FH Aachen, Campus Jülich

Fachbereich 09
Medizintechnik und Technomathematik
Studiengang Angewandte Mathematik und Informatik

Seminararbeit

Approaches to Create a SYNOP Station Dataset for Machine Learning in the WeatherGenerator model

Till Hauer
Matr.Nr.: 3622274
till.hauer@alumni.fh-aachen.de

1. Prüfer: **Prof. Dr. rer. nat. Matthias Grajewski**

2. Prüfer: **Prof. Dr. Martin Schultz**

Jülich, December 15, 2025

Declaration

I hereby declare that I have written this seminar thesis independently and have used no other sources and aids than those indicated.

Name: Till Hauer

Jülich, date: December 15, 2025

Signature

The work was carried out at Forschungszentrum Jülich GmbH in the Jülich Supercomputing Center (JSC).

Abstract

Machine learning-based weather forecasting models such as the WeatherGenerator and RAINA require high-quality, standardised observational data for training and operational deployment. This work presents a comprehensive approach to create a SYNOP (Surface Synoptic Observations) station dataset from raw DWD (Deutscher Wetterdienst) observations at 10-minute temporal resolution. We developed a fully automated, reproducible data processing pipeline that transforms over three decades (1990-2024) of heterogeneous observation data from hundreds of weather stations across Germany into four progressively refined dataset versions. The pipeline included data quality checks including gap filling, duplicate resolution and physical constraint validation. In addition, we applied a comprehensive quality assessment that evaluates distribution characteristics, variance homogeneity across stations, inter-annual variability analysis and feature completeness. Inter-annual variability analysis revealed solid measurement quality across most variables, with observed variability largely reflecting genuine atmospheric processes rather than sensor issues. Even though precipitation already has 30% missing values, it demonstrates relatively high data completeness compared to at least 70% missing values for all other features, making the dataset especially valuable for RAINA's focus on extreme precipitation and flood events. Spatial variance heterogeneity analysis confirmed expected geographical differences, validating the need for station-specific normalisation in machine learning applications. While feature sparsity necessitated removal of extreme wind and most solar radiation variables, the final dataset retains eleven essential meteorological features covering temperature, humidity, wind, precipitation and sky radiation. The modular pipeline architecture supports continuous updates and is publicly available, enabling reproducible high-resolution atmospheric data processing for the German region and could provide a valuable input stream for the WeatherGenerator and RAINA.

1 Introduction

Global climate and weather models, particularly those leveraging advanced machine learning (ML) techniques such as the WeatherGenerator (*Lessig* et al., 2025) and RAINA (*Forschungszentrum Jülich GmbH*, 2025), are critically dependent on consistent, high-quality observational input data. The operational success and predictive accuracy of these models depend upon receiving data standardised to specific meteorological formats. The SYNOP (Surface Synoptic Observations) format, governed by the World Meteorological Organization (WMO), serves as the universal standard for reporting surface weather conditions. It conventionally encodes measurements—including temperature, pressure, wind, and precipitation—into a compact numerical code (FM-12) to facilitate rapid global transmission. Yet, raw atmospheric observation data acquired from diverse sources often lacks the necessary consistency and quality control required for direct usage in the training and deployment of ML models. This disparity between available raw data and required standardised input presents a bottleneck for reliable model training.

This work describes the planning, creation and testing of a reliable input stream for the WeatherGenerator by utilising DWD observation data (*DWD*, 2024a). DWD collects the data in three phases: Historical (updated yearly with quality checks), Recent (updated daily), and Hourly (updated hourly). We pick up at this point with data acquisition by scraping all the data we need and additional quality assessment to standardise the

SYNOP conversion. A primary scientific objective is the evaluation and comparison of different conversion methodologies. The study assesses these approaches based on their accuracy in mapping observation variables to the SYNOP standard, their computational efficiency, and their overall robustness against missing or inconsistent measurements. Furthermore, comprehensive data quality assessments, including variance analysis and an inter-annual variability analysis is applied to the processed data to ensure their reliability and consistency for machine learning applications.
Beyond the scientific evaluation of conversion methods, the work aims to deliver a tangible, operational asset: a fully functional, reproducible data pipeline. This automated pipeline is designed to process new observation data daily, continuously appending it to the existing SYNOP station dataset. The implementation includes robust logging, monitoring and basic reporting functionalities to track data quality over time and automatically detect potential anomalies. The reproducibility and maintainability of the final pipeline structure are critically assessed to guarantee its ability to support sustained, continuous updates necessary for long-term operational use within the WeatherGenerator ecosystem.

2 Related Work

Recent works show that ML-based (machine learning-based) weather models become more accurate than NWP (numerical weather prediction) models and are way faster at inference. The combination of lowering compute requirements for machine learning and rapid advancements in HPC (high-performance computing) and GPU acceleration were the central piece for this big progress in only a few years. Nevertheless, as a result of that, many researchers and engineers reached the point of lacking high quality data to train their models. For image classification (*Wang* et al., 2025) scaled up training data for VLMs (Vision Language Models) from 10 billion examples to 100 billion and compared the results. They showed that even though model performance overall saturated, it still led to improvements in rare cases. Low-resource languages as well as diversity of the output benefited the most. We expect the same behaviour when applying their technique to atmospheric data. Utilising high resolution data in multiple modalities over many features in a smaller region may not have an impact on model performance overall. But it might lead to the model being trained more diverse and therefore having the ability to forecast extreme events better.
For weather forecasts, (*Zhang* et al., 2025) shows that numerical models outperform AI in predicting record-breaking extremes. While AI excels at average conditions, it struggles with unprecedented events, likely due to an implicit cap around extreme training observations. In contrast, physical principles allow numerical models to extrapolate beyond the training domain. Although newer probabilistic AI models reduce smoothing, they likely face similar challenges with out-of-distribution extremes.
(*Mirowski* et al., 2024) addresses the challenge of managing the massive volume of atmospheric data, noting that recent climate research generates 260TB of data every 16s. While higher resolution offers better modeling opportunities, the sheer size of atmospheric states poses a significant bottleneck for training and serving models. Furthermore, they highlight that naive compression methods focusing solely on minimising MSE (Mean Squared Error) can lead to scientifically problematic distortions, such as the erasure of extreme events. This underscores the importance of our ap-

proach: by focusing on high-resolution station data for only Germany, we can provide the necessary granularity to capture local extremes without the prohibitive data volume or compression artifacts associated with global high-resolution states.
(*Sha* et al., 2021) did something similar by only focusing on British Columbia in Canada and utilising CNNs (Convolutional Neural Networks) to classify bad precipitation observation data for Quality Control. Even though we do not use deep learning for QC (Quality Control) they also addressed the challenge of data sparsity arising from missing precipitation measurements. We already plan to use the quality flags from DWD to plug in to our model, so it can learn correlations between measurement quality and the provided quality flags by DWD. Nevertheless (*Sha* et al., 2021) prioritised to minimise Type 2 errors (failing to detect bad data) over Type 1 errors (incorrectly flagging good data as bad). In theory this is a good procedure as it ensures that most low quality measurements are removed by deleting some good measurements as well, which results in a dataset of truly high quality. But the good quality data that is removed are often the rare cases that we care about. Therefore, we focused on minimising Type 1 errors at the cost of Type 2 errors resulting in low quality data still being in the dataset. But as we only remove highly unrealistic patterns and measurements, we ensure that we keep the extreme events we care about to predict catastrophes.
(*Nguyen* et al., 2023) introduced ClimateLearn, an open-source PyTorch library designed to standardise the training and evaluation of machine learning models for data-driven climate science. By offering accessible pipelines for processing major gridded datasets, ClimateLearn has facilitated the application of deep learning techniques by domain experts, supporting the broader adoption of data-driven methods in weather and climate science.
Building on this goal of accessibility, we address the need for standardised station-based data. We develop a comprehensive processing pipeline for DWD SYNOP observations, transforming raw reports into standardised machine learning datasets that complement existing gridded resources and support the development of models capable of capturing local meteorological phenomena.

3 Data and Experimental Setup

We used the air temperature (*DWD*, 2024b), extreme temperature (*DWD*, 2024c), wind (*DWD*, 2024g), extreme wind (*DWD*, 2024d), precipitation (*DWD*, 2024e) and solar (*DWD*, 2024f) dataset from the German observations data from DWD in 10 minutes time resolution. The datasets always contain the DWD station id and time of measurement in the beginning of every entry. While station id is constant for every entry in a file, time of measurement is incremented by ten minutes. For all datasets the first measurement was tracked in 1989. In the DWD datasets -999 is always used for missing values in the measurements. The data is publicly available on the DWD Open Data server. We selected a 10-minute temporal resolution, optimising the trade-off between temporal granularity and data volume. The server structure organises data by resolution and variable type. Within each variable category, data is further divided into 'historical', 'recent', and 'now' periods, alongside metadata. Each subdirectory contains individual compressed files for every station that can be downloaded. The data is collected at stations from DWD and qualitatively equal partner networks (*DWD*, 2024b; *DWD*, 2024c; *DWD*, 2024g; *DWD*, 2024d; *DWD*, 2024e; *DWD*, 2024f). Multiple quality tests are ap-

plied to the data (*Kaspar* et al., 2013). First the quality flag is applied directly at the stations. Subsequently, the data undergoes two stages of automated quality control in the central database using the QualiMet software (*Spengler*, 2002). The second check applies broad climatological limits to identify gross errors, while the third check imposes more strict, variable-dependent thresholds. This stage includes comprehensive consistency tests: temporal consistency to detect unrealistic jumps, internal consistency to ensure physical plausibility between related variables (e.g., dew point temperature cannot exceed air temperature) and spatial consistency to compare measurements against neighboring stations. A final check is applied after an entire month of data is available and then the quality flags are adjusted accordingly. All the data is provided as observed by the instruments at the stations and the only adjustment is the quality flag. The files are all named in the same pattern. First the time resolution, then the variable category, then the station id, then the time range and finally the file type. A typical filename for an air temperature dataset follows this structure:

`10minutenwerte_TU_00003_19930428_19991231_hist.zip`

In this section we describe those datasets, explain why we used them and how we pre-processed the datasets.

3.1 Datasets summaries

As shown in Fig. 3.1 the data is not evenly distributed between variable categories and files. As a result, precipitation is measured at more stations than other variables.

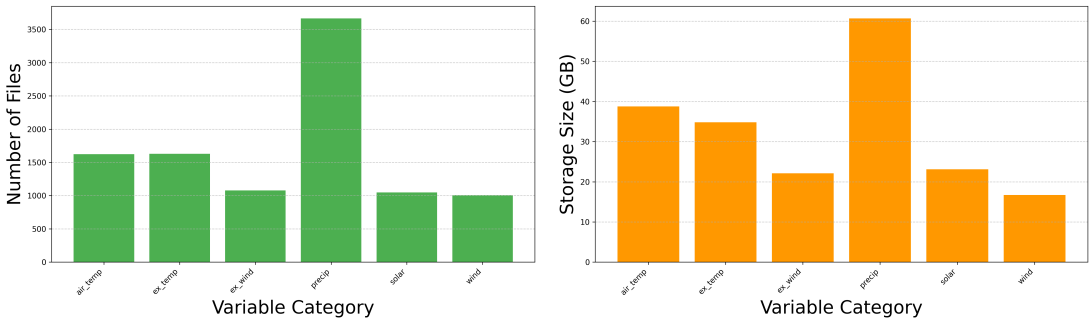

Figure 3.1: Storage and Inode usage of raw downloaded DWD data.

air temperature (*DWD*, 2024b): Relative humidity, air pressure at station level, air temperature near ground level (5cm), air temperature at 2m and dew point temperature are the measurements that are stored in the air temperature dataset. The dew point temperature is calculated with the air temperature at 2m and the relative humidity.
extreme temperature (*DWD*, 2024c): Maximum and minimum air temperature near ground and maximum and minimum air temperature at 2m are the measurements that are stored in the extreme temperature dataset.
wind (*DWD*, 2024g): Standard deviation of longitude and latitude windspeed as well as mean wind speed and wind direction are the measurements that are stored in the wind dataset.
extreme wind (*DWD*, 2024d): Wind direction of maximum wind velocity, minimum and maximum wind velocity are the measurements that are stored in the extreme wind dataset.

precipitation (*DWD*, 2024e): Precipitation height, duration of precipitation and a precipitation fallen indicator as either 0 or 1 is part of the precipitation dataset. In old measurements the precipitation fallen indicator could have been 2 or 3 as well for signaling that heating is in operation. Missing values are represented as -999 here as well.
solar (*DWD*, 2024f): Diffuse sky radiation, global radiation, sunshine duration and longwave radiation are the measurements that are stored in the solar dataset.

3.2 Why we chose the Data

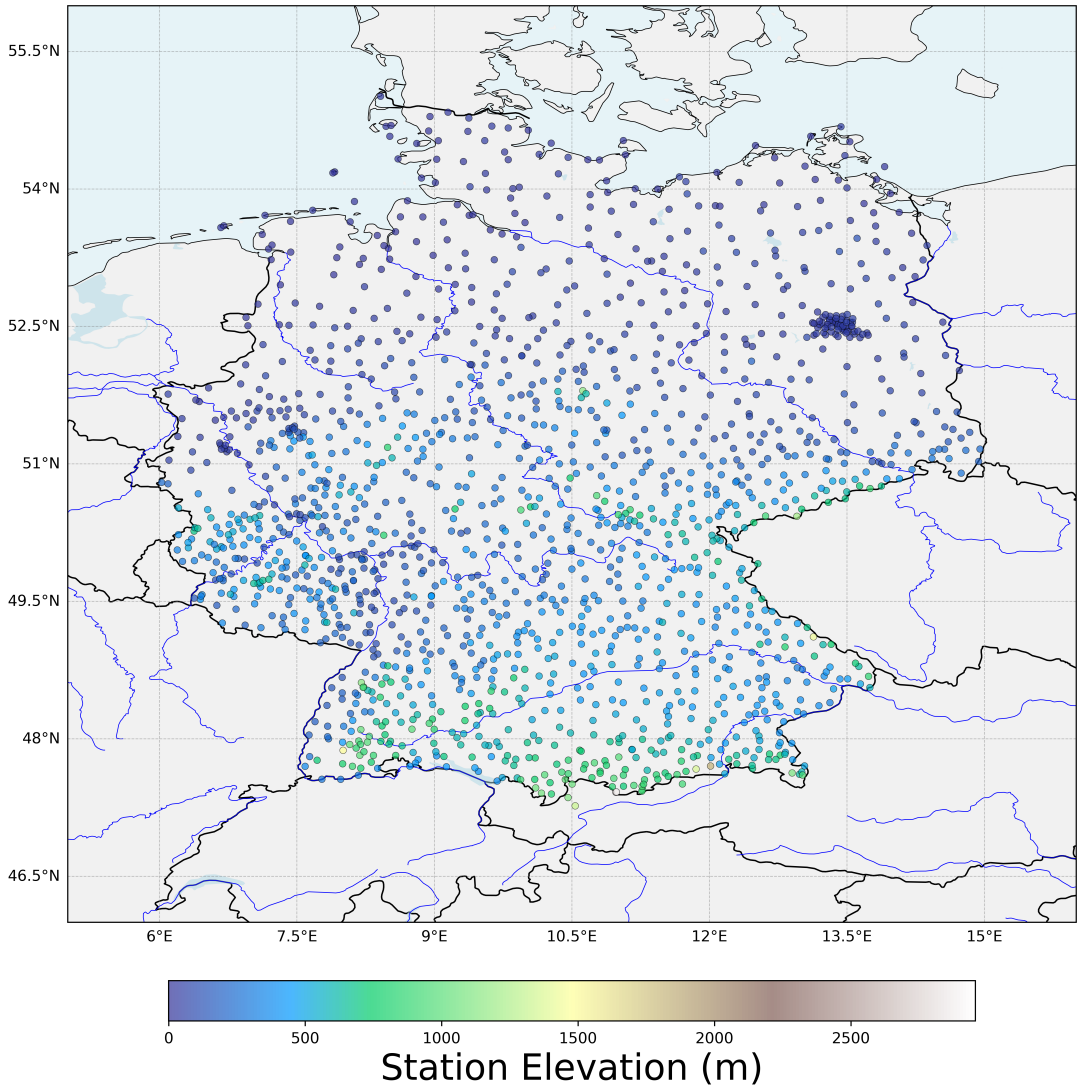

Figure 3.2: Distribution of weather stations in Germany.

The open-source SYNOP data provided by the DWD in high spatial and temporal resolution is particularly valuable in the context of RAINA (*Forschungszentrum Jülich GmbH*, 2025). Leveraging the WeatherGenerator foundation model, RAINA aims to predict extreme weather events in the short term by utilising data at resolutions as fine as 1km. In addition to reanalysis and remote sensing datasets, SYNOP observations serve as a crucial input source, providing detailed measurements of temperature, wind and precipitation, which are the variables RAINA also tries to predict.

These variables are essential for accurately forecasting extreme weather events. The high-resolution DWD data does not only capture their fine-scale variability but also records extreme values of temperature and wind, which are critical for improving predictive performance over Germany.
Fig. 3.2 shows the distribution of weather stations in Germany. The stations are distributed relatively evenly over the country, with a slightly higher concentration in the south and west. Nevertheless we still have high resolution data for the north and east as well. We can also see that in the north most stations are at surface level and in the south towards the Alps more stations have higher altitudes often reaching over 1000m. Additionally, we can see that there is a dense cluster of stations in Berlin and the surrounding area. The DWD stations also cover some islands in the North Sea and Baltic Sea along the northern coast of Germany.
Precipitation presents a particular challenge, as its height and duration are highly irregular, extreme events are difficult to predict and many measurements are zero, representing no precipitation. The inclusion of a "precipitation fallen" indicator helps the model better understand precipitation dynamics, providing context that can improve learning despite these challenges. In addition to that we also include solar data, which helps the model to predict temperature and precipitation as well as it correlates with cloud cover and also contains sunshine duration values, which should help the model to predict temperature developments.
The solar data can even help with the WeatherGenerator (*Lessig* et al., 2025), when trying to predict solar production, which is a downstream application of the model. Overall, the combination of high-resolution SYNOP observations with RAINA lays the foundation for accurate and localised short-term predictions in Germany.

3.3 Data pre-processing

DWD uses -999 to represent missing values. However, we convert all missing values to np.nan to ensure compatibility with xarray and NetCDF, which require consistent data types. We also use np.nan to fill temporal gaps between and within files for the same station, ensuring a continuous and consistent time series. The data pre-processing was parallelised over almost 100 CPUs as its calculations are highly independent and therefore the speed up was almost the number of CPUs.

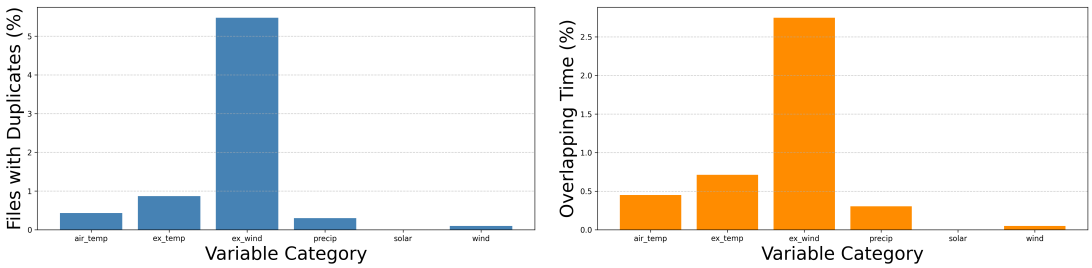

Figure 3.3: Files that have duplicate sections with other files and the duplicated time per variable category.

We identify duplicated files and missing data gaps by analysing the start and end dates encoded in the file names. A duplicated file is defined as one that has a temporal overlap with another file from the same station and variable category. This occurs when

the start date of a file is earlier than the end date of the preceding file in the sorted sequence for that station. To resolve this, we compare the duration of the overlapping files and delete the one with the shorter time span to maximise data retention. Fig. 3.3 differentiates between the prevalence of this issue and its temporal impact. The left panel shows the proportion of files affected by overlaps, while the right panel displays the percentage of total time that is duplicated. We can deduce that while up to 15% of extreme wind files, which has the most duplicates, are affected, the actual duplicated time remains negligible. This indicates that overlaps are frequent but typically short in duration. Solar data has no duplicates at all.

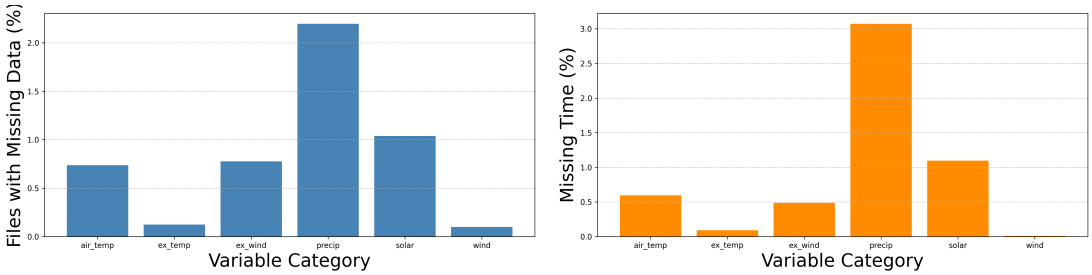

Figure 3.4: Files that can't be connected without missing values and the total missing time per variable category.

We identify missing data as gaps between consecutive files for the same station that exceed 24 hours. These gaps represent periods where no data files were available to form a continuous time series. To ensure a consistent time grid, we create new files for these gaps and fill them with the missing value indicator (-999), which is later converted to NaN. Therefore Fig. 3.4 shows that solar has relatively more missing periods between files, while precipitation reaches over three percent of its total time missing between files.

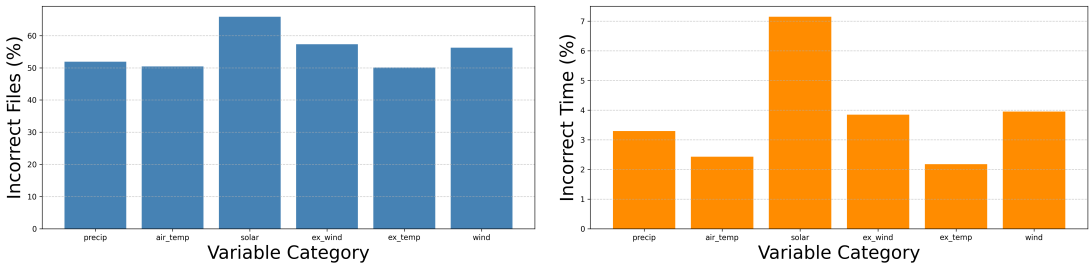

Figure 3.5: Files that do not contain the time values given by file names and the total wrong time per variable category.

We also verify the internal consistency of each file by comparing its actual content with the time range specified in its filename. A file is flagged as incorrect if its start or end time does not match the filename, or if there are missing 10-minute intervals within the file. To resolve these inconsistencies, we pad the missing time steps whether at the beginning, end or internally with -999. This ensures that every file strictly adheres to the 10-minute temporal resolution and matches its designated time range. Fig. 3.5 shows that while nearly half of the files required such corrections, the total amount of missing time added is generally low, often spanning only a few hours to days. Solar data, how-

ever, exhibited the highest proportion of missing time, accounting for approximately five percent of its total measurements.

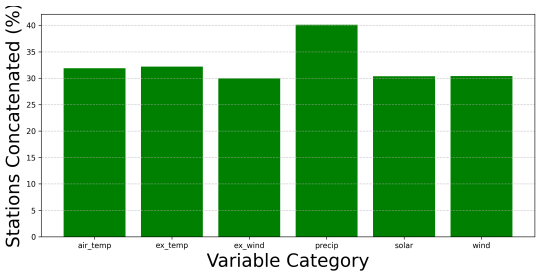

Figure 3.6: Concatenated Files per file in the beginning

Ultimately, all files were concatenated. Fig. 3.6 shows that only about one-third of the initial files remained. Therefore, one stations measurements were on average split over three files.

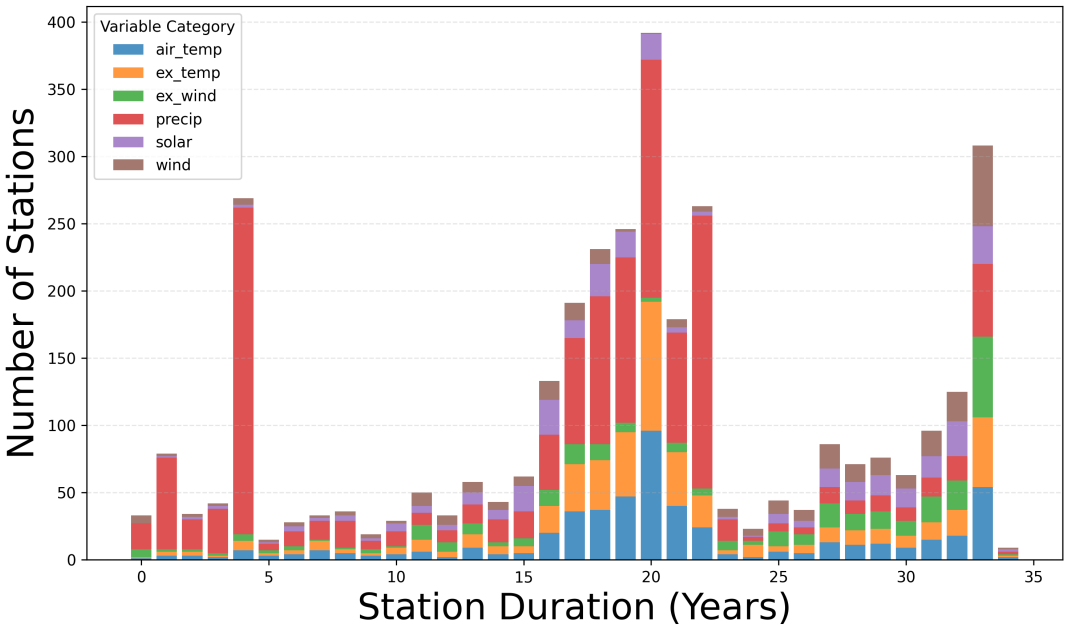

Figure 3.7: Years of duration for every station id categorised by dataset.

Fig. 3.7 visualises that precipitation has a lot of stations with short durations and a lot with around 17 to 22 years. Even though precipitation has the most stations it has the least when it comes to stations with consistent measurements for over 25 years. Air temperature and extreme temperature has a lot of stations with long term measurements. When looking at the durations without precipitation most stations have more than 10 years of total duration.

4 Methodology

Our data processing pipeline transforms raw DWD observations into an analysis-ready dataset through four progressive stages. The first stage consolidates the preprocessed

data into yearly NetCDF files, creating continuous time series from fragmented source files. The second stage applies comprehensive quality control procedures to identify and handle problematic values while preserving extreme weather events. The third stage addresses feature sparsity by removing features with excessive missing data and retaining only those features with sufficient completeness for reliable model training. The final stage applies interpolation and extrapolation techniques to fill small temporal gaps in the retained features. Each stage produces a versioned dataset (v1-v4), allowing users to select the appropriate level of processing for their specific applications. We do not detail the basic NetCDF conversion process here, as the preprocessing steps (file concatenation, duplicate resolution, temporal alignment) were already described in the Data and Experimental Setup section and represented the main work to get to the first version of the dataset (v1). In the following subsections, we focus on the quality checks, statistical analyses, and gap-filling procedures that constitute the core of our methodology.

4.1 Quality checks

For the quality checks we have performed multiple checks to remove problematic values and analyse data integrity. The first check analysed NaN and inf (infinity) values as a standard procedure. We'll go more over the NaN values in the next subsection when addressing feature sparsity. The standard procedure would have been to replace inf values with NaN values, but as expected we didn't find any so those checks did not have an impact on the data. The next step is the analysis and removal of physically unreasonable values for German observation data. In this case we of course ignored the metadata (longitude, latitude, height and QN (quality flag)), when removing problematic values, but still analysed them to see if they fit for German stations. We didn't analyse the quality flags by DWD as they are also based on station measurement devices.

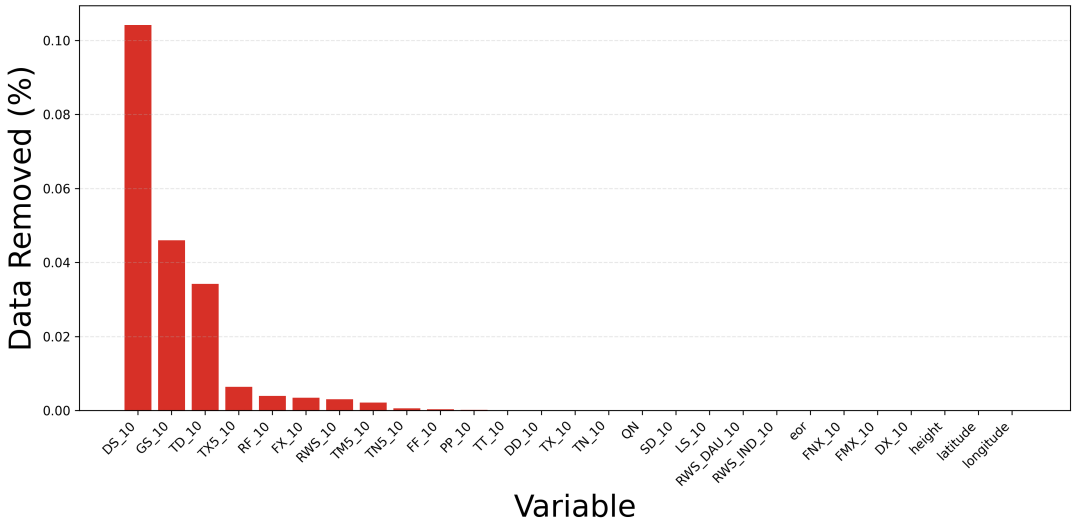

Figure 4.1: Cleaning Impact

As we can see in Fig. 4.1 the other metadata values are all correct either way. Sky radiation (DS_10) relatively has the most values that need to be removed according to physical consistency [0, 350], but this still only represents 0.1% of the entire fea-

ture, which is then replaced with NaN. The sky radiation (DS_10) and global radiation (GS_10) performed more poorly than the other values as the measurement of radiation is not as precise as temperature or precipitation, because it is only measured indirectly and the sensors can be affected by other weather conditions. After that we only analyse the time series for quality checks but do not remove the values as this would impact the data unnecessarily.
Either way users can still go with the first version of the data easily, when they want the raw data, that was just pre-processed without any adjustments with the values. However, since the infinite value check had no impact, the removal due to physical constraints is the only replacement with NaN values that is performed to create the next version of the dataset and therefore also the entire cleaning impact in Fig. 4.1. This way other developers are still able to include other quality checks that they want to use to filter the data.

Table 4.1: Physical constraints used for quality control. Range indicates valid min/max values. Max Change indicates the maximum allowed absolute difference between consecutive 10-minute measurements.

Feature	Description	Value Ranges	Max Change
DD_10	Wind direction	[0, 360]	-
DS_10	Diffuse radiation	[0, 350]	-
DX_10	Wind direction at max gust	[0, 360]	-
FF_10	Wind speed average	[0, 45]	30.0
FMX_10	Max wind gust average	[0, 55]	40.0
FNX_10	Max wind gust	[0, 55]	40.0
FX_10	Max wind gust	[0, 55]	40.0
GS_10	Global radiation	[0, 500]	-
LS_10	Long-wave radiation	[0, 500]	-
PP_10	Station pressure	[600, 1100]	5.0
RF_10	Relative humidity	[5, 100]	30.0
RWS_10	Precipitation amount	[0, 100]	-
RWS_DAU_10	Precipitation duration	[0, 10]	-
RWS_IND_10	Precipitation indicator	[0, 3]	-
SD_10	Sunshine duration	[0, 600]	-
TD_10	Dew point temperature	[-20, 40]	5.0
TM5_10	Temperature at 5cm	[-50, 45]	5.0
TN5_10	Min temperature at 5cm	[-50, 45]	5.0
TN_10	Min temperature	[-50, 45]	5.0
TT_10	Air temperature	[-50, 45]	5.0
TX5_10	Max temperature at 5cm	[-50, 45]	5.0
TX_10	Max temperature	[-50, 45]	5.0

The physical constraints used for these checks are detailed in Table 4.1. We have decided to go with rather weak constraints to minimise Type 1 errors. For example the temperature constraints were chosen in a way that the upper bound is higher than the max temperature ever measured in Germany and the lower bound is below the lowest temperature ever measured in Germany. Regarding temperature we only chose a different range for the dew point temperature, as it generally does not reach such low and high values. Additionally, it has a strong correlation with humidity and is therefore

also highly seasonal. During winter the dew point is around 0 °C and during summer it is around 18 °C and a bit more volatile than in the colder seasons. Therefore, we went with a range of [-20, 40] °C for the dew point temperature to truly only filter out unrealistic measurements. For humidity we went with a range of [5%, 100%], because humidity often reaches 100%. This can even occur on summer nights. But humidity below 5% was never reached in Germany. These conditions occur only in desert regions along the Tropic of Cancer and the Tropic of Capricorn, but they are most commonly found near the Tropic of Cancer. For station pressure we went with a range of [600hPa, 1100hPa], because station pressure can reach low values at high altitude. But it rarely reaches values over the normal pressure at surface, because it is hard to go down from that level without surpassing sea level. We also applied a buffer for the upper bound as there is still some seasonality in station pressure.
Max Change in Table 4.1 refers to the maximum allowed absolute difference between two consecutive 10-minute measurements. We also applied generous thresholds for the maximum change per 10 minute interval. This threshold helps identify unrealistic jumps in the time series, such as those caused by sensor errors, while still preserving legitimate extreme weather events like cloudbursts or wind gusts. There are some features with physical constraints, but without any max change per 10 minute interval. We indicated this in the table by using a dash. All metadata is ignored as the changes between two stations have no correlation, except that they both have to be in Germany, which is fully addressed in the physical constraints. Additionally we ignore all sudden shifts in radiation and only use the physical constraints as radiation can shift in seconds through the movement of clouds and shadows in general. We also ignore sudden shifts in precipitation, as heavy rain can also start in short times without any prior developments of rain in a location. Lastly we ignore changes in the duration of the rain, the rain indicator flag as well as the radiation duration.
We also performed distribution analysis to assess whether the atmospheric variables follow normal distributions, which is relevant for selecting appropriate statistical methods and understanding data characteristics. We applied the Kolmogorov-Smirnov (KS) test for normality to each variable. The KS test compares the empirical cumulative distribution function (ECDF) of the observed data against the theoretical cumulative distribution function (CDF) of a normal distribution with the same mean and standard deviation as the sample. The test statistic D is defined as:

$$D = \sup_x |F_n(x) - F(x)| \tag{1}$$

where $F_n(x)$ is the empirical distribution function and $F(x)$ is the theoretical normal distribution function. We used a significance level of $\alpha = 0.05$, meaning that variables with $p > 0.05$ are considered approximately normally distributed. To ensure statistical validity, we required a minimum of 20 valid (non-NaN) data points per variable for the test to be performed. Variables with insufficient data were excluded from the normality assessment.
The KS test revealed that none of the atmospheric variables exhibited normal distributions at the 0.05 significance level. This finding is consistent with the physical nature of atmospheric measurements, which often display skewed distributions due to physical constraints (e.g., precipitation cannot be negative, wind speed has a lower bound of zero) and the presence of extreme weather events that create heavy-tailed distributions. Additionally, we calculated skewness and excess kurtosis for each variable

to quantify the degree and nature of deviation from normality. These distribution characteristics inform preprocessing decisions for machine learning applications, such as whether to apply transformations (e.g., log-transform for highly skewed variables) or use robust scaling methods that are less sensitive to outliers. Unfortunately we could not create any insights from skewness and kurtosis, as the distributions are all skewed.
To assess spatial consistency across measurement stations, we performed variance homogeneity tests using Levene's test (*Levene*, 1960). This analysis evaluates whether different weather stations exhibit similar variability for each atmospheric variable, which is important for ensuring data quality and identifying potential calibration issues or local effects. Levene's test compares the variances across multiple groups (in our case, stations) by testing the null hypothesis that all groups have equal variance. The test statistic is computed by performing an ANOVA (analysis of variance) (*Fisher*, 1925) on the absolute deviations from group medians. We required a minimum of 10 valid measurements per station and at least 2 stations with sufficient data for the test to be valid. A significance level of $\alpha = 0.05$ was used, where $p > 0.05$ indicates homogeneous variance across stations.
As shown in Fig. 4.2, most variables exhibit heterogeneous variance across stations, with variance ratios (max/min variance) often exceeding 5. This heterogeneity is expected for atmospheric data due to geographical differences. Coastal stations experience different wind patterns than inland stations, mountainous regions show greater temperature variability than plains and urban heat islands affect local temperature measurements. The high variance ratios for radiation variables (DS_10, GS_10) and wind variables (FF_10, FX_10) reflect the strong influence of local topography and exposure conditions on these measurements. Temperature variables show more moderate variance ratios, indicating relatively consistent measurement conditions across Germany. These findings validate the need for station-specific normalisation or robust scaling methods that account for location-dependent variability when preprocessing data for machine learning models.

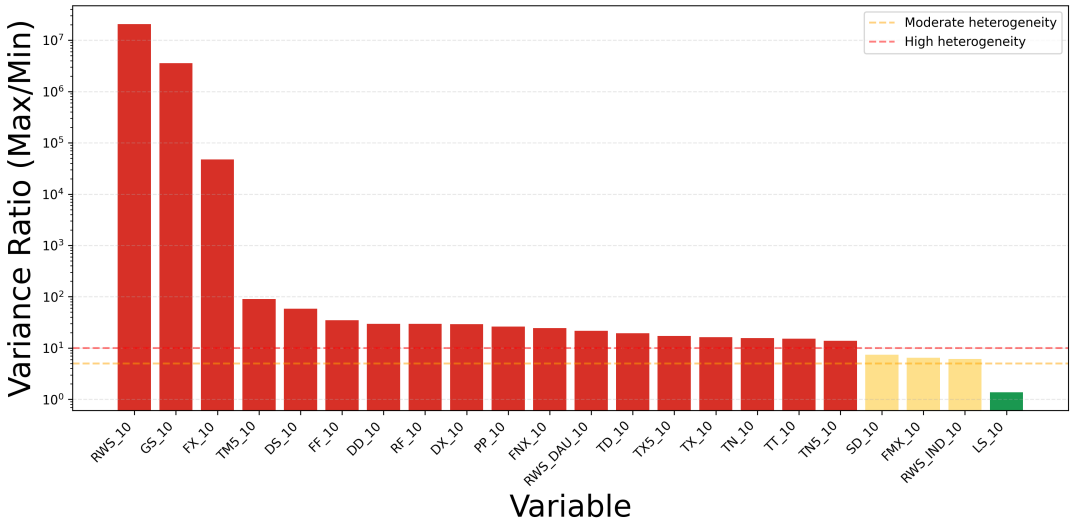

Figure 4.2: Variance Homogeneity

To evaluate temporal consistency of measurements across different time periods, we performed inter-annual variability analysis by computing the coefficient of variation (CV)

of yearly statistics for each variable and station. Unlike stationarity tests (e.g., Augmented Dickey-Fuller or KPSS (Kwiatkowski-Phillips-Schmidt-Shin) tests) that assume same conditions over the entire time series, our approach accounts for the seasonal nature of atmospheric data by analysing year-to-year variability. For each station and variable, we calculated the mean and standard deviation for each year, requiring a minimum of 30 valid measurements per year for statistical reliability. We then aggregated these statistics across all years for each station and computed the coefficient of variation: $\text{CV} = \sigma/|\mu|$, where σ is the standard deviation of the means (or standard deviations) across years and μ is the mean of those values. The CV quantifies relative variability, with lower values indicating more consistent measurements over time.

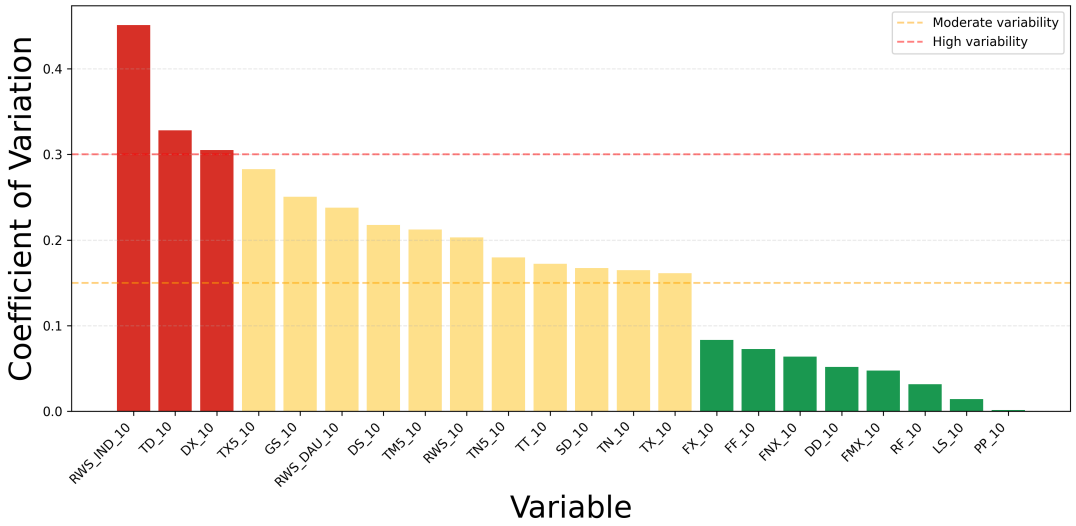

Figure 4.3: Inter-annual Variability Analysis Results

As illustrated in Fig. 4.3, the atmospheric variables exhibit a range of coefficients of variation (CV), reflecting different degrees of temporal consistency across the measurement period. Most variables show moderate CVs in the range of 0.15 to 0.30, indicating some year-to-year variability while maintaining overall consistency. Temperature variables such as air temperature (TT_10, TM5_10) show moderate CVs around 0.17 and 0.21 respectively, which is expected as temperature is more volatile near the surface level due to direct solar heating, energy absorption/re-emission by ground/urban surfaces and less atmospheric insulation, causing rapid warming/cooling cycles. Humidity (RF_10) exhibits low CV around 0.04. This really low CV is a result of its physical constraints being between 0% and 100%. Therefore no trends can establish over multiple years and therefore humidity mostly reverts to the means of the prior years. We can see the worst results are for dew point temperature (TD_10), because RWS_IND_10 can be neglected, as it is only the precipitation indicator. The lower inter-annual variability for dew point temperature is expected though as it is more sensitive to changes in atmospheric conditions, leading to more year-to-year variability. The directional wind variables show more varied behavior. Wind direction at max gust (DX_10) has one of the highest CV values, while regular wind direction has one of the lowest (DD_10). This is expected as average wind tends to be consistent over a measurement period, while wind gusts are instantaneous events associated with atmospheric turbulence and chaotic, localised air movements. Precipitation (RWS_10) shows moderate CV values, consistent with the high inter-annual variability of precipitation patterns in Germany.

While the CVs are not uniformly low, most variables remain below 0.30, suggesting that sensor calibration and measurement protocols have remained reasonably stable over the multi-decade observation period.

4.2 Removal of Sparse Features

We have decided that we then remove features that are too sparse after applying the numerical checks. We have removed all features with more than 90% NaN values. This high threshold was chosen to balance the need for data completeness with the risk of introducing significant imputation bias to a variable that is almost entirely missing. We are also able to go with this high threshold, due to the fact that we have a lot of data and therefore we can still use features with up to 90% NaN values, because that should be enough for the model to learn patterns.

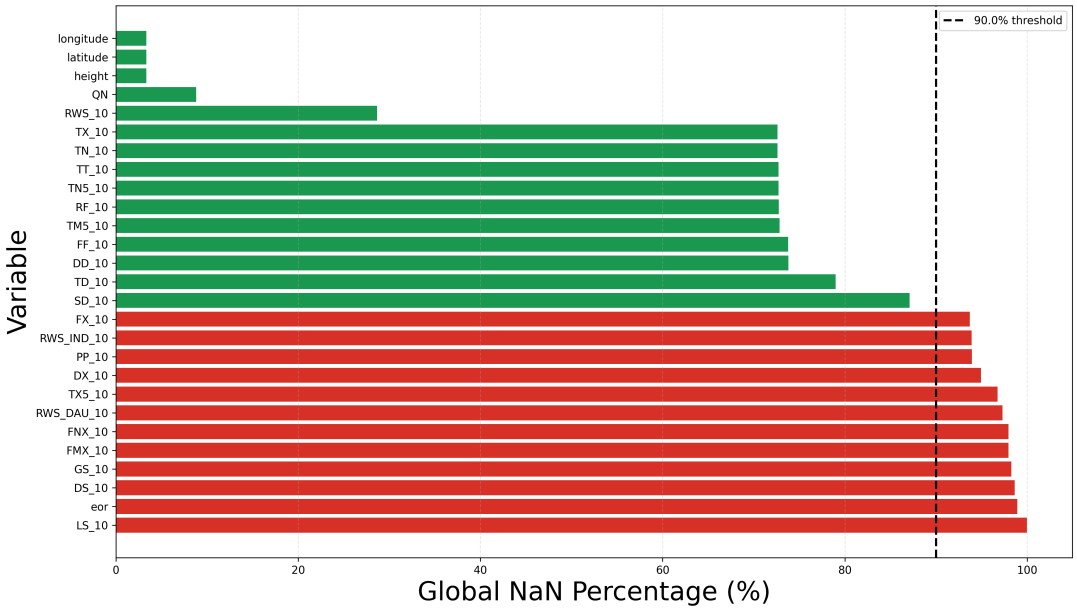

Figure 4.4: Sparsity of Features and Metadata

Fig. 4.4 shows the sparsity of the features and highlights the removed features in red as well as the removal threshold, which is represented in a dashed black line. The first thing that is visible is that all extreme wind features are removed. This is really unfortunate as they play a crucial role in extreme events. Additionally, every solar feature except sky radiation is removed as well as station pressure, the rain indicator and rain duration. The last feature that is removed is the maximum air temperature at 5cm above the ground. Even though the regular maximum temperature at 2m above the ground is still in the dataset and correlates with the one at 5cm, it's still a loss as the 5cm one is more volatile and therefore brings extra atmospheric details that are not present in the 2m one. Therefore, all non extreme air temperature and all wind features are retained. Nevertheless, they still suffer from sparsity. Every feature except the precipitation sum over all minutes has a NaN percentage of at least 70%. Fortunately the precipitation sum only has a NaN percentage of 30%, because this is the most important feature for the RAINA project, where we are explicitly interested in extreme rain and flood events (*Forschungszentrum Jülich GmbH*, 2025). We can also see that there

are around 3% of stations with missing metadata, which could be removed entirely, but we decided to keep them as they can still be used for model training. We should make sure though that they are ignored in validation and testing as especially height is an important metadata to forecast other metrics. Due to only having German stations longitude and latitude are not as crucial, because the conditions do not differ heavily through the country. The most significant difference is probably between the wind in the south and the north.

4.3 Interpolation and Extrapolation of Remaining Features

The last changes we made to the data for a fourth version is to interpolate and extrapolate the remaining features. We have used a combination of linear interpolation and extrapolation to fill small gaps in the data as well as appending to the edges of the existing data. We use linear interpolation first before considering linear extrapolation, as it should perform better on comparable cases. Therefore we also are more confident with linear interpolation and enable bigger gaps to be interpolated than extrapolated. While we enable interpolation to fill gaps up to a size of six NaN values, extrapolation can only fill gaps of two values. For linear interpolation, we require the valid value immediately preceding the gap, (x_A, y_A), and the valid value immediately succeeding the gap, (x_B, y_B). For extrapolation, we require two valid, consecutive values to define the linear function, which we denote as (x_1, y_1) and (x_2, y_2). As we first apply interpolation, all intermediate gaps containing up to six missing values ($n \in \{1, 2, \ldots, 6\}$) are filled.

$$Y(x_n) = y_A + \frac{(x_n - x_A)(y_B - y_A)}{x_B - x_A}$$

After that, extrapolation can be applied in three cases:

1. **Missing values at the beginning of the file:** The last two valid points are used to extrapolate backwards, provided there are at least two valid values available.
2. **Missing values at the end of the file:** The first two valid points are used to extrapolate forward, provided there are at least two valid values available.
3. **Intermediate gaps larger than 6:** The first two and last two missing values of the gap are extrapolated if the surrounding valid segments are at least two values long.

Let (x_1, y_1) and (x_2, y_2) be the two consecutive valid points used to define the linear function. The extrapolation is performed using the general form of the line:

$$m = \frac{y_2 - y_1}{x_2 - x_1}$$
$$Y(x) = y_2 + m(x - x_2)$$

This function is applied for both forward and backward extrapolation based on the case.

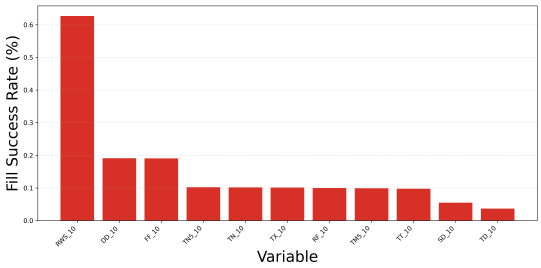

Figure 4.5: Percentage of Gaps Filled.

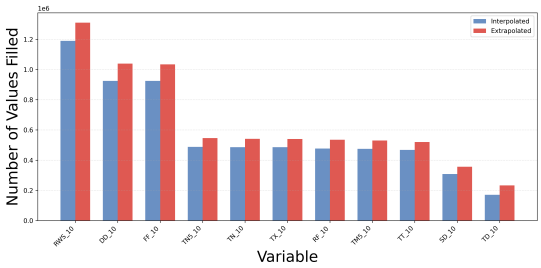

Figure 4.6: Impact of Inter- and Extrapolation.

Fig. 4.5 shows the percentage of gaps filled for each variable, while Fig. 4.6 shows the impact of each method on the overall percentage of gaps filled. It is visible that both methods contribute equally to the reduction and that there is no difference between variables. The impact of interpolation in comparison to extrapolation stays consistent over all features. Fig. 4.7 visualises the total reduction of NaN values. It does not only show us that inter- and extrapolation had almost no impact on the total amount of NaN values, but also that values exist consistently over long periods followed by longer periods of NaN values.

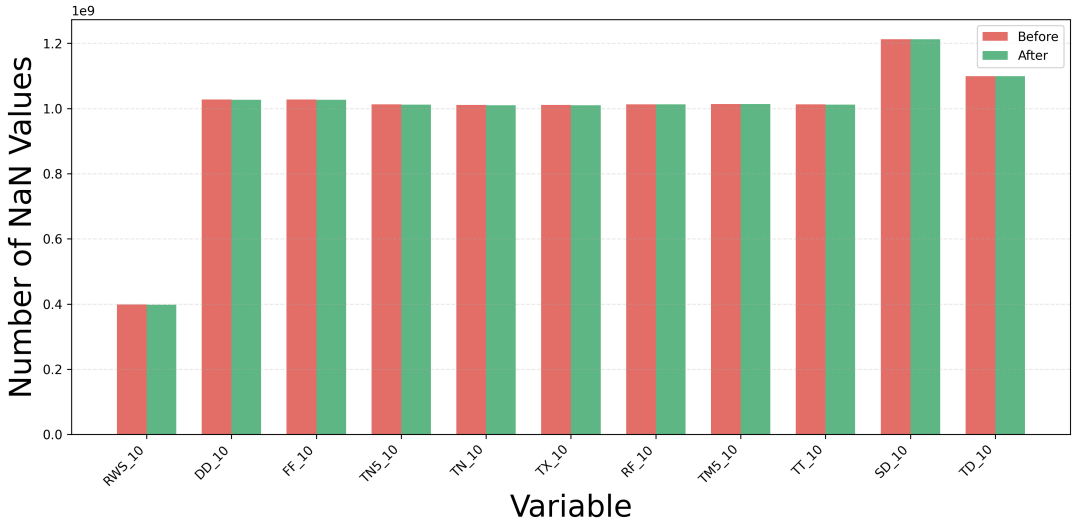

Figure 4.7: NaNs Before and After Gap Filling.

This is partially due to our attempt to create continuous time series by filling them with NaN values. Often these gaps that we filled spanned over multiple years, which have a massive impact on overall NaN percentage in the datasets. Nevertheless, this is also a consequence of long periods without measurements by DWD, where they just used -999 to fill the gaps. Even though this hinders us in trying to recreate the original values, it is an advantage, when training models, as they are dependent on continuous time series with existing values to learn representations of the physical relationships between the features and upcoming states of the atmosphere.

5 Results

The implementation of our data processing pipeline resulted in the creation of four progressively refined versions of the SYNOP station dataset, each addressing specific data quality and completeness objectives. All code, processing scripts, and pipeline infrastructure are publicly available in our GitHub repository (*Hauer*, 2025).

5.1 Dataset Versions

Version 1 (v1): The initial version focused on basic preprocessing and file concatenation. Starting from the raw DWD data comprising thousands of individual files across six variable categories (air temperature, extreme temperature, wind, extreme wind, precipitation, and solar), we successfully concatenated files by station and variable type. As shown in Fig. 3.6, we reduced the original file count by a factor of three. This version established continuous time series from 1990 to 2024 for hundreds of stations across Germany, with temporal resolution of 10 minutes. The concatenation process handled duplicate file detection (affecting up to 15% of extreme wind files, Fig. 3.3), missing data gap identification (exceeding 3% of total time for precipitation, Fig. 3.4), and internal file consistency verification (requiring corrections in nearly 50% of files, Fig. 3.5).
Version 2 (v2): This version implemented comprehensive quality control procedures. Physical constraints validation removed 0.1% of measurements from the most affected variable (diffuse sky radiation DS_10), with minimal impact on other variables (Fig. 4.1). The quality checks identified and flagged physically unreasonable values while deliberately using generous thresholds to minimise Type 1 errors and preserve extreme weather events. Statistical analyses revealed that no atmospheric variables follow normal distributions (Kolmogorov-Smirnov test, $p < 0.05$), variance heterogeneity across stations is significant for most variables (Fig. 4.2), and temporal consistency varies by variable with coefficients of variation ranging from 0.04 (humidity) to over 0.30 (dew point temperature and wind direction at maximum gust) (Fig. 4.3).
Version 3 (v3): Feature selection based on sparsity thresholds was applied in this version. We removed all features exceeding 90% NaN values, resulting in the elimination of all extreme wind variables (FMX_10, FNX_10, FX_10, DX_10), most solar radiation features (GS_10, LS_10, SD_10), station pressure (PP_10), precipitation indicators (RWS_IND_10, RWS_DAU_10), and maximum temperature at 5cm (TX5_10). The retained features (Fig. 4.4) include: air temperature variables (TT_10, TM5_10, TN_10, TN5_10, TX_10, TD_10), humidity (RF_10), wind variables (DD_10, FF_10), precipitation amount (RWS_10), and diffuse sky radiation (DS_10). Notably, precipitation amount (RWS_10) exhibits only approximately 30% NaN values, making it the most complete feature and particularly valuable for the RAINA project's focus on extreme precipitation events.
Version 4 (v4): The final version applied interpolation and extrapolation to fill small temporal gaps. Linear interpolation filled gaps up to six consecutive missing values (60 minutes), while linear extrapolation addressed gaps of up to two values (20 minutes) at file boundaries and within larger gaps. As shown in Fig. 4.5 and Fig. 4.6, both methods contributed approximately equally to gap filling, with consistent impact across all variables. However, the overall reduction in NaN percentage was modest (Fig. 4.7), as the majority of missing data occurs in extended multi-year gaps that cannot be reliably

filled through simple linear methods.

5.2 Final Dataset Characteristics

The final dataset (v4) encompasses measurements from more than 1600 weather stations distributed across Germany (Fig. 3.2), with relatively even spatial coverage and elevation ranging from sea level to over 1000m in the Alpine regions. Station duration analysis (Fig. 3.7) reveals that while precipitation has the highest number of stations, air temperature and extreme temperature datasets contain the most stations with consistent measurements exceeding 25 years. Most non-precipitation variables show station durations concentrated between 10 and 25 years.
The dataset retains eleven meteorological features plus four metadata fields (station_id, latitude, longitude, height) across all stations and time periods. Despite the sparsity challenges, the retained features cover the essential variables for weather forecasting: temperature (six features), humidity (one feature), wind (two features), precipitation (one feature), and solar radiation (one feature). The 10-minute temporal resolution provides fine-scale variability crucial for capturing rapid weather transitions and extreme events.
All dataset versions are stored in NetCDF format with comprehensive metadata, enabling efficient access and integration with existing climate science workflows. The modular pipeline architecture supports continuous updates as new DWD data becomes available, with automated quality control and reporting ensuring ongoing data integrity.

5.3 Data Quality Insights

Our quality analyses revealed that observed variability largely reflects genuine atmospheric processes rather than measurement issues. Precipitation (RWS_10) showed a high CV but demonstrates solid quality. The variability stems from precipitation's inherently volatile nature (localised, rapid onset/termination) rather than sensor problems. This is confirmed by minimal values removed during quality control and the lowest sparsity (30% NaN) among retained features.
Temperature variables exhibited moderate CVs (0.17-0.21), potentially reflecting both natural climate variability and long-term warming trends in Germany. Surface temperature (TM5_10) showed higher variability than air temperature at 2m (TT_10) due to direct solar heating and reduced atmospheric insulation. Humidity (RF_10) displayed the lowest CV (0.04), explained by its physical bounds (0-100%) preventing long-term trends.
Wind patterns revealed physically meaningful contrasts: regular wind direction (DD_10) showed a low CV reflecting consistent prevailing patterns, while gust direction (DX_10) exhibited a high CV due to the stochastic nature of turbulent events. Dew point temperature (TD_10) showed the highest CV among meteorological variables, as expected given its sensitivity to simultaneous changes in moisture and temperature.
Spatial variance heterogeneity was highest for radiation (DS_10) and wind variables (DD_10, FF_10), with variance ratios often exceeding five, reflecting topography and exposure differences between coastal, inland, mountainous, and urban stations. This validates the need for station-specific normalisation in ML preprocessing.

6 Evaluation and Analysis

Having presented our methodology and results, we now evaluate the dataset's contributions and limitations within the broader context of atmospheric data processing research. We first compare our approach to related work in neural compression, precipitation quality control, and standardised ML workflows, highlighting how our regional focus and design choices address complementary challenges. We then provide a critical assessment of the final dataset, examining both its strengths for extreme weather prediction applications and the significant limitations that constrain its utility.

6.1 Comparison to Related Work

Compared to the work of (*Mirowski* et al., 2024) on managing massive atmospheric data volumes, our regional focus on Germany provides a complementary solution. Rather than compressing global high-resolution data, we maintain full 10-minute resolution for a focused geographical region. This approach avoids lossy compression, which can smooth out extreme events while still providing the fine-scale temporal variability essential for short-term extreme weather prediction. Our dataset directly supports applications like RAINA that require high-resolution local observations without the computational overhead of global datasets.

Our quality control philosophy differs significantly from the approach to precipitation data in (*Sha* et al., 2021) in British Columbia. While they prioritised minimising Type 2 errors (accepting low-quality data) through CNN-based classification, we deliberately minimise Type 1 errors (rejecting valid data) to preserve extreme events. This design choice is critical given the findings of (*Zhang* et al., 2025), who show that AI models struggle to extrapolate beyond their training distribution. By retaining extreme values that might otherwise be discarded as outliers, we maximize the observational range available to the model, helping to mitigate the implicit performance cap on record-breaking events. Our physical constraint validation removed only 0.1% of measurements for the most affected variable, demonstrating that generous thresholds successfully preserve rare extreme values while still identifying clearly unrealistic measurements. The resulting 30% missing data for precipitation compares favorably to the sparsity challenges they encountered, validating our region-specific approach.

Similar to the goal of (*Nguyen* et al., 2023) of standardising ML workflows for climate science, we provide accessible, reproducible infrastructure for atmospheric data processing. However, while their work focuses on gridded reanalysis datasets, we address the complementary need for high-resolution station observations. Our publicly available pipeline enables researchers to process DWD data without developing custom infrastructure, lowering barriers to entry for station-based weather forecasting research in the German region. The modular architecture and comprehensive documentation support both direct dataset usage and adaptation for other regional meteorological services.

6.2 Dataset Evaluation

The final dataset successfully already achieves its primary objective of providing observational input for WeatherGenerator and RAINA. The 10-minute temporal resolution captures rapid weather transitions and convective events that are critical for extreme weather prediction but often missed in hourly or coarser data. The 30+ year temporal

span (1990-2024) allows for the inclusion of a wide range of historical weather patterns. The dataset's long duration increases the total number of valid observations potentially available for training, which compensates partially for the high data sparsity.
Precipitation data quality represents a particular strength, with only 30% missing values compared to over 70% for most other variables. This completeness is especially valuable for RAINA's focus on flood and extreme precipitation events, where reliable observational data is essential for model validation and training. The inter-annual variability analysis confirmed that observed variability in precipitation reflects genuine atmospheric processes rather than sensor drift, providing confidence in the data's suitability for long-term trend analysis and extreme event detection.
However, limitations remain. Feature sparsity necessitated removal of all extreme wind variables and most solar radiation measurements, limiting the dataset's utility for applications requiring these features. The high missing data percentages (>70%) for retained temperature and wind variables may challenge some ML architectures, though the continuous time series structure enables temporal imputation strategies.
Despite these limitations, the dataset can provide value for regional weather forecasting research. The combination of high temporal resolution, multi-decade coverage, quality assessments, and public availability addresses the need for open-source available ML training data for the German region. The modular pipeline architecture ensures the dataset can grow as new DWD observations become available.

7 Conclusion

This work presents a comprehensive, automated pipeline for transforming raw DWD observational data into a high-quality dataset suitable for machine learning-based weather forecasting applications. We successfully processed over three decades (1990-2024) of heterogeneous atmospheric observations from more than 1600 weather stations across Germany, producing four progressively refined dataset versions that address different user needs and processing requirements.
Our comprehensive quality assessment framework validated the reliability of the processed data. Inter-annual variability analysis revealed solid measurement quality across most variables, with observed variability largely reflecting genuine atmospheric processes. A key achievement is the preservation of high-quality precipitation data. This reliability positions the dataset as a valuable resource for modeling extreme precipitation and flood events.
The pipeline's design philosophy prioritised preservation of extreme weather events over aggressive quality filtering. By minimising Type 1 errors through generous physical constraint thresholds, we retained rare extreme values essential for catastrophic event prediction while removing only clearly unrealistic measurements (0.1% for the most affected variable). This approach distinguishes our work from traditional climatological quality control and aligns with the specific needs of extreme weather forecasting applications.
Despite achieving its core objectives, the work reveals important limitations that guide future development. Feature sparsity necessitated removal of all extreme wind variables and most solar radiation measurements, limiting coverage of some meteorological phenomena. We hope that the sparsity will be reduced in the future as higher quality data is introduced to the DWD data every day. The pipeline is then able to address this

automatically and adding features again to the third and fourth version of the dataset once the threshold is reached.
The primary next step for this work is to apply the dataset to RAINA and WeatherGenerator for comprehensive evaluation of extreme weather prediction capabilities in Germany. This will involve training and fine-tuning these models using our high-resolution observational data, with particular focus on extreme precipitation and flood events where our dataset's strengths are most valuable. Systematic evaluation will assess model performance on historical extreme events, comparing predictions against actual observations to validate the dataset's utility for catastrophic event forecasting. This application-focused evaluation will provide concrete evidence of the dataset's practical value and identify any remaining data quality issues that emerge during operational model training.
The modular pipeline architecture supports continuous updates and is publicly available (*Hauer*, 2025), enabling reproducible high-resolution atmospheric data processing for Germany. By addressing the critical gap between raw observational data and ML-ready formats, this work contributes to the broader goal of improving extreme weather prediction through data-driven methods.

8 Tools and Technologies

The entire data processing pipeline was implemented in Python 3.10.11, leveraging a carefully selected stack of scientific computing libraries optimised for handling large-scale atmospheric datasets. For data manipulation and analysis, we utilised NumPy (*Harris* et al., 2020) for efficient array operations, Pandas (*McKinney* et al., 2010) for tabular data handling and xarray (*Hoyer* and *Hamman*, 2017) for labeled multi-dimensional arrays with built-in support for NetCDF operations. The NetCDF4 library (*Whitaker* et al., 2024) provided the interface for reading and writing NetCDF files, which served as our primary data storage format due to its self-describing nature, compression capabilities and widespread adoption in the atmospheric sciences community. Statistical analysis and quality control procedures were implemented using SciPy (*Virtanen* et al., 2020) for scientific computing functions including Levene's test for variance homogeneity and Kolmogorov-Smirnov tests for distribution analysis. Visualisation of data quality metrics and pipeline outputs was accomplished using Matplotlib (*Hunter*, 2007), generating publication-quality figures for completeness analysis, range violations and temporal consistency checks.
Data acquisition from the DWD servers was automated using the requests library (*Reitz* et al., 2024) for HTTP operations and BeautifulSoup4 (*Richardson*, 2024) for HTML parsing, enabling systematic scraping of meteorological observations and station metadata. The pipeline architecture was designed as a modular system with seven distinct processing stages, orchestrated through a central `pipeline.py` script that manages dependencies between stages and provides comprehensive logging and error handling.
To handle the computational demands of processing over 260 TB of raw atmospheric data, the pre-processing stage was parallelised across nearly 100 CPU cores on the Forschungszentrum Jülich computing infrastructure. This parallelisation strategy was particularly effective given the independent nature of station-level processing operations, achieving near-linear speedup proportional to the number of allocated cores. All

processed datasets, intermediate results and quality control reports were stored on high-performance storage systems provided by Forschungszentrum Jülich, ensuring efficient access for subsequent analysis and model training workflows.

9 Acknowledgments

We would like to thank the DWD for providing the data and Forschungszentrum Jülich for providing the compute and storage. We also thank Sabine Schröder (JSC), Ilaria Luise (ECMWF), Carsten Hinz (JSC), Christian Lessig (ECMWF), Sabrina Wahl (DWD), Michael Langguth (former JSC, now QUADRA Energy), Wael Almikaeel (JSC) and Timothee Hunter (ECMWF) for their support and help.
For coding and writing support we have used Claude Sonnet 4.5 (Anthropic), Gemini 2.5 and 3 (Google) and GPT-4o (OpenAI). LLMs played an especially crucial role in working with matplotlib to find the best ways to visualise the data.

References

DWD (2024a). *10-minute station observations for Germany, Version v24.03*. Version v24.03. [Dataset]. URL: https://opendata.dwd.de/climate_environment/CDC/observations_germany/climate/.

– (2024b). *10-minute station observations of air temperature for Germany, Version v24.03*. Version v24.03. [Dataset]. URL: https://opendata.dwd.de/climate_environment/CDC/observations_germany/climate/10_minutes/air_temperature/historical/.

– (2024c). *10-minute station observations of extreme temperatures for Germany, Version v24.03*. Version v24.03. [Dataset]. URL: https://opendata.dwd.de/climate_environment/CDC/observations_germany/climate/10_minutes/extreme_temperature/historical/.

– (2024d). *10-minute station observations of extreme wind for Germany, Version v24.03*. Version v24.03. [Dataset]. URL: https://opendata.dwd.de/climate_environment/CDC/observations_germany/climate/10_minutes/extreme_wind/historical/.

– (2024e). *10-minute station observations of precipitation for Germany, Version v24.03*. Version v24.03. [Dataset]. URL: https://opendata.dwd.de/climate_environment/CDC/observations_germany/climate/10_minutes/precipitation/historical/.

– (2024f). *10-minute station observations of solar and sunshine for Germany, Version v24.03*. Version v24.03. [Dataset]. URL: https://opendata.dwd.de/climate_environment/CDC/observations_germany/climate/10_minutes/solar/historical/.

– (2024g). *10-minute station observations of wind for Germany, Version v24.03*. Version v24.03. [Dataset]. URL: https://opendata.dwd.de/climate_environment/CDC/observations_germany/climate/10_minutes/wind/historical/.

Fisher, Ronald A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Forschungszentrum Jülich GmbH (2025). *RAINA*. Project Website. URL: https://raina-project.de/.

Harris, Charles R. et al. (Sept. 2020). “Array programming with NumPy”. In: *Nature* 585.7825, pp. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2.

Hauer, Till (2025). *SYNOP Dataset Pipeline*. GitHub repository. URL: https://github.com/TillHae/synop_dataset_pipeline.

Hoyer, Stephan and Joseph *Hamman* (2017). “xarray: N-D labeled arrays and datasets in Python”. In: *Journal of Open Research Software* 5.1, p. 10. DOI: 10.5334/jors.148.

Hunter, John D. (2007). “Matplotlib: A 2D graphics environment”. In: *Computing in Science & Engineering* 9.3, pp. 90–95. DOI: 10.1109/MCSE.2007.55.

Kaspar, F. et al. (2013). “Monitoring of climate change in Germany – data, products and services of Germany’s National Climate Data Centre”. In: *Advances in Science and Research* 10.1, pp. 99–106. DOI: 10.5194/asr-10-99-2013. URL: https://asr.copernicus.org/articles/10/99/2013/.

Lessig, Christian et al. (2025). *The WeatherGenerator: a foundation model for weather and climate*. ECMWF Project. URL: https://weathergenerator.eu/.

Levene, Howard (1960). “Robust tests for equality of variances”. In: *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Ed. by Ingram *Olkin* et al. Stanford University Press, pp. 278–292.

McKinney, Wes et al. (2010). “Data structures for statistical computing in python”. In: *Proceedings of the 9th Python in Science Conference*. Vol. 445. Austin, TX, pp. 51–56.

Mirowski, Piotr et al. (2024). “Neural Compression of Atmospheric States”. In: *arXiv preprint arXiv:2407.11666*.

Nguyen, Tung et al. (2023). “ClimateLearn: Benchmarking Machine Learning for Weather and Climate Modeling”. In: *arXiv preprint arXiv:2307.01909*.

Reitz, Kenneth et al. (2024). *Requests: HTTP for Humans 2.32.3*. Computer software. URL: https://requests.readthedocs.io.

Richardson, Leonard (2024). *Beautiful Soup 4.12.3*. Computer software. URL: https://www.crummy.com/software/BeautifulSoup/.

Seabold, Skipper and Josef *Perktold* (2010). “statsmodels: Econometric and statistical modeling with python”. In: *9th Python in Science Conference*.

Sha, Yingkai et al. (Apr. 2021). “Deep-learning-based precipitation observation quality control”. In: *Journal of Atmospheric and Oceanic Technology*. DOI: 10.1175/JTECH-D-20-0081.1.

Spengler, R. (2002). “The new Quality Control- and Monitoring System of the Deutscher Wetterdienst”. In: *Proceedings of the WMO Technical Conference on Meteorological and Environmental Instruments and Methods of Observation*. Bratislava.

Virtanen, Pauli et al. (2020). “SciPy 1.0: fundamental algorithms for scientific computing in Python”. In: *Nature Methods* 17.3, pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

Wang, Xiao et al. (2025). “Scaling Pre-training to One Hundred Billion Data for Vision Language Models”. In: *arXiv preprint arXiv:2502.00000*.

Whitaker, Jeffrey et al. (2024). *netcdf4-python: Python/numpy interface to the netCDF C library*. Version 1.6.5. URL: https://unidata.github.io/netcdf4-python/.

Zhang, Zhongwei et al. (2025). *Numerical models outperform AI weather forecasts of record-breaking extremes*. arXiv: 2508.15724 [physics.ao-ph]. URL: https://arxiv.org/abs/2508.15724.