 143 | 47 |
| 6–6.9 | 53 | 12 | 9 | 88 | 31 |
| 7–7.9 | 37 | 15 | 8 | 56 | 25 |
| 8–8.9 | 7 | 4 | 1 | 25 | 15 |
| 9–9.9 | 3 | 2 | 2 | 15 | 7 |
| >10 | 6 | 5 | 2 | 22 | 14 |
| Total | 1584 | 647 | 91 | 512 | 166 |
| Hail size (cm) | Vehicle body | Vehicle windows | Small animals killed | Large animals killed | People injured |
| <2 | 0 | 0 | 0 | 0 | 0 |
| 2–2.9 | 17 | 0 | 0 | 0 | 3 |
| 3–3.9 | 50 | 0 | 3 | 0 | 5 |
| 4–4.9 | 64 | 0 | 2 | 1 | 5 |
| 5–5.9 | 104 | 7 | 2 | 4 | 14 |
| 6–6.9 | 58 | 19 | 2 | 4 | 8 |
| 7–7.9 | 54 | 39 | 4 | 3 | 8 |
| 8–8.9 | 19 | 16 | 5 | 5 | 7 |
| 9–9.9 | 11 | 12 | 0 | 1 | 1 |
| >10 | 19 | 17 | 0 | 1 | 14 |
| Total | 396 | 110 | 18 | 19 | 65 |

Figure 1. This figure is from [26], and shows the type of damage caused by hail, the size of hail and how many times this type of event was reported.

- The most short lived thunderstorms have **singlecellular** convection and are typically associated with small values of deep layer wind shear. Singlecellular thunderstorms are usually short lived because the downdraft of air from higher in the atmosphere falls into the updraft, thus the thunderstorm loses its supply of warm, moist air and decays, see Section 2.1.6.1 of [28].

- Moderate values of deep layer wind shear increase the chance of **multicellular** convection occurring. This is the result of the storm becoming tilted and the downdraft sliding under the updraft, which then triggers new storms as the previous ones decay. Multicellular thunderstorms typically have a longer lifetime than singlecellular thunderstorms, see Section 2.1.6.2 of [28].

- **Supercellular** convection is much less common than the two previously mentioned types of convection, however it is associated with producing the most hazardous and long-lived storms, see Section 2.1.6.3 of [28]. This type of convection is associated with large values of deep layer wind shear and has one rotating updraft, see [33] and [28]. Supercellular storms are known to produce large (hail stone diameter $\geq 2\,\text{cm}$ [26]) and very large (hail stone diameter $\geq 5\,\text{cm}$ [26]) hail along with tornadoes and other extreme weather phenomena, see [2].

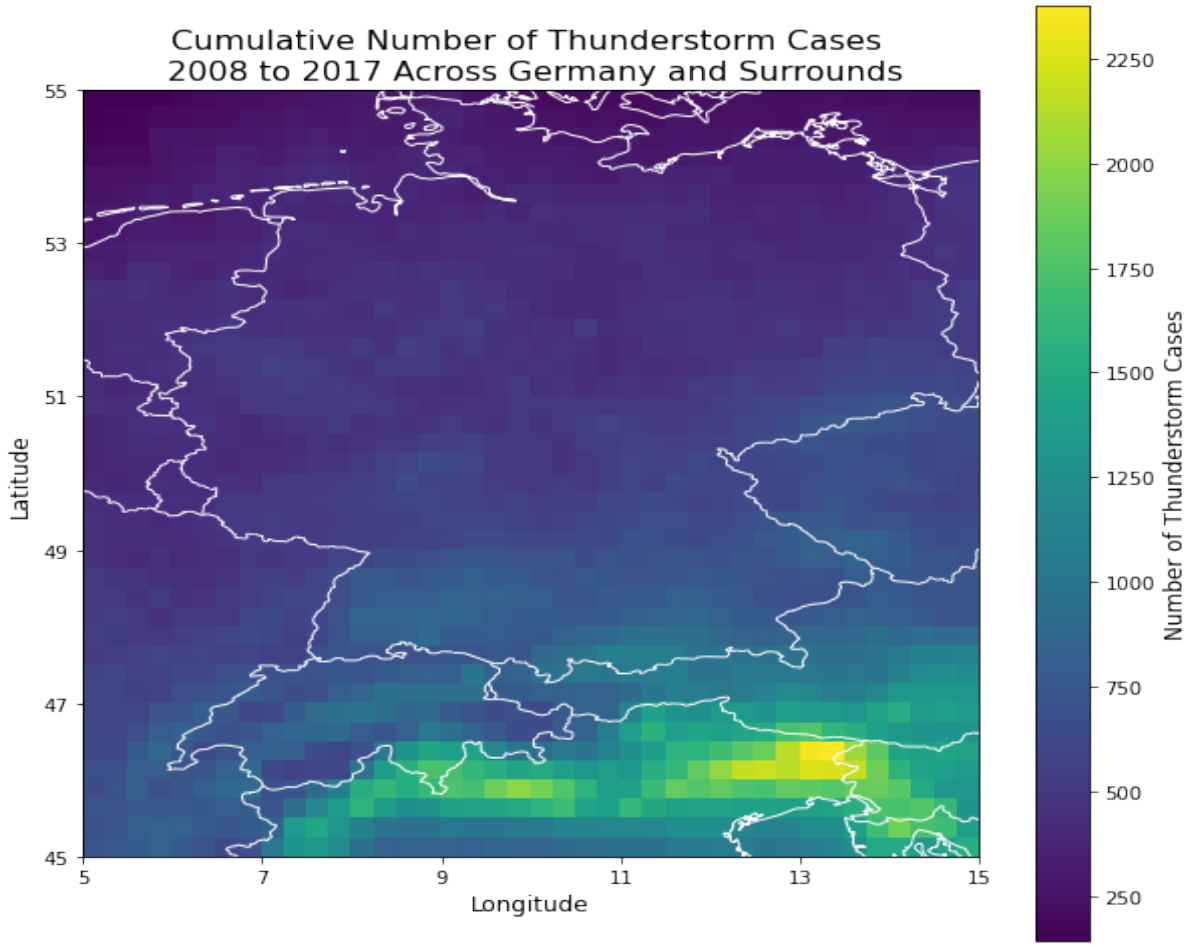


Figure 2. The number of thunderstorm cases detected across Germany and surrounds between 2008 and 2017. The cloud to ground lightning was detected by EUCLID with a grid resolution of one hour by 0.25° by 0.25°.

## 2.3 Hail Formation

The microphysical process by which hail forms is detailed in Section 2 of [2], and the following section draws on this paper to explain the formation of hail. Hail requires an 'embryo', which is most often a small frozen particle, to begin growth. Hail can undergo two types of growth: 'wet' or 'dry', both of which require the air to be sufficiently moist. Wet growth, called 'accretion', occurs when supercooled liquid water accumulates on the surface of the hail stone before freezing at a later point in time. This type of growth leads to clear ice because the supercooled water fills in any air bubbles or gaps on the surface of the hail stone. Dry growth, called 'riming', occurs when supercooled liquid water freezes as soon as it accumulates on the hail stone. This leads to a more opaque colour because air bubbles are trapped. Hail stones can go through multiple periods of each type of growth leading to a layered looking hail stone. This only means that it has experienced different types of growth, not necessarily that it has circulated through a cloud multiple times, see [2].
Hail formation also requires time, hence embryos which rise too quickly in a strong updraft rise out of the growth zone before they grow much larger than the initial nucleus. For a hail stone to continue growing, force balance must occur between gravity acting on the stone and the updraft of the storm, resulting in the stone remaining in the growth zone for an extended period of time, see [5].
The size of a hailstone is also affected by the zero degree level, the height in the atmosphere where the air temperature is 0 °C. This is the altitude below which a hail stone will melt, thus the lower the zero degree level the less time a hail stone has to melt before it reaches the ground, see Page 15 of [28]. The number of severe hail cases across Germany and surrounds from 2008 to 2017 are shown in Figure 3, using data from the ESWD. Here we can see several regions which have higher numbers of reports, and very few reports over the centre of the European Alps. This is expected due to not only the relatively low population density, but mostly because the cold, dry air above the year round snow over the highest parts of the Alps does not provide suitable conditions for the development of severe storms or hail.

## 2.4 Convective Available Potential Energy

'Convective available potential energy' (CAPE) relates to the strength of the updraft of a storm, and is the amount of buoyancy a parcel of 1 kg of air has. It can be calculated using the following formula from Equation 9.47 in [12]

$$CAPE = \int_{h_{fcl}}^{h_{eq}} g \frac{T_{parcel} - T_{env}}{T_{env}} dh \, [\mathrm{J\,kg^{-1}}], \tag{2.4.1}$$

where $g = 9.8067\,\mathrm{m\,s^{-2}}$ is the acceleration due to gravity, $h_{fcl}$ is the height of the free convection layer, $h_{eq}$ is the height of the equilibrium layer, $T_{parcel}$ is the temperature of the parcel and $T_{env}$ is the temperature of the environment, both of which are measured in Kelvin. CAPE is a measure of the amount of energy which could be released if a parcel of air was lifted to the free convection layer, where the parcel would become buoyant and begin rising in the atmosphere.
CAPE is related to the 'maximum updraft speed' ($w_{max}$) via the following equation, using

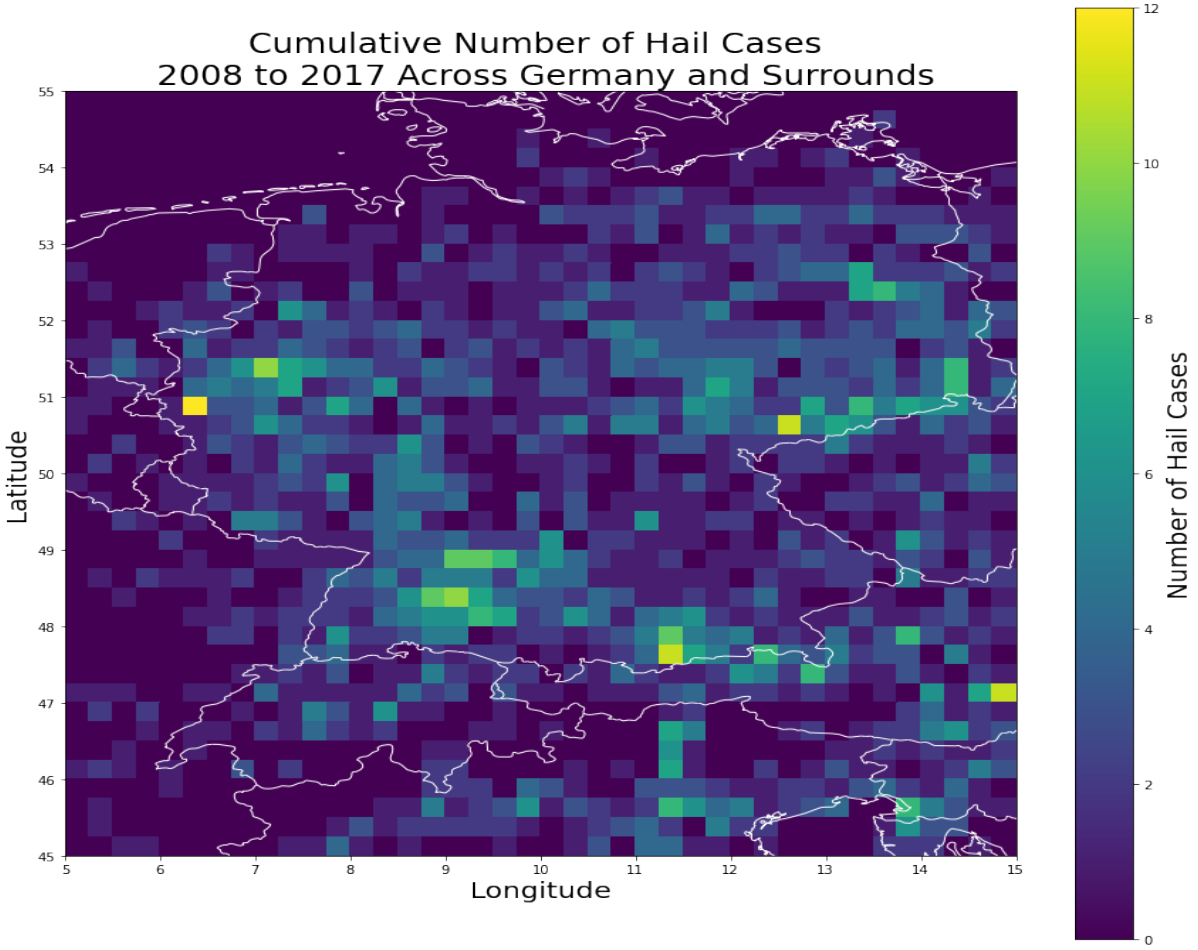


Figure 3. The number of severe hail cases reported across Germany and surrounds between 2008 and 2017. The reports were gathered by the ESWD and are gridded with a resolution of one hour by 0.25° by 0.25°.

Equation 9.47 in [12]:

$$w_{max} = \sqrt{2 \times CAPE}\,[\mathrm{m\,s^{-1}}]. \tag{2.4.2}$$

The updraft speed is a key element in large hail formation because this is the force which counteracts the force of gravity on a hail stone thus preventing it from falling to the earth's surface.

## 2.5 Deep Layer Wind Shear

'Deep layer wind shear' (DLS) is related to the type and severity of thunderstorms. The DLS has been calculated using the following formula from Equation 2.15 from [28]:

$$DLS = \sqrt{(u_{sfc} - u_{6\,\mathrm{km}})^2 + (v_{sfc} - v_{6\,\mathrm{km}})^2}\,[\mathrm{m\,s^{-1}}], \tag{2.5.1}$$

where $u$ is the eastward wind component and $v$ the northward wind component. DLS is sometimes approximated using the 500 hPa pressure level as in Equation (2.5.2),

$$DLS_{approx} = \sqrt{(u_{ground} - u_{500\,\mathrm{hPa}})^2 + (v_{ground} - v_{500\,\mathrm{hPa}})^2}\,[\mathrm{m\,s^{-1}}], \tag{2.5.2}$$

instead of the 6 km level, for example in [27]. This paper focused on severe hail under multiple climate scenarios until the year 2100. Here the 500 hPa approximation was used due to insufficient pressure levels being available in weather simulations based on the future climate scenarios.

## 2.6 Standard Deviation of Orography

The 'standard deviation of orography' represents the ruggedness or steepness of features of the terrain which are on a smaller scale than the resolution of the data. Hills, valleys and mountains are examples of this, and the steeper these features are, the larger the value of the standard deviation of orography. Orography can act as a trigger for the initiation of a thunderstorm as detailed in Section 2.1.3 of [28].

## 2.7 Zero Degree Level

The 'zero degree level' is the height in atmosphere at which the air temperature is 0 °C. Frozen water will melt when the temperature is higher than approximately 0 °C. This means that hail stones will begin melting at the zero degree level, and continue melting until they fall to the ground as discussed in Section 2.2.2 of [28].

## 2.8 Relative Humidity

The 'relative humidity' is the saturation of water in the air, where water in both solid and liquid phase is considered. This is expressed as a percentage and has a maximum value of 100 percent. Here we consider the average relative humidity between the 500 hPa and 850 hPa pressure levels. The importance of moisture for the initiation of a thunderstorm is discussed in Section 3.4 of [33].

# 3 Binary Regression

This section details methods which can be used to fit regression models to binary data. We first discuss the exponential family of distributions, followed by the procedures of fitting a Generalised Linear Model (GLM) and a Generalised Additive Model (GAM).

## 3.1 Exponential Family Distributions

A distribution belongs to the exponential family if its probability density function can be written in the form:

$$f(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right), \tag{3.1.1}$$

see Section 4.1 of [31]. Here $a$, $b$ and $c$ are arbitrary scalar functions, while $\theta$ and $\phi$ are the canonical and dispersion parameters, respectively. Consider an independent binomial random variable $Y_i$ with realisation $y_i$ for $i = 1, ..., n$, where $n$ is the number of responses, which has a binomial distribution:

$$Y_i \sim Bin(n_i, p_i), \tag{3.1.2}$$

where $n_i$ is the number of trials and $p_i$ the probability of success. The probability mass function is:

$$f(n_i, p_i) = \binom{n_i}{p_i} p_i^{y_i}(1-p_i)^{n_i - y_i}, \tag{3.1.3}$$

see, for example, Equation 4.1 of [35]. The distribution of $Y_i^s \coloneqq \frac{Y_i}{n_i}$, with realisations $y_i^s$, is called the scaled binomial distribution, and is a member of the exponential family, see Page 30 of [4]. Following [4] and using Equation (3.1.3) we can show that the scaled binomial distribution of the $Y_i^s$ is a member of the exponential family by writing the equation in the same form as Equation (3.1.1) with canonical parameter $\theta_i$ and functions $a_i$, $b$ and $c_i$ which are expressed as follows:

$$\theta_i = \ln\frac{p_i}{1-p_i}, \tag{3.1.4}$$

$$a_i(\phi) = \frac{\phi}{n_i},\ \phi = 1, \tag{3.1.5}$$

$$b(\theta_i) = \ln(e^{\theta_i} + 1), \tag{3.1.6}$$

and

$$c_i(y_i^s;\phi) = \ln\binom{n_i}{n y_i^s}. \tag{3.1.7}$$

The expected value of $Y_i^s$ is calculated as in [4] as follows:

$$\mu_i \coloneqq E[Y_i^s] = n_i p_i, \tag{3.1.8}$$

and the variance of $Y_i^s$ is calculated as in [4] as follows:

$$Var(Y_i^s) = n_i p_i (1-p_i). \tag{3.1.9}$$

## 3.2 Generalised Linear Model Construction

In the following section we define the components of which a Generalised Linear Model (GLM) is comprised. Let $Y_i$ and $y_i$ for $i = 1, ..., n$ denote the $i^{th}$ response variable and its realisations respectively. The $i^{th}$ corresponding covariate is denoted by $\boldsymbol{x}_i = (1, x_{i1}, ..., x_{ik})^T$, where the number of regression parameters, $p$, is defined by $p \coloneqq k + 1$.

### 3.2.1 GLM Components

We define a GLM as in Section 4.4 of [15] to have a random, systematic and link component.
The random component requires that the responses, or $Y_i$, are independent. Furthermore $Y_i$ has a probability mass function from the exponential family with a canonical parameter $\theta_i$ and dispersion parameter $\phi > 0$, as defined in Equation (3.1.1).
The systematic component is defined in Equation (3.2.1) for the linear predictor, $\eta_i(\boldsymbol{\beta})$:

$$\eta_i(\boldsymbol{\beta}) \coloneqq \boldsymbol{x}_i^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik}, \tag{3.2.1}$$

where $\boldsymbol{\beta} = (\beta_0, ..., \beta_k)^T$ are p unknown regression parameters.
The link component consists of the link function, defined in Equation (3.2.2):

$$g(\mu_i) \coloneqq \eta_i(\boldsymbol{\beta}), \tag{3.2.2}$$

which defines the relationship between the linear predictor, $\eta_i$, defined in Equation (3.2.1), and $\mu_i$, where $\mu_i$ is the mean of the response $Y_i$.

### 3.2.2 Properties of a GLM

Using the above definitions we now discuss some further definitions and properties associated with GLMs, along with the maximum likelihood estimate method for fitting regression coefficients.
We first define the log likelihood for a GLM with observed data $\boldsymbol{y} = (y_1, y_2, ..., y_n)^T$ as in Definition 3.7 of [4]:

$$l(\boldsymbol{\beta}, \phi|\boldsymbol{y}) \coloneqq \sum_{i=1}^{n} \ln\Big(f(y_i|\theta_i, \phi)\Big) = \sum_{i=1}^{n} l_i(\mu_i, \phi_i|y_i), \tag{3.2.3}$$

for a response variable, $Y_i$, which has a distribution from the exponential family.
Using Equation (3.1.1), Equation (3.2.3) becomes:

$$l(\boldsymbol{\beta}, \phi|\boldsymbol{y}) = \sum_{i=1}^{n} \left(\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c_i(y_i, \phi)\right), \tag{3.2.4}$$

for $Y_i$ with canonical parameter $\theta_i$ and dispersion parameter $\phi$.
The expectation and variance function are, as shown in Theorem 3.3 and Section 3.2 of [4]:

$$E[Y_i] = b'(\theta_i) = \mu_i, \tag{3.2.5}$$

and

$$V(\theta_i) \coloneqq b''(\theta_i). \tag{3.2.6}$$

To determine the maximum likelihood estimates of $\boldsymbol{\beta}$, we need to, as the name suggests, maximise the log likelihood, $l$. Therefore we maximise Equation (3.2.3) with respect to $\boldsymbol{\beta}$:

$$\frac{\partial}{\partial \beta_j} l = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \mu_i} \frac{d\mu_i}{d\eta_i} \frac{\partial \eta_i}{\partial \beta_j}. \tag{3.2.7}$$

Using Equation (3.2.2) we can see that:

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}. \tag{3.2.8}$$

Using Equations (3.2.5) and (3.2.4) the following can be shown:

$$\frac{\partial l_i}{\partial \mu_i} = \frac{\frac{\partial l_i}{\partial \theta_i}}{\frac{\partial \mu_i}{\partial \theta_i}} = \frac{y_i - b'(\theta_i)}{a(\phi) b''(\theta_i)}. \tag{3.2.9}$$

Thus using Equations (3.2.8) and (3.2.9), Equation (3.2.7) becomes:

$$\sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{a(\phi) b''(\theta_i)} \frac{d\mu_i}{d\eta_i} x_{ij} = 0. \tag{3.2.10}$$

We define the weights for a generalised linear model as in Definition 3.8 of [4] as:

$$W_i \coloneqq W_i(\beta) \coloneqq \frac{(\frac{d\mu_i}{d\eta_i})^2}{b''(\theta_i)}. \tag{3.2.11}$$

Combining Equations (3.2.10) and (3.2.11) we define the unscaled score equations in Equation (3.2.12), as in Definition 3.9 of [4]:

$$s_j(\boldsymbol{\beta}, \boldsymbol{y}) \coloneqq \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_j} = \sum_{i=1}^{n} W_i (y_i - \mu_i) \frac{d\eta_i}{d\mu_i} x_{ij} = 0, \tag{3.2.12}$$

for $j = 1, .., p$, where $\eta_i$ and $x_{ij}$ are defined in Equation (3.2.1). We now solve to find $\hat{\boldsymbol{\beta}}$, the value of $\boldsymbol{\beta}$ such that the log likelihood is maximised.
Typically these equations need to be solved numerically, for example, by using the Fisher scoring method, see Page 35 of [4]. We define the unscaled Hessian as in Definition 3.10 of [4] with entries:

$$H_{jk} = \frac{\partial s_j(\boldsymbol{\beta}, \boldsymbol{y})}{\partial \beta_k}, \tag{3.2.13}$$

for $j, k = 1, ..., p$. The Fisher information matrix, $I$, is defined as in Definition 3.10 of [4] as:

$$I(\boldsymbol{\beta}) \coloneqq E[-H(\boldsymbol{\beta}, \boldsymbol{Y})] \tag{3.2.14}$$

It is shown on Page 36 of [4] that the $(j, k)^{th}$ element of $I(\boldsymbol{\beta})$ from Equation (3.2.14) can be expressed as:

$$I_{j,k}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{1}{b''(\theta_i)} \left( \frac{d\mu_i}{d\eta_i} \right)^2 x_{i,k} x_{i,j} = \sum_{i=1}^{n} W_i x_{i,k} x_{i,j}, \tag{3.2.15}$$

where $W_i$ is defined in Equation (3.2.11) for $j, k = 1, ..., p$. Expressing this in matrix form we have the following:

$$I(\boldsymbol{\beta}) = X^T W(\boldsymbol{\beta}) X, \tag{3.2.16}$$

where $X^T = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n) \in \mathbb{R}^{p \times n}$ and $W(\boldsymbol{\beta}) = diag(W_i(\boldsymbol{\beta})) \in \mathbb{R}^{n \times n}$ where $diag$ denotes the diagonal matrix as follows:

$$diag(W_i(\boldsymbol{\beta})) = \begin{pmatrix} W_1(\boldsymbol{\beta}) & 0 & \cdots & 0 \\ 0 & W_2(\boldsymbol{\beta}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & W_n(\boldsymbol{\beta}) \end{pmatrix}.$$

We then use this to construct the Fisher scoring algorithm as in Algorithm 3.12 of [4] as follows:

1. Select initial values $\boldsymbol{\beta}^0$ and $\epsilon$.

2. For k $> 0$ we define the following:

$$\boldsymbol{\beta}^{k+1} \coloneqq \boldsymbol{\beta}^k + (I(\boldsymbol{\beta}^k)^{-1} \boldsymbol{s}(\boldsymbol{\beta}^k, \boldsymbol{y}) \tag{3.2.17}$$

   where $\boldsymbol{s}(\boldsymbol{\beta}, \boldsymbol{y}) = (s_1(\boldsymbol{\beta}, \boldsymbol{y}), ..., s_p(\boldsymbol{\beta}, \boldsymbol{y}))^T$.

3. If $||\boldsymbol{\beta}^{k+1} - \boldsymbol{\beta}^k|| < \epsilon$, define estimates of $\boldsymbol{\beta}$ as follows: $\hat{\boldsymbol{\beta}} \coloneqq \boldsymbol{\beta}^{k+1}$

This can also be expressed as an iterative weighted least squares algorithm. For this, we first express Equation (3.2.17) as in Equation 3.17 of [4] as:

$$I(\boldsymbol{\beta}^k)\boldsymbol{\beta}^{k+1} \coloneqq I(\boldsymbol{\beta}^k)\boldsymbol{\beta}^k + \boldsymbol{s}(\boldsymbol{\beta}^k, \boldsymbol{y}). \tag{3.2.18}$$

Following Page 37 of [4] this can be expressed as:

$$\sum_{i=1}^{n} W_i(\boldsymbol{\beta}^k) x_{ij} Z_i^k = \sum_{i=1}^{n} W_i(\boldsymbol{\beta}^k) x_{ij} \eta_i^{k+1}, \tag{3.2.19}$$

for $j = 1, ..., p$, where $\eta_i^k = \boldsymbol{x}_i^T \boldsymbol{\beta}^k$, $\mu_i^k = g^{-1}(\eta_i^k)$ and $Z_i^k = \eta_i^k + (y_i - \mu_i^k)\frac{d\eta_i^k}{d\mu_i^k}$, which is called the adjusted dependent variable. We then define the inverse mean function as in Definition 3.6 of [4] as the inverse of $b'(\cdot)$:

$$h(\cdot) \coloneqq [b']^{-1}(\cdot), \tag{3.2.20}$$

Now we can construct the iterative weighted least squares algorithm as in Algorithm 3.12 of [4]:

1. Select initial values $\boldsymbol{\beta}^0$ and $\epsilon$.

2. For the current estimate of $\boldsymbol{\beta}$, $\boldsymbol{\beta}^k$, calculate the following for $i = 1, ..., n$: $\hat{\eta}_i^k = \boldsymbol{x}_i^T \boldsymbol{\beta}^k$, $\hat{\mu}_i^k = g^{-1}(\hat{\eta}_i^k)$, $\hat{\theta}_i^k = h(\hat{\mu}_i^k)$,

$$Z_i^k = \hat{\eta}_i^k + (y_i - \hat{\mu}_i^k)\left(\frac{d\eta_i^k}{d\mu_i^k}\right)\Bigg|_{\mu_i^k = \hat{\mu}_i^k}, \tag{3.2.21}$$

$$W_i^k = \left[b''(\theta_i)\Big|_{\theta_i = \hat{\theta}_i^k}\left(\frac{d\eta_i^k}{d\mu_i^k}\right)\Bigg|_{\mu_i^k = \hat{\mu}_i^k}\right]^{-1}. \tag{3.2.22}$$

3. Regress $Z_i^k$ on $x_{i1}, ..., x_{ip}$ using weights $[W_i^k]^{-1}$ to calculate the updated the estimates $\boldsymbol{\beta}^{k+1}$.

4. Repeat from step 2 until $||\boldsymbol{\beta}^{k+1} - \boldsymbol{\beta}^k|| < \epsilon$, then define estimates of $\boldsymbol{\beta}$ as follows: $\hat{\boldsymbol{\beta}} \coloneqq \boldsymbol{\beta}^{k+1}$.

### 3.2.3 Goodness of Fit of GLMs

We now look at a measure of goodness of fit, namely the deviance, to determine how well the model fits the data. We first define the fitted means as on Page 40 of [4] as follows:

$$\hat{\mu}_i \coloneqq g^{-1}(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}), \tag{3.2.23}$$

where the function $g$ is defined in Equation (3.2.2) and $\hat{\boldsymbol{\beta}}$ are the estimated regression parameters.
We can now define the scaled deviance as in Definition 3.17 of [4]:

$$D^s(\hat{\boldsymbol{\mu}}, \boldsymbol{y}, \phi) = -2\Big(l(\hat{\boldsymbol{\mu}}, \phi|\boldsymbol{y}) - l(\boldsymbol{y}, \phi|\boldsymbol{y})\Big), \tag{3.2.24}$$

which, using Equation (3.2.4) can also be written as:

$$= 2\sum_{i=1}^{n} \frac{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)}{a(\phi)}, \tag{3.2.25}$$

where $a$ and $b$ are the functions defined in Equation (3.1.1), $\hat{\theta}_i = h(\hat{\mu}_i)$ and $\tilde{\theta}_i = h(y_i)$. We can also assess the goodness of fit of a GLM using the Pearson residuals. Following Section 4.4 of [4], we consider the fitted probability of success:

$$\hat{p}_i = \frac{e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}{e^{\boldsymbol{x}_i^T \boldsymbol{\beta}} + 1}. \tag{3.2.26}$$

The Pearson residuals are defined as in Definition 3.23 of [4] as:

$$e_i^P = \frac{y_i - \hat{\mu}_i}{(V(\hat{\mu}_i))^{0.5}}, \tag{3.2.27}$$

for $i = 1, ..., n$, where $\hat{\mu}_i \coloneqq n_i \hat{p}_i$ are the fitted means and $V(\cdot)$ is the variance function.

### 3.2.4 Logistic Regression Example

In this section we look at logistic regression, an example of a GLM, and some of the above equations in this case. Here we consider the logit link from Equation 3.7 of [4] as defined in Equation (3.2.28):

$$g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right). \tag{3.2.28}$$

Using Section 3.1 where we showed that the scaled binomial distribution belongs to the exponential family, as in Definition 4.16 of [4], we can show that the log likelihood from Equation (3.2.4) becomes:

$$l(\boldsymbol{\beta}, \boldsymbol{y}^s) = \sum_{i=1}^{n} n_i \left( y_i^s \boldsymbol{x}_i^T \boldsymbol{\beta} - \ln(e^{\boldsymbol{x}_i^T \boldsymbol{\beta}} + 1) \right) + C, \tag{3.2.29}$$

where $C \in \mathbb{R}$ is independent of $\boldsymbol{\beta}$. We can also calculate that Equation (3.2.6) can be written as:

$$V(\theta_i) = \frac{e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}{(e^{\boldsymbol{x}_i^T \boldsymbol{\beta}} + 1)^2}, \tag{3.2.30}$$

where $\theta_i = p_i = (e^{\boldsymbol{x}_i^T \boldsymbol{\beta}} + 1)^{-1}$ The scaled score equations are defined as in Definition 4.17 of [4] as:

$$s(\boldsymbol{\beta}, \boldsymbol{y}^s)_j^s \coloneqq \frac{\partial l(\boldsymbol{\beta} | \boldsymbol{y}^s)}{\partial \beta_j}, \tag{3.2.31}$$

which can also be written as:

$$s(\boldsymbol{\beta}, \boldsymbol{y}^s)_j^s = \sum_{i=1}^{n} n_i x_{ij} \left( y_i^s - \frac{e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}{e^{\boldsymbol{x}_i^T \boldsymbol{\beta}} + 1} \right), \tag{3.2.32}$$

for $j = 1, ..., p$. This allows us to define the scaled Hessian, as on Page 67 of [4]:

$$H^s \coloneqq \frac{\partial \boldsymbol{s}^s(\boldsymbol{\beta}, \boldsymbol{y}^s)}{\partial \boldsymbol{\beta}} \in \mathbb{R}^{p \times p}, \tag{3.2.33}$$

the $(k, l)^{th}$ element of which can be expressed as follows:

$$H_{kl}^s = -\sum_{i=1}^{n} n_i p(\boldsymbol{x}_i)(1 - p(\boldsymbol{x}_i)) x_{ik} x_{il}. \tag{3.2.34}$$

We will be interested in examining the empirical logits as part of the exploratory data analysis later in this thesis. We define the empirical logit, $l_i$, as in Definition 4.15 of [4]:

$$l_i = \ln\left(\frac{y_i + \frac{1}{2}}{n_i - y_i + \frac{1}{2}}\right), \tag{3.2.35}$$

for $i = 1, ..., n$, where $n$ is the number of responses, $y_i$ is the number of successes and $n_i$ is the number of trials. One half has been added to both the numerator and denominator in Equation (3.2.35) to account for the case where $y_i = 0$ or $y_i = n_i$. This continuity correction is explained on Page 61 of [4].

As shown in Definition 4.13 of [4], for the category $i$ with *score*, $s(i)$, the logistic model can be rewritten with $s(i)$ as the covariate,

$$p(i) = P(Y = 1|X = i) = \frac{e^{\beta_0+\beta_1 s(i)}}{1 + e^{\beta_0+\beta_1 s(i)}}, \tag{3.2.36}$$

where it is assumed that the logits are linear in terms of the scores as in Equation (3.2.37)

$$\text{logit}(p(i)) = \beta_0 + \beta_1 s(i). \tag{3.2.37}$$

### 3.2.5 Overdispersion

So far we have discussed a method for fitting regression coefficients for a GLM, along with techniques to asses the goodness of fit. We now turn our attention to the case where there is overdispersion present in the data.
One indicator for the presence of overdispersion in a data set is the condition $D > n - p$, where $D$ is the deviance, $n$ the number of responses and $p$ the number of regression parameters, see Remark 5.2 of [4].
We now define the latent random variable, $v_i$, as in definition 5.4 of [4]: for $Y_i$ successes in $n_i$ trials we define the random success probability, $v_i$, also called the latent random variable, for $i = 1, ..., n$. Additionally, we assume the $Y_i$ given $v_i$ are independent, with:

$$E[v_i] = p_i, \tag{3.2.38}$$

and

$$Var(v_i) = \phi_i p_i(1 - p_i), \tag{3.2.39}$$

where $\phi$ is the unknown scale parameter and $p_i \coloneqq p(\boldsymbol{x}_i)$. Section 5.3 of [4] shows that the expectation and variance of $Y_i$ are:

$$E[Y_i] = n_i p_i, \tag{3.2.40}$$

for $i = 1, ..., n$ and

$$Var(Y_i) = n_i p_i(1 - p_i)(1 + (n_i - 1)\phi), \tag{3.2.41}$$

for $i = 1, ..., n$. We now define the beta-binomial logistic model, where $v_i \sim Beta(a_i, b_i)$ as in Section 5.4 of [4]. The expectation and variance of $v_i$ are given by:

$$E[v_i] = \frac{a_i}{a_i + b_i} = p_i, \tag{3.2.42}$$

$$Var(v_i) = \frac{a_i b_i}{(a_i + b_i)^2(a_i + b_i + 1)} = \frac{p_i(1 - p_i)}{(a_i + b_i + 1)}. \tag{3.2.43}$$

For the case $\frac{1}{(a_i+b_i+1)} = \phi$ for all $i = 1, ..., n$ we obtain the beta-binomial logistic model, where the $v_i$ are independent and follow the $Beta(a_i, b_i)$ distribution for $a_i = \frac{1-\phi}{\phi} p_i$ and $b_i = \frac{1-\phi}{\phi}(1 - p_i)$. Note that the $p_i$ remain unchanged from the previously defined logistic model. To fit the regression coefficients for our beta-binomial model using maximum likelihood estimation we follow Page 106 of [4] and express the likelihood as follows:

$$L(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{i=0}^{n} \int_0^1 f(y_i|v_i) f(v_i|a_i, b_i) dv_i, \tag{3.2.44}$$

which can also be written as:

$$L(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{i=0}^{n} \binom{n_i}{y_i} \frac{B(y_i + a_i, n_i - y_i + b_i)}{B(a_i, b_i)}. \tag{3.2.45}$$

Page 106 of [4] explains that the optimisation of Equation (3.2.44) must be performed numerically.

## 3.3 Construction of GAM model

We now consider the construction of a GAM. This type of model can either be used for exploratory data analysis, or in the case that a model with more flexibility than a GLM is required can be used as a model in its own right, see Section C of [4]. The use of a GAM requires choosing a basis, fitting model coefficients and selecting a smoothing parameter. Once a basis has been chosen we can reformulate the GAM to have the same form as a GLM.
Here we define a GAM as in Section 6.1 of [37]:

$$g(\mu_i) = \boldsymbol{A}_i\boldsymbol{\gamma} + \sum_{j=1}^{k} f_j(x_{ij}), \tag{3.3.1}$$

for $j = 1, ..., k$ and $i = 1, ..., n$, where $n$ is the number of responses, $\mu_i := E[Y_i]$ and the probability density function of the response $Y_i$ belongs to the exponential family distribution. The $y_i$ are assumed to be independent given $\mu_i$. $\boldsymbol{A}_i$ is the $i^{th}$ row of a parametric model matrix with corresponding parameter vector $\boldsymbol{\gamma}$ and $f_j$ is a smooth function of the covariate vector $\boldsymbol{x_j}$, where $x_{ij}$ is the $i^{th}$ element of $\boldsymbol{x_j}$ and $k$ is the number of smooth functions.
This allows for more flexibility when compared to a GLM; however this model requires a basis to be chosen and the selection of a smoothing parameter, in addition to the fitting of model coefficients.

## 3.4 Basis Selection

In this section we want to choose a basis, or way of representing the $f_j$ from Equation (3.3.1). There are many different options to select a basis ranging from polynomials through to splines. One particular method suited to large and complex datasets is the tensor product spline covered later in this section. The goal of all of these methods is to choose a basis which compromises between goodness of fit and the curvature of the function or over-fitting. We first consider the polynomial basis example from Section 4.2 of [37].

### 3.4.1 Polynomial Basis

Consider a model with only one covariate, $\boldsymbol{x}$:

$$y_i = f(x_i) + \epsilon_i, \text{ with } \epsilon_i \overset{ind.}{\sim} \mathcal{N}(0, 1), \tag{3.4.1}$$

for $i = 1, ..., n$, where $y_i$ is a dependent variable, $x_i$ is the $i^{th}$ element of $\boldsymbol{x}$, the covariate, and $f$ is a smooth function. To estimate $f$, we want to express it so that Equation (3.4.1) becomes a linear model. We do this by choosing a basis consisting of $M$ basis functions and approximating $f$ by $f_{approx}$,

$$f_{approx}(x) = \sum_{l=1}^{M} b_l(x)\beta_l, \tag{3.4.2}$$

for $l = 1, ..., M$, where $f_{approx}$ is an element of the function space spanned by the basis. In this case $b_l$ is the $l^{th}$ basis function and the $\beta_l$s are coefficients which will be fitted in Section 3.6.
For this example we will consider a cubic polynomial:

$$b_1(x) = 1,\ b_2(x) = x,\ b_3(x) = x^2,\ b_4(x) = x^3. \tag{3.4.3}$$

Equation (3.4.2) becomes:

$$f_{approx}(x) = \sum_{l=1}^{4} x^{l-1}\beta_l, \tag{3.4.4}$$

and Equation (3.4.1) becomes:

$$y_i = \sum_{l=1}^{4} x_i^{l-1}\beta_l + \epsilon_i. \tag{3.4.5}$$

The trade off for the simplicity of this model comes from Taylor's Theorem as discussed in Section 4.2.1 of [37]. Taylor's Theorem implies that this method is only useful for very localised areas, not for the whole domain of $f$.

### 3.4.2 Cubic Spline

The choice of a spline as a basis results in a method more useful across a larger domain of the smooth $f$. The most commonly used spline is the cubic spline, see Section 1.8 and the Preface of [22], providing a good compromise between approximation power and computational speed. Here we follow Section 5.1.1 of [37] and Section C of [4]. Our representation of $f_j$ from Equation (3.3.1) is as follows:

$$f_j(x) = \beta_j x + s_j(x), \tag{3.4.6}$$

for $j = 1, ..., k$, where $\beta_j$ are the regression coefficients for the parametric component. The non parametric component, $s_j(x)$, is a spline which consists of $m - 1$ piece-wise polynomials, $p_l(x)$, which are joined at a set of knots $x'_1, ..., x'_m$ as shown in [4]:

$$s_j(x) = p_l(x), \tag{3.4.7}$$

where $x'_l \leq x \leq x'_{l+1}$ for $l = 1, ..., m - 1$, where the $p_l$ polynomials of degree $d$ fulfill the following:

$$p_l^h(x'_l) = p_{l+1}^h(x'_l), \tag{3.4.8}$$

for $h = 1, .., d-1$. The case of a cubic spline corresponds to $d = 3$. Consider the responses $y_i$ and covariate elements $x_i$ for $i = 1, ..., n$. The parameter $\lambda$ controls the weighting between the fit of the function $f$ to the data and penalising the curvature of $f$, we will show how this parameter is chosen in Section 3.7. For knots $a \leq x'_1 \leq x'_2 \leq ... \leq x'_m \leq b$ we choose $f$ such that the following equation, as in Equation C.5 of [4], is minimised:

$$\mathcal{A}(f) = \sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int_b^a f''(x)^2 dx. \tag{3.4.9}$$

We now consider the case where we have $k$ smooth functions, $s_j(x)$, for $j = 1, ..., k$. The penalty, $P'_j$, for $s_j$ is defined as follows:

$$P'_j \coloneqq \int_{v_j}^{u_j} s''_j(x)^2 dx, \tag{3.4.10}$$

where $u_j = \min(x_{ij}, 1 \leq i \leq n)$ and $v_j = \max(x_{ij}, 1 \leq i \leq n)$ for $j = 1, ..., k$. Since we have defined $s_j(x)$ to be a cubic polynomial, we can expand $s_j(x)$ as in Equation C.3 of [4] as follows:

$$s_j(x) = \sum_{k=1}^{q_j} \beta_{j,k} b_{jk}(x), \tag{3.4.11}$$

for $j = 1, ..., q_j$, where the $\beta_{j,k}$ are coefficients and the $b_{jk}$ are the basis functions. Using this, as in Section C.3 of [4], we can then write:

$$P'_j = \boldsymbol{\beta}_j^T P_j \boldsymbol{\beta}_j, \tag{3.4.12}$$

where $\boldsymbol{\beta}_j$ is a vector of length $q_j$ containing the coefficients from Equation (3.4.11) and $P_j \in \mathbb{R}^{q_j \times q_j}$ defined as follows:

$$(P_j)_{lk} = \int_{u_j}^{v_j} b_{jl}(x) b_{jk}(x) dx. \tag{3.4.13}$$

The overall penalty matrix, $P_\lambda$ is defined as in Section C.3 of [4] as follows:

$$P_\lambda \coloneqq diag(\lambda_1 P_1, ..., \lambda_k P_k), \tag{3.4.14}$$

where the $\lambda > 0$ are the smoothing parameters, which will be fit in Section 3.7.

### 3.4.3 Tensor Product Smooths

We now consider a more complex basis best suited for cases with multiple covariates. This section follows directly from Section 5.6 of [37], which recommends using tensor product smooths for large multidimensional data sets due to their computational favourability. First, consider a smooth function $f(x, z)$ with marginal bases $f_x$, $f_z$ of order N and M respectively, for covariates $x$ and $z$:

$$f_x(x) = \sum_{j=1}^{N} \alpha_j a_j(x), \tag{3.4.15}$$

$$f_z(z) = \sum_{k=1}^{M} \beta_k b_k(z), \tag{3.4.16}$$

where $a_j$ is the $j^{th}$ basis function and a smooth function of $x$ for $j = 1, ..., N$ and $b_k$ is the $k^{th}$ basis function and a smooth function of $z$ for $k = 1, ..., M$. The marginal penalties $J_x$ and $J_z$ can then be expressed as in Section 5.6.2 of [37]:

$$J_x(f_x(x)) = \boldsymbol{\alpha}^{\boldsymbol{T}} \boldsymbol{S_x} \boldsymbol{\alpha} \tag{3.4.17}$$

$$J_z(f_z(z)) = \boldsymbol{\beta}^{\boldsymbol{T}} \boldsymbol{S_z} \boldsymbol{\beta} \tag{3.4.18}$$

where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_N)^T \in \mathbb{R}^N$ and $\boldsymbol{\beta} = (\beta_1, ..., \beta_M)^T \in \mathbb{R}^M$ are vectors of the coefficients $\alpha_j$ and $\beta_k$ from Equations (3.4.15) and (3.4.16) and $\boldsymbol{S}_x$ and $\boldsymbol{S}_z$ are penalty coefficient matrices defined as on Page 226 of [37] as follows:

$$(S_x)_{ij} = \int_{u_x}^{v_x} a_i(x) a_j(x), \tag{3.4.19}$$

where $v_x$ and $u_x$ are $\max(x)$ and $\min(x)$ respectively. Similarly for $z$:

$$(S_z)_{ij} = \int_{u_z}^{v_z} b_i(x) b_j(x), \tag{3.4.20}$$

where $v_z$ and $u_z$ are $\max(z)$ and $\min(z)$ respectively. Now we want to make $f_x(x)$ vary smoothly in $z$. We can use Equation (3.4.16) to express a function varying smoothly in $z$, which we now use to express $\alpha_j$ as a function of $z$. We must now take into account that each $\alpha_j$ will have a coefficient vector. We now define a coefficient matrix for all $\alpha_j$,

$$\boldsymbol{B} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_N) \in \mathbb{R}^{M \times N}, \tag{3.4.21}$$

where each of the $N$ $\boldsymbol{\beta}_j \in \mathbb{R}^M$ represent the vector coefficient for the corresponding $\alpha_j$. Note that these are not $\beta_j$, which we used to denote the element of a single coefficient vector. Thus each $\alpha_j(z)$ can be expressed as follows:

$$\alpha_j(z) = \sum_{l=1}^{M} B_{lj} b_l(z), \tag{3.4.22}$$

for $j = 1, ..., N$, where $B_{lj}$ is the $l^{th}$ element of the $j^{th}$ column of $\boldsymbol{B}$, or the $l^{th}$ coefficient of the function $\alpha_j$, which depends on the covariate z. Using Equations (3.4.22) and (3.4.15), $f(x, z)$ can be expressed as in Section 5.6.1 of [37] as:

$$f(x, z) = \sum_{j=1}^{N} \sum_{l=1}^{M} B_{lj} b_l(z) a_j(x). \tag{3.4.23}$$

Now that we have constructed the tensor product basis, we measure the contribution of higher order terms to $f$, or 'wiggliness', as in [37] by using penalties induced from marginal penalties. We measure this by considering first the wiggliness in $f$ with respect to $x$ with fixed $z$, then with respect to $z$ with fixed $x$. We denote $f$ where $z$ is fixed as $f_{x|z}$ and $f$ where $x$ is fixed as $f_{z|x}$. $f_{x|z}$ is defined as follows:

$$f_{x|z}(x) = \sum_{j=1}^{N} \alpha_j(z) a_j(x), \tag{3.4.24}$$

and $f_{z|x}$ is defined similarly. Let $J_{x|z}$ and $J_{z|x}$ denote the marginal penalties with fixed $z$ and $x$ respectively. Let $J_{x|z}(f_{x|z}(x))$ measure the wiggliness of $f(x, z)$ for fixed $z$.

Thus, $\int J_{x|z}(f_{x|z}(x))dz$ is proportional to the average wiggliness at $x$. Using the same argument for $z$, the penalty can be written as:

$$J(f(x,z)) = \lambda_x \int J_{x|z}(f_{x|z}(x))dz + \lambda_z \int J_{z|x}(f_{z|x}(z))dx, \tag{3.4.25}$$

where $\lambda_x$ and $\lambda_z$ are coefficients, see Section 5.6.2 of [37]. If we consider the marginal bases to be cubic splines as in the previous section, then the penalty becomes:

$$J(f(x,z) = \int \left(\lambda_x \left(\frac{\partial^2 f}{\partial x^2}\right)^2 + \lambda_z \left(\frac{\partial^2 f}{\partial z^2}\right)^2\right) dxdz \tag{3.4.26}$$

The reader is referred to [37] for further detail. Using the design matrix $\boldsymbol{X}$, basis $f(x,z)$, and penalty $J(f(x,z))$, the coefficients and smoothing parameters can be estimated.

## 3.5 Equivalence of a GAM to a GLM

Once the basis and penalty have been chosen the model can be reformulated as a GLM as shown in Section C of [4]. This allows us to use model fitting techniques discussed under the GLM framework, however, this can lead to overfitting due to a high number of parameters as argued in Section C of [4].
Consider smooth components $s_j(x)$ with $q_j$ basis functions $b_{jk}(x)$:

$$s_j(x) = \sum_{k=1}^{q_j} \beta_{j,k} b_{jk}(x), \tag{3.5.1}$$

for $j = 1, ..., q_j$ where the $\beta_{j,k}$ are the regression coefficients. We can define the transformed $i^{th}$ observation as follows:

$$\boldsymbol{z_i} = (1, x_{i1}, ..., x_{ik}, b_{11}(x_{i1}), ..., b_{1q_1}(x_{i1}), ..., b_{k1}(x_{ik}), ..., b_{kq_k}(x_{ik})) \in \mathbb{R}^P, \tag{3.5.2}$$

where the number of parameters, $P$, is defined as:

$$P = 1 + k + \sum_{j=1}^{k} q_j. \tag{3.5.3}$$

Similarly we can define the regression coefficients as:

$$\boldsymbol{\beta} = (\beta_0, ..., \beta_k, \beta_{1,1}, ..., \beta_{1,q_1}, ..., \beta_{k,1}, ..., \beta_{k,q_k}) \in \mathbb{R}^P. \tag{3.5.4}$$

Then the design matrix can be written as:

$$\boldsymbol{Z} = (\boldsymbol{z_1}, ..., \boldsymbol{z_n})^T \in \mathbb{R}^{n \times P}, \tag{3.5.5}$$

and the regression coefficients fitted, as shown in Section C of [4].

## 3.6 Fit Using Penalised Iterative Least Squares

In this section our aim is to fit the regression coefficients, $\boldsymbol{\beta}$. Before defining the algorithm which we will use to fit $\boldsymbol{\beta}$, namely the penalised iterative least squares (PIRLS) algorithm shown in Section 3.6.1, we first cover all definitions and assumptions required for the algorithm.
The following algorithm from Section 6.1 of [37] fits a logistic GAM with penalised likelihood maximisation, using penalised iterative least squares (PIRLS). Thus, we want to maximise the penalised log likelihood as defined in [37]. Using [4], Equations (3.5.4) and (3.4.14) and the log likelihood defined in Equation (3.2.3) we define the penalised log likelihood, $l_{pen}$ as follows:

$$l_{pen}(\boldsymbol{\beta}|\boldsymbol{y}) = l(\boldsymbol{\beta}|\boldsymbol{y}) - \frac{1}{2}\boldsymbol{\beta}^T P_\lambda \boldsymbol{\beta}. \tag{3.6.1}$$

Using Equation (3.2.2), the following equations are used to initialise and update $\hat{\mu}$ and $\hat{\eta}$:

$$\hat{\eta}_i = \boldsymbol{z}_i^T \hat{\boldsymbol{\beta}}_i \tag{3.6.2}$$

$$\hat{\mu}_i = g^{-1}(\hat{\eta}_i) \tag{3.6.3}$$

Define $\alpha$ as in Equation (3.6.4):

$$\alpha(\mu_i) \coloneqq 1 + (y_i - \mu_i)\left(\frac{V'(\mu_i)}{V(\mu_i)} + \frac{g''(\mu_i)}{g'(\mu_i)}\right), \tag{3.6.4}$$

where $g$ is the link function as defined in Equation (3.2.2), and $V$ is defined in Equation (3.2.6).

### 3.6.1 PIRLS Algorithm

1. Using Equations (3.6.2), (3.6.3), (3.2.6) and (3.6.4) calculate:

$$w_i = \frac{\alpha(\hat{\mu}_i)}{V(\hat{\mu}_i)g'(\hat{\mu}_i)^2} \tag{3.6.5}$$

and

$$z_i = \frac{g'(\hat{\mu}_i)(y_i - \hat{\mu}_i)}{\alpha(\hat{\mu}_i)} + \hat{\eta}. \tag{3.6.6}$$

2. Use $\boldsymbol{W} \coloneqq diag(w_i)$ and $||a||_W^2 \coloneqq \boldsymbol{a^T W a}$, to choose $\boldsymbol{\beta}$ so that the following equation is minimised:

$$||\boldsymbol{z} - \boldsymbol{X\beta}||_W^2 + \boldsymbol{\beta^T} P_\lambda \boldsymbol{\beta}, \tag{3.6.7}$$

3. Update $\hat{\eta}_i$ and $\hat{\mu}_i$ using Equations (3.6.2), (3.6.3).

The $\boldsymbol{\beta}$ which is chosen to minimise Equation (3.6.7) is defined as the fitted regression coefficients, $\hat{\boldsymbol{\beta}}$. Note that $\alpha(\mu_i)$ can be set to 1 to correspond to the Fisher scoring case, where the Hessian of the log likelihood is replaced by its expectation, see Section 6.1.1 of [37].

## 3.7 Smoothing Parameter Selection

The smoothing parameter $\lambda$ should be chosen such that a balance is achieved between a sufficiently smooth function and the fitted function $\hat{f}(x)$ being as close as possible to the true function $f(x)$. Ordinary cross validation (OCV), defined in [37] Section 4.2.1, can be used to choose $\lambda$ to minimise the cross validation score:

$$\nu_o = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}^{[-i]} - y_i)^2$$

where $\hat{f}^{[-i]}$ is the fit excluding $(y_i, x_i)$, where $\mathrm{X} = (\boldsymbol{x}_1^T, ..., \boldsymbol{x}_n^T)^T \in \mathbb{R}^{n,k}$. The OCV score can be expressed as

$$\nu_o = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{f}_i)^2}{(1 - \boldsymbol{A}_{ii})^2}, \tag{3.7.1}$$

as shown in Section 4.2.3 of [37]. The hat matrix $A$ is defined as $\hat{\boldsymbol{y}} = A\boldsymbol{y}$ where $\hat{\boldsymbol{y}}$ are the fitted values. Generalised cross validation (GCV) is often used in practise since it is less computationally expensive than OCV, see [37], this involves replacing the $A_{ii}$s with their mean and Equation (3.7.1) becomes:

$$\nu_g = n\sum_{i=1}^{n}\frac{(y_i - \hat{f}_i)^2}{(n - tr(A))^2}, \tag{3.7.2}$$

where $tr(A)$ is the trace of $A$. See Pages 169 to 171 of [37] for the derivation of the ordinary and generalised cross validation scores. These methods are computationally expensive for binary data, see Page 7 of [16], thus Section 6.2.5 of [37] advises to adjust the GCV method to use the deviance, which is defined in Equation (3.2.25). Thus the cross validation score becomes:

$$\nu_g(\boldsymbol{\lambda}) = \frac{nD^s(\hat{\boldsymbol{\mu}}, \boldsymbol{y}, \phi)}{(n - \gamma\tau)^2}, \tag{3.7.3}$$

where $\tau = \sum_{i=1}^{n}\frac{\partial\hat{y}_i}{\partial y_i}$ is the model effective degrees of freedom and $\gamma$ is the parameter for the smoothness of the model, which is usually set to 1 [37].
We have now discussed the theory of fitting a GLM and GAM to binomial data, along with how to account for an overdispersed data set. This allows us to move on to preprocessing our data, so that we may fit and select the hail and thunderstorm models.

# 4 Data Analysis: Preprocessing

In this chapter we prepare the raw data to be ready for the statistical analysis in the next chapter. First we ensure that all of the data has the same time and location grid. We then define our dependent variables for the hail and thunderstorm models, and calculate the considered covariates. Finally, we investigate whether we can increase the ratio of successes to failures for both the thunderstorm and hail models, without excluding a large number of successes. The aim of this is to improve the stability of the model fitting algorithms we use later, see [17] for an argument as to the difficulties associated with models for very rare events.

## 4.1 Short Description of Data Sets and Their Indices

Recall the goal to fit two models, one for the probability of a thunderstorm, and a second for the probability of severe hail given that a thunderstorm is occurring.
This thesis combines the following data sources as given in Table 1 to construct the covariates for the thunderstorm and hail models.

| Data set | Description | Source |
|---|---|---|
| Lightning data | The number of cloud to ground lightning detections per grid box | EUCLID `www.euclid.org` |
| Hail data | Reports of hail stones with a diameter larger than 2 cm | ESWD `www.eswd.eu` |
| Reanalysis data | Reanalysis atmospheric conditions within each grid box | ECMWF `www.ecmwf.int` |

Table 1. Description of agencies which provided the data sets used in this thesis.

| Index | Description | Domain |
|---|---|---|
| $i$ | latitude 45°N to 55°N | $45 + 0.25l$, for $l = 0, 1, ..., 40$ |
| $j$ | longitude 5°E to 15°E | $5 + 0.25k$, for $k = 0, 1, ..., 40$ |
| $t$ | time 00:00 01/01/2008 until 23:00 31/12/2017, expressed as hours since 00:00 01/01/2008 | $m = 0, 1, ..., 87672$ |
| $p$ | pressure level 200 hPa to 1000 hPa | $200 + 50n$ for $n = 0, 1, ..., 11$ and $775 + 25q$ for $q = 0, 1, ..., 9$ |

Table 2. All indices and their domains which are discussed in this thesis.

The data frames used in this thesis have indices latitude, longitude and time denoted by $i$, $j$ and $t$ respectively. Some variables used in intermediate calculations also have the index $p$ which represents pressure level, as shown in Table 2. A grid box refers to a unique $i$ (latitude), $j$ (longitude) and $t$ (time), for example 51.0°N, 11.75°E, 15:00 31/1/2008. Table 3 shows all variables used in this thesis. The northward and eastward wind components used to calculate $D_{ijt}$ and the relative humidity levels used to calculate $R_{ijt}$ also have the index pressure level which ranges from 200 hPa to 1000 hPa, measured in 50 hPa intervals from 200 hPa to 750 hPa, then in 25 hPa intervals to 1000 hPa. The variables $S_{ijt}$ and $H_{ijt}$ are the dependent variables for the thunderstorm and hail models respectively. We have defined the detection of a thunderstorm to be the detection of one or more cloud to ground lightning detections within a grid box. The value of the dependent variable $H_{ijt}$ (hail) is 1 if there is a hail case in a grid box where $S_{ijt} = 1$ (thunderstorm case), and 0 if there is no hail case in a grid box where the value of $S_{ijt} = 1$. All grid boxes where $S_{ijt} = 0$ are excluded from the hail model.
An example from the combined, preprocessed data sets is shown in Table 4. Only data points with $S_{ijt}$ (thunderstorm) = 1 and $C_{ijt}$ (CAPE) larger than 100 J kg$^{-1}$ are included in the hail model.

## 4.2 Number of Data Points Within Data Sets

The total number of data points for the thunderstorm and hail models are shown in Table 5. The cutoff values for $C_{ijt}$ (CAPE), below which data points have been excluded, have been calculated and are explained in Section 4.4. Here we note that we have a large data set, however, $s_{hail}$ in particular is small compared to $n_{hail}$, which can be seen in Table 5.

| Variable | Variable Type | Domain | Description |
|---|---|---|---|
| $S_{ijt}$ | Binary | $\{0,1\}$ | Detection of a thunderstorm at latitude $i$, longitude $j$ and time $t$. |
| $H_{ijt}$ | Binary | $\{0,1\}$ | Report of severe hail at latitude $i$, longitude $j$ and time $t$, where $S_{ijt} = 1$ . |
| $C_{ijt}$ | Continuous | $\in \mathbb{R}_{\geq 0}$ | The value of CAPE in $\mathrm{J\,kg^{-1}}$ at latitude $i$, longitude $j$ and time $t$. |
| $D_{ijt}$ | Continuous | $\in \mathbb{R}_{\geq 0}$ | The value of DLS in $\mathrm{m\,s^{-1}}$ at latitude $i$, longitude $j$ and time $t$. |
| $O_{ij}$ | Continuous | $\in \mathbb{R}_{\geq 0}$ | The value of the standard deviation of orography at latitude $i$, longitude $j$ and time $t$. |
| $Z_{ijt}$ | Continuous | $\in \mathbb{R}_{\geq 0}$ | The height in the atmosphere where the temperature is 0 °C, in m at latitude $i$, longitude $j$ and time $t$. |
| $R_{ijt}$ | Continuous | $\in \mathbb{R}_{\geq 0}$ | The average relative humidity in the atmosphere between 500 hPa and 850 hPa at latitude $i$, longitude $j$ and time $t$. |
| $r_{ijtp}$ | Continuous | $\in \mathbb{R}_{\geq 0}$ | The relative humidity in the atmosphere at latitude $i$, longitude $j$, time $t$ and pressure level $p$. |
| $u_{ijtp}$ | Continuous | $\in \mathbb{R}_{\geq 0}$ | The eastward wind speed at latitude $i$, longitude $j$, time $t$ and pressure level $p$. |
| $v_{ijtp}$ | Continuous | $\in \mathbb{R}_{\geq 0}$ | The northward wind speed at latitude $i$, longitude $j$, time $t$ and pressure level $p$. |
| $G_{ijtp}$ | Continuous | $\in \mathbb{R}_{\geq 0}$ | The geo potential at latitude $i$, longitude $j$, time $t$ and pressure level $p$. |
| $h_{ijtp}$ | Continuous | $\in \mathbb{R}_{\geq 0}$ | The height above ground level at latitude $i$, longitude $j$, time $t$ and pressure level $p$. |

Table 3. This table gives an overview of all variables used in this thesis. The first section contains the dependent variables, the second section contains the independent variables used in the thunderstorm and hail models and the third section contains all intermediate variables used for calculations.

| Variable | Name | Value 1 | Value 2 | |
|---|---|---|---|---|
| i | latitude | 51.0 | 51.0 | ... |
| j | longitude | 11.75 | 11.75 | ... |
| t | time | 5:00 1/1/2008 | 6:00 1/1/2008 | ... |
| $S_{ijt}$ | thunderstorm | 1 | 0 | ... |
| $H_{ijt}$ | hail | 0 | 0 | ... |
| $C_{ijt}$ | CAPE [$\mathrm{J\,kg^{-1}}$] | 11.01 | 16.75 | ... |
| $D_{ijt}$ | DLS [$\mathrm{m\,s^{-1}}$] | 39.22 | 37.83 | ... |
| $O_{ij}$ | std. orography | 39.3 | 39.3 | ... |
| $Z_{ijt}$ | zero degree level [m] | 0 | 0 | ... |
| $R_{ijt}$ | relative humidity [%] | 51.85 | 52.03 | ... |

Table 4. This is an example from the data sets after they have been combined for the thunderstorm model and shows some values for the independent and dependent variables.

**Number of Data Points and Successes For The Two Models**

| | |
|---|---|
| $n_{storm}$ = 55,002,099: | The total number of $S_{ijt}$ or data points used for the thunderstorm model. |
| $s_{storm}$ = 1,039,367: | The total number of $S_{ijt} = 1$ or thunderstorm cases reported within $n_{storm}$. |
| $n_{hail}$ = 727,615: | The total number of $H_{ijt}$ or data points used for the hail model . |
| $s_{hail}$ = 2230: | The total number of $H_{ijt} = 1$ or hail cases within $n_{hail}$. |

Table 5. The number of data points and successes for both the thunderstorm and hail models, after preprocessing.

## 4.3 Data Preprocessing

The lightning, reanalysis and hail data sets were preprocessed using python, predominantly so that the packages xarray [14] and iris [21] could be used. These packages have functions built to specifically deal with large files of type network common data form (netcdf), and have the advantage of not immediately loading the entire data frame into working memory. This makes selecting specific parts of the data frame quicker and simpler.

### 4.3.1 Lightning Data Preprocessing

The lightning detection data was provided by the European Cooperation for Lightning Detection (EUCLID) which has a network of sensors to detect cloud to ground lightning strikes. The number of lightning strikes per hour within each 0.25° by 0.25° grid box is detected. We were only interested in whether a thunderstorm occurred or not, so the counts of lightning detections were converted to a binary format to indicate the presence of at least one cloud to ground lightning detection, and therefore a thunderstorm. Thunderstorms are also associated with intra-cloud lightning, typically producing a combination of both types of lightning, however, we assume that the number of thunderstorms with only intra-cloud lightning is negligible. We see in Section 4.3.2 that this assumption is consistent with our hail data.

### 4.3.2 Hail Data Preprocessing

The hail report data was provided by the European Severe Weather Database (ESWD) as a comma separated values (csv) file with the report location in latitude and longitude, location accuracy, observation quality, time, the time accuracy of the report and hail stone diameter or hail depth. The ESWD have four grades for the quality of reports, see [18], which are listed below in order of lowest to highest quality:

- *QC0* : received but unverified report
- *QC0+* : report is plausible, for example weather reports show thunderstorm activity in the area at the time of the reported incident.
- *QC1* : report has been confirmed by a reliable source
- *QC2* : report is part of a scientific study

In this thesis events with the quality grade QC0 were excluded due to the lack of reliability in the reports. It is also worth noting that very few observations satisfy the QC2 requirements. Reports with a time accuracy of plus minus thirty minutes or less were considered and the rest excluded. These measures were taken to ensure that the hail reports corresponded to the correct storm and atmospheric conditions. The ESWD records events with either hail stones larger than 2 cm in diameter, or very large quantities of small hail which result in a layer of hail lying on the ground which is 10 cm or deeper. Events with only depth recorded were excluded because this thesis is specifically interested in the occurrence of severe hail events where the diameter of hail stones is 2 cm or larger. Table 6 shows these requirements for hail reports to be included in tabular form,

**Requirements for Hail Report Inclusion**

| Report Quality | Time Accuracy | Hail Stone Diameter [cm] |
|---|---|---|
| QC0+, QC1 or QC2 | $\leq \pm 30$ minutes | $\geq 2$ cm |

Table 6. The requirements for hail reports to be included in this thesis.

**Examples of Included and Excluded Hail Reports**

| included / excluded | quality | time accuracy | hail stone diameter | depth on ground |
|---|---|---|---|---|
| included | QC1 | $\pm 15$ minutes | 3 cm | na |
| excluded (diameter) | QC0+ | $\pm 5$ minutes | 1 cm | 11 cm |
| excluded (time accuracy) | QC0+ | $\pm 60$ minutes | 4 cm | na |

Table 7. Examples of hail reports which would be included or excluded from analysis in this thesis.

and Table 7 shows some examples of hail reports, their report information and whether they would be included or excluded.

The goal of the hail data preprocessing was to create a data frame in the same format as the lightning data set. The ESWD requirement for hail reports with quality grades QC0+, QC1 and QC2 is that thunderstorm activity is present in the general vicinity of the report, which is usually checked using radar records. Furthermore, it is unphysical for hail to occur without the presence of a thunderstorm. This means we expect each report of severe hail in the hail data set to coincide in time and space with a cloud to ground lightning detection in the lightning data set. It can be seen in Table 8 that only 16 hail reports were excluded because they did not coincide with a lightning detection. Thus we conclude that our definition of hail occuring during a thunderstorm is consistent with our data.

Many hail reports had time or location uncertainties which made it unclear which grid box the hail report took place in. For this reason a matching algorithm was developed to determine the closest grid box to the hail report which also had a thunderstorm. The grid box in which the hail event was reported, and the surrounding 26 ($3^3$ - 1 = 26) grid boxes with plus or minus one grid box in terms of time, latitude and or longitude were investigated. The distance from the hail report to the centre of each grid box was calculated for time, latitude and longitude, in terms of grid boxes. These were sorted and the grid box with the smallest distance to the hail report which also had a lightning report

| | |
|---|---|
| Number of reports not matched to a grid box with a thunderstorm | 16 |
| Number of reports matched to a grid box with a thunderstorm | 3777 |
| Number of cases in total | 2445 |

Table 8. The number of hail reports from the ESWD which have and have not been matched to a grid box where there is a thunderstorm ($S_{ijt} = 1$), and the total number of $H_{ijt}$ or hail cases. Note that mutiple reports matched to a single grid box are counted as one case.

was defined to be the location of the hail report.

### 4.3.3 Disagreement Between Lightning and Hail

Despite attempts to match hail reports with lightning detections within the vicinity of each hail report, not all hail reports could be matched. Some of these hail reports were investigated in detail. The case in Figure 4 shows a hail report with surrounding cloud to ground lightning activity, but outside of the neighbouring grid cells which were investigated. Other reports also showed lightning activity in the general vicinity of unmatched hail reports. This could also occur when only cloud to cloud lightning is present, when high winds blow hail a long way before it reaches the ground, or for storm clouds which are strongly tilted, resulting in hail landing on the ground at a different location to the lightning activity. A successfully matched case is shown in Figure 5.

### 4.3.4 Atmospheric Data Preprocessing

The data for atmospheric conditions was downloaded from European Centre for Medium-Range Weather Forecasts (ECMWF). The variables $C_{ijt}$ (CAPE), $Z_{ijt}$ (zero degree level) and $O_{ij}$ (standard deviation of orography) were downloaded with no additional preprocessing.
The variable $D_{ijt}$ (DLS) was calculated as in Equation (2.5.1) using $v_{ijtp}$ (the northward wind component) and $u_{ijtp}$ (the eastward wind component) downloaded from ECMWF. Furthermore, the pressure level corresponding to a height of 6 km needed to be calculated. The height was calculated using Equation (4.3.1):

$$h_{ijtp} = \frac{G_{ijtp}}{g}\,\mathrm{m}, \tag{4.3.1}$$

where the acceleration due to gravity, $g = 9.8067\,\mathrm{m\,s^{-2}}$.
After determining the $h_{ijtp}$ for each unique combination of $i$, $j$, $t$ and $p$, we then needed to calculate, for each unique combination of $i$, $j$ and $t$, for which $p$ is the value of $h_{ijtp}$ closest to 6000 (6 km).
The variable $R_{ijt}$ (relative humidity) was calculated as an average of $r_{ijtp}$ (relative humidity at each pressure level) over the pressure levels from 500 hPa to 8500 hPa, as shown in Equation (4.3.2)

$$R_{ijt} = \frac{\sum_{p=500}^{850} r_{ijtp}}{P}, \tag{4.3.2}$$

where $P = 10$ is the number of pressure levels between 500hPa and 850hPa.

## 4.4 Reducing The Number of Non Hail Cases Using Physical Justifications

In this section we want to increase the ratio of successes to failures for our hail model. We know from common experience that hail stones larger than 2cm in diameter are very rare. From a societal perspective this is good because the damage caused by severe hail is often expensive to repair, see [25], and at its most extreme can lead to loss of life, see

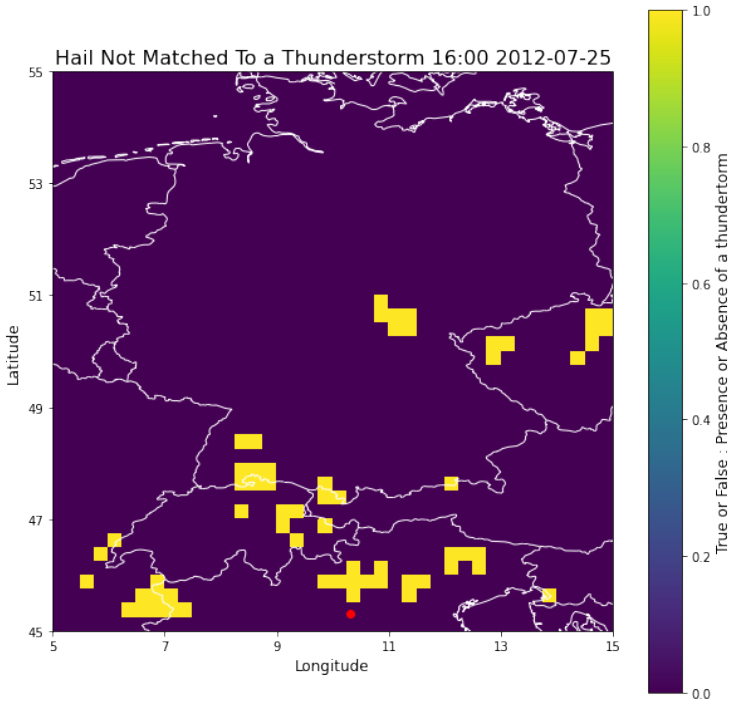


Figure 4. The presence of thunderstorms is shown in yellow, and a hail report is marked in red. This shows one of the 16 hail reports which was not matched to a thunderstorm. Thunderstorm activity is visible in the image, however it is further away from the report than the matching algorithm takes into account, and it cannot be known with 100 percent accuracy why this hail report did not coincide with a thunderstorm, therefore it is excluded from the analysis.

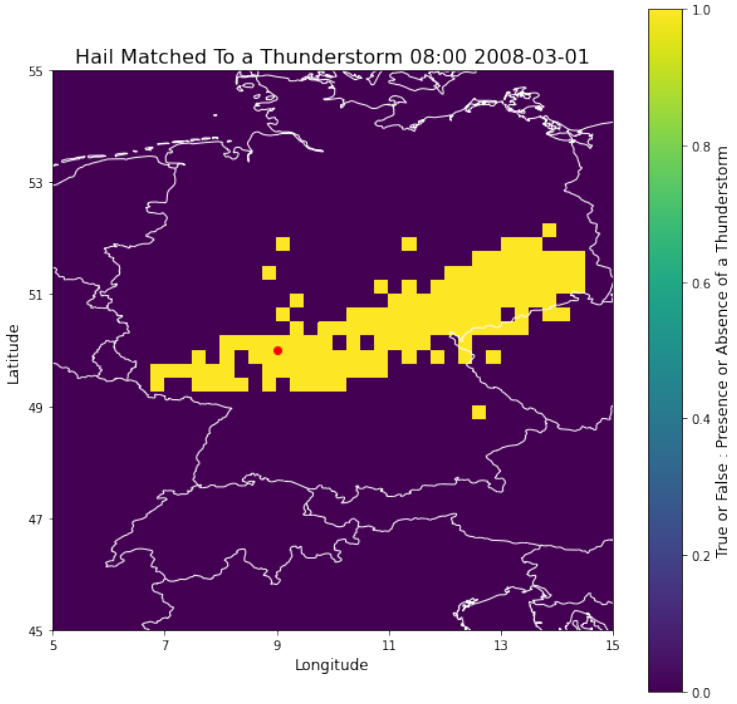


Figure 5. The presence of thunderstorms is shown in yellow, and a hail report is marked in red. This shows a hail report which was matched to a thunderstorm, and recorded as a case.

[3]. However, as we discussed at the beginning of this chapter, we want to increase the ratio of successes to failures to improve the stability of the model fitting algorithm used next chapter. Thus we use our knowledge from Section 2 to exclude data which represents physical conditions not favourable for the growth of hail stones.

Atmospheric conditions with very low CAPE are not favourable conditions for hail formation and growth. This is because the maximum updraft speed provides the effect which opposes the effect of gravity acting on a hail stone as seen on Page 226 of [5]. We look for values of $C_{ijt}$ and $D_{ijt}$ below which we see very few $H_{ijt} = 1$ (hail cases), in this case less than 10 percent of the total number of $H_{ijt} = 1$. $C_{ijt}$ and $D_{ijt}$ were investigated because these two covariates were selected first by stepwise regression shown in Table 16 of the Appendix 7, however, we expect $D_{ijt}$ to show little effect because it is physically possible for severe hail to grow in conditions of very low DLS as seen in Section 2.1.5 of [28].
We constructed heatmap style plots to investigate the frequency of $H_{ijt} = 1$ and the ratio of $H_{ijt} = 1/H_{ijt} = 0$, whilst excluding data points with progressively larger values of $C_{ijt}$ and $D_{ijt}$.
Figure 6 shows the frequency of $H_{ijt} = 1$ excluding data points below the values of $C_{ijt}$ and $D_{ijt}$. The red contour lines showing this as a fraction of the total number of $H_{ijt} = 1$. For example the top right red point ($C_{ijt}$, $D_{ijt}$) = (140,20) includes only data points with $C_{ijt} > 140$ and $D_{ijt} > 20$. We can see that approximately 1000 $H_{ijt} = 1$ satisfy this condition.
The bottom left red point at ($C_{ijt}$ , $D_{ijt}$ ) = (50,0) includes only data points with $C_{ijt} > 50$ and $D_{ijt} > 0$. Approximately 2300 $H_{ijt} = 1$ (hail cases) satisfy this condition. This point is also close to the contour line of 0.95, which shows where 95 percent of $H_{ijt} = 1$ remain and 5 percent of $H_{ijt} = 1$ have been excluded.
Figure 7 shows the factor by which the ratio of $H_{ijt} = 1$ to $H_{ijt} = 0$ is increased for given $C_{ijt}$ and $D_{ijt}$ values, below which all data is excluded in the same way as above.
We use Figure 8 to select $C_{ijt}$ and $D_{ijt}$ 'cutoff' values below which all data is excluded. We want to select values which lead to the largest improvement in the ratio of $H_{ijt} = 1$ to $H_{ijt} = 0$ whilst excluding at most 10 percent of $H_{ijt} = 1$. This means we choose $C_{ijt}$ and $D_{ijt}$ values close to the red 0.90 contour with the largest value shown by the colour map. Values of $C_{ijt} = 100$ and $D_{ijt} = 0$ were chosen because meteorologists often work with CAPE in steps of $50\,\mathrm{J\,kg^{-1}}$, the ratio of $H_{ijt} = 1$ to $H_{ijt} = 0$ was improved and less than 10 percent of $H_{ijt} = 1$ were excluded.

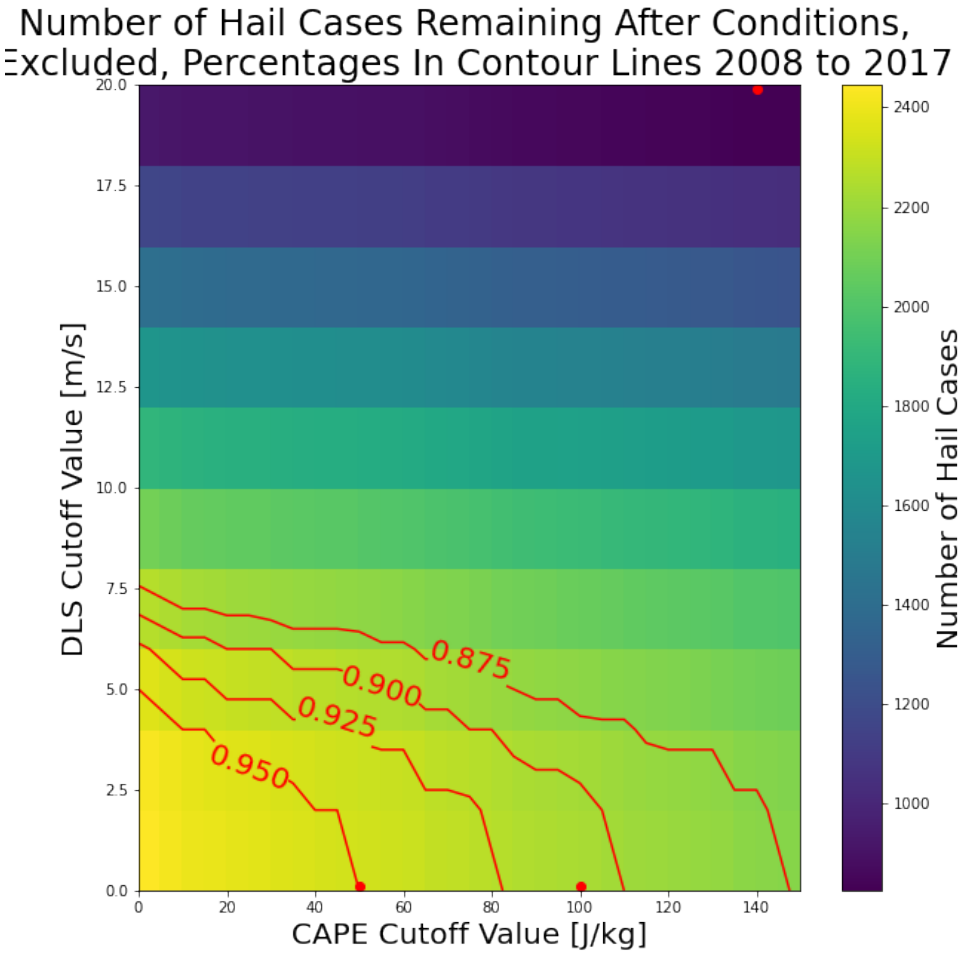


Figure 6. This plot shows the frequency of $H_{ijt} = 1$ when excluding all data points below the values of $C_{ijt}$ and $D_{ijt}$, with the contour lines showing the fraction relative to the total number of $H_{ijt}$.

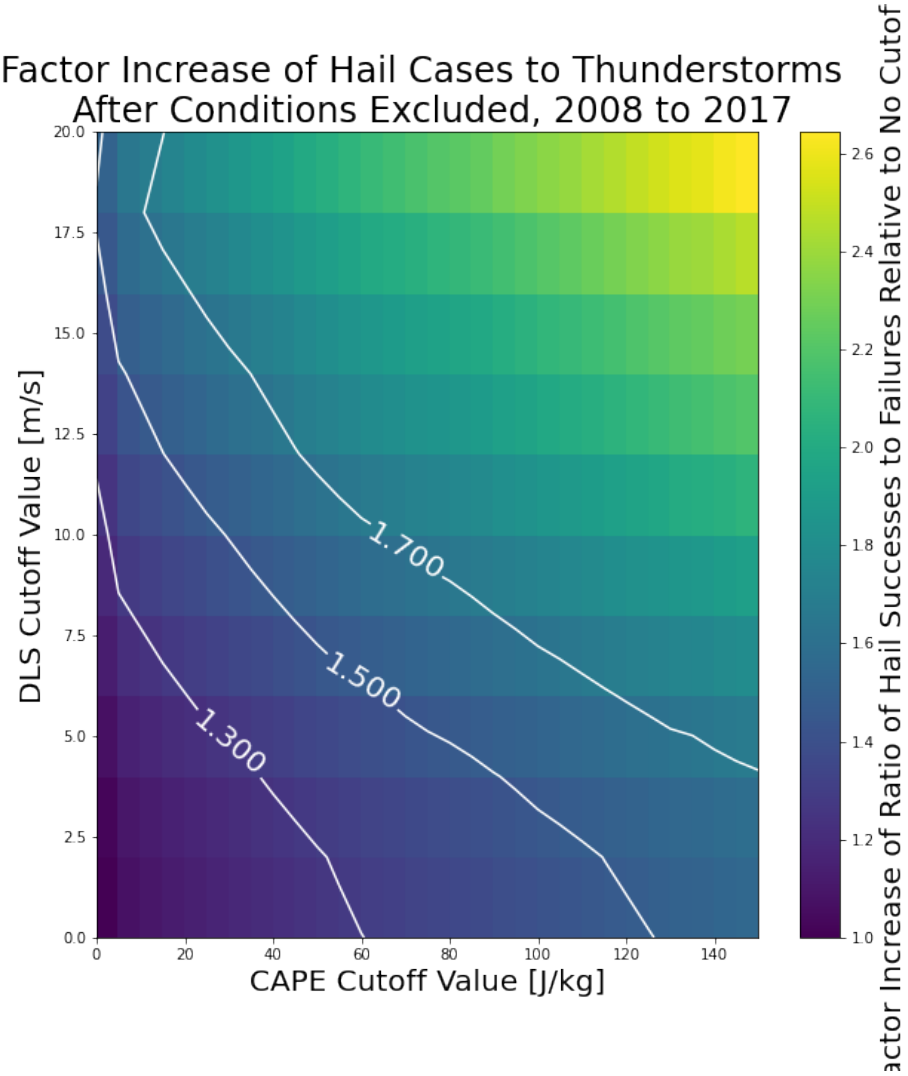


Figure 7. This plot shows the factor increase in the ratio of $H_{ijt} = 1$ to $H_{ijt} = 0$ using the same cutoffs for $C_{ijt}$ and $D_{ijt}$ as the above figure, with contours are overlaid in white.

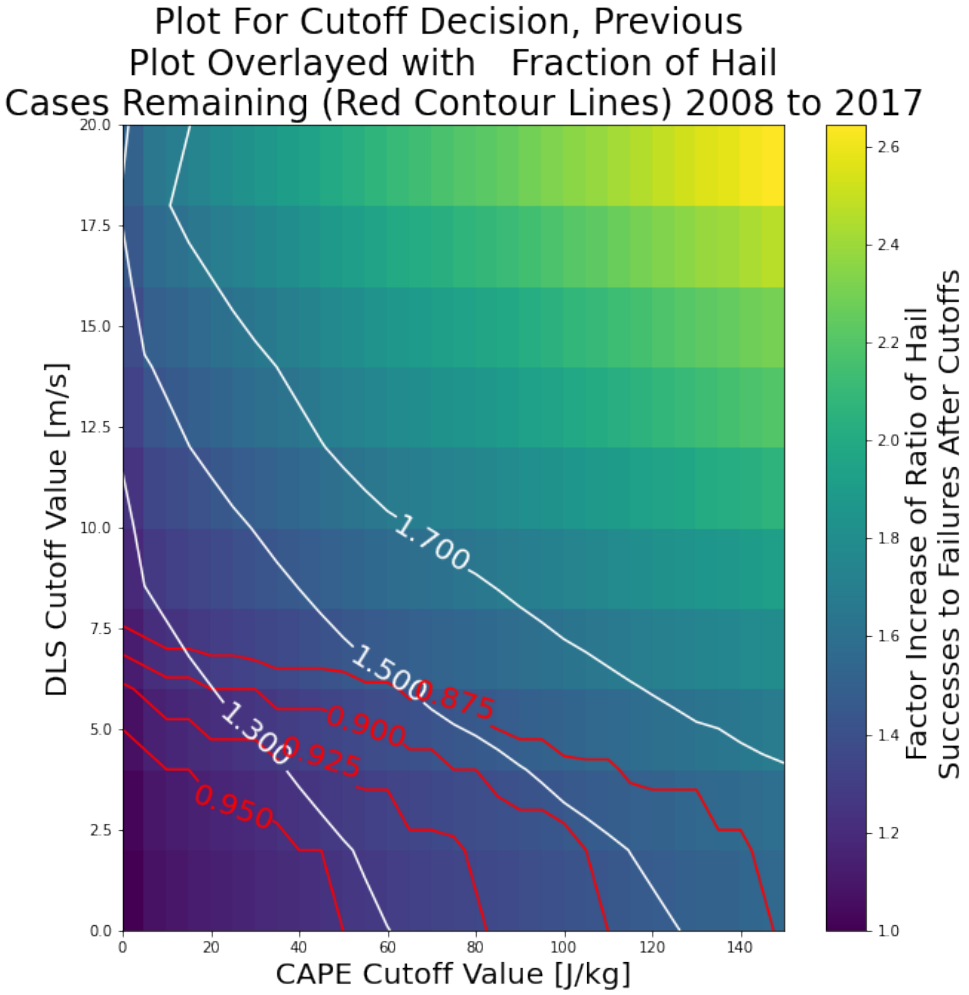


Figure 8. This plot shows Figure 7 overlaid with the red contour lines from Figure 6. The x axis shows $C_{ijt}$, and the y axis shows $D_{ijt}$. This plot was used to decide on the best choice of cutoff values for $C_{ijt}$ and $D_{ijt}$ to increase the ratio of $H_{ijt} = 1$ to $H_{ijt} = 0$ with at least 90 percent of the $H_{ijt} = 1$ remaining.

## 4.5 Reducing The Number of Non Thunderstorm Cases Using Physical Justifications

In this section we want to increase the ratio of successes to failures for our thunderstorm model as we did for the hail model. Thunderstorms require warm, moist air to initiate as discussed in Section 2.2. As expected, Table 18 in the Appendix 7 shows that $C_{ijt}$ and $R_{ijt}$ are the first two covariates chosen when performing a step wise regression.

We constructed heatmap style plots as in the previous section to investigate the frequency of $S_{ijt} = 1$ and the ratio of $S_{ijt} = 1/S_{ijt} = 0$ whilst excluding all data points with increasing values of $C_{ijt}$ and $R_{ijt}$. Using Figure 9 $C_{ijt} = 5$ and $R_{ijt} = 0$ were chosen as cutoff values. Additional plots for the number of $S_{ijt} = 1$ excluded and the increase factor for the ratio of $S_{ijt} = 1$ to $S_{ijt} = 0$ can be seen in the the Appendix 7.1. In Chapter 5 exploratory data analysis will be performed using the points remaining after all of the data preprocessing steps have been completed.
We have now formatted and processed our data such that we are ready to perform our statistical analysis and choose our thunderstorm and hail models in the following chapter.

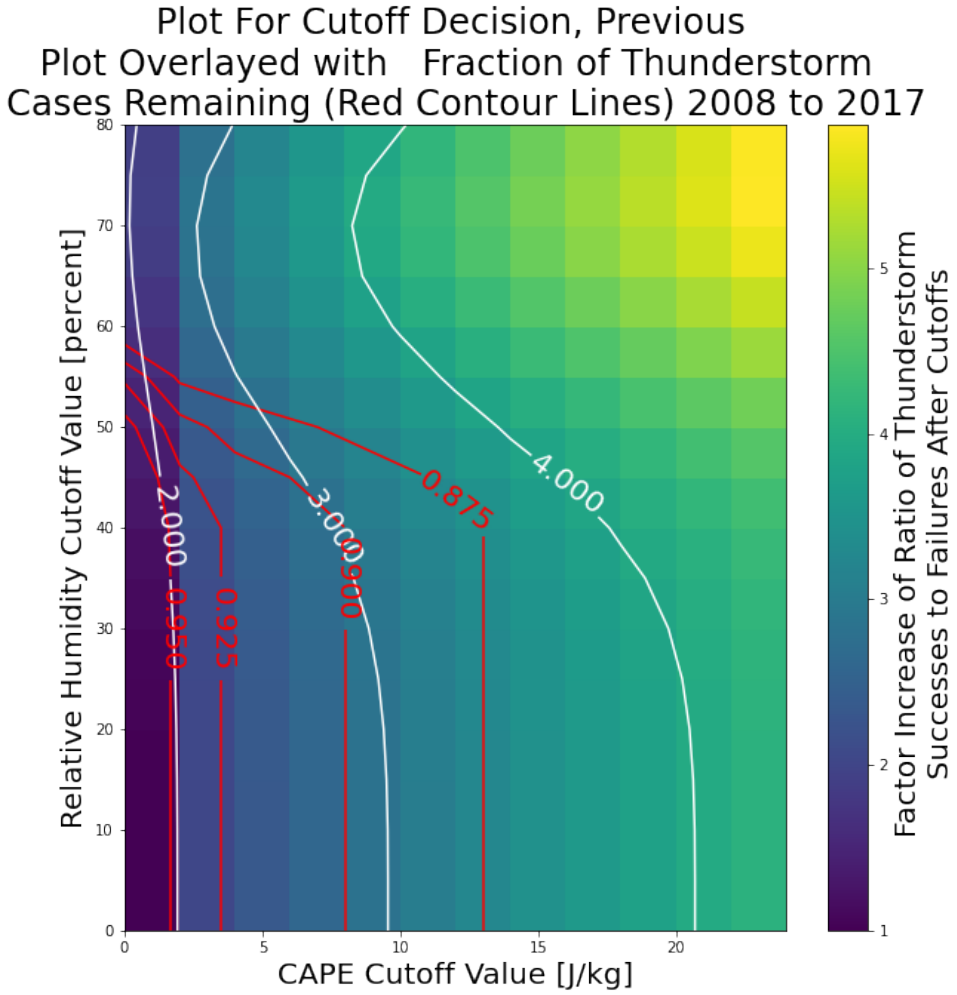


Figure 9. This plot is analogous to Figure 8, but for the thunderstorm model. The x axis shows $C_{ijt}$, and the y axis shows $R_{ijt}$. This plot was used to decide on the best choice of cutoff values for $C_{ijt}$ and $R_{ijt}$ to increase the ratio of $S_{ijt} = 1$ to $S_{ijt} = 0$ with at least 90 percent of the $S_{ijt} = 1$ remaining.

# 5 Exploratory Data Analysis

The Sections 5.1 and 5.3 show the results from the exploratory data analysis (EDA) for the thunderstorm and hail models with covariates $C_{ijt}$ (CAPE), $D_{ijt}$ (DLS), $O_{ij}$ (standard deviation of orography), $Z_{ijt}$ (zero degree level) and $R_{ijt}$ (relative humidity). Pairs plots, histograms, box plots and empirical logit plots were calculated after the preprocessing in Section 4 was complete. All plots except for the surface plots were constructed using the ggplot2 package, see [34], in R. The upper surface plots were constructed using the package MGCV from [36] and the lower surface plots were constructed using the package plot3D [29].
Models have been suggested for both the hail and thunderstorm cases using data which has been grouped, and are discussed in Sections 5.2 and 5.4.

## 5.1 EDA Plots For Hail Model

This section contains the pairs plots, box plots, histograms, empirical logit plots, transformation analysis and interaction effect plots for the hail model. These were calculated after the data preprocessing was complete, so all data points have $C_{ijt} \geq 100$. The first set of plots we look at are pairs plots in Figure 10 and 11. We see no strong relationship between any of the covariates, however, the large number of data points make these plots inconclusive.

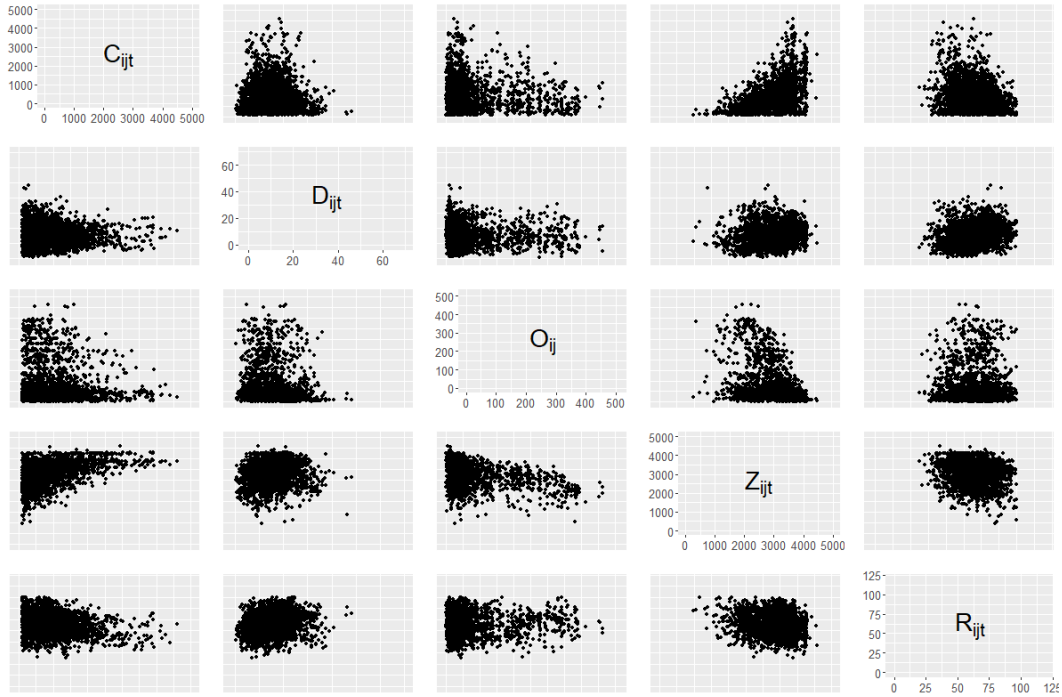


Figure 10. **Hail Data**: Pairs plots of the covariates ($C_{ijt}$, $D_{ijt}$, $O_{ij}$, $Z_{ijt}$ and $R_{ijt}$) calculated from the reanalysis data for data points only with $H_{ij} = 1$.

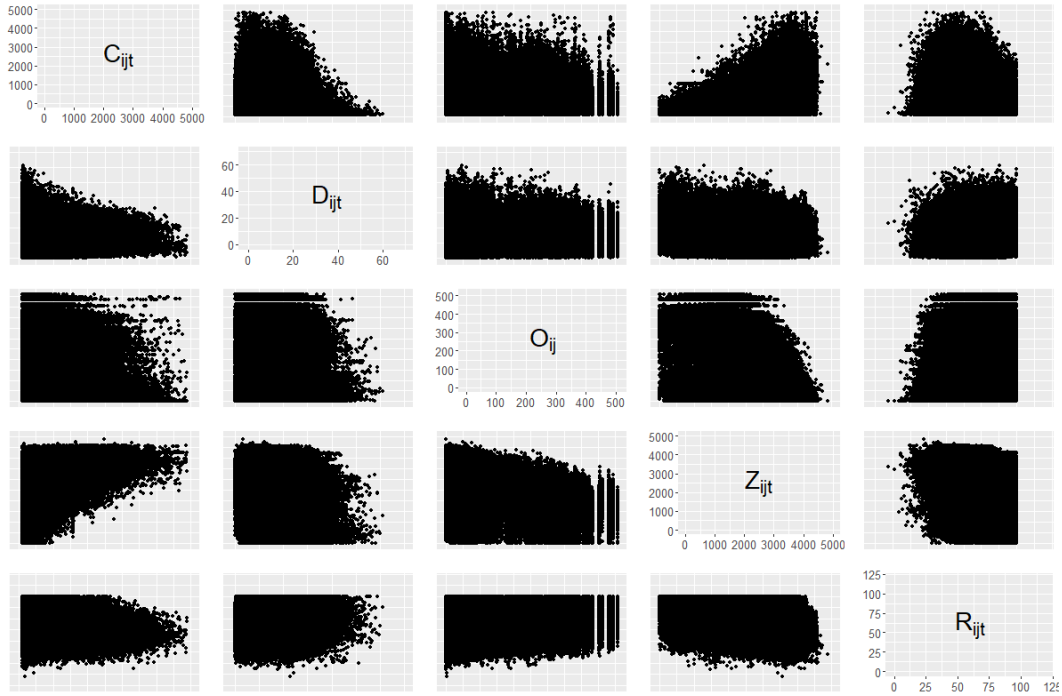


Figure 11. **Hail Data**: Pairs plots of the covariates ($C_{ijt}$, $D_{ijt}$, $O_{ij}$, $Z_{ijt}$ and $R_{ijt}$) calculated from the reanalysis data for data points only with $H_{ij} = 0$.

### 5.1.1 Analysis of Individual Considered Covariates

We now want to analyse the covariates individually. The covariates have been grouped with respect to their quantiles, and the mean defined as the score for each interval so that the logits could be plotted against the scores and the assumption of linearity checked. The two most extreme groups have been excluded because these represent very rare meteorologial conditions. For the covariates $O_{ij}$ and $R_{ijt}$ an additional two groups of the smallest values were removed. This was because many data points had the same value, so multiple groups had the same score which caused issues computationally. The first plots for the individual covariate analysis that we look at are the box plots. In Figure 12 we see that $C_{ijt}$ has a moderate effect, with the plots of remaining covariates being inconclusive. Following this we look at the histograms shown in Figure 13, where we see that our data is not uniformly distributed over all covariate values, as expected. Using this information we decided to calculate covariate groups with respect to quantiles. We decided on 50 groups as this provided a compromise between enough groups to see relationships in the plots of the empirical logits, which are discussed next and shown in Figure 14, and causing numerical issues due to multiple groups having the same covariate value.
We now investigate the empirical logits of the grouped covariates in Figure 14. Here we see the original empirical logits, the empirical logits after a transformation of the covariate and the studentised residuals of a polynomial fit of the transformed variables as covariates and the empirical logits as response. All covariates appear to have non-linear relationships with the response variable empirical logit. Log transforms were chosen for the covariates $C_{ijt}$, $D_{ijt}$ and $O_{ij}$ based on the shape of the empirical logit plots. The covariates $Z_{ijt}$ and $R_{ijt}$ were not transformed.

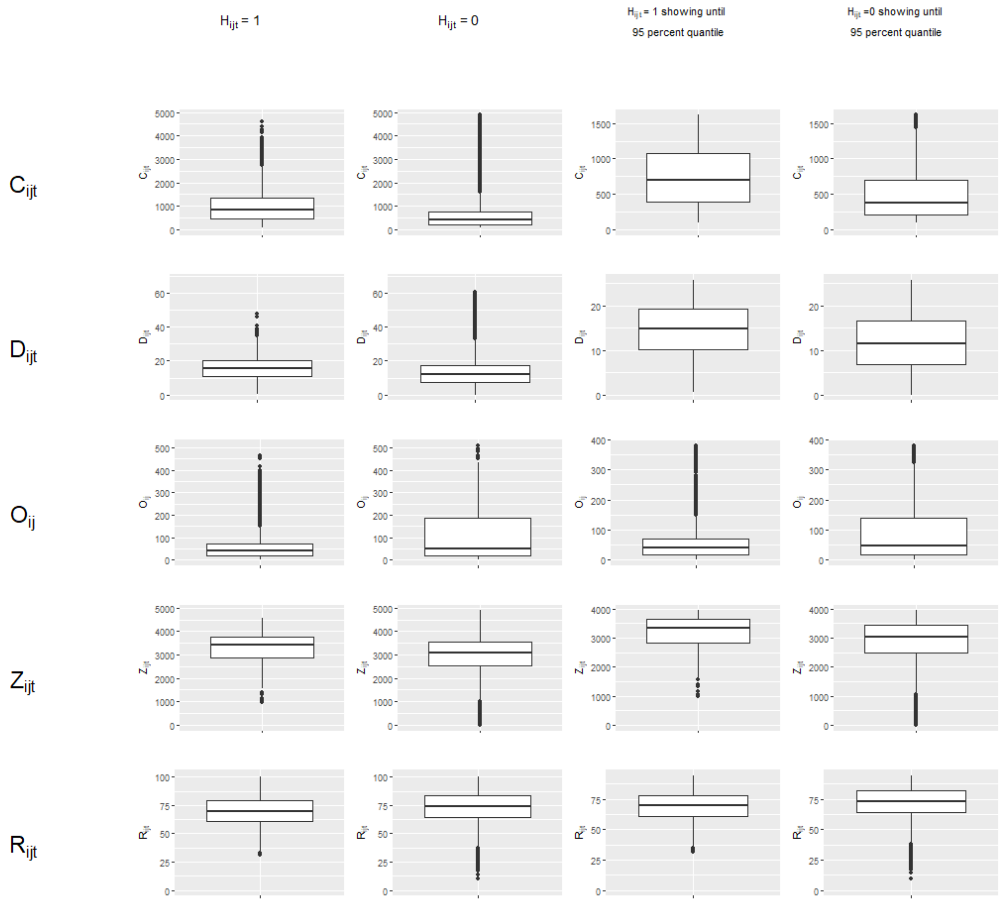


Figure 12. **Hail Data**: Box plot of the covariates ($C_{ijt}$, $D_{ijt}$, $O_{ij}$, $Z_{ijt}$ and $R_{ijt}$). The box plots were calculated separately for data points with $H_{ijt} = 1$ and $H_{ijt} = 0$. The two left most columns show the extremes, and the two right most columns show the whiskers of the same box plots.

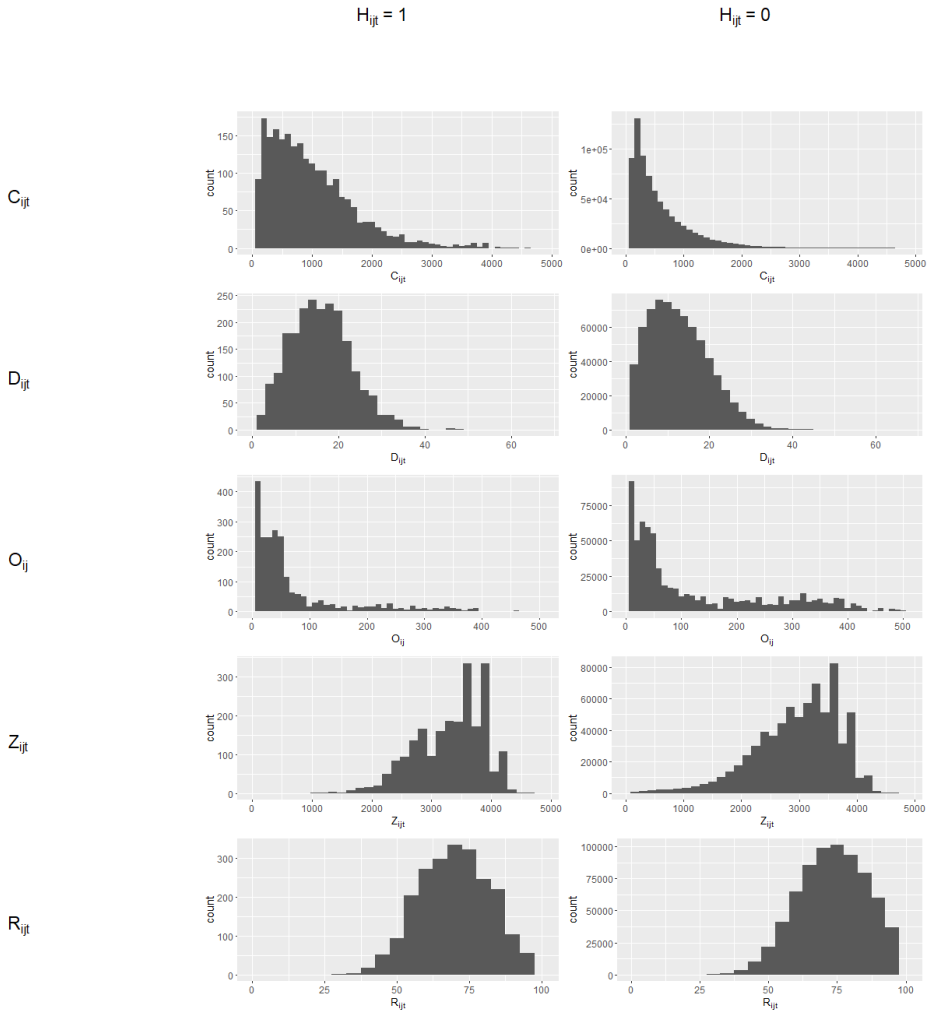


Figure 13. **Hail Data**: Histograms of the covariates ($C_{ijt}$, $D_{ijt}$, $O_{ij}$, $Z_{ijt}$ and $R_{ijt}$). Histograms were calculated separately for data points with $H_{ijt} = 1$ (left column) and $H_{ijt} = 0$ (right column).

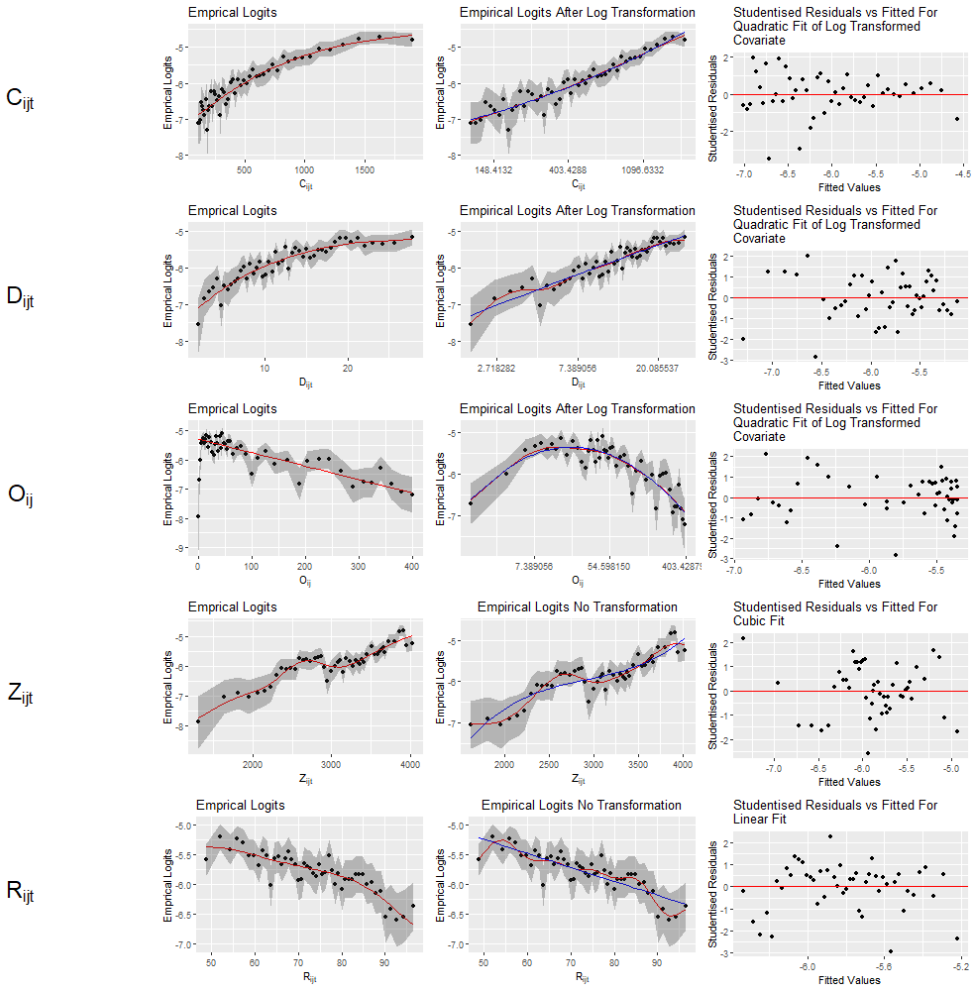


Figure 14. **Hail Data**: These plots show the empirical logits calculated for all grouped covariates ($C_{ijt}$, $D_{ijt}$, $O_{ij}$, $Z_{ijt}$ and $R_{ijt}$), before and after transformations have been applied. The empirical logit points calculated with the covariates prior to transformation are shown with their 95 percent confidence band. The red line shows a GAM fit. $C_{ijt}$, $D_{ijt}$ and $O_{ij}$ have been transformed with the log function. The empirical logit points calculated with the transformed covariates are shown with their 95 percent confidence band. The red line shows a GAM fit and the blue line shows a polynomial fit, the equations for which can be seen in Models (5.2.1),(5.2.2), (5.2.3), (5.2.4) and (5.2.5). Additionally, the studentised residuals of polynomial fits using the empirical logits, $l$, as response and the transformed variables, have been plotted against their fitted values, with the zero line shown in red. For an $i^{th}$ order polynomial of covariate $x$ with coefficient vector $\boldsymbol{\beta} = (\beta_0, ..., \beta_i)^T$ the residuals of the following fit would be plotted: $l = \boldsymbol{\beta} \times poly(x, i)$.

### 5.1.2 Analysis of Interaction Effects

We now investigate for any possible interaction effects using the transformed covariates from the previous section, in Figure 15. For each interaction effect we investigate we must now consider a grouping with respect to two covariates. Thus each covariate is divided into 5 groups and the empirical logits calculated for 25 groups for each intereaction effect plot. This number of groups was chosen as a compromise between the number of groups being large enough such that any relationships could be seen whilst ensuring that the number of observations in each group was large enough (or the number of groups small enough) so that the empirical logits were smooth enough and not dominated by noise.
The plots showing the interactions of $C_{ijt}$ and $D_{ijt}$, $D_{ijt}$ and $O_{ij}$ and $C_{ijt}$ and $Z_{ijt}$ showed some cross over between groups. Interaction effects between these covariates are also of interest from a meteorological perspective based on consultation with meteorologists at Munich Re and from the theory on supercellular thunderstorms discussed in Chapter 2. The fact that the surfaces in Figures 16, 17 and 18 are not planes suggests that there could be interaction effects. Furthermore we see that the shapes of the GAM tensor and empirical logit surfaces are in general agreement.

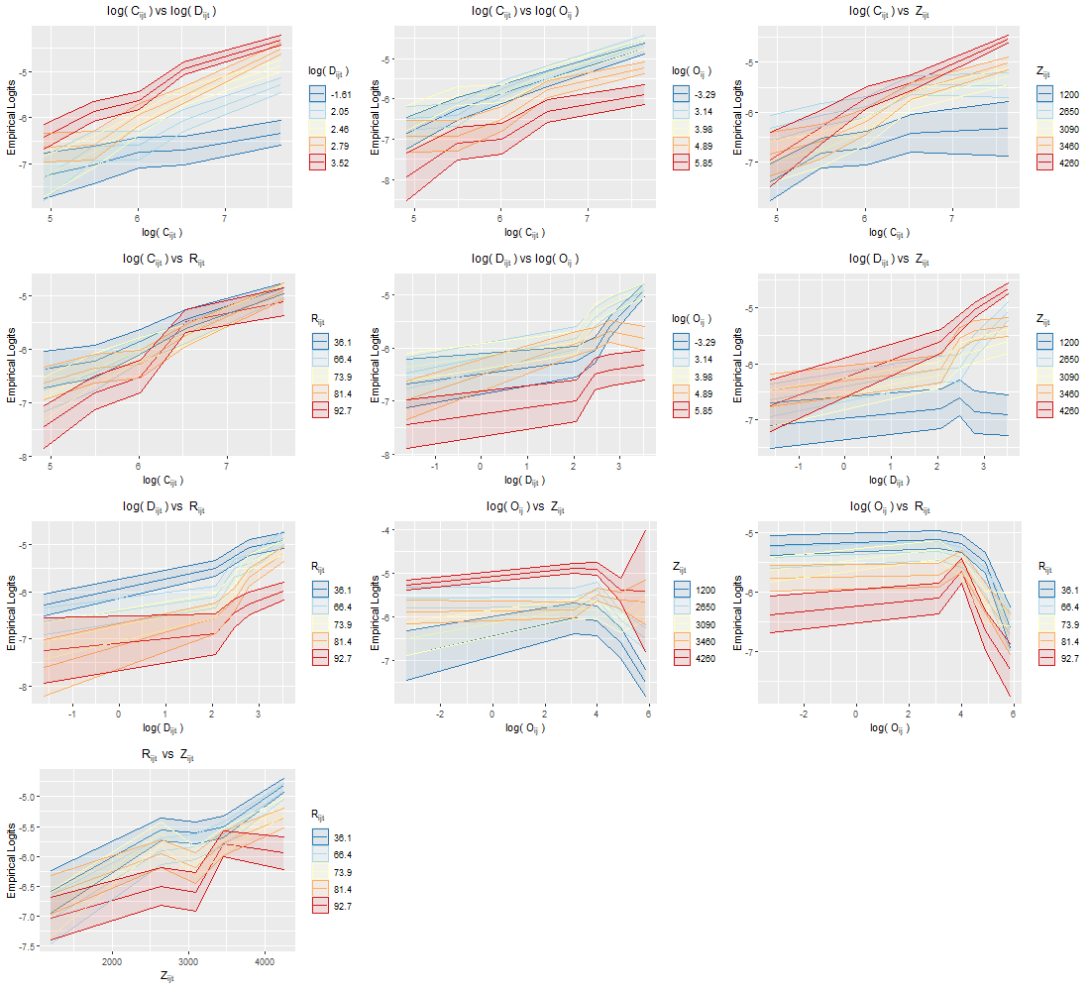


Figure 15. **Hail Data**: Interaction effects between the transformed covariates ($C_{ijt}$, $D_{ijt}$, $O_{ij}$, $Z_{ijt}$ and $R_{ijt}$). $C_{ijt}$, $D_{ijt}$ and $O_{ij}$ have been transformed with the log function. The empirical logits have been calculated where the covariates have been grouped into 25 (5 x 5 = 25) groups with respect to their quantiles and are shown along with their confidence limits.

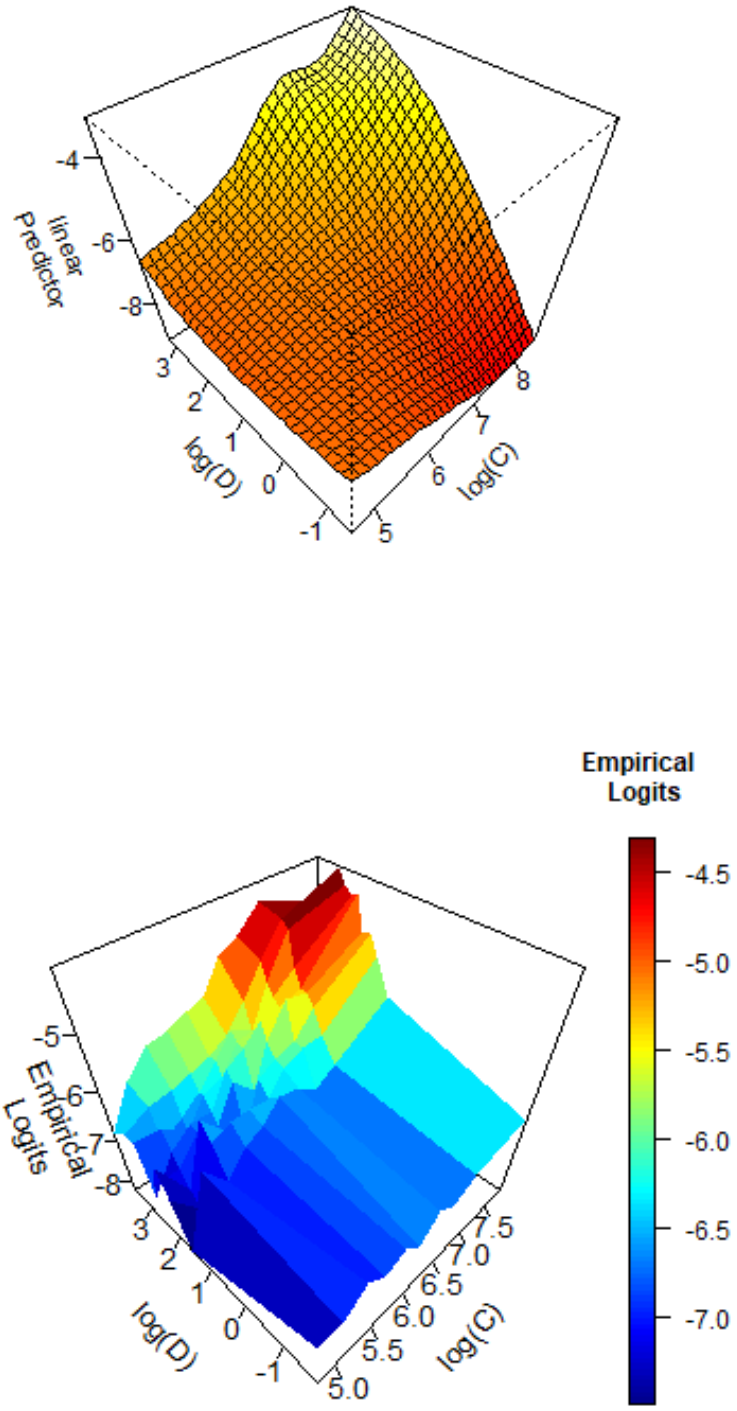


Figure 16. **Hail Data**: Interaction effects between the covariates $\log(C_{ijt})$ and $\log(D_{ijt})$ and the response $H_{ijt}$ using a GAM tensor product surface (above) and the empirical logits (below).

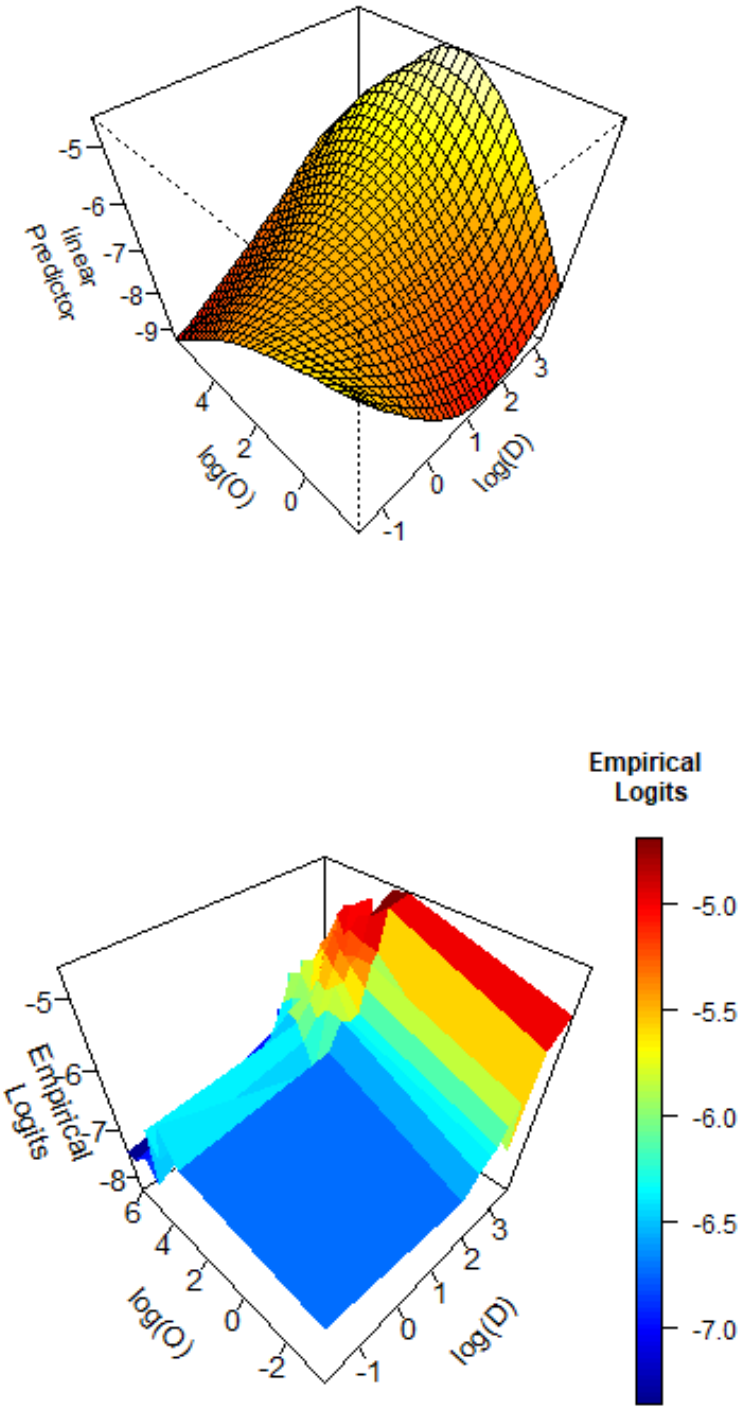


Figure 17. **Hail Data**: Interaction effects between the covariates $\log(D_{ijt})$ and $\log(O_{ijt})$ and the response $H_{ijt}$ using a GAM tensor product surface (above) and the empirical logits (below).

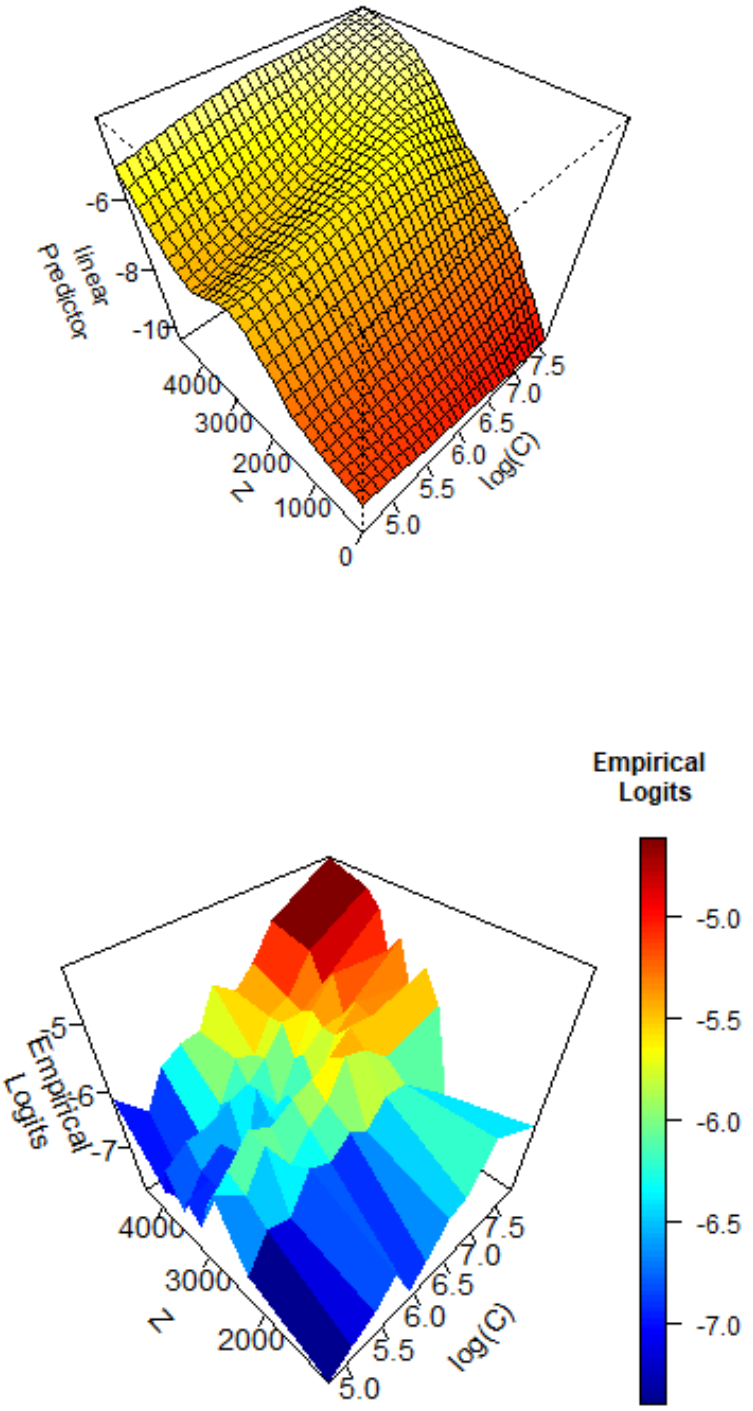


Figure 18. **Hail Data**: Interaction effects between the covariates $\log(C_{ijt})$ and $Z_{ijt}$ and the response $H_{ijt}$ using a GAM tensor product surface (above) and the empirical logits (below).

## 5.2 Fitting of Models on Grouped Data For The Hail Model Case

We now investigate several proposed models for grouped data in the hail model case. We continue using the groups from the previous section where each covariate was divided into 5 levels. This provides a trade off between computational speed for model fitting and retaining the information contained in, or not over simplifying, the full data set. This results in a reduction of the hail data from 727615 to 3125 ($5^5 = 3125$) data points. Thus, $C_{ijt}$, $D_{ijt}$, $O_{ij}$, $Z_{ijt}$ and $R_{ijt}$ now correspond to $C_k$, $D_k$, $O_k$, $Z_k$ and $R_k$ respectively, for $k = 1, ..., 3125$. Note that where higher order effects have been included, the function 'poly' from the software R has been used due to its computational stability. This is denoted $poly(X, a)b$ where $X$ is the covariate, $a$ the highest order of the polynomial and $b$ the order of the specific term.

### 5.2.1 Univariate Binomial Models

For each covariate we now consider a univariate binomial regression: a logistic model with a quadratic effect for the log transform of $C_k$,

$$M^H_{logC2} : \text{ logit}(p_k) = \beta_0 + \beta_1 poly(\log(C_k), 2)1 + \beta_2 poly(\log(C_k), 2)2 \tag{5.2.1}$$

a logistic model with a quadratic effect for the log transform of $D_k$,

$$M^H_{logD2} : \text{ logit}(p_k) = \beta_0 + \beta_1 poly(\log(D_k), 2)1 + \beta_2 poly(\log(D_k), 2)2, \tag{5.2.2}$$

a logistic model with a quadratic effect for the log transform of $O_k$,

$$M^H_{logO2} : \text{ logit}(p_k) = \beta_0 + \beta_1 poly(\log(O_k), 2)1 + \beta_2 poly(\log(O_k), 2)2, \tag{5.2.3}$$

a logistic model with a cubic effect for $Z_k$,

$$M^H_{Z3} : \text{ logit}(p_k) = \beta_0 + \beta_1 poly(Z_k, 3)1 + \beta_2 poly(Z_k, 3)2 + \beta_3 poly(Z_k, 3)1, \tag{5.2.4}$$

and a logistic model with a linear effect for $R_k$,

$$M^H_R : \text{ logit}(p_k) = \beta_0 + \beta_1 R_k, \tag{5.2.5}$$

for $k = 1, ..., 3125$, where the $\beta_k$ are the regression coefficients. The models are named using superscript to indicate the dataset, in this case $H$ to denote hail data, and subscript to denote the covariates included in the model. The summaries for Models (5.2.1), (5.2.2), (5.2.3), (5.2.4) and (5.2.5) are shown in Table 9 for the grouped data. We see that for each of the univariate models the residual deviance is much larger than the residuals degrees of freedom, thus showing a lack of fit for all of these models. This leads us to next consider a model with multiple covariates. As the highest order term is significant in all of the univariate models we need to include both the highest order and all lower order terms for each of the covariates. Summaries for the ungrouped data are shown in the Appendix 7 in Table 17.

| $M^H_{logC2}$ | Estimate | Std. Error | z value | Pr(> \|z\|) |
|---|---|---|---|---|
| (Intercept) | -6.0033 | 0.0264 | -226.97 | 0.0000 |
| poly(log(C),2)1 | 37.8838 | 1.3765 | 27.52 | 0.0000 |
| poly(log(C),2)2 | -4.1683 | 1.4100 | -2.96 | 0.0031 |
| Null deviance: 4930.8 on 3068 degrees of freedom | | | | |
| Residual deviance: 3890.3 on 3066 degrees of freedom | | | | |

| $M^H_{logD2}$ | Estimate | Std. Error | z value | Pr(> \|z\|) |
|---|---|---|---|---|
| (Intercept) | -5.8846 | 0.0241 | -243.91 | 0.0000 |
| poly(log(D),2)1 | 28.3184 | 1.5145 | 18.70 | 0.0000 |
| poly(log(D),2)2 | -5.0257 | 1.3396 | -3.75 | 0.0002 |
| Null deviance: 4930.8 on 3068 degrees of freedom | | | | |
| Residual deviance: 4471.3 on 3066 degrees of freedom | | | | |

| $M^H_{logO2}$ | Estimate | Std. Error | z value | Pr(> \|z\|) |
|---|---|---|---|---|
| (Intercept) | -5.8446 | 0.0231 | -253.26 | 0.0000 |
| poly(log(O),2)1 | -17.2385 | 1.3974 | -12.34 | 0.0000 |
| poly(log(O),2)2 | -17.8958 | 1.2774 | -14.01 | 0.0000 |
| Null deviance: 4930.8 on 3068 degrees of freedom | | | | |
| Residual deviance: 4589.4 on 3066 degrees of freedom | | | | |

| $M^H_{Z3}$ | Estimate | Std. Error | z value | Pr(> \|z\|) |
|---|---|---|---|---|
| (Intercept) | -5.9188 | 0.0249 | -238.04 | 0.0000 |
| poly(Z,3)1 | 29.9815 | 1.6628 | 18.03 | 0.0000 |
| poly(Z,3)2 | 0.2552 | 1.3199 | 0.19 | 0.8467 |
| poly(Z,3)3 | 5.4733 | 1.2179 | 4.49 | 0.0000 |
| Null deviance: 4930.8 on 3068 degrees of freedom | | | | |
| Residual deviance: 4367.5 on 3065 degrees of freedom | | | | |

| $M^H_R$ | Estimate | Std. Error | z value | Pr(> \|z\|) |
|---|---|---|---|---|
| (Intercept) | -5.7998 | 0.0218 | -265.55 | 0.0000 |
| R | -14.0753 | 1.0926 | -12.88 | 0.0000 |
| Null deviance: 4930.8 on 3068 degrees of freedom | | | | |
| Residual deviance: 4771.3 on 3067 degrees of freedom | | | | |

Table 9. **Hail Data**: Summaries of univariate binomial models for the grouped data for each variable after transformation.

### 5.2.2 Full Models

We see that for each univariate regression in Section 5.2.1 the highest order term is significant. Therefore all lower order terms must be kept in each model with higher order terms. A binomial regression was performed on the grouped data using all covariates investigated in the univariate models of Section 5.2.1. The following binomial model was fit:

$$\begin{aligned} M^H_{full}\ \text{logit}(p_k) = \beta_0 &+ \beta_1 poly(\log(C_k), 2)1 + \beta_2 poly(\log(C_k), 2)2 + \beta_3 poly(\log(D_k), 2)1 \\ &+ \beta_4 poly(\log(D_k), 2)2 + \beta_5 poly(\log(O_k), 2)1 + \beta_6 poly(\log(O_k), 2)2 \\ &+ \beta_7 poly(Z_k, 3)1 + \beta_8 poly(Z_k, 3)2 + \beta_9 poly(Z_k, 3)3 + \beta_{10} R_k \\ &+ \beta_{11} \log(C_k) : \log(D_k) + \beta_{12} \log(D_k) : \log(O_k) + \beta_{13} \log(C_k) : Z_k, \end{aligned} \tag{5.2.6}$$

for $k = 1, ..., 3125$, where $\beta_l$ for $l = 0, ..., 13$ are the regression coefficients. This was compared to the following GAM:

$$\begin{aligned} M^H_{fullGAM}\ \text{logit}(p_k) = &\ s(\log(C_k)) + s(\log(D_k)) + s(\log(O_k)) + s(Z_k) + s(R_k) \\ &+ te(\log(C_k), \log(D_k)) + te(\log(D_k), \log(O_k)) + te(\log(C_k), Z_k), \end{aligned} \tag{5.2.7}$$

for $k = 1, ..., 3125$, where $s$ and $te$ denote the smooths used to fit a GAM using the package MGCV [36].
Smaller models were chosen using the forward selection technique based on AIC, defined in Equation (5.2.8), using the software R and package MASS [30]. The AIC is defined as in [4] as follows:

$$AIC := -2loglik \pm 2p, \tag{5.2.8}$$

where $loglik$ is the log likelihood and $p$ is the number of parameters of the model. These models with one to four covariates are shown in the Appendix 7.
The summary outputs for Models (5.2.6) and (5.2.7) are shown in Tables 10 and 11. We see that for Model (5.2.6) the residual deviance has a value close to the residual degrees of freedom, giving no indication of a lack of fit. The log likelihood is slightly smaller than that of Model (5.2.7). In both models the ratio of residual deviance to null deviance is larger than 0, suggesting that both models are explaining some of the variability in the data. The number of trials, number of successes and number of failures for each group in the above binomial regressions are shown in Figure 19. The left most plot shows that most groups contain only $H_{ijt} = 0$, as expected. The right most plot shows that most groups had between 0 and approximately 500 observations. There are a few groups with a very large number of observations, however, we are satisfied that the grouping used for the hail model has found a suitable trade off to select number of groups.

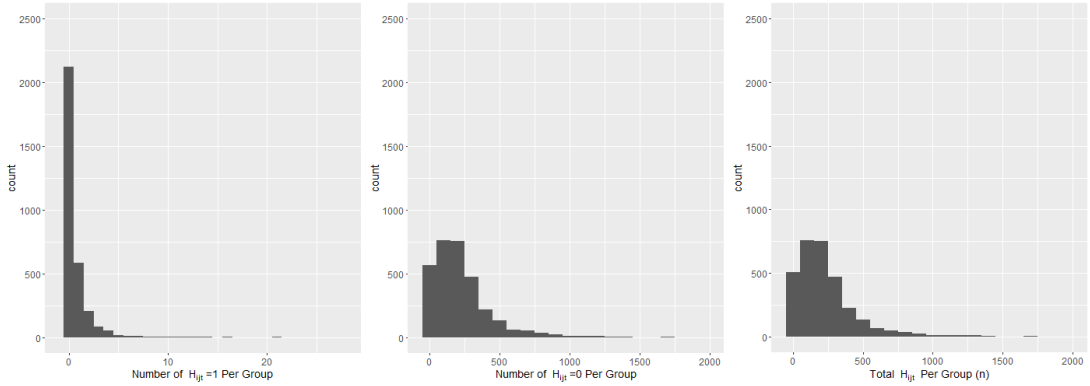


Figure 19. **Hail Data**: The number of trials, number of successes and number of failures for each group within the grouped data used for the binomial models. The modes for the plots from left to right are (x,z) = (0,2067), (100,763), (100,758)

| $M^H_{full}$ | Estimate | Std. Error | z value | Pr($>\|z\|$) |
|---|---|---|---|---|
| (Intercept) | -8.0181 | 0.7079 | -11.33 | 0.0000 |
| poly(log(C), 2)1 | 3.5574 | 6.9194 | 0.51 | 0.6072 |
| poly(log(C), 2)2 | -6.7900 | 1.5302 | -4.44 | 0.0000 |
| poly(log(D), 2)1 | 33.1609 | 10.2391 | 3.24 | 0.0012 |
| poly(log(D), 2)2 | -5.1951 | 1.3801 | -3.76 | 0.0002 |
| poly(log(O), 2)1 | 6.7364 | 5.5114 | 1.22 | 0.2216 |
| poly(log(O), 2)2 | -19.2809 | 1.3842 | -13.93 | 0.0000 |
| poly(Z, 3)1 | -43.8793 | 11.1352 | -3.94 | 0.0001 |
| poly(Z, 3)2 | -7.3231 | 1.5549 | -4.71 | 0.0000 |
| poly(Z, 3)3 | 9.7184 | 1.2323 | 7.89 | 0.0000 |
| R | -0.0086 | 0.0011 | -7.88 | 0.0000 |
| log(C):log(D) | 0.0495 | 0.0273 | 1.81 | 0.0699 |
| log(D):log(O) | -0.0928 | 0.0251 | -3.70 | 0.0002 |
| log(C):Z | 0.0001 | 0.0000 | 4.38 | 0.0000 |
| Null deviance: 4930.8 on 3068 degrees of freedom | | | | |
| Residual deviance: 2841.1 on 3055 degrees of freedom | | | | |
| Log-likelihood: -2634.782 (df = 14) | | | | |

Table 10. **Hail Data**: Summary of the full binomial model for data grouped into 5 levels for each transformed covariate.

| $M^H_{fullGAM}$ | Estimate | Std. Error | z value | Pr(> \|z\|) |
|---|---|---|---|---|
| (Intercept) | -6.27744 | 0.03323 | -188.9 | <2e-16 |
| | edf | Ref.df | Chi.sq | p-value |
| s(log(C)) | 2.689 | 2.919 | 206.37 | <2e-16 |
| s(log(D)) | 2.310 | 2.438 | 336.27 | <2e-16 |
| s(log(O)) | 2.380 | 2.438 | 86.70 | <2e-16 |
| s(Z) | 3.000 | 3.000 | 29.35 | 1.92e-06 |
| s(R) | 2.026 | 2.351 | 66.51 | 3.74e-14 |
| te(log(C),log(D)) | 2.248 | 11.000 | 23.46 | 1.27e-06 |
| te(log(D),log(O)) | 3.210 | 11.000 | 26.08 | 9.43e-09 |
| te(log(C),Z) | 2.858 | 11.000 | 28.81 | 1.26e-06 |
| | | $\frac{residual\ deviance}{null\ deviance} = 0.44$ | | |
| Log-likelihood: -2599.427 (df = 21.72225) | | | | |

Table 11. **Hail Data**: Summary of the full GAM model for data grouped into 5 levels for each transformed covariate.

### 5.2.3 Model Selection and Assessing Goodness of Fit

We now have two proposed models, one GLM and one GAM. We need to decide which model is best based on a trade off between goodness of fit, simplicity and the time it takes to fit the model. In Figure 20 we see plots of the Pearson residuals, fitted values against empirical estimates and the fitted values of the GAM and binomial models plotted against each other. The residuals show no large outliers or strong pattern or relationship. Most fitted values cluster around the diagonal line when plotted against the empirical estimates, and similarly the GAM and logistic fitted values cluster around the diagonal line when plotted against each other. In addition, the log likelihoods of the two models have a relative difference of approximately 1 percent. The use of a logistic model as opposed to a GAM is much more favourable computationally thus we select the logistic model, $M^H_{full}$, defined in Model (5.2.6).
Finally, we plot look up tables of our results and compare them to the results of [28]. In Figures 21, 22 and 23 we consider the covariates C and D whilst holding O, Z and R constant at their $5^{th}$, $50^{th}$ and $95^{th}$ percentile values to produce look up tables for Models 5.2.6 and 5.2.7. We can compare this qualitatively to plots (c) and (d) of Figure 24, where we see that the plots have the same general shape.

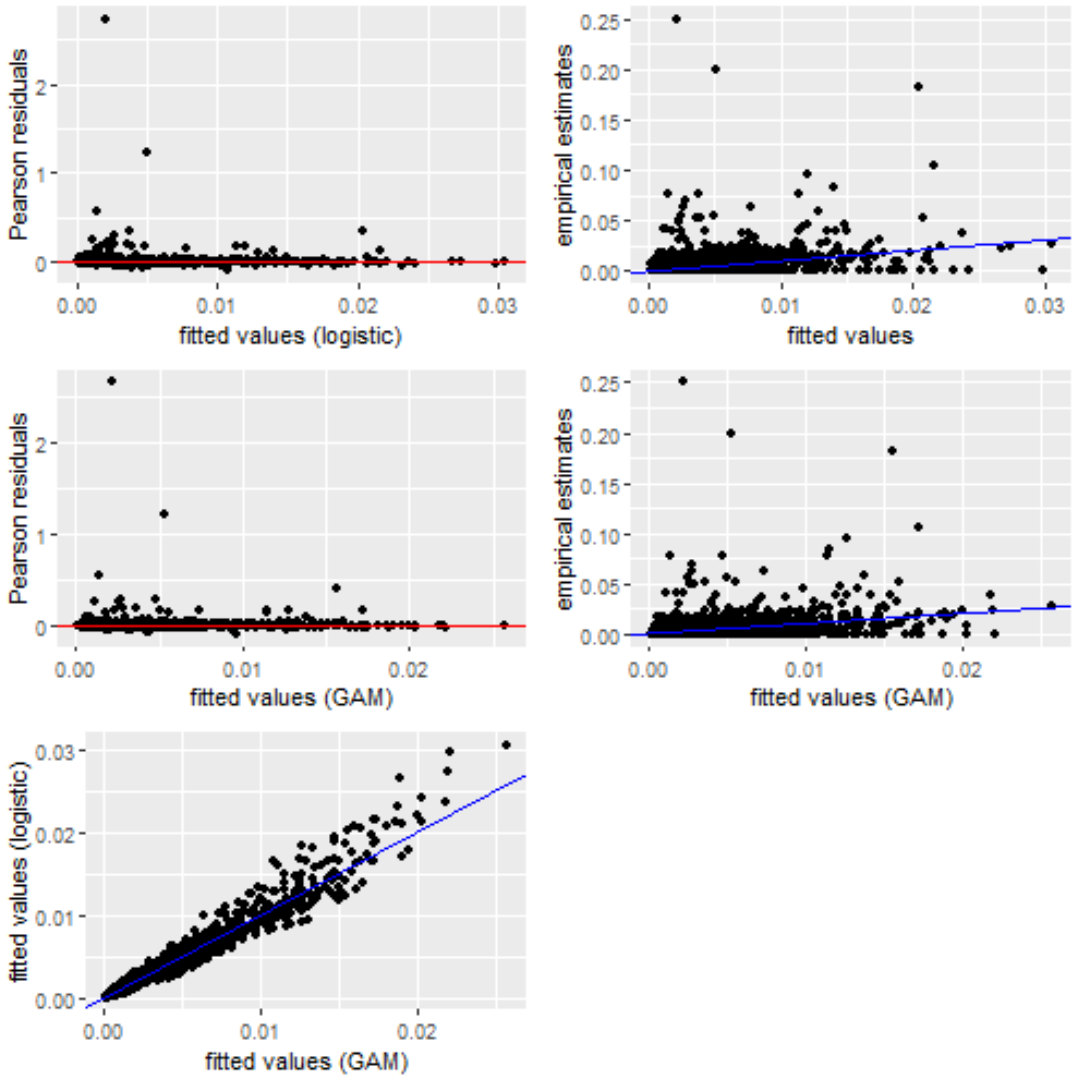


Figure 20. **Hail Data**: Plots of the fitted values vs empirical estimates and fitted values vs the Pearson residuals for Models 5.2.6 and 5.2.7, $M^H_{full}$ and , $M^H_{fullGAM}$. The Pearson residuals were calculated as in [4] and the fitted values calculated using the R function prediction. Additionally the fitted values for both models have been plotted against each other. The red lines show the zero line and the blue lines show the diagonal line.

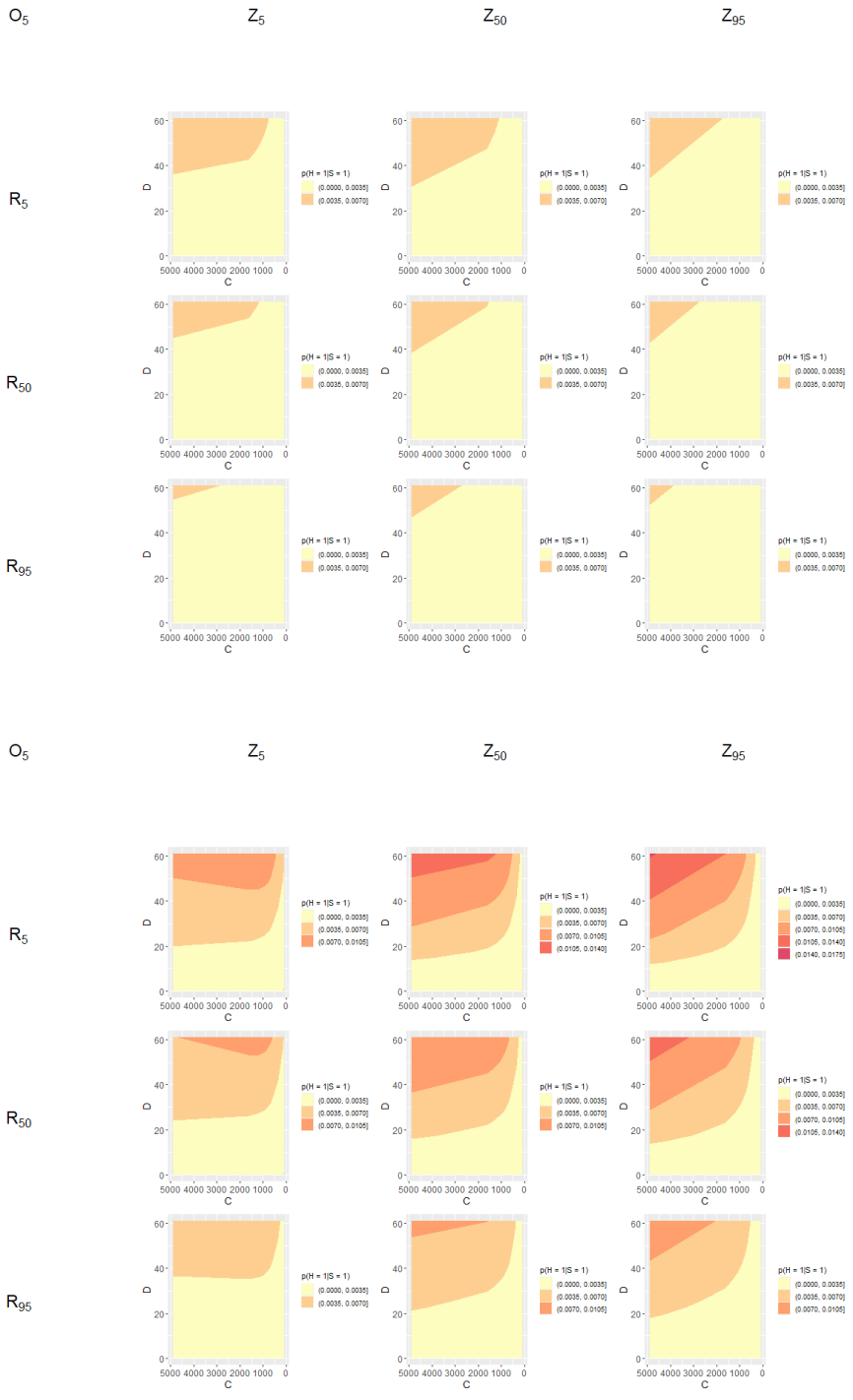


Figure 21. **Thunderstorm Data**: lookup tables for the probability of $H = 1|S = 1$ for values of C and D, whilst holding O constant at the $5^{th}$ percentile and R and Z constant at the $5^{th}$, $50^{th}$ and $95^{th}$ for the logistic model above and the GAM below.

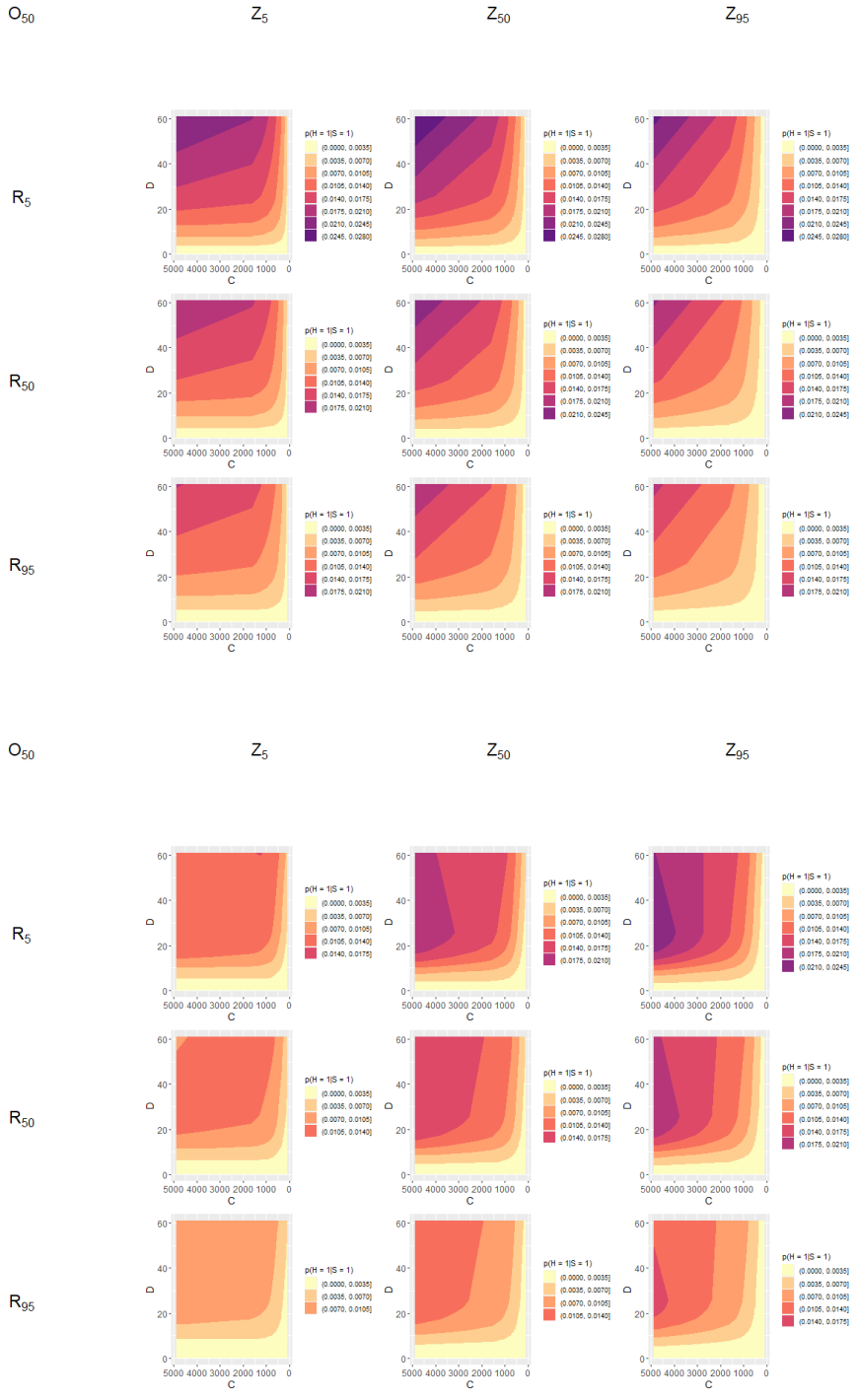


Figure 22. **Thunderstorm Data**: lookup tables for the probability of $H = 1|S = 1$ for values of C and D, whilst holding O constant at the $50^{th}$ percentile and R and Z constant at the $5^{th}$, $50^{th}$ and $95^{th}$ for the logistic model above and the GAM below.

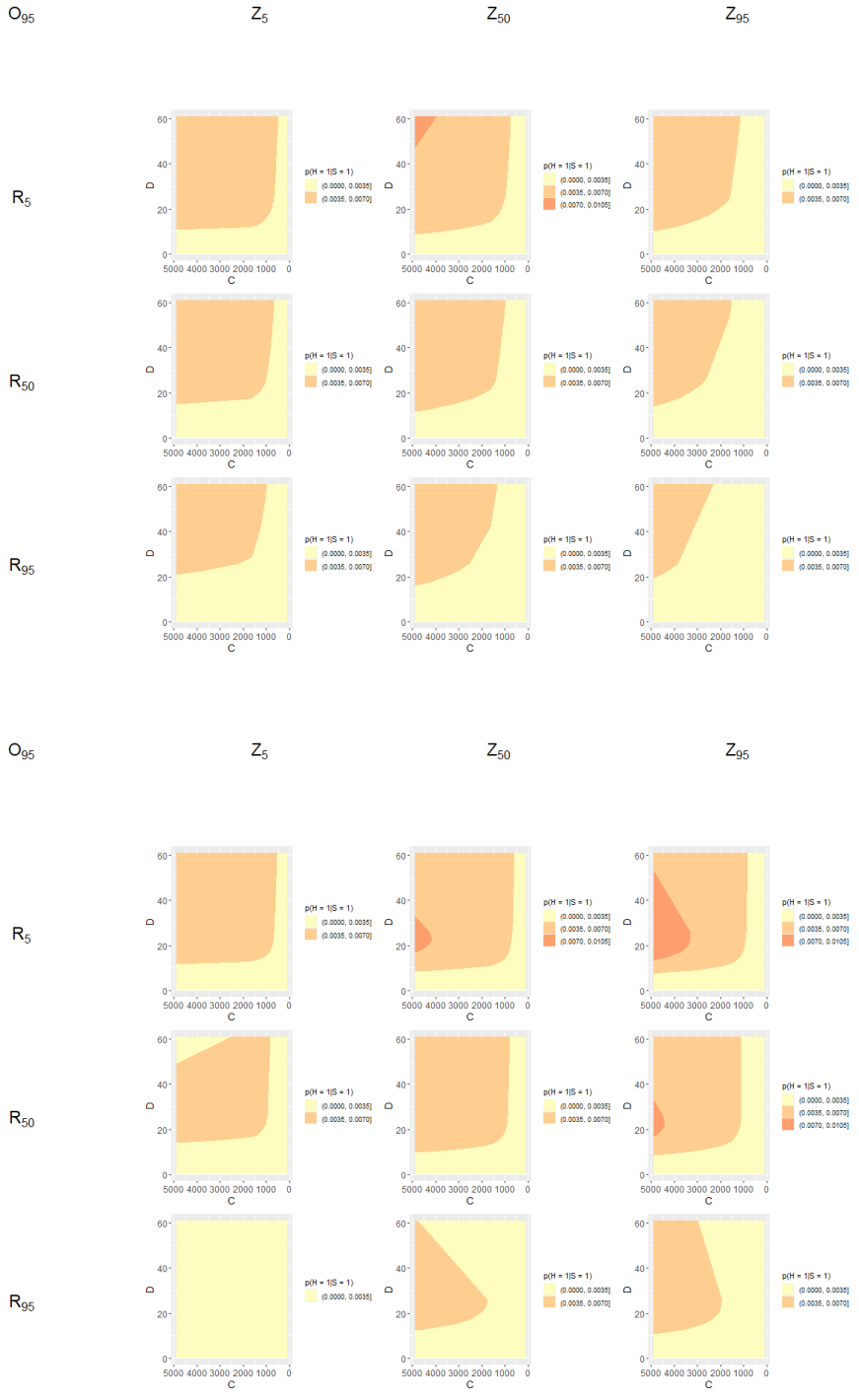


Figure 23. **Thunderstorm Data**: lookup tables for the probability of $H = 1|S = 1$ for values of C and D, whilst holding O constant at the $95^{th}$ percentile and R and Z constant at the $5^{th}$, $50^{th}$ and $95^{th}$ for the logistic model above and the GAM below.

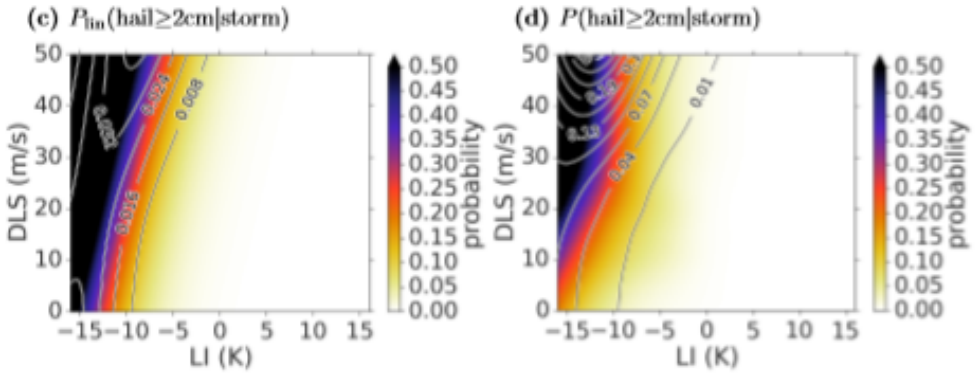


Figure 24. **Hail Data**: This figure shows the lookup table for LI and D whilst holding O, Z and R constant and LI and R whilst holding D, O and Z constant for the generalised additive and logistic hail models from [28]. LI is qualitatively similar to C, and was used because the available data was too coarse to calculate C. Larger values of C correspond to more negative values of LI.

## 5.3 EDA Plots For Thunderstorm Model

We now consider the analysis for the thunderstorm model, following the same structure as for the hail model. This section contains the pairs plots, box plots, histograms, empirical logit plots, transformation analysis and interaction effect plots for the thunderstorm model. These were calculated after the data preprocessing was complete, so that for all data points $C_{ijt} \geq 5$. In the pairs plots in Figures 25 and 26 we again see no strong relationship between any of the covariates, however, as for the hail model, the large number of data points make these plots inconclusive.

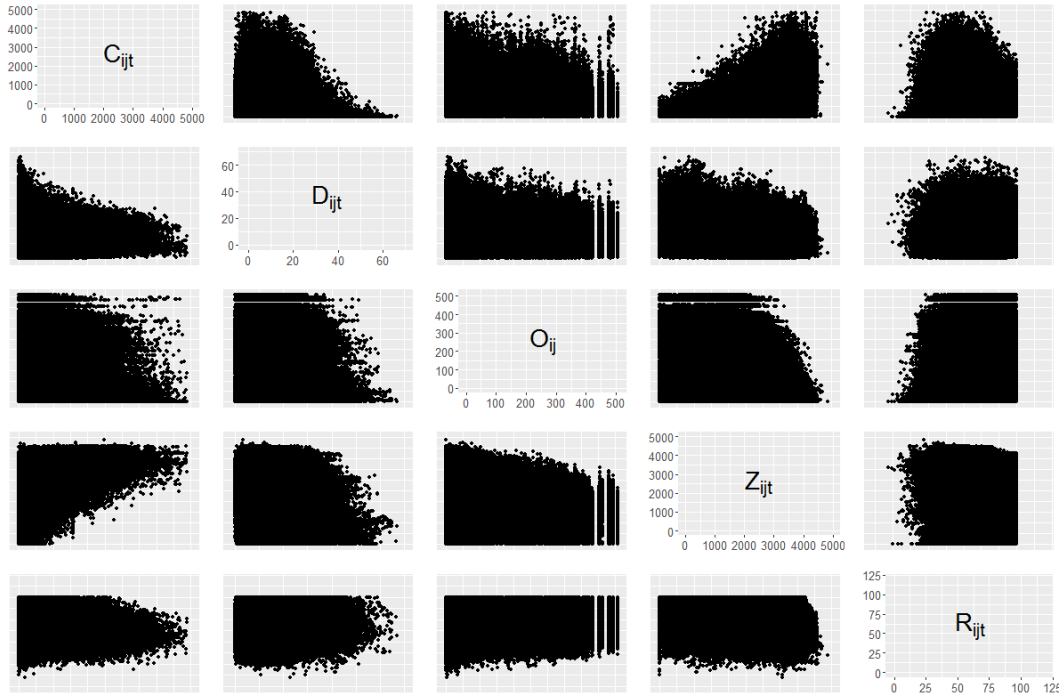


Figure 25. **Thunderstorm Data**: Pairs plots of the covariates ($C_{ijt}$, $D_{ijt}$, $O_{ij}$, $Z_{ijt}$ and $R_{ijt}$) for data points only with $S_{ijt} = 1$.

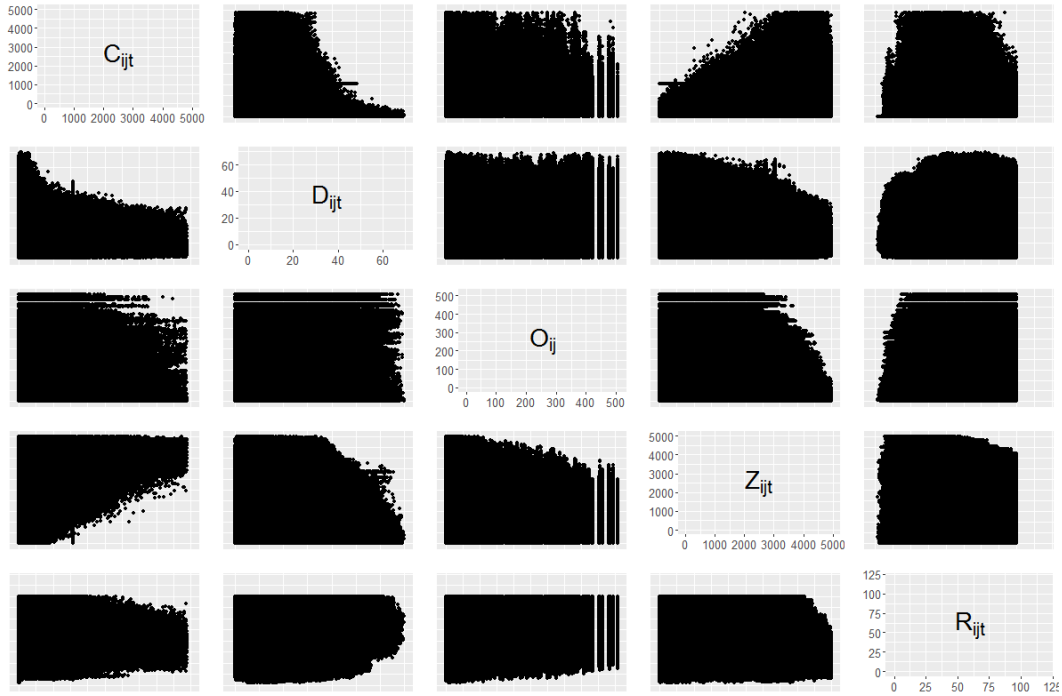


Figure 26. **Thunderstorm Data**: Pairs plots of the covariates ($C_{ijt}$, $D_{ijt}$, $O_{ij}$, $Z_{ijt}$ and $R_{ijt}$) for data points with only $S_{ijt} = 0$.

### 5.3.1 Analysis of Individual Considered Covariates

We now want to analyse the covariates individually. Again, the covariates have been grouped with respect to their quantiles, and the mean defined as the score for each interval. Looking at the box plots shown in Figure 27 we see that $C_{ijt}$ has a moderate effect, with the plots of remaining covariates being inconclusive. Next we look at the histograms in Figure 28, where we again see that our data is not uniformly distributed over all covariate values, as expected, thus the covariate groups were calculated with respect to quantiles. For the same reasons as in the exploratory data analysis for the hail model we decide to work with 50 groups.
We now examine the empirical logits in Figure 29, which, as in the hail model EDA, show the original empirical logits, the empirical logits after the transformation of the covariates and the studentised residuals of a polynomial fit of the transformed variable as covariates and the empirical logits as response. Again, all covariates appear to have non-linear relationships with the response variable empirical logit. Log transforms were chosen for the covariates $C_{ijt}$ and $O_{ij}$ based on the shape of the empirical logit plots and the covariates $D_{ijt}$, $Z_{ijt}$ and $R_{ijt}$ were not transformed.

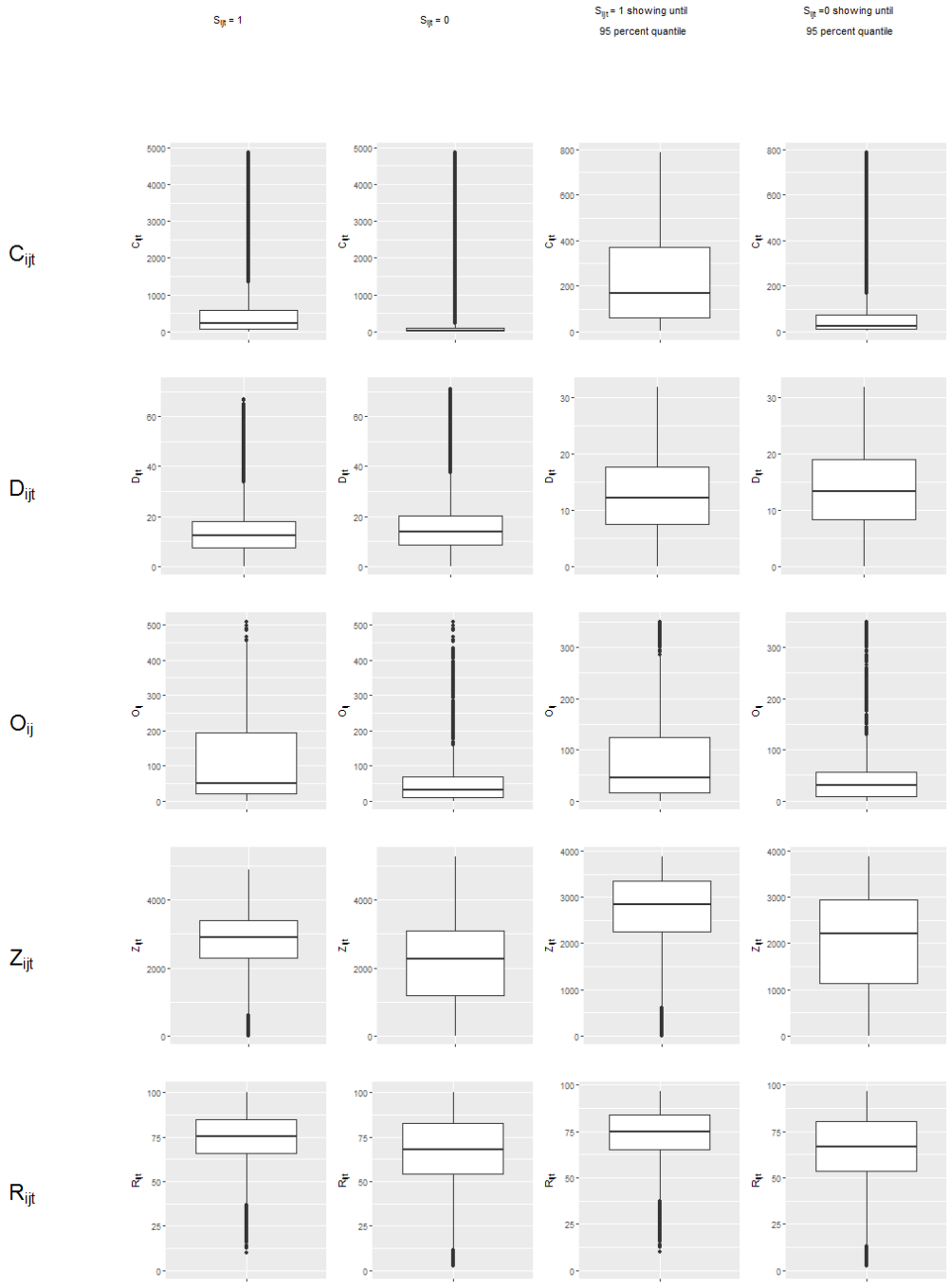


Figure 27. **Thunderstorm Data**: Boxplots of the covariates ($C_{ijt}$, $D_{ijt}$, $O_{ij}$, $Z_{ijt}$ and $R_{ijt}$). The boxplots were calculated separately for data points with $S_{ijt} = 1$ and $S_{ijt} = 0$. The two left most columns show the extremes and the two right most columns show the whiskers of the same box plots.

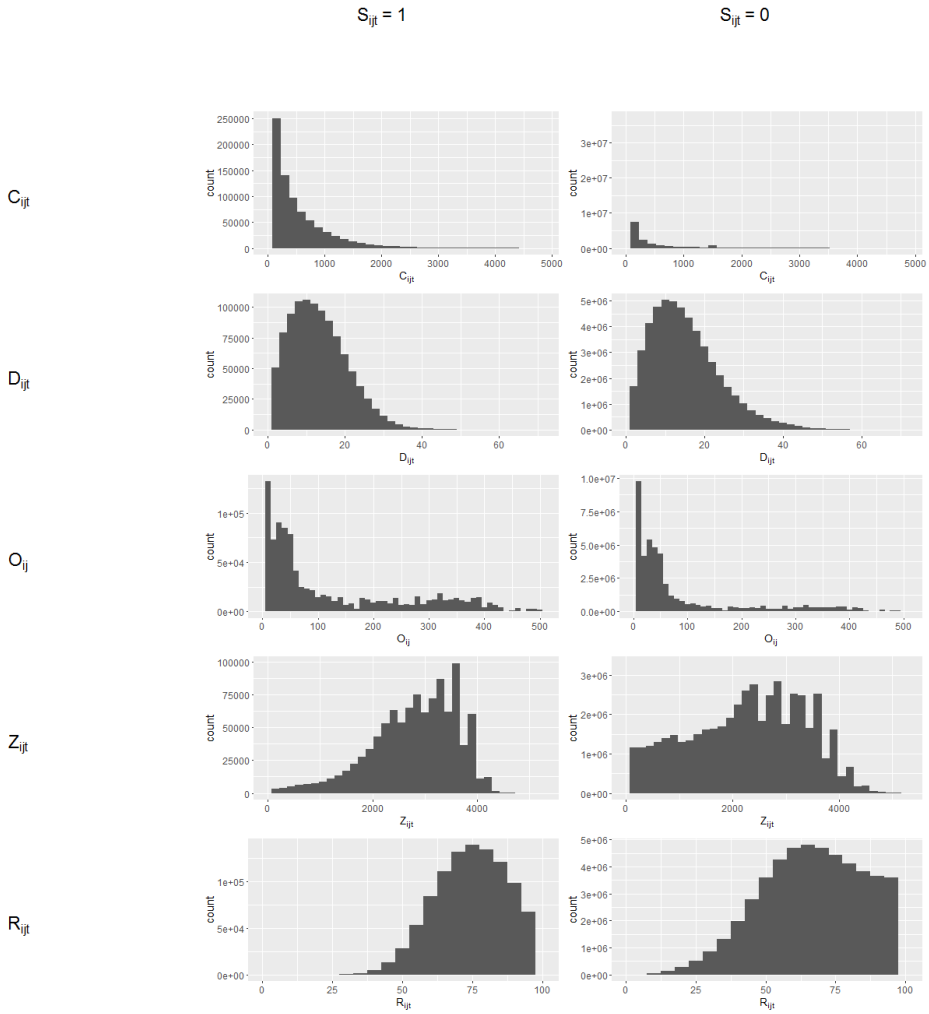


Figure 28. **Thunderstorm Data**: Histograms and empirical logit plots of the covariates ($C_{ijt}$, $D_{ijt}$, $O_{ij}$, $Z_{ijt}$ and $R_{ijt}$). Histograms were calculated separately for data points with $S_{ijt} = 1$ and $S_{ijt} = 0$.

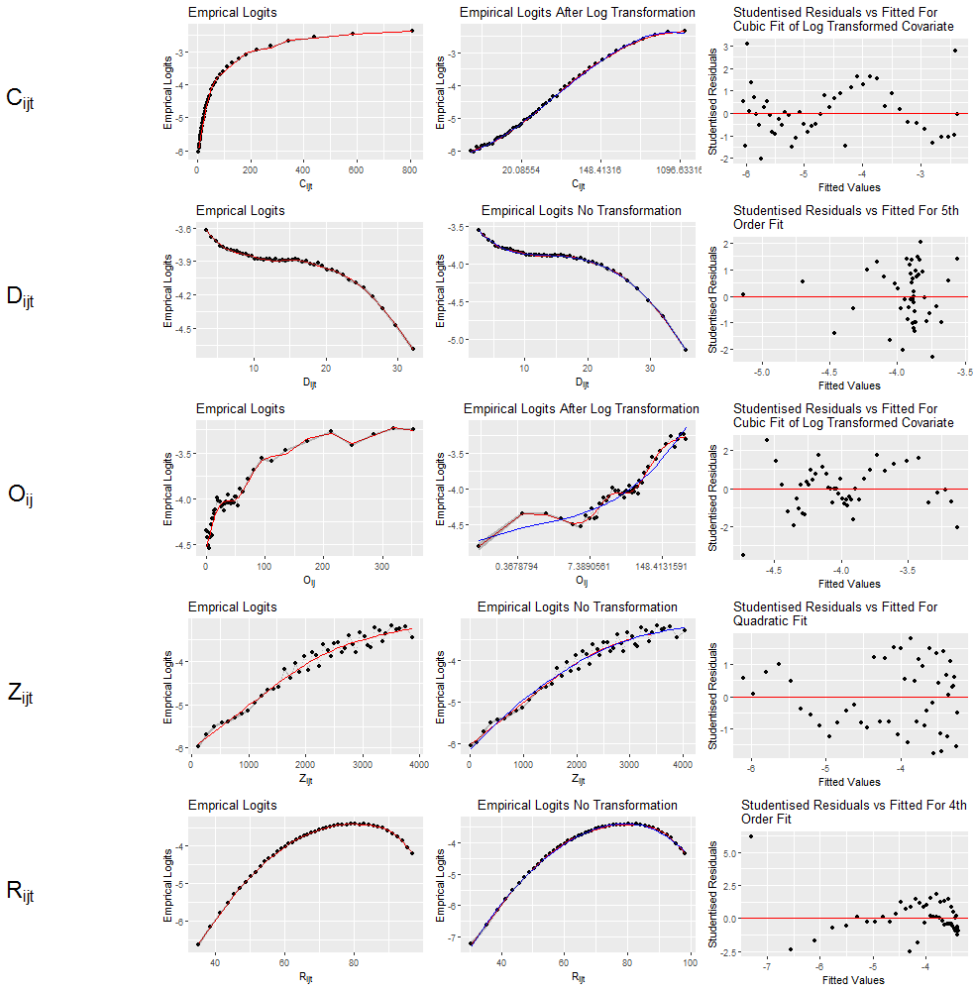


Figure 29. **Thunderstorm Data**: These plots show the empirical logits calculated for all covariates ($C_{ijt}$, $D_{ijt}$, $O_{ij}$, $Z_{ijt}$ and $R_{ijt}$), before and after the transformations have been applied. The left column shows the empirical logit points calculated with the covariates prior to transformation, which have been grouped into 50 groups with respect to their quantiles, and are shown with their 95 percent confidence band. The red line shows a GAM fit. $C_{ijt}$ and $O_{ij}$ have been transformed with the log function. The middle column shows the empirical logit points calculated with the transformed covariates which have again been grouped into 50 groups with respect to their quantiles, and are shown with their 95 percent confidence band. The red line shows a GAM fit and the blue line shows a polynomial fit, the equations for which can be seen in Models (5.4.1),(5.4.2), (5.4.3), (5.4.4) and (5.4.5). Additionally, the studentised residuals of polynomial fits of the empirical logits, $l$, and the transformed covariates have been plotted against their fitted values, shown in the right column. For an $i^{th}$ order polynomial of covariate $x$ with coefficient vector $\boldsymbol{\beta} = (\beta_0, ..., \beta_i)^T$ the residuals of the following fit would be plotted: $l = \boldsymbol{\beta} \times poly(x, i)$.

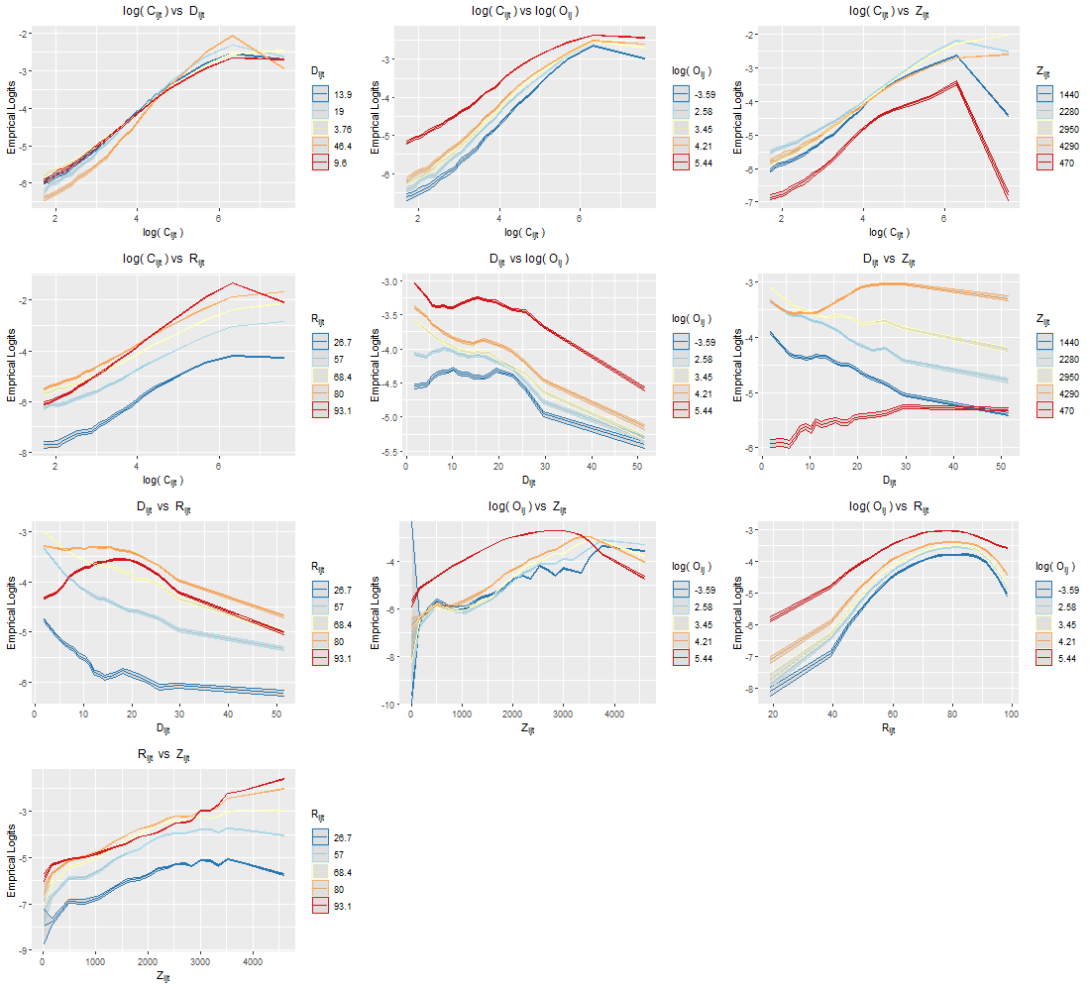


Figure 30. **Thunderstorm Data**: Interaction effects between all covariates ($C_{ijt}$, $D_{ijt}$, $O_{ij}$, $Z_{ijt}$ and $R_{ijt}$). The empirical logit points have been calculated with the covariates grouped into 100 (20 x 5 = 100) groups with respect to their quantiles, and are shown with their 95 percent confidence band.

### 5.3.2 Analysis of Interaction Effects

After completing the analysis of each individual covariate we now investigate for interaction effects using the transformed covariates from the previous section. As in the case of the hail model, for each plot we must consider a grouping with respect to two covariates. As this data set is larger than that for the hail model, we increase the number of groups, but only for the covariate shown on the x-axis of each plot. Again, this number of groups provides a compromise between the smoothness of the empirical logit plots, and any relationships being visible.

Consider Figure 30, we see that the plots showing the interactions of $C_{ijt}$ and $D_{ijt}$, $D_{ijt}$ and $Z_{ijt}$ and $D_{ijt}$ and $R_{ijt}$ show some cross over between groups. Thus we investigate further by looking at surface plots in Figures 31, 32 and 33 for the covariates $C_{ijt}$ and $D_{ijt}$, $D_{ijt}$ and $Z_{ijt}$ and $D_{ijt}$ and $R_{ijt}$. As in the hail model EDA we see that the shapes of the GAM tensor and empirical logit surfaces are in agreement, but not planes, thus suggesting possible interaction effects between these covariates.

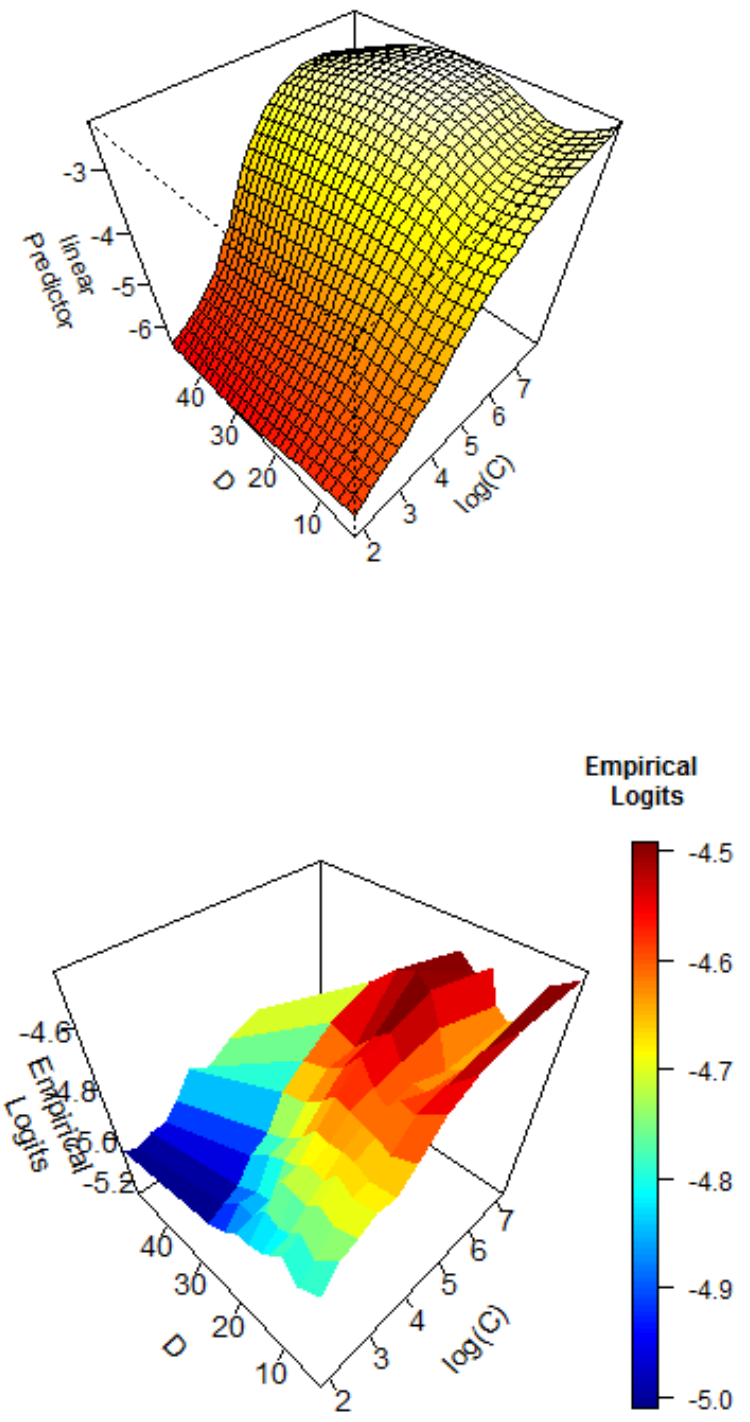


Figure 31. **Thunderstorm Data**: Interaction effects between the covariates $\log(C_{ijt})$ and $D_{ijt}$ and the response $S_{ijt}$ using a GAM tensor product surface (above) and the empirical logits (below)

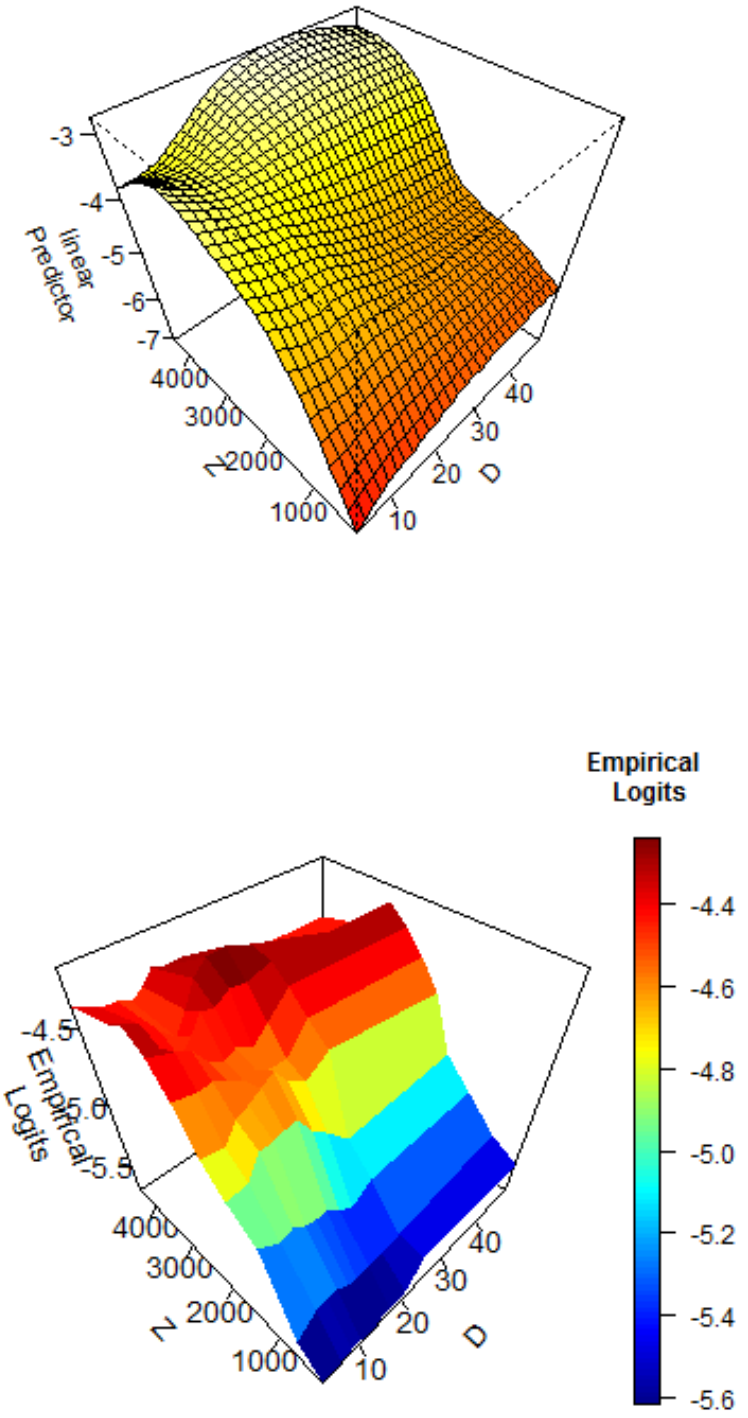


Figure 32. **Thunderstorm Data**: Interaction effects between the covariates $D_{ijt}$ and $Z_{ijt}$ and the response $S_{ijt}$ using a GAM tensor product surface (above) and the empirical logits (below).

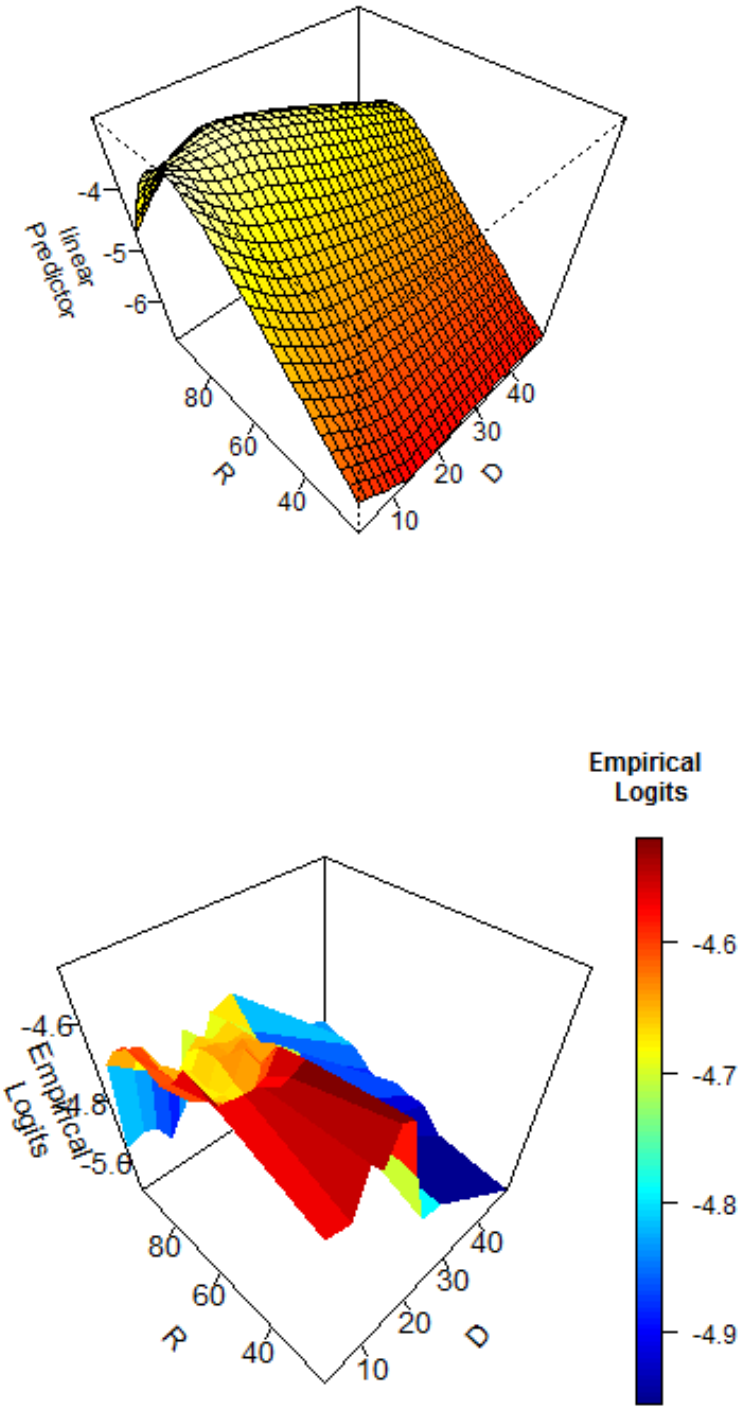


Figure 33. **Thunderstorm Data**: Interaction effects between the covariates $D_{ijt}$ and $R_{ijt}$ and the response $S_{ijt}$ using a GAM tensor product surface (above) and the empirical logits (below).

## 5.4 Fitting of Models on Grouped Data For The Thunderstorm Model Case

We now investigate several proposed models for grouped data in the thunderstorm model case. Each of the 5 covariates were grouped into 10 levels reducing the thunderstorm data from 55,002,099 to 100,000 ($10^5 = 100{,}000$) data points. Thus, $C_{ijt}$, $D_{ijt}$, $O_{ij}$, $Z_{ijt}$ and $R_{ijt}$ now correspond to $C_m$, $D_m$, $O_m$, $Z_m$ and $R_m$ respectively, for $m = 1, ..., 100000$.

### 5.4.1 Univariate Binomial Models

For each covariate we consider a univariate binary regression: a logistic model with a cubic effect for the log transform of $C_m$,

$$M^S_{logC3} : \text{logit}(p_m) = \beta_0 + \beta_1 poly(\log(C_m), 3)1 + \beta_2 poly(\log(C_m), 3)2 + \beta_3 poly(\log(C_m), 3)3, \tag{5.4.1}$$

a logistic model with a cubic effect for the log transform of $D_m$,

$$\begin{aligned} M^S_{D5} : \text{logit}(p_m) = \beta_0 &+ \beta_1 poly(D_m, 5)1 + \beta_2 poly(D_m, 5)2 + \beta_3 poly(D_m, 5)3 \\ &+ \beta_4 poly(D_m, 5)4 + \beta_5 poly(D_m, 5)5, \end{aligned} \tag{5.4.2}$$

a logistic model with a quadratic effect for the log transform of $O_m$,

$$M^S_{logO3} : \text{logit}(p_m) = \beta_0 + \beta_1 poly(\log(O_m), 3)1 + \beta_2 poly(\log(O_m), 3)2 + \beta_3 poly(\log(O_m), 3)3, \tag{5.4.3}$$

a logistic model with a quadratic effect for $Z_m$,

$$M^S_{Z2} : \text{logit}(p_m) = \beta_0 + \beta_1 poly(Z_m, 2)1 + \beta_2 poly(Z_m, 2)2, \tag{5.4.4}$$

and a logistic model with a quadratic effect for $R_m$,

$$M^S_{R4} \text{logit}(p_m) = \beta_0 + \beta_1 poly(R_m, 4)1 + \beta_2 poly(R_m, 4)2 + \beta_3 poly(R_m, 4)3 + \beta_4 poly(R_m, 4)4, \tag{5.4.5}$$

for $m = 1, ..., 100000$, where the $\beta_l$ are the regression coefficients. The summaries for Models (5.4.1), (5.4.2), (5.4.3), (5.4.4) and (5.4.5) are shown in Table 19. The residual deviance is much larger than the residual degrees of freedom for each of these models, thus showing a lack of fit and suggesting we should consider a model with multiple covariates. The highest order term is significant in all of the univariate models meaning all lower order terms for each of the covariates must be included. Summaries for the ungrouped data are shown in the Appendix 7 in Table 19.

| $M^S_{logC3}$ | Estimate | Std. Error | z value | Pr(> \|z\|) |
|---|---|---|---|---|
| (Intercept) | -5.1785 | 0.0025 | -2078.85 | 0.0000 |
| log(C)1 | 383.3814 | 0.6845 | 560.12 | 0.0000 |
| log(C)2 | -96.2823 | 0.6917 | -139.19 | 0.0000 |
| log(C)3 | -40.0898 | 0.5394 | -74.32 | 0.0000 |
| Null deviance: 1600768 on 99318 degrees of freedom | | | | |
| Residual deviance: 753471 on 99315 degrees of freedom | | | | |
| $M^S_{D3}$ | Estimate | Std. Error | z value | Pr(> \|z\|) |
| (Intercept) | -4.4369 | 0.0013 | -3368.87 | 0.0000 |
| D1 | -92.5459 | 0.5465 | -169.35 | 0.0000 |
| D2 | -16.2035 | 0.4414 | -36.71 | 0.0000 |
| D3 | -9.4235 | 0.3967 | -23.75 | 0.0000 |
| D4 | 11.1923 | 0.3892 | 28.75 | 0.0000 |
| D5 | -1.7759 | 0.3855 | -4.61 | 0.0000 |
| Null deviance: 1600768 on 99318 degrees of freedom | | | | |
| Residual deviance: 1562159 on 99313 degrees of freedom | | | | |
| $M^S_{logO3}$ | Estimate | Std. Error | z value | Pr(> \|z\|) |
| (Intercept) | -4.4848 | 0.0014 | -3292.52 | 0.0000 |
| log(O)1 | 121.4215 | 0.4482 | 270.90 | 0.0000 |
| log(O)2 | 32.5787 | 0.4381 | 74.36 | 0.0000 |
| log(O)3 | -10.1596 | 0.4189 | -24.25 | 0.0000 |
| Null deviance: 1600768 on 99318 degrees of freedom | | | | |
| Residual deviance: 1491908 on 99315 degrees of freedom | | | | |
| $M^S_{Z2}$ | Estimate | Std. Error | z value | Pr(> \|z\|) |
| (Intercept) | -4.7939 | 0.0020 | -2400.92 | 0.0000 |
| Z1 | 314.1897 | 0.7463 | 420.97 | 0.0000 |
| Z2 | -131.7958 | 0.5480 | -240.48 | 0.0000 |
| Null deviance: 1600768 on 99318 degrees of freedom | | | | |
| Residual deviance: 1281221 on 99316 degrees of freedom | | | | |
| $M^S_{R4}$ | Estimate | Std. Error | z value | Pr(> \|z\|) |
| (Intercept) | -4.7088 | 0.0021 | -2219.48 | 0.0000 |
| R1 | 269.2659 | 1.1626 | 231.61 | 0.0000 |
| R2 | -152.8784 | 1.0657 | -143.46 | 0.0000 |
| R3 | -63.6848 | 0.6290 | -101.24 | 0.0000 |
| R4 | 5.0942 | 0.4598 | 11.08 | 0.0000 |
| Null deviance: 1600768 on 99318 degrees of freedom | | | | |
| Residual deviance: 1366052 on 99314 degrees of freedom | | | | |

Table 12. **Thunderstorm Data**: Summaries of univariate binomial models for each covariate after transformation.

### 5.4.2 Full Models

All highest order terms were significant in the univariate regressions in Section 5.4.1, therefore no lower order terms could be removed from the model. A binomial regression was performed on the grouped data using all covariates investigated in the univariate models of Section 5.4.1 as in the hail model EDA. The following binomial model was fit

$$\begin{aligned} M^S_{full}\text{logit}(p_m) = \beta_0 &+ \beta_1 poly(\log(C_m), 3) + \beta_2 poly(\log(C_m), 3)2 + \beta_3 poly(\log(C_m), 3)3 \\ &+ \beta_4 poly(D_m, 5)1 + \beta_5 poly(D_m, 5)2 + \beta_6 poly(D_m, 5)4 \\ &+ \beta_7 poly(D_m, 5)5 + \beta_8 poly(\log(O_m), 3)1 + \beta_9 poly(\log(O_m), 3)2 \\ &+ \beta_{10} poly(\log(O_m), 3)3 + \beta_{11} poly(Z_m, 2)1 + \beta_{12} poly(Z_m, 2)2 \\ &+ \beta_{13} poly(R_m, 4)1 + \beta_{14} poly(R_m, 4)2 + \beta_{15} poly(R_m, 4)3 \\ &+ \beta_{16} poly(R_m, 4)4 + \beta_{17} \log(C_m) : D_m + \beta_{18} D_m : Z_m, \end{aligned} \tag{5.4.6}$$

for $m = 1, ..., 100000$, where $\beta_l$ for $l = 1, ..., 18$ are the regression coefficients. This was also compared to the following GAM:

$$\begin{aligned} M^S_{fullGAM}\ \text{logit}(p_m) = &\ s(\log(C_m)) + s(D_m) + s(\log(O_m)) + s(Z_m) + s(R_m) \\ &+ te(\log(C_m), \log(D_m)) + te(D_m, Z_m), \end{aligned} \tag{5.4.7}$$

for $m = 1, ..., 100000$, where $s$ and $te$ denote the smooths used to fit a GAM using the package MGCV [36].

As in the hail model case, smaller models for one to four covariates were chosen using forward selection based on the AIC, defined in Equation (5.2.8) using the software R and package MASS [30], and are defined in the Appendix 7. The number of trials, number of successes and number of failures for each group in the above binomial regressions are shown in Figure 34, where we see most groups have between 0 and 2000 observations, with a very small number of groups having a larger number of observations, thus we are satisfied that our chosen grouping has found an acceptable trade off to select number of groups. The summary output for the logistic model, Model (5.4.6), $M^S_{full}$, is shown in Table 13. The residual deviance is much larger than the residual degrees of freedom indicating possible overdispersion. Thus a beta-binomial model was fit, the summary of which can be seen in Table 14. Here we see that for the overdispersion parameter, $\phi$, the ratio of its estimate, $5.766 \times 10^{-3}$, to standard error, $4.363 \times 10^{-5}$, is approximately 130. Furthermore, looking at the right most histogram in Figure 34, we see that the number of observations, $n_m$ for $m = 1, ..., 100000$, in most groups ranges between 0 and 2000. Recall Equation (3.2.41), and in particular the overdispersion factor $(1 + (n_m - 1)\phi)$. For most groups $(n_m - 1) \times \phi$ is not negligible, suggesting overdisperson is present in the data. The package aod was used to fit the beta-binomial model, see [19]. The summary output for Model (5.4.7), $M^S_{fullGAM}$ is shown in Table 15. We see that the ratio of the residual to null deviance is 0.85 suggesting the model explains some of the variability in the data, however, log likelihood is somewhat smaller than that of the beta-binomial model, with a relative difference of approximately 20 percent.

### 5.4.3 Model Selection and Goodness of Fit

We now have two proposed models for the the thunderstorm model: a GAM and a beta-binomial logistic model. As in the case of the hail model we want to choose the best

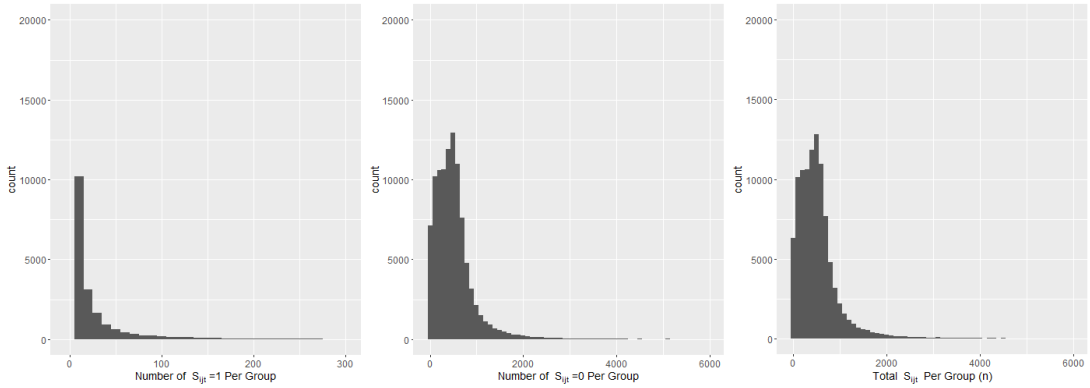


Figure 34. **Thunderstorm Data**: The number of trials, number of successes and number of failures for each group within the grouped data used for the binomial models. The modes for the plots from left to right are (x,z) = (10,14591), (500,13002), (500,12818).

| $M^S_{full}$ | Estimate | Std. Error | z value | Pr(> \|z\|) |
|---|---|---|---|---|
| (Intercept) | -6.8563 | 0.0177 | -387.22 | 0.0000 |
| poly(log(C),3)1 | 316.3599 | 0.9770 | 323.81 | 0.0000 |
| poly(log(C),3)2 | -52.6994 | 0.6895 | -76.43 | 0.0000 |
| poly(log(C),3)3 | -36.5564 | 0.5402 | -67.68 | 0.0000 |
| poly(D,5)1 | -278.5434 | 4.7919 | -58.13 | 0.0000 |
| poly(D,5)2 | -8.6570 | 0.4782 | -18.10 | 0.0000 |
| poly(D,5)3 | -10.5874 | 0.4088 | -25.90 | 0.0000 |
| poly(D,5)4 | 7.0541 | 0.3985 | 17.70 | 0.0000 |
| poly(D,5)5 | -3.2509 | 0.3940 | -8.25 | 0.0000 |
| poly(log(O),3)1 | 100.1817 | 0.4842 | 206.92 | 0.0000 |
| poly(log(O),3)2 | 40.8011 | 0.4614 | 88.42 | 0.0000 |
| poly(log(O),3)3 | 19.7247 | 0.4279 | 46.10 | 0.0000 |
| poly(Z,2)1 | 224.0155 | 1.2404 | 180.60 | 0.0000 |
| poly(Z,2)2 | -93.9203 | 0.5614 | -167.28 | 0.0000 |
| poly(R,4)1 | 305.7326 | 1.5186 | 201.32 | 0.0000 |
| poly(R,4)2 | -64.8773 | 1.0809 | -60.02 | 0.0000 |
| poly(R,4)3 | -51.6417 | 0.6371 | -81.06 | 0.0000 |
| poly(R,4)4 | 4.1589 | 0.4690 | 8.87 | 0.0000 |
| log(C):D | 0.0010 | 0.0001 | 12.73 | 0.0000 |
| D:Z | 0.0000 | 0.0000 | 42.92 | 0.0000 |
| D:R | 0.0006 | 0.0000 | 52.34 | 0.0000 |
| Null deviance: 1600768 on 99318 degrees of freedom | | | | |
| Residual deviance: 257077 on 99298 degrees of freedom | | | | |

Table 13. **Thunderstorm Data**: Summary of the full binomial model for data grouped into 10 levels for each transformed covariate. The residual deviance is much larger than the residual degrees of freedom indicating overdispersion.

| $M^S_{full,betabin}$ | Estimate | Std. Error | z value | Pr(> \|z\|) |
|---|---|---|---|---|
| (Intercept) | -5.210e+00 | 8.042e-03 | -6.478e+02 | 0.000e+00 |
| poly(log(C), 3)1 | 2.970e+02 | 1.265e+00 | 2.347e+02 | 0.000e+00 |
| poly(log(C), 3)2 | -5.224e+01 | 8.763e-01 | -5.962e+01 | 0.000e+00 |
| poly(log(C), 3)3 | -3.430e+01 | 7.736e-01 | -4.434e+01 | 0.000e+00 |
| poly(D, 3)1 | -3.929e+01 | 2.206e+00 | -1.781e+01 | 0.000e+00 |
| poly(D, 3)2 | 7.737e-01 | 7.090e-01 | 1.091e+00 | 2.752e-01 |
| poly(D, 3)3 | -1.150e+01 | 6.841e-01 | -1.682e+01 | 0.000e+00 |
| poly(log(O), 2)1 | 9.247e+01 | 7.349e-01 | 1.258e+02 | 0.000e+00 |
| poly(log(O), 2)2 | 4.280e+01 | 7.146e-01 | 5.990e+01 | 0.000e+00 |
| poly(Z, 2)1 | 2.153e+02 | 1.056e+00 | 2.039e+02 | 0.000e+00 |
| poly(Z, 2)2 | -8.171e+01 | 8.668e-01 | -9.427e+01 | 0.000e+00 |
| poly(R, 4)1 | 3.333e+02 | 1.804e+00 | 1.847e+02 | 0.000e+00 |
| poly(R, 4)2 | -5.845e+01 | 1.651e+00 | -3.539e+01 | 0.000e+00 |
| poly(R, 4)3 | -5.087e+01 | 1.065e+00 | -4.779e+01 | 0.000e+00 |
| poly(R, 4)4 | 6.068e+00 | 7.934e-01 | 7.648e+00 | 2.043e-14 |
| log(C):D | -4.082e-04 | 1.009e-04 | -4.045e+00 | 5.224e-05 |
| D:Z | 4.009e-06 | 2.000e-13 | 2.004e+07 | 0.000e+00 |
| Overdispersion coefficients: | | | | |
| | Estimate | Std. Error | z value | Pr(> \|z\|) |
| phi.(Intercept) | 5.766e-03 | 4.363e-05 | 1.322e+02 | 0e+00 |
| Log-likelihood statistics: | | | | |
| Log-lik | nbpar | df res. | Deviance | AIC |
| -2.234e+05 | 18 | 99301 | 2.108e+05 | 4.468e+05 |

Table 14. **Thunderstorm Data**: Summary of the full beta-binomial model for data grouped into 10 levels for each transformed covariate.

| $M^S_{full,GAM}$ | Estimate | Std. Error | z value | Pr(> \|z\|) |
|---|---|---|---|---|
| (Intercept) | -5.145610 | 0.002509 | -2051 | <2e-16 |
| | edf | Ref.df | F | p-value |
| s(log(C)) | 6.966 | 6.972 | 23498.9 | <2e-16 |
| s(D) | 6.098 | 6.161 | 478.4 | <2e-16 |
| s(log(O)) | 6.924 | 6.997 | 98668.7 | <2e-16 |
| s(Z) | 6.957 | 6.965 | 8061.7 | <2e-16 |
| s(R) | 6.979 | 7.000 | 309746.7 | <2e-16 |
| te(D,Z) | 52.198 | 53.000 | 15654.8 | <2e-16 |
| te(log(C),D) | 50.435 | 53.000 | 5132.3 | <2e-16 |
| | | $\frac{residual\ deviance}{null\ deviance} = 0.85$ | | |
| Log Likelihood: -280382.9 (df=137.5574) | | | | |

Table 15. **Thunderstorm Data**: Summary of the full GAM for data grouped into 10 levels for each transformed covariate.

model weighing between fitting the data well, simplicity, computational speed and ease of application in the setting of a company. In Figure 35 we see plots of the Pearson residuals, fitted values against empirical estimates and the fitted values of the GAM and beta-binomial models plotted against each other. The GLM residuals show no large outliers or strong pattern or relationship. The large residuals from the GAM fit were investigated more closely, and all were found to have very small group sizes. Most fitted values cluster around the diagonal line when plotted against the empirical estimates, and similarly the GAM and beta-binomial fitted values cluster around the diagonal line when plotted against each other. Thus both Models (5.4.6) and (5.4.7), $M^S_{full}$ and $M^S_{fullGAM}$, can be considered as potential models for the thunderstorm data. Since this thesis has been written in collaboration with a company with the goal that the hail and thunderstorm models can be applied to business problems in the future, we take particular note of the ease of application criterion. The team with which we collaborated for this thesis are familiar with the use of GAMs in modelling and coding, and a model of this type can be implemented correctly into operational code with greater ease. Additionally the package MGCV [36], which was used for all coding involving GAMs in this thesis, has thorough function documentation, which may prove very helpful for the case that the code used for this thesis is used or updated for future applications or projects. Whilst the beta binomial logistic model has a log likelihood somewhat larger than the GAM model with a relative difference between the two log likelihoods of approximately 20 percent, we place more importance on the time it would take for our industry partners to become familiar with beta binomial models and the software packages needed to apply these models to new data sets. Thus we select the GAM as our final model.

Finally, we verify our results with those of [28] by comparing look up table plots. In Figures 36, 37 and 38 we consider the covariates $C$ and $R$ whilst holding $D$, $O$ and $Z$ constant at their $5^{th}$, $50^{th}$ and $95^{th}$ percentile values to plot look up tables for Models 5.4.6 and 5.4.7. As for the hail model, we see that both the above plots and those from plots (a) and (b) of Figure 39 have the same general shape.

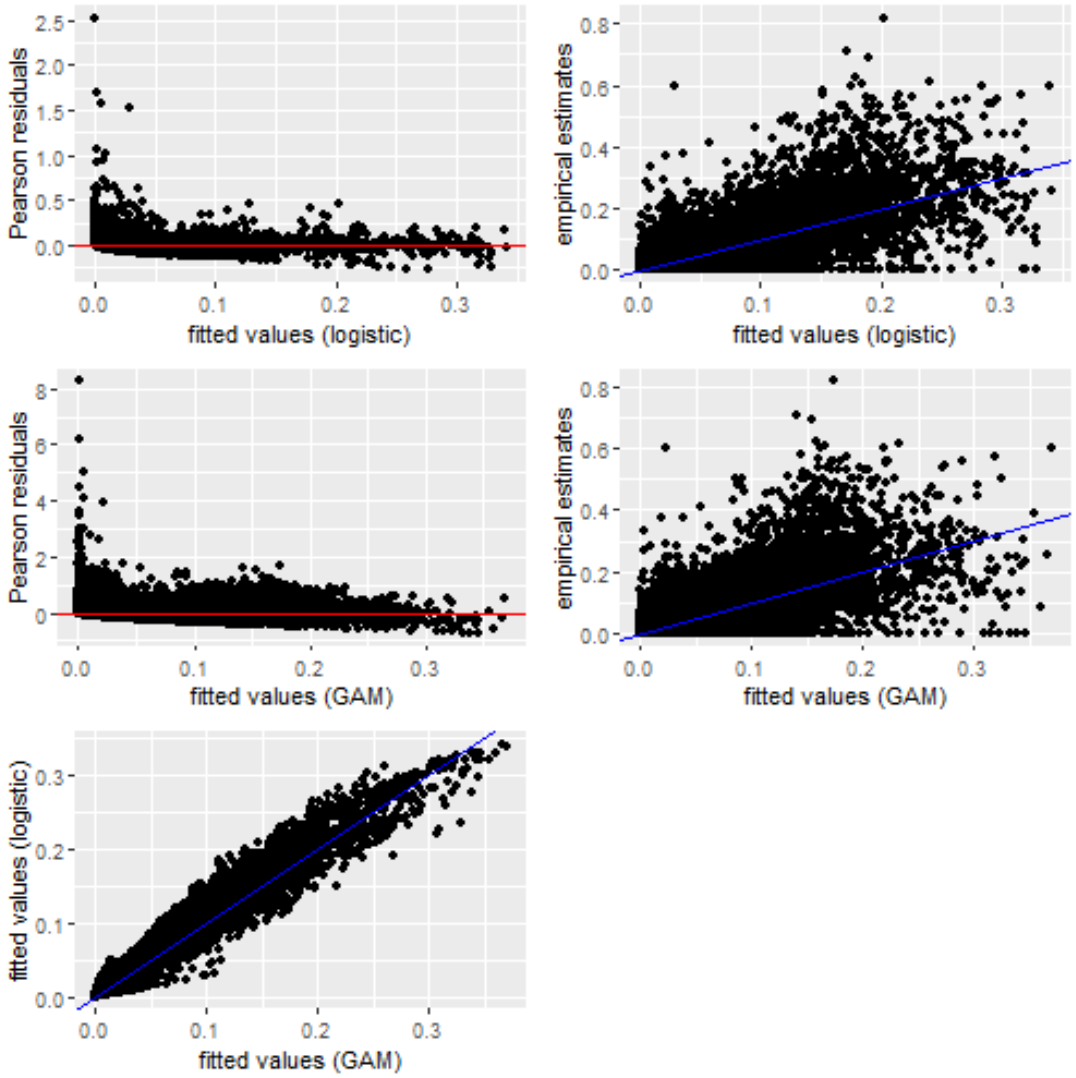


Figure 35. **Thunderstorm Data**: Plots of the fitted values vs empirical estimates and fitted values vs the Pearson residuals for Models 5.4.6 and 5.4.7, $M^S_{full}$ and , $M^S_{fullGAM}$. The Pearson residuals were calculated as in [4] and the fitted values calculated using the R function prediction. Additionally the fitted values for both models have been plotted against each other with the diagonal shown in blue. We see that the GAM residuals show some very large values. These groups all have very small sample sizes and extreme covariate values thus we are not concerned about our model fitting our data poorly in this region of our data.

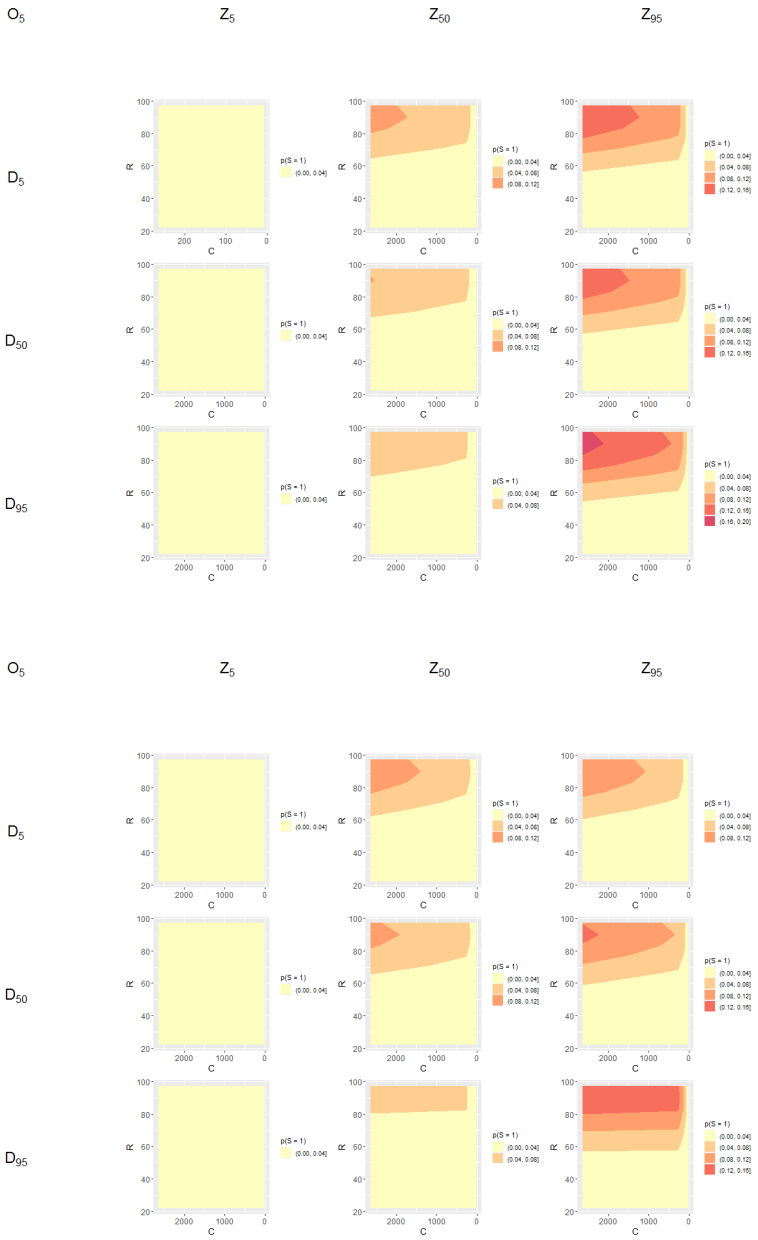


Figure 36. **Thunderstorm Data**: lookup tables for the probability of S = 1 for values of C and R, whilst holding O constant at the $10^{th}$ percentile and D and Z constant at the $10^{th}$, $50^{th}$ and $90^{th}$ for the beta-binomial logistic model and the GAM below.

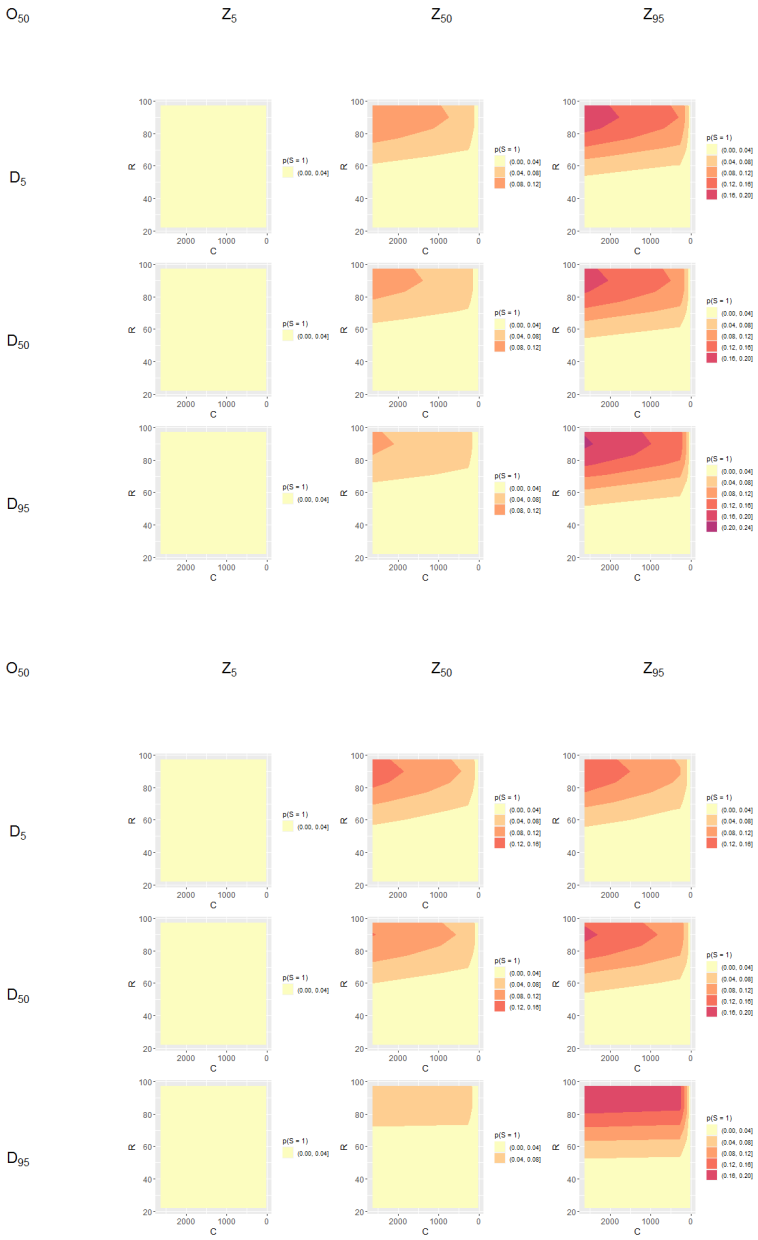


Figure 37. **Thunderstorm Data**: lookup tables for the probability of S = 1 for values of C and R, whilst holding O constant at the $50^{th}$ percentile and D and Z constant at the $10^{th}$, $50^{th}$ and $90^{th}$ for the beta-binomial logistic model above and the GAM below.

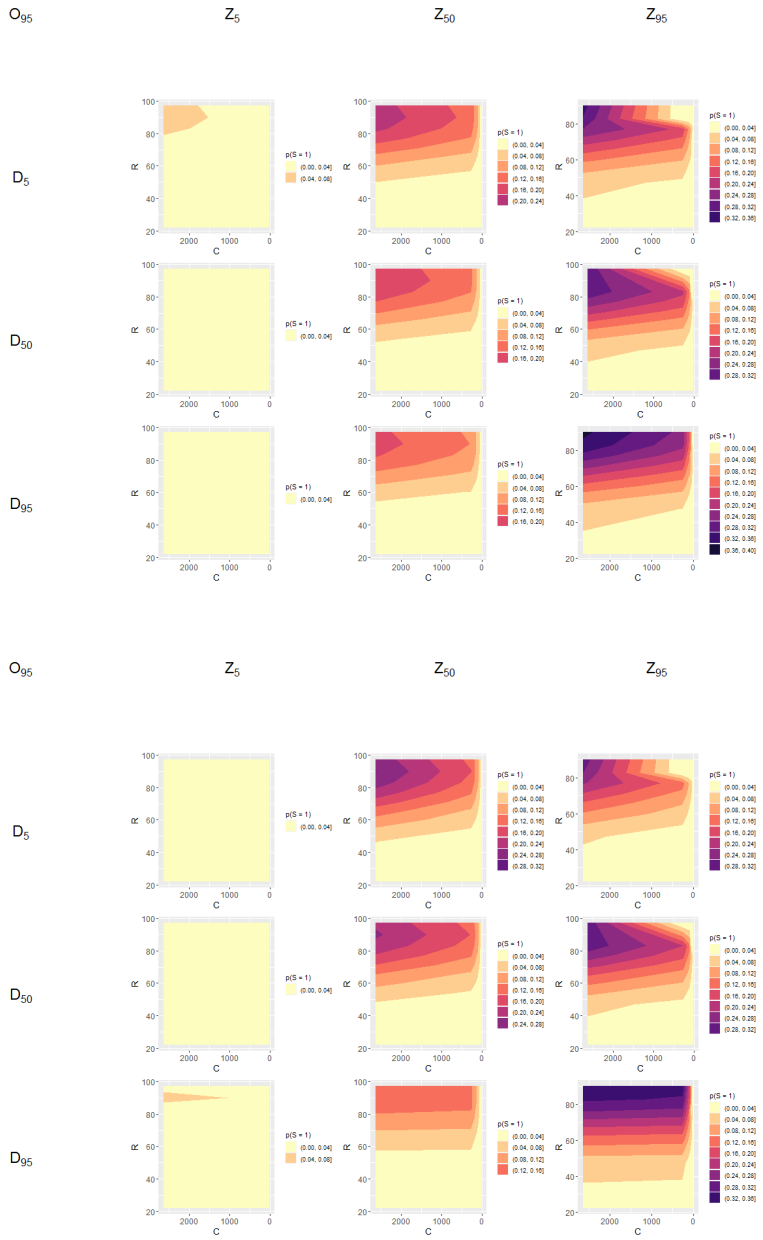


Figure 38. **Thunderstorm Data**: lookup tables for the probability of S = 1 for values of C and R, whilst holding O constant at the $90^{th}$ percentile and D and Z constant at the $10^{th}$, $50^{th}$ and $90^{th}$ for the beta-binomial logistic model and the GAM below.

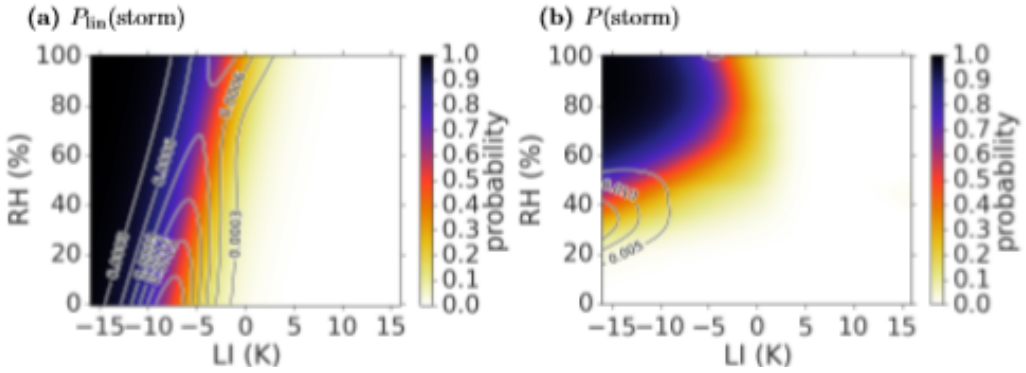


Figure 39. **Thunderstorm Data**: This figure shows the lookup table for LI and R whilst holding D, O and Z constant for the thunderstorm generalised additive and logistic models from [28]. LI is qualitatively similar to C, and was used because the available data was too coarse to calculate C. Larger values of C correspond to more negative values of LI.

# 6 Conclusion

The aim of this thesis was to develop two models for the probability of a thunderstorm and severe hail given that there is a thunderstorm. We first covered the meteorological background of these severe weather phenomena, and why we expected to see a relationship between the covariates and the probability of a thunderstorm or severe hail. We then discussed the statistical theory for binary regression, which we would be applying later in this thesis, including the fitting of a generalised linear model, specifically the case of logistic regression, overdispersion and the fitting of a generalised additive model. The following chapter detailed the extensive preprocessing applied to the data set used in this thesis, prepared so as to be ready for exploratory data analysis techniques to be applied. During the exploratory data analysis both models showed non-linear relationships to the covariates, consequently, for both models a logistic model was compared to a GAM. In the case of the thunderstorm model a beta-binomial logistic model was considered, to allow for overdispersion in the data.

We see from the analysis conducted in Sections 5.1 and 5.2 that the GLM and GAM approach lead to hail models with very similar results. The log likelihood values from the two summary outputs have a relative difference of approximately 1 percent, and the fitted values from the two models agree with each other, see Figure 20. In addition we see that Figures 21, 22 and 23 are in agreement with Figure 24, suggesting that our hail models are consistent with those in [28]. Thus we choose to recommend Model (5.2.6) as our hail model, because a GLM is simpler and faster to calculate than a GAM.

As with the hail model, the analysis conducted in Sections 5.3 and 5.4 showed that the GAM and beta-binomial logistic model approaches lead to thunderstorm models with similar results. Whilst the log likelihood of the beta-binomial logistic regression model is larger than that of the GAM with a relative difference of approximately 20 percent, Figure 35 shows that the fitted values from the two models agree with each other. In addition we see that Figures 36, 37 and 38 are in agreement with Figure 39, suggesting that our thunderstorm models are consistent with those in [28]. We also take into account that this thesis has been written together in partnership with Munich RE, and that it is a major goal that in the future the work from this thesis will be applied by the meteorologists with whom we worked. The team with which we collaborated for this work are familiar with the use of GAMs, along with the way in which these models can be applied to new data sets using the software packages discussed in this thesis. The simplicity of choosing a model type with which the industry partners of this thesis have experience applying provides the greatest benefit as it will dramatically aid the application of the results to current projects. Thus we choose to recommend Model (5.4.7) as our thunderstorm model.

# 7 Appendix

Additional plots for the selection of the thunderstorm cutoff value have been included, along with additional model fitting analysis.

## 7.1 Additional Thunderstorm Cutoff Plots

This section contains the two plots which were combined to make Figure 9, which was used to choose the CAPE cutoff value for the thunderstorm model.

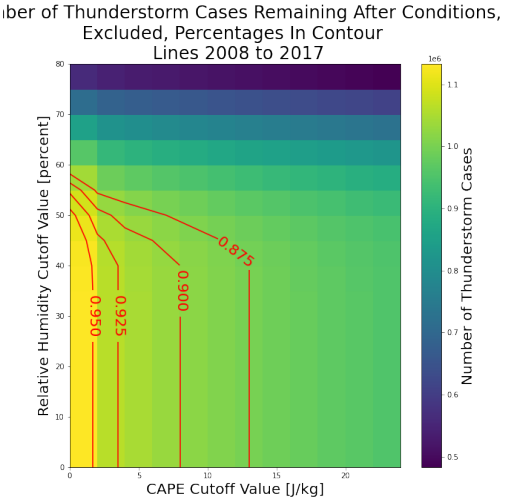


Figure 40. The number of thunderstorm cases remaining when all data points with values for CAPE ($C_{ijt}$) and relative humidity ($R_{ijt}$) less than the cutoff values which are shown on the x and y axes are excluded. The contour lines show the fractions remaining (0.95, 0.925, 0.90 and 0.875)

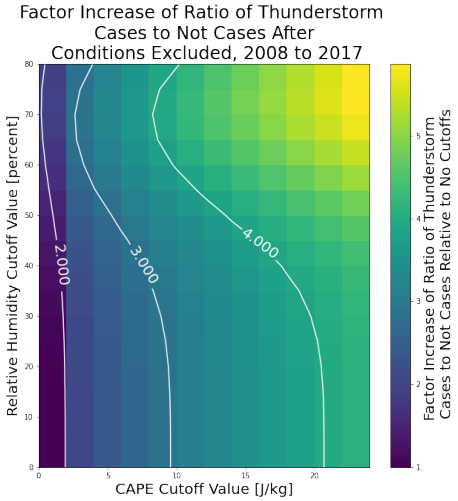


Figure 41. The factor by which the ratio of thunderstorm successes to failures increases when all data points with values for CAPE ($C_{ijt}$) and relative humidity ($C_{ijt}$) less than the cutoff values which are shown on the x and y axes are excluded.

## 7.2 Additional Smaller Models for One to Four Covariates for the Hail Model

Here additional fits for the hail model are shown. The smaller models for the hail data are defined below: the binomial model for one covariate, $C$,

$$M^H_{lnC2}: \ \text{logit}(p_i) = \beta_0 + \beta_1 \ln(C_i) + \beta_2(\ln(C_i))^2, \tag{7.2.1}$$

the binomial model for two covariates, $C$ and $D$,

$$M^H_{lnCD2} : \text{logit}(p_i) = \beta_0 + \beta_1 \ln(C_i) + \beta_2(\ln(C_i))^2 + \beta_3 \ln(D_i) + \beta_4(\ln(D_i))^2, \tag{7.2.2}$$

the binomial model for three covariates, $C$, $D$ and $O$,

$$\begin{aligned} M^H_{lnCDO2} : \text{logit}(p_i) =& \beta_0 + \beta_1 \ln(C_i) + \beta_2(\ln(C_i))^2 + \beta_3 \ln(D_i) + \beta_4(\ln(D_i))^2 + \\ & \beta_5 \ln(O_i) + \beta_6(\ln(O_i))^2, \end{aligned} \tag{7.2.3}$$

and the binomial model for four covariates, $C$, $D$, $O$ and $Z$,

$$\begin{aligned} M^H_{lnCDO2,Z3} : \text{logit}(p_i) =& \beta_0 + \beta_1 \ln(C_i) + \beta_2(\ln(C_i))^2 + \beta_3 \ln(D_i) + \beta_4(\ln(D_i))^2 + \\ & \beta_5 \ln(O_i) + \beta_6(\ln(O_i))^2 + \beta_7 Z_i + \beta_8 Z_i^2 + \beta_9 Z_i^3, \end{aligned} \tag{7.2.4}$$

where $\beta_i$ are the regression coefficients. The summary outputs for Models (7.2.1), (7.2.2), (7.2.3) and (7.2.4) are shown in Table 16. Additionally the summary output for univariate binary logistic regressions of the ungrouped data are shown in Table 17.

| $M^H_{lnC2}$ | Estimate | Std. Error | z value | Pr(> \|z\|) |
|---|---|---|---|---|
| (Intercept) | -6.0033 | 0.0264 | -226.97 | 0.0000 |
| ln(C)1 | 37.8838 | 1.3765 | 27.52 | 0.0000 |
| ln(C)2 | -4.1683 | 1.4100 | -2.96 | 0.0031 |
| Null deviance: 4930.8 on 3068 degrees of freedom | | | | |
| Residual deviance: 3890.3 on 3066 degrees of freedom | | | | |

| $M^H_{lnCD2}$ | Estimate | Std. Error | z value | Pr(> \|z\|) |
|---|---|---|---|---|
| (Intercept) | -6.1430 | 0.0292 | -210.24 | 0.0000 |
| ln(C)1 | 39.9216 | 1.3804 | 28.92 | 0.0000 |
| ln(C)2 | -4.9004 | 1.4121 | -3.47 | 0.0005 |
| ln(D)1 | 31.2359 | 1.5235 | 20.50 | 0.0000 |
| ln(D)2 | -4.7732 | 1.3456 | -3.55 | 0.0004 |
| Null deviance: 4930.8 on 3068 degrees of freedom | | | | |
| Residual deviance: 3326.1 on 3064 degrees of freedom | | | | |

| $M^H_{lnCDO2}$ | Estimate | Std. Error | z value | Pr(> \|z\|) |
|---|---|---|---|---|
| (Intercept) | -6.2112 | 0.0305 | -203.60 | 0.0000 |
| ln(C)1 | 39.2703 | 1.3803 | 28.45 | 0.0000 |
| ln(C)2 | -4.7564 | 1.4125 | -3.37 | 0.0008 |
| ln(D)1 | 30.8308 | 1.5283 | 20.17 | 0.0000 |
| ln(D)2 | -4.6605 | 1.3457 | -3.46 | 0.0005 |
| ln(O)1 | -14.7472 | 1.4132 | -10.44 | 0.0000 |
| ln(O)2 | -18.5136 | 1.2957 | -14.29 | 0.0000 |
| Null deviance: 4930.8 on 3068 degrees of freedom | | | | |
| Residual deviance: 3030.1 on 3062 degrees of freedom | | | | |

| $M^H_{lnCDO2,Z3}$ | Estimate | Std. Error | z value | Pr(> \|z\|) |
|---|---|---|---|---|
| (Intercept) | -6.2391 | 0.0313 | -199.34 | 0.0000 |
| ln(C)1 | 38.9610 | 1.4965 | 26.03 | 0.0000 |
| ln(C)2 | -4.5687 | 1.4198 | -3.22 | 0.0013 |
| ln(D)1 | 30.7938 | 1.5473 | 19.90 | 0.0000 |
| ln(D)2 | -4.4355 | 1.3472 | -3.29 | 0.0010 |
| ln(O)1 | -13.8458 | 1.6189 | -8.55 | 0.0000 |
| ln(O)2 | -19.5445 | 1.3538 | -14.44 | 0.0000 |
| Z1 | 4.4597 | 1.8981 | 2.35 | 0.0188 |
| Z2 | -4.1568 | 1.3625 | -3.05 | 0.0023 |
| Z3 | 10.8704 | 1.2258 | 8.87 | 0.0000 |
| Null deviance: 4930.8 on 3068 degrees of freedom | | | | |
| Residual deviance: 2945.4 on 3060 degrees of freedom | | | | |

Table 16. **Hail Data**: Summaries of the binomial models for one through to four covariates for data grouped into 5 levels for each transformed covariate. The covariates were chosen using step-wise regression.

| $M^H_{lnC2}$ | Estimate | Std. Error | z value | Pr(> \|z\|) |
|---|---|---|---|---|
| (Intercept) | -6.0198 | 0.0267 | -225.46 | 0.0000 |
| ln(C)1 | 582.5054 | 22.8438 | 25.50 | 0.0000 |
| ln(C)2 | 41.7480 | 16.7434 | 2.49 | 0.0127 |
| Null deviance: 30194 on 715817 degrees of freedom | | | | |
| Residual deviance: 29036 on 715815 degrees of freedom | | | | |
| $M^H_{lnD2}$ | Estimate | Std. Error | z value | Pr(> \|z\|) |
| (Intercept) | -5.8999 | 0.0250 | -235.62 | 0.0000 |
| ln(D)1 | 483.0693 | 32.7193 | 14.76 | 0.0000 |
| ln(D)2 | -15.1913 | 31.3023 | -0.49 | **0.6275** |
| Null deviance: 30194 on 715817 degrees of freedom | | | | |
| Residual deviance: 29712 on 715815 degrees of freedom | | | | |
| $M^H_{lnO2}$ | Estimate | Std. Error | z value | Pr(> \|z\|) |
| (Intercept) | -6.2053 | 0.0481 | -129.14 | 0.0000 |
| ln(O)1 | 1472.8223 | 154.7045 | 9.52 | 0.0000 |
| ln(O)2 | -1333.6531 | 106.9105 | -12.47 | 0.0000 |
| Null deviance: 30194 on 715817 degrees of freedom | | | | |
| Residual deviance: 29744 on 715815 degrees of freedom | | | | |
| $M^H_{Z3}$ | Estimate | Std. Error | z value | Pr(> \|z\|) |
| (Intercept) | -5.9739 | 0.0309 | -193.44 | 0.0000 |
| Z1 | 657.5119 | 55.2061 | 11.91 | 0.0000 |
| Z2 | -255.1010 | 71.6961 | -3.56 | 0.0004 |
| Z3 | 198.7689 | 40.9344 | 4.86 | 0.0000 |
| Null deviance: 30194 on 715817 degrees of freedom | | | | |
| Residual deviance: 29565 on 715814 degrees of freedom | | | | |
| $M^H_R$ | Estimate | Std. Error | z value | Pr(> \|z\|) |
| (Intercept) | -5.8078 | 0.0220 | -263.76 | 0.0000 |
| R | -236.9935 | 17.6622 | -13.42 | 0.0000 |
| Null deviance: 30194 on 715817 degrees of freedom | | | | |
| Residual deviance: 30015 on 715816 degrees of freedom | | | | |

Table 17. **Hail Data**: Summaries of univariate logitstic models for the ungrouped data for each variable after transformation, where $ln$ denotes the log transformation. We see that the coefficient for the quadratic effect of D is not significant, which is highlighted in bold.

## 7.3 Additional Smaller Models for One to Four Covariates for the Thunderstorm Model

Here additional fits for the thunderstorm model are shown. The smaller models for the thunderstorm model are defined below: the binomial model for one covariate, $C$,

$$M^S_{lnC3} : \text{logit}(p_i) = \beta_0 + \beta_1 \ln(C_i) + \beta_2(\ln(C_i))^2 + \beta_3(\ln(C_i))^3, \tag{7.3.1}$$

the binomial model for two covariates, $C$ and $D$,

$$M^S_{lnC3,R4} : \text{logit}(p_i) = \beta_0+\beta_1 \ln(C_i)+\beta_2(\ln(C_i))^2+\beta_3(\ln(C_i))^3+\beta_4 R_i+\beta_5 R_i^2+\beta_6 R_i^3+\beta_7 R_i^4, \tag{7.3.2}$$

the binomial model for three covariates, $C$, $D$ and $O$,

$$\begin{aligned} M^S_{lnC3,R4,Z2} : \text{logit}(p_i) =&\beta_0 + \beta_1 \ln(C_i) + \beta_2(\ln(C_i))^2 + \beta_3(\ln(C_i))^3 + \beta_4 R_i + \beta_5 R_i^2+ \\ &\beta_6 R_i^3 + \beta_7 R_i^4 + \beta_8 Z_i + \beta_9 Z_i^2 + \beta_{10} Z^3 + \beta_{11} Z^4, \end{aligned} \tag{7.3.3}$$

and the binomial model for four covariates, $C$, $D$, $O$ and $Z$,

$$\begin{aligned} M^S_{lnC3,R4,Z2,lnO2} : \text{logit}(p_i) =&\beta_0 + \beta_1 \ln(C_i) + \beta_2(\ln(C_i))^2 + \beta_3(\ln(C_i))^3 + \beta_4 R_i + \beta_5 R_i^2+ \\ &\beta_6 R_i^3 + \beta_7 R_i^4 + \beta_8 Z_i + \beta_9 Z_i^2 + \beta_{10} Z^3 + \beta_{11} Z^4 \\ &+ \beta_{12} \ln(O_i) + \beta_{13}(\ln(O_i))^2 + \beta_{14}(\ln(O_i))^3, \end{aligned} \tag{7.3.4}$$

where $\beta_i$ are the regression coefficients. The summary outputs for Models (7.3.1), (7.3.2), (7.3.3) and (7.3.4) are shown in Table 18. Additionally the summary output for univariate binary logistic regressions of the ungrouped data are shown in Table 19.

| $M^S_{lnC3}$ | Estimate | Std. Error | z value | Pr(> \|z\|) |
|---|---|---|---|---|
| (Intercept) | -5.1785 | 0.0025 | -2078.85 | 0.0000 |
| ln(C)1 | 383.3814 | 0.6845 | 560.12 | 0.0000 |
| ln(C)2 | -96.2823 | 0.6917 | -139.19 | 0.0000 |
| ln(C)3 | -40.0898 | 0.5394 | -74.32 | 0.0000 |
| Null deviance: 1600768 on 99318 degrees of freedom | | | | |
| Residual deviance: 753471 on 99315 degrees of freedom | | | | |

| $M^S_{lnC3,R4}$ | Estimate | Std. Error | z value | Pr(> \|z\|) |
|---|---|---|---|---|
| (Intercept) | -5.6689 | 0.0032 | -1774.85 | 0.0000 |
| ln(C)1 | 406.3911 | 0.6868 | 591.68 | 0.0000 |
| ln(C)2 | -69.3715 | 0.6866 | -101.03 | 0.0000 |
| ln(C)3 | -45.9254 | 0.5393 | -85.16 | 0.0000 |
| R1 | 411.1570 | 1.1879 | 346.13 | 0.0000 |
| R2 | -152.6116 | 0.9637 | -158.36 | 0.0000 |
| Null deviance: 1600768 on 99318 degrees of freedom | | | | |
| Residual deviance: 449317 on 99313 degrees of freedom | | | | |

| $M^S_{lnC3,R4,Z2}$ | Estimate | Std. Error | z value | Pr(> \|z\|) |
|---|---|---|---|---|
| (Intercept) | -5.8217 | 0.0033 | -1742.22 | 0.0000 |
| ln(C)1 | 354.6892 | 0.7003 | 506.47 | 0.0000 |
| ln(C)2 | -51.8639 | 0.6869 | -75.51 | 0.0000 |
| ln(C)3 | -41.7084 | 0.5387 | -77.43 | 0.0000 |
| R1 | 419.2269 | 1.1777 | 355.98 | 0.0000 |
| R2 | -130.1974 | 0.9573 | -136.01 | 0.0000 |
| Z1 | 207.1160 | 0.7618 | 271.88 | 0.0000 |
| Z2 | -108.1027 | 0.5510 | -196.18 | 0.0000 |
| Null deviance: 1600768 on 99318 degrees of freedom | | | | |
| Residual deviance: 342900 on 99311 degrees of freedom | | | | |

| $M^S_{lnC3,R4,Z2,lnO2}$ | Estimate | Std. Error | z value | Pr(> \|z\|) |
|---|---|---|---|---|
| (Intercept) | -5.9272 | 0.0034 | -1721.13 | 0.0000 |
| ln(C)1 | 324.6230 | 0.7087 | 458.07 | 0.0000 |
| ln(C)2 | -52.3752 | 0.6868 | -76.26 | 0.0000 |
| ln(C)3 | -38.3078 | 0.5392 | -71.04 | 0.0000 |
| R1 | 435.0256 | 1.1963 | 363.65 | 0.0000 |
| R2 | -129.2099 | 0.9665 | -133.68 | 0.0000 |
| Z1 | 263.4182 | 0.8103 | 325.09 | 0.0000 |
| Z2 | -95.6833 | 0.5512 | -173.59 | 0.0000 |
| ln(O)1 | 100.8414 | 0.4736 | 212.91 | 0.0000 |
| ln(O)2 | 48.0208 | 0.4374 | 109.78 | 0.0000 |
| Null deviance: 1600768 on 99318 degrees of freedom | | | | |
| Residual deviance: 271251 on 99309 degrees of freedom | | | | |

Table 18. **Thunderstorm Data**: Summaries of the binomial models for one through to four covariates for data grouped into 5 levels for each transformed covariate.

| $M^S_{lnC3}$ | Estimate | Std. Error | z value | Pr(> \|z\|) |
|---|---|---|---|---|
| (Intercept) | -5.1822 | 0.0025 | -2099.17 | 0.0000 |
| ln(C)1 | 8928.0792 | 16.3876 | 544.81 | 0.0000 |
| ln(C)2 | -1263.8584 | 14.2384 | -88.76 | 0.0000 |
| ln(C)3 | -1382.6578 | 10.2701 | -134.63 | 0.0000 |
| Null deviance: 6909968 on 52693121 degrees of freedom | | | | |
| Residual deviance: 6044957 on 52693118 degrees of freedom | | | | |

| $M^S_{D5}$ | Estimate | Std. Error | z value | Pr(> \|z\|) |
|---|---|---|---|---|
| (Intercept) | -4.4559 | 0.0014 | -3165.19 | 0.0000 |
| D1 | -2644.8376 | 17.5260 | -150.91 | 0.0000 |
| D2 | -1801.3218 | 24.3006 | -74.13 | 0.0000 |
| D3 | -1019.1025 | 19.5650 | -52.09 | 0.0000 |
| Null deviance: 6909968 on 52693121 degrees of freedom | | | | |
| Residual deviance: 6867755 on 52693118 degrees of freedom | | | | |

| $M^S_{lnO2}$ | Estimate | Std. Error | z value | Pr(> \|z\|) |
|---|---|---|---|---|
| (Intercept) | -4.4832 | 0.0014 | -3295.81 | 0.0000 |
| ln(O)1 | 2215.1431 | 11.8632 | 186.72 | 0.0000 |
| ln(O)2 | 2006.4976 | 9.7985 | 204.78 | 0.0000 |
| Null deviance: 6909968 on 52693121 degrees of freedom | | | | |
| Residual deviance: 6802569 on 52693119 degrees of freedom | | | | |

| $M^S_{Z2}$ | Estimate | Std. Error | z value | Pr(> \|z\|) |
|---|---|---|---|---|
| (Intercept) | -4.8419 | 0.0022 | -2220.95 | 0.0000 |
| Z1 | 8116.5961 | 20.5560 | 394.85 | 0.0000 |
| Z2 | -3075.2982 | 14.3971 | -213.61 | 0.0000 |
| Null deviance: 6909968 on 52693121 degrees of freedom | | | | |
| Residual deviance: 6583958 on 52693119 degrees of freedom | | | | |

| $M^S_{R2}$ | Estimate | Std. Error | z value | Pr(> \|z\|) |
|---|---|---|---|---|
| (Intercept) | -4.7776 | 0.0020 | -2355.51 | 0.0000 |
| R1 | 7034.6489 | 20.0688 | 350.53 | 0.0000 |
| R2 | -5767.2379 | 18.7481 | -307.62 | 0.0000 |
| Null deviance: 6909968 on 52693121 degrees of freedom | | | | |
| Residual deviance: 6667705 on 52693119 degrees of freedom | | | | |

Table 19. **Thunderstorm Data**: Summaries of univariate logitstic models for each covariate after transformation, where *ln* denotes the log transformation.

# References

[1] C. Donald Ahrens. *Meteorology Today: An Introduction to Weather, Climate and The Environment*. Brooks / Cole, Cengage Leanring, Belmont, ninth edition, 2009, 2007.

[2] John T. Allen, Ian M. Giammanco, Matthew R. Kumjian, Heinz Jurgen Punge, Qinghong Zhang, Pieter Groenemeijer, Michael Kunz, and Kiel Ortega. Understanding Hail in the Earth System. *Reviews of Geophysics*, 58(1), mar 2020.

[3] S. A. Changon, D. Changon, and S. Hilberg. *Hailstorms Across the Nation: An Atlas about Hail and Its Damages*. Illinois State Water Survey, Champaign, Illinois, United States of America, 2009. https://www.ideals.illinois.edu/bitstream/handle/2142/15156/ISWSCR2009-12.pdf?sequence=4.

[4] C. Czado, E. Brechmann, and N. Kraemer. Gernalized Linear Models with Applications, 2018.

[5] Charles A. Doswell, editor. *Severe Convective Storms*. American Meteorological Society, Boston, 2001.

[6] ECMWF. Era-interim. URL = https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era-interim, accessed 03.01.2020.

[7] ECMWF. Era5. URL = https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5, accessed 03.01.2020.

[8] Copernicus Program ECMWF. Climate reanalysis: Era5. URL = https://climate.copernicus.eu/climate-reanalysis, accessed 03.01.2020.

[9] E. Garcia-Ortega, L. Fita, R. Romero, L. Lopez, C. Ramis, and J.L. Sanchez. Numerical simulation and sensitivity study of a severe hailstorm in northeast Spain. *Atmospheric Research*, 83:225–241, 2007. doi:10.1016/j.atmosres.2005.08.004.

[10] Karl Hennermann. Era5: surface elevation and orography. URL = https://confluence.ecmwf.int/display/CKB/ERA5accessed 03.01.2020, last modified: 2020.

[11] H. et. al. Hersbach. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, pages 1999–2049, 2020. DOI: 10.1002/qj.3803.

[12] James R. Holton. *An Introduction to Dynamic Meteorolgy*. Elsevir Academic Press, fourth edition, 2004.

[13] Á. Hovráth, I. Geredesi, P. Németh, K. Csirmaz, and F. Dombai. Numerical modeling of severe convective storms occurring in the Carpathian Basin. *Atmospheric Research*, 93:221–237, 2009.

[14] S. Hoyer and J. Hamman. xarray: N-D labeled arrays and datasets in Python. *Journal of Open Research Software*, 5(1), 2017.

[15] M.A. Islam and R.I. Chowdhury. *Generalized Linear Models. In: Analysis of Repeated Measures Data*. Springer, Singapore, 2017.

[16] Kelvyn Jones and Neil Wrigley. Generalized Additive Models, Graphical Diagnostics, and Logistic Regression. *Geographical Analysis*, 27(1):1–18, 1995.

[17] L. King, G.and Zeng. Logistic regression in rare events data. *Political Analysis*, 9:137–163, 2001.

[18] European Severe Storms Laboratory. Eswd quality control. URL = https://www.essl.org/cms/european-severe-weather-database/eswd-quality-control/, accessed 14.01.2020.

[19] Lesnoff, M., Lancelot, and R. *aod: Analysis of Overdispersed Data*, 2012. R package version 1.3.1.

[20] Paul Markowski and Yvette Richardson. *Mesoscale Meteorology in Midlatitudes*. Wiley-Blackwell, Chichester, 2010.

[21] Met Office. *Iris: A Python library for analysing and visualising meteorological and oceanographic data sets*. Exeter, Devon, v1.2 edition, 2010 - 2013.

[22] G. Micula and S. Micula. *Handbook of Splines*. Springer, Romania and U.S.A, 1999.

[23] S. Mohr, M. Kunz, and B. Geyer. Hail potential in Europe based on a regional climate model hindcast. *Geophysical Research Letters*, 42:904–912, 2015. doi:10.1002/2015GL067118.

[24] W. Pilorz, I. Laskowski, E. Łupikasza, and M. Taszarek. Wind Shear and the Strength of Severe Convective Phenomena—Preliminary Results from Poland in 2011–2015. *Climate*, 51:7, 2007. doi:10.3390/cli4040051.

[25] A. F. Prein and G. J. Holland. Global estimates of damaging hail hazard. *Weather and Climate Extremes*, 22:10–23, 2018. https://doi.org/10.1016/j.wace.2018.10.004.

[26] Tomáš Púčik, Christopher Castellano, Pieter Groenemeijer, Thilo Kühne, Anja T. Rädler, Bogdan Antonescu, and Eberhard Faust. Large Hail Incidence and Its Economic and Societal Impacts across Europe. *Monthly Weather Review*, 147(11):3901–3916, nov 2019.

[27] Anja T. Rädler, Pieter Groenemeijer, Eberhard Faust, and Robert Sausen. Detecting Severe Weather Trends Using an Additive Regressive Convective Hazard Model (ARCHaMo). *Journal of Applied Meteorology and Climatology*, 57(3):569–587, mar 2018.

[28] Anja Theresa Raedler. *Modeling of convective storm hazard occurrence , taking convective initiation explicitly into account*. PhD thesis, Ludwig Maximilians University, 2018.

[29] Karline Soetaert. Package 'plot3d', 2019.

[30] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

[31] E. F. Vonesh. *Generalized Linear and Nonlinear Models for Correlated Data: Theory and Applications Using SAS*. SAS Institute Inc, Cary, NC, United States of America, 2012.

[32] J. M. Wallace and P. V. Hobbs. *Atmospheric Science: An Introductory Survey*. Elsevir, Canada, second edition, 2006.

[33] Anja T. Westermayer, Pieter Groenemeijer, Georg Pistotnik, Robert Sausen, and Eberhard Faust. Identification of favorable environments for thunderstorms in reanalysis data. *Meteorologische Zeitschrift*, 26(1):59–70, feb 2017.

[34] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[35] D. S. Wilks. *Statistical Methods in the Atmospheric Sciences*. Elsevir, United States of America, second edition, 2006.

[36] Simon Wood. Package ‘ mgcv ’. Technical report, 2017.

[37] Simon N. Wood. *Generalized Additive Models : An Introduction with R, Second Edition*. CRC Press LLC, second edition, 2017.