

Response to Reviewer 2

We would like to thank the reviewer for reading the revised manuscript and providing a positive assessment. The remaining two points have been addressed as follows:

1. Climatological plausibility versus independent validation:

We agree that the correspondence between the reconstructed seasonal cycle and modern or historical reference series at an aggregated level provides evidence of a plausible climatic signal. However, this does not validate the classification of individual events. We have therefore consistently distinguished between these two levels of quality assessment throughout the manuscript.

The evidence classes (C1–C3) and the confidence index (see Sections 3.2 and 4.2) are the components that constitute validation, since only these can assess the classification at the level of individual measurements. Conversely, the comparison with the DWD normal periods and the reconstructions by Lenke (1960) and Camuffo et al. (2000) are referred to as plausibility checks.

2. Independence of the machine learning evaluation

The reviewer rightly asks whether quotations from the same source can appear in both the training and test sets. This is indeed the case: the split is only stratified by class, with no grouping by source, and is performed at the level of individual quotes. We have added a paragraph to Section 3.5 that explicitly states this, clarifying that the reported performance describes how consistently the models reproduce the classification scheme within this corpus. It should not be taken as evidence of generalisation to unseen sources.

We consider the risk of the models learning source-specific patterns to be limited, since the vocabulary used to describe thunderstorm and hail events is narrow and highly formulaic across all sources in the corpus. This is also what makes these phenomena suitable for text-based classification. However, given the source-specific reporting patterns shown in Section 4.4, we cannot rule out the possibility that the models have partly learned such patterns. Testing this would require a source-grouped evaluation on a larger corpus. We are currently creating such a corpus by automatically transcribing historical newspapers, which can be classified using the existing framework.