# Digital innovations in historical climatology: Classifying weather and climatic extremes and their impacts on societies using machine learning on written documents

MICHAEL KAHLE, RÜDIGER GLASER

University Freiburg, Faculty of Environment and Natural Resources, Department of Physical Geography, Germany; e-mail: michael.kahle@geographie.uni-freiburg.de, ruediger.glaser@geographie.uni-freiburg.de

ABSTRACT This article explores how digital innovations – particularly machine learning and natural language processing – can streamline and enhance workflows in historical climatology. Traditionally reliant on time-consuming manual analysis of historical documents, the field now benefits from modern digital tools at each research stage, from source discovery to publication. Focusing on classifying large, unstructured textual data, the study examines methods ranging from manual keyword searches and Bayesian models to advanced large language models. Using the tambora.org corpus, it extracts and categorizes references to weather extremes like thunderstorms and heavy rainfall and their impacts on mobility. The paper compares these approaches in terms of accuracy, resource demands such as runtime performance and memory, and their ability to interpret historical language. It argues that digital methods – especially AI – can transform the extraction and classification of climate data from historical texts, offering significant advantages by assisting researchers in historical climatology.

KEY WORDS historical climatology – mobility – machine learning – natural language processing

KAHLE, M., GLASER, R. (2025): Digital innovations in historical climatology: classifying weather and climatic extremes and their impacts on societies using machine learning on written documents. Geografie, 130, 3, 271–297.
https://doi.org/10.37040/geografie.2025.014
Received April 2025, accepted August 2025.

## 1. Introduction

In recent years, the use of artificial intelligence in science has tremendously increased with the application of machine learning and deep learning algorithms. It is used for weather forecasting (Bi et al. 2023; Yang et al. 2025), drought impact prediction (Stephan, Stahl, Dormann 2023), land surface forecasting (Wesselkamp et al. 2025), and simulating hurricane tracks (Bose, Pintar, Simiu 2023). It helps to reconstruct summer temperatures from tree ring data (Kuhl et al. 2024), assess the impacts of droughts (Madruga de Brito, Kuhlicke, Marx 2020) or floods (Kahle et al. 2022), or reconstruct ancient hollow roads from LiDAR data (Verschoof-van Der Vaart et al. 2025) as well as predict heat stress on street level (Briegel et al. 2025).

Using the classical workflow for classifying historical documents regarding climate information and its impact on societies has been and still is a time-consuming task (Brázdil et al. 2005; White et al. 2023). Therefore, applying digital innovations to the field of historical climatology could accelerate the entire workflow from finding sources to publishing the generated results.

To transform unstructured written sources into condensed and standardized climatological information, numerous work steps with different expert knowledge are required (Riemann et al. 2015). Efforts had been made to ease this process, not only by guiding scientists through this workflow, but also by assisting (or even replacing) them during each single step. This is mainly driven by digital innovations and machine learning methods. Figure 1 shows, from left to right, the successive steps and the typical tools and methods used either in a more classical way of working (lower brownish row) or the digital counterparts (upper blueish row), revolutionizing the way of working (Kahle 2025a).

Typical sources include chronicles, early newspapers, diaries, and many others. Nowadays, an increasing number of them are available online in digitized formats, which gives a broad range of users access to rare material. On the one hand, recent sources like social media content or online newspaper articles, allow the calibration of methods with modern instrumental measurement or reanalysis data. Hologa and Glaser analyzed the storm event 'Friederike' by comparing Twitter tweets with data from the Global Forecast System (Hologa, Glaser 2021). On the other hand, the pure numerical measurement data can be enriched with aspects of causes and effects and societal consequences by analyzing these texts, as performed for the 'Ahrtal' flood (Kahle et al. 2022).

The search for relevant material required manual effort in libraries or archives. Catalogs and finding aids rarely used climatic keywords, so researchers often had to undertake a laborious, page-by-page examination of documents in potentially distant physical locations. Today, scientists can access and search entire digitized collections, such as the Monumenta Germaniae Historica (MGH) from their

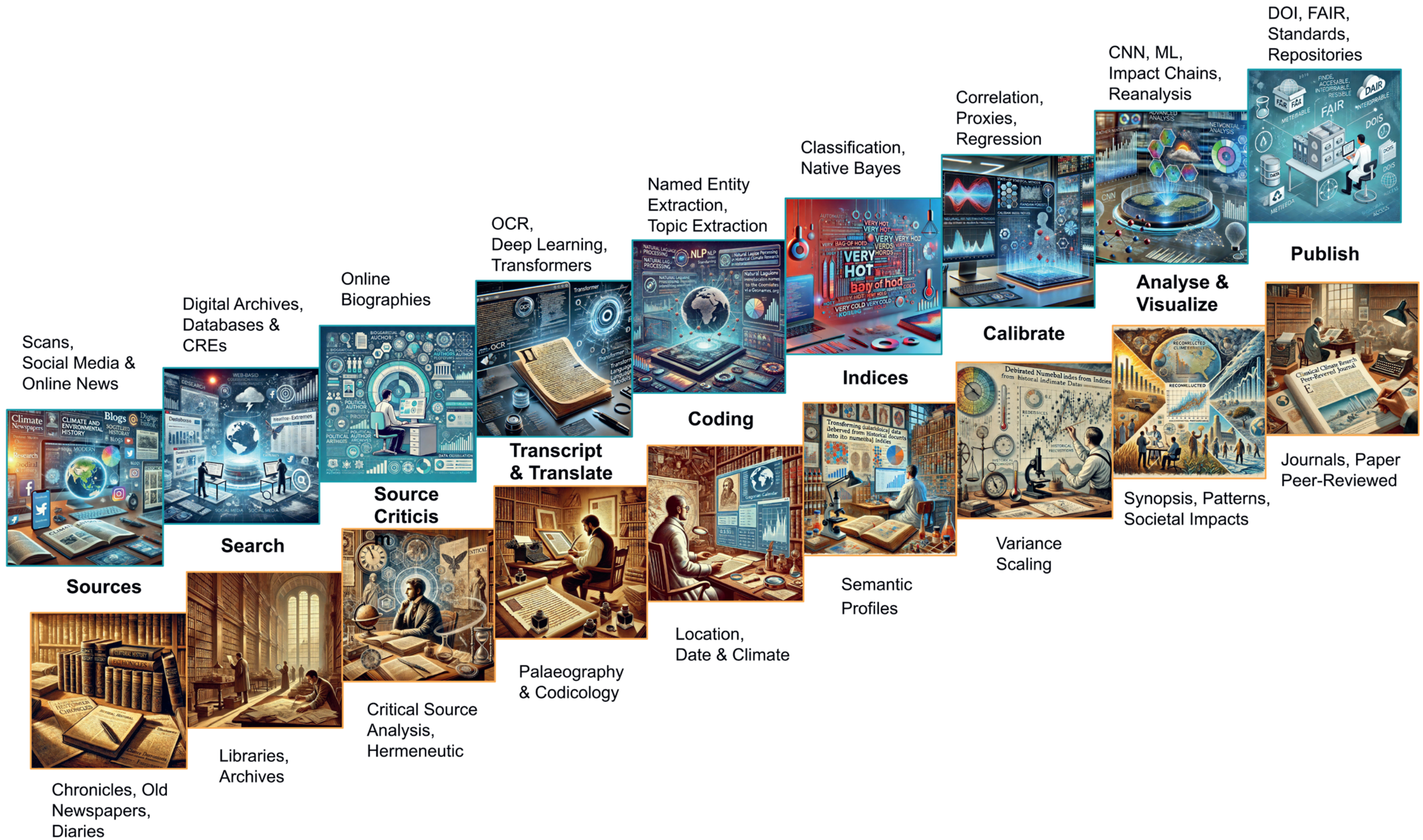


**Fig. 1** – Classical work steps (Riemann et al. 2015) and their digital counterparts in the field of historical climatology, created with Dalle3. The article provides a detailed explanation of each step and how it revolutionizes the workflow. Prompts used for the creation of the tiles are given in the Supplements.

desktops (Sahle, Vogeler 2013). Search engines like Google Books and platforms like archive.org (Gratzinger 2021, Kahle 2007) provide access to a wide range of digitized books and newspapers. Portals of libraries and archives facilitate access, such as the Bavarian State Library (Bayrische Staatsbibliothek) or the German Digital Library (Deutsche Digitale Bibliothek). Data that once was entered into a database or a Collaborative Research Environment (CRE) can easily be reused for other projects with different scopes.

Evaluating the reliability of historical documents by critical source analysis involves a thorough analysis of the author's background concerning the education, profession, eyewitness status, and intentions. It also covers the socio-historical context in which the document was created, as well as the living conditions, conflicts, and prevailing belief systems. This process aimed to understand the author's perspective, forming part of the Hermeneutic Circle, and required considerable background knowledge (Glaser 2013). While the core principles of source criticism remain, digital methods can aid in gathering and organizing information relevant to this process. Access to online biographical databases, historical encyclopedias, and digitized historical literature can expedite the collection of information about authors and their contexts. Moreover, formalization and partial automation of source criticism have been explored, potentially leveraging digital tools for analyzing textual features (Schätz 2023; WBIS 2005).

Transcribing handwritten or printed texts into a readable format required expertise in paleography and codicology to decipher various handwriting styles. Deciphering printed scripts like Fraktur was somewhat less challenging but still a very time-consuming and demanding task. The optical character recognition (OCR) technology enables the automatic conversion of scanned documents into digital text. Although OCR is highly effective for well-preserved printed materials, especially in Fraktur, it often struggles with documents that are faded, damaged, or handwritten. Advanced methods such as neural networks, trained on manually transcribed texts, offer potential solutions for transcribing challenging historical handwriting (Kahle et al. 2017; Marchant 2023). The complete diaries of Ignaz Speckle and Carl Hugo Hahn could be imported to tambora.org within weeks rather than years using Tesseract (Smith 2007; Hahn, Moritz 1998; Speckle 1968).

Translating texts written in languages such as Latin or historical German dialects into modern German or English is necessary for analysis and dissemination. This task required linguistic expertise and an understanding of historical language nuances. Machine translation, particularly using transformer architectures, has made significant strides and can be readily applied to some historical languages like Latin, and various web services are available (Vaswani et al. 2017). However, the translation of historical German dialects remains challenging due to their non-standardized nature and variability in spelling and grammar (Schätz 2023,

p. 39). While automatic translation can often provide a general understanding, human expertise is still required for accurate and nuanced interpretation.

Location, time and climatic incidents information is manually extracted and transferred to a standardized format for coding the text. The transition from qualitative text data to quantitative, numerical data forms a bridge between the humanities and natural sciences (Matuschek 2014). This process requires knowledge of historical place names, past calendar systems, and the development of thematic, hierarchical classification schemes (Grotefend 2015; Grässe, Benedict 1909). Modern algorithms in the field of natural language processing (NLP) allow the recognition of named-entities like persons, locations or date and time phrases (named-entity recognition NER) and many software libraries for various computer and human languages are freely available (Honnibal et al. 2020). The extracted place names could be linked to coordinates via geocoding services, such as geonames.org (Wick 2015). This procedure had been proved possible for digital newspapers as the hotspots of the flooding 2021 in Germany could be reproduced (Kahle et al. 2022), and even works for older, printed newspapers such as the 'Freiburger Zeitung'. Pattern matching and more sophisticated algorithms can extract and standardize temporal information, accounting for different calendar systems and historical dating conventions (Reingold, Dershowitz 2001). Regularly recurring topics can be automatically extracted using statistical analysis such as the term frequency-inverse document frequency (tf-idf), followed by a non-negative matrix factorization (NMF) and assigned via keywords, although the delivered topics do not always fit the desired focus (Khoo Khyou Bun, Ishizuka 2002; Hoyer 2004; Kahle et al. 2022).

Classes used for coding can differ in their characteristics: while some are independent and exclusive, others are linked (i.e., to parent classes), are ordered and follow given probability distributions or can occur simultaneously. Ships, wagons, railway, i.e. belong to the parent class "means of transportation" and thus form a nominal scale. The classes below "impacts on transportation" (Enabled, Possible, Restricted, Impossible) can be sorted, form an ordinal scale, and are often referred to as indices. Other classes, such as the standardized temperature index with its Gaussian distribution even fulfills the criteria for an interval scale, and the precipitation index with its natural point of zero validates for a ratio scale (Stevens 1946). Semantic profiles can help determine which phrases in a text belong to which index class. Therefore, the extracted phrases are ordered and groups can be formed by considering their frequency (Glaser 2013, Matuschek 2014). Modern machine learning approaches offer various classification methods. The scale of measurement, underlying distribution, and desired output as single- or multi-label reduce the selection of sophisticated algorithms. Some methods require (and take advantage of) the grammatical structure of sentences, whereas Bag-of-Words (BoW) models ignore them and therefore may adapt better to a changing language

(Harris 1954). The proper assignment to a class is ambiguous for some use cases, thus methods using fuzzy logic are promising (Bösmeier 2020). When further processing demands probabilities for each label, the naive Bayes classification is a good option. This paper compares the advantages and disadvantages of some of these algorithms by classifying several parameters describing the impacts of climatic hazards on mobility.

In a further step, the numerical data obtained from the classification into indices must be mapped to real climatic parameters with common measurement units. This calibration is performed using statistical transfer functions. Given that the monthly temperature values are gaussian distributed and the temperature index is an interval scale, a variance scaling can be used (Riemann 2010, p. 44ff). The authors map the instrumental-based standardized precipitation index (SPI) to the historical precipitation index (HPI) derived from historical documents to obtain the most severe drought periods in Central Europe since 1500 CE (Glaser, Kahle 2020). The classical statistical regression models were supplemented by machine learning algorithms up to deep learning approaches using neural networks. Steadily decreasing climate reconstructions or reanalysis data using different proxies allow cross-validation or even enhance the time period usable for calibration (Warren et al. 2024, Cook et al. 2015). For example, we compared the drought periods extracted since 1500 CE with drought indices derived from tree ring data (Glaser, Kahle 2020).

A deeper analysis of the obtained climate values typically includes the creation of time series or synoptic spatial representations like temperature fields or pressure maps. Longer periods with outstanding deviation from usual climate conditions can thus be identified and linked to extraordinary societal reactions or environmental impacts. The Little Ice Age (LIA), a period of temperatures below average, is, for example, linked to witch hunting and growing glaciers in Europe (Behringer 1995, Zasadni 2007). Dry and hot phases triggered a chain of impacts with decreasing harvest amounts, higher prices, and finally waves of migration in Germany (Glaser, Himmelsbach, Bösmeier 2017). Today's neural networks trained with measured climate data allow the interpolation of obtained data in time and space. Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), or Diffusion Models can reconstruct time series and maps that are consistent with physical principles (Wesselkamp et al. 2025; Raissi et al. 2024; Brunton, Kutz 2019; Bi et al. 2023). The Bayesian theorem had been used to combine reanalysis data with indices derived manually from text documents (Camenisch et al. 2021). Ongoing, interactive data visualizations can be used to explore existing coverage and relations and refine and fill in areas of weak data coverage.

The classical way to disseminate collected data and derived results is to publish papers in peer-reviewed journals and present them at scientific conferences. The raw, underlying data were often kept only locally and, in the best case, were available for reuse by other researchers only on request. The Open Science movement (Molloy 2011), which tambora.org is committed to, opens new ways of sharing knowledge. Results, but also raw data, methods, software, etc, are published in repositories more and more (De Vos et al. 2020; Thöle, Wegmann 2024). Established practices, likewise the FAIR principles, allow them to be Findable, Accessible, Interoperable, and Reusable (Lamprecht et al. 2020, Mons et al. 2020). Using Digital Object Identifiers (DOI) makes data and results citable, while legal constructs, as the Creative Common licenses, encourage and allow their use. Standards for basic entry types, likewise UTF8 for text, ISO 8601 for dates, WGS84 for coordinates, or SI-units for physical values, simplify their further processing. Standards for structured data collections as comma-separated-file format (CSV), JSON or GeoJSON data, and the format for Linked Paleo Data (LiPD) ease the reuse further (McKay, Emile-Geay 2016; Khider et al. 2019; Mitlohner et al. 2016; Butler et al. 2014). Repositories are publicly available and ensure long-term storage of contributed data. FreiDok plus from the university library Freiburg, as such a repository, is used to publish data from tambora.org in a special data series that provides files in both human-readable and machine-readable formats and secures copyright and rights of ownership (Glaser, Kahle, Hologa 2016).

Overall, every step of the workflow in historical climatology can benefit by using innovative digital methods. Particular advantages today include time savings in researching sources and their availability, as well as increased computing power, which allows complex statistical analyses of large amounts of data. For this paper, we will look further into several approaches to code and classify weather and climatic extremes and their impacts on society, and compare them.

## 2. Data and methods

### *2.1. Data*

The largest amount of data used in this study is from existing projects in the Collaborative Research Environment tambora.org (Riemann et al. 2015). The data before 1500 were already used and published in a study on the impacts of climatic hazards on mobility (Kahle, Glaser 2022, 2025). Sources from 1500 onwards had been extracted and classified using the same methods (Kahle 2025b). We coded the quotes regarding the four parameter groups of climatic hazard, the mean and zone of transportation, and the impact (Table 1). For the hazards 1489 quotes could be extracted, for the zones 1186 quotes, for the impacts 1083 quotes and for the means 958 quotes. Multiple classes can occur in the same quote (multi-label classification). Some additional sources have been added to increase the coverage for some periods and demonstrate the ability of advanced source collection methods,

**Table 1** – The parameter groups, their classes and their frequency regarding extremes and mobility

| Parameters | Hazards: 1,489 | Means: 958 | Zones: 1,186 | Impacts: 1,083 |
|---|---|---|---|---|
| Classes/Labels | Floods: 476<br>Storm: 365<br>Drought: 119<br>Cold: 463<br>Heat: 19<br>Snow: 209<br>Rain: 19 | Foot: 499<br>Ships: 294<br>Horse: 336<br>Wagon: 322<br>Sled: 55<br>Rail: 95<br>Plane: 14<br>Balloon:2 | Streets: 205<br>Rivers: 466<br>Bridges: 273<br>Lakes: 35<br>Sea: 198<br>Plains: 48<br>Rails: 70<br>Air: 9 | Impossible: 373<br>Restricted: 516<br>Possible: 29<br>Enabled: 303 |

likewise the diary of Ignaz Speckle (Speckle 1968) or recent online newspaper articles on floods and droughts (Kahle et al. 2022; Zhang, Glaser, Kahle 2025).

The following three quotes give an example of the content, its translation to English, the source, as well as the manual coding, and we will refer to them later when we review the results. More examples can be found in the mentioned study (Kahle, Glaser 2025).

Quote Id:7080

Anno domini MCCClxv Do was eyn so kalt wynter dat der Rijn eyn gantz veirdel jairs bestanden was. dat men vp sent Pauwels dach ètzo Riell ouer den Rijn ginck Ind tzo allen iiij wechen was groiss marckkk vp dem ijss, In dem seluen jair was eyn groiss sterffde.

*(In the year of our Lord MCCClxv, it was such a cold winter that the Rhine remained frozen for a whole quarter of a year. On St. Paul's Day, people crossed the Rhine at Riell, and for four weeks there was a large market on the ice. In that same year, there was a great mortality.)*

Koelhoff, Johann: Die Cronica van der// hilliger Stat Coellen: Sancta Colonia diceris. quia sanguine tincta// Sanctorum. meritis quo[rum] stas vndiq[ue] cincta (1306–1443)

Hazard: Cold, Mean: Foot & Wagon, Zone: River, Impact: Enabled

Quote Id:22036

Anno 1546. fiel der Winter im December ein mit ungewöhnlicher Kälte/ und bestunde nach dem ersten Tauen wieder biß Ostern 2 Monate lang/ daß alle Schiffreiche Wasser überfroren. In Nord=Ländern hat man auf der Ost=See aus Hollstein in Dänemarck/ aus Fühnen in Seeland/ von Coppenhagen nach Rostock aufm Eiß fahren und kommen können.

*(In the year 1546, winter arrived in December with unusual cold weather and, after the first thaw, lasted for two months until Easter, so that all navigable waters froze over. In northern countries, it was possible to travel on the ice across the Baltic Sea from Hollstein in Denmark, from Fühnen in Zealand, and from Copenhagen to Rostock.)*

Lehmann, Christian (1699): Christian Lehmanns Sen. weiland Pastoris zu Scheibenberg Historischer Schauplatz derer natürlichen Merckwürdigkeiten in dem Meißnischen Ober-Ertzgebirge/ : Darinnen Eine außführliche Beschreibung dieser gantzen gebirgischen und angräntzenden Gegend/ Nach ihrem Lager/ Gestalt/ Bergen/ Thälern ... enthalten / Weiland von dem seel. Autore ... zusammen getragen ... Nun aber Mit schönen Kupfern ... gezieret/ und ... aufgethan von dessen Hinterlassenen Erben

Hazard: Cold, Mean: Wagon, Zone: Sea, Impact: Enabled

Quote Id:22149

Ein, mit Waaren beladenes, Schiff auf der Trave ward in einem Gewitter vom Blitze angezündet, und verbrannte.

*(A ship loaded with goods on the Trave was struck by lightning during a thunderstorm and burned down.)*

Kuss, Christian: Jahrbuch denkwürdiger Naturereignisse in den Herzogtümern Schleswig und Holstein vom 11.–19. Jahrhundert. 2 Teile

Hazard: Storm, Mean: Ships, Zone: River, Impact: Restricted

The machine learning methods used in this paper belong to the group of supervised learning and therefore require manually coded examples for first learning or fitting the relation between text and labels, and then testing the accuracy of the derived classification. The accuracy of labeled data is usually measured by comparing right and wrong predictions, i.e. the fraction of number of correctly predicted labels to the number of all labels, as this metric is defined for the binary accuracy. The loss of a prediction does not only measure, if a label is wrong or right, but also how far it is away from the true value. This avoids steps in measuring the quality of a model and therefore is better suited for optimization. Common methods are the mean squared error, but here we use Binary Focal Crossentropy (Lin et al. 2018) as it performs better for multi-label data. The loss of the model decreases in small incremental steps (gradient descent), so that several rounds (so-called epochs) of training on the same data are necessary. Some models tend to overfit, meaning that they can classify the trained examples quite accurately but lose their ability to classify untrained examples with similar accuracy (Fig. 2).

To prevent overfitting, the data is further split up to validate the accuracy during the training phase on unseen data and stop refining the model further when the validation degrades. The number of training epochs differs with the diversity of the data, the amount of data and the complexity of the model and cannot be determined beforehand. The quotes in this study cover the period from the 11$^{th}$ to the 21$^{st}$ century and are divided according to these centuries. Additionally, the data of each century are divided into three segments that can be used for fitting or training, validating, and testing each method (Xu, Goodacre 2018). To check

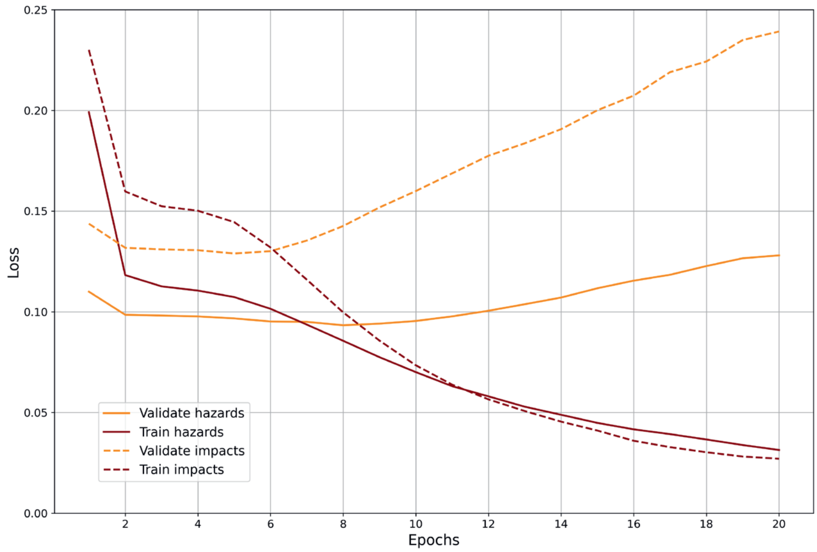


**Fig. 2** – Development of losses for train and test data during the epochs (here for LSTM model). The loss on training data typically decreases continuously, while the loss on test data first decreases, but then due to overfitting increases again.

the transferability of each method to different centuries and thereby the effect of variations of the German language, in each of the eleven rounds, one century is completely excluded from training and validating and used exclusively for testing the model (see Fig. 3). The classes in each group are not balanced, as some occur more often than others and their fractions change with the centuries, as well as the number of quotes and their number of words (Kahle, Glaser 2025, p. 128). All preprocessing is done using the python library pandas (Reback et al. 2021).

### *2.2. Models*

The models tested are implemented using the deep learning library Keras (Chollet, others 2015) using the Adam optimizer (Kingma, Ba 2017) and running for 20 epochs. The Adam optimizer is used to reduce the loss and thereby increase the accuracy by an adaptive gradient descent. The text is vectorized into up to 20,000 tokens (i.e., words) using the Count Vectorizer from the scikit-learn library (Pedregosa et al. 2011). The text is not stemmed or lemmatized, that is bringing words to their root or base form, as this task is difficult for the German language used in historical times. The text is not converted to lowercase, as it contains additional information; stopwords (frequently occurring words with little information)

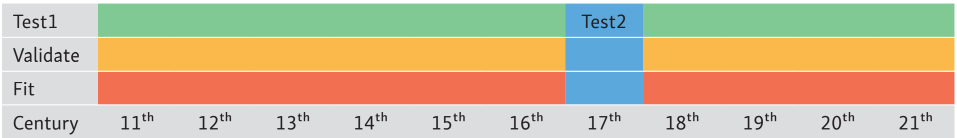


**Fig. 3** – Splitting of data in centuries and segments for fitting (red), validating (orange) and testing (green and blue)

are not removed. In each round, one century is completely left out for fitting and validating, but only used for separate testing (Fig. 3). The remaining ones are split into three sections: one used for fitting, one for validating, and the third one for testing ("test1") on that model, which led to the smallest loss during validation. The excluded century is also cross-validated on this optimized model ("test2"). All six permutations on the order of fit, test, and validate are performed, so that for each of the four parameters (hazard, mean, zone, impact) 66 variants are processed for each model type. To analyze the influence of the amount of training data, only a fraction of the fitting data could be used, while the rest is then added to the testing data. To determine the influence of the quality of the training data, a given fraction of the classification could be corrupted, so that incorrect labels are introduced in some cases. This degradation is done by randomly swapping the classifications of two labels, scrambling the parameter classes of one quote, flipping one class in one quote or inverting all classes of one quote.

We choose models to represent different types of architecture, such as the Bag-of-Words (BoW) approaches (Naive Bayes and Conservative Bayes), models using unsupervised Topic Extraction algorithms (TfIdf, LDA, NMF), Recurrent Networks (LSTM, GRU, RNN), and the Transformer model.

The Naive Bayes classification is a Bag-of-Words algorithm. It cannot take advantage of the order of tokens, but also does not depend on it. It assumes that all tokens are independent – which is obviously not the case for language. Nonetheless, Bayes' Theorem is commonly used for text categorization by calculating the probability of a category given a word (token) occurring from out of the probability that a word occurs within a given category, the overall probability of the category and the overall probability of a word in the text corpus. It makes use of the fact that some words are more frequently used in context than others. For example, names of rivers like 'Rhein' or 'Trave' occur more frequently when the zone of transportation is a river:

$$P(Category|Token) = P(Category) \cdot \frac{P(Token|Category)}{P(Token)} \tag{1}$$

For each word or token, the probabilities are multiplied and the estimation gets refined further:

$$P(Category|Tokens) = P(Category) \cdot \prod_{Tokens} \frac{P(Token|Category)}{P(Token)} \tag{2}$$

Using logarithms, the calculation transforms from multiplications to additions and the numbers get more manageable.

$$\begin{aligned} &logP(Category|Tokens) \\ &\qquad = logP(Category) + \sum_{Tokens} logP(Token|Category) - \sum_{Tokens} logP(Token) \end{aligned} \tag{3}$$

As the last term is equal for all categories (and the final probabilities are normalized via softmax), it can be dropped. The remaining probabilities are calculated by the ratio of absolute numbers of tokens in each category and token:

$$P(Category) = \frac{\#Tokens_{*,Category}}{\#Tokens_{*,*}}$$

and (4)

$$P(Token|Category) = \frac{Tokens_{Token,Category}}{Tokens_{*,Category}}$$

With * defined as counting over all tokens and/or categories. By using logarithms:

$$logP(Category) = log(\#Tokens_{*,Category}) - log(\#Tokens_{*,*}) \tag{5}$$

$$logP(Token|Category) = log(\#Tokens_{Token,Category}) - log(\#Tokens_{*,Category}) \tag{6}$$

By dropping the last term of equation (6) and inserting (5) and (6) into (3) the result is:

$$\begin{aligned} &logP(Category|Tokens) \\ &\qquad = log(\#Tokens_{*,Category}) \\ &\qquad + \sum_{Tokens} log(\#Tokens_{Token,Category}) - \sum_{Tokens} log(\#Tokens_{*,Category}) \end{aligned} \tag{7}$$

The first term represents the prior probability and given there are enough tokens used, even this term can be neglected. The model uses the two remaining terms as fixed embedding layers. To avoid log(0) the numbers of tokens are increased by a small amount $\varepsilon = 10^{-12}$.

The Naive Bayes fails, when tokens do not appear in one category in the training data, as the probabilities degenerate to zero (called zero inflation) and can never rise again. When this token now appears in the test data, it will force the probability of this category to zero (or the very small number ε) ignoring the votes of all other tokens. Therefore, we developed the Conservative Bayes algorithm, which uses the assumption that some tokens are accidentally missing. Their absolute number therefore is corrected by adding the standard deviation assuming a Poisson distribution for values greater than zero and the overall probability of the category (additive smoothing):

$$\#Tokens_{\widehat{Token},Category} = \#Tokens_{Token,Category} + \sqrt{\#Tokens_{Token,Category}} + \frac{\#Tokens_{*,Category}}{\#Tokens_{*,*}} + \epsilon \tag{8}$$

Other Bag-of-Words algorithms are based on the term frequency-inverse document frequency (TF-IDF; Spärck Jones 2004), the Latent Dirichlet Allocation (LDA; Hoffman et al. 2013; Hoffman, Blei, Bach 2010) or an TF-IDF followed by an Non-negative Matrix Factorization (NMF; Hoyer 2004). They are used to extract a given number of topics by unsupervised learning and perform well, when the intended classification is inherent in the documents itself (Kahle et al. 2022). A feedforward neural network then processes the derived topic weights.

As text is sequential data, Recurrent Neural Networks (RNN) are often used, as they can capture dependencies and patterns within sequences and so are able to extract information hidden in the grammar of language. To better learn long-range dependencies inside the text, the Long Short-Term Memory (LSTM) architecture was invented (Hochreiter, Schmidhuber 1997) and later the Gated Recurrent Units (GRU) approach was implemented for better runtime performance (Cho et al. 2014). For large language models LLM today, usually the transformer architecture is applied, that uses the method of multi-head attention to concentrate on the most relevant parts of a text (Vaswani et al. 2017). Here we use a smaller variant of it, with fewer variables and less required training time.

For each model architecture, 264 variants are tested (4 parameters, 6 Permutations of fit, validate and test, 11 centuries). The runtime needed for training and prediction is measured, as well as the accuracy of the results and the size of the model by counting the number of parameters it uses. The collected results are fitted to a Generalized Linear Model (GLM) with shrinkage using the python library statsmodels (Seabold, Perktold 2010). Thereby the importance and the significance of the different categories are determined (see Supplements).

## 3. Results

### *3.1. Accuracy of prediction*

The accuracy of the prediction is not only a function of the model, but also is driven by the parameter, the century and the amount and quality of training data. We will look into the most significant factors and visualize their influence.

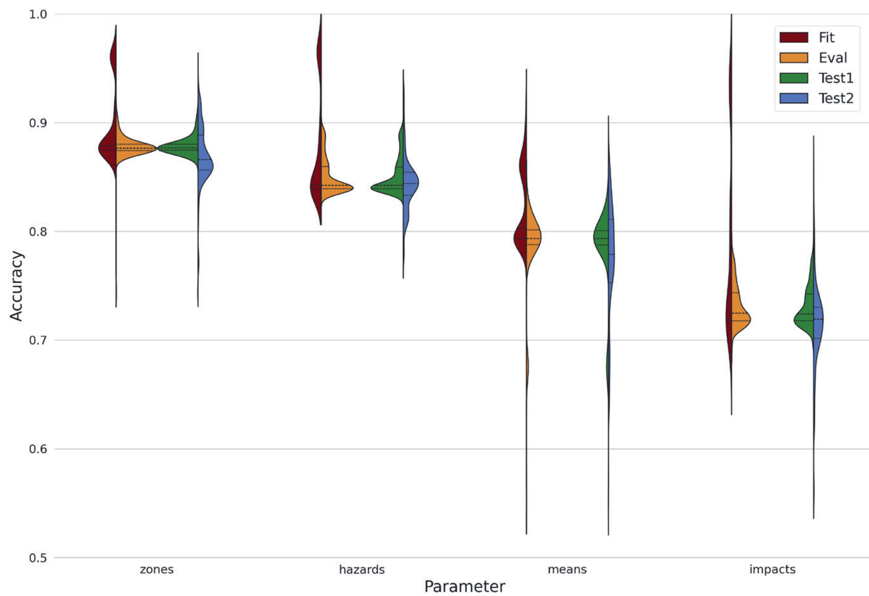


**Fig. 4** – The distribution of the accuracy for different parameters for all models combined

The largest factor influencing the accuracy is the parameter (hazard, mean, zone, impact) to be categorized (Fig. 4). The zone of traffic could be predicted best, the hazards follow closely behind. Their categories are usually well defined and explicitly given in the quotes. Rivers, bridges, roads, lakes, and the sea are distinguishable as well as most of the hazards like cold waves, floods, storms, droughts, and snow. Only some of them could overlap in a fuzzy way, as is the case for storm surges. The means of transportation is harder to predict, while the impact is the worst. Often both are not explicitly mentioned in the text, but need some interpretation. For example, a destroyed bridge affects multiple means or a frozen lake can be crossed in different ways. The impact categories "Restricted" and "Impossible" are harder to assign, as they need to consider incidents like how many people are affected in which way and how long a possible delay is. A ship affected by a storm might be classified as restricted as long as it does not sink or people die.

Some examples of single predicted results are shown in Figure 5 using the quotes introduced in the data section. The hazard for quote #22036 is predicted well as "cold" by the Bayes and LSTM models, while the TfIdf and the transformer model are undistinguished between floods, storm, cold, and snow and only down vote heat and rain. The traffic mean is quite blurred for all models, but one can argue, that on the frozen sea it is possible to move by foot, horse, wagon or sled in the same way (but not with ships or balloons). For the zone of travel in quote

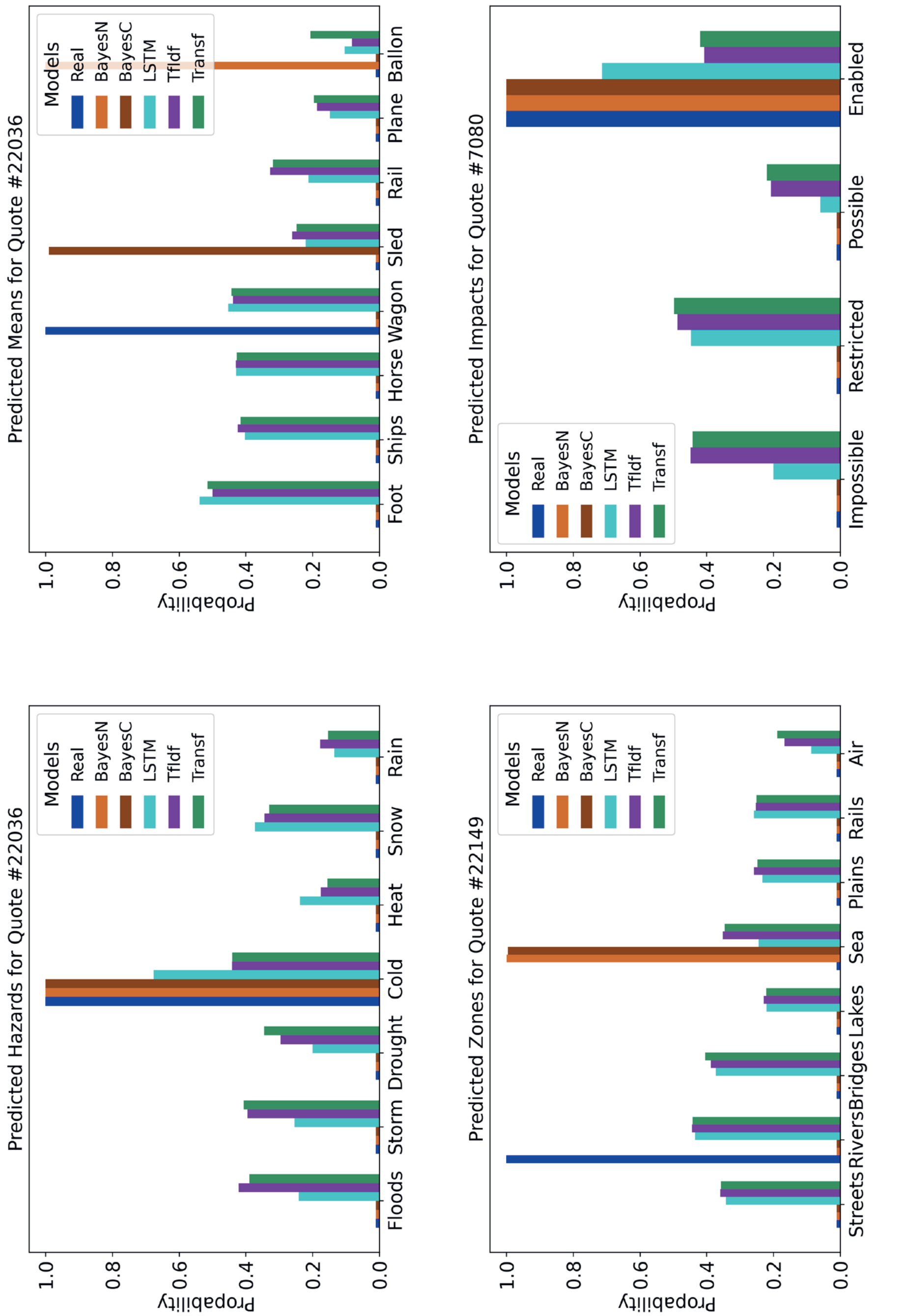


**Fig. 5** – Predicted parameters for selected quotes and parameters

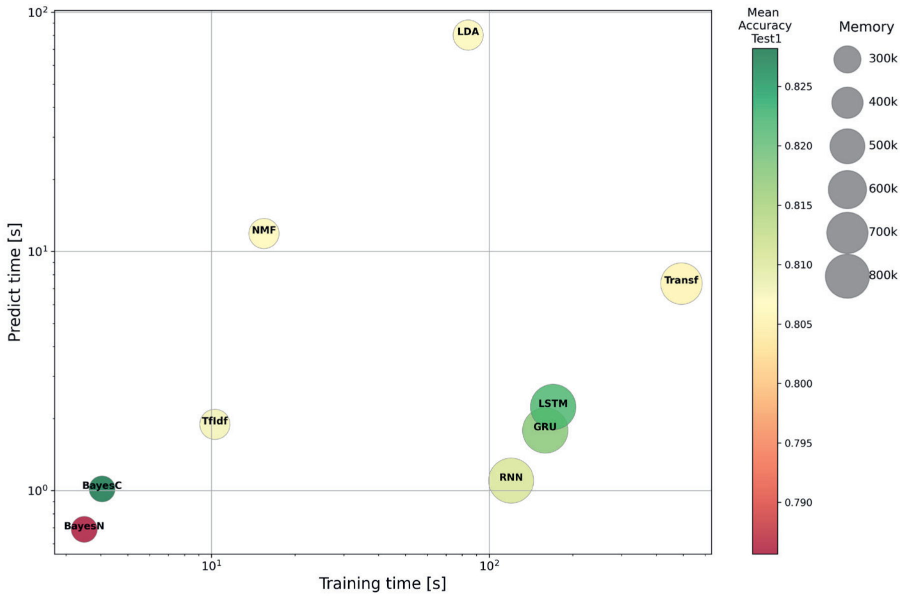


**Fig. 6** – Performance of models regarding training time, predict time, model size and accuracy

#22149 the Bayes models both wrongly assume the sea instead of the river, which might be caused because ships are mentioned more often on the sea and the river name (Trave) does not occur as often. The other models prefer the river, but cannot be clearly distinguished from streets or bridges. The predicted impacts for quote #7080 are accurate for the Bayes and the LSTM model, while the TfIdf and the transformer model tend more to the incorrect negative impacts (impossible and restricted).

### *3.2. Performance of the models*

For further observations we will concentrate on the mean accuracy of all test cases using the first test case (test1). It is also very similar to the evaluation data (validation). The models do not differ so much in the second test case that is performed on the skipped centuries. For comparing the models, we will measure the mean accuracy of the predictions, the time needed for training as well as for prediction and the memory size of the resulting model.

The most important difference among the models is the runtime, as it varies by two orders of magnitude (Fig. 6). Using the Bayes classifications results in very fast (and small) models, both for training and predicting. The Transformer model is the most expensive to train, followed by the model using LSTM. Regarding

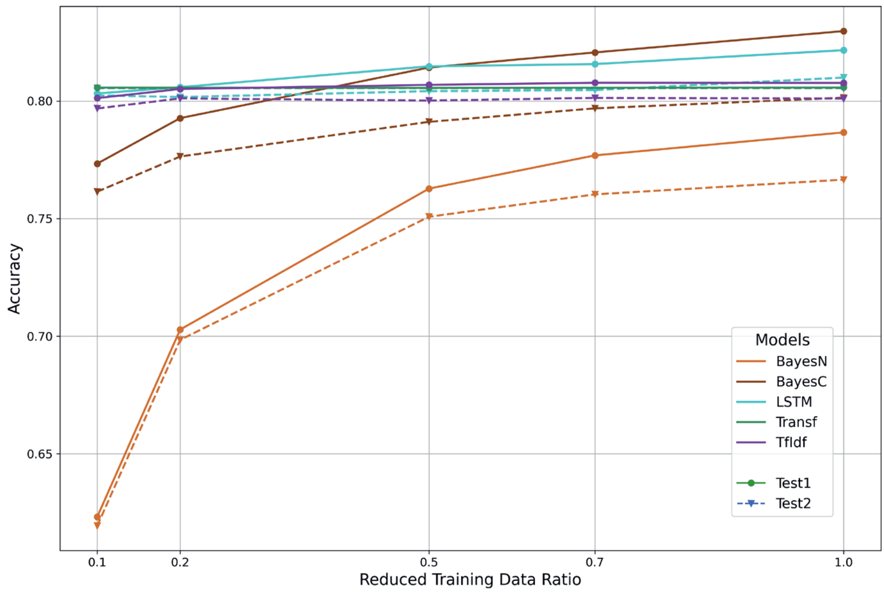


**Fig. 7** – Accuracy of the models for reduced training data amount. The ratio determines the fraction of quotes used for training, (i.e. 0.5 only uses every second quote).

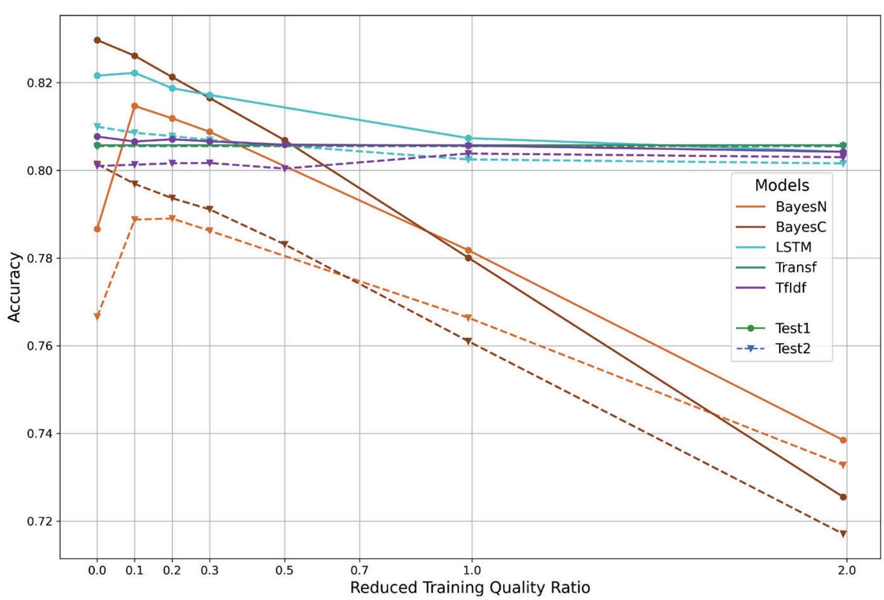


**Fig. 8** – Accuracy of the models for reduced training quality. The ratio describes how many quotes are mislabeled. A ratio of 2 mean a quote is distorted by two methods in average.

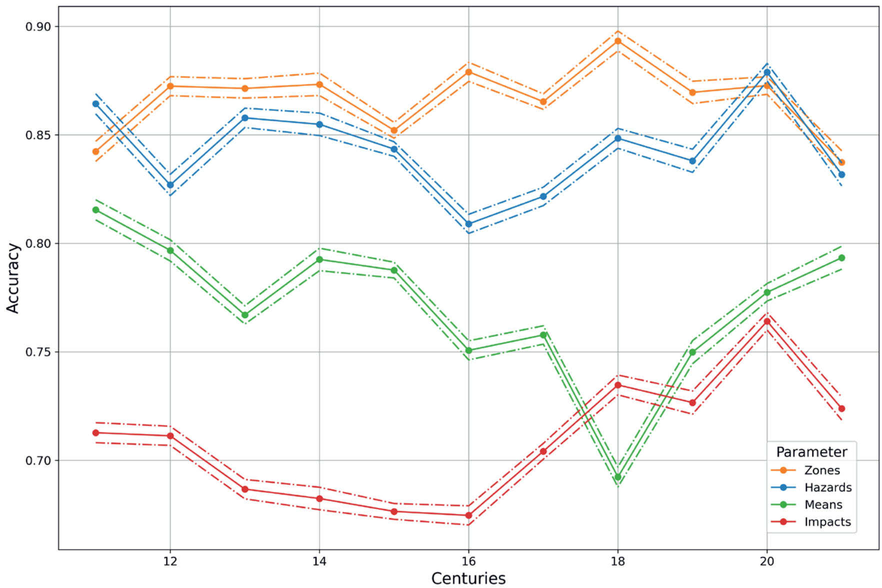


**Fig. 9** – Accuracy of the parameters across the centuries

the accuracy, the Conservative Bayes Classifier performs best, even better than the LSTM (and the other recurrent networks). The architectures based on Topic Extraction (TfIdf, NMF, LDA) are less accurate, as well as the Transformer and the Naive Bayes Classifier. The Transformer might get better with further training, as we finished early after 20 epochs without running into overfitting, but then runtime performance would be even worse.

Figure 7 and 8 visualize the results for reduced training data in terms of quantity or quality. Overall, all models tend to lose accuracy for less data of poorer quality. The largest effect on amount of training data is observed for the naive Bayes model, while the conservative Bayes model is clearly better. The smallest effect appears for the complex models LSTM and transformer) and the model using the topic extraction (TfIdf). The unsupervised topic extraction has the advantage, that it can perform on unlabeled data and uses the reduced training data only to refine these pre-classifications to the desired ones via few-shot learning. Looking at the effect on the quality of the training data used for fitting, the decline is largest for the two Bayes models. The LSTM model is also affected by a small decrease. The transformer and the TfIdf model both stay on the same level, but overall start at a lower accuracy.

For explaining the accuracy along the eleven centuries, the most important attribute besides the century itself is the parameter (Fig. 9). The influence of the model is smaller and only significant for some combinations. The dominant

**Table 2** – Summarized results for the different models compared to manual classification. The models used for the comparison are printed in italics. Indicators highlights performance from very good (++) to very poor (––).

| Method | Fully manual from scratch | Keyword search on corpus | Bayesian Classification | Topic Extractions & Neural Network | Recurrent Neural Network | Large Language Models |
|---|---|---|---|---|---|---|
| Models | – | – | bayesNF, *bayesCF* | *TfIdf*, NMF, LDA | RNN, *LSTM*, GRU | *Transformer* |
| Training Time | – | – | ++ 4.3s | + 10.3s | – 170s | –– 493s |
| Predict Time | hours | hours | ++ 1.2s | + 1.9s | + 2.2s | o 7.3s |
| Memory Usage | – | – | ++ 280k | + 378k | o 850k | o 715k |
| Data Amount needed | ++ | ++ | –– | ++ | o | ++ |
| Data Quality needed | ++ | ++ | – | o | – | o |
| Use of Grammar & Syntax | Yes | Yes | No | No | Yes | Yes |
| Accuracy min<br>Accuracy avg | (99%)<br>(99%) | (99%)<br>(99%) | 72%<br>83.0% | 69%<br>81% | 70%<br>81.6% | 71%<br>81% |

influence of the data suggests that improving quality and quantity of labeled data would be the first option to consider for improvements.

The relevant results of the different models are summarized in Table 2. It uses the determined values from Figure 6 for the training and prediction time, the memory usage and the accuracy. The influence of the training data amount and its quality is determined using Figure 7 and 8. In addition, we estimate the performance of manual classification. A simple indicator distinguishes between very good (++), good (+), average (o), poor (–) and very poor (––) performances. A perfect model is fast, small and accurate, even for limited, poor quality training data. The first workflow considers collecting data entirely from scratch, which involves searching for feasible documents first. The time to find and classify quotes is immense and could take several minutes or even up to hours to find and code a single quote. While you may need some experience, which somewhat reflects the training time, only a few examples are needed to understand the idea of the categorization. In addition, usually the grammar and syntax could be used to extract the exact information from the text. The performance increases, when a searchable corpus already exists.

The summary shows, that smaller models can even outperform more complex ones in classifying documents in the field of historical climatology as long as the training data is of good quality and its amount is sufficient. This result is remarkable, especially considering their training time is faster by two orders of magnitude.

## 4. Discussion

While the use of digital methods already assists different steps in the workflow of historical climatology, there are still many opportunities to accelerate this workflow further. The development of the German language through time is nonetheless challenging for today's algorithms and pretrained models. Usually, they are trained on large amounts of modern texts and less frequently on historical documents and cannot be used for historical language out of the box. Therefore, models need to be adapted or trained from scratch for these kinds of texts. On one hand, this approach demands a minimum amount of manually labeled quotes. On the other hand, the time needed for training must not exceed practical limits to avoid waisting resources. The critical characteristics of feasible algorithms are the training time needed, the accuracy achieved, and the amount and quality of training data. To train large language models from scratch, such as the LSTM or Transformer, for historical documents would require a lot of resources regarding the number of historical sources with annotated quotes, but also computation time. Therefore, most of the text analysis research in the area of climatology or natural hazards is done on recent sources, such as social media or online newspapers. This focus also explains, why there is no well-established, commonly available tool to classify historical documents, and using these methods requires some programming skills and an adaptation to the specific task, even if some digital methods are applied in the field of digital humanities, theology or history (Nunn et al. 2024; Antenhofer, Kühberger, Strohmeyer 2023).

The architectures using Bayes classification and Topic Extraction are promising for processing historical sources and might even be improved further. For example, the inner components of the models and the settings used during training could be tuned further, or the two might even be combined. While the loss function used is well accepted for multi-label classification, adapted functions can better reflect the needs of the climatological classification. Furthermore, Bayes methods are already used in the field of climate indices, although not in the context of text handling (Camenisch et al. 2021).

In addition, the preprocessing might be optimized by applying normalization techniques. Transforming all words to lowercase merges different tokens into one and so increases the absolute count numbers and thereby their significance. Stemming or Lemmatizing takes this idea further, and reduces inflectional forms to a common base form. As they rely on grammar rules or dictionaries, the change of non-standardized language in historical times cannot be handled completely by modern libraries. The downside of normalization is that nuances hidden in the token variants may get lost. On one hand, words that appear regularly in every kind of text (stopwords) could be dropped as well, as they might lead to noise.

On the other hand, they contain information about relations between words and algorithms cannot use these after removal.

While the overall accuracy with approximately 80% today is not sufficient to use the algorithms autonomously, they still can save a vast amount of time when used in a copilot way to assist human researchers.

## 5. Conclusions

In conclusion, digital methods can profoundly transform the workflow of historical climatology. By automating or enhancing many of the traditionally manual and time-consuming steps, digital tools enable researchers to process larger volumes of historical information, apply more sophisticated analytical techniques, and facilitate greater collaboration and data sharing within the scientific community. Some of the methods are established well, i.e. the search for sources in web portals, while others such as the classification of climate related information are still in the development phase and not commonly used yet, but offer great potential in future research. This digital revolution is crucial for unlocking the wealth of climate information embedded in historical documents and for improving our understanding of long-term climate variability and its impact on human societies.

As shown in this paper, the classification of written text regarding different climatic parameters can be assisted by methods of machine learning. They will hopefully revolutionize the discovery and classification of climatic information in large text corpora in the future.

## References

ANTENHOFER, C., KÜHBERGER, C., STROHMEYER, A. (2023): Digital Humanities in den Geschichtswissenschaften. Böhlau, Wien. https://doi.org/10.36198/9783838561165

BEHRINGER, W. (1995): Weather, Hunger and Fear: Origins of the European Witch-Hunts in Climate, Society and Mentality. German History, 13, 1–27. https://doi.org/10.1177/026635549501300101

BI, K., XIE, L., ZHANG, H., CHEN, X., GU, X., TIAN, Q. (2023): Accurate medium-range global weather forecasting with 3D neural networks. Nature, 7970, 619, 533–538. https://doi.org/10.1038/s41586-023-06185-3

BOSE, R., PINTAR, A.L., SIMIU, E. (2023): Simulation of Atlantic Hurricane Tracks and Features: A Coupled Machine Learning Approach. Artificial Intelligence for the Earth Systems, 2, 2, 220060. https://doi.org/10.1175/AIES-D-22-0060.1

BÖSMEIER, A. (2020): Exploring and analyzing data for reconstruction, modeling, and hazard assessment of historical floods in the Kinzig catchment, Upper Rhine area, https://freidok.uni-freiburg.de/data/222531.

BRÁZDIL, R., PFISTER, C., WANNER, H., STORCH, H. V., LUTERBACHER, J. (2005): Historical Climatology In Europe – The State Of The Art. Climatic Change, 3, 70, 363–430. https://doi.org/10.1007/s10584-005-5924-1

BRIEGEL, F., SCHRODI, S., SULZER, M., BROX, T., PINTO, J.G., CHRISTEN, A. (2025): Deep learning enables city-wide climate projections of street-level heat stress. Urban Climate, 62. https://doi.org/10.1016/j.uclim.2025.102564

BRUNTON, S.L., KUTZ, J.N. (2019): Data-driven science and engineering: machine learning, dynamical systems, and control. Cambridge University Press, Cambridge New York, NY Port Melbourne New Delhi Singapore.

BUTLER, H., DALY, M., DOYLE, A., GILLIES, S., SCHAUB, T., SCHMIDT, C. (2014): GeoJSON. Electronic, http://geojson.org.

CAMENISCH, C., JAUME-SANTERO, F., WHITE, S., PEI, Q., HAND, R., ROHR, C., BRÖNNIMANN, S. (2021): A Bayesian Approach to Historical Climatology for the Burgundian Low Countries in the 15<sup>th</sup> Century. preprint. Climate Modelling/Historical Records/Centennial-Decadal. https://doi.org/10.5194/cp-2021-169

CHO, K., VAN MERRIENBOER, B., GULCEHRE, C., BAHDANAU, D., BOUGARES, F., SCHWENK, H., BENGIO, Y. (2014): Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. arXiv. https://doi.org/10.3115/v1/D14-1179

CHOLLET, F., OTHERS (2015): Keras, https://github.com/fchollet/keras.

COOK, E. R., SEAGER, R., KUSHNIR, Y., BRIFFA, K.R., BÜNTGEN, U., FRANK, D., KRUSIC, P.J., TEGEL, W., VAN DER SCHRIER, G., ANDREU-HAYLES, L., BAILLIE, M., BAITTINGER, C., BLEICHER, N., BONDE, N., BROWN, D., CARRER, M., COOPER, R., ČUFAR, K., DITTMAR, C., ESPER, J., GRIGGS, C., GUNNARSON, B., GÜNTHER, B., GUTIERREZ, E., HANECA, K., HELAMA, S., HERZIG, F., HEUSSNER, K.-U., HOFMANN, J., JANDA, P., KONTIC, R., KÖSE, N., KYNCL, T., LEVANIČ, T., LINDERHOLM, H., MANNING, S., MELVIN, T.M., MILES, D., NEUWIRTH, B., NICOLUSSI, K., NOLA, P., PANAYOTOV, M., POPA, I., ROTHE, A., SEFTIGEN, K., SEIM, A., SVARVA, H., SVOBODA, M., THUN, T., TIMONEN, M., TOUCHAN, R., TROTSIUK, V., TROUET, V., WALDER, F., WAŻNY, T., WILSON, R., ZANG, C. (2015): Old World megadroughts and pluvials during the Common Era. Science Advances, 10, 1, e1500561. https://doi.org/10.1126/sciadv.1500561

DE VOS, M.G., HAZELEGER, W., BARI, D., BEHRENS, J., BENDOUKHA, S., GARCIA-MARTI, I., VAN HAREN, R., HAUPT, S.E., HUT, R., JANSSON, F., MUELLER, A., NEILLEY, P., VAN DEN OORD, G., PELUPESSY, I., RUTI, P., SCHULTZ, M.G., WALTON, J. (2020): Open weather and climate science in the digital era. Geoscience Communication, 2, 3, 191–201. https://doi.org/10.5194/gc-3-191-2020

GLASER, R. (2013): Klimageschichte Mitteleuropas: 1200 Jahre Wetter, Klima, Katastrophen. WBG, Wiss. Buchges, Darmstadt.

GLASER, R., HIMMELSBACH, I., BÖSMEIER, A. (2017): Climate of migration? How climate triggered migration from southwest Germany to North America during the 19th century. Climate of the Past, 11, 13, 1573–1592. https://doi.org/10.5194/cp-13-1573-2017

GLASER, R., KAHLE, M. (2020): Reconstructions of droughts in Germany since 1500 – combining hermeneutic information and instrumental records in historical and modern perspectives. Climate of the Past, 4, 16, 1207–1222. https://doi.org/10.5194/cp-16-1207-2020

GLASER, R., KAHLE, M., HOLOGA, R. (2016): The tambora.org data series edition. Albert-Ludwigs-Universität Freiburg, https://doi.org/10.6094/TAMBORA.ORG/2016/SERIESNOTES.PDF.

GRÄSSE, J.G.T., BENEDICT, F. (1909): Orbis Latinus: R. C. Schmidt & Company, https://books.google.de/books?id=oJt-AAAAMAAJ.

GRATZINGER, O. (2021): The Internet Archive: Founded by Brewster Kahle, https://archive.org/about/. American Journalism, 2, 38, 249–251. https://doi.org/10.1080/08821127.2021.1912531

GROTEFEND, H. (2015): Handbuch der historischen Chronologie des deutschen Mittelalters und der Neuzeit. Vero Verlag, Norderstedt.

HAHN, C.H., MORITZ, W. (1998): Erkundungsreise ins Ovamboland 1857: Tagebuch. Lempp, Schwäbisch Gmünd.

HARRIS, Z.S. (1954): Distributional Structure. WORD, 2–3, 10, 146–162. https://doi.org/10.1080/00437956.1954.11659520

HOCHREITER, S., SCHMIDHUBER, J. (1997): Long Short-Term Memory. Neural Computation, 8, 9, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

HOFFMAN, M.D., BLEI, D.M., BACH, F. (2010): Online learning for Latent Dirichlet Allocation. In: Proceedings of the 24th International Conference on Neural Information Processing Systems, 1. Curran Associates Inc., Red Hook, NY, USA, 856–864.

HOFFMAN, M.D., BLEI, D.M., WANG, C., PAISLEY, J. (2013): Stochastic Variational Inference. Journal of Machine Learning Research, 40, 14, 1303–1347, http://jmlr.org/papers/v14/hoffman13a.html.

HOLOGA, R., GLASER, R. (2021): The Societal Echo of Severe Weather Events: Ambient Geospatial Information (AGI) on a Storm Event. ISPRS International Journal of Geo-Information, 12, 10, 815. https://doi.org/10.3390/ijgi10120815

HONNIBAL, M., MONTANI, I., VAN LANDEGHEM, S., BOYD, A. (2020): Spacy: Industrial-strength Natural Language Processing in Python.

HOYER, P. O. (2004): Non-negative matrix factorization with sparseness constraints. Journal of machine learning research, 9, 5.

KAHLE, B. (2007): Universal Access to All Knowledge. The American Archivist, 1, 70, 23–31, http://www.jstor.org/stable/40294448 (30.12.2024). https://doi.org/10.17723/aarc.70.1.u114006770252845

KAHLE, M. (2025a): Innovative, digitale Verfahren zur maschinellen Erfassung und Klassifizierung von Texten zur historischen und rezenten Klimaanalyse und deren gesellschaftliche Folgen. Universität Freiburg. https://doi.org/10.6094/UNIFR/268947

KAHLE, M., GLASER, R. (2025): Climatic conditions and mobility from 1000 to 1500. Hermeneutic and statistical approaches. In: Sarti, L., Trott zu Stolz, von, H. (eds.): Mobility in the Early Middle Ages, and Beyond – Mobilität im Frühmittelalter und darüber hinaus. https://doi.org/10.1515/9783111166698-009

KAHLE, M., KEMPF, M., MARTIN, B., GLASER, R. (2022): Classifying the 2021 'Ahrtal' flood event using hermeneutic interpretation, natural language processing, and instrumental data analyses. Environmental Research Communications, 5, 4, 051002. https://doi.org/10.1088/2515-7620/ac6657

KAHLE, P., COLUTTO, S., HACKL, G., MÜHLBERGER, G. (2017): Transkribus – A Service Platform for Transcription, Recognition and Retrieval of Historical Documents. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). 19–24. https://doi.org/10.1109/ICDAR.2017.307

KHIDER, D., EMILE-GEAY, J., MCKAY, N. P., GIL, Y., GARIJO, D., RATNAKAR, V., ALONSO-GARCIA, M., BERTRAND, S., BOTHE, O., BREWER, P., BUNN, A., CHEVALIER, M., COMAS-BRU, L., CSANK, A., DASSIÉ, E., DELONG, K., FELIS, T., FRANCUS, P., FRAPPIER, A., GRAY, W., GORING, S., JONKERS, L., KAHLE, M., KAUFMAN, D., KEHRWALD, N.M., MARTRAT, B., MCGREGOR, H., RICHEY, J., SCHMITTNER, A., SCROXTON, N., SUTHERLAND, E., THIRUMALAI, K., ALLEN, K., ARNAUD, F.,

AXFORD, Y., BARROWS, T., BAZIN, L., PILAAR BIRCH, S. E., BRADLEY, E., BREGY, J., CAPRON, E., CARTAPANIS, O., CHIANG, H.-W., COBB, K.M., DEBRET, M., DOMMAIN, R., DU, J., DYEZ, K., EMERICK, S., ERB, M.P., FALSTER, G., FINSINGER, W., FORTIER, D., GAUTHIER, N., GEORGE, S., GRIMM, E., HERTZBERG, J., HIBBERT, F., HILLMAN, A., HOBBS, W., HUBER, M., HUGHES, A.L.C., JACCARD, S., RUAN, J., KIENAST, M., KONECKY, B., LE ROUX, G., LYUBCHICH, V., NOVELLO, V.F., OLAKA, L., PARTIN, J.W., PEARCE, C., PHIPPS, S.J., PIGNOL, C., PIOTROWSKA, N., POLI, M. -S., PROKOPENKO, A., SCHWANCK, F., STEPANEK, C., SWANN, G.E.A., TELFORD, R., THOMAS, E., THOMAS, Z., TRUEBE, S., VON GUNTEN, L., WAITE, A., WEITZEL, N., WILHELM, B., WILLIAMS, J., WILLIAMS, J.J., WINSTRUP, M., ZHAO, N., ZHOU, Y. (2019): PaCTS 1.0: A Crowdsourced Reporting Standard for Paleoclimate Data. Paleoceanography and Paleoclimatology, 10, 34, 1570–1596. https://doi.org/10.1029/2019PA003632

KHOO KHYOU BUN, ISHIZUKA, M. (2002): Topic extraction from news archive using TF*PDF algorithm. In: Proceedings of the Third International Conference on Web Information Systems Engineering, 2002. WISE 2002. IEEE Comput. Sci, Singapore, 73–82. https://doi.org/10.1109/WISE.2002.1181645

KINGMA, D.P., BA, J. (2017): Adam: A Method for Stochastic Optimization. arXiv.

KUHL, E., ESPER, J., SCHNEIDER, L., TROUET, V., KUNZ, M., KLIPPEL, L., BÜNTGEN, U., HARTL, C. (2024): Revising Alpine summer temperatures since 881 CE. Climate Dynamics. https://doi.org/10.1007/s00382-024-07195-1

LAMPRECHT, A.-L., GARCIA, L., KUZAK, M., MARTINEZ, C., ARCILA, R., MARTIN DEL PICO, E., DOMINGUEZ DEL ANGEL, V., VAN DE SANDT, S., ISON, J., MARTINEZ, P.A., MCQUILTON, P., VALENCIA, A., HARROW, J., PSOMOPOULOS, F., GELPI, J. LL., CHUE HONG, N., GOBLE, C., CAPELLA-GUTIERREZ, S. (2020): Towards FAIR principles for research software. Data Science, 1, 3, 37–59. https://doi.org/10.3233/DS-190026

LIN, T.-Y., GOYAL, P., GIRSHICK, R., HE, K., DOLLÁR, P. (2018): Focal Loss for Dense Object Detection. arXiv. https://doi.org/10.1109/ICCV.2017.324

MADRUGA DE BRITO, M., KUHLICKE, C., MARX, A. (2020): Near-real-time drought impact assessment: a text mining approach on the 2018/19 drought in Germany. Environmental Research Letters, 10, 15, 1040a9. https://doi.org/10.1088/1748-9326/aba4ca

MARCHANT, J. (2023): AI reads text from ancient Herculaneum scroll for the first time. Nature, d41586-023-03212–1. https://doi.org/10.1038/d41586-023-03212-1

MATUSCHEK, O. (2014): Data-Mining in den Reisetagebüchern James Silk Buckinghams 1815/1816: neu entwickelte Suchalgorithmen und effiziente Arbeitsmethoden in der Historischen Klimatologie, https://freidok.uni-freiburg.de/data/9623 (7.1.2025).

MCKAY, N.P., EMILE-GEAY, J. (2016): Technical note: The Linked Paleo Data framework – a common tongue for paleoclimatology. Climate of the Past, 4, 12, 1093–1100. https://doi.org/10.5194/cp-12-1093-2016

MITLOHNER, J., NEUMAIER, S., UMBRICH, J., POLLERES, A. (2016): Characteristics of Open Data CSV Files. In: 2016 2nd International Conference on Open and Big Data (OBD). IEEE, Vienna, 72–79. https://doi.org/10.1109/OBD.2016.18

MOLLOY, J.C. (2011): The Open Knowledge Foundation: Open Data Means Better Science. PLoS Biology, 12, 9, e1001195. https://doi.org/10.1371/journal.pbio.1001195

MONS, B., SCHULTES, E., LIU, F., JACOBSEN, A. (2020): The FAIR Principles: First Generation Implementation Choices and Challenges. Data Intelligence, 1–2, 2, 1–9. https://doi.org/10.1162/dint_e_00023

NUNN, C.A. et al. (2024): Introducing Digital Humanities to Theology, heiBOOKS.

PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M., DUCHESNAY, E. (2011): Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

RAISSI, M., PERDIKARIS, P., AHMADI, N., KARNIADAKIS, G.E. (2024): Physics-Informed Neural Networks and Extensions, arXiv.

REBACK, J., MCKINNEY, W., JBROCKMENDEL, BOSSCHE, J.V.D., AUGSPURGER, T., CLOUD, P., GFYOUNG, HAWKINS, S., SINHRKS, ROESCHKE, M., KLEIN, A., TERJI PETERSEN, TRATNER, J., SHE, C., AYD, W., NAVEH, S., GARCIA, M., SCHENDEL, J., HAYDEN, A., SAXTON, D., PATRICK, JANCAUSKAS, V., MCMASTER, A., BATTISTON, P., SKIPPER SEABOLD, GORELLI, M., KAIQI DONG, CHRIS-B1, H-VETINARI, HOYER, S. (2021): pandas-dev/pandas: Pandas 1.2.1. Zenodo.

REINGOLD, E.M., DERSHOWITZ, N. (2001): Calendrical calculations. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781107051119

RIEMANN, D. (2010): Methoden zur Klimarekonstruktion aus historischen Quellen am Beispiel Mitteleuropas, https://freidok.uni-freiburg.de/data/7904 (7.1.2025).

RIEMANN, D., GLASER, R., KAHLE, M., VOGT, S. (2015): The CRE tambora.org – new data and tools for collaborative research in climate and environmental history. Geoscience Data Journal, 2, 2, 63–77. https://doi.org/10.1002/gdj3.30

SAHLE, P., VOGELER, G. (2013): Digital Monumenta Germaniae Historica (dMGH). Digital Philology: A Journal of Medieval Cultures, 1, 2, 135–139. https://doi.org/10.1353/dph.2013.0006

SCHÄTZ, F. (2023): Voraussetzungen und Grenzen der Auswertung klimarelevanter Informationen historischer Textquellen mit Hilfe von Automatisierungsprozessen. Albert-Ludwigs-Universität Freiburg.

SEABOLD, S., PERKTOLD, J. (2010): statsmodels: Econometric and statistical modeling with python. In: 9th Python in Science Conference. https://doi.org/10.25080/Majora-92bf1922-011

SMITH, R. (2007): An Overview of the Tesseract OCR Engine. In: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007). 629–633. https://doi.org/10.1109/ICDAR.2007.4376991

SPÄRCK JONES, K. (2004): A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 5, 60, 493–502. https://doi.org/10.1108/00220410410560573

SPECKLE, I. (1968): Das Tagebuch von Ignaz Speckle. Veröffentlichungen der Kommission für geschichtliche Landeskunde in Baden-Württemberg, Stuttgart.

STEPHAN, R., STAHL, K., DORMANN, C.F. (2023): Drought impact prediction across time and space: limits and potentials of text reports. Environmental Research Letters, 7, 18, 074004. https://doi.org/10.1088/1748-9326/acd8da

STEVENS, S.S. (1946): On the Theory of Scales of Measurement. Science, 2684, 103, 677–680. https://doi.org/10.1126/science.103.2684.677

THÖLE, L.M., WEGMANN, M. (2024): Open climate science is brave climate science. Environmental Research Letters, 12, 19, 122001. https://doi.org/10.1088/1748-9326/ad893f

VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A.N., KAISER, L., POLOSUKHIN, I. (2017): Attention Is All You Need, arXiv. https://doi.org/10.48550/arXiv.1706.03762

VERSCHOOF-VAN DER VAART, W., KAPTIJN, E., BOURGEOIS, Q., LAMBERS, K. (2025): 13 Using Citizen Science to map hollow roads in LiDAR data from the Netherlands. Results from the Heritage Quest project. In: Sarti, L., Von Trott Zu Solz, H. (eds.): Mobility in the Early

Middle Ages, and Beyond – Mobilität im Frühmittelalter und darüber hinaus. De Gruyter, 271–286. https://doi.org/10.1515/9783111166698-013

WARREN, R., BARTLOME, N.E., WELLINGER, N., FRANKE, J., HAND, R., BRÖNNIMANN, S., HUHTAMAA, H. (2024): ClimeApp: data processing tool for monthly, global climate data from the ModE-RA palaeo-reanalysis, 1422 to 2008 CE. Climate of the Past, 12, 20, 2645–2662. https://doi.org/10.5194/cp-20-2645-2024

WESSELKAMP, M., CHANTRY, M., PINNINGTON, E., CHOULGA, M., BOUSSETTA, S., KALWEIT, M., BÖDECKER, J., DORMANN, C.F., PAPPENBERGER, F., BALSAMO, G. (2025): Advances in land surface forecasting: a comparison of LSTM, gradient boosting, and feed-forward neural networks as prognostic state emulators in a case study with ecLand. Geoscientific Model Development, 4, 18, 921–937. https://doi.org/10.5194/gmd-18-921-2025

WHITE, S., PEI, Q., KLEEMANN, K., DOLÁK, L., HUHTAMAA, H., CAMENISCH, C. (2023): New perspectives on historical climatology. WIREs Climate Change, 1, 14, e808. https://doi.org/10.1002/wcc.808

WICK, M. (2015): Geonames Ontology, http://www.geonames.org/about.html (22.4.2015).

WBIS (2005): World Biographical Information System. Reference Reviews, 2, 19, 54–55. https://doi.org/10.1108/09504120510591828

XU, Y., GOODACRE, R. (2018): On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. Journal of Analysis and Testing, 3, 2, 249–262. https://doi.org/10.1007/s41664-018-0068-2

YANG, S., NAI, C., LIU, X., LI, W., CHAO, J., WANG, J., WANG, L., LI, X., CHEN, X., LU, B., XIAO, Z., BOERS, N., YUAN, H., PAN, B. (2025): Generative assimilation and prediction for weather and climate, arXiv.

ZASADNI, J. (2007): The Little Ice Age in the Alps: its record in glacial deposits and rock glacier formation. Studia Geomorphologica Carpatho-Balcanica, 41, 117–137.

ZHANG, D., GLASER, R., KAHLE, M. (2025): Stable yet dynamic: a cross-era comparative case study of drought impacts and social responses in Germany and Jing-Jin-Ji Region (China). https://doi.org/10.5194/cp-21-1481-2025

## Data

KAHLE, M., GLASER, R. (2022): Climatic conditions and mobility in the Middle Ages in Europe. Albert-Ludwigs-Universität Freiburg. https://doi.org/10.6094/TAMBORA.ORG/2022/C567/CSV.ZIP

KAHLE, M. (2025b): Climatic conditions and mobility since 1500 in Europe. Albert-Ludwigs-Universität Freiburg. https://doi.org/10.6094/TAMBORA.ORG/2025/C583/CSV.ZIP

### ACKNOWLEDGEMENTS

This work was supported in part by the DEMUR Project within the DFG Priority Program “On the Way to the Fluvial Anthroposphere” (DFG-SPP 2361). This paper benefited from the participation in the Climate Reconstruction and Impacts from the Archives of Societies (CRIAS) working group of the Past Global Changes (PAGES) project.

**ORCID**

MICHAEL KAHLE
https://orcid.org/0000-0001-8571-2821

RÜDIGER GLASER
https://orcid.org/0000-0001-6819-2764