Digitization and calibration of historical solar absorption infrared spectra from the Jungfraujoch site

Jamal Makkor[1], Mathias Palm[1], Matthias Buschmann[1], Emmanuel Mahieu[2], Martyn P. Chipperfield[3], and Justus Notholt[1]

[1]Institute of Environmental Physics (IUP), University of Bremen, Bremen, 28359, Germany
[2]Institute of Astrophysics and Geophysics, UR SPHERES, University of Liège, Liège, 4000, Belgium
[3]School of Earth and Environment, University of Leeds, UK

Correspondence: Jamal Makkor (makkor@uni-bremen.de)

Abstract. This study describes the digitization and calibration of historically significant solar absorption spectra recorded at the Jungfraujoch International Scientific Station in the 1950s. Using a homemade Pfund-type grating spectrometer, these spectra were recorded on paper rolls to study the solar spectrum which was then used to compile a solar atlas between 2.8 and 23.7 ~~microns (~421 to $3571cm^{-1}$) that~~ μm (421 to $3571cm^{-1}$) that, in particular, later contributed to the development of the HITRAN database. We have now digitized these ~~old~~ analogue recorded spectra to make them available for atmospheric studies. Our approach involves image processing techniques, including colour masking for digitization and peak detection for accurate wavenumber calibration against a synthetic spectrum.

We have also developed a validation method by re-digitizing degraded ~~FTIR~~ Fourier transform infrared (FTIR) spectra to the same resolution as the old spectra to evaluate the digitization accuracy. Furthermore, we have studied the influence of line thickness on the digitization error.

The number of spectra transformed into a machine-readable format is ~~108~~ 106 (freely available for download), with an average digitization error of 1.55% and a wavenumber shift standard ~~diviation~~ deviation of $0.075cm^{-1}$. These digitized and calibrated spectra now offer a valuable resource for atmospheric studies, providing essential historical data for atmospheric research. This work not only helps to preserve scientific heritage but also enhances the utility of historical data in contemporary research.

1 Introduction

The Jungfraujoch station is located in the Swiss Alps (46.55N, 7.98E) at an altitude of 3580m, on a saddle between the Mönch (4107m) and the Jungfrau (4158m) summits. The Sphinx observatory, built in 1931 (ISS, 1931), focuses on atmospheric studies (among many other scientific disciplines) and provides continuous measurements of various species using state-of-the-art technologies (Zander et al., 2008). Its strategic location at high altitude, coupled with minimal interference from pollution and water vapour, renders it an optimal setting for such studies.

The first atmospheric measurements on this site started in the 1940s using ~~grating based~~ grating-based spectrometers, including the one-metre focal length Pfund-type prism-grating instrument described in this paper, which operated between 1950 and 1951. Later on, from the late ~~1950 to 1980~~ 1950s until 1980s, a seven-metre focal length Ebert/Fastie-type prism grating

spectrometer, designed at the University of Liège (U. Liège), was installed for studies in the near-UV to near-IR range, covering wavelengths from 300 to 1200 nm. In 1984, the observatory became aware of the prevalence and efficiency of Fourier Transform Infrared based spectrometers (FTIR) compared to their grating counterparts, and therefore augmented its capabilities by integrating a homemade FTIR spectrometer. Then starting from the early 1990s, an FTIR manufactured by the German company Bruker (IFS 120HR) and modified by the U. Liège team was added and operated in parallel with the homemade spectrometer. The homemade FTIR was retired in 2008 (Zander et al., 2008), while the Bruker instrument has been in use (Prignon et al., 2019) until the summer of 2024. It has now been replaced by a Bruker IFS125HR instrument.

The solar atlas that resulted from the spectral rolls and produced by Migeotte et al. (1956), extending from 2.8 to 23.7 μm, has been a historically scientific achievement and a valuable resource for scientists for decades. It includes spectra of atmospheric gases such as carbon dioxide, ozone, and water vapour, covering a wide spectral range. The process of identifying spectral features in this atlas was carried out by Migeotte from 1950 to 1956. Researchers from various fields, including climate science (Zander et al., 1994), atmospheric physics (Ehhalt et al., 1983), and chemistry (Murcray et al., 1978), have utilized this information to understand the behaviour and interactions of these gases in Earth's atmosphere. Migeotte used the instrument to detect CO at the top of the Alps, a location previously assumed to be unpolluted (Migeotte and Neven, 1950). Zander et al. (1989) then calculated the total column of this gas using these spectra. Additionally, Zander et al. (2008) investigated the existence of dichlorodifluoromethane (CCl_2F_2, commonly known as CFC-12 or Freon12™) in the old spectra through visual analysis of spectral lines recorded in 1951, alongside those obtained from a homemade FTIR spectrometer. Additionally, the spectra measured by a Bruker 120 HR Fourier Transform Spectrometer (FTS) at a resolution of $0.0035 cm^{-1}$ were included in the analysis. All measurements were taken under comparable solar zenith angle conditions (SZA) in the year 2000.

Despite their potential value, the spectra have not been fully utilized by the scientific community, primarily because they are not available in a digital format that can be easily shared. In this work, we have developed a method to digitize and calibrate these spectra to the correct wavenumber range. To achieve this, a digitization/calibration software has been developed. This software transformes the original paper spectra into a machine-readable and universally accessible ASCII format. The digitization process employs image processing techniques to detect plotted lines and extract the spectrum, while the calibration phase involves curve-fitting the produced spectrum (pixel-mapped) to a synthetic one covering well-known atmospheric lines on a known wavenumber scale. The current paper describes this process in the sections below.

2 Methods

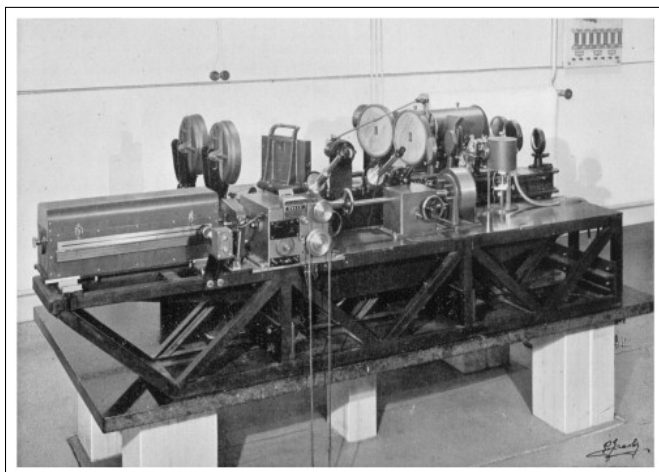
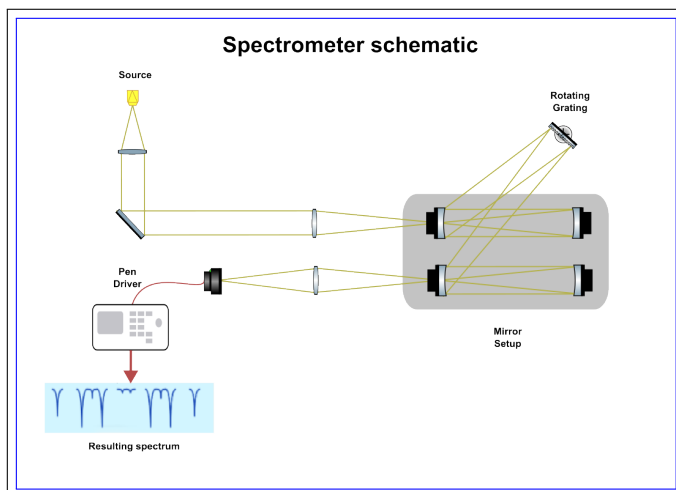
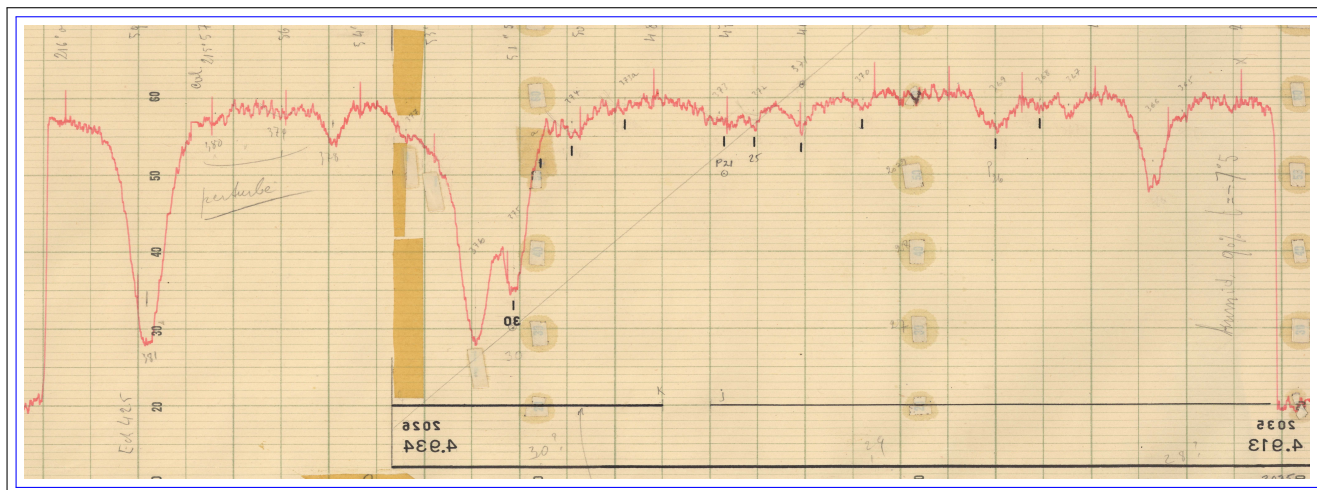
The paper spectra were scanned using an Epson A3 (WORKFORCE DS-50000 (EDC)) professional scanner. The resulting high-resolution uncompressed TIFF files were then preprocessed, corrected for any misalignment, and combined into a single file. To digitize and calibrate these spectra, a software was developed using Python and the Tkinter graphical library.

(a)

(b)

(c)

Figure 1. The spectrometer used to produce the historical spectra. **(a)** The instrument was a home-made one-meter focal length Pfund-type grating spectrometer. It was described by Migeotte when he reported the solar atlas in 1956 (Photograph from Zander et al. (2008)). **(b)** A simplified schematic of the Pfund-type grating spectrometer (adapted from Miller and Thompson (1949)). **(c)** Example of a resulting spectrum showing hand-written notes and traces of adhesive tape.

The spectra were stored in a machine-readable format not only for easier analysis but also to preserve them for posterity. The Pfund-type grating that was used to record these spectra (shown in Figure 1a and illustrated in Figure 1b) used a thermocouple as a detector (Table 1). Thermocouples have commonly been used in the past as detectors of infrared radiation due to their cost-effectiveness, simplicity, and wide temperature range. However, a significant drawback of thermocouple-based spectra is the high level of thermal noise. This causes these types of detectors to have a lower ~~signal to noise~~ signal-to-noise (S/N)

60

Table 1. Technical specifications of the Pfund-type ~~Grating Instrument~~ grating instrument and the produced spectra.

Parameter	Value
Operation years	1950 to 1951
Resolution [~~cm^{-1}~~ cm^{-1}]	0.12 to 0.4
Signal to noise ratio	40 to 80
Spectral range (at various intervals) [cm^{-1}]	520 to 3565
The instrument lineshape (ILS) function	$sinc^2$
Apodisation function	Triangular
Detector	Perkin Elmer Thermocouple

ratio compared to the recent ones. Additionally, an important quantity calculated during the extraction of the spectra is the ~~solar zenith angle (SZA)~~ SZA which is often used in atmospheric retrievals. The SZA was calculated based on date, time and location. However, the measurements using the grating took a relatively long time, up to 2.5 hours (1.5h average recording

65 time). This influences the value of SZA and subsequently the airmass. This also needed to be taken into account ~~as well~~ when producing the spectra. Therefore SZA/airmass was not a single value, instead it was calculated using the start and end time of the recording at a regular interval of 1 minute.

2.1 Digitization

This section describes the method used to extract spectral data from the spectral rolls. ~~They~~ The rolls were marked with

70 additional information, such as the date, start and end recording time, high and low wavenumber, and occasionally surface temperature and humidity (which were gathered from notes on the papers during the scanning process). The recorded lines were plotted using red ink, and an algorithm was developed to detect these lines using image processing.

2.1.1 Scanning and Preprocessing

The spectral rolls presented a range of conditions, some of which contained annotations from various scientists who had previ-

75 ously engaged in the analysis of these spectra. Others were repaired using adhesive tape (see Figure 1c). Notably, the presence of tape did not impede the digitization process, as its coloration was distinct from that of the spectral lines. A critical preparatory step involved ensuring that the spectral rolls were aligned as horizontally as possible prior to scanning, a measure taken to mitigate potential distortions. However, very small misalignment that cannot be observed by the naked eye, proved to also have an effect on the total digitization error. This was observed during the validation phase (a slope in the difference between

80 calibrated and original spectrum caused by a small rotation in the image). To mitigate this, an automatic alignment and cropping algorithm was also developed to correctly adjust the images.

The ~~employed~~ scanner offers A3 format capabilities and a resolution of 600 ~~dpi~~ dots-per-inch (DPI) for uncompressed im-

ages, which were notably large in file size (thus offering higher digitization accuracy, but needing longer processing periods). However, due to the extended length of some rolls, it was necessary to perform multiple scans to adequately capture the entire

85   spectral range of a single roll. Subsequently, these individual scans were automatically merged into a full spectrum using image editing software (GIMP (2019) and Hugin (2020)). Once the spectral images were obtained, they were processed for plot extraction using the designed algorithm.

2.1.2   Digitization algorithm

Figure 2a shows the GUI (Graphical User Interface) used to digitize the paper spectra. The digitization algorithm used to turn

90   the plots in the image files into a readable text format runs as follows:

- Image reading and colour space conversion:

  Python ~~OpenCV~~OpenCV's library (Zelinsky, 2009) is used to read and process the images. The colour space of the image is converted from BGR (Blue, Green, Red) to HSV (Hue, Saturation, Value) to facilitate colour-based filtering.

- Straightening and cropping:

95   Before the digitization, an automated straightening step was applied to the image. This is achieved by fitting a rectangle to the image and then applying a rotation matrix to it. The angle used for this operation is calculated between the fitted rectangle and the horizontal line in the image canvas. Following this, the resulting image is then cropped.

- Colour ~~Filtering~~filtering:

  A mask is created to filter out specific colour ranges, in this case, using a minimum and maximum threshold. This step

100   isolates parts of the image relevant to the spectral data, this threshold can be adapted to detect any given colour. Figure 2b shows the spectrum taken from the detected line. The visible discontinuities in the spectrum come from the ruled lines which interfere with the colour detection.

- Binary ~~Thresholding~~thresholding:

  The masked image is converted to greyscale. Following this, a binary threshold is applied to create a black and white

105   image, further isolating the spectral lines (Figure 2c).

- Spectral ~~Data Extraction:~~data extraction:

  The spectral data is extracted by calculating the mean of the pixel values along the vertical axis of the columns of the binary image then linearly interpolated to account for any missing pixels. This step translates the visual spectrum into numerical values.

110   Since the line has an inherent thickness coming from the pen, the mean of the thickness of the detected line was taken as the resulting spectrum. However, the true spectrum lies in between the border of this line and the thickness needs to be accounted in the calculation of error. The standard deviations of the detected plot points are shown in Figure 2d. The calculated mean deviation is about 9.3 pixels, taking into account an ~~image~~ scanner resolution of 600 ~~dpi~~DPI.

This gives us a relative mean digitization error of about: $\epsilon = \frac{9.3*100}{600} = 1.55\%$ This error is dependent on the scanner resolution used. However, when taking the the height of the image into consideration, calculating the error needs to take into account the full height of the image in pixels. Therefore, for an image height of 3000 pixels the error is approximately 5 times smaller.

2.1.3 Accompanying metadata

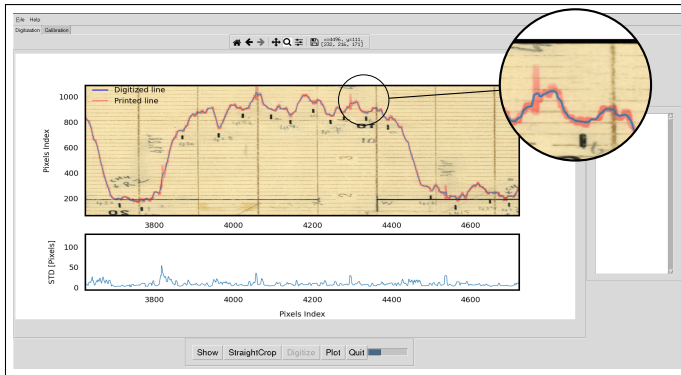
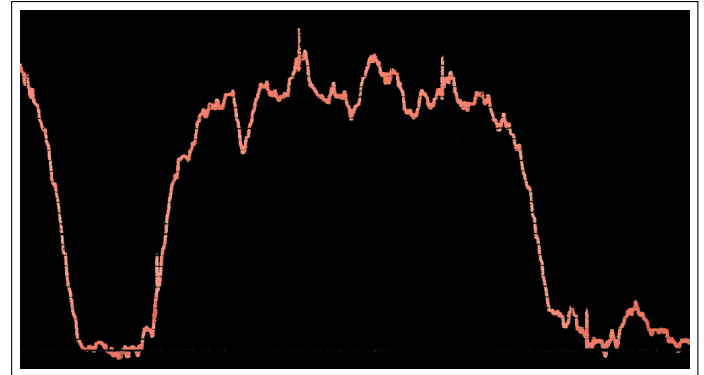
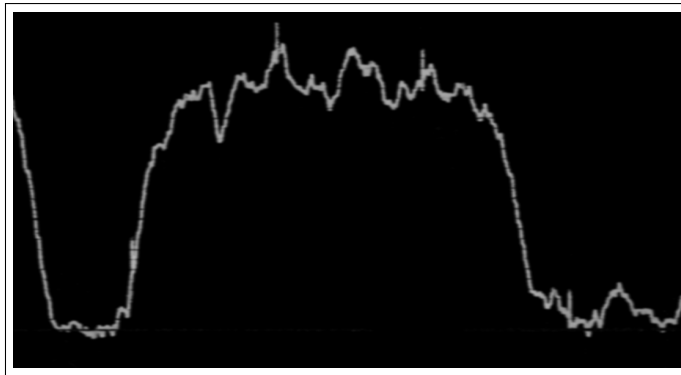
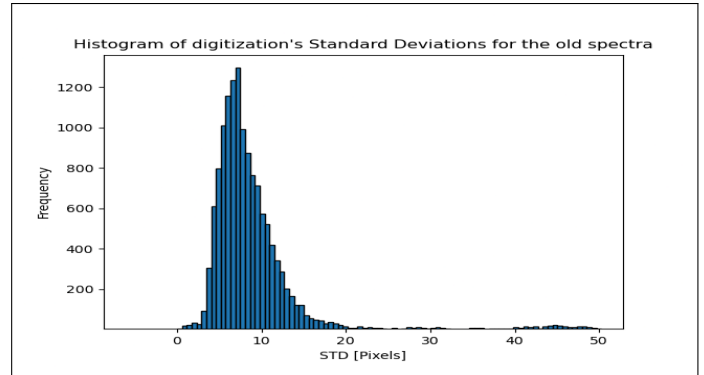
In addition to the spectral data, housekeeping information was also recorded and stored alongside the generated spectrum. This data encompassed the lowest and highest wavenumber, ~~solar zenith angle~~ SZA range, and airmass. When documented on the rolls, temperature and humidity values were also included.

(a)

(b)

(c)

(d)

Figure 2. Digitization of a scanned old spectrum. **(a)** The digitized spectrum (in blue) extracted from the pen plotted line (in red). The circled ~~zoomed in~~ zoomed-in view shows sudden peaks that appear throughout the spectra. **(b)** Resulting image after applying the colour detection mask. ~~this~~ This method allows the detection of a line despite the handwritten annotation in the image. **(c)** The resulting image after binary threshholding of the grayscale image. **(d)** Histogram of the standard ~~diviations~~ deviations of each digitized point in the spectrum. The mean of the calculated deviation is about 9.3 pixels.

120

2.2 Calibration

The digitized spectrum, saved only as pixel positions, lacks a definition in the correct wavenumber range. After digitization, calibration to this range is necessary. This was achieved by first producing a synthetic spectrum with a known wavenumber range. Then, after choosing appropriate calibration points, fitting the digitized data-points to the synthetic spectrum ~~thus defining the appropriate wavenumber scale.~~ wavenumber range using 2nd-degree polynomial least squares curve fitting algorithm. Details of this calibration procedure are explained below.

2.2.1 Synthetic spectrum preparation

The synthetic spectrum was generated using the SFIT4 forward model, leveraging known spectroscopy from the High-Resolution Transmission Molecular Absorption Database (HITRAN) (Gordon et al., 2017) and specific instrumental parameters, such as the apodization function and resolution (https://wiki.ucar.edu/display/sfit4/). The synthetic spectrum was calculated using an averaged resolution of $0.25\,\mathrm{cm}^{-1}$ ~~and a triangular apodization according to the properties of the Pfund spectrometer~~ (see Table 1).

~~The choice of the apodization function for the interferograms is connected to the instrumental properties of a grating spectrometer, since for a diffraction limited resolution , where the image of the spectral line is no longer the same shape as the entrance slit but controlled by the diffraction pattern, the instrumental line function in the spectral domain corresponds to a $sinc^2$ function (Griffiths, 2002) (therefore justifying the use of a triangular apodization in the interferogram domain when producing the synthetic spectrum).~~ Additionally, due to the limited resolution of the spectrometer and the fact that the ILS is dominated by the aperture, triangular apodisation function is used to create the synthetic spectrum (Griffiths, 2002). The spectral lines of the synthetic spectrum do not perfectly match the calibrated spectrum due to line broadening and line strength differences, however, their ~~center~~ centre positions remain unchanged allowing for the calibration.

The synthetic spectrum was produced to cover the whole range of the spectral rolls (from $500cm^{-1}$ to $5000cm^{-1}$) using a fixed $0.25cm^{-1}$ resolution. However, it is worth mentioning that for a grating spectrometer the resolution is ~~waevnumber dependent~~ waevnumber-dependent and the averaged resolution was thus used for practical reasons. ~~This spectrum is then read and plotted by the software to be used in the subsequent calibration step.~~

2.2.2 ~~Calibration methodology~~

~~The calibration algorithm used to map the digitized~~

2.2.2 Calibration methodology

150 The least squares fitting algorithm (using the Python library `lmfit`) uses matching pixel points from the digitized spectrum and wavenumber points from the synthetic one to correctly fit the digitized spectrum to the appropriate wavenumber scale. The reason for choosing this method to calibrate the spectrum to the correct wavenumber range was due to the fact that the wavenumbers given in the spectral rolls were not very precise. This calibration method has been checked for larger spectral intervals, and the fitted spectra matched the digitized ones really well (see Section 3). The calibration is performed as follows:

155 – Metadata such as date, time, minimum and maximum wavenumber and location coordinates are read from a configuration file saved during the spectrum extraction phase (see Section 2.1.2). They are then used to calculate the ~~solar zenith angle~~SZA, azimuth and various other parameters which are saved in the final spectrum or provided in the accompanying housekeeping file.

160 – Both the synthetic spectrum and the digitized pixel values are loaded side by side (see Figure 3a). The synthetic spectrum is shown at the appropriate wavenumber range read from the configuration file, and the peak detection ~~boundary~~ prominence is defined (horizontal blue line).

– The relative maxima search method is the peak detection algorithm used here to identify calibration points. This method identifies a data point as a peak if it is greater than its ~~neighbors~~ neighbours on both sides, effectively pinpointing the maximum values of each spectral line. While this approach is typically effective, there may be instances where the peak
165 detection identifies more data points in either the digitized spectrum or the synthetic one. In such cases, manual removal of misidentified peaks is also possible.

– In case the peak detection algorithm fails (if the ~~paper based~~ paper-based spectrum is too noisy, for example), it is possible to manually select calibration points.

– The calibration points are then saved and used to curve fit the digitized data points to the synthetic spectrum using a
170 standard least square fitting ~~method and a~~ using the following 2nd degree ~~polynomial model (which was sufficient to give a good fit without risking overfitting at higher degree ones)~~equation $y = a \times x^2 + b \times x + c$ and the fitting library `lmfit`. The spectrum is then saved ~~as a text~~in a text file format with the appropriate wavenumber range, ready for use.

– The slope of the spectrum has been left unchanged and the product is simply a normalized spectrum using the zero level and the maximum signal. The spectrum itself is a transmission spectrum, and the 100% background signal is unknown.
175 However, the latter is assumed constant for trace gas retrievals at smaller spectral windows.

2.2.3 Choice of curve fitting function for wavenumber calibration

In our analysis of curve fitting models, we assessed both χ^2 values (which represents an insightful overview of the goodness of the fit) and the standard deviation in wavenumber shift. The 2nd~~degree~~ -degree polynomial emerged as the optimal model, balancing fit quality and wavenumber shift precision. While the 3rd~~degree~~ -degree polynomial showed a lower χ^2 value (~~more
180 than half of that for~~ approximately third of the 1st~~degree~~ -degree polynomial), it suffered from a higher standard deviation

(a)   (b)

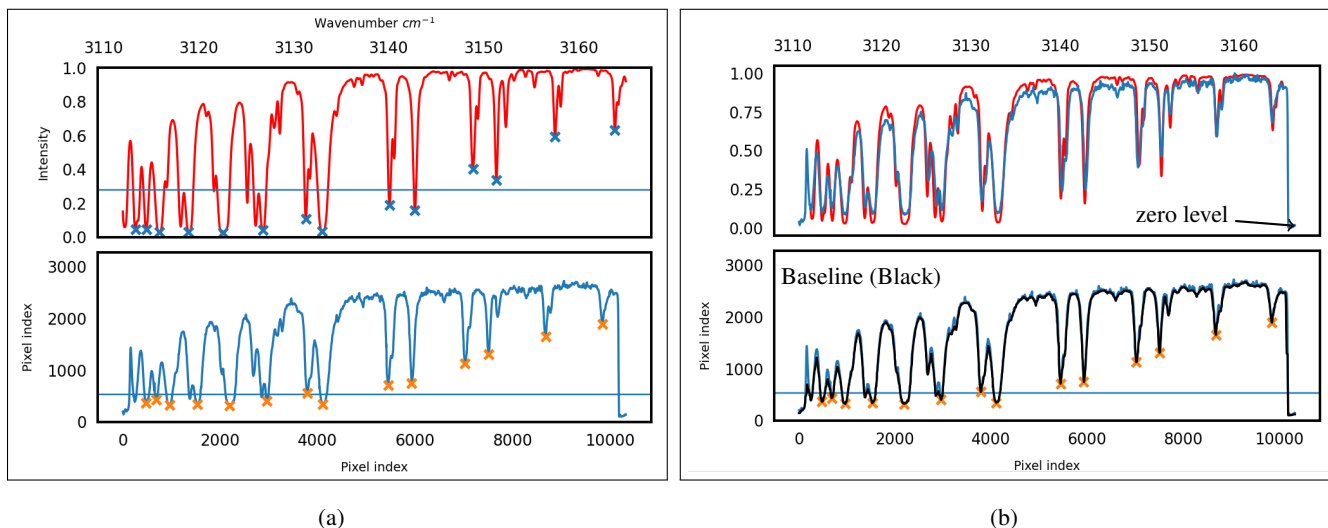



Figure 3. Calibration of an old spectrum. **(a)** The lower ~~plot~~ panel shows the digitized spectrum and the upper ~~plot~~ panel shows the normalized synthetic spectrum (generated using SFIT4). Both spectra show the detected calibration points by using a chosen detection limit (blue horizontal line). **(b)** The lower ~~plot~~ panel shows the digitized (blue) spectrum with its baseline (black). The upper ~~plot~~ panel shows both the calibrated (blue) and simulated (red) spectra.

(STD) in wavenumber shift (more than twice that of the 2nd degree polynomial). The 1st ~~degree~~ degree polynomial, though simpler, had the highest χ^2 value and a relatively high standard error, indicating a poorer fit and less reliability. Given these observations, the 2nd ~~degree~~ degree polynomial was selected for its better balance between a lower χ^2 value (only slightly higher than the 3rd ~~degree~~ degree polynomial) and a significantly more stable standard deviation of the wavenumber shift.

185   Figure 4 shows a plot of the wavenumber versus the pixel calibration points detected by the peak detection algorithm and fitted to a 2nd degree polynomial ~~and the~~. The mean values of different statistics for three polynomial models are detailed in ~~the~~ Table 2.

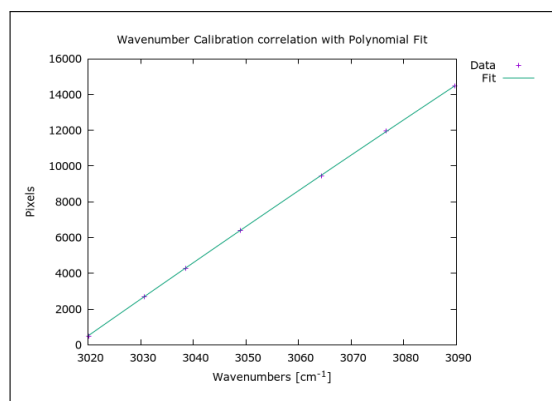



Figure 4. Pixel calibration to the correct wavenumber fitted to a 2nd degree polynomial.

Table 2. Comparison of polynomial models used in curve fitting the digitized spectrum.

Statistic	N>3 Polynomial	N3 Polynomial	N2 Polynomial	N1 Polynomial
χ^2	No Fit	0.03	0.05	~~0.1~~ 0.11
Wavenumber Shift ~~Std~~ STD (cm^{-1})	-	± 0.165	± 0.065	± 0.127

The spectra displayed abrupt vertical spikes, or upticks, throughout the recording, the origin of which remains unclear (see the ~~zoomed in~~ zoomed-in circle in Figure 2a). These upticks posed challenges in saving a normalized spectrum at the accurate

190 zero level. Fortunately, the zero level was consistently indicated in the recorded spectra. To mitigate this, each spectrum was fitted to a baseline, effectively ignoring sudden peaks and determining the maxima and minima for normalization. Each normalized spectrum was saved as an ASCII text file.

The error at the zero level can be explained by the fact that even in the absence of a signal (i.e. no solar radiation) at the detector there are still some voltage fluctuations (noise) recorded by the pen on the paper. This adds some additional uncertainty to

195 retrievals performed by the spectra. This offset error was estimated by Zander et al. (1994) to be around 3% (referred to as zero transmission offset).

During the extended recording durations, the SZA shows significant variations, potentially leading to inaccuracies in retrieval values. The `pvlib` package was utilized to ~~calculate~~ compute the SZA array ~~at different recording times, the results were compared against NOAA's solar zenith angle calculator (https://gml.noaa.gov/grad/ antuv/SolarCalc.jsp).~~ where each SZA

200 value was calculated at 1-minute increments from start to end recording time.

3 Method validation

The original spectra were produced in the absence of comparable datasets or spectra for validation, requiring unique validation approach for our digitization and calibration method. To verify the efficacy of the digitization method we utilized high-resolution FTIR spectra from the Jungfraujoch site ~~which~~. These were first artificially lowered to a resolution of $0.25cm^{-1}$ by

205 truncating the Fourier transform of the high-resolution spectrum to match the required resolution. Subsequently, the spectra were subjected to triangular apodization (the reason for this is explained in Section 2.2.1), a process facilitated by the Bruker Opus Software (OPUS, 2017).

Following this, we plotted the ~~low resolution~~ low-resolution FTIR spectrum (using a similar pen colour) on a background with a texture comparable to the original paper used to produce the old spectra. We also added similar annotation to this back-

210 ground. This spectrum was then printed on physical paper to try and replicate the effects of printing, like the physical aging of the paper, and non-uniform colour of the plotted line. It was then rescanned back to a digital format at 600 ~~dpi~~ DPI resolution. Finally, the resulting image was digitized and compared to the original spectrum. Figure 5b illustrates a comparison between an example of the artificially degraded ~~high resolution~~ high-resolution (HR) FTIR spectrum and the resultant digitized spectrum.

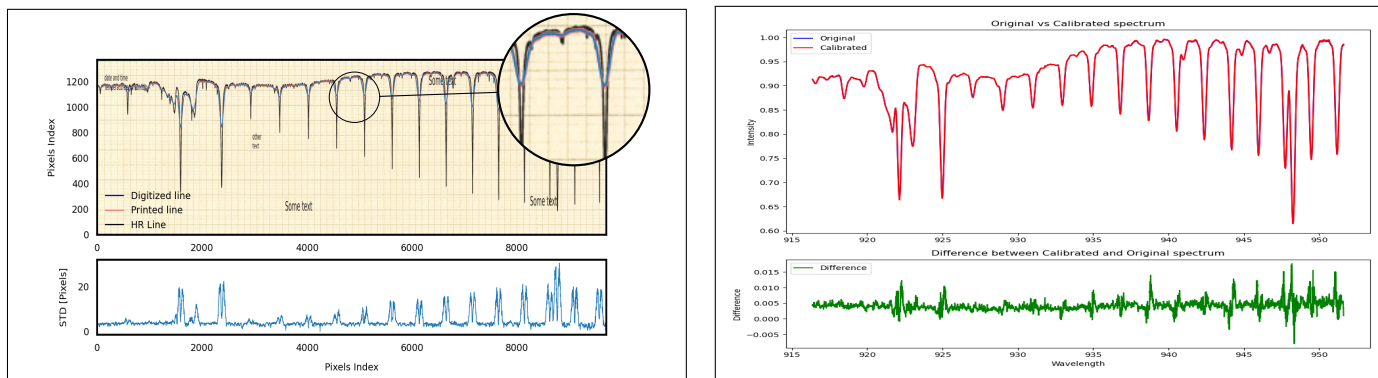
The residuum in this case is about 1.5%. This could be attributed to multiple ~~reasons~~causes. One of ~~which could be caused by~~ these could be printer accuracy, where printers might not reproduce the colours and details of the original spectrum. Another ~~reason could also~~ factor could be attributed to scanning error where the colour might be modified by the scanner.

(a)   (b)

Figure 5. Lower resolution FTIR spectrum printed on paper and scanned. **(a)** The high-resolution (HR) FTIR spectrum (in black) was reduced to a lower resolution and ~~appodized~~ apodized using a triangular function (plotted in red) before being digitized (the digitized line in blue is plotted over the low resolution line). **(b)** Comparison between the original spectrum and the digitized one is showing a good alignment. A digitization residuum of 1.5% is observed.

The observed increase in residuum especially at deeper lines can be explained by an amplified error at higher slopes, where small discrepancies between the digitized spectrum and the original one can produce larger differences thus influencing the error. This can also be observed in the digitization standard ~~diviation~~ deviation before calibration.

3.1 Influence of line thickness

The original ISSJ spectra had an inherent line thickness coming from the pen. The line thickness of the printed line has an influence on the digitization error. Therefore, quantifying the efficiency of the digitization algorithm at different line thicknesses is crucial.

Figure 6 displays the same spectrum (~~zoomed in~~ zoomed-in view) digitized at three different line thicknesses: the same as the old paper spectrum, half, and double the thickness), calibrated and then compared to the original spectrum. The spectrum was produced at a fixed spectral resolution of $0.25cm^{-1}$.

In our analysis, we observed a notable impact of line thickness on the accuracy of spectra digitization, as demonstrated in Figure 6. Specifically, we found that the discrepancy between the original spectrum and its digitized counterpart increases with the thickness of the plot lines. This result can be attributed to an inherent limitation in the digitization process when handling thicker lines. In the case of spectral lines with significant thickness, the digitization algorithm is designed to calculate the mean position of the detected line. However, the actual location of the true spectral line within this thicker plot remains ambiguous.

When a plot line is sharply defined and narrow, the original spectrum position is usually closer to the mean digitized line. Conversely, as the line thickness broadens, the line's precise location becomes increasingly obscured, merging into the overall thickness of the line. This leads to greater uncertainty in determining the exact position of the spectral data during digitization.

235 Therefore, thicker lines introduce an additional layer of uncertainty in this digitization process. The algorithm's reliance on averaging across the line's ~~width~~ thickness means that any deviation from the true line position within the thickness of the line directly contributes to increased digitization error. Consequently, the accuracy of digitization is inversely related to the thickness of the spectral lines, with thicker lines resulting in a higher likelihood of deviation from the true spectral data.

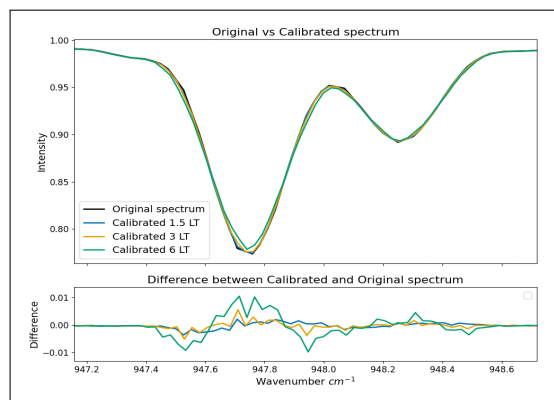



Figure 6. A zoomed in view of an original degraded FTIR spectrum and the redigitized version of it at different line thicknesses (LT). A 3LT corresponds to 25 pixels thickness.

4 Error Study

240 The spectral lines produced by the grating spectrometer have an inherent thickness caused by the use of the recording pen. During the digitization process, the mean of the detected line for a given pixel array was taken. The digitization process of the spectral lines results in an estimated 1.55% error due to the digitization alone.

The 2nd ~~degree~~ degree polynomial proved a better choice for curve fitting the spectra, although having an overall higher χ^2 value than the 3rd ~~degree~~ degree one, it has a lower wavenumber shift STD. Additionally, when calibrating the digitized

245 spectrum to the synthetic wavenumber-mapped spectrum, we need to account for wavenumber fitting shift error of about $\pm 0.065\,\mathrm{cm}^{-1}$. The difference between the detected peak (or ~~minima~~ minimum) of the line and the true peak needs to be quantified as well. The true peak is the ~~center~~ centre of the line calculated from spectroscopic measurements by HITRAN and the error coming from the calibration using peak detection can be calculated as $\epsilon_{peak} = \nu_{true} - \nu_{detected}$. This is done by comparing the line ~~center~~ centre from HITRAN lookup table to the peak detected wavenumber. After fitting the spectra and

250 comparing the wavenumber difference an additional estimated error of $0.01 cm^{-1}$ was calculated.

The error from wavenumber calibration can be expressed as follows $\epsilon_{cm^{-1}} = \sqrt{\epsilon_{peak}^2 + \epsilon_{fit}^2}$. This error needs to be added to the spectroscopic error from the chosen linelist when using these spectra. ~~It was~~ We have shown that the line thickness plays a

role in influencing the error. Where the digitization error doubles when doubling the line thickness.

Additionally, it is worth mentioning that this digitization method only works if the colour of the plot on the paper is different from the rest of the paper (including annotations). Otherwise, in case there are artifacts of the same colour as the spectrum, the digitization will fail. It is then imperative, when possible, to remove any artifacts that might cause the algorithm to misidentify lines.

5 Summary and Conclusion

This paper describes the process of digitizing historical atmospheric spectra from the 1950s, originally recorded at the Jungfraujoch International Scientific Station using a Pfund-Type grating spectrometer. The digitization utilized high-resolution scanning and a colour masking technique in image processing to accurately capture the spectral data. Calibration was achieved using a synthetic spectrum based on HITRAN data, aligning the digitized spectra with the correct wavenumber range at an estimated digitization error of about 1.55% and a wavenumber calibration STD of about $0.075cm^{-1}$. The study revealed a relationship between the plotting line thickness and the corresponding digitization error, which unsurprisingly increased with increasing line thickness.

In conclusion, the successful digitization and calibration of these historical spectra have preserved valuable scientific data, facilitating future atmospheric research and comparisons with modern datasets. This work will hopefully contribute to the field of atmospheric science and potentially other relevant fields.

Data records

The data records are saved as individual spectra (ranging from $520cm^{-1}$ to $3565cm^{-1}$), each containing relevant information for data analyses. Each spectrum has a header that contains the minimum and maximum wavenumber, the number of points, and the resolution, among other details (SZA, apodization type etc...). Accompanying the spectra is a housekeeping file that contains additional data. To facilitate the visualization of the digitized/calibrated spectra, a web portal was created using the Python Flask framework (Grinberg, 2018). The produced spectra can be visualized on the web portal https://iup.uni-bremen.issj.spectra.makkor.de . The files are saved in plain text format, making them easily accessible to users. Additionally, the solar zenith angle (SZA) and airmass ranges are saved as a text array.

Code availability. A snapshot of the digitization and calibration code can be downloaded from the following source https://zenodo.org/records/11204115. The web portal code can also be accessed from the same source under https://zenodo.org/records/11058350 and can be run using Python. The software is published under the GNU public license.

Data availability. The data is freely available at (Makkor et al., 2024) under the download link https://zenodo.org/records/14537672

Author contributions. M.B provided the base code for the calibration/digitization modeling and E.M gave us access to the paper spectra. J.N, M.P.C, M.P and E.M supervised this work. J.M further developed the Digitization/Calibration software, scanned, digitized, and calibrated the spectra.

Competing interests. At least one of the (co-)authors is a member of the editorial board of Atmospheric Measurement Techniques.

285   *Acknowledgements.* We are grateful to the staff of the University of ~~Liege~~ Liège who have kept these historical spectra in relatively good condition throughout the years. The simulated spectrum was produced using SFIT4, which is one of the first advanced atmospheric retrieval algorithms developed through the collaboration between the University of Colorado and the University of Bremen. The software utilizes a line-list generated by HITRAN. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within the projects NO 404/24-1 and PA 1714/8-1. EM is a senior research associate with F.R.S. - FNRS (Brussels, Belgium). His 6-month
290   research stay at Universität Bremen was supported by a grant from the University of Liège.

References

EPSON SCANNER WORKFORCE DS-50000/60000/70000, https://download.epson.com.sg/product_brochures/scanner/EPC/Epson%20WorkForce%20DS-50K_60K_70K(Dec%202019)NoAddress.pdf?t=4.

The Jungfraujoch Scientific Station, Nature, 128, 817–820, https://doi.org/10.1038/128817a0, 1931.

Ehhalt, D. H., Zander, R. J., and Lamontagne, R. A.: On the temporal increase of tropospheric CH4., Journal of Geophysical Research, 88, 8442–8446, https://doi.org/10.1029/JC088IC13P08442, 1983.

GIMP: The GIMP Development Team, https://www.gimp.org, 2019.

Gordon, I., Rothman, L., Hill, C., Kochanov, R., Tan, Y., Bernath, P., Birk, M., Boudon, V., Campargue, A., Chance, K., Drouin, B., Flaud, J.-M., Gamache, R., Hodges, J., Jacquemart, D., Perevalov, V., Perrin, A., Shine, K., Smith, M.-A., Tennyson, J., Toon, G., Tran, H., Tyuterev, V., Barbe, A., Császár, A., Devi, V., Furtenbacher, T., Harrison, J., Hartmann, J.-M., Jolly, A., Johnson, T., Karman, T., Kleiner, I., Kyuberis, A., Loos, J., Lyulin, O., Massie, S., Mikhailenko, S., Moazzen-Ahmadi, N., Müller, H., Naumenko, O., Nikitin, A., Polyansky, O., Rey, M., Rotger, M., Sharpe, S., Sung, K., Starikova, E., Tashkun, S., Auwera, J. V., Wagner, G., Wilzewski, J., Wcisło, P., Yu, S., and Zak, E.: The HITRAN2016 molecular spectroscopic database, Journal of Quantitative Spectroscopy and Radiative Transfer, 203, 3–69, https://doi.org/https://doi.org/10.1016/j.jqsrt.2017.06.038, hITRAN2016 Special Issue, 2017.

Griffiths, P. R.: Resolution and Instrument Line Shape Function, https://doi.org/10.1002/0470027320.s0111, 2002.

Grinberg, M.: Flask web development: developing web applications with python, " O'Reilly Media, Inc.", 2018.

Hugin: The HUGIN Development Team, https://hugin.sourceforge.io, 2020.

Makkor, J., Palm, M., Buschmann, M., Mahieu, E., Chipperfield, M., and Notholt, J.: Digitized paper baed grating spectra from 1950/51., https://doi.org/10.5281/zenodo.14537672, 2024.

Migeotte, M., Neven, L., and Swensson, J.: The Solar Spectrum from 2.8 to 23.7 Microns, Part I, Photometric Atlas, Mem. Soc. Roy. Sci, Liege, Special Volume 1, 1956.

Migeotte, M. V. and Neven, L.: Détection du monoxyde de carbone dans l'atmosphère terrestre, à 3580 mètres d'altitude, Physica D: Non-linear Phenomena, 16, 423–424, https://api.semanticscholar.org/CorpusID:122714199, 1950.

Miller, C. H. and Thompson, H. W.: Vibration-Rotation Bands of Allene, Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences, Vol. 200, No. 1060, https://www.jstor.org/stable/98385, 1949.

Murcray, D. G., Goldman, A., Bradford, C. M., Cook, G. R., Allen, J. W. V., Bonomo, F. S., and Murcray, F. H.: Identification of the v2 vibration-rotation band of ammonia in ground level solar spectra, Geophysical Research Letters, 5, 527–530, https://doi.org/10.1029/GL005I006P00527, 1978.

OPUS: Bruker, "OPUS spectroscopy software," 2017,, https://www.bruker.com/products/infrared-near-infrared-and-raman-spectroscopy/opus-spectroscopy-software.html., 2017.

Prignon, M., S, C., D, M., S, O., C, S., and G, S.: Improved FTIR retrieval strategy for HCFC-22 (CHClF2), comparisons with in situ and satellite datasets with the support of models, and determination of its long-term trend above Jungfraujoch., 19:12309–24, https://doi.org/10.5194/acp-19-12309-2019., 2019.

Zander, R., Demoulin, P., Ehhalt, D. H., Schmidt, U., and Rinsland, C. P.: Secular increase of the total vertical column abundance of carbon monoxide above central Europe since 1950, Journal of Geophysical Research: Atmospheres, 94, 11 021–11 028, https://doi.org/10.1029/JD094ID08P11021, 1989.

Zander, R., Ehhalt, D. H., Rinsland, C. P., Schmidt, U., Mahieu, E., Rudolph, J., Demoulin, P., Roland, G., Delbouille, L., and Sauval, A. J.: Secular trend and seasonal variability of the column abundance of NeO above the Jungfraujoch station determined from IR solar spectra, JOURNAL OF GEOPHYSICAL RESEARCH, 99, 745–761, https://doi.org/10.1029/94JD01030, 1994.

330 Zander, R., Mahieu, E., Demoulin, P., Duchatelet, P., Roland, G., Servais, C., Mazière, M. D., Reimann, S., and Rinsland, C. P.: Our changing atmosphere: Evidence based on long-term infrared solar observations at the Jungfraujoch since 1950, Science of The Total Environment, 391, 184–195, https://doi.org/10.1016/J.SCITOTENV.2007.10.018, 2008.

Zelinsky, A.: Learning OpenCV—Computer Vision with the OpenCV Library, IEEE Robotics and Automation, Magazine 16, no. 3, 100, https://doi.org/https://doi.org/10.1109/MRA.2009.933612., 2009.