

From manual classification to transformer-based language models: assessing the quality and consistency of historical convective event records

Franck Schätz¹ and Rüdiger Glaser¹

¹University of Freiburg, Institute of Environmental Social Sciences and Geography, Chair of Physical Geography, Stefan-Meier-Strasse 76, 79104 Freiburg, Germany

Correspondence: Franck Schätz (franck.schaetz@geographie.uni-freiburg.de)

Abstract. ~~Historical text sources represent a central, yet methodologically challenging basis for the reconstruction~~ This article investigates whether Transformer-based language models could replace the labour-intensive, manual classification of convective weather events. This study examines the extent to which historical reports on thunderstorms and hailstorms contain reliable climatological information, despite heterogeneous sources, varying degrees of detail and linguistic diversity. Based on a corpus prepared using source criticism, qualitative descriptions are converted into structured evidence levels and intensity classes and analysed using statistical methods and a multilingual BERT language model ~~phenomena found in written texts from the pre-instrumental measurement era. The training set is based on a corpus of 6,999 written observations from 494 Central European sources, spanning the period from 1000 to 1817. This corpus has been linguistically normalised, ranging from Middle High German to Contemporary German.~~

For the training set, the text sources containing observations of thunderstorm and hail events are first classified manually using a formalised procedure. Quality assurance is initially carried out at the level of the existing textual information. To this end, evidence classes are introduced. These are based on linguistic evidence regarding the associated phenomena of thunderstorm and hail events. In the next step, the plausibility of the classified thunderstorm and hail events is verified by checking their consistency with the DWD normal periods (1961–1990, 1991–2020).

~~The reconstructed time series show a distinctly stable seasonal signal with a dominant summer maximum that occurs independently of fluctuations in source density and is consistent both in the overall series and in a dense observation window. A comparison with modern observation data from the German Weather Service and with independent historical reconstructions shows a high degree of agreement in seasonal patterns despite different survey methods and time periods. Analysis of the intensity classes also shows that historical sources do not primarily document extreme events, but rather reflect a physically plausible ranking of event strengths.~~

~~The results of the automated classification prove that the language model reliably reproduces seasonal and intensity-related patterns and implicitly captures source-specific reporting patterns without levelling them. Overall, the study shows that AI-supported methods can extract robust climatological information from historical texts when processed using rigorous methods, thus opening up new perspectives for quantitative historical climate research~~ seasonal signal is preserved across four language stages and nine source types, from the early 15th to the early 19th century. The classified thunderstorm and

hail events exhibit a plausible physical signal and show strong correlations with current observations. The trained models “ThunderstormBERT” and “HailBERT” achieve macro-F1 scores of 0.83 and 0.93 respectively. Misclassifications primarily impact the middle thunderstorm class (moderate thunderstorm), where the classification scheme is least clear-cut.

1 Introduction

Written observations provide a wealth of direct information on weather phenomena. Before the introduction of instrumental and radar measurements, textual sources were often the only records of mesoscale weather events such as thunderstorm and hail events, with the sequence, intensity, spatial extent and effects of individual events often documented in detail. They are indispensable not only for reconstructing past weather and extreme events, but also for addressing current issues, as modern measurement networks and radar data series do not go back far enough in time to reliably capture the variability and long-term trends of convective processes. Historical observations also provide information on the seasonal distribution, frequency and intensity of documented events and their consequences and thus form a central basis for extreme weather analyses and applications in risk management. Against this background, texts represent an essential source of data, as they enable consistent classification of mesoscale event dynamics over long periods of time (Martius et al., 2018; Kahraman et al., 2024). They describe the course, intensity and extent of an event, as well as the damage it causes (e.g. Punge and Kunz, 2016; Taszarek et al., 2019; Hawkins et al., 2023; Luterbacher et al., 2024) and remain the most important basis for investigating the variability and risk management of extreme weather events (Brázdil et al., 2016b, a; Diodato et al., 2019; Giordani et al., 2024; Hulton and Schultz, 2024; Brönnimann et al., 2019; Glaser, 2013; Stahl et al., 2016; Erfurt et al., 2020; Cutter, 2021).

Despite this potential, historical texts have only been used to a limited extent so far, as climate-related information must be identified and evaluated manually for the most part. Each. The use of written weather observations is limited not so much by the availability of sources as by the effort involved in processing them: each event must be recognised individually manually identified, contextualised and classified in terms of its meteorological significance. With large text corpora, there is therefore not a lack of sources, but a lack of processing capacity. Modern developments in the field of natural language processing and, in particular, large. Consequently, only parts of large corpora have been analysed to date. Transformer-based language models such as BERT (Devlin et al., 2019) are opening up the possibility for the first time of systematic and scalable thematic potentially capable of carrying out high-quality and scalable automated analysis of historical texts on the topics of climate, weather and risk events risks (see, e.g., Webersinke et al., 2022; Zhou et al., 2022; Sakaji and Kaneda, 2023).

However, creating Training such models requires a linguistically and climatologically validated data basis data that has been validated from both a linguistic and a climatological perspective. Historical climate records are linguistically heterogeneous, idiomatic and semantically strongly tied to the time of their origin. They therefore written in the language of their era (Glaser, 2013; Schätz, 2023) and differ fundamentally from modern scientific weather reports weather observations (Grzega, 2022). Furthermore, they are not standardised measurement data, but subjective observational descriptions that vary greatly in terms. They consist of descriptions whose level of detail, temporal and spatial precision, and meteorological interpretation. Without quality control, there is a risk that a language model will learn and reproduce stylised or traditional forms of description instead

of actual weather events. Therefore, a procedure is required that ensures both linguistic and climatological plausibility before such data can be used for fine-tuning a large language model vary considerably (Brázdil et al., 2010; Glaser, 2013; Brönnimann et al., 2019). If these characteristics, as well as the physical plausibility of the manual classifications, are not taken into account when generating the model, the results cannot be interpreted. Consistent and validated data, in terms of both their linguistic and climate plausibility, are therefore a prerequisite for automation using Transformer-based language models.

There is a particular need for research into thunderstorms and hail: previous work has mainly recorded historical events using simple classifications (such as The analysis of thunderstorm and hail events based on historical textual sources is particularly under-represented. In existing studies, historical events are classified according to frequency, intensity or damage), but (see Lenke, 1960; Camuffo et al., 2000; Gudd, 2004; Rohr, 2009; Brázdil et al., 2016a, b; Huang et al., 2022), none of them provides a formally defined and reproducible classification scheme for text sources is still lacking. This means that there is no approach that systematically takes into account both, reproducible framework that links the linguistic variability of historical descriptions and with meteorological interpretation.

The aim of this study is to develop a scalable method for quality-assured, automated classification of historical To bridge this gap, we are developing a text-based classification method that enables climate-relevant information on thunderstorm and hail events to be systematically recorded and categorised. To this end, a formally defined classification procedure is being developed to serve as the basis for an AI-supported classification procedure. Using the period 1000–1817 as an example, historical event descriptions will first be systematically classified and statistically validated. The text data validated classification scheme based on associated phenomena is defined ex ante and applied to a corpus of 6,999 written observations from Central European sources dating from 1000 to 1817. The data are checked at the text level for consistency and quality and verified for physical plausibility against the DWD normal periods (1961–1990 and 1991–2020).

The dataset prepared in this way will form the basis for fine-tuning a BERT model, which will then be used for the automated identification and classification of further event descriptions. The study thus demonstrates the potential of modern large language models for the structured evaluation of historical text sources without undermining established scientific standards of climatological data quality. To promote follow-up research, all is then used to train Transformer-based language models. These models can analyse further text sources on the basis of the classifications made. All models are made publicly available via Hugging Face.

2 Data basis and sources

The analyses are based on a systematically structured data set containing metadata on the sources and characteristics from text excerpts.

This study draws on corpora that have already been normalised and temporally and geographically referenced, and focuses on their classification. Below, we provide an overview of the data and the process by which it was generated.

The dataset comprises text extracts (quotes) relating to thunderstorms thunderstorm and hail events. This is based on the HSKLID data collection in Central Europe between 1000 and 1817 (Schätz and Glaser, 2025). It is based mainly

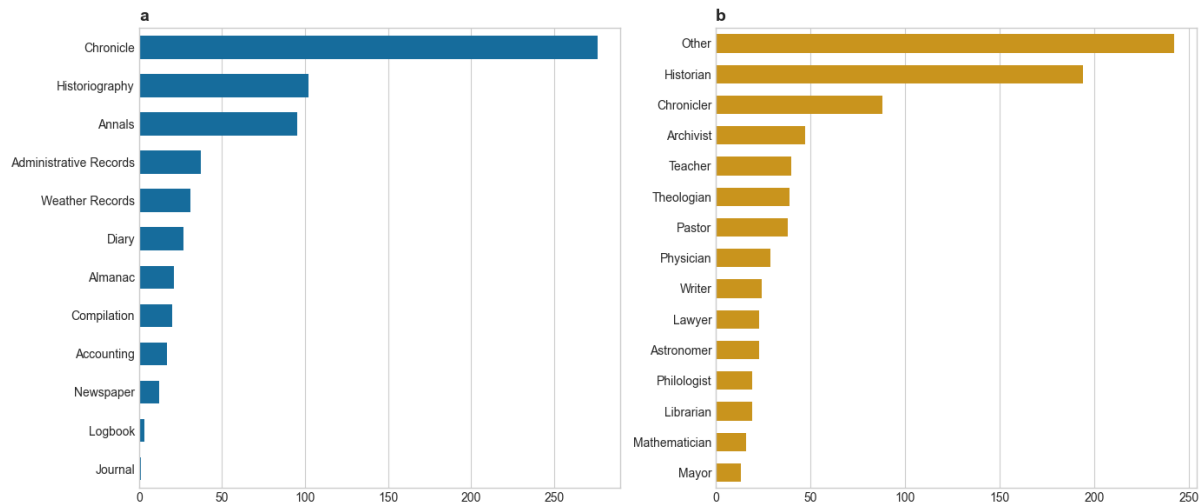


Figure 1. a Distribution of source types in the corpus: chronicles, historiographies, annals and administrative records predominate. b Authors' professional groups: chroniclers, historians, theologians, teachers and administrative officials. This reflects the profile of observers in the early systematic meteorological networks of the 18th and 19th centuries, which were primarily run by highly educated individuals (cf. Bayerische Akademie der Wissenschaften, 1789; Lamont, 1844; Preußisches Meteorologisches Institut, 1897).

on the HISKID2 dataset (Glaser, 2014)), which has been ~~integrated into~~ compiled since the 1980s on the virtual research ~~environment~~ platform [tambora.org](https://www.tambora.org)¹, ~~primarily for the reconstruction of temperature and precipitation time series.~~

Quotes referring to thunderstorm and hail events were extracted from this collection. The dataset used (version 2.2) goes
 95 beyond the original HISKID2 corpus and additionally includes linguistically normalised texts, standardised time data and
spatial georeferencing. The complete dataset is publicly available on GitLab². ~~The data set comprises 7002 entries with 6678~~
~~quotations-~~

In total, the dataset contains 6,999 quotes from 494 sources ~~for Central Europe in the period from 1000 to 1817. The~~
~~citations are mainly taken from chronicles, historiographies, annals and administrative records (see Figure 1). They document~~
 100 ~~6155 thunderstorm and 2011 hail events and are presented in tabular form. Each row represents a single citation with the~~
~~associated characteristics and event parameters). The characteristic values are collected and enriched in 21 steps from the~~
~~source references and citations,~~ documenting 6,157 thunderstorm events and 2,006 hail events. 99.1 % of the quotes come from
HISKID2 (483 out of 494 sources), with the remainder from eleven additional sources, primarily 18th-century newspapers
and a diary. The corpus thus represents a sample drawn from a general climate database and not a collection specifically
 105 compiled for extreme events. Of the 51,833 data records in HISKID2, 44,834 (86.5 %) mention neither thunderstorm nor hail
events, meaning that a selective choice favouring extreme events is structurally ruled out at the corpus level.

¹<https://www.tambora.org>

²https://gitlab.com/reservoirdog/hist_thunderstorm_hail_central_europe

In addition to bibliographic details, the source references also contain biographical information about the authors, including their place of work and professional activities. In addition to primary sources, secondary sources are also included, which means that there can sometimes be considerable time gaps between observation and publication. The bulk of the corpus consists of chronicles, historiographies, annals and administrative records, followed by weather records, almanacs and compilations (see Figure 2). The increasing number of sources and recorded observations reflects technological and social developments, such as the introduction of printing around 1450, the Reformation from 1517 onwards, the Thirty Years' War from 1618 to 1648, and the university reforms at the beginning of the 1a). In their role as observers of local environmental conditions, the authors show parallels with the early instrumental observers of the 18th and early 19th century.

centuries (see Figure 1b). This continuity points to comparable social conditions underpinning observational activity, although the nature of data collection differs fundamentally.

Since the text sources cover different stages of the German language (The quotes span four language stages: Middle High German, Early New High German, New High German, contemporary German), all quotations were manually normalised in terms of language.

and Contemporary German. Latin quotes were translated into Contemporary German prior to normalisation. When standardising the German-language quotes, grammatical modernisation was avoided; instead, lexical and orthographic normalisation was carried out. Idiomatic expressions were left in their original form in order to preserve semantic authenticity (Schätz, 2023). Obsolete or dialect-specific vocabulary was resolved using the Wörterbuchnetz³. This normalisation is a prerequisite for a consistent lexical basis, and therefore for applying and thus for the application of modern language models (Ehrmanntraut, 2025). The same applies to the creation of semantic word lists ('Silver Labels') and the coding of linguistic indicators, the quality of which depends directly on a homogeneous linguistic basis

Table 1 provides a quantitative overview of the distribution of documented events by century and language stage. The data illustrate the linguistic heterogeneity of the corpus, as well as the transition from Latin and Middle High German sources of the Middle Ages to Early and New High German sources from the 16th century onwards. The total number of events in the corpus increases in parallel with the number of textual sources. According to Ernst (2021), the transmission of textual sources was shaped by the invention of printing (1445), the Reformation (1517) and Humboldt's educational reform (1810). A comparison of the corpus with these historical events can be found in Figure 2.

The dates given in the textual sources, including those based on religious or regional calendars, were deciphered using Grotefend (Grotefend, 2007). The place names were georeferenced using GeoNames⁴ and historical place name directories (e.g. Eichler and Walther, 2001; Reitzenstein, 2009, 2013). The spatial distribution of the events is concentrated in German-speaking countries, although the location where a source originated from which a source originates does not necessarily correspond to the reported observation location (Figure ??). Newspapers in particular sites. In particular, newspapers report on events across in several regions, often beyond across territorial and linguistic boundaries. The As shown in Figure 3, the observations cover the most important landscapes climatically most relevant zones of Central Europe (??). This means that the

³<https://woerterbuchnetz.de>

⁴<https://geonames.org>

Table 1. Number of documented convective events by century and language stage. MHG = Middle High German, ENHG = Early New High German, NHG = New High German, CG = Contemporary German.

Century	Latin	MHG	ENHG	NHG	CG	Total
10.	1	0	0	0	0	1
11.	71	0	1	19	2	93
12.	146	0	0	32	2	180
13.	106	1	2	18	17	144
14.	30	1	35	43	10	119
15.	46	0	128	104	18	296
16.	242	0	972	662	43	1919
17.	24	0	851	1519	93	2487
18.	3	0	3	1589	143	1738
19.	0	0	0	0	22	22
Total	669	2	1992	3986	350	6999

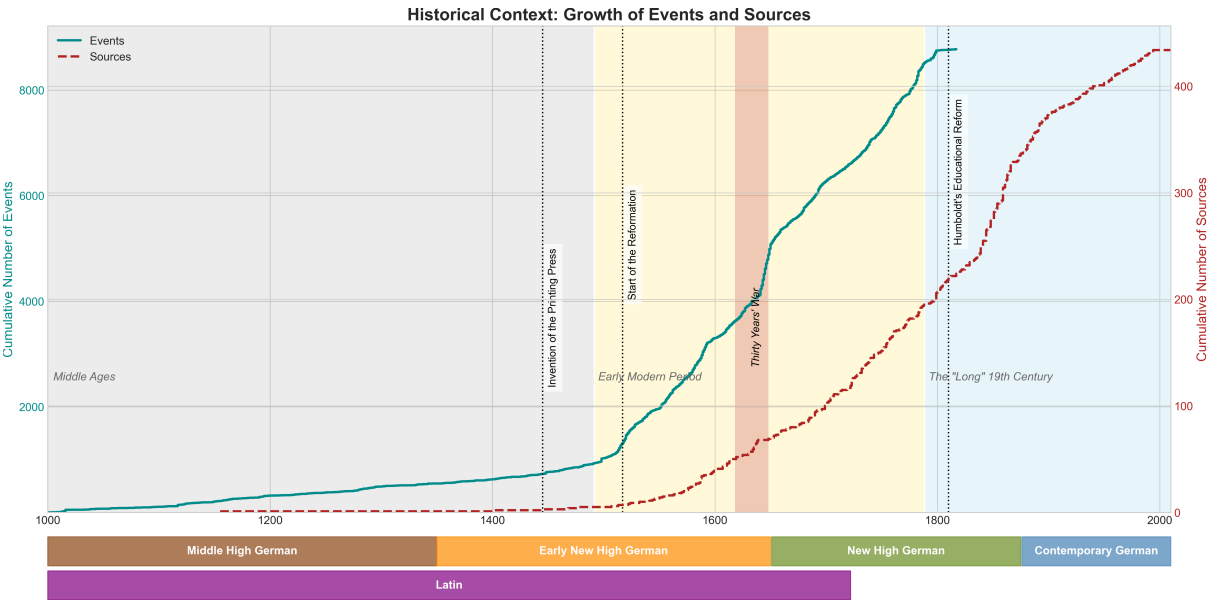


Figure 2. The cumulative number of events (left axis, solid line) and sources (right axis, dotted line) over time. The shaded areas indicate the Middle Ages, the Early Modern Period and the ‘long’ 19th century. Key historical events that have influenced the transmission of textual sources serve as points of reference. The coloured bar at the bottom shows the language stages of the corpus according to Ernst (2021). Latin continued to be used well into the early 18th century. A quantitative breakdown by century and language stage can be found in Table 1.

140 corresponding climate regions and mesoscale climatic effects such as maritime and continental influences, orography, etc., are represented according to the Köppen–Geiger classification (Beck et al., 2023). The dataset thus spans several climatic regimes, including the temperate oceanic climate (Cfb), the warm-summer continental climate (Dfb) and the Mediterranean-influenced climate (Csa, Csb), as well as their maritime, continental and orographic influences.

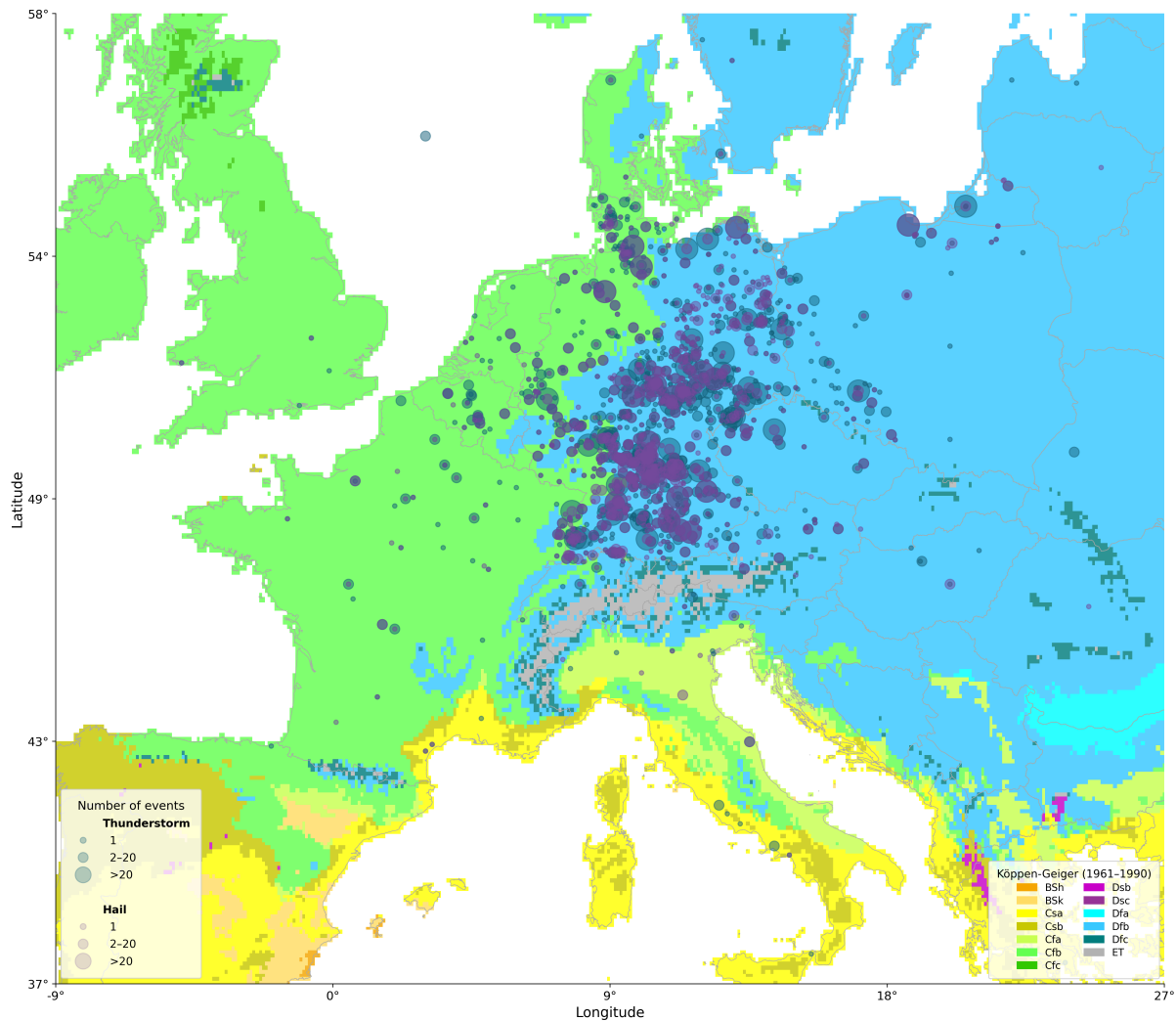


Figure 3. The spatial distribution of thunderstorm and hail frequency in Europe is based on the Köppen–Geiger climate classification system. The background shows the climate zones for the 1961–1990 reference period at a resolution of 0.1°, with each zone representing a specific temperature and precipitation regime (e.g. Cfb: temperate climate without a dry season and with a warm summer, Dfb: cold climate without a dry season and with a warm summer; ET: polar tundra climate). Köppen–Geiger climate classification data courtesy of (Beck et al., 2023) (<https://www.gloh2o.org/koppen>).

3 Methods

145 ~~This study builds~~ The classification method described in this study, including its validation, forms part of a comprehensive workflow for source evaluation (see Figure 4). Each level consists of formally defined steps, which are documented in detail in Schätz (2023).

The normalised quotes, together with the attributes “time” (T) and “location” (L), form the basis for the classification process, which is followed by a plausibility check. The quotes and the classified thunderstorm and hail events (EI) make up
150 the training set.

The underlying conceptual and formal model of the classification is set out in more detail below. It is based on the approach of outlined in Schätz (2023), which describes weather and ~~elimatic-climate~~ events using a ~~quadruple-consisting-of~~ quadruplet comprising event type (~~ET~~ET), time (~~TT~~T), location (~~PL~~L) and event intensity (~~EI~~), where ~~EI~~ is understood as a classification of textual sources ~~EI~~. In this study, ~~EI~~ corresponds to the classification of thunderstorm and hail events. Formally, ~~EI~~ ~~EI~~ is
155 the result of a mapping ~~f_{ET}~~, which maps a citation ~~z~~ $z \in Z$ from the set of all citations ~~Z~~ f_{EI} that maps a quote $q \in Q$ to a class $c \in C$ ~~c~~ C from a finite set of uniquely defined event intensity classes C :

$$f_{EI} : Q \rightarrow C, \quad EI = f_{EI}(q) \quad (1)$$

Classification is carried out in ~~a multi-stage process in which linguistic references to accompanying and subsequent several stages: linguistic indications of associated~~ phenomena are systematically analysed and recorded at the level of semantically
160 ~~coherent word groups recorded and analysed for their causal relationships~~. On this basis, ~~thunderstorm and hail events are divided the events are categorised~~ into intensity classes ~~and then statistically evaluated, including comparison with observation~~. Finally, the results are linguistically validated and their physical plausibility is statistically verified using observational data from the ~~Deutscher Wetterdienst~~ German Weather Service (DWD) ~~for the normal period~~.

3.1 Classification procedure

165 The classification ~~scheme-consists-of~~ procedure is based on a classification scheme comprising five mutually exclusive thunderstorm classes and four hail classes (see Tables A1 and A2). It is based on the classification and warning system of the Deutscher Wetterdienst (DWD) (2025) and the Tornado and Storm Research Organisation (TORRO) (2025).

The analysis of the ~~citations-quotes~~ takes into account the cause-and-effect ~~relationship as explicitly or implicitly formulated relationships set out~~ in the sources. ~~For the systematic re-recording of~~, which are referred to below as ‘impact pathways’.
170 ~~To systematically record the~~ observations, the ~~accompanying-associated~~ phenomena mentioned are ~~classified-categorised~~ according to their physical relevance into primary, secondary, tertiary and quaternary ~~characteristics-associated phenomena~~ (see Figure?? 5). This ~~classification-serves-the-purpose-of-structured-source-evaluation-categorisation~~ structures the analysis of the sources and forms the basis for the subsequent classification.

Classification is carried out in stages. First, ~~it is checked whether the quotation mentions accompanying a check is made to~~
175 ~~see whether the quote describes associated~~ phenomena that can be clearly attributed to a thunderstorm or ~~hailstorm~~ hail event

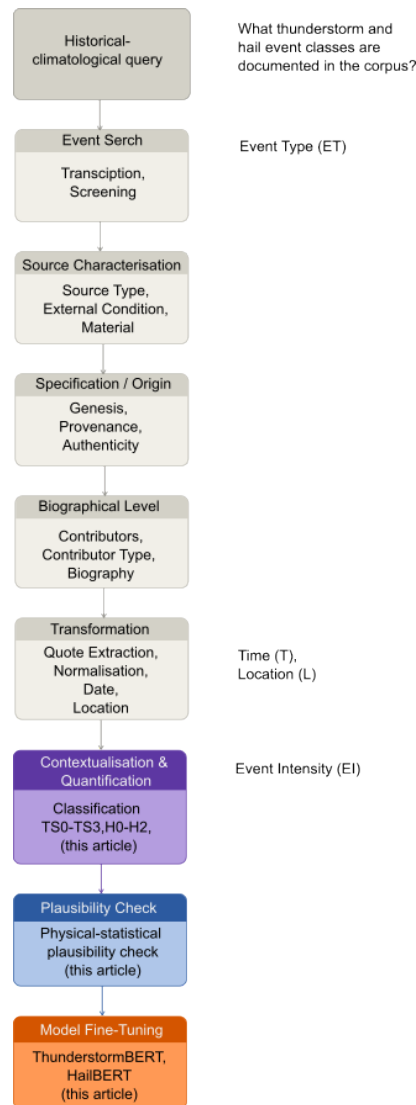


Figure 4. A schematic overview of the workflow for source analysis based on Schätz (2023). Each level comprises several formally defined steps. The processes described in this article are colour-coded. The attributes generated at each level are shown on the right (ET = event type, T = time, L = location, EI = event intensity).

(linguistic indicators: Table B1). Only if such indications are present ~~is the event considered further. The next step is to classify the intensity of the event based on the~~ are these associated phenomena taken into account and classified further. To this end, their intensity is determined on the basis of the criteria defined in ~~tables A3 – A5 . The effects described are systematically classified and explicitly included in the assessment.~~

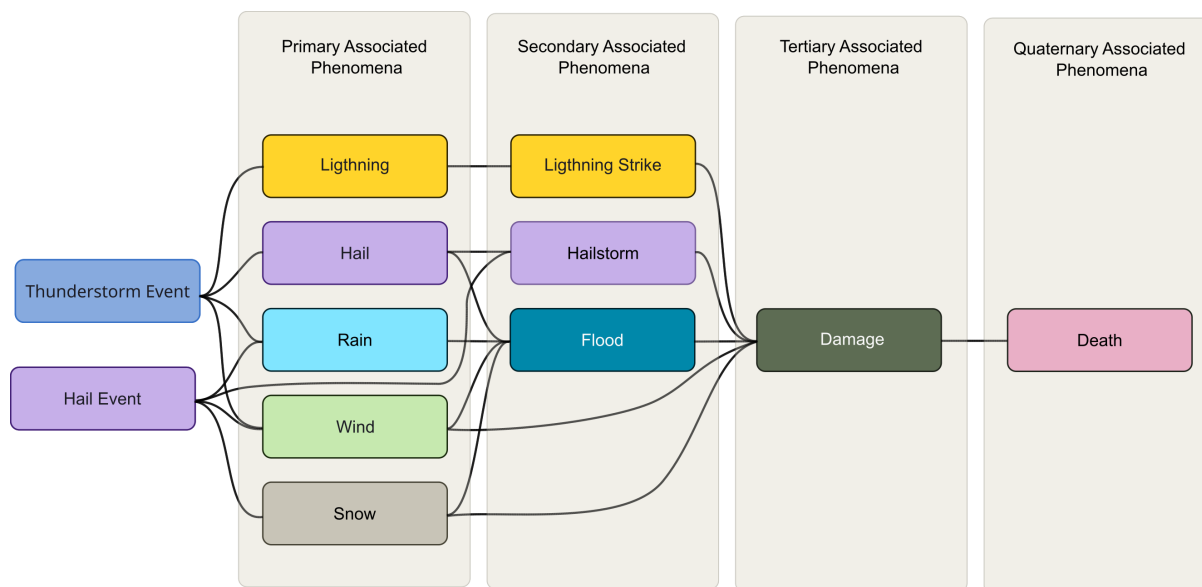


Figure 5. Impact pathways of thunderstorm and hail events. The diagram illustrates the cause-and-effect relationship between thunderstorm and hail events (left) and their associated phenomena (right). Primary associated phenomena include rain, hail, lightning, snow and wind. These can trigger secondary effects such as flooding, hail damage or lightning strikes. These, in turn, can lead to tertiary effects in the form of damage. In severe cases, these can trigger quaternary effects, which may include fatalities. The classification scheme thus reflects a causal chain of meteorological processes and their ecological and societal impacts.

180 ~~The classification begins with the highest intensity class (class 2 for hail, class 3 for all other phenomena considered) in order to avoid underestimating particularly severe events. If this class is not met, the lowest intensity class (class 1) is examined and then, if applicable, the middle class. The highest intensity class is assigned exclusively on the basis of explicit damage descriptions. For Tables A3 to A5 (rain, snow, wind) and in Table A6 (hailstone size). The damage described is also classified and systematically assigned to the respective associated phenomena along the impact pathways. In the case of hail events,~~

185 ~~qualitative information on grain hailstone size is also taken into account (see Table A6). The low intensity class is characterised by the fact that the event is not described in further detail. Intermediate classes are assigned when the criteria for the extreme classes are not met. Table 2 shows five classified examples with Early New High German and New High German quotes.~~

3.2 Validation through linguistic evidence

190 ~~To ensure consistent and comprehensible classification of thunderstorm and hail events, class-specific vocabularies are derived from the examined quotations. These contain typical linguistic indicators of the respective accompanying phenomena and are based on frequently used word forms and phrases in the historical sources (e.g. for thunderstorms: “flash”, “thunder”, “storm”). The vocabularies were developed inductively from the source material and checked iteratively. An overview is provided in Table B1 in the appendix.~~

Table 2. Five examples illustrating the classification of thunderstorm and hail events. The columns show the original historical text, the normalised German form, the English translation, the language stage, the rationale for the classification and the assigned intensity class. A detailed explanation of the word-level classification procedure can be found in Figure 6.

Original	Normalised	Translation	Language stage	Explanation	Class
<i>Diese nacht hatts zu Hanauw einen gewaltigen sturmwindt gehabt, darauf umb 12 Uhr ein donnerwetter mitt schloößen gefolget, welches großen schaden an fenstern, hatt sich auch ein feurig meteorom über dem schloßthurm daselbst sehen lassen.</i>	Diese Nacht hat es zu Hanau einen gewaltigen Sturmwind gehabt, darauf um 12 Uhr ein Donnerwetter mit Schloößen gefolgt, welches großen Schaden an Fenstern, hat sich auch ein feuriger Meteor über dem Schlossturm daselbst sehen lassen.	This night there was a violent storm wind in Hanau, followed at 12 o'clock by a thunderstorm with hailstones, which caused great damage to windows; a fiery meteor was also seen over the castle tower.	ENHG	Wind force 3 (gale) and hailstorms causing window damage (C2). These two criteria are sufficient for TS3. The “fiery meteor” shows how atmospheric phenomena were understood in a particular era.	TS3
<i>Wesel, vom 7. Aug. Diesen Abend umb 7. Uhr entstund allhier ein so schweres Ungewitter, dergleichen bey Menschen Gedencken nicht geschehen [...] Es hat drittelhalb Stundenlang so starck geblitzet [...] Dieses Gewitter endigte sich mit einem sehr starcken Hagel und Regen.</i>	Wesel, vom 7. August. Diesen Abend um 7 Uhr entstand allhier ein so schweres Ungewitter, dergleichen bei Menschen Gedenken nicht geschehen [...] Es hat dreieinhalb Stunden lang so stark geblitzt [...] Dieses Gewitter endigte sich mit einem sehr starken Hagel und Regen.	Wesel, 7 August. This evening at 7 o'clock there arose a thunderstorm the like of which had not been seen in living memory [...] For three and a half hours the lightning flashed as if the sky itself were on fire [...] The storm ended with very heavy hail and rain.	NHG	Intense lightning, as well as heavy hail and rain, has been recorded without causing any damage. While the TS1 threshold has been exceeded, the TS3 criteria have not been met.	TS2
<i>Am 28sten blitzte es des Abends.[...]</i>	Am 28. blitzte es des abends.	On the 28th, there was a flash of lightning that evening.	NHG	A simple thunderstorm, with no further details.	TS1
<i>1772, den 27. Juni kam gegen Abend ein entsetzliches Donnerwetter [...] Die Kissel schlossen von durchgehengs wie Hühnereier, ja viele so dick wie eine Mannsfaust, daß sie die Layen auf den Dächern, die Fenster in Kirchen und Häusern alle zerschlagen, rundherum in 38 Dörfern, Felder und Weinberge völlig verwüstet.</i>	1772, den 27. Juni kam gegen Abend ein entsetzliches Donnerwetter [...] Die Kieselgeschloßen von durchgehend wie Hühnereier, ja viele so dick wie eine Männerfaust, dass sie die Ziegel auf den Dächern, die Fenster in Kirchen und Häusern alle zerschlagen, rundherum in 38 Dörfern, Felder und Weinberge völlig verwüstet.	On 27 June 1772 towards evening a terrible thunderstorm [...] The hailstones were consistently the size of hen's eggs, many as thick as a man's fist, shattering roof tiles and windows in churches and houses, devastating fields and vineyards in 38 villages.	NHG	Direct measurements of the size of the hailstones are available: chicken's egg (C1) and fist (C1). This has caused damage to buildings, fields and vineyards in 38 villages (C2). Hailstorm H2 meets both criteria.	H2
<i>Gewitter zogen auf und zertheilten sich wieder [...] Nun wechselten Gewitter, Regen, Stürme und Donner mit Sonnenschein ab, wobei es warm und fruchtbar war. Gewitterwolken zogen umher, und ein paarmal kam das Gewitter zum Ausbruch.</i>	Gewitter zogen auf und zerteilten sich wieder [...] Nun wechselten Gewitter, Regen, Stürme und Donner mit Sonnenschein ab, wobei es warm und fruchtbar war. Gewitterwolken zogen umher, und ein paarmal kam das Gewitter zum Ausbruch.	Thunderstorms gathered and dispersed again [...] Now thunderstorms, rain, storms and thunder alternated with sunshine, while it was warm and fertile. Thunderclouds moved around, and a few times the storm broke out.	NHG	This is a general description of the weather over an extended period. Several events are mentioned, but not described in detail. No specific effects or damage resulting from any individual event are documented.	—99

~~In addition to~~ For historical weather observations, there is no independent benchmark against which the classification could be directly verified. In order to be able to assess the quotes critically, the evidence presented in the text is evaluated.

To this end, individual words or phrases are assigned to the classification of content, the strength of evidence of the linguistic descriptions is recorded in order to transparently evaluate the significance of the underlying text sources previously defined evidence classes C1, C2 and C3 (see Table 3). ~~A distinction is made between three classes of evidence. Evidence class C1 (direct and specific) has the highest reliability, as it provides concrete information, such as /measurable/ comprises data containing specific and measurable information, for example on hailstone size, wind strength or precipitation levels, which correspond to measurable variables. These data have the highest level of evidence, as they are most comparable to instrumental measurements. C2 (indirect) includes descriptions in which damage indicators such as hail, wind or rain damage clearly point to an event/damage) comprises damage indicators that plausibly and unambiguously point to a thunderstorm or hail event, or its intensity. Examples include destroyed crops, damaged buildings or uprooted trees. C3 (relative/qualitative) has the lowest reliability, as purely qualitative adjectives (e.g. level of evidence. It comprises non-specific qualitative descriptions without explicit reference values, such as ‘strong’, ‘violent’, ‘violent’ or ‘very large/terrible’) without reference values allow for a wide range of interpretations.~~

Table 3. Examples of silver labels for the phenomenon groups hail, rain and wind. Each expression is assigned to a phenomenon category and an evidence class (C1–C3), which reflects the reliability of the underlying linguistic evidence. C1 entries are based on concrete, measurable information, C2 entries are based on indirect damage indicators, and C3 entries are based on qualitative descriptors only. The complete silver label lists, including all recorded variants, can be found in Appendices B2 – B4

Phenomenon	Category	Representative Examples	Class
Hail	Size descriptions	Hazelnut-sized, walnut-sized, hen’s egg-sized	C1
	Hail impact	Crops beaten down, vineyards destroyed	C2
	Intensity	Terrible, severe hailstorm	C3
Rain	Precipitation height	Water was foot-deep, one cubit high	C1
	Damage	Bridges broken, cellars flooded	C2
	Intensity	Flood, inundation, deluge, washed away	C3
Wind	Speed	Only indirect descriptions (see Damage/Intensity)	C1
	Damage	Trees uprooted, roofs blown off	C2
	Intensity	Strong wind, hurricane-like, gusts	C3

~~This evidence-based differentiation supplements the classification with an internal, source-critical assessment of linguistic significance. This increases the traceability of the classifications and explicitly takes into account the interpretative nature of historical text sources~~All words and phrases describing associated phenomena are extracted from the quotes. Each expression

is assigned to one of the three evidence classes. This results in three lists, known as “silver labels”, which serve as linguistic indicators for the individual classes (see Table 3). A complete list of the silver labels can be found in the Appendix B2–B4.

Figure 6 illustrates the classification process using a historical quote referring to a thunderstorm in Hanau. The example shows how the classification is derived from the text and how the silver labels are assigned to the evidence classes.

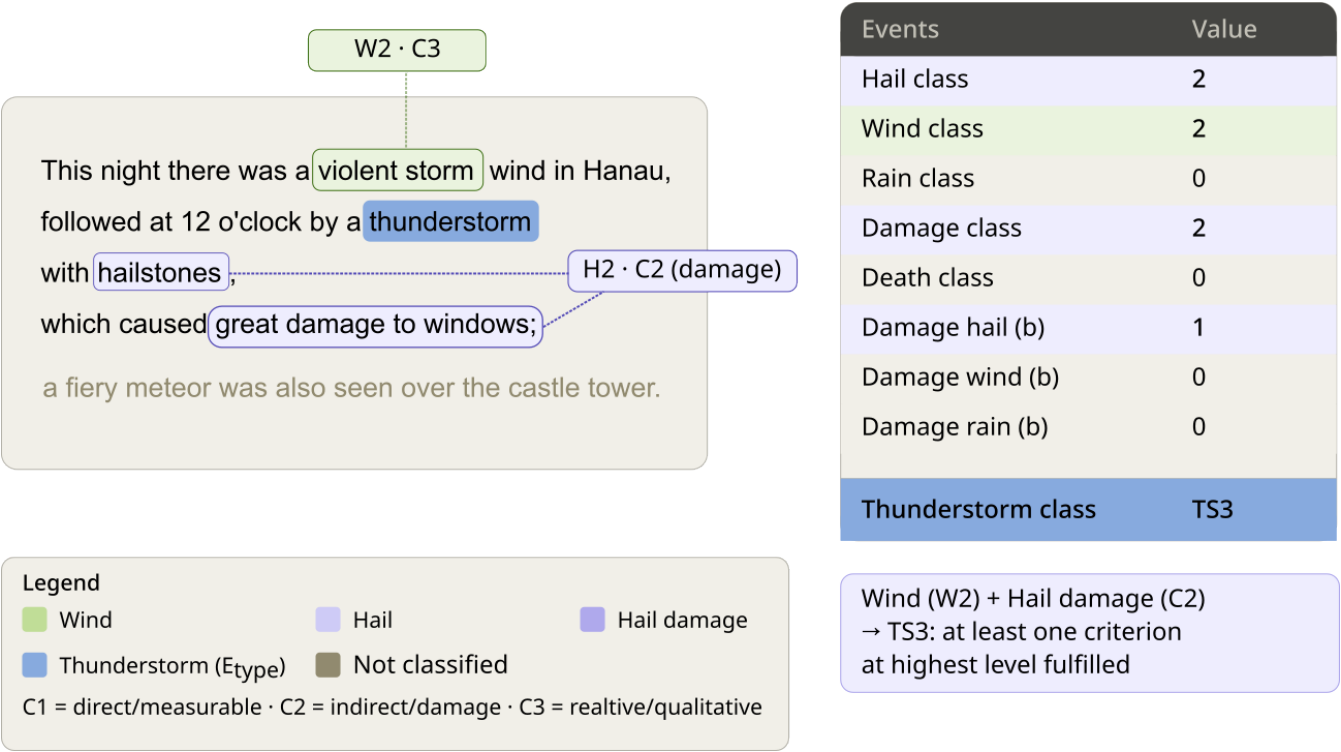


Figure 6. Annotated classification of a thunderstorm event in Hanau. The associated phenomena are colour-coded: green = wind, purple = hail, blue = thunderstorm (event type), brown = unclassified. The evidence classes are specified for each annotation (C1 = direct/measurable, C2 = indirect/damage, C3 = relative/qualitative). On the right-hand side, the individual events and the hail event are coded as follows: events marked with (b) are binary (0 = absent, 1 = present). The final thunderstorm class (TS3) is derived from the wind class (W2, C3) and the hail event (H2, C2). Further text examples can be found in Table 2.

215 From these evidence classes, a four-level confidence index is derived for each quote as a measure of the uncertainty in interpretation. This is based on the number of robust evidence classes (C1 and C2) identified for each associated phenomenon along the impact pathways. If at least two such evidence classes are present, or if a single associated phenomenon is accompanied by classes C1 and C2, level 3 (“multiple evidence”) is assigned. A single evidence class C1 or C2 results in level 2 (“confirmed”). If only an evidence class C3 is present, level 1 (“qualitative”) is assigned. If no applicable evidence

220 class can be assigned, level 0 (“no silver label”) is assigned. Table 4 summarises the confidence index and its derivation.

Table 4. Four-level index for assessing the confidence of the classification.

Index level	Designation	Criterion
0	no silver label	no assignable evidence class
1	qualitative	exclusively evidence class C3
2	confirmed	exactly one evidence class C1 or C2
3	multiple evidence	at least two evidence classes C1/C2, either via multiple associated phenomena or via a single C1 associated phenomenon with a complete impact pathway

3.3 Validation by statistical-physical evidence

3.3.1 ~~Annual~~ Seasonal distribution of events

To ~~verify~~ assess the quality of the ~~historical climatological data sets~~ classified thunderstorm and hail events, we examine whether the ~~historical event corpus exhibits~~ data exhibit a physically plausible ~~seasonal signal that is consistent with~~ signal that
225 corresponds to the known seasonality of convective events(~~summer maximum~~). This analysis serves to ~~qualitatively~~ validate the corpus with regard to its suitability as a training ~~basis for language model-based dataset for Transformer-based~~ classification tasks. ~~Since language models learn statistical regularities from the~~ As Transformer-based language models adopt the statistical patterns of their training data, ~~the presence of a source-related biases in the annotations would be directly incorporated into the model.~~ A realistic seasonal pattern is a key prerequisite for ensuring that the model reflects real physical relationships and
230 ~~not primarily transmission artefacts~~ therefore demonstrates that the annotated classes capture real physical events and that the corpus provides a robust training basis.

The analysis is based on monthly aggregated count data ~~of~~ for historical thunderstorm and hail events ~~for covering~~ the period 1000–1817. ~~Since historical~~ Historical sources predominantly document only positive events ~~and since,~~ meaning that there are no systematic zero observations ~~are available, a direct reconstruction of.~~ As the observation density is ~~not possible.~~
235 ~~To reduce distortions caused by different source densities, the annual progression is normalised. To do this therefore unknown, the analysis is based on relative monthly distributions. To this end,~~ the monthly event ~~numbers~~ counts for each year are normalised to an annual total of 1,

$$p_{m,y} = \frac{E_{m,y}}{\sum_{j=1}^{12} E_{j,y}}, \quad (2)$$

so that each year is described as a distribution ~~over~~ across the calendar months.

240 The ~~final annual curve is derived from the aggregation of the monthly shares, standardised on an annual basis, across all years. In principle, the median~~ median generally provides a robust description of the typical seasonal pattern, as it ~~limits the influence of individual years with a wealth of historical data~~ is insensitive to outliers. However, in the case of highly incomplete

time series with predominantly positive event reports, ~~the median leads to it is distorted, as~~ months with rarely documented events ~~being systematically underrepresented and often have zero values. This leads to a systematic downward bias and thus~~
 245 ~~to an artificial flattening of the seasonal pattern appearing artificially thinned out.~~

~~For this reason, the mean value.~~ The mean of the normalised monthly proportions is ~~therefore~~ used as the central estimator for reconstructing the seasonal ~~annual-cycle~~ pattern over the year,

$$\bar{p}_m = \text{mean}_y(p_{m,y}), \quad (3)$$

as it consistently ~~reflects~~ represents the relative frequency of events across all years and ~~maintains~~ preserves stable seasonal
 250 structures even ~~with fragmentary data when data are incomplete~~. Due to the ~~upstream~~ prior annual normalisation, all years contribute equally to the ~~analysis~~ estimate, so that the mean ~~value~~ is not dominated by years with a high ~~event density~~ density of events, but primarily reflects the shape of the seasonal pattern.

The uncertainty of the reconstructed ~~annual~~ seasonal cycle is quantified over the years using bootstrap resampling (e.g. Wilks, 2019). ~~95% confidence~~ %confidence intervals are derived from the resulting distributions. Stability is ~~additionally~~
 255 ~~also~~ tested for a time window with ~~high source density~~ a high density of sources (1624–1654), which, due to its comparatively ~~complete transmission~~ dense record, allows for independent validation of the seasonal structure.

For external validation, the reconstructed ~~historical annual~~ seasonal cycle is compared with ~~modern observation data from the DWD~~ DWD observational data for the normal periods 1961–1990 and 1991–2020. The ~~1961–1990 normal period is based on visual and auditory observations in accordance with the DWD Observer’s Handbook (Deutscher Wetterdienst, 2014).~~
 260 ~~The~~ monthly totals of convective events serve as a reference. These are normalised ~~on an annual basis after~~ annually following spatial aggregation, ~~thus ensuring thereby ensuring the~~ direct comparability of the relative monthly proportions. The agreement of the seasonal patterns is quantified using Spearman’s rank correlation (ρ) and the root mean square error (RMSE) ~~of the normalised monthly proportions~~.

In addition ~~to the standardised annual progression, a summer–winter, the summer–winter~~ ratio is calculated ~~for each year.~~
 265 ~~This is based on the on the basis of the~~ monthly aggregated event counts ~~without further normalisation. For each year, the sum of events in figures. To this end, the total number of events for~~ the summer months of June to August (JJA) and ~~in the~~ winter months of December to February (DJF) is ~~calculated, with December being determined for each year. December is~~ assigned to the winter of the following year. ~~The summer–winter~~ Years with no documented winter events are excluded. The ~~summer–winter~~ ratio is calculated as the quotient of ~~the annual summer and winter event sums. Years without documented~~
 270 ~~winter events are excluded these totals.~~ To improve ~~the comparability of the time series and reduce the influence of extreme~~ quotients, ~~the summer–winter ratio is additionally~~ comparability and stabilise the variance, the quotient is logarithmised. The resulting ~~logarithmised~~ time series is ~~smoothed by a moving 10-year average then smoothed using a 20-year moving average~~ (e.g. Wilks, 2019).

3.3.2 Class-specific annual distribution

275 ~~The~~ In order to assess the stability of class-specific seasonal patterns, the seasonal distribution of thunderstorm (~~TS1–TS3~~ events (TS1–TS3) and hail (~~H1–H2~~) events is analyzed to assess the stability of class-specific seasonal patterns events (H1–H2) is analysed. The aim ~~of this analysis is to examine whether a consistent seasonal hierarchy emerges for the individual classes is to determine whether the classes exhibit a consistent ordering relation~~ and whether this ~~order remains stable over the course of~~ remains stable throughout the year.

280 Only events lasting a maximum of 31 days are taken into account. Each event is assigned to all calendar months that it affects (“presence-per-month”) Events are assigned to the calendar months in which they take place. If an event ~~exceeds a monthly~~ limit spans the end of one month and the start of the next (e.g. from 31 July into the following day), it is counted once in each affected month. Events of very long duration that can no longer be clearly assigned to a specific season are excluded of the months concerned. Events with unspecified dates or those lasting a very long time (e.g. ‘in the summer’) are not taken into account.

285 For each year y , each month m and each class c , ~~the events assigned in this way c , the associated events~~ are aggregated by class. The analysis is based on ~~relative proportions within a month~~ monthly relative proportions, so that only the class distribution is considered and no assumptions ~~need to be made~~ about absolute event frequencies or observation ~~density are required~~ densities. The monthly class proportion is defined as

$$290 \quad p_{c,m,y} = \frac{E_{c,m,y}}{\sum_{c'} E_{c',m,y}}, \quad (4)$$

where $(E_{c,m,y})$ denotes the number of events ~~of class in class (c in month m and year y)~~ in month (m and year y). The summation index ~~c' runs over c covers~~ all classes of the respective event type considered (e.g. ~~TS1–TS3 or H1–H2~~ (TS1–TS3, H1–H2)). Thus, $p_{c,m,y}$ describes the proportion of a class ~~in all events reported in a given out of all events observed in a specific~~ month of a given year.

295 In order to limit the influence of individual years with ~~high data availability~~ a high density of events, the monthly class shares are first determined on an annual basis and then aggregated ~~equally across all years~~. The typical across the years as an unweighted mean. The seasonal class structure is estimated as the mean ~~value~~ of the annual class ~~proportions~~ shares,

$$\bar{p}_{c,m} = \text{mean}_y (p_{c,m,y}), \quad (5)$$

300 where ~~the upstream~~ normalisation ensures that each year ~~contributes equally to the estimate is included in the estimate with~~ the same weighting, regardless of its event ~~density~~ frequency.

The statistical uncertainty of the reconstructed class distributions is quantified over the years using bootstrap resampling. ~~For each bootstrap sample~~ To this end, the monthly class distribution is recalculated ~~; mean for each bootstrap sample~~. Mean values and 95% confidence intervals (2.5th % and 97.5th percentiles % quantiles) are determined from the resulting distributions.

To assess the stability of the seasonal ~~class hierarchy~~ ordering relation, the proportion of bootstrap samples in which the expected ~~class order is fulfilled~~ ordering relation is satisfied is also determined for each month. For thunderstorm events, the

~~strict-order~~ordering relation

$$TS1 > TS2 > TS3 \quad (6)$$

is used~~, and~~. Similarly, for hail events, the ordering relation

$$H1 > H2 \quad (7)$$

310 is applied. The resulting proportion (~~strict support~~) describes the robustness of the complete ~~class hierarchy~~ordering relation with respect to sample ~~variations~~variability.

In addition, a weighted support ordering index is calculated for the thunderstorm events, which ~~distinguishes between fully and partially fulfilled class order~~. A complete order ($TS1 > TS2 > TS3$) is weighted higher than a partial order, in which only the dominance of the weakest over the stronger classes is fulfilled ($(TS1 > TS2)$ or $(TS2 > TS3)$). Weighted support
315 ~~thus allows for a more refined~~evaluates the two sub-relations of the ordering relation separately: the relation $TS1 > TS2$ is weighted by 0.7 and the relation $TS2 > TS3$ by 0.3. A fully satisfied ordering relation ($TS1 > TS2 > TS3$) thus yields a value of 1, whilst partially satisfied orders yield values of 0.7 and 0.3 respectively. The index is averaged across the bootstrap samples and enables a nuanced assessment of the seasonal class structure, ~~especially~~particularly in months with low event density or sparsely populated classes.

320 3.4 Influence of source type on classification

~~Potential selection or transmission~~Any potential selection or reporting biases are tested for independence using a chi-square test
~~, which examines chi-squared test~~ (e.g. Wilks, 2019). This involves analysing the relationship between source type and ~~storm~~
~~class ($TS1-TS3$)~~thunderstorm class ($TS1-TS3$) or hail class ($H1-H2$). This $H1-H2$). The analysis is based on a contingency
325 of source type and class,~~thereby taking appropriate account of multiple~~. Multiple mentions of individual sources are taken into account accordingly.

Under the null hypothesis (H_0), it is assumed that source type and classification are independent of ~~each other~~one another. In addition to the ~~chi-square statistic (χ^2)~~, the effect size is calculated according to chi-squared test statistic (χ^2). Cramér's ~~V~~
~~in order's V~~ is calculated as a measure of effect to quantify the strength of the ~~correlation~~. ~~Standardised residuals are also~~
330 ~~evaluated in order to identify association~~. In addition, the standardised residuals are analysed to identify the combinations of source type and class that contribute most strongly significantly to the deviation from independence.

3.5 Fine-tuning the ~~large~~transformer-based language model

The ~~validated data sets~~classified and validated datasets form the basis for fine-tuning the Transformer-based language model. The pre-trained language model mDeBERTa V3 Base, which is based on multilingual text data and ~~has~~exhibits a high degree
335 of context sensitivity, is used to classify descriptions of thunderstorm and hail ~~descriptions~~events. The models are adapted to the classification task (~~$TS0-TS3$ and $H0-H2$~~) by means of $TS0-TS3$ and $H0-H2$ through fine-tuning. The ~~data set is divided~~

~~dataset is split~~ into training, validation and test ~~portions-sets~~ in a ratio of 70:10:20, with the ~~division-being-stratified-in-order-to~~ ~~maintain-the-class-ratios~~split being stratified to preserve the class proportions.

A two-stage hyperparameter search is ~~performed-carried out~~ to identify suitable training parameters. In ~~the-phase~~ Phase A, an
340 exploratory random search strategy is applied. ~~For this purpose~~To this end, a combinatorial search space is defined, consisting
of learning rates ~~$1 \times 10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}$~~ $\{1 \times 10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}\}$, maximum sequence length ~~128, 256~~ $\{128, 256\}$,
label smoothing factor ~~0.0, 0.05~~ $\{0.0, 0.05\}$ and learning rate ~~scheduler-linear, cosine~~ schedulers $\{linear, cosine\}$, resulting in a
total of 24 possible hyperparameter combinations.

Twelve combinations are randomly selected from this search space and trained ~~with-using~~ a fixed seed. This sample corres-
345 ponds to half of the entire ~~hyperspace-search space~~ and serves as an efficient exploratory coverage without ~~having-the need to~~
fully evaluate all possible configurations. For all training runs, a batch size of ~~16~~ 16, a weight decay of ~~0.01~~ 0.01, a warm-up
rate of ~~0.06~~ 0.06 and a training duration of five epochs are used. Training is ~~performed-with-mixed-precision-carried out using~~
mixed-precision (fp16) on GPU hardware. The models are evaluated after each epoch using the validation data, with the best
model ~~of-a-configuration-being-for each configuration~~ selected based on the F1 score.

350 In ~~the-phase~~ Phase B, the three ~~hyperparameter-configurations-with-the-best-ratings-in~~ highest-performing hyperparameter
configurations from Phase A are selected and retrained ~~with-using~~ three different random seeds ~~.-The aim of this phase is~~
~~to-check-for each. This allows~~ the stability and reproducibility ~~of-the-results-against-random-fluctuations-to be quantified in~~
relation to stochastic influences in the training process. For each configuration, the mean ~~value~~ and standard deviation of the
F1 score are calculated across the three ~~training~~ runs.

355 ~~The models from phase B are finally~~ Finally, the models from Phase B are evaluated on the ~~basis-of-the-F1-score-and~~
~~accuracy-on-the~~ validation and test ~~data-set~~ datasets using F1 scores and accuracy. For qualitative analysis, a confusion matrix
is also used to ~~reveal-systematic-misclassifications-between-the~~ identify systematic misclassifications of thunderstorm and hail
~~intensity-classes~~ events.

4 Results

360 4.1 ~~Annual-distribution~~ Seasonal Cycle

The ~~aggregated-historical-records-of-thunderstorms~~ normalised monthly proportions of thunderstorm and hail events show a
pronounced seasonal cycle ~~in-the-normalised-monthly-proportions~~ (Figure ~~(Figure~~ 7). For both types of ~~events, the highest~~
~~proportions-occur-in-event, the peak occurs during~~ the summer months of June to August (JJA). The ~~increase-in-monthly~~
~~proportions-begins-in-spring, reaches-its-maximum-in-summer-and-decreases-continuously-in-the-autumnmonths~~ annual pattern
365 is characterised by an increase in spring, a summer peak and a steady decline in autumn.

During the winter months (~~DJF~~ December, January, February), the ~~standardised-monthly-proportions-remain~~ proportions are
low but consistently ~~show-positive~~ values positive. The transitional seasons ~~are-characterised-by-lower-proportions-compared-to~~
show reduced proportions compared with the summer half-year.

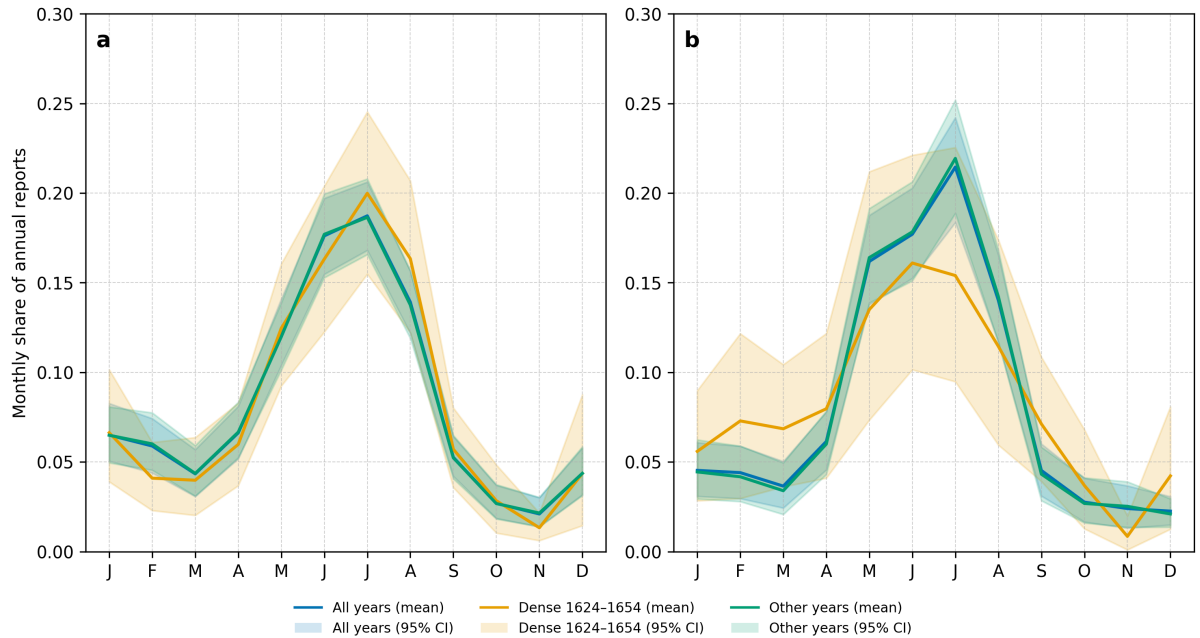


Figure 7. Seasonal cycle of thunderstorm and hail events. Monthly values represent the mean fraction of annual activity (annual total = 1).

The annual ~~cycle calculated for the most densely documented pattern for the~~ period (1624–1654) ~~follows the same basic~~
 370 ~~structure as the annual cycle for~~ corresponds structurally to that of the entire period. The ~~form of the monthly distribution~~
~~remains the same, while the relative amplitude between the summer and winter months differs from that of the entire period.~~
~~A comparison with the other years shows distribution pattern remains unchanged, whilst differences in seasonal amplitude~~
~~(summer–winter contrast) are evident. Overall, this indicates~~ a high degree of ~~consistency in the monthly proportions~~
~~temporal stability in the relative monthly distribution.~~

375 ~~For hail events, the highest standardised monthly proportions are also concentrated in~~ Hail events also peak during the sum-
 mer months. Compared ~~to with~~ thunderstorm events, ~~increased proportions already occur in the late spring months, especially~~
~~in however, the seasonal increase begins earlier, with higher frequencies already evident in late spring (particularly~~ March and
 April).

A comparison with ~~observation data from the DWD for the~~ reference DWD observational data for the normal periods
 380 1961–1990 and 1991–2020 (~~Figure 8) shows see Figure 8) reveals~~ a correlation between the seasonal patterns of normalised
 monthly thunderstorm frequencies. ~~The~~ For the entire study period, the Spearman rank correlations between ~~the historical data~~
~~(1624–1654) and the DWD references are between $p = 0.78$ and $p = 0.89$ for the combinations of time~~ thunderstorm events and
~~the DWD reference data range from 0.66 to 0.78 ($RMSE \approx 0.05$), depending on the normal period and search radius~~ considered.
~~The mean square errors are $RMSE \approx 0.03$ to 0.04 .~~ For the period 1624–1654, the correlations are higher, ranging from 0.74
 385 to 0.86, whilst the deviations are smaller ($RMSE \approx 0.03$ to 0.04). For hail events, the correlations range from 0.78 to 0.93 for

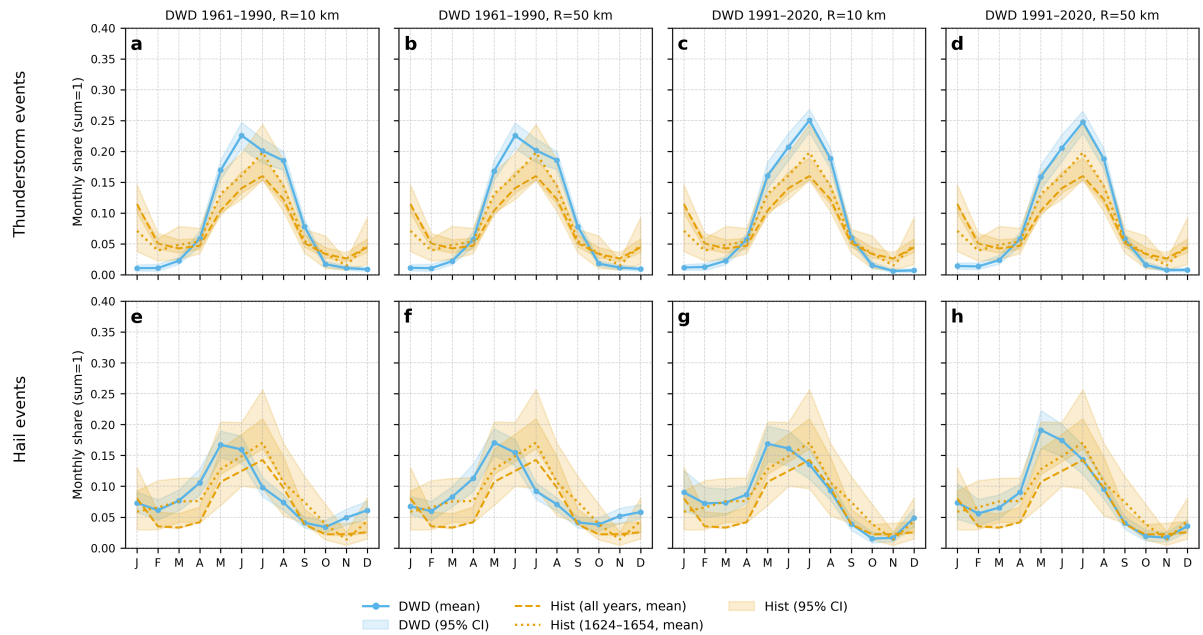


Figure 8. Seasonal cycle of normalised monthly frequencies of thunderstorm and hail events from the study, compared with DWD observations for the normal periods 1961–1990 and 1991–2020. The DWD stations are selected based on their distance from the historical event locations (radius of 10 km or 50 km). All values are normalised to the respective annual total, so that the relative monthly proportions are directly comparable.

the entire period (RMSE $\approx$ 0.03 to 0.04) and from 0.83 to 0.92 for the period 1624–1654 (RMSE $\approx$ 0.02 to 0.03). The results differ only slightly between the selection of DWD stations within a radius of 10 km and 50 km around the historical event locations. The key figures are comparable for both reference periods. Compared with the normal period 1991–2020, the correlations are consistently higher than when compared with 1961–1990.

390 The logarithmic summer–winter–summer–winter ratio (JJA/DJF) shows predominantly positive values for thunderstorm events over the period under investigation (Figure 9), indicating a dominance of the summer months over the winter months, thus indicating that activity is predominantly concentrated in the summer. The smoothed time series (40-year average) lies 20-year average) remains above the zero line for over long periods and mostly ranges generally fluctuates between approximately 1.0 and 2.0. Elevated values occur particularly in the late 16th and early 17th centuries, while phases with reduced summer dominance whilst phases of reduced summer activity are evident in the mid-16th century and early 17th century. Negative individual years occur sporadically, but centuries. Individual years are significantly smoothed out by the averaging process and have only a minor impact on the smoothed curve influence on the long-term trend.

For hail events, the seasonal signal is weaker overall. The logarithmic summer–winter–summer–winter ratio shows greater dispersion variation and, in the smoothed curve, lies predominantly in the range between approximately 0.3 and 1.0. Values close to or below zero occur at times, especially zero or below occur from time to time, particularly in the early phase of the

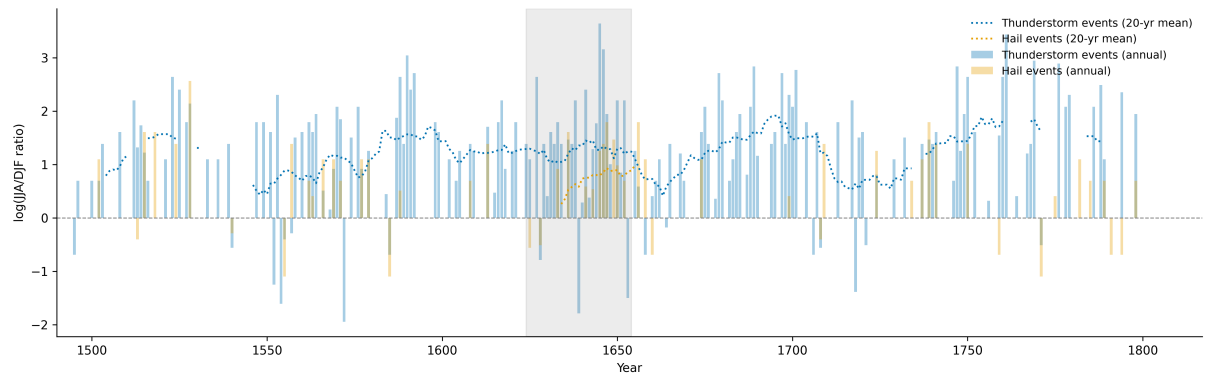


Figure 9. Annual logarithmic summer-winter ratio (JJA/DJF) of historical thunderstorm and hail events from 1490 onwards. The dotted lines show a centred 20-year moving average. Gaps indicate periods with fewer than ten valid annual values. The dashed horizontal line marks a balanced ratio between summer and winter events. The shaded area indicates the period 1624–1654.

time series. From the second half of the 16th century onwards, there is a phase of elevated values, followed by an overall moderate and comparatively stable seasonal ratio ~~until~~ well into the 18th century.

The strength of the seasonal signal varies ~~depending on~~ with the number of documented events (Figure 10). When ~~looking at the~~ individual event types are considered separately, the monthly time series are characterised by numerous months ~~without~~ with no documented events. ~~Increased~~ Elevated monthly values occur sporadically and are often limited to individual years. In these cases, the seasonal distribution is only ~~recognisable~~ discernible to a limited extent.

When ~~thunderstorms~~ thunderstorm and hail events are ~~evaluated~~ analysed together, the number of events recorded each month increases. The ~~monthly series derived from this resulting monthly time series~~ shows a clear ~~difference between~~ distinction between the summer and winter months, with ~~higher proportions in the months of June to August and very low values in the winter half-year~~ a higher proportion of events occurring between June and August, and a very low number of events occurring during the winter months. This pattern ~~is~~ remained consistent throughout much of the ~~period under review~~ study period.

Even when the corpus is stratified, the seasonal cycle of thunderstorm events remains intact. The seasonal profiles show a very high degree of consistency across the four largest language groups (Latin, Early New High German, New High German and Contemporary German) (Spearman's $\rho = 0.82$ – 0.94). There is also a high degree of agreement between the source types (median $\rho = 0.82$). The seasonal signal is therefore not tied to any particular language stage or source type. For hail events, the agreement between source types is lower, which is consistent with the relationship between source type and hail class described in section 4.4.

4.2 ~~Annual distribution~~ Validation of thunderstorm and hail classes linguistic evidence

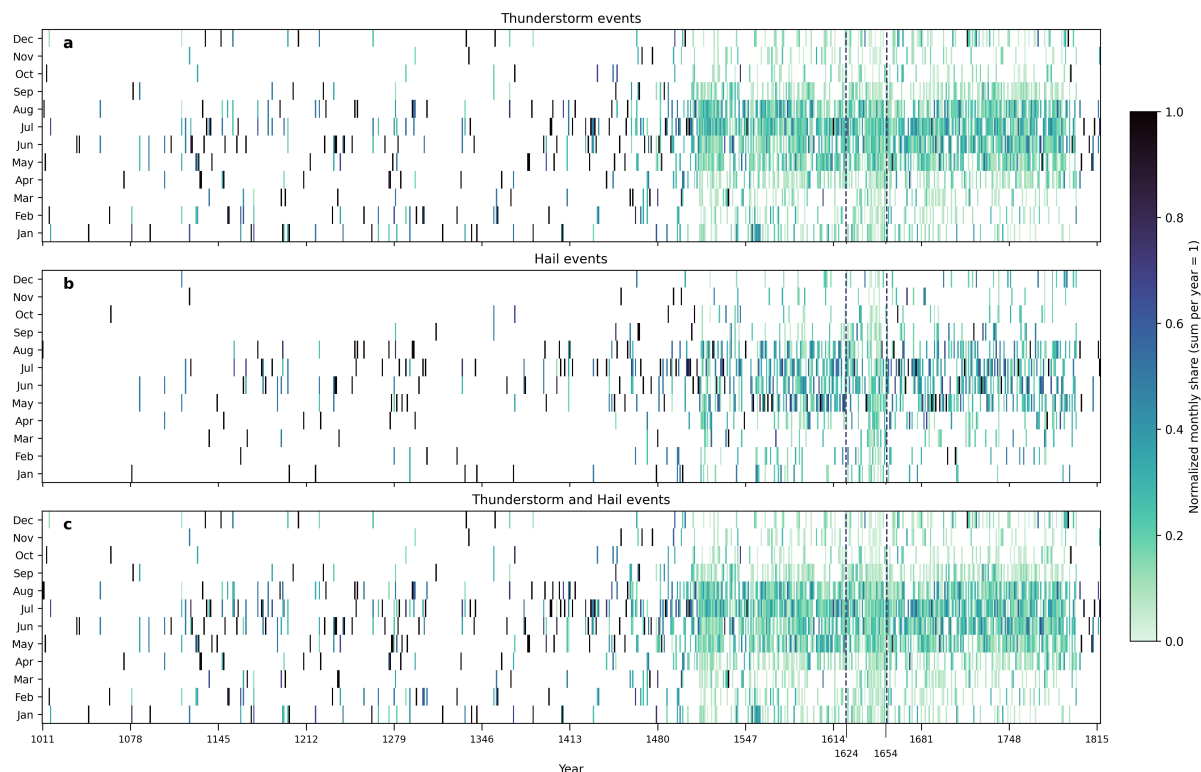


Figure 10. Calendar heatmaps showing the annualised monthly proportions of thunderstorm and hail events for the entire study period. The relative monthly proportions are shown, with the number of events in each year normalised to an annual total of 1, so that the seasonal distribution can be compared independently of the absolute frequency of events. Separate plots show thunderstorm events, hail events and the combination of both event types. The colour intensity corresponds to the normalised monthly proportion. The dotted vertical lines mark the boundaries of the period with a particularly high density of sources (1624–1654).

420 The confidence index clearly distinguishes the classes from one another and rises monotonically between classes TS1 to TS3 and H1 to H2. Classes TS3 and H2 consistently show high values. As expected, the TS1 and H1 classes are dominated by level 0. In the TS2 thunderstorm class, the distribution of evidence classes is mixed, with levels 0 and 1 dominating. This is the class with the greatest uncertainties (see Table 5).

425 Evidence class C1, which is characterised primarily by the size of the hailstones, contributes significantly to the increase in the confidence index. This proportion rises with the thunderstorm or hail class. It occurs in 25.2 % of quotes in class TS3 and 35.1 % in class H2, whilst it is virtually never found in the lower classes (TS1: 0.0 %, H1: 2.4 %).

A small proportion of the TS3 and H2 classes must be classified as uncertain (levels 0 and 1). For TS3, this amounts to 10.5 %, and for H2, 13.6 %. Thunderstorm events are predominantly described in qualitative terms. In 2.9 % of thunderstorm events, there is no clear link between the event and the reported damage. The uncertain hail events are described in qualitative

430 terms, but there is no information on their scale or any clearly identifiable damage.

Table 5. Distribution of the confidence index by thunderstorm and hail class. Higher levels indicate a better, more reliable classification. The “robust” column combines levels 2 and 3 of the index. Evidence class C1 is highlighted. $\emptyset$ is the average of the evidence classes per quote.

Class	n	Index level (%)				robust	Evidence class C1	$\emptyset$
		0	1	2	3			
TS1	3540	90.7	7.5	1.9	0.0	1.9	0.0	0.12
TS2	917	46.9	30.2	20.5	2.4	22.9	1.3	0.99
TS3	966	2.9	7.6	55.9	33.6	89.5	25.2	2.40
H1	735	57.0	38.6	4.2	0.1	4.3	2.4	0.96
H2	1121	0.0	13.6	62.7	23.6	86.3	35.1	2.38

If we look at the confidence index over the centuries, we can see that the proportions of the index levels remain relatively constant (see Figure 11).

4.3 Seasonal cycle of thunderstorm and hail event classes

The weakest thunderstorm class, ~~TS1, dominates in all months~~ TS1, dominates throughout the entire study period in every month. The average monthly proportions ~~are predominantly between approximately 55% and 70%. Class TS2 contributes approximately 11–34% depending on the month, while range from around 53 % to 71 %. The proportion of class TS2 ranges from 12–34 %, whilst~~ the strongest class, ~~TS3, reaches proportions~~ TS3, accounts for between around 10 % and just under ~~30%. Increased proportions of TS3 29 %. Higher proportions of TS3~~ occur particularly in the summer months, ~~with a maximum in the months of June to August. In the winter and transitional months, the ranking TS1 > TS2 > TS3 applies predominantly, while in the summer months, constellations in which~~ peaking in June (28.8 %). From November to April, the order is TS1 > TS2 > TS3. From May to October, TS3 achieves higher proportions than TS2 occur more frequently similar or higher average proportions than TS2.

The ~~bootstrap-based 95% confidence intervals~~ %-confidence interval for the monthly class proportions ~~range from approximately ± 5–10%, with slightly wider intervals in the summer months. Strict rank support (TS1 > TS2 > TS3) is very high in the winter months, remaining is approximately ±4–11 %.~~ It is narrowest during the eventful summer months and widest in October and November. The ordering relation TS1 > TS2 > TS3 remains almost consistently above 95 % from November to April, but falls to almost zero in the summer, indicating systematic violations of the ordering relation. In the transitional months of September and October, it stands at just under 50 %almost throughout. In the summermonths, this value drops significantly and sometimes reaches very low values, indicating frequent breaches of the strict class hierarchy. The weighted support, which ~~also takes into account cases with TS1 > TS3 > TS2, remains significantly higher across all months and is predominantly in the range of approximately 70–85%, even in summer.~~ ordering index, by contrast, remains high throughout the year, standing

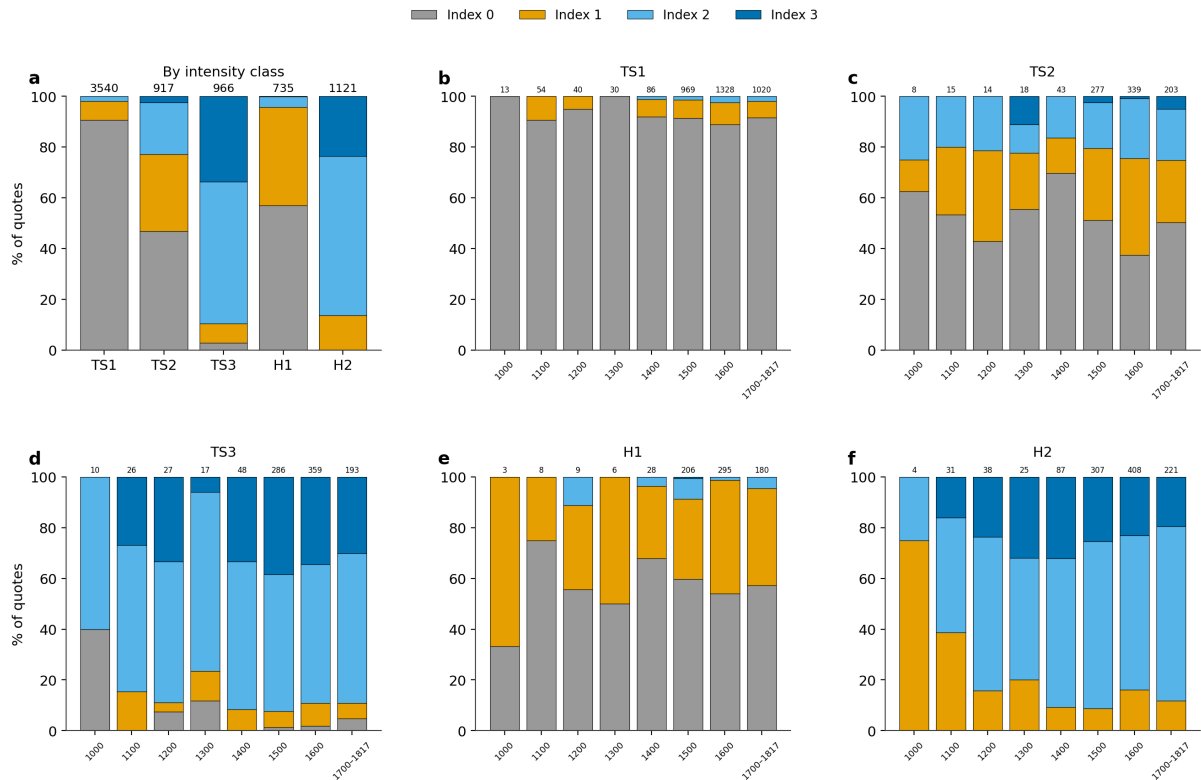


Figure 11. Distribution of the confidence index (levels 0–3) by intensity class and over time. (a) Proportions of index levels by class (TS1–TS3, H1–H2), (b–f) temporal development of the index levels by class over the centuries. The final interval combines the 1700s and 1800s (up to 1817), as there are only a few quotes available for this period. Level 0 = no silver label, 1 = qualitative only (C3), 2 = confirmed (one evidence class C1 or C2), 3 = confirmed multiple times (at least two evidence classes C1/C2).

at around 70 % during the summer months and at 83–85 % in September and October. In summer, the ordering relation $TS1 > TS2$ is almost always satisfied, whilst $TS2 > TS3$ is almost entirely absent.

For the period 1624–1654, the seasonal structure of the storm-classes is similar overall. $TS1$ remains the dominant class in all months except February, with average shares between approximately 50% and over 75%. Classes $TS2$ and $TS3$ show stronger monthly fluctuations proportions ranging from around 52 % to 79 %. In February, however, $TS2$ reaches its highest proportion at 51 %. $TS2$ and $TS3$ show greater monthly variation than in the overall data set, especially in dataset, particularly during the winter months. The 95% bootstrap intervals are partially wider in %-confidence interval is significantly wider in some cases during this period, which is due to the smaller number of years. Strict rank support Support for the ordering relation is significantly reduced in during the summer half-year and falls, falling below 20% in individual months, while weighted support is also predominantly % in July and August. Even in February, support is only slightly higher at 16 %, due to the high proportion of $TS2$. The

weighted ordering index predominantly lies in the range of ~~about 75–90% here~~approximately 70–95 %, with the exception of February (41 %).

465 ~~Hail events also show a pronounced~~ A distinct seasonal pattern in ~~terms of class distribution~~ the class distribution is also evident for hail events. Over the entire period, weaker hail events (~~of class H1~~) dominate in dominate during the winter and transitional months, accounting for ~~more than 65–85 per cent of the total~~around 65–86 %. In the summer months, the distribution shifts significantly ~~in favour of stronger events H2, whose proportions reach values of around 75–85 towards~~ stronger events (H2), whose proportion reaches around 77–82 % in the months of June to August. The ~~95% bootstrap intervals~~
470 ~~are mostly in the range of ± 5 –10%.~~

~~In the period 1624–1654, the basic ranking $H1 > H2$ remains unchanged in the winter months, while high proportions of class H2 occur. %-bootstrap intervals range from approximately ± 5 –14 %, being narrowest in the summer months. The strict ranking support ($H1 > H2$) is very high~~

This pattern persists over the period 1624–1654. The ordering relation $H1 > H2$ is satisfied 95 % of the time in the winter
475 months, ~~usually above 95%, while it~~ drops significantly in ~~summer and almost disappears in some months~~ the summer months and is virtually non-existent in August. This reflects the strong seasonal shift in ~~class proportions~~ the class proportions (see Figure 12).

4.4 Influence of source types on event classes

The ~~chi-square~~ chi-squared test for independence ~~shows~~ reveals a highly significant ~~dependence~~ correlation between source
480 type and intensity class for thunderstorm events (~~$\chi^2 = 529.2$~~ $\chi^2 = 479.1$, $df = 22$, $p < 0.001$). The effect size is in the ~~medium~~ moderate range (Cramér's $V = 0.21$).

The standardised residuals show ~~that weather records in particular are characterised by an overrepresentation of class TS1~~ clear differences between the source types: the weather records show a strong over-representation of class TS1 ($R = +9.0$) and an ~~underrepresentation of class TS3~~ under-representation of TS3 ($R = -11.6$). ~~In chronicles, Chronicles, on the other~~
485 ~~hand, show higher proportions of~~ the strongest class, ~~TS3 is overrepresented~~ ($R = +7.6$), ~~while TS1 is underrepresented~~ alongside an under-representation of TS1 ($R = -5.6$). ~~Almanacs show a clear underrepresentation of TS3~~ Handwritten diaries are characterised by a marked under-representation of TS3 ($R = -6.2$). ~~Newspapers, whilst newspapers~~ and historiographical works show increased proportions of ~~class TS3~~ TS3 ($R = +4.5$ and $R = +3.6$, respectively). These ~~cells~~ deviations contribute significantly to the ~~total overall~~ χ^2 value (Figure 13).

490 For hail events, ~~the chi-square test also shows too, there is~~ a highly significant correlation between source type and intensity class (~~$\chi^2 = 390.6$~~ $\chi^2 = 373.2$, $df = 10$, $p < 0.001$). The effect size is significantly ~~more pronounced~~ greater here than for thunderstorm events (Cramér's $V = 0.44$ 0.43).

The ~~standardised~~ residuals show a ~~strong overrepresentation of class H1 in weather records~~ similar, yet more differentiated pattern: the weather records are characterised by a marked over-representation of H1 ($R = +10.4$) and a simultaneous underrepresentation
495 of class H2 ($R = -8.3$). ~~In chronicles, class H2 is overrepresented~~ an under-representation of H2 ($R = -8.2$). Chronicles show increased proportions of H2 ($R = +4.8$), ~~while H1~~ whilst H1 is underrepresented ($R = -6.0$). Almanacs also show

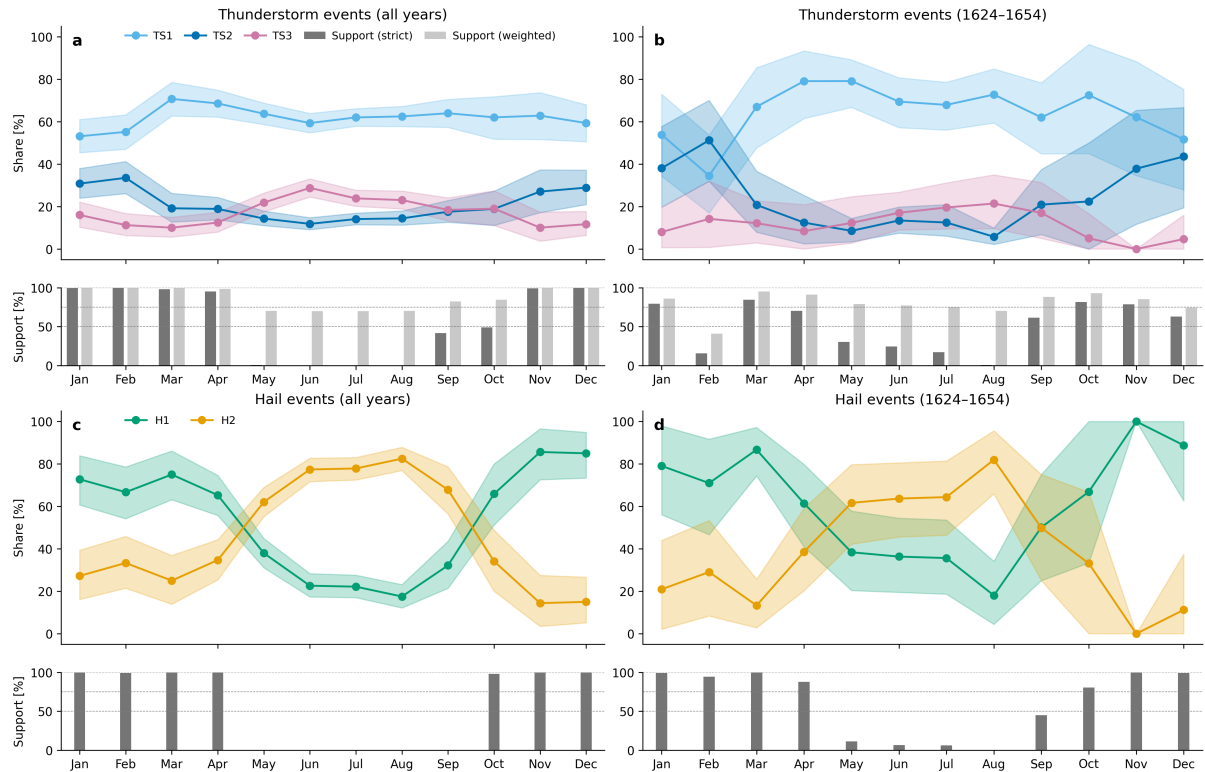


Figure 12. Distribution of class proportions for thunderstorm and hail events over the entire study period and the period 1624–1654. The monthly class proportions (lines) are shown with 95 %-confidence intervals (shading), calculated from monthly distributions normalised on an annual basis. The lower panels show the proportion of bootstrap samples in which the expected ordering relation, based on the resampled monthly proportions, is satisfied.

increased proportions of $H1$ ($R = +6.6$) and reduced proportions of $H2$ ($R = -5.2$). Administrative files also show an overrepresentation of $H1$ ($R = +4.6$). These deviations explain a large part of the observed χ^2 value (Figure 13).

4.5 Transformer-based language model

Both classification models (ThunderstormBERT, HailBERT) are based on the same transformer architecture (mDeBERTa-v3-base). The hyperparameters specified in Table 6 correspond to the best settings identified during model development for the respective classification task. Despite having an identical architecture, there are differences in the training setup, particularly in with regard to the learning rate scheduler.

The learning curves of ThunderstormBERT show stable convergence after approximately five epochs with no evidence of relevant overfitting around five epochs, with no signs of significant overfitting (Figure 14).

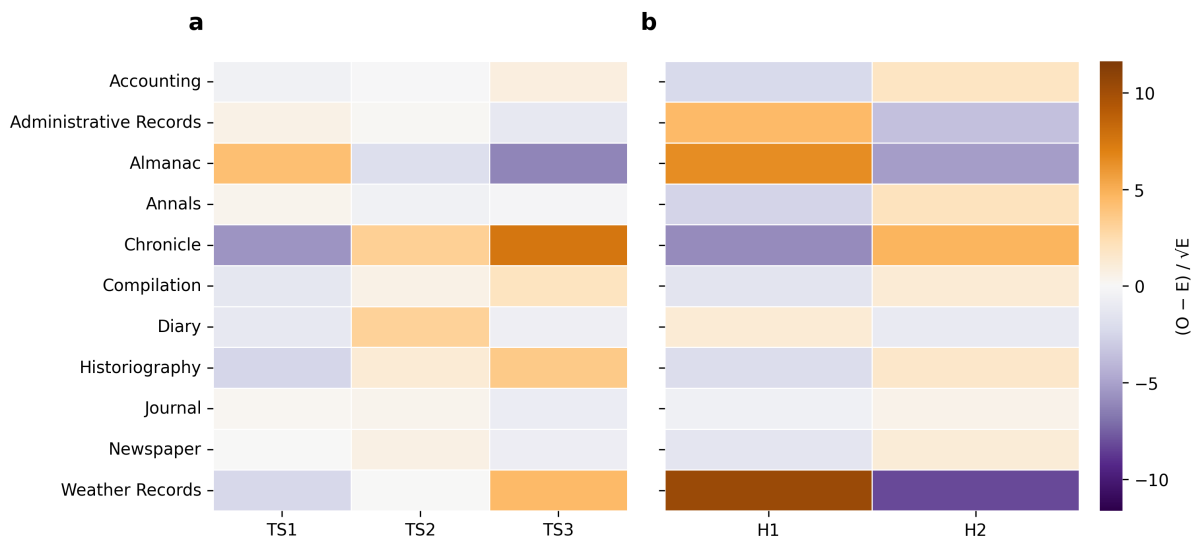


Figure 13. Standardised residuals from the chi-squared tests for the relationship between source type and event class for thunderstorm events (left, TS1–TS3) and hail events (right, H1–H2). The residuals $(O - E)/\sqrt{E}$ are shown for each combination of source type and class. Positive values indicate over-represented combinations, whilst negative values indicate under-represented combinations relative to the frequency expected if there were no dependence.

Table 6. Key metrics of the language models for classifying historical thunderstorm and hail events: ThunderstormBERT and HailBERT.

Model	Thunderstorm events (4 classes)	Hail events (3 classes)
Architecture	mDeBERTa-v3-base	mDeBERTa-v3-base
Max. token length	256	256
Learning rate	3×10^{-5}	3×10^{-5}
Label smoothing	0.05	0.05
Scheduler	linear	cosine
Batch size	16	16
Epochs	5	5
F1 (macro), validation	0.841	0.907
F1 (macro), test	0.826	0.931
Accuracy (test)	0.871	0.961
Test samples (n)	1377	1399
Error rate (test)	12.9 %	3.9 %

ThunderstormBERT achieves a macro-averaged F1 ~~performance-score~~ of 0.826 on the independent test dataset, with an accuracy of 0.871 (corresponding to an error rate of 12.9 %). ~~Frequent-Common~~ classes such as *no thunderstorm* and *light thunderstorm* are ~~detected-recognised~~ with high precision and high-recall (F1 > 0.91), ~~while-whilst~~ the rarer intensity classes, as

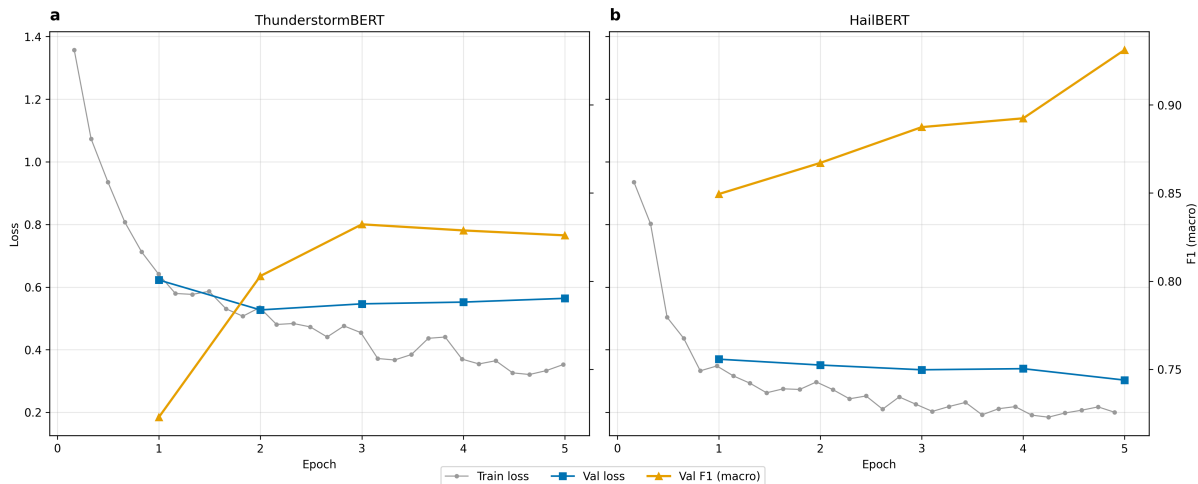


Figure 14. The figure shows the trend in the training and validation metrics for ThunderstormBERT and HailBERT over 5 epochs.

expected, have exhibit lower but consistent F1 values (Tablescores (Table 7). Misclassifications occur predominantly between neighbouring intensity levels (Figure 1), suggesting 15), which suggests gradual semantic transitions in the historical descriptions, where intensity gradations are often formulated implicitly or contextually.

Table 7. Class-specific test performance of the thunderstorm model.

Class	Precision	Recall	F1	Support
TS0	0.894	0.930	0.911	199
TS1	0.948	0.932	0.940	760
TS2	0.642	0.677	0.659	201
TS3	0.806	0.783	0.794	217

The learning curves for HailBERT show a continuous increase in the F1 score for HailBERT with alongside a simultaneous decrease in loss (Figure 14). Validation stabilises from the fourth epoch onwards, indicating efficient convergence with no signs of relevant-significant overfitting. The normalised confusion matrix shows very high discrimination separation between classes, with only a few misclassifications outside neighbouring intensity levels (Figure 15).

HailBERT achieves a very high overall classification performance with a macro-averaged F1 score of 0.931 and an accuracy of 0.961 (corresponding to an error rate of 3.9 %). All classes are classified with high and balanced scores (Table 8). Here, too, misclassifications are predominantly concentrated in-on neighbouring intensity classes, which indicates-suggests a consistent internal representation of the class-order. The ordering relation. HailBERT's higher overall performance of HailBERT must should also be interpreted in the context of the smaller number of classes and the clearer semantic distinguishability of hail events.

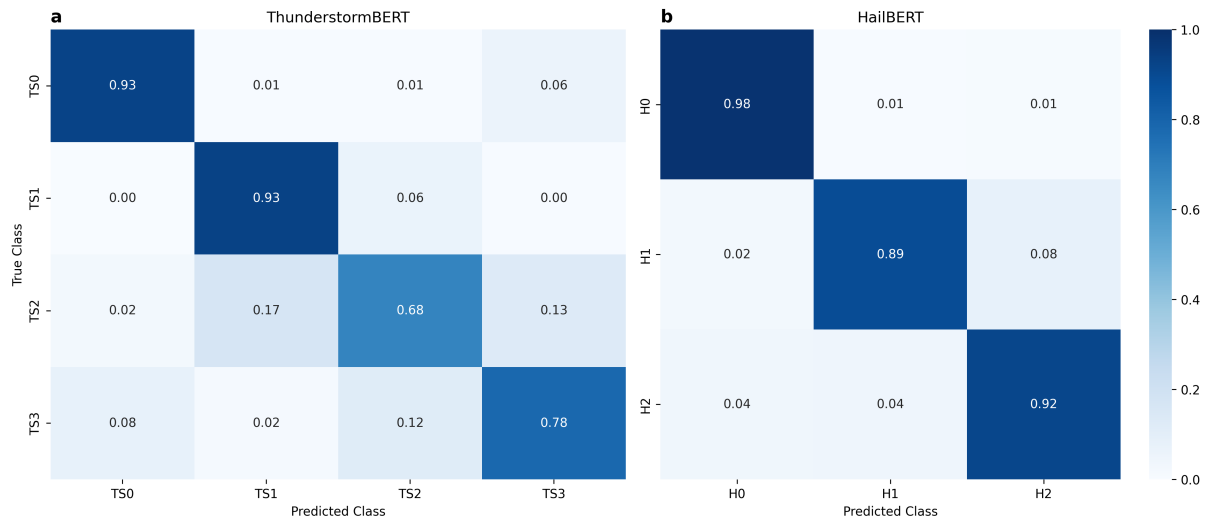


Figure 15. Normalised confusion matrix for ThunderstormBERT and HailBERT on the independent test dataset. For each true class (row), the probability that the model will predict this class correctly (diagonally) or as a different class is shown.

Table 8. Class-specific test performance of the hail model.

Class	Precision	Recall	F1	Support
H0	0.984	0.984	0.984	964
H1	0.889	0.894	0.892	161
H2	0.919	0.916	0.918	274

5 Discussion

525 The ~~analysis shows that~~historical study shows that, despite widely varying source density and a heterogeneous source base, ~~historical reports of~~ thunderstorm and hail ~~event reports~~ events exhibit a remarkably stable seasonal signal~~despite highly variable source density and a heterogeneous source base~~. The reconstructed annual ~~cycle, which is based on monthly proportions normalised on an annual basis,~~ pattern is dominated by a pronounced summer maximum. The maximum ~~determined shows significant agreement with modern observational data~~ identified closely matches the observational data from the DWD normal

530 ~~periods~~ in terms of location, shape and ~~relative~~ amplitude. This pattern ~~manifests itself consistently both in the overall series and in the densest observation window (1624–1654). The stability of the seasonal pattern is evident regardless of fluctuations in the density of records and is therefore not attributable to individual phases of increased documentation or the distribution of sources~~ remains consistent for both the entire period and the period with the highest density of observations (1624–1654).

~~Despite a time lag of several centuries and fundamentally different survey methodologies, a comparison with current~~

535 ~~observation data from the DWD shows a high degree of consistency in seasonal patterns.~~ The high Spearman rank correlations and ~~the low deviations~~ small variations in the normalised monthly proportions indicate that the historical observations

~~consistently refer examined in this study are consistently related~~ to real meteorological processes. Furthermore, the seasonal ~~signature proves to be pattern is~~ largely independent of ~~the~~ observation density: ~~even pronounced significant~~ fluctuations in the number of ~~active sources only slightly quotes only minimally~~ alter the shape of the annual ~~eye~~pattern.

540 ~~The annual trends presented in this study show significant agreement~~ A comparison with independent studies further supports these findings. The seasonal cycle shows a clear correspondence with the independent reconstructions by Lenke (1960) and Camuffo et al. (2000). This agreement is evident both for the entire time series and for the ~~densest window from 1624 to 1654, as shown in Figure 16.~~ The period with the highest observation density, 1624–1654 (Figure 16). The correlation analysis supports this agreement (Hist vs. Lenke: Spearman 0.86; Hist vs. Camuffo: 0.71), with almost identical results yielded for the ~~most densely observed period (1624–1654).~~ This points to a robust seasonal signature of thunderstorm activity across different ~~time periods, regions and source types, and highlights the plausibility of the reconstructed annual pattern.~~

545 Differences are evident for the winter: the ratio of summer to winter ~~proportions components~~ (JJA vs. DJF) in the ~~historical series is approximately 3.0–3.5, while reconstructed time series of this study is approximately 3.0–3.5, whilst the values according to Camuffo et al. (2000) (6.6) and , in particular, in Lenke (1960) (13.2) show a significantly stronger summer-winter contrast.~~ Winter activity is ~~therefore more pronounced in the historical data set than in the comparative series, but remains structurally within the expected seasonal pattern.~~ One potential are more pronounced. One possible explanation for the increased observation of ~~thunderstorms~~ thunderstorm events in winter could be ~~due to source-typical source-specific~~ perceptions of winter ~~thunderstorms~~ thunderstorm and hail events, which were ~~perceived as unusual.~~

550 ~~The correlation analysis supports the agreement (Hist vs. Lenke: Spearman 0.86; Hist vs. Camuffo: 0.71), with the period 1624–1654 yielding almost identical results. This indicates a robust seasonal signature of thunderstorm activity across different periods, regions and source types, and underlines the plausibility of the reconstructed annual cycle~~ regarded as unusual (e.g. Camuffo et al., 2000; Glaser, 2013). Whether this is attributable to the composition of the corpus cannot be conclusively determined here.

~~The temporal development of the summer–winter ratio~~

560 The ratio between summer and winter (JJA/DJF) ~~also confirms this robustness in a longitudinal analysis. Over remains stable above the unit line throughout the entire observation period: over~~ more than four centuries, ~~the ratio has remained predominantly above the unit line and shows there are~~ no systematic trends that would ~~indicate suggest~~ changes in source density or documentation practices. The consistently pronounced summer maximum suggests that the observed seasonal ~~signature is primarily determined by real pattern is primarily attributable to actual~~ convective processes and is not ~~generated by transmission or selection artefacts.~~

565 ~~This is further supported by the calendar heat maps. Periods of increased activity are concentrated in the summer months, while the winter months show very little activity. The due to effects of transmission or the selection of quotes.~~

570 The discernible, physically consistent seasonality ~~that can be observed~~ is a key criterion for the plausibility of historical records and ~~speaks for supports~~ the reliability of the reconstructed thunderstorm and hail time series, as it demonstrates that the data are reliable when seasonality and climate are taken into account. The individual signals from the thunderstorm and hail classes, on the other hand, paint a somewhat more nuanced picture.

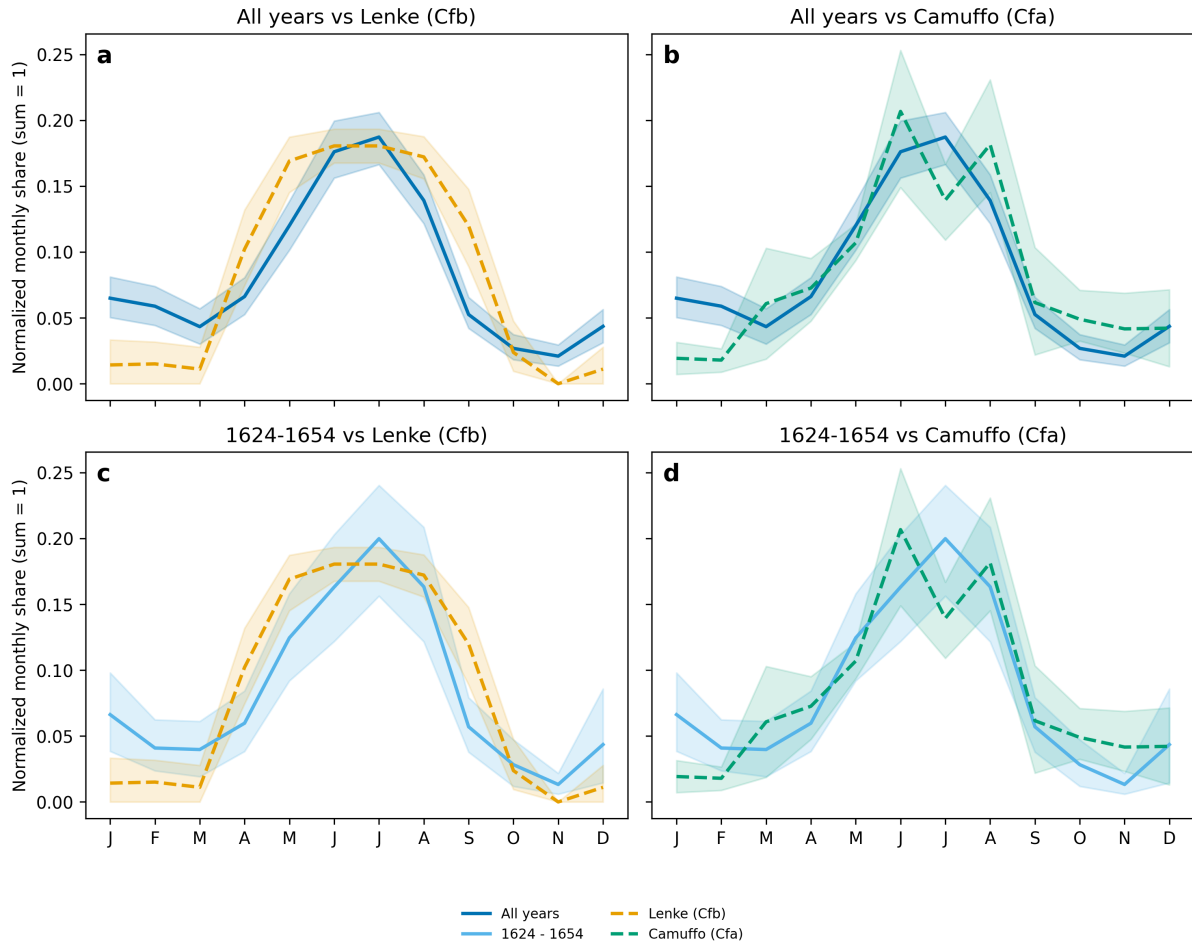


Figure 16. Comparison of the seasonal cycle of thunderstorm events using two independent time series. The seasonal distributions are shown, each with 95 % bootstrap confidence intervals. The top panel shows a comparison of the complete time series from this study with Lenke (1960) (Hesse, Köppen Cfb) and Camuffo et al. (2000) (Padua, Köppen Cfa). The bottom panel shows the corresponding comparison for the period with the highest observation density (1624–1654). Despite differing climate regimes, namely the oceanic Cfb climate of central Germany and the humid subtropical Cfa climate of the Po Valley, there is a high degree of consistency in the shapes of the curves. This underlines the robustness of the seasonal pattern, characterised by a pronounced summer maximum and moderately increased winter activity, across different datasets, regions and methodological approaches.

Analysis of the monthly class proportions shows a clear dominance of the weakest thunderstorm class (*TS1*) over the entire study period, with the highest average proportions in all months. Class *TS2* occurs with First, we will examine the linguistic evidence. This shows that the classified hail events are of a consistently high quality across all periods. The majority of hail events classified as H2 have a confidence index of ≥ 2 . The distinction between classes H1 and H2 can be clearly made on the basis of hailstone size or the intensity of the hail, which leaves little room for interpretation. Hail events with an index level of 0 are found only in class H1. In such quotes, the term 'hail' stands alone and leaves no room for interpretation.

The quality of the classification of thunderstorm events differs from that of hail events: whilst the *TS1* and *TS3* classes exhibit a stable composition of evidence classes, *TS2* shows significantly greater variability and thus greater uncertainties in interpretation. Individual *TS1* events with evidence class C3 are difficult to classify due to a lack of descriptions, but do not significantly affect the overall signal. Furthermore, for a small proportion of thunderstorm events (2.9 %), the causal link between the event and the reported damage is absent, meaning that the damage does not contribute to validating the classification. Overall, it is clear that the classification of thunderstorm events is more challenging, particularly in the middle class *TS2*. The classification is therefore asymmetric: the high-intensity classes (*TS3* and H2) are based on direct, verifiable evidence, whereas the threshold-free middle class (*TS2*) remains transitional with weaker evidential support. As with hail events, the composition of the confidence index remains stable across the entire period and across all classes.

The analysis of the ordering relation shows a clear dominance of thunderstorm class *TS1* throughout the seasonal cycle. Class *TS2* occurs at a lower but stable proportions over proportion throughout much of the year, while the strongest class, *TS3*, achieves higher relative proportions, especially in class *TS3* achieves a higher proportion, particularly during the summer months. In During the transitional seasons, the ranking is predominantly $TS1 > TS2 > TS3$. In ordering relation is predominantly $TS1 > TS2 > TS3$. However, in the summer months, however, the average class proportions repeatedly show a changed secondary ranking of $TS1 > TS3 > TS2$ the secondary ranking repeatedly shifts to $TS1 > TS3 > TS2$ in the medium-range class proportions. An exception is February, during the dense observation period 1624–1654, when *TS2* dominates at 51 %. However, given the small number of winter events and the correspondingly wide confidence intervals, this single outlier is not statistically significant.

These findings conflict with contradict the assumption frequently expressed put forward in the literature that historical sources primarily record extreme weather events (cf. Lenke, 1960; Camuffo et al., 2000; Brázdil et al., 2016a, b; Burgdorf, 2021). However, such a general dominance of extreme classes cannot be confirmed for thunderstorm and hail events. Under this assumption, a ranking of $TS3 > TS2 > TS1$ an ordering relation of $TS3 > TS2 > TS1$ would have been expected. In fact, the observed class proportions in historical records also observations in this study predominantly follow the physical frequency distribution, with a clear dominance of weaker events ($TS1 > TS2 > TS3$) thunderstorm events (*TS1*). This suggests that, at least for thunderstorm events, there is no systematic overrepresentation over-representation of extreme intensities.

The analysis of the dependence between source type and intensity classification shows for thunderstorms ($\chi^2 = 479.66$, $df = 22$, $p < 0.001$, Cramér's $V = 0.21$) and hail events ($\chi^2 = 376.23$, $df = 10$, $p < 0.001$, Cramér's $V = 0.43$). While the correlation for thunderstorm events can be classified as small to moderate (corresponding to an explained variance of about 4%), hail

events show a significantly stronger, medium effect (variance explained about 18%). These differences point to source-specific divergent reporting patterns, which are much more pronounced for hail events than for thunderstorm events overall.

The seasonal contrast in the ranking differing seasonal cycles of hail classes H1 and H2 can be explained by a combination of physical and source-related effects. In During the summer months, high convective energies mean energy means that hail events tend to be more intense when they occur, while weak hail is either meteorologically less frequent or less frequently documented whilst light hail either occurs less frequently meteorologically or is documented less often due to its lower visibility. In winter and transitional seasons, on the other hand, hail occurs predominantly by contrast, hail events predominantly occur in weaker forms, but is are recorded even at low intensity intensities due to the overall generally lower event density and increased perception sensitivity sensitivity to observation. The transitional months of April and September consistently mark the seasonal change in this dynamic change of season. This is consistent with the fact that the seasonal increase in hail events begins earlier than that for thunderstorm events, with higher proportions already evident in March and April. Noteworthy is the high correlation between the H1 data and the DWD observational data ($\rho(\text{H1}, \text{DWD}) = 0.87\text{--}0.92$), which generally only record light hail, whilst H2 shows a typical peak in summer. This underlines both the data quality and the plausibility of real-world observations.

These physical differences are further reinforced by source-specific reporting patterns. Narrative sources in particular, The relationship between source type and intensity class (Section 4.4) is significant for both thunderstorm and hail events, although to varying degrees: for thunderstorm events, the effect is small (explained variance approximately 4 %), whereas for hail events it is moderate (approximately 18 %). The reporting patterns of the source types influence the classification. However, they are not strong enough to be interpreted as a primary effect. In particular, narrative sources such as chronicles , tend to disproportionately tend to document more intense hail events (H2) , while weather diaries disproportionately, whilst weather records more frequently record weaker hail events (H1). Hail appears to be particularly newsworthy in narrative sources when it causes damage or takes In documentary sources, hail events appear to be considered newsworthy primarily when they cause damage or take on unusual forms, while continuous observation formats also systematically record . Continuous observation formats, by contrast, systematically record even lower intensities.

It is noteworthy that these source-specific distortions are implicitly reflected in the hail classification model, even though the source type was not explicitly integrated as a training feature. The high classification quality of the model with only minor misclassifications suggests that the semantic patterns contained in the historical texts are sufficient to automatically map and take into account such source-related differences. However, the stability of the confidence index throughout the entire observation period shows that the description of thunderstorm and hail events hardly differs in terms of type and composition. This suggests that only a synoptic analysis across all sources leads to plausible results.

The results of the language model fit in with the findings previously obtained transformer-based language model can be contextualised within the previously obtained findings on the seasonal structure of convective events. They The high data quality of the hail events and the uncertainties in thunderstorm classification influence the quality of the model. Whilst the automatic classification of hail events and thunderstorm classes TS1 and TS3 is accurate, slightly lower F1 scores are achieved for thunderstorm class TS2. The results of the model training thus reflect the results of the confidence index.

However, the results also show that the patterns observed in historical sources can also be consistently reproduced at the semantic-linguistic level. Although the model was trained exclusively on texts and does not include any additional meteorological variables, it reproduces the characteristic ~~class hierarchies~~ ordering relation (e.g. ~~$TS1 > TS2 > TS3$ over the course of the year~~ $TS1 > TS2 > TS3$ during the seasonal cycle) as well as the summer intensification of ~~stronger~~ more severe events. The model thus confirms the coherence of the corpus in terms of content: the historical texts contain sufficient meteorologically relevant information to reliably capture ~~seasonal and intensity-related differences~~.

The class structure used shows stable results across all classes. This illustrates that the model does not primarily memorise formulations, but rather recognises recurring semantic patterns. ~~The few both seasonal differences and differences in the severity of the events.~~

The remaining misclassifications occur almost exclusively at ~~transitions between neighbouring intensity levels~~ the boundaries between neighbouring classes. This pattern is ~~known from both~~ recognised both in meteorological practice and in source-critical analysis. ~~Serious mix-ups~~ This study is no exception in this regard, although serious misclassifications are rare. This indicates illustrates that the model does not primarily learn formulations by heart (overfitting), but rather recognises recurring semantic patterns, and speaks in favour of a robust separation of the relevant linguistic signals and the resilience of the classification scheme.

~~This source-specific effect was not explicitly integrated into the language model as an input variable, but is reproducibly reflected in the classification results. This suggests that the model learns implicit structures of historical tradition — including source-specific selection mechanisms — rather than levelling them out. The models thus work consistently with the characteristics of the corpus, which is of central importance for applications in historical climate research.~~

~~Both models show~~ Both models produce stable and reproducible results and are suitable for the automated classification of historical weather descriptions. ~~The robustness of the results is supported by consistent seasonal patterns, narrow confidence intervals and stable rankings of intensity classes~~ In doing so, they learn the implicit structures of the historical record, including source-specific selection mechanisms, rather than smoothing them out, and thus operate in a manner consistent with the characteristics of the corpus. The observed ~~performance differences~~ differences in performance between thunderstorm and hail classification ~~are significantly greater than differences in the training setup and~~ can be explained primarily by the different class ~~structure and semantic selectivity~~ structures and semantic discriminative power of the respective target phenomena.

6 Conclusions

~~The~~ Despite their varying source types and linguistic heterogeneity, the study shows that historical ~~text sources have~~ textual sources offer great potential for the quantitative reconstruction of convective weather events ~~despite linguistic heterogeneity. By~~ . A consistent database was created by combining source-critical analysis, rule-based classification ~~and statistical validation, it was possible to create a consistent database whose event patterns correspond to modern convective climatology. The conversion of,~~ and linguistic and physical-statistical validation. Converting qualitative descriptions into structured evidence ~~levels and semantic labels enabled~~ classes and developing a confidence index made it possible to assess the data quality. The reliability

is very high for hail events, as well as for light, severe, and extreme thunderstorm events. However, uncertainties remain
675 for moderate thunderstorm classes and winter thunderstorm events, which are mentioned too frequently. The qualitative
composition has remained remarkably stable for over 800 years. This demonstrates that historical observations of thunderstorm
and hail events exhibit minimal linguistic variation, allowing typical patterns to be identified in the descriptions. This makes
automated classification all the more relevant, as demonstrated by the successful use of a multilingual BERT model ~~that reliably~~
~~recognises and classifies~~ for detecting and classifying thunderstorm and hail events. ~~Overall, the work shows that~~ Both models
680 can correctly classify new data.

Overall, this study demonstrates the ability of AI-based methods ~~can only to~~ extract reliable climatological information
from ~~historically data that has been methodically prepared~~ systematically processed historical data. The formally defined
classification procedure with an evidence-based quality assessment provides a reproducible framework that explicitly considers
both the variability in the language used to describe historical events and the meteorological interpretation required for
685 climatological analysis. This extends existing approaches in a useful way.

~~Future work should expand the dataset to systematically investigate regional differences between large landscapes~~ The high
quality of the observations can be explained by the significant social role of thunderstorm and hail events in pre-industrial
agrarian societies, among other things. Such events posed an existential threat in these societies and were therefore documented
in great detail. This dataset provides new insights into the study of convective weather in the context of climate change. As
690 instrumental and radar-based observation series only span a few decades, long-term reference data from the pre-industrial
climate regime is necessary to contextualise the natural variability of thunderstorm and hail activity. The period covered here,
which extends back to the year 1000, provides such a basis. However, a key question remains: how does the internal variability
of convective activity in the natural climate regime compare with today's anthropogenically influenced conditions? This
dataset and the classification tools it provides form a basis for answering this question, which can be further developed through
695 expanded spatial and temporal coverage in future work. The empirical results apply to the German-speaking source corpus of
Central Europe. The classification method itself is designed to be language-independent. However, it must be recreated using
the same corpus-based procedure for each new target language. The next step is therefore to apply the method to comparable
corpora in other languages, such as French or English.

Code and data availability. Historical Climate Observations of Thunderstorms and Hail in Central Europe (1000–1900):
700 <https://doi.org/10.60493/834bd-mww13>,
ThunderstormBERT-de-v1: 10.57967/hf/6982 or <https://huggingface.co/Stickmu/ThunderstormBERT-de-v1>,
HailBERT-de-v1: 10.57967/hf/6989 or <https://huggingface.co/Stickmu/HailBERT-de-v1>

Appendix A: Classification schemes

Table A1. Classification of thunderstorm intensity

Class	Description
–99	The thunderstorm’s characteristics cannot be clearly determined.
0	No thunderstorm: This class includes events that do not meet the basic criteria for a thunderstorm and serves as a baseline category for differentiation.
1	Light thunderstorm: Events that do not meet the criteria for TS2 are classified as light if all associated phenomena exhibit only low intensity (rain $\leq$ level 1, wind $\leq$ level 1, hail $\leq$ level 1).
2	Moderate thunderstorm: This class includes thunderstorm events that do not reach the thresholds for a severe thunderstorm (TS3) but exceed the intensity of a light thunderstorm (TS1).
3	Severe thunderstorm: An event is assigned to this highest level if at least one of the following criteria is met: – Hail diameter $>$ 2.0 cm – Wind speed reaches predefined level 3 (severe storm/hurricane-force gusts) – Rain intensity reaches predefined level 3 (heavy rain)

Table A2. Classification of hail intensity

Class	Description
–99	The hail’s characteristics cannot be clearly determined.
0	No hail: Quote without hail events.
1	Light hail: Quote mentioning hail without hail impact. Hail size is less than 2 cm.
2	Hail impact or large accumulation: Hail impact occurs at sizes of 2 cm or larger. Frequent comparisons include cherry-, walnut-, pigeon-, or hen’s egg-sized hail. Damage may also be described indirectly (e.g., damaged leaves, branches, fruit, window panes, roofs, or metal cladding). Quantity descriptions such as “hail lay knee-deep in the alleys” also fall under this category.

Table A3. Classification of rain intensity during thunderstorm and hail events

Class	Description
–99	The rain’s characteristics cannot be clearly determined.
0	No rain: Thunderstorm or hail without accompanying precipitation.
1	Rain: Thunderstorm or hail with ordinary rain.
2	Heavy rain without flooding: Thunderstorm or hail with heavy rain but without immediate consequential damage. Typical historical descriptions include <i>cloudburst</i> or <i>torrential rain</i> .
3	Heavy rain and flooding: Thunderstorm or hail with heavy rain and severe impacts such as floods, landslides, damage to mills, or inundated fields. Unlike class 2, this class involves immediate geomorphological and socio-economic consequences.

Table A4. Classification of snow intensity during thunderstorm and hail events

Class	Description
–99	The snow’s characteristics cannot be clearly determined.
0	No snow: Thunderstorm or hail without accompanying snowfall.
1	Light snowfall: Thunderstorm or hail with snow. Fresh snow depth below 10 cm in lowland areas or below 20 cm in mountainous regions, without significant consequences.
2	Heavy snowfall: Thunderstorm or hail with heavy snowfall. Fresh snow depth of 10 cm or more in lowland areas or 20 cm or more in mountainous regions, without descriptions of significant consequences. Typical historical descriptions include “a lot of snow,” “large amounts of snow,” or “piling snow.”
3	Heavy snowfall with consequences: Any snowfall that explicitly causes disruptions or damage, regardless of the amount. Typical indicators include snow breakage in trees, collapsed roofs, impassable roads, snowed-in persons, or avalanches.

Table A5. Classification of wind intensity during thunderstorm and hail events

Class	Description
–99	The wind’s characteristics cannot be clearly determined.
0	No wind: Thunderstorm or hail without significant wind.
1	Wind: Thunderstorm or hail with wind; convective gusts up to approx. 7 Beaufort.
2	Strong wind with minor damage: Convective storm gusts (8–9 Beaufort). Typical indicators: Large trees sway, shutters open, branches break, significant difficulty walking, minor damage to houses (individual roof tiles or chimney pots lifted).
3	Strong wind with severe damage: Convective severe storm gusts, hurricane-force gusts, or tornadoes (≥ 10 Beaufort). Typical indicators: Uprooted trees, snapped trunks, windthrow in forests, severe damage to buildings (roofs blown off, thick walls damaged), walking impossible, widespread devastation.

Table A6. Standardised terms and size specifications for historical hail descriptions. The table summarises typical expressions from historical sources used to describe hail events; by assigning them to standardised size ranges and energy levels, it enables consistent and comparable classification of hail intensity within the classification scheme. The thresholds follow established meteorological and climatological references. The decisive 2 cm threshold for hailstones follows European Severe Storms Laboratory (ESSL) (2025) and Tornado and Storm Research Organisation (TORRO) (2025), above which the kinetic energy of hailstones increases disproportionately, raising the probability of substantial damage to agricultural crops, buildings and infrastructure.

Term (original)	Example (database ID)	∅ cm	Class
<i>Schnellkugel</i> (musket ball)	“...wie große Schnellkugeln...” (5638)	0.9	H1;TS2
<i>Bohne</i> (bean)	“...wie kleine Bohnen...” (479)	1.0–2.0	H1;TS2
<i>Haselnuss</i> (hazelnut)	“...so groß wie Haselnüsse...” (61)	1.2–2.0	H1;TS2
<i>Ital. Nuss</i> (Italian nut)	“...Größe einer ital. Nuss...” (2073)	1.5–2.0	H1–H2;TS2–TS3
<i>Schoßkeule</i> (pistol ball)	“...so groß wie Schoßkeulen...” (496)	1.5–2.0	H2;TS3
<i>Märmel</i> (marble)	“...teils so groß wie eine Märmel...” (33897)	1.6–2.0	H1;TS2
<i>Viertelguld.</i> (quarter gulden)	“...¼ Gulden schwer...” (4240)	1.8–2.2	H2;TS3
<i>Daumen</i> (thumb)	“...als ein Daumen dick...” (707)	2.0–2.5	H2;TS3
<i>Kastanie</i> (chestnut)	“...so groß wie Kastanien...” (1010)	2.0–4.0	H2;TS3
<i>Baumnuss</i> (walnut)	“...so groß wie Baumnüsse...” (2959)	2.1–3.0	H2;TS3
<i>Walnuss</i> (walnut)	“...so groß wie Walnüsse...” (354)	2.1–3.0	H2;TS3
<i>Böhm. Grosch.</i> (Bohemian groschen)	“...wie böhmische Groschen...” (928)	2.7–3.0	H2;TS3
<i>Taubenei</i> (pigeon’s egg)	“...wie Taubeneier groß...” (86)	3.1–4.0	H2;TS3
<i>Taler</i> (thaler coin)	“...so groß wie halbe Taler...” (696)	4.0–4.5	H2;TS3
<i>Hühnerei</i> (hen’s egg)	“...so groß wie ein Hennenei...” (60)	5.1–6.0	H2;TS3
<i>Apfel (halb)</i> (half apple)	“...Größe halber Äpfel...” (6676)	6.0–7.5	H2;TS3
<i>Brot (halb)</i> (half loaf)	“...so groß wie ein halbes Brot...” (2787)	7.5–10.0	H2;TS3
<i>Gänseei</i> (goose egg)	“...wie Gänseeier groß...” (281)	8.0	H2;TS3
<i>Faust</i> (fist)	“...wie eine Faust groß...” (1053)	10.0–12.0	H2;TS3
<i>Kürbis</i> (pumpkin)	“...wie Kürbis...” (2994)	20+	H2;TS3
<i>Hagelstein</i> (hailstone)	“...ungewöhnlichen Größe...” (6)	0.5–5.0	H1–H2;TS2–TS3
<i>Anderthalb Pfund</i> (one and a half pounds)	“...anderthalb Pfund...” (361)	n.a.	H2;TS3
Damage (C2)	“...schlug die Fenster ein...” (385)	n.a.	H2;TS3
Accumulated (C1)	“...fast schuhhoch...” (2140)	n.a.	–

Appendix B: Linguistic indicators

Table B1. Linguistic indicators for thunderstorm and hail events and their associated phenomena (wind, rain, snow) in historical sources

Phenomenon	Linguistic indicators (nouns, verbs, adjectives)
Thunderstorm	blitz, blitzeinschlag, blitzen, blitzgewitter, blitzn, blitzschlag, blitzstrahl, blitzte, donner, donnereinschlag, donneren, donnergewitter, donnergrollen, donnerhall, donnerkeil, donnerknall, donnerknalle, donnerkrach, donnern, donnerndes, donnerregenwetter, donnerschlag, donnerstein, donnerstrahl, donnerstreich, donnerwetter, feuerblick, feuerdrach, feuererscheinung, feuerflamme, feuerhagel, feuerklumpe, feuerklumpen, feuerkugel, feuerregen, feuerschlag, feuerstahl, feuerstrahl, feuerwerk, feuerwolke, feuerzeichen, fromalwetter, geblitz, gedonnert, gewitt, gewitter, gewitterheiß, gewitterleuchte, gewitterleuchten, gewitterläuten, gewittern, gewitterregen, gewitterschaden, gewitterschauer, gewittersturm, gewitterwind, gewitterwolke, gewittrig, hagelunwetter, himmelblitzen, hochgewitter, hochwetter, kanonenschuss, knall, krachen, krachend, lichtwolke, niedergestreckt, platzregen, regen-gewitter, regengewitter, regenstrahl, schlagwetter, strichgewitter, sturmdonner, ungewitter, unwetter, wassergewitter, wasserstrahl, wasserstrahle, wasserstreich, wettergeleuchten, wettergeleuchtet, wetterleuchte, wetterleuchten, wetterleuchteten, wetterläuten, wettern, wetterregen, wetterschein, wetterschlag, wetterstrahl, wetterstreich, wetterwolke, wildfeuer, wildstürmend, wirbelwind, wittergeleuchten
Hail	baumnuss, baumnüs, donnerstein, ei, eishagel, eisklumpe, eisschelle, eisstein, eisstück, enteneier, erbe, faust, faustgroß, gehageln, graupel, gänseeier, hagel, hagel-, hagel-regen, hagel-wetter, hagelaufstieg, hagelgewitter, hagelkorn, hageln, hagelregen, hagelregenwetter, hagelschauer, hagelschlag, hagelstein, hagelsturm, hagelsturz, hagelstück, hagelunwetter, hagelwolke, haselnuss, haselnuß, haselnüsse, henneneier, hühnerei, hühnereier, kiesel, kieseln, kieselregen, kieselschlag, kieselschloßen, kieselstein, kieselwetter, nuß, nüsse, schlagwetter, schnellkugel, schoßkeule, steinschlag, streifkiesel, strich, strichgewitter, taubeeier, taubenei, taubeneie, taubeneier, taubeneigroß, verhageln, walnuss, walnussgröße, walnüs, walnüss, welsch, welschen, welschen-nüsse, zerhageln
Wind	braus, brausen, brausend, böe, luftmasse, lüftlein, nord-ost, nord-ost-wind, nord-osten, nord-west, nord-west-wind, nord-westen, nordastwind, norden, nordost, nordosten, nordostwind, nordwesten, nordwestwind, nordwind, nordöstlich, nördlich, orkan, orkansturm, ostwind, schneegeäst, schneegeäst, schneesturm, starkwind, sturm, sturm-balken, sturmbalken, sturmdonner, sturmwetter, sturmwind, sturmzeichen, stürmen, stürmend, stürmisch, süd, süd-ost, süd-ost-süd, süd-ost-wind, süd-süd-ost, süd-west-wind, süd-westen, süd-wind, südost, südostwind, südwest, südwesten, südwind, südwärts, unterwind, vorüberziehend, wehen, west, west-nord-west, west-süd-west, west-südwind, west-wind, westlich, westwind, wind, windbrause, windbrausen, windbö, windböe, winde, winden, windfahne, windig, windrauschen, windstill, windstille, windstoß, windsturm, windstöße, windwirbel, wirbel, wirbelsturm, wirbelwind, östlich
Rain	donnerregenwetter, ergießen, ergießung, ergoss, feuerregen, flut, fromalwetter, frühlingwetter, frühlingwitterung, geregnet, gewaltwasser, gewalzt, gewitterregen, gewitterschauer, grobwetter, grundregen, gussregen, hagel-regen, hagelregen, hagelregenwetter, hochwasser, hochwetter, landregen, nass, naß, niederschlag, nieselregen, pegel, platzregen, prassel, regen, regen-gewitter, regenbogen, regenfall, regengewitter, regenguss, regenguß, regengüssen, regenmasse, regenschauer, regenschauerwetter, regenstrahl, regenwasser, regenwetter, regenwolke, regnen, regnerisch, reißend, schauer, schlagregen, schlagwetter, schlamm, schwefel, sindfluth, springflut, sprührege, sprühregen, spülung, starkregen, strich, strichgewitter, strichregen, strom, strömen, strömung, sündflut, tropfen, verschemung, wasser, wasser-wolke, wasserergüsse, wasserfluss, wasserflut, wassergefahr, wassergewitter, wasserguss, wasserguß, wassergüsse, wassergüssen, wassermasse, wassermenge, wassernebel, wassernot, wasserrinne, wasserspiegelbogen, wasserstrahl, wasserstrahle, wasserstreich, wasserstrom, wassersturz, wassersäule, wasserwirbel, wegschwemmen, wetterregen, wolkenbruch, überfloss, überflutung, überschwemmen, überschwemmung
Snow	alpenschne, beschneien, eingeschneit, schnee, schneeflocke, schneegeäst, schneegeäst, schneesturm, schneewasser, schneeweiß, schneewetter, schneewolke, schneien, triebschnee

Table B2. Vocabulary for semantic identification of hail in the corpus used (silver labels).

Group	Description	Complete vocabulary
C1	Direct and specific: Concrete measurements or physical references (see also table A6).	6 zentnern · apfel · äpfel · baumnuss · baumnüsse · bleikugel · bohne · bohnen · brot · daumen · dotter von eiern · drei männerfäuste · ei · eier · enteneier · erbse · erbsen · faust · fingerlang · gänseei · groschen · groß wie bälle · haselnuss · hennenei · hühnerlei · hühnerlei · hühnerlei · italienischen nuss · kastanien · kieselstein · kleine kürbisse · kugel im feuerrohr · kürbis · lot · märmel · männerfaust · nuss · nüsse · orangen · pfund · pfundstein · schnellkugel · schoßkeule · schoßkugel · sperling · spielkugeln · steingröße · taler · taubenei · taubeneier · walnuss · walnussgröße · walnüsse · welsch nuss · welschen nüsse · werben und berkelinge
C2	Indirectly: Indications of damage or effects.	beschädigt · erschlag · erschlagen · feldfrüchte ausgeschlagen · fenster eingeschlagen · flurschaden · früchte beschädigt · große verwüstung · großer hagelschlag · hagel abgeschlagen · hagelschaden · hagelschlag · häuserschaden · eingeschlagen · geschlagen · geschädigt · kieselschlag · mäßiges unheil · niedergeschlagen · ruiniert · schadete · schaden · schädigung durch hagel · schädlich · schädliche hagelwetter · schlug · strich · verdorben · verhagelt · verwüstet · weggeschlagen · wein von den stöcken schlug · wetterschlag · zerknickte · zerstörten · zerschlagen · zerschmettert · zerrieb · zerschlug
C3	Relative (qualitative): Linguistic intensifiers or names for strong hail, intensity and severity.	böse [hagelwetter] · furchtbaren [hagel] · furchtbares [hagelwetter] · gewaltigem [hagel] · gewaltigen [schloßen] · gewaltiger [hagel] · grausam [hagel] · grausam [gekieselt] · grausamen [hagel] · grausamer [hagel] · grausames [hagelwetter] · große [hagel] · große [kiesel] · große [schloßen] · großem [hagel] · großen [hagels] · großen [kieselsteinen] · großen [schloßen] · großen [steinwürfen] · große [steine] · großes [hagelwetter] · [hagel] von erstaunlicher größe · hartes [kieselwetter] · heftig [hagel] · heftiger [hagelwetter] · [schloßen] so groß · [schloßen] ungeheuer groß · schreckliche [hagelwetter] · schrecklichem [hagel] · schrecklichen [hagelwetter] · schrecklicher [hagel] · schreckliches [hagelwetter] · schwere [hagelwetter] · schwere [kiesel] · schweres [hagelwetter] · schwersten [hagelwetter] · stark [gehagelt] · starke [schloßen] · starkem [hagel] · starken [kieseln] · starken [schloßen] · starker [hagel] · starkes [hagelwetter] · übergroßer [hagel] · ungewöhnliche [schloßen] · ungewöhnlicher größe · verheerende [hagelwetter] · verheerender [hagel] · viele [schloßen] · große [hagelsteine] · eisengleichen [kugeln] · großer menge · ungemeiner größe · große ungeheure [schloßen] · grausame [hagel] · großer und schrecklicher vereister [hagel] · erstaunlichem [hagel] · solch einer größe · schrecklich [zu hageln] · heftigen [kieselwetter]

Table B3. Vocabulary for the semantic identification of rain in the corpus used (silver labels).

Group	Description	Complete vocabulary
C1	Direct and specific: Direct, measurable, or quantitatively determinable information.	(no direct information on specific measurements available)
C2	Indirect: Indications of damage or effects.	weggerissen · wasser riss · erde fortriss · vernichtete · verwüstete · weggeschwemmt · weggespült · ausgewaschen · überschwemmte · überflutete · überspült · verschlemmte · verschemmung · weggeführt · bäume geführt · weggetrieben · wasser riss häuser · keller voll · brücke zerbrochen · schaden mühlen · mühle verdarb · über deiche · nicht ernten · heu verfault · feldfrüchte schaden · land verschoben · schlamm bedeckt · ertranken · ersäufen · versoffen · untergingen · starken regen schaden · wassergefahr · wassernot · wolkenbruch großer gefahr
C3	Relative (qualitative): Linguistic descriptions of intensity, amount of water, height, or extent.	hochwasser · überschwemmung · flut · wasserfluten · feldfluten · gewässer · großes [wasser] · viel [wasser] · großmächtiges [wasser] · gewaltige [wassermenge] · regenmassen · schreckliches [wasser] · [wasser] so stark · gewaltigem [regen] · wilde [wasser] · [wasser] grausam wuchsen · hoch stand · sehr hoch angelaufen · hohes [wasser] · [wasser] hoch gestanden · steigen des [wassers] · [bäche] anschwellen · halben mann hoch · [wasser] zwei ellen hoch · [bäche] über ihre ufer gingen · [Fluss/Ort] ausgetreten · [Radaune] ausbrach · [flüsse] ergossen · schrecklicher [ergießung] · übergelaufen · gänzlich überströmt · zusammenfließende [wasser] · sündflut · unpassierbar · unter [wasser] · kahn fahren · schwammen · flößte

Table B4. Vocabulary for semantic identification of wind in the corpus used (silver labels).

Group	Description	Complete vocabulary
C1	Direct and specific: Direct, measurable, or quantitatively determinable information.	(no direct information on specific measured values available)
C2	Indirect (damage indicators): Effects on objects, buildings, vegetation, or infrastructure.	gehobelt · bäume gebrochen · bäume riss · bäume entwurzelte · linden zerstörte · schutzmauern umstürzte · türme eingerissen · häuser geworfen · mauern verdarben · häuser abdeckte · kirchenfenster übel zurichtete · dächer abgetragen · mauer eingestürzt · schiffe vergangen · fenster zerschlagen · getreide niedergeschlagen · wurzel gerissen · winde gewütet · großen schaden · schaden häusern
C3	Relative (qualitative): Linguistic intensifiers or names for strong winds, intensity and severity.	sturm · sturmwind · sturmwinden · stürmte · sturmdonner · sturmwetter · windsturm · gewittersturm · gewalt des [windes] · großer stärke · starker wind · sehr starken [wind] · heftiger wind · böiger [wind] · heftigen [böen] · mächtigen [wind] · schrecklicher [wind] · ungestümer [wind] · grausamer [wind] · stürmischer [wind] · orkan · tornado · wirbelwind · starker reißender [wind]

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

References

- Bayerische Akademie der Wissenschaften: Der Baierischen Akademie der Wissenschaften in München meteorologische Ephemeriden, Der Baierischen Akademie der Wissenschaften in München meteorologische Ephemeriden, 8, type: Zeitschriftenband, 1789.
- Beck, H. E., McVicar, T., Vergopolan, N., Berg, A., Lutsko, N., Dufour, A., Zeng, Z., Jiang, X., Dijk, A. v., and Miralles, D.: High-resolution (1 km) Köppen-Geiger maps for 1901–2099 based on constrained CMIP6 projections, *Scientific Data*, 10, 724, <https://doi.org/10.1038/s41597-023-02549-6>, 2023.
- Brázdil, R., Dobrovolný, P., Luterbacher, J., Moberg, A., Pfister, C., Wheeler, D., and Zorita, E.: European climate of the past 500 years: new challenges for historical climatology, *Climatic Change*, 101, 7–40, <https://doi.org/10.1007/s10584-009-9783-z>, 2010.
- Brázdil, R., Chromá, K., Valášek, H., Dolák, L., and Řezníčková, L.: Damaging hailstorms in South Moravia, Czech Republic, in the seventeenth to twentieth centuries as derived from taxation records, *Theoretical and Applied Climatology*, 123, 185–198, <https://doi.org/10.1007/s00704-014-1338-1>, 2016a.
- Brázdil, R., Chromá, K., Valášek, H., Dolák, L., Řezníčková, L., Zahradníček, P., and Dobrovolný, P.: A long-term chronology of summer half-year hailstorms for South Moravia, Czech Republic, *Climate Research*, 71, 91–109, <https://www.jstor.org/stable/24897492>, 2016b.
- Brönnimann, S., Martius, O., Rohr, C., Bresch, D. N., and Lin, K. E.: Historical weather data for climate risk assessment, *Annals of the New York Academy of Sciences*, 1436, 121–137, <https://doi.org/10.1111/nyas.13966>, 2019.
- Burgdorf, A.-M.: A global inventory of historical documentary evidence related to climate since the 15th century, preprint, Proxy Use-Development-Validation/Historical Records/Decadal-Seasonal, <https://doi.org/10.5194/cp-2021-165>, 2021.
- Camuffo, D., Cocheo, C., and Enzi, S.: Seasonality of instability phenomena (hailstorms and thunderstorms) in Padova, northern Italy, from archive and instrumental sources since AD 1300, *The Holocene*, 10, 635–642, <https://doi.org/10.1191/095968300666845195>, 2000.
- Cutter, S. L.: The Changing Nature of Hazard and Disaster Risk in the Anthropocene, *Annals of the American Association of Geographers*, 111, 819–827, <https://doi.org/10.1080/24694452.2020.1744423>, 2021.
- Deutscher Wetterdienst: Beobachterhandbuch für Wettermeldestellen des synoptisch-klimatologischen Mess- und Beobachtungsnetzes, Vorschriften und Betriebsunterlagen 3 (VuB 3 BHB), Deutscher Wetterdienst, Offenbach am Main, 2014.
- Deutscher Wetterdienst (DWD): Offizielle Webseite des Deutschen Wetterdienstes, <https://www.dwd.de/>, 2025.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, <http://arxiv.org/abs/1810.04805>, arXiv:1810.04805 [cs], 2019.
- Diodato, N., Ljungqvist, F. C., and Bellocchi, G.: A millennium-long reconstruction of damaging hydrological events across Italy, *Scientific Reports*, 9, 9963, <https://doi.org/10.1038/s41598-019-46207-7>, 2019.
- Ehrmantraut, A.: Historical German Text Normalization Using Type- and Token-Based Language Modeling, <https://doi.org/10.48550/arXiv.2409.02841>, arXiv:2409.02841 [cs], 2025.
- Eichler, E. and Walther, H., eds.: Historisches Ortsnamenbuch von Sachsen, no. Bd. 21 in Quellen und Forschungen zur sächsischen Geschichte, Akademie, Berlin, ISBN 978-3-05-003728-8, oCLC: ocm54419084, 2001.
- Erfurt, M., Skiadaresis, G., Tisdeman, E., Blauhut, V., Bauhus, J., Glaser, R., Schwarz, J., Tegel, W., and Stahl, K.: A multidisciplinary drought catalogue for southwestern Germany dating back to 1801, *Natural Hazards and Earth System Sciences*, 20, 2979–2995, <https://doi.org/10.5194/nhess-20-2979-2020>, 2020.
- Ernst, P.: Deutsche Sprachgeschichte: eine Einführung in die diachrone Sprachwissenschaft des Deutschen, no. 2583 in utb Sprachwissenschaft, Facultas, Wien, 3. auflage edn., ISBN 978-3-8252-5532-9, 2021.

- European Severe Storms Laboratory (ESSL): European Severe Storms Laboratory - Official Website, <https://www.essl.org/cms/>, 2025.
- Giordani, A., Kunz, M., Bedka, K. M., Punge, H. J., Paccagnella, T., Pavan, V., Cerenzia, I. M. L., and Di Sabatino, S.: Characterizing hail-prone environments using convection-permitting reanalysis and overshooting top detections over south-central Europe, *Natural Hazards and Earth System Sciences*, 24, 2331–2357, <https://doi.org/10.5194/nhess-24-2331-2024>, 2024.
- Glaser, R.: *Klimageschichte Mitteleuropas : 1200 Jahre Wetter, Klima, Katastrophen*, Primus, Darmstadt, Sonderausg. 2013 3., unveränd. Aufl. edn., ISBN 3-86312-350-6, http://deposit.d-nb.de/cgi-bin/dokserv?id=4215166&prov=M&dok_var=1&dok_ext=htm&dhttp://www.ub.unibas.ch/tox/IDSBB/006075550/PDF, 2013.
- Glaser, R.: HISKLID: Historische Klimadatenbank, <https://freidok.uni-freiburg.de/proj/3720>, 2014.
- Grotefend, H.: *Taschenbuch der Zeitrechnung des deutschen Mittelalters und der Neuzeit*, Hahn, Hannover, 14. Aufl. edn., ISBN 978-3-7752-5177-8, 2007.
- Grzega, J.: Climatic Conditions and Lexis: Some Diachronic Notes on Weather-Related Words in English and Other European Languages, *Transactions of the Philological Society*, 120, 320–331, <https://doi.org/10.1111/1467-968X.12243>, 2022.
- Gudd, M.: *Gewitter und Gewitterschäden im südlichen hessischen Berg- und Beckenland und im Rhein-Main-Tiefland 1881 bis 1980*, Ph.D. thesis, Johannes Gutenberg-Universität Mainz, <https://doi.org/10.25358/OPENSOURCE-3426>, 2004.
- Hawkins, E., Brohan, P., Burgess, S. N., Burt, S., Compo, G. P., Gray, S. L., Haigh, I. D., Hersbach, H., Kuyper, K., Martínez-Alvarado, O., McColl, C., Schurer, A. P., Slivinski, L., and Williams, J.: Rescuing historical weather observations improves quantification of severe windstorm risks, *Natural Hazards and Earth System Sciences*, 23, 1465–1482, <https://doi.org/10.5194/nhess-23-1465-2023>, 2023.
- Huang, S.-Y., Wu, S.-Y., Chen, Y.-J., Tsai, R. T.-H., and Fan, I.-C.: Climate event classification based on historical meteorological records and its presentation on a Spatio-Temporal research platform, *Digital Scholarship in the Humanities*, 37, 1022–1032, <https://doi.org/10.1093/lc/fqab099>, 2022.
- Hulton, F. and Schultz, D. M.: Climatology of large hail in Europe: characteristics of the European Severe Weather Database, *Natural Hazards and Earth System Sciences*, 24, 1079–1098, <https://doi.org/10.5194/nhess-24-1079-2024>, 2024.
- Kahraman, A., Kendon, E. J., and Fowler, H. J.: Climatology of severe hail potential in Europe based on a convection-permitting simulation, *Climate Dynamics*, <https://doi.org/10.1007/s00382-024-07227-w>, 2024.
- Lamont, J. v.: *Gewitterbeobachtungen von Ansbach, Burglengenfeld, Dillingen, Gunzenhausen, Hof, Hohenpeißenberg, Neustadt a.d.A., Würzburg von 1842 u. 1843*, *Annalen für Meteorologie, Erdmagnetismus und verwandte Gegenstände*, 1844, 147–167, <https://www.digitale-sammlungen.de/en/view/bsb10133287>, 1844.
- Lenke, W.: Klimadaten von 1621-1650 Nach Beobachtungen des Landgrafen Hermann IV. von Hessen (Uranophilus Cyriandrus), *Tech. Rep. 63*, Deutscher Wetterdienst, Offenbach am Main, https://www.dwd.de/DE/leistungen/pbfb_verlag_berichte/pdf_einzelbaende/63_pdf.pdf?__blob=publicationFile&v=3, 1960.
- Luterbacher, J., Allan, R., Wilkinson, C., Hawkins, E., Teleti, P., Lorrey, A., Brönnimann, S., Hechler, P., Velikou, K., and Xoplaki, E.: The Importance and Scientific Value of Long Weather and Climate Records; Examples of Historical Marine Data Efforts across the Globe, *Climate*, 12, 39, <https://doi.org/10.3390/cli12030039>, 2024.
- Martius, O., Hering, A., Kunz, M., Manzato, A., Mohr, S., Nisi, L., and Trefalt, S.: Challenges and Recent Advances in Hail Research, *Bulletin of the American Meteorological Society*, 99, ES51–ES54, <https://doi.org/10.1175/BAMS-D-17-0207.1>, 2018.
- Preußisches Meteorologisches Institut: *Ergebnisse der Gewitter-Beobachtungen : im Jahre .. ; 1892/94, Ergebnisse der Gewitter-Beobachtungen, 1892/94*, backup Publisher: Bayerische Staatsbibliothek, München Type: Zeitschriftenband, 1897.

- 780 Punge, H. and Kunz, M.: Hail observations and hailstorm characteristics in Europe: A review, *Atmospheric Research*, 176-177, 159–184, <https://doi.org/10.1016/j.atmosres.2016.02.012>, 2016.
- Reitzenstein, W.-A. F. v.: *Lexikon fränkischer Ortsnamen. Herkunft und Bedeutung. Oberfranken, Mittelfranken, Unterfranken*, C.H. Beck, München, ISBN 978-3-406-59131-0, 2009.
- Reitzenstein, W.-A. F. v.: *Lexikon schwäbischer Ortsnamen. Herkunft und Bedeutung. Bayerisch-Schwaben*, C.H. Beck, München, ISBN 785 978-3-406-65208-0, 2013.
- Rohr, C.: *Der Umgang mit Naturkatastrophen im Mittelalter*. In: Christian Rohr (Koordinator), *Krisen, Kriege, Katastrophen. Zum Umgang mit Angst und Bedrohung im Mittelalter*. 9. Interdisziplinäre Ringvorlesung des Interdisziplinären Zentrums für Mittelalter-Studien, Salzburg, Wintersemester 2009/10, type: Other, 2009.
- Sakaji, H. and Kaneda, N.: Indexing and Visualization of Climate Change Narratives Using BERT and Causal Extraction, in: 2023 IEEE International Conference on Big Data (BigData), pp. 5674–5683, <https://doi.org/10.1109/BigData59044.2023.10386320>, arXiv:2408.01745 [cs], 2023.
- 790 Schätz, F.: *Voraussetzungen und Grenzen der Auswertung klimarelevanter Informationen historischer Textquellen mit Hilfe von Automatisierungsprozessen*, Monographie, Universität Freiburg, Freiburg im Breisgau, <https://freidok.uni-freiburg.de/data/244079>, 2023.
- Schätz, F. and Glaser, R.: *Historical Climate Observations of Thunderstorms and Hail in Central Europe (1000–1900)*, <https://doi.org/10.60493/834BD-MWW13>, 2025.
- 795 Stahl, K., Kohn, I., Blauhut, V., Urquijo, J., De Stefano, L., Acácio, V., Dias, S., Stage, J. H., Tallaksen, L. M., Kampragou, E., Van Loon, A. F., Barker, L. J., Melsen, L. A., Bifulco, C., Musolino, D., de Carli, A., Massarutto, A., Assimacopoulos, D., and Van Lanen, H. A. J.: Impacts of European drought events: insights from an international database of text-based reports, *Natural Hazards and Earth System Sciences*, 16, 801–819, <https://doi.org/10.5194/nhess-16-801-2016>, 2016.
- 800 Taszarek, M., Allen, J., Púčik, T., Groenemeijer, P., Czernecki, B., Kolendowicz, L., Lagouvardos, K., Kotroni, V., and Schulz, W.: A Climatology of Thunderstorms across Europe from a Synthesis of Multiple Data Sources, *Journal of Climate*, 32, 1813–1837, <https://doi.org/10.1175/JCLI-D-18-0372.1>, 2019.
- Tornado and Storm Research Organisation (TORRO): *The TORRO Hail Intensity Scale (H-Scale)*, <https://www.torro.org.uk/research/hail/hscale>, 2025.
- 805 Webersinke, N., Kraus, M., Bingler, J. A., and Leippold, M.: ClimateBert: A Pretrained Language Model for Climate-Related Text, <https://doi.org/10.48550/arXiv.2110.12010>, arXiv:2110.12010 [cs], 2022.
- Wilks, D. S.: *Statistical methods in the atmospheric sciences: an introduction*, Elsevier, Amsterdam, 4th ed edn., ISBN 978-0-12-816527-0 978-0-12-815823-4, 2019.
- Zhou, B., Zou, L., Mostafavi, A., Lin, B., Yang, M., Gharaibeh, N., Cai, H., Abedin, J., and Mandal, D.: VictimFinder: Harvest-
810 ing rescue requests in disaster response from social media with BERT, *Computers, Environment and Urban Systems*, 95, 101824, <https://doi.org/10.1016/j.compenvurbsys.2022.101824>, 2022.