Response to Reviewer’s comments:

This manuscript presents MeteoSaver v1.0, an open-source, machine-learning based pipeline for the transcription, quality control, and structuring of historical meteorological records. The work is technically strong, well motivated, and relevant for climate data rescue, particularly in data-scarce regions.

The system is a valuable contribution to the field of historical climate data rescue, and the open-source, modular design is commendable.. The paper represents a valuable contribution at the interface of climate science, machine learning, and data engineering.

To strengthen the manuscript, I recommend expanding or more clearly contextualizing the validation, clarifying accuracy requirements for climate applications, and addressing potential biases introduced by rule-based QC. Addressing these points will significantly enhance the practical usefulness of the software.

The manuscript reports several performance metrics (e.g., transcription match rates, MAE, quality flags). The reported median match rate of approximately 74% between MeteoSaver outputs and manual transcription is relatively low for climate data rescue applications, where accuracy requirements are often stringent. While the authors also report a median MAE of 0.3 °C for temperature, the relationship between these metrics and their implications for downstream climate analyses is not sufficiently discussed. The paper should sufficiently explain:

What level of transcription error is acceptable for climate or meteorological analysis,

How the reported errors might affect climatologies, trend analyses, or extreme-event detection,

Whether the current performance is adequate for the intended use cases.

The authors should clearly link their validation results to the types of climate analyses for which MeteoSaver outputs are (or are not) suitable, and explicitly discuss limitations.

I recommend “minor revision”.

We thank the reviewer for their time and valuable suggestions to improve the manuscript. Below, we respond to the individual comments and illustrate modifications to the manuscript to accommodate the concerns raised. We believe that the manuscript has benefited from these modifications. The following convention is used in this document to illustrate the text modifications in the original manuscript: modified text.

We have expanded our discussion section to clarify on the reported accuracy metrics for climate or meteorological analyses as well as the potential use cases for the transcribed data. Below is the modification to the discussion section on the Quality Assessment and Quality Control:

Section 5.3 Quality Assessment and Quality Control

The QA/QC results in this release, as outlined in sect. 3.5, demonstrate that our current pipeline is robust in identifying transcription errors by employing (i) user-defined data thresholds, (ii) multi-day totals and averages, (iii) logic checks, such as Tmin < Tavg < Tmax, (iv) iterative comparisons between related variables, including the relationship between Tmin, Tavg, Tmax, and the Diurnal Temperature Range (DTR), and (v) reiteration across the different checks (Fig. 9). Notably, the presence of pentad totals and averages in our sheets proved particularly valuable for QA/QC, as they simplified the process of recalculating and confirming transcribed values within each pentad. In contrast, the more common monthly totals and/or averages found in many international archived weather data sheets would present greater challenges, particularly when multiple daily values are transcribed incorrectly.

An evaluation of the transcribed values that achieved the highest quality flag revealed a median confirmation rate of 74.4% of the transcribed temperature data across the sheets, either confirmed or corrected during QA/QC. This means that 25.6% of the transcribed values were excluded from the final output timeseries. Among these excluded values, a substantial fraction is correctly transcribed but could not be validated by the QA/QC framework. This may occur, for example, when incorrectly transcribed values appear within the same pentad, preventing confirmation of the correctly transcribed values through the QA/QC checks, or in rare cases when extreme but valid observations fall outside the predefined threshold criteria. While the QA/QC checks are used to validate and refine the transcribed data, as demonstrated in Fig. 11, their effectiveness heavily relies on the initial transcription quality, which depends on the OCR/HTR model, the variability in handwriting styles, and the maintenance condition of the paper sheets in the archives. For instance, when the paper condition is well-preserved (as in Fig. 1) and the initial transcription is nearly accurate across most cells, only the first few QA/QC checks are typically sufficient to confirm all daily temperature values in a pentad, as illustrated in Fig. 11 i - j. On the other hand, in cases where the initial transcription contains multiple errors, all QA/QC checks and iterative re-evaluations in our framework are necessary to confirm the temperature values in that pentad, as seen in Fig. 11 a - h. The latter could, in rare cases, lead to incorrectly confirmed values if the originally transcribed data contains errors that still meet multiple QA/QC checks, or may result in values falling outside the user-defined uncertainty margin, which would therefore remain unconfirmed. Users should be aware that the current QA/QC framework may exclude some valid extreme observations and therefore additional manual verification is advised for applications focusing on rare extremes.

In our study, the final confirmed temperature values showed a match rate of 52.4–100% with manually transcribed records, yielding a median accuracy of 74% and a median mean absolute error (MAE) of 0.3°C (see Table 2). While the accuracy indicates the proportion of automatically transcribed values that match the manually transcribed values with a predefined uncertainty margin, the MAE provides an indication of the magnitude of transcription deviations. The median MAE of 0.3°C observed in these sample sheets is comparable to typical uncertainties associated with historical thermometer measurements of 0.2°C (1σ) (Morice et al., 2012; Brohan et al., 2006; Folland et al., 2001). Because many climatological analyses rely on aggregated statistics derived from combined station data, such as spatial averages and long-term trends, transcription deviations of this magnitude are unlikely to substantially affect the resulting climatological interpretations (Brohan et al., 2006). However, for analyses requiring precise daily values such as extreme-event detection, additional manual verification may still be advisable.

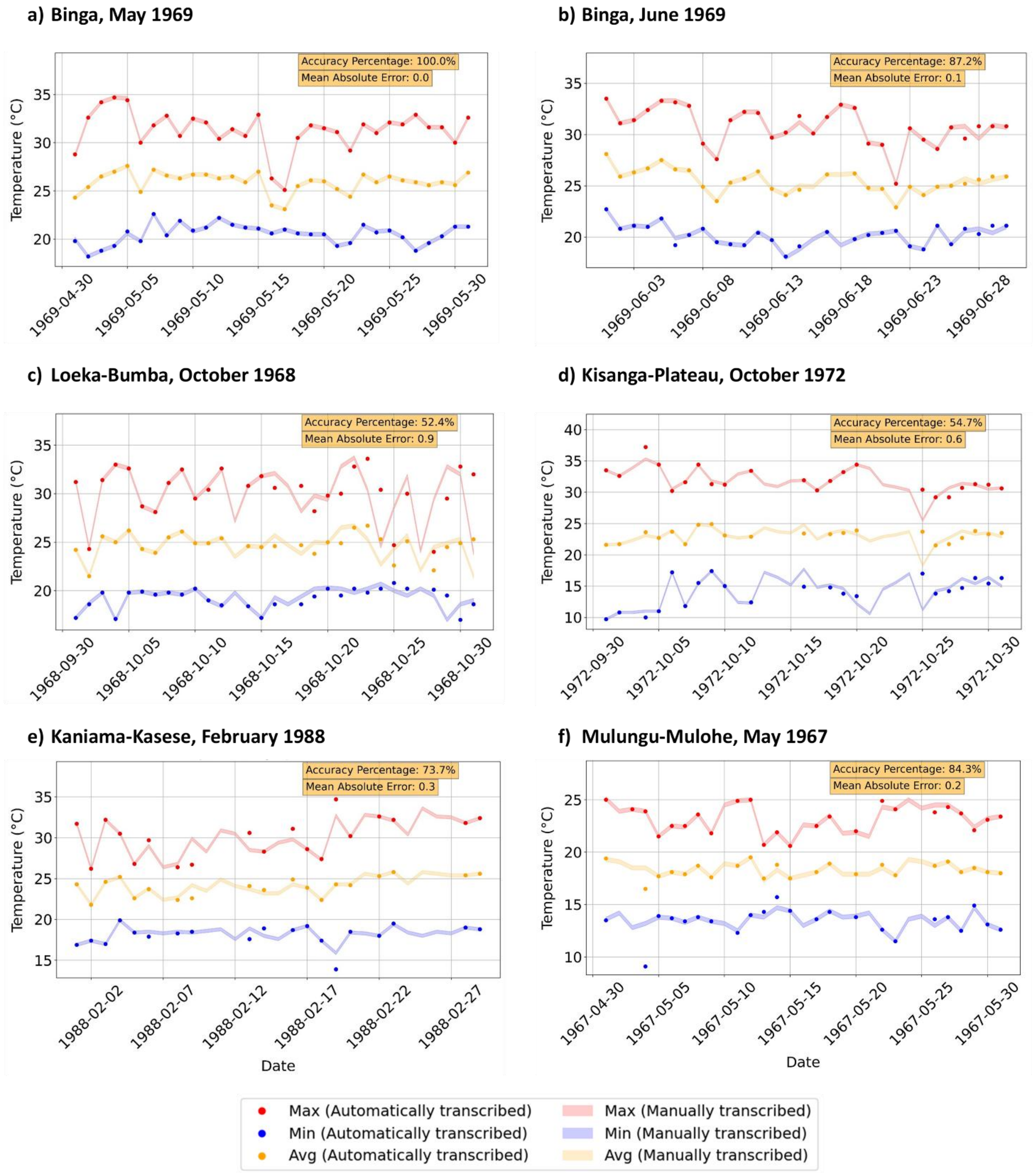
These reported performance metrics are specific to this study's sample weather sheet formats, input image quality, handwriting styles on these sheets, paper maintenance conditions, and manual transcription quality. Consequently, the reported performance may differ when MeteoSaver is applied to other historical datasets with different table structures, handwriting styles or image quality. Additionally, the variability of climate variables should be considered, for example, while temperature values in the DRC exhibit relatively small annual ranges, extratropical regions often experience much larger seasonal variations (on the order of tens of degrees). In such contexts, transcription errors may be larger and more difficult to detect through the applied QA/QC procedures. Nevertheless, the modular design of MeteoSaver allows users to adapt the configuration and retrain the OCR/HTR models to accommodate different table layouts and handwriting styles.

Minor comments:

1. The resolution of figures 13 & 14 should be improved. Some validation figures would benefit from clearer guidance on how they should be interpreted by climate scientists.

Response:

We thank the reviewer for this suggestion. The resolution and font size of the labels in Figures 13 and 14 have been increased in the revised manuscript to improve readability. In addition, we have revised the figure captions (see below) and added explanatory text in the manuscript, as described above, to provide clearer guidance on how these validation figures should be interpreted by climate scientists, particularly with respect to the transcription accuracy metrics and their implications for the reliability of the transcribed climate observations.



a) Binga, May 1969

b) Binga, June 1969

c) Loeka-Bumba, October 1968

d) Kisanga-Plateau, October 1972

e) Kaniama-Kasese, February 1988

f) Mulungu-Mulohe, May 1967

Figure 13. Time series plot of the daily maximum (red), average (orange), and minimum (blue) temperatures for the respective stations. Each variable shows automatically transcribed values as solid markers, while manually transcribed values are displayed as lighter time series bands with a 0.2°C uncertainty margin applied during QA/QC checks. The accuracy percentage and mean absolute error (MAE) between the automatically and manually transcribed values are noted in the upper right corner of the plot. The accuracy percentage denotes the percentage of confirmed, automatically transcribed values that fall within 0.2°C of the manually transcribed value. Together, these metrics quantify the agreement between the automatically transcribed values using MeteoSaver v1.0 (markers) and their corresponding manually transcribed values (bands), providing an indication of the reliability of the automatically transcribed observations (with respect to manual transcriptions) for subsequent climatological analyses. The analysis assumes that the manually

transcribed values are correct; however, this may not always be the case, as manual transcription is also subject to errors depending on the methods applied.

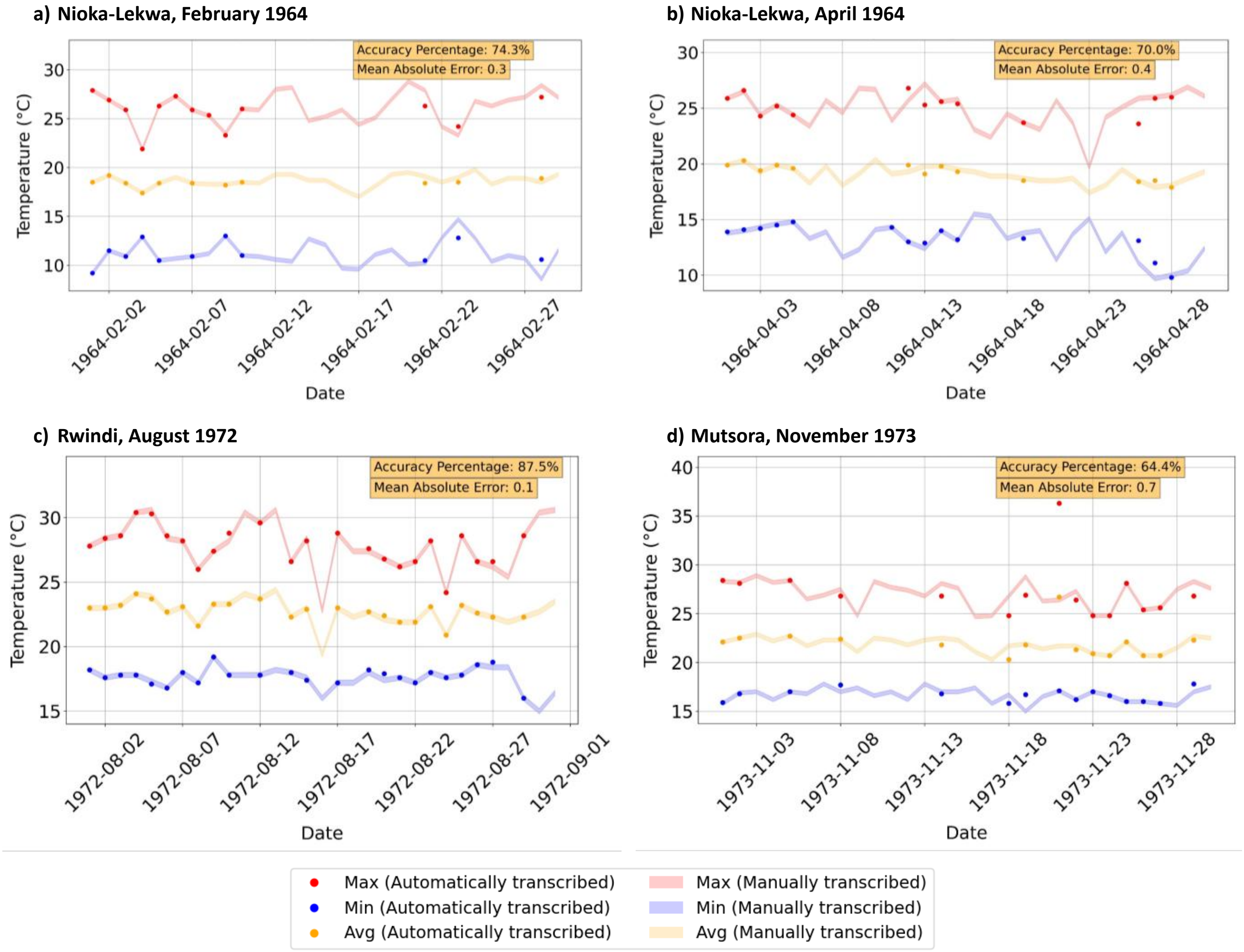


Figure 14. Time series plot of the daily maximum (red), average (orange), and minimum (blue) temperatures for the respective stations. Each variable shows automatically transcribed values as solid markers, while manually transcribed values are displayed as lighter time series bands with a 0.2°C uncertainty margin applied during QA/QC checks. The accuracy percentage and mean absolute error (MAE) between the automatically and manually transcribed values are noted in the upper right corner of the plot. The accuracy percentage denotes the percentage of confirmed, automatically transcribed values that fall within 0.2°C of the manually transcribed value. Together, these metrics quantify the agreement between the automatically transcribed values using MeteoSaver v1.0 (markers) and their corresponding manually transcribed values (bands), providing an indication of the reliability of the automatically transcribed observations (with respect to manual transcriptions) for subsequent climatological analyses. The analysis assumes that the manually transcribed values are correct; however, this may not always be the case, as manual transcription is also subject to errors depending on the methods applied.

2. The discussion would benefit from a short paragraph situating MeteoSaver within the broader ecosystem of data rescue tools.

Response:

We thank the reviewer for this suggestion. We have added the paragraph below in the discussion *Section 5.4 : Potential of the software*.

> Our study demonstrates the flexibility of MeteoSaver in transcribing historical tabular weather data across a range of handwriting styles, table dimensions, paper sizes, and maintenance conditions, highlighting its potential contribution to ongoing climate data rescue efforts. Throughout our model development, we focused on reusability for similar case studies, equipping MeteoSaver with numerous configurable settings within the configuration module (see sect. 3.1) to allow users to tailor the software for specific table formats. Given this flexibility, MeteoSaver has substantial potential for transcribing millions of rescued archived weather records, such as those available on the C3S Data Rescue Service Portal. It is important to note, however, that in this initial release, users may need to make further adjustments when applying the software to data sheets with complex tabular formats or variables or data types beyond temperature and precipitation, particularly within the QA/QC checks.
>
> MeteoSaver therefore complements existing efforts in historical climate data rescue. Many recent efforts rely on manual transcription workflows, including citizen science initiatives (e.g., Noone et al.,2024, Hawkins et al., 2019, 2022; Craig and Hawkins, 2020), while other approaches explore open-source and commercial OCR/HTR models to directly transcribe individual values in scanned historical documents (Vercruysse et al., 2024; Nockels et al., 2022). MeteoSaver, on the other hand, contributes to these existing data rescue efforts by providing an open-source end-to-end workflow that integrates machine-learning into image processing, table and cell detection, and transcription, as well as QA/QC and data formatting ready for upload. In addition, MeteoSaver can potentially make use of quality-controlled manually transcribed data as training data with multiple handwriting styles for the OCR models used in the transcription module. By automating these key steps of the data rescue process, following the digitization (imaging) of paper-based records, MeteoSaver aims to substantially reduce the manual effort required for climate data rescue while integrating QA/QC procedures into OCR/HTR-based transcription workflows. Furthermore, its modular and open-source framework allows for continuous improvement of the machine-learning components as additional training data becomes available.
>
> While we showcase its application on tabular weather data, we also envision MeteoSaver's potential in transcribing other historical environmental records in tabular and numerical form, spanning fields like hydrology, biology, ecology, and oceanography. For instance, de Smeth et al. (2024) recently highlighted data rescue efforts for historical river flow records from Irish catchments, recorded from the early 1940s and recently transcribed manually; a process where MeteoSaver could potentially have saved numerous hours of manual work.

3. The evaluation focuses primarily on temperature variables. Please clarify whether MeteoSaver currently supports other common variables (precipitation, humidity, pressure, radiation) and whether different QC logic would be required.

Response:

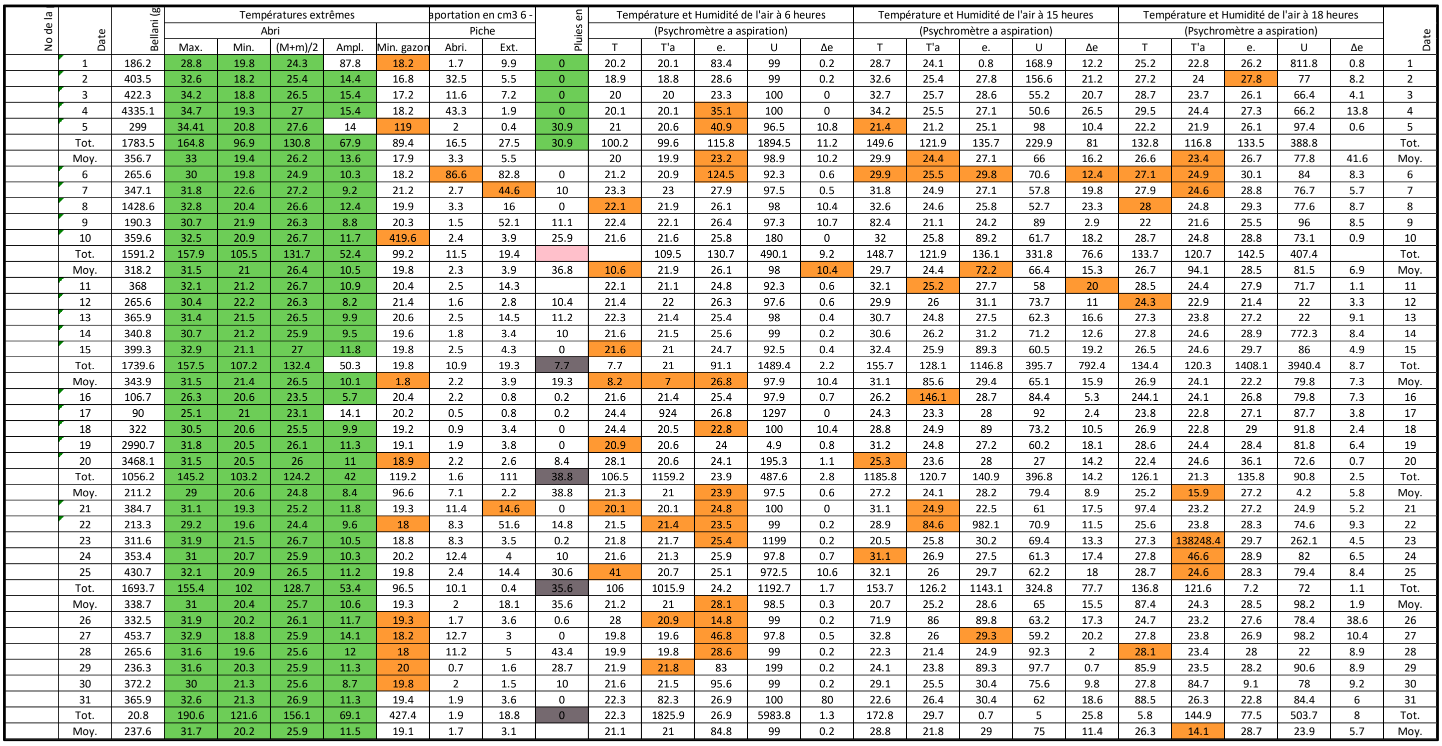
We thank the reviewer for this clarification request. MeteoSaver is designed to transcribe multiple meteorological variables and diagnostics recorded in tabular structures beyond the temperature, including but not limited to total incoming solar radiation, evaporation, precipitation, vapor pressure, and humidity as shown in our sheets (Fig 1).

In the first manuscript, however, the validation focuses on temperature variables (daily maximum, minimum and average), We chose this focus because these variables allow the illustration of a larger set of QA/QC checks such as the pentad (multi-day) totals and averages, maximum and minimum thresholds, and logic checks (e.g., Tmin < Tavg < Tmax, and their relation to the diurnal temperature range).

Other variables have fewer QA/QC checks that can be applied on our sheets. For example, daily precipitation records include the pentad total as the only direct QA/QC. This is a constraint for identifying and correcting erroneously transcribed daily precipitation values (the more checks passed, the higher the confidence in the automatically transcribed value). As a result, fewer transcribed daily precipitation values can be confirmed through the QA/QC framework compared with daily temperature values (Fig. 12 and Appendix Fig. B1 below).

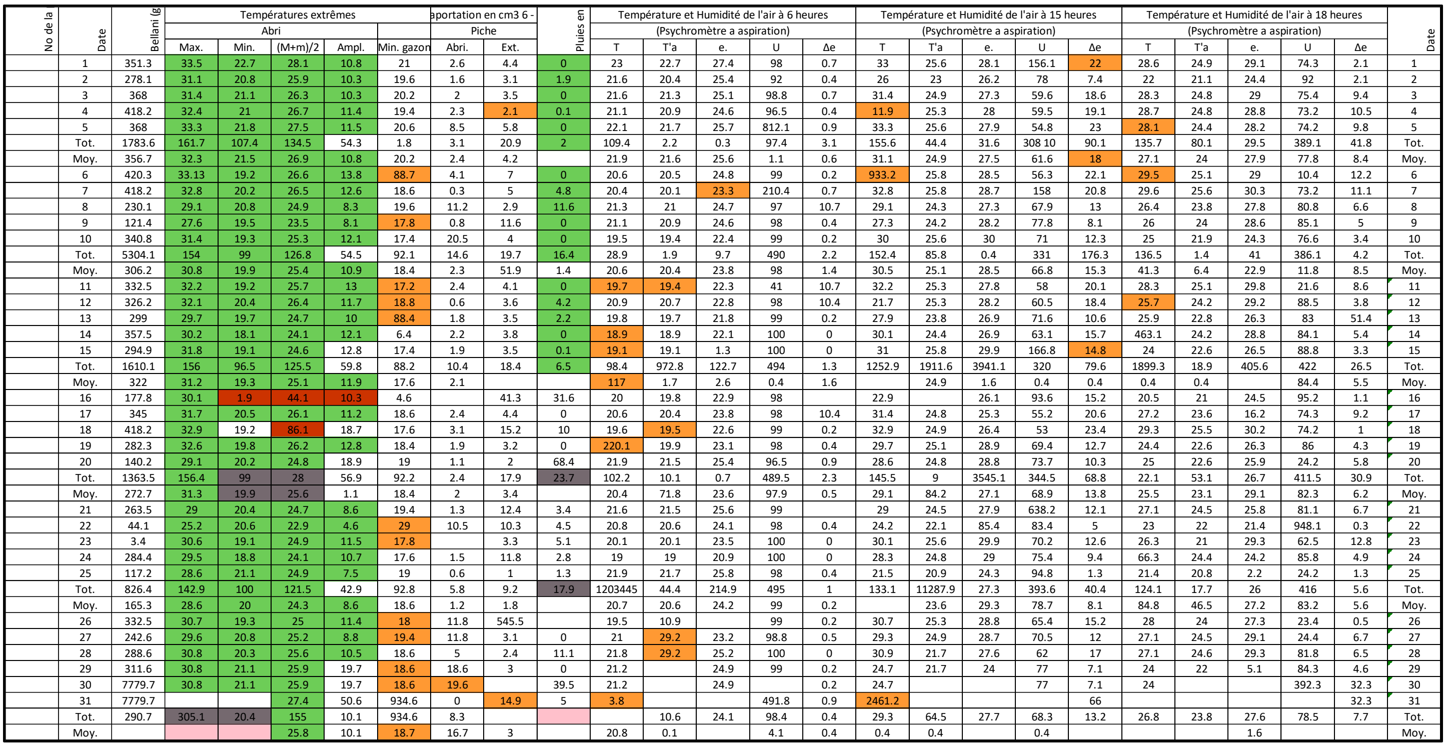
To illustrate MeteoSaver's capability to handle additional variables, we have now incorporated a precipitation pentad consistency check into the QA/QC module in the revised manuscript (Sect 3.5). We also updated the figures 12 and Appendix Figs. B1-B9 to incorporate the daily precipitation data that we transcribed and confirmed in these selected sheets. We note that in follow-up research, we aim to transcribe both temperature, precipitation and humidity from thousands of sheets digitized for DR Congo, and that we will therefore extend the analysis beyond temperature in future studies building on MeteoSaver.

> Section 3.5 Module 5: Quality Assessment and Quality Control
>
> In the second check, daily temperature values are tested to ensure they fall within the set maximum and minimum thresholds and are flagged if they do not (see Fig. 11 c-d, with flagged values in d shown in dark red). If a daily temperature value exceeds the maximum threshold and is also greater than 100, a specific adjustment is applied by dividing the value by 10. For values within the thresholds, the multi-day (here, pentad) totals and averages for Tmax, Tmin, and Tavg are calculated and compared with the transcribed multi-day totals and averages. Similarly, the multi-day totals for daily precipitation values (P) are calculated and compared with the transcribed multi-day totals. If the transcribed values match (or are within the set uncertainty margin of) the calculated totals or averages, both the multi-day values and their respective daily values are flagged as confirmed or not confirmed accordingly (see Fig. 11 c-d, with unconfirmed pentad total and average values in d shown in grey).



No de la	Date	Bellani (g	Températures extrêmes					aportation en cm3 6 -		Pluies en	Température et Humidité de l'air à 6 heures					Température et Humidité de l'air à 15 heures					Température et Humidité de l'air à 18 heures					Date
			Abri					Piche			(Psychromètre a aspiration)					(Psychromètre a aspiration)					(Psychromètre a aspiration)					
			Max.	Min.	(M+m)/2	Ampl.	Min. gazon	Abri.	Ext.		T	T'a	e.	U	Δe	T	T'a	e.	U	Δe	T	T'a	e.	U	Δe	
	1	186.2	28.8	19.8	24.3	87.8	18.2	1.7	9.9	0	20.2	20.1	83.4	99	0.2	28.7	24.1	0.8	168.9	12.2	25.2	22.8	26.2	811.8	0.8	1
	2	403.5	32.6	18.2	25.4	14.4	16.8	32.5	5.5	0	18.9	18.8	28.6	99	0.2	32.6	25.4	27.8	156.6	21.2	27.2	24	27.8	77	8.2	2
	3	422.3	34.2	18.8	26.5	15.4	17.2	11.6	7.2	0	20	20	23.3	100	0	32.7	25.7	28.6	55.2	20.7	28.7	23.7	26.1	66.4	4.1	3
	4	4335.1	34.7	19.3	27	15.4	18.2	43.3	1.9	0	20.1	20.1	35.1	100	0	34.2	25.5	27.1	50.6	26.5	29.5	24.4	27.3	66.2	13.8	4
	5	299	34.41	20.8	27.6	14	119	2	0.4	30.9	21	20.6	40.9	96.5	10.8	21.4	21.2	25.1	98	10.4	22.2	21.9	26.1	97.4	0.6	5
	Tot.	1783.5	164.8	96.9	130.8	67.9	89.4	16.5	27.5	30.9	100.2	99.6	115.8	1894.5	11.2	149.6	121.9	135.7	229.9	81	132.8	116.8	133.5	388.8		Tot.
	Moy.	356.7	33	19.4	26.2	13.6	17.9	3.3	5.5		20	19.9	23.2	98.9	10.2	29.9	24.4	27.1	66	16.2	26.6	23.4	26.7	77.8	41.6	Moy.
	6	265.6	30	19.8	24.9	10.3	18.2	86.6	82.8	0	21.2	20.9	124.5	92.3	0.6	29.9	25.5	29.8	70.6	12.4	27.1	24.9	30.1	84	8.3	6
	7	347.1	31.8	22.6	27.2	9.2	21.2	2.7	44.6	10	23.3	23	27.9	97.5	0.5	31.8	24.9	27.1	57.8	19.8	27.9	24.6	28.8	76.7	5.7	7
	8	1428.6	32.8	20.4	26.6	12.4	19.9	3.3	16	0	22.1	21.9	26.1	98	10.4	32.6	24.6	25.8	52.7	23.3	28	24.8	29.3	77.6	8.7	8
	9	190.3	30.7	21.9	26.3	8.8	20.3	1.5	52.1	11.1	22.4	22.1	26.4	97.3	10.7	82.4	21.1	24.2	89	2.9	22	21.6	25.5	96	8.5	9
	10	359.6	32.5	20.9	26.7	11.7	419.6	2.4	3.9	25.9	21.6	21.6	25.8	180	0	32	25.8	89.2	61.7	18.2	28.7	24.8	28.8	73.1	0.9	10
	Tot.	1591.2	157.9	105.5	131.7	52.4	99.2	11.5	19.4			109.5	130.7	490.1	9.2	148.7	121.9	136.1	331.8	76.6	133.7	120.7	142.5	407.4		Tot.
	Moy.	318.2	31.5	21	26.4	10.5	19.8	2.3	3.9	36.8	10.6	21.9	26.1	98	10.4	29.7	24.4	72.2	66.4	15.3	26.7	94.1	28.5	81.5	6.9	Moy.
	11	368	32.1	21.2	26.7	10.9	20.4	2.5	14.3		22.1	21.1	24.8	92.3	0.6	32.1	25.2	27.7	58	20	28.5	24.4	27.9	71.7	1.1	11
	12	265.6	30.4	22.2	26.3	8.2	21.4	1.6	2.8	10.4	21.4	22	26.3	97.6	0.6	29.9	26	31.1	73.7	11	24.3	22.9	21.4	22	3.3	12
	13	365.9	31.4	21.5	26.5	9.9	20.6	2.5	14.5	11.2	22.3	21.4	25.4	98	0.4	30.7	24.8	27.5	62.3	16.6	27.3	23.8	27.2	22	9.1	13
	14	340.8	30.7	21.2	25.9	9.5	19.6	1.8	3.4	10	21.6	21.5	25.6	99	0.2	30.6	26.2	31.2	71.2	12.6	27.8	24.6	28.9	772.3	8.4	14
	15	399.3	32.9	21.1	27	11.8	19.8	2.5	4.3	0	21.6	21	24.7	92.5	0.4	32.4	25.9	89.3	60.5	19.2	26.5	24.6	29.7	86	4.9	15
	Tot.	1739.6	157.5	107.2	132.4	50.3	19.8	10.9	19.3	7.7	7.7	21	91.1	1489.4	2.2	155.7	128.1	1146.8	395.7	792.4	134.4	120.3	1408.1	3940.4	8.7	Tot.
	Moy.	343.9	31.5	21.4	26.5	10.1	1.8	2.2	3.9	19.3	8.2	7	26.8	97.9	10.4	31.1	85.6	29.4	65.1	15.9	26.9	24.1	22.2	79.8	7.3	Moy.
	16	106.7	26.3	20.6	23.5	5.7	20.4	2.2	0.8	0.2	21.6	21.4	25.4	97.9	0.7	26.2	146.1	28.7	84.4	5.3	244.1	24.1	26.8	79.8	7.3	16
	17	90	25.1	21	23.1	14.1	20.2	0.5	0.8	0.2	24.4	924	26.8	1297	0	24.3	23.3	28	92	2.4	23.8	22.8	27.1	87.7	3.8	17
	18	322	30.5	20.6	25.5	9.9	19.2	0.9	3.4	0	24.4	20.5	22.8	100	10.4	28.8	24.9	89	73.2	10.5	26.9	22.8	29	91.8	2.4	18
	19	2990.7	31.8	20.5	26.1	11.3	19.1	1.9	3.8	0	20.9	20.6	24	4.9	0.8	31.2	24.8	27.2	60.2	18.1	28.6	24.4	28.4	81.8	6.4	19
	20	3468.1	31.5	20.5	26	11	18.9	2.2	2.6	8.4	28.1	20.6	24.1	195.3	1.1	25.3	23.6	28	27	14.2	22.4	24.6	36.1	72.6	0.7	20
	Tot.	1056.2	145.2	103.2	124.2	42	119.2	1.6	111	38.8	106.5	1159.2	23.9	487.6	2.8	1185.8	120.7	140.9	396.8	14.2	126.1	21.3	135.8	90.8	2.5	Tot.
	Moy.	211.2	29	20.6	24.8	8.4	96.6	7.1	2.2	38.8	21.3	21	23.9	97.5	0.6	27.2	24.1	28.2	79.4	8.9	25.2	15.9	27.2	4.2	5.8	Moy.
	21	384.7	31.1	19.3	25.2	11.8	19.3	11.4	14.6	0	20.1	20.1	24.8	100	0	31.1	24.9	22.5	61	17.5	97.4	23.2	27.2	24.9	5.2	21
	22	213.3	29.2	19.6	24.4	9.6	18	8.3	51.6	14.8	21.5	21.4	23.5	99	0.2	28.9	84.6	982.1	70.9	11.5	25.6	23.8	28.3	74.6	9.3	22
	23	311.6	31.9	21.5	26.7	10.5	18.8	8.3	3.5	0.2	21.8	21.7	25.4	1199	0.2	20.5	25.8	30.2	69.4	13.3	27.3	138248.4	29.7	262.1	4.5	23
	24	353.4	31	20.7	25.9	10.3	20.2	12.4	4	10	21.6	21.3	25.9	97.8	0.7	31.1	26.9	27.5	61.3	17.4	27.8	46.6	28.9	82	6.5	24
	25	430.7	32.1	20.9	26.5	11.2	19.8	2.4	14.4	30.6	41	20.7	25.1	972.5	10.6	32.1	26	29.7	62.2	18	28.7	24.6	28.3	79.4	8.4	25
	Tot.	1693.7	155.4	102	128.7	53.4	96.5	10.1	0.4	35.6	106	1015.9	24.2	1192.7	1.7	153.7	126.2	1143.1	324.8	77.7	136.8	121.6	7.2	72	1.1	Tot.
	Moy.	338.7	31	20.4	25.7	10.6	19.3	2	18.1	35.6	21.2	21	28.1	98.5	0.3	20.7	25.2	28.6	65	15.5	87.4	24.3	28.5	98.2	1.9	Moy.
	26	332.5	31.9	20.2	26.1	11.7	19.3	1.7	3.6	0.6	28	20.9	14.8	99	0.2	71.9	86	89.8	63.2	17.3	24.7	23.2	27.6	78.4	38.6	26
	27	453.7	32.9	18.8	25.9	14.1	18.2	12.7	3	0	19.8	19.6	46.8	97.8	0.5	32.8	26	29.3	59.2	20.2	27.8	23.8	26.9	98.2	10.4	27
	28	265.6	31.6	19.6	25.6	12	18	11.2	5	43.4	19.9	19.8	28.6	99	0.2	22.3	21.4	24.9	92.3	2	28.1	23.4	28	22	8.9	28
	29	236.3	31.6	20.3	25.9	11.3	20	0.7	1.6	28.7	21.9	21.8	83	199	0.2	24.1	23.8	89.3	97.7	0.7	85.9	23.5	28.2	90.6	8.9	29
	30	372.2	30	21.3	25.6	8.7	19.8	2	1.5	10	21.6	21.5	95.6	99	0.2	29.1	25.5	30.4	75.6	9.8	27.8	84.7	9.1	78	9.2	30
	31	365.9	32.6	21.3	26.9	11.3	19.4	1.9	3.6	0	22.3	82.3	26.9	100	80	22.6	26.4	30.4	62	18.6	88.5	26.3	22.8	84.4	6	31
	Tot.	20.8	190.6	121.6	156.1	69.1	427.4	1.9	18.8	0	22.3	1825.9	26.9	5983.8	1.3	172.8	29.7	0.7	5	25.8	5.8	144.9	77.5	503.7	8	Tot.
	Moy.	237.6	31.7	20.2	25.9	11.5	19.1	1.7	3.1		21.1	21	84.8	99	0.2	28.8	21.8	29	75	11.4	26.3	14.1	28.7	23.9	5.7	Moy.

Figure 12. Post-quality controlled table using MeteoSaver, showing confirmed values of daily maximum, minimum and average temperature, diurnal temperature range, and daily precipitation (highlighted in green) for the weather sheet shown in Fig. 1. The description of the colors in the post-quality controlled table is given in Fig. 10.



No de la	Date	Bellani (g	Températures extrêmes					aportation en cm3 6 -		Pluies en	Température et Humidité de l'air à 6 heures					Température et Humidité de l'air à 15 heures					Température et Humidité de l'air à 18 heures					Date
			Abri					Piche			(Psychromètre a aspiration)					(Psychromètre a aspiration)					(Psychromètre a aspiration)					
			Max.	Min.	(M+m)/2	Ampl.	Min. gazon	Abri.	Ext.		T	T'a	e.	U	Δe	T	T'a	e.	U	Δe	T	T'a	e.	U	Δe	
	1	351.3	33.5	22.7	28.1	10.8	21	2.6	4.4	0	23	22.7	27.4	98	0.7	33	25.6	28.1	156.1	22	28.6	24.9	29.1	74.3	2.1	1
	2	278.1	31.1	20.8	25.9	10.3	19.6	1.6	3.1	1.9	21.6	20.4	25.4	92	0.4	26	23	26.2	78	7.4	22	21.1	24.4	92	2.1	2
	3	368	31.4	21.1	26.3	10.3	20.2	2	3.5	0	21.6	21.3	25.1	98.8	0.7	31.4	24.9	27.3	59.6	18.6	28.3	24.8	29	75.4	9.4	3
	4	418.2	32.4	21	26.7	11.4	19.4	2.3	2.1	0.1	21.1	20.9	24.6	96.5	0.4	11.9	25.3	28	59.5	19.1	28.7	24.8	28.8	73.2	10.5	4
	5	368	33.3	21.8	27.5	11.5	20.6	8.5	5.8	0	22.1	21.7	25.7	812.1	0.9	33.3	25.6	27.9	54.8	23	28.1	24.4	28.2	74.2	9.8	5
	Tot.	1783.6	161.7	107.4	134.5	54.3	1.8	3.1	20.9	2	109.4	2.2	0.3	97.4	3.1	155.6	44.4	31.6	308 10	90.1	135.7	80.1	29.5	389.1	41.8	Tot.
	Moy.	356.7	32.3	21.5	26.9	10.8	20.2	2.4	4.2		21.9	21.6	25.6	1.1	0.6	31.1	24.9	27.5	61.6	18	27.1	24	27.9	77.8	8.4	Moy.
	6	420.3	33.13	19.2	26.6	13.8	88.7	4.1	7	0	20.6	20.5	24.8	99	0.2	933.2	25.8	28.5	56.3	22.1	29.5	25.1	29	10.4	12.2	6
	7	418.2	32.8	20.2	26.5	12.6	18.6	0.3	5	4.8	20.4	20.1	23.3	210.4	0.7	32.8	25.8	28.7	158	20.8	29.6	25.6	30.3	73.2	11.1	7
	8	230.1	29.1	20.8	24.9	8.3	19.6	11.2	2.9	11.6	21.3	21	24.7	97	10.7	29.1	24.3	27.3	67.9	13	26.4	23.8	27.8	80.8	6.6	8
	9	121.4	27.6	19.5	23.5	8.1	17.8	0.8	11.6	0	21.1	20.9	24.6	98	0.4	27.3	24.2	28.2	77.8	8.1	26	24	28.6	85.1	5	9
	10	340.8	31.4	19.3	25.3	12.1	17.4	20.5	4	0	19.5	19.4	22.4	99	0.2	30	25.6	30	71	12.3	25	21.9	24.3	76.6	3.4	10
	Tot.	5304.1	154	99	126.8	54.5	92.1	14.6	19.7	16.4	28.9	1.9	9.7	490	2.2	152.4	85.8	0.4	331	176.3	136.5	1.4	41	386.1	4.2	Tot.
	Moy.	306.2	30.8	19.9	25.4	10.9	18.4	2.3	51.9	1.4	20.6	20.4	23.8	98	1.4	30.5	25.1	28.5	66.8	15.3	41.3	6.4	22.9	11.8	8.5	Moy.
	11	332.5	32.2	19.2	25.7	13	17.2	2.4	4.1	0	19.7	19.4	22.3	41	10.7	32.2	25.3	27.8	58	20.1	28.3	25.1	29.8	21.6	8.6	11
	12	326.2	32.1	20.4	26.4	11.7	18.8	0.6	3.6	4.2	20.9	20.7	22.8	98	10.4	21.7	25.3	28.2	60.5	18.4	25.7	24.2	29.2	88.5	3.8	12
	13	299	29.7	19.7	24.7	10	88.4	1.8	3.5	2.2	19.8	19.7	21.8	99	0.2	27.9	23.8	26.9	71.6	10.6	25.9	22.8	26.3	83	51.4	13
	14	357.5	30.2	18.1	24.1	12.1	6.4	2.2	3.8	0	18.9	18.9	22.1	100	0	30.1	24.4	26.9	63.1	15.7	463.1	24.2	28.8	84.1	5.4	14
	15	294.9	31.8	19.1	24.6	12.8	17.4	1.9	3.5	0.1	19.1	19.1	1.3	100	0	31	25.8	29.9	166.8	14.8	24	22.6	26.5	88.8	3.3	15
	Tot.	1610.1	156	96.5	125.5	59.8	88.2	10.4	18.4	6.5	98.4	972.8	122.7	494	1.3	1252.9	1911.6	3941.1	320	79.6	1899.3	18.9	405.6	422	26.5	Tot.
	Moy.	322	31.2	19.3	25.1	11.9	17.6	2.1			117	1.7	2.6	0.4	1.6		24.9	1.6	0.4	0.4	0.4	0.4		84.4	5.5	Moy.
	16	177.8	30.1	1.9	44.1	10.3	4.6		41.3	31.6	20	19.8	22.9	98		22.9		26.1	93.6	15.2	20.5	21	24.5	95.2	1.1	16
	17	345	31.7	20.5	26.1	11.2	18.6	2.4	4.4	0	20.6	20.4	23.8	98	10.4	31.4	24.8	25.3	55.2	20.6	27.2	23.6	16.2	74.3	9.2	17
	18	418.2	32.9	19.2	86.1	18.7	17.6	3.1	15.2	10	19.6	19.5	22.6	99	0.2	32.9	24.9	26.4	53	23.4	29.3	25.5	30.2	74.2	1	18
	19	282.3	32.6	19.8	26.2	12.8	18.4	1.9	3.2	0	220.1	19.9	23.1	98	0.4	29.7	25.1	28.9	69.4	12.7	24.4	22.6	26.3	86	4.3	19
	20	140.2	29.1	20.2	24.8	18.9	19	1.1	2	68.4	21.9	21.5	25.4	96.5	0.9	28.6	24.8	28.8	73.7	10.3	25	22.6	25.9	24.2	5.8	20
	Tot.	1363.5	156.4	99	28	56.9	92.2	2.4	17.9	23.7	102.2	10.1	0.7	489.5	2.3	145.5	9	3545.1	344.5	68.8	22.1	53.1	26.7	411.5	30.9	Tot.
	Moy.	272.7	31.3	19.9	25.6	1.1	18.4	2	3.4		20.4	71.8	23.6	97.9	0.5	29.1	84.2	27.1	68.9	13.8	25.5	23.1	29.1	82.3	6.2	Moy.
	21	263.5	29	20.4	24.7	8.6	19.4	1.3	12.4	3.4	21.6	21.5	25.6	99		29	24.5	27.9	638.2	12.1	27.1	24.5	25.8	81.1	6.7	21
	22	44.1	25.2	20.6	22.9	4.6	29	10.5	10.3	4.5	20.8	20.6	24.1	98	0.4	24.2	22.1	85.4	83.4	5	23	22	21.4	948.1	0.3	22
	23	3.4	30.6	19.1	24.9	11.5	17.8		3.3	5.1	20.1	20.1	23.5	100	0	30.1	25.6	29.9	70.2	12.6	26.3	21	29.3	62.5	12.8	23
	24	284.4	29.5	18.8	24.1	10.7	17.6	1.5	11.8	2.8	19	19	20.9	100	0	28.3	24.8	29	75.4	9.4	66.3	24.4	24.2	85.8	4.9	24
	25	117.2	28.6	21.1	24.9	7.5	19	0.6	1	1.3	21.9	21.7	25.8	98	0.4	21.5	20.9	24.3	94.8	1.3	21.4	20.8	2.2	24.2	1.3	25
	Tot.	826.4	142.9	100	121.5	42.9	92.8	5.8	9.2	17.9	1203445	44.4	214.9	495	1	133.1	11287.9	27.3	393.6	40.4	124.1	17.7	26	416	5.6	Tot.
	Moy.	165.3	28.6	20	24.3	8.6	18.6	1.2	1.8		20.7	20.6	24.2	99	0.2		23.6	29.3	78.7	8.1	84.8	46.5	27.2	83.2	5.6	Moy.
	26	332.5	30.7	19.3	25	11.4	18	11.8	545.5		19.5	10.9		99	0.2	30.7	25.3	28.8	65.4	15.2	28	24	27.3	23.4	0.5	26
	27	242.6	29.6	20.8	25.2	8.8	19.4	11.8	3.1	0	21	29.2	23.2	98.8	0.5	29.3	24.9	28.7	70.5	12	27.1	24.5	29.1	24.4	6.7	27
	28	288.6	30.8	20.3	25.6	10.5	18.6	5	2.4	11.1	21.8	29.2	25.2	100	0	30.9	21.7	27.6	62	17	27.1	24.6	29.3	81.8	6.5	28
	29	311.6	30.8	21.1	25.9	19.7	18.6	18.6	3	0	21.2		24.9	99	0.2	24.7	21.7	24	77	7.1	24	22	5.1	84.3	4.6	29
	30	7779.7	30.8	21.1	25.9	19.7	18.6	19.6		39.5	21.2		24.9		0.2	24.7			77	7.1	24			392.3	32.3	30
	31	7779.7			27.4	50.6	934.6	0	14.9	5	3.8			491.8	0.9	2461.2				66					32.3	31
	Tot.	290.7	305.1	20.4	155	10.1	934.6	8.3				10.6	24.1	98.4	0.4	29.3	64.5	27.7	68.3	13.2	26.8	23.8	27.6	78.5	7.7	Tot.
	Moy.				25.8	10.1	18.7	16.7	3		20.8	0.1		4.1	0.4	0.4	0.4		0.4				1.6			Moy.

Figure B1. Post-quality controlled table using MeteoSaver, showing confirmed values of daily maximum, minimum and average temperature, diurnal temperature range, and daily precipitation (highlighted in green) for the Station Binga in June 1969 (Fig. A1). The description of the colors in the post-quality controlled table is given in Fig. 10.

We further acknowledge this in discussion sections 5.3 and 5.4 that additional meteorological variables and diagnostics may require the incorporation of variable-specific equations or logical constraints within the QA/QC module of the software given its flexibility.

Section 5.3

> While our demonstration focuses exclusively on QA/QC checks for daily temperature and precipitation values, the evaluation of this first version of the software primarily focuses on temperature variables (daily maximum, minimum, and average temperatures). This choice was made because temperature allows the illustration of a broader set of QA/QC procedures available in the sheets, including pentad totals and averages, logical consistency checks (e.g., Tmin < Tavg < Tmax), diurnal temperature range checks, and threshold tests. Together, these checks provide a comprehensive demonstration of the QA/QC framework implemented in MeteoSaver. In contrast, precipitation values presented in the sample sheets include only one QA/QC constraint: the pentad total, which limits the ability to identify and correct erroneously transcribed daily precipitation values. As a result, fewer daily precipitation values can be confirmed through the QA/QC framework compared with daily temperature values (Fig.12 and Appendix Figs. B1-B9).
>
> Nevertheless, future software versions could expand these checks to include additional variables and diagnostics. For instance, in our sheets, the columns under Température et Humidité contain vapor pressure (e) and relative humidity (U) recorded at specific times, which are calculated using the observed dry-bulb temperature (T) and wet-bulb temperature (T'a). This would allow us to incorporate more equations within the QA/QC module to validate an even broader range of transcribed values across the sheets. Therefore, for this initial release, we provide a detailed description of the current QA/QC checks to guide software users and illustrate the framework's flexibility.

Section 5.4

> Our study demonstrates the flexibility of MeteoSaver in transcribing historical tabular weather data across a range of handwriting styles, table dimensions, paper sizes, and maintenance conditions, highlighting its potential contribution to ongoing climate data rescue efforts. Throughout our model development, we focused on reusability for similar case studies, equipping MeteoSaver with numerous configurable settings within the configuration module (see sect. 3.1) to allow users to tailor the software for specific table formats. Given this flexibility, MeteoSaver has substantial potential for transcribing millions of rescued archived weather records, such as those available on the C3S Data Rescue Service Portal. It is important to note, however, that in this initial release, users may need to make further adjustments when applying the software to data sheets with complex tabular formats or variables or data types beyond temperature and precipitation, particularly within the QA/QC checks.

4. The reported processing time of approximately 8 minutes per sheet on a laptop raises questions. Can you provide estimates for batch processing on HPC or cloud infrastructure and discuss expected performance for thousands of sheets and potential bottlenecks.

Response:

We thank the reviewer for raising this concern. The reported processing time of approximately 8 minutes per sheet corresponds to execution on a standard local machine using a single processing thread. However, as described in Sect. 3.1, MeteoSaver is designed to also run on HPC infrastructure (a setting that can be selected in the first module: Configuration) such that the tasks can be parallelized across multiple CPU cores and thus process multiple sheets simultaneously, significantly reducing the total processing time for large archives.

We have added the following sentences in the discussion (Sect.5) to describe the expected processing performance for batch transcription using HPC infrastructure:

> **5 Discussion**
>
> In this demonstration, we illustrate the application of MeteoSaver v1.0 on ten sample sheets, where machine learning algorithm are used both to detect tables and cells and to transcribe the data within them. The software also performs QA/QC checks to flag confirmed values and formats the data into Station Exchange Format (SEF) for upload to open-access repositories. The ten sheets, with various handwriting styles, paper sizes, and maintenance conditions, are used to evaluate its flexibility and accuracy in transcribing historical weather data.
>
> Processing each sheet on a local machine equipped with an 11th Gen Intel® Core™ i7-1165G7 and 16.0 GB of RAM takes under 8 minutes. Because MeteoSaver processes individual sheets independently, and because it can also be executed on HPC infrastructure through the configuration settings (Sect 3.1), the transcription process can be parallelized across multiple CPU cores to allow multiple sheets to be processed simultaneously, significantly reducing the total processing time for large archives. For example, processing 1,000 sheets sequentially on a local machine would require approximately 130 hours, whereas distributing the workload across 20 parallel CPU cores with the same specifications on HPC infrastructure would reduce the processing time to under 7 hours. The parallel processing of individual sheets also means that computationally intensive steps such as image pre-processing, table and cell detection and transcription can take advantage of increased processing power and larger dedicated memory on HPC infrastructure.
>
> While the initial transcription results from this sample are promising, with a median accuracy of 74%, there are limitations in this first version of the software. In the following subsections, we discuss the strengths and weaknesses of this version, highlight developments that were excluded from the initial release, and provide ideas for future improvements in each module.

5. Either extend the validation to at least one additional archive (different country or table layout), or explicitly acknowledge and discuss the limitations of the current validation, clearly stating that the reported performance metrics may not generalize to other historical datasets.

Response:

We thank the reviewer for highlighting this limitation. The validation presented in this manuscript is based on historical weather records from one archive (INERA) collected from different stations, but with relatively consistent table structures. While MeteoSaver is designed to be flexible and adaptable to other archived datasets, its performance may vary depending on factors such as table layout, handwriting style, paper maintenance condition, and the quality of the images (scans). To this, we are currently applying the software to other test sheets from other archives and have so far added an example of two extra layouts from UK data on our GitHub repository for the next improved version.

To address this, we have added a paragraph in the discussion (Sect. 5.3) explicitly acknowledging that the reported performance metrics may not directly generalize to other historical datasets, and that additional retraining or configuration may be required when applying MeteoSaver to different archives with different table formats or handwriting characteristics.

> In our study, the final confirmed temperature values showed a match rate of 52.4–100% with manually transcribed records, yielding a median accuracy of 74% and a median mean absolute error of 0.3°C (see Table 2). While the accuracy indicates the proportion of automatically transcribed values that match the manually transcribed values with a predefined uncertainty margin, the MAE provides an indication of the magnitude of transcription deviations. The median MAE of 0.3°C observed in these sample sheets is comparable to typical uncertainties associated with historical thermometer measurements of 0.2°C (1σ) (Morice et al., 2012; Brohan et al., 2006; Folland et al., 2001).). Because many climatological analyses rely on aggregated statistics derived from combined station data, such as spatial averages and long-term trends, transcription deviations of this magnitude are unlikely to substantially affect the resulting climatological interpretations (Brohan et al., 2006). However, for analyses requiring precise daily values such as extreme-event detection, additional manual verification may still be advisable.
>
> It is important to note that these reported performance metrics are specific to this study's sample weather sheet formats, input image quality, handwriting styles on these sheets, and paper maintenance conditions. Consequently, the reported performance may differ when MeteoSaver is applied to other historical datasets with different table structures, handwriting styles or image quality. Additionally, the variability of climate variables should be considered, for example, while temperature values in the DRC exhibit relatively small annual ranges, extratropical regions often experience much larger seasonal variations (on the order of tens of degrees). In such contexts, transcription errors may be larger and more difficult to detect through the applied QA/QC procedures. Nevertheless, the modular design of MeteoSaver allows users to adapt the configuration and retrain the OCR/HTR models to accommodate different table layouts and handwriting styles.
>
> To enhance transcription accuracy, we recommend further training of the OCR/HTR model on a wider range of handwriting styles, specifically for handwritten digits. This would improve transcription accuracy even prior to the QA/QC step, subsequently enhancing the accuracy of QA/QC-verified values. Moreover, in future software versions, we suggest incorporating a feedback loop where corrected and confirmed values from the QA/QC process serve as additional training data for the OCR/HTR models. This iterative approach would enable the OCR/HTR model to continuously learn from past corrections and improve its ability to transcribe specific handwriting styles over time. This "on-the-fly" learning capability would progressively increase the model's transcription accuracy with each batch of post-processed data.

p.14: "Following the transcription of the data, quality assessment and quality control (QA/QC) is carried out to ensure the final output data is highly accurate with reference to the original handwritten daily temperature records (see Fig. 9)."

>> The phrase “highly accurate” is not operationally defined. It would strengthen the methodology to clarify whether “accuracy” here refers to:

internal consistency (logical constraints, totals, relationships),

agreement with manually transcribed values,

or conformity with physical bounds.

Clarifying this will help readers understand what the QA/QC module is designed to guarantee.

Response:

We thank the reviewer for this clarification request. Regarding MeteoSaver, the QA/QC module performs multiple checks, including logical checks, thresholds (conformity within physical bounds), and consistency with pentad totals and averages. However when referring to “accuracy” of the final output data, we mean the agreement between the automatically transcribed and QA/QC confirmed values and their corresponding manually transcribed values.

We have revised the manuscript to clarify this definition of accuracy in the highlighted sentence above.

> Section 3.5 Module 5: Quality Assessment and Quality Control.
>
> Following the transcription of the data, quality assessment and quality control (QA/QC) is carried out to ensure the final output data are highly accurate with reference to the original handwritten daily temperature records (see Fig. 9). Here, accuracy refers to the agreement between the automatically transcribed and QA/QC confirmed values and their corresponding manually transcribed values, evaluated within a set uncertainty margin to account for small numerical differences arising from rounding or other minor discrepancies. This assumes that the manually transcribed values are correct, which may not always be the case, as manual transcription is also subject to errors depending on the methods applied. This assumption means that the resulting inferred error rates are a conservative estimate.

We iterate this definition of accuracy in the software evaluation and results (Sect 4), where the comparison with manually transcribed records is used to quantify the performance metrics.

> Section 4 Software Evaluation and Results
>
> ….. Here, accuracy is defined as the proportion of automatically transcribed and QA/QC confirmed values that match manually transcribed values, considering a set uncertainty margin of 0.2°C.

p.16:

“If this condition is not met, a specific adjustment, unique to our sheets, is applied: the first digit is removed from the value, and the cell is flagged to indicate this manipulation (see Fig. 11 a-b, with manipulated values in b shown in orange).”

>> This is a data transformation rule, not only a quality check. It would help to explicitly describe this as a correction operation and to specify its assumptions (e.g., why the first digit is assumed to be erroneous, and under what conditions this may fail).

"However, if the check is passed, the transcribed temperature values are then adjusted to match the required decimal places, set to one in this case (see Fig. 11 b–c)."

>> This step modifies the data but is not mathematically described. Please clarify:

whether this is rounding, truncation, or scaling,

and how uncertainty introduced by this step is handled.

"For the daily maximum temperature threshold, we use 40°C. For the daily minimum temperature threshold, we use 5°C."

>> The manuscript would benefit from a brief discussion of how sensitive the results are to these fixed thresholds, and whether they are intended to be region-specific or globally applicable.

Response:

We thank the reviewer for these comments and clarification requests.

We agree that this step represents an additional data correction rather than a single quality check. Regarding the specific adjustment to our sheets that entails manipulating the incorrectly transcribed values (i.e., temperature values with more than four digits), we now include the following explanations and assumptions of this rule in Sect 3.5:

> The first check involves verifying that the transcribed values for Tmax, Tmin, and Tavg contain fewer than four digits. This check is specific to daily temperature values recorded in °C units with one decimal place, where the decimal place is deliberately not recognized by the OCR/HTR model. For example, a value of 27.8 °C would be correctly transcribed as "278" (Fig 11 a-b). Therefore, if more than three digits are detected (e.g. MeteoSaver reads "1278"), it is likely that a wrong transcription was made. If this condition is not met, a specific adjustment—unique to our sheets—is applied: the first digit is removed from the value (i.e. "1278" becomes "278" in our example through this data transformation step), and the cell is flagged to indicate this manipulation (see Fig. 11 a-b, with manipulated values in b shown in orange). This adjustment addresses cases where the OCR/HTR system mistakenly interprets a cell boundary line as an extra digit, such as "1" (as in Fig. 11 a-b). This data transformation assumes that the first digit of the wrongly transcribed value is erroneous which may not always be true, for example, if an extra digit occurs in the middle or at the end of the value.

Regarding the following step that entails modifying the transcribed data to the defined decimal places, we have clarified in Sect. 3.5 that this operation corresponds to a scaling procedure rather than rounding or truncation. Because the OCR/HTR model was restricted to recognize only digits (0–9) to reduce noise from extraneous characters such as dotted table lines that could be misinterpreted as decimal points (Sect. 3.4), the decimal position is restored after transcription. Below is the updated paragraph:

> However, if the check is passed, the transcribed temperature values are then adjusted to match the required decimal places, set to one in this case (see Fig. 11b-c, “278” becomes “27.8” in our example through this postprocessing step). This step corresponds to a scaling operation based on the number of decimal places specified in the configuration settings (Table 1). This is because the original observations were recorded to one decimal place (Fig. 1; Figs. A1–A9), whereas the OCR/HTR model was restricted to recognize only digits (0–9) to avoid misinterpreting dotted table lines as decimal points.

Lastly, regarding the daily maximum temperature threshold of 40°C and the daily minimum threshold of 5°C, we have added a brief discussion in Sect 5.3 addressing the sensitivity of the results to these fixed thresholds as shown below:

> Section 5.3 Quality Assessment and Quality Control
>
> The QA/QC results in this release, as outlined in sect. 3.5, demonstrate that our current pipeline is robust in identifying transcription errors by employing **(i) user-defined data thresholds**, (ii) multi-day totals and averages, (iii) logic checks, such as Tmin < Tavg < Tmax, (iv) iterative comparisons between related variables, including the relationship between Tmin, Tavg, Tmax, and the Diurnal Temperature Range (DTR), and (v) reiteration across the different checks (Fig. 9).
>
> It is important to note that some QA/QC checks, specifically the user-defined thresholds (here, maximum and minimum temperature thresholds), are region-specific and should be informed by expert knowledge or prior climatological studies. However, while these regional thresholds are generally effective for identifying incorrectly transcribed values, they may potentially flag correctly transcribed observations associated with extreme events, potentially excluding these undocumented local extremes in the final output dataset.
>
> Notably, the presence of pentad totals and averages in our sheets proved particularly valuable for QA/QC, as they simplified the process of recalculating and confirming transcribed values within each pentad. In contrast, the more common monthly totals and/or averages found in many international archived weather data sheets would present greater challenges, particularly when multiple daily values are transcribed incorrectly.

p.19:

“Only the confirmed (green) daily temperature values are passed to the next module, Data Formatting and Upload (sect. 3.6).”

>> This implies that a large portion of transcribed data may be excluded. Please indicate the proportion of discarded values and discuss potential impacts on time series completeness. Here

the manuscript transitions from checking to correcting. Explicitly distinguishing these two roles would improve conceptual clarity.

“Two examples … illustrate the sequence of QA/QC checks perfored on the initial transcribed values, leading to the final confirmed values (flagged in green).”

>> Figure 11 shows the propagation of flags and value states, but the underlying equations and replacement rules are not visible in the figure. Consider annotating the panels with the rule names (threshold, digit removal, Eq. 1-4, etc.) to make the logic traceable.

Response:

We thank the reviewer for this comment. We agree that a portion of automatically transcribed data that cannot be confirmed by the QA/QC is excluded from the final output data. To address this concern, we now include information on the portion of discarded (unconfirmed) temperature values in this case study in Sect 4. In our case study, a median of 74.4% of the transcribed values are confirmed by the QA/QC procedure, while 25.6% remain unconfirmed and are therefore excluded from the final output time series. We added a discussion in Sect. 5.3 addressing the potential impacts of these exclusions on time-series completeness and clarifying the distinction between validation checks and correction operations within the QA/QC module as shown below:

> Section 4: Software Evaluation and Results
>
> The results indicate that between 95-100% of the handwritten temperature records (daily maximum, minimum, and average temperatures) from the 10 sheets were successfully detected using the Table and Cell Detection module, with a median of 100% of cells identified across the sheets. These detected cells were then automatically transcribed by our software. Of these transcribed values, a median of 74.4% across the sheets achieved the highest quality flag and were therefore confirmed by the QA/QC. This means that 25.6% of the transcribed values were excluded from the final output timeseries because they could not be confirmed by the QA/QC. The confirmed temperature values showed a median match rate of 74% with manually transcribed records (see Table 2). Here, accuracy is defined as the proportion of automatically transcribed values that match manually transcribed values, considering a set uncertainty margin of 0.2°C. Additionally, we calculate the mean absolute error (MAE) of these automatically transcribed temperature values compared to the manually transcribed ones. The MAE across these transcribed sheets ranged from 0.0-0.9°C, with a median of 0.3°C (see Table 2).
>
> Section 5.3 Quality Assessment and Quality Control
>
> The QA/QC results in this release, as outlined in sect. 3.5, demonstrate that our current pipeline is robust in identifying transcription errors by employing (i) user-defined data thresholds, (ii) multi-day totals and averages, (iii) logic checks, such as Tmin < Tavg < Tmax, (iv) iterative comparisons between related variables, including the relationship between Tmin, Tavg, Tmax, and the Diurnal Temperature Range (DTR), and (v) reiteration across the different checks (Fig. 9).
>
> It is important to note that some QA/QC checks, specifically the user-defined thresholds (here, maximum and minimum temperature thresholds), are region-specific and should be informed by prior climatological studies. However, while these regional thresholds are generally effective for identifying incorrectly transcribed values, they may flag correctly

transcribed observations associated with extreme events, potentially excluding these undocumented local extremes in the final output dataset.

Notably, the presence of pentad totals and averages in our sheets proved particularly valuable for QA/QC, as they simplified the process of recalculating and confirming transcribed values within each pentad. In contrast, the more common monthly totals and/or averages found in many international archived weather data sheets would present greater challenges, particularly when multiple daily values are transcribed incorrectly.

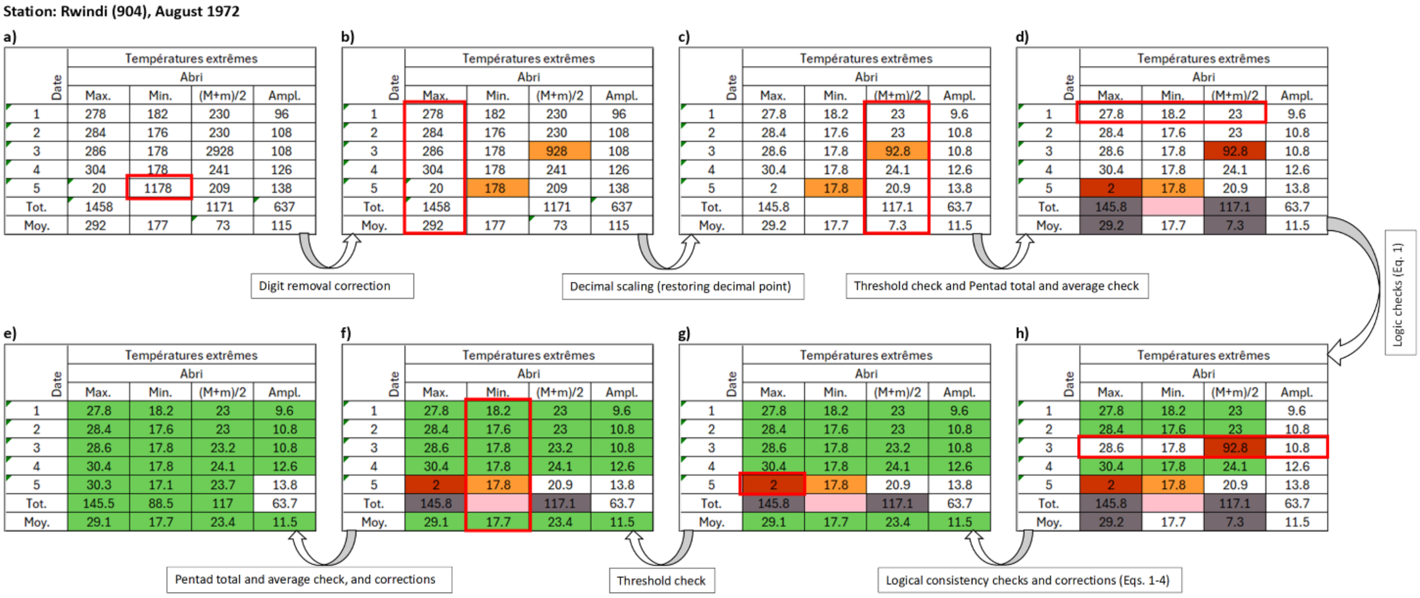
An evaluation of the transcribed values that achieved the highest quality flag revealed a median confirmation rate of 74.4% of the transcribed temperature data across the sheets, either confirmed or corrected during QA/QC. This means that 25.6% of the transcribed values were excluded from the final output timeseries. Among these excluded values, a substantial fraction is correctly transcribed but could not be validated by the QA/QC framework. This may occur, for example, when incorrectly transcribed values appear within the same pentad, preventing confirmation of the correctly transcribed values through the QA/QC checks, or in rare cases when extreme but valid observations fall outside the predefined threshold criteria. While the QA/QC checks are used to validate and refine the transcribed data, as demonstrated in Fig. 11, their effectiveness heavily relies on the initial transcription quality, which depends on the OCR/HTR model, the variability in handwriting styles, and the maintenance condition of the paper sheets in the archives. For instance, when the paper condition is well-preserved (as in Fig. 1) and the initial transcription is nearly accurate across most cells, only the first few QA/QC checks are typically sufficient to confirm all daily temperature values in a pentad, as illustrated in Fig. 11 i - j. On the other hand, in cases where the initial transcription contains multiple errors, all QA/QC checks and iterative re-evaluations in our framework are necessary to confirm the temperature values in that pentad, as seen in Fig. 11 a - h. The latter could, in rare cases, lead to incorrectly confirmed values if the originally transcribed data contains errors that still meet multiple QA/QC checks, or may result in values falling outside the user-defined uncertainty margin, which would therefore remain unconfirmed.

3.5 Module 5: Quality Assessment and Quality Control

Following the transcription of the data, quality assessment and quality control (QA/QC) is carried out to ensure the final output data is highly accurate with reference to the original handwritten daily temperature records (see Fig. 9). Here, accuracy refers to the agreement between the automatically transcribed and QA/QC confirmed values and their corresponding manually transcribed values, evaluated within a set uncertainty margin to account for small numerical differences arising from rounding or other minor discrepancies. This assumes that the manually transcribed values are correct, which may not always be the case, as manual transcription is also subject to errors depending on the methods applied. This assumption means that the resulting inferred error rates are a conservative estimate.

This module performs two complementary roles: (i) validation checks that assess whether the transcribed values are logically and physically consistent, and (ii) correction operations that adjust specific transcription errors where the correct value can be inferred from the structure of the table such as totals or averages, or from related variables.

We also thank the reviewer for the suggestion to annotate the panels with the QA/QC rule names and/or equations. We have updated Fig. 11 such that the QA/QC steps between the panels are labelled with their corresponding rule names (e.g., digit removal correction, threshold checks, pentad total and average checks), allowing readers to more easily trace the logic of the QA/QC procedure described in Sect. 3.5:



Station: Rwindi (904), August 1972

a)

Date	Températures extrêmes			
	Abri			
	Max.	Min.	(M+m)/2	Ampl.
1	278	182	230	96
2	284	176	230	108
3	286	178	2928	108
4	304	178	241	126
5	20	1178	209	138
Tot.	1458		1171	637
Moy.	292	177	73	115

b)

Date	Températures extrêmes			
	Abri			
	Max.	Min.	(M+m)/2	Ampl.
1	278	182	230	96
2	284	176	230	108
3	286	178	928	108
4	304	178	241	126
5	20	178	209	138
Tot.	1458		1171	637
Moy.	292	177	73	115

c)

Date	Températures extrêmes			
	Abri			
	Max.	Min.	(M+m)/2	Ampl.
1	27.8	18.2	23	9.6
2	28.4	17.6	23	10.8
3	28.6	17.8	92.8	10.8
4	30.4	17.8	24.1	12.6
5	2	17.8	20.9	13.8
Tot.	145.8		117.1	63.7
Moy.	29.2	17.7	7.3	11.5

d)

Date	Températures extrêmes			
	Abri			
	Max.	Min.	(M+m)/2	Ampl.
1	27.8	18.2	23	9.6
2	28.4	17.6	23	10.8
3	28.6	17.8	92.8	10.8
4	30.4	17.8	24.1	12.6
5	2	17.8	20.9	13.8
Tot.	145.8		117.1	63.7
Moy.	29.2	17.7	7.3	11.5

e)

Date	Températures extrêmes			
	Abri			
	Max.	Min.	(M+m)/2	Ampl.
1	27.8	18.2	23	9.6
2	28.4	17.6	23	10.8
3	28.6	17.8	23.2	10.8
4	30.4	17.8	24.1	12.6
5	30.3	17.1	23.7	13.8
Tot.	145.5	88.5	117	63.7
Moy.	29.1	17.7	23.4	11.5

f)

Date	Températures extrêmes			
	Abri			
	Max.	Min.	(M+m)/2	Ampl.
1	27.8	18.2	23	9.6
2	28.4	17.6	23	10.8
3	28.6	17.8	23.2	10.8
4	30.4	17.8	24.1	12.6
5	2	17.8	20.9	13.8
Tot.	145.8		117.1	63.7
Moy.	29.1	17.7	23.4	11.5

g)

Date	Températures extrêmes			
	Abri			
	Max.	Min.	(M+m)/2	Ampl.
1	27.8	18.2	23	9.6
2	28.4	17.6	23	10.8
3	28.6	17.8	23.2	10.8
4	30.4	17.8	24.1	12.6
5	2	17.8	20.9	13.8
Tot.	145.8		117.1	63.7
Moy.	29.1	17.7	23.4	11.5

h)

Date	Températures extrêmes			
	Abri			
	Max.	Min.	(M+m)/2	Ampl.
1	27.8	18.2	23	9.6
2	28.4	17.6	23	10.8
3	28.6	17.8	92.8	10.8
4	30.4	17.8	24.1	12.6
5	2	17.8	20.9	13.8
Tot.	145.8		117.1	63.7
Moy.	29.2	17.7	7.3	11.5

Digit removal correction
Decimal scaling (restoring decimal point)
Threshold check and Pentad total and average check
Logic checks (Eq. 1)
Pentad total and average check, and corrections
Threshold check
Logical consistency checks and corrections (Eqs. 1-4)

Figure 11. Two examples of pentad transcribed temperature values (Top: Station Rwindi [S 0°47’, E 29°17’], August 1972, and Bottom: Station Binga [N 2°18’, E 20°30’], May 1969) illustrate the sequence of QA/QC checks performed on the initial transcribed values, leading to the final confirmed values (flagged in green). The arrows, along with their respective labels between each panel, indicate the specific QA/QC checks applied at each stage for the pentad, corresponding to the procedures and equations described in Sect. 3.5 and illustrating the progression of the QA/QC workflow. Red-bordered cells, rows, or columns highlight examples where unconfirmed transcribed values were identified and corrected during QA/QC, with changes reflected in the subsequent panel. See Key for all colors (quality flags) in Fig. 10.

p.20:

“At this stage, an additional check is performed, which was not included in the QA/QC module due to the availability of longer temperature series at this point.”

>> This introduces a new methodological step after the main QA/QC description. For structural clarity, it may be preferable to describe this earlier as an optional extension of Module 5.

Response:

We thank the reviewer for this suggestion. We agree that introducing this additional check within the QA/QC module would improve the structural clarity of this manuscript. Therefore, we have moved the description of this step to the end of Sect. 3.5, and now describe it as an optional extension of the QA/QC module.

3.5 Module 5: Quality Assessment and Quality Control

…..

In addition to the checks described above, an optional time series consistency check can be applied once a longer temperature series is available for a given station. This step requires the consolidation of daily observations across multiple sheets and therefore cannot be applied during the initial QA/QC stage. It involves identifying outliers in the temperature time series for each station by using standard deviation to detect unusual patterns in the transcribed temperature records (as in (Chauhan and Parashar, 2020)). This check within our framework unfolds in two steps, following the creation of a distribution of all values across the station's time series: (i) Temperature values that deviate more than three standard deviations from the mean are identified, flagged, and removed (until confirmed by an expert). (ii) we identify abrupt transitions in daily temperature by examining the standard deviation differences of consecutive days. Specifically, we check for cases where a large deviation (e.g., less than -4 standard deviations from the mean) on one day is followed by a large deviation in the opposite direction (e.g., more than +4 standard deviations) on the next day, and then a return to a similar deviation on the third day. When this pattern is observed, we flag the middle day as an outlier and remove it from the time series (until confirmed by an expert). For example, if a sequence of days shows a temperature that deviates significantly below the mean, followed by a sharp increase above the mean, and then returns to a lower deviation on the following day, the middle day is flagged. This approach allows us to capture rapid shifts in temperature that may indicate transcription errors or anomalies, even if the individual values do not exceed the fixed ±3 standard deviation threshold used in the first method. Although included in our framework, the first step is not applied in this demonstration because we illustrate with a single month's data (a short series), where it could lead to mistakenly removing extreme but valid values.

Added References:

Brohan, P., Kennedy, J. J., Harris, I., Tett, S. F. B., and Jones, P. D.: Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850, Journal of Geophysical Research: Atmospheres, 111, https://doi.org/https://doi.org/10.1029/2005JD006548, 2006

Folland, C. K., Rayner, N. A., Brown, S. J., Smith, T. M., Shen, S. S. P., Parker, D. E., Macadam, I., Jones, P. D., Jones, R. N., Nicholls, N., and Sexton, D. M. H.: Global temperature change and its uncertainties since 1861, Geophysical Research Letters, 28, 2621–2624, https://doi.org/https://doi.org/10.1029/2001GL012877, 2001

Morice, C. P., Kennedy, J. J., Rayner, N. A., and Jones, P. D.: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set, Journal of Geophysical Research: Atmospheres, 117, https://doi.org/https://doi.org/10.1029/2011JD017187, 2012