

Response to Reviewer 1

We would like to thank the reviewer for their constructive criticism. We are very pleased that the strengths of the method and the results have been recognised.

1 — Classification and uncertainties

1. We argue that uncertainties can only be quantified on the basis of a measurement standard. For historical textual sources that record convective events, the usual physical references, such as historical measurements, tree rings or varves, are lacking. Even if two or more trained analysts reach similar conclusions when evaluating historical texts, it is impossible to say with certainty whether these conclusions accurately reflect meteorological reality. In climatology, where measurement is paramount, it is often criticised that the physical truth behind a text cannot be negotiated intersubjectively. We have therefore taken the opposite approach, asking ourselves whether we can develop a general, transparent, verifiable and internally consistent procedure based exclusively on the text itself.

Our classification procedure is based on the text itself and the classification scheme. This enables us to achieve a consistent classification of thunderstorm and hail events in terms of content. Additionally, we have introduced the evidence class and the silver labels. These document interpretative inaccuracies for all events equally. Thus, the uncertainty of text interpretation is recorded at the point at which it arises.

We validate the quality of the classification for each quote using the evidence classes along the impact pathways (C1: direct magnitude; C2: damage; C3: qualitative assessment). As the significance of the evidence classes and silver labels is not immediately apparent, we have developed a confidence index using the reviewers' comments.

A clear pattern emerges: hail events and light or severe/extreme thunderstorms in classes TS1 and TS3 have a stable classification, whereas thunderstorms in class TS2 are subject to greater uncertainty. To illustrate this more clearly, we have made the following changes to the manuscript:

1. We have included a detailed example of an annotated quote with evidence classes and a silver label.
2. We have introduced a confidence index and conducted an evaluation over the entire period.
3. The uncertainties have been highlighted more clearly.
4. The consistency checks introduced during the revision identified annotation errors in 25 of the 6,999 quotes (0.4 per cent). These were reclassified and all statistics were recalculated.
5. Discussion: Reference to the asymmetric interpretation (H2/TS3 supported vs. TS2 transitional class)

2 — Source type and representativeness

- Background: 99.1 per cent of the quotes are taken from HISKID2, a historical climate database that has been compiled since the 1980s in order to record temperature and precipitation data. HISKID2 was not specifically compiled for the purposes of this study, nor was it designed for research into thunderstorms or hail. Of the total 51,833 quotes in HISKID2, 44,834 mention neither thunderstorms nor hail. Therefore, a selective focus on extreme weather at corpus level was ruled out from the outset. Rather than searching for dramatic events, we extracted convective quotes from a corpus that had been compiled for other purposes.

We considered adjusting for source type as a model feature, but subsequently rejected this idea. The reasons are:

- Differences in source type constitute information in themselves. Furthermore, Weather Records disproportionately record light hail (H1, standardised residual $R = +10.4$), while chronicles disproportionately record heavy hail (H2, $R = +4.8$; see Figure 13). Uniform weighting would level out precisely those records of damage that define H2.
- Furthermore, a source type feature would tell the model more about where a quote comes from than what it describes. This would limit the model's generalisability.
- However, the model classifies reliably across all source types. The patterns specific to each source type are part of the text. The test dataset comprises 20 per cent of the quotes and all source types are represented. As can be seen in Tables 6 to 8, the F1 scores are very high.
- Analysis of the influence of source type on classification revealed only a secondary effect. We have emphasised these results more strongly, as they might otherwise be misinterpreted.

We have made the following amendments and additions to the manuscript:

1. Section 'Data and sources': paragraph 'Corpus as a sample' added (99.1% from HISKLID2, originally created for temperature and precipitation, not for thunderstorms/hail)
2. Results (section 'Influence of source types on the classes') and discussion: Classification is carried out on a per-quote basis, independently of source type. Effect sizes and residuals quantify the influence of source type. The discussion notes that it is only by considering all sources together that plausible results can be obtained.
3. A clearer presentation in the discussion of the secondary influence of source type.

3 — Geographical and linguistic transferability

We distinguish between the transferability of the framework and the transferability of the results.

The framework is language-independent. This means that both the framework and the methodology are transferable. The model itself, which was trained on German, is not transferable and would need to be retrained for each target language. The language-specific components, namely normalisation and the silver label word lists, must also be recreated for each target language using the same corpus-based procedure.

For example, normalisation takes into account individual, cultural, and regional characteristics. This is particularly important for the German language. The model learns sentence structure and style from the examples, which are not removed during normalisation. This means that we are dealing with a historical and cultural unit within the German-speaking world. The situation in other linguistic regions is beyond the scope of this study.

The results (seasonal phase, intensity distribution and stability at the century level) only apply to the Central European corpus and its climatic region, and cannot be generalised. Within this region, the corpus is heterogeneous in terms of the factors that determine the quality of the evidence, including four linguistic stages, nine source types, a period of approximately 800 years, and both confessional spheres. Transferability across linguistic boundaries has not been tested. We regard this as the main limitation of generalisability and the next step to take.

Changes to the manuscript:

Conclusion: Transferability is identified as a linguistic limitation. The classification procedure is language-independent, but its language-specific components must be recreated for each target language. The next step is to apply it to comparable corpora in other languages, such as French and English.

4 — Value Added by the Model

The aim is controlled automation, not pattern recognition. Determining the intensity of convective events based on quotes is laborious and requires specialist knowledge. Classifying the nearly 7,000 quotes alone took more than 112 working days.

The intended outcome is to reproduce the classification, which ensures greater efficiency by applying the trained method to corpora that are orders of magnitude larger than would be possible manually. It has been demonstrated that the model can be trained using linguistic features. The model is most uncertain where the scheme itself is least unambiguous, particularly at the boundary between TS2 and TS3. This also provides insight into the quality of the schema and the corpus.

The decision criteria can be verified explicitly. The word lists in Appendices B1–B4 specify which expressions correspond to which evidence class. The protocol (see Figures 4 and 6 and Tables A1–A6) assigns these expressions to intensity classes according to transparent rules. As each classification follows rules that can be verified retrospectively, the automation process remains controllable.

By 'new contexts', we mean whether the model generalises beyond the annotated examples. The model can correctly classify unknown quotes, as demonstrated by the F1 scores. For training purposes, the quotes were divided into three sets in the following ratio: training: validation: test = 70:10:20. This means that 20 per cent of the quotes are unknown to the model. Tables 6 to 8 show how well the model can correctly classify the remaining 20 per cent of quotes. This reveals linguistically novel combinations, sentence lengths, word orders and contextualisations. At first glance, it is not immediately obvious that the model automatically learns text patterns. These text patterns reflect source types, turns of phrase and word orders, among other things, and become part of the model — or, more precisely, part of the optimisation of the weights.

Changes to the manuscript:

Discussion: The remaining model uncertainty is linked to the uncertainty of the schema (the TS2/TS3 boundary; the misclassifications reflect the confidence index).

5 — Circularity and external validation

The validation process is not circular. Classification is carried out in stages and the final intensity class is only assigned once the indicators (consequences, damage and fatalities) have been annotated independently (see Figure 6). The model is only trained on a subset of the corpus (the training set). The test set is separated from the training set prior to training. Consequently, the model cannot have prior knowledge of the quotes in the test set.

External validation at the individual event level is discussed below. Aggregate framing is made explicit (no single quote substantiates a convective event in itself; the evidence lies in the aggregate), and is limited in pre-instrumental periods, as we point out in the above section. We would add that the contrast between subjective historical sources and objective instrumental data for thunderstorm and hail events is sometimes overstated. Well into the 20th century, thunderstorms at DWD stations were recorded by trained observers who noted

thunder or lightning, and hailstone size was estimated visually using reference scales (e.g. peas, hazelnuts, walnuts and hen's eggs). This practice continues in modern severe weather reporting. The validation chain runs from observer to observer. The main difference lies in how the observations were recorded.

Changes to the manuscript:

1. In the 'Classification procedure' section of the Methods section, the stepwise classification is described. The intensity class is only assigned after the associated phenomena and impact pathways have been annotated (Figure 6).
2. Data and sources and methods: observer continuity between historical sources and DWD station observations. The 1961–1990 reference period is based on visual and auditory observations, in accordance with the DWD Observer's Handbook.

6 — Seasonal plausibility as evidence of physical validity

We agree that a single correspondence between historical and modern seasonal patterns would be weak evidence, as it could be attributed to cultural factors or distortions in data transmission. However, the argument here is based on several independent seasonal characteristics, each of which requires a separate explanation:

1. The temporal distribution of thunderstorms and hail differs: both exhibit seasonal cycles that coincide with the normal periods of the DWD. This typical seasonal cycle is evident throughout the entire period and across different linguistic and cultural regions (see also Figure 10). Even in years with limited documentation, this pattern remains discernible in the combined signal from thunderstorms and hail. To be culture-dependent, the signal would have to explain why the probability of a thunderstorm or hail event being recorded in writing is greater in summer, while also ruling out the physical explanation. It would have to vary depending on the phenomenon, historical period, and cultural region yet still apply equally to all types of source. The two phenomena also differ in their seasonal timing. The increase in hail events begins earlier in the year (in March or April), which is consistent with convective climatology and the timing of thunderstorm events. A reporting convention would need to provide onset dates specific to each phenomenon.
2. The reconstruction is consistent with three independent data series from different perspectives: the modern hail data from the DWD (instrumental measurements), Lenke (an independent historical source base) and Camuffo (written sources from a different linguistic area and a different climate zone).
3. The seasonal concentration is more pronounced in the higher-intensity classes (TS3 and H2) than in the lower-intensity classes. A transmission bias would be independent of intensity. The fact that the signal strengthens with increasing intensity is a distinguishing physical feature.
4. Cross-source consistency: The seasonal cycle remains consistent across the four linguistic stages (median pairwise $\rho = 0.82$ – 0.94 for thunderstorms) and the nine source types (median pairwise $\rho = 0.82$).
5. From the late 15th to the early 19th century, summer reports consistently outnumbered winter reports (see Figure 9). Over the course of this period, the sources changed almost entirely. The written language evolved from Early New High German to modern German, and the recorders shifted from chroniclers to early weather station observers. However, the seasonal balance remained unchanged. This is to be expected physically, as there are more thunderstorms and hailstorms in summer, and accordingly more records originate from this season. Only a reporting convention that kept the language unchanged throughout all these shifts could mimic this. However, we are not aware of any such convention.

6. The agreement with the DWD reference improves where the record is densest. For the period 1624–1654, for example, the correlations rise from 0.66–0.78 to 0.74–0.86, while the RMSE decreases. Unlike a real signal obscured by sparse data, a cultural artefact would have no reason to converge towards the modern physical reference as documentation density increases.
7. Severe hail in winter only occurs with the low frequency expected on meteorological grounds. In the case of winter thunderstorms, the higher proportions in narrative sources suggest a source-type effect. We refer to this as 'residual uncertainty' in the manuscript.

The varying evidence makes a physical signal more likely than a culturally determined signal.

Changes to the manuscript:

Results/Discussion:

1. H1↔DWD comparison ($p = 0.87\text{--}0.92$) and stratified analyses of seasonality by language stage and source type added.
2. Results/Discussion: H1/DWD result and differentiation between the two types of hail (H1 winter and transitional months, H2 summer)
3. Discussion: Lenke and Camuffo comparison categorised by climate zone (Hesse Cfb, Padua Cfa) as evidence of a split seasonal cycle across three climate zones ($p = 0.86 / 0.71$, Figure 16)

7 — Are we reconstructing climate, or how people described climate?

This question goes far beyond the scope of this study and has already been addressed by researchers such as Le Roy Ladurie. In this context, the treatise on historical theory by Tschiggerl, Walach and Zahlmann (2019) is also relevant as it explores the question of which historical reconstructions are actually possible through the use of historical sources.

Nevertheless, we would like to attempt to answer this question from the perspective of this study. The following findings suggest a physical, rather than purely linguistic or cultural, origin.

1. The class distribution follows the physical frequency distribution rather than narrative logic. The weakest class, TS1, dominates every month (53–71 %). If sources had primarily preserved memorable extremes, the reverse ordering ($TS3 > TS2 > TS1$) would be expected.
2. There is consistency with other independent observations. The reconstruction is consistent with three series of observations which differ in terms of language, climate zone, and the purpose for which they were documented. These are data from the DWD standard periods, Lenke, and Camuffo. If we were simply reproducing how people describe thunderstorms and hail events, we would expect to see greater variability.
3. The two hail classes, defined by hailstone diameter and damage, exhibit two meteorologically significant seasonal cycles. H1 corresponds to light hail recorded by DWD stations, dominating the winter and transitional months ($p = 0.87\text{--}0.92$). H2 corresponds to convective summer hail and accounts for 77–82 per cent of events from June to August.
4. The density of linguistic evidence aligns with the class definitions. H2 is defined by strict physical criteria, which is confirmed by the silver labels extracted independently of class assignment: every H2 quote contains at least one silver label, with 35.1 per cent containing a direct measurement (C1). This composition remains stable across the centuries (see Figure 11). By contrast, 46.9 per cent of quotes in

the threshold-free middle class TS2 lack a silver label. The texts provide verifiable physical data where the schema requires it. Where the descriptions remain non-specific, the confidence index indicates this rather than over-interpreting it.

5. Of the 51,833 quotes in HSKLID2, we only considered those that clearly identified a thunderstorm or hail event and had been manually annotated previously. It quickly became apparent that the range of language used to describe convective events is limited. This is also evident in the composition of the evidence classes, which we have developed into a confidence index. The composition of the individual classes remained remarkably stable throughout the entire period. Written observations consistently follow the same linguistic pattern. These patterns remain constant across all periods and source types, showing little variability in the choice of words used to describe the observations. Some descriptions contain precise details regarding the sequence of downdrafts and the onset of rain, which are typical of frontal thunderstorms.

The reverse conclusion also applies, of course (see above): a culturally determined signal would have to explain why thunderstorms and hail events are more likely to be recorded in writing in summer, while ruling out the physical explanation. It would have to vary depending on the phenomenon, era and cultural region, yet still apply equally to all types of sources.