Data rescue of historical wind observations in Sweden since the 1920s

John Erik Engström et al.

Open discussion – review.

RC1: 'Comment on essd-2023-2', Simon Noone, 15 Feb 2023

This paper has created the most comprehensive long-term historical monthly climate dataset using Guidelines on Best Practices for Climate Data Rescue. This quality-controlled dataset is an invaluable resource to the scientific community and I see that it has been made openly available on the Zenodo data repository and the SMHI data web portal. The dataset and corresponding information is also openly shared and will be very useful for climate reconstructions and historical wind analysis. I would like to commend the authors on their hard work in producing this dataset. I would most certainly recommend this article for publication.

Comments from referees:

Line 235: For the future work you could look at engaging with local universities to introduce data rescue into the classroom and encourage undergraduate students to get involved in digitising small amounts of data but "many hands make light work". Students can digitise large volumes of meteorological observations quickly and effectively. We have just successful integrated a project into the 2nd Undergraduates Geography curriculum at Maynooth University Ireland. Students helped digitise some unique historical Sub-Saharan African climate data from imaged sheets. Ryan et al 2021 also previously ran a successful data rescue project with undergraduate students.

Author response:

Line 235: Yes it is an interesting method to engage the public and students in citizen science project to get help with digitization. We have not tried that yet at SMHI but we follow other such initiatives and we discuss this opportunity.

Author's changes in manuscript:

None

Comments from referees:

Line 105: How many of these stations were able to be joined to contemporary station records providing longer extended records? Was this not included in your criteria for digitisation? Also did the archive sheets contain any other variables? If so are there any plans to digitise these data?

Author response:

It is common that we join stations to get longer time series. In this project the focus time period for digitization was around 1920-1940, and during that time period joining of stations was not needed. In the time period from 1940 until now stations were joined in some cases. Of the thirteen stations that were digitized, ten stations are still active. The Holmögadd station have a coupling established for temperature and precipitation, which

would be suitable also for wind observations. The stations Väderöbod and Torslanda have not prior coupling established so that will be investigated more in coming projects. The archive sheets did contain other variables that will be digitized according to the priority of stations and variables outside this project.

Author's changes in manuscript:

First, an inventory of suitable and available observations in Sweden on wind speed and direction was performed in the archives of SMHI. Only stations with wind speed data in meter per second (m s^{-1}) were selected. These stations were considered to have been equipped with anemometers so the wind speed could be expected to have been measured. At stations where wind speed was only reported in the Beaufort scale, the wind speed was expected to have been estimated from signs in the surrounding environment due to the information from the journals and the wind series were discarded for the digitization. This resulted in a list of 13 stations across Sweden (Fig. 5 and Table 2). The stations were located in different parts of the country, mostly along the coast. Although it was not a criterion for selection, ten of the stations are still active making analyses over a full century possible. The archive sheets did contain other variables that will be digitized according to the priority of stations and variables in coming years.

Comments from referees:

Line 195: Do you have any metrics on the scale of qc issues found in the digitised data? What percentage of missing data was present in each station after qc was conducted?

Author response:

During digitization the data was directly stored in our database and during QC the values were corrected in the database. During the time period of digitization 1920 to 1940, totally 242364 values were digitized. Out of these 242364, values 392 values were corrected during the QC. The ratio of missing monthly values from 1920 to now was 9%, 975 out of 11222 values.

Author's changes in manuscript:

The performed quality control is, however, not complete. When starting to analyze data in more detail, additional suspicious values will probably appear. Then there is a possibility to check the original value in the scanned paper journal. If the value in the database turns out to be wrong then the database can be updated with the correct value and the quality of the analyzis will improve. During the time period of digitization 1920 to 1940, totally 242364 values were digitized. Out of these 242364 values, 392 values were corrected during the QC. For the ten stations with century-long records the ratio of missing monthly values from 1920 to now was 9%, 975 out of 11222 values.

RC2: 'Comment on essd-2023-2', Simon Noone, 15 Feb 2023

Comments from referees:

correction
first line of review should read; "This paper has created a long-term historical wind dataset using Guidelines on Best Practices for Climate Data Rescue"

Author response:

Okey

Author's changes in manuscript:

None

RC3: 'Comment on essd-2023-2', Anonymous Referee #2, 26 Feb 2023

This work has added new and valuable digitized instrumental measurements of wind speed and direction covering 1920s-1940s, to some extent, many observations only stored in paper journals have been rescued. It is meaningful for the researchers to well understand wind changes and multidecadal variability enveloping one full century. And this work has also made some contributions to the international data rescue. However, there are still some shortcomings in this paper need to be improved.

Specific comments:

Comments from referees:

L183: It is recommended to delete "section 4.2.1 Scanning and digitization" or incorporate this section into section 3.2 or 3.3, simply describe the contents of 4.2.1 in section 3.2 or 3.3.

Author response:

Yes we will consider your recommendation.

Author's changes in manuscript:

3.2 Scanning

Each book lookup was scanned in a dedicated Bookeye 4 scanner (Fig. 7). Each image of a book lookup was compiled to a pdf file containing one year of observations. The scanner could be operated both by hand using buttons or a computer mouse, or by foot using a pedal. The scanning had to be done in the institute while the original journals were not allowed to leave the building. The paper journals were mostly in good shape and were handled with care. In a few cases, however, data were missing in the paper journals or the handwriting was difficult to read.

3.3 Digitization

A template for the digitization was designed and then used in the project (Fig. 8). The scanned image of the paper journal was viewed on one computer screen and the values from the paper journals were typed into the template on another computer screen. When the values were typed into the template, they were automatically stored in the observational MORA database at SMHI. The scanning and digitization were static work and the staff performing these duties were encouraged to often take short breaks not to get stuck in one working position for too long.

4.2.1 ~~Scanning and digitization~~

~~The corona pandemic has affected the whole world in many different ways. The staff at SMHI had to work from home as much as possible, which posed a challenge as the paper journals had to be handled physically at SMHI. The solution we chose was to scan the paper journals at the institute and then do the digitization at home using the electronic image files. The scanning and digitization were static work and the staff performing these duties were encouraged to often take short breaks not to get stuck in one working position for too long. A few times the scanning had to be redone several times to get acceptable quality. It was very good to have a pedal control for the scanner to free the hands to shift pages in the paper journals during the scanning. The paper journals were mostly in good shape and were handled with care. In a few cases, however, data were missing in the paper journals or the handwriting was difficult to read.~~

Comments from referees:

In the whole paper, the section of homogenizing wind data has not been introduced, and some key techniques such as establishing reference series, detecting and adjusting shifts are also missing. The data connection provided in this paper is not enough to support the complete data processing for wind observations.

Author response:

This project was divided into three work packages. Digitization, homogenisation and evaluation and attribution of trends. This paper describes work package 1, digitization. The other two work packages, homogenisation and evaluation of trends are presented in two other publications:

Work package 2, homogenization:

Zhou, C., Azorin-Molina, C., Engstrom, E., Minola, L., Wern, L., Hellstrom, S., Lonn, J., and Chen, D.: HomogWS-se: A century-long homogenized dataset of near-surface wind speed observations since 1925 rescued in Sweden, Earth System Science Data Discussions, 2022, 1–24, https://doi.org/10.5194/essd-2022-29

Work package 3, Evaluation and attribution of trends:
The contribution of large-scale atmospheric circulation to variations of observed near-surface wind speed across Sweden since 1926
Lorenzo Minola et al. https://www.researchsquare.com/article/rs-2255253/v1

Author's changes in manuscript:

Change in end of section “Introduction”:

SMHI holds valuable wind records only available in paper journals prior to the 1940s as the digitization of climate series at SMHI systematically started only in the 1950s–1960s. To rescue historical wind observations in Sweden from the 1920s to 1940s, the project "Assessing centennial wind speed variability from a historical weather data rescue project in Sweden" (WINDGUST) was initiated, aimed at filling the gap of data prior to 1939 and temporal inhomogeneity of wind datasets in Sweden, with the ultimate goal of bringing more wind data into the hands of scientists who seek to improve the limited knowledge on the causes of wind variability and change including that of “stilling” and “reversal” as discussed in global climate science.

Here we focus the results, challenges and lessons learned from the first part on data rescue of the WINDGUST project. Methods and results of the homogenisation are

presented by and Zhou et al. (2022) and the evaluation and attribution of trends are presented by Minola et al. (Submitted).

Comments from referees:

The article is to support the publication of the wind dataset, so it should display some outstanding results in the processing of quality control and homogenization, rather than a reference Zhou et al. (2022) mentioned in “section 7 conclusions” (L240) can represent. It is recommended to analyze some contents which can be presented by tables or figures: (1) number of suspected error data before and after quality control, cause statistics, and how to deal with them, etc. ; (2) the series of wind observations raw and adjusted; (3) assessment of the climate characteristics of the newly wind data, etc.

Author response:

The analysis of the digitized wind data is presented in the section 4.1 Data screening and additional figures in the appendix. During the time period of digitization 1920 to 1940 totally 242364 subdaily values were digitized. Out of these 242364 values 392 values were corrected during the QC. The ratio of missing monthly values from 1920 to now was 9%, 975 out of 11222 values. We will add this statistic to the manuscript.

Author's changes in manuscript:

The performed quality control is, however, not complete. When starting to analyze data in more detail, additional suspicious values will probably appear. Then there is a possibility to check the original value in the scanned paper journal. If the value in the database turns out to be wrong then the database can be updated with the correct value and the quality of the analyzis will improve. During the time period of digitization 1920 to 1940, totally 242364 values were digitized. Out of these 242364 values, 392 values were corrected during the QC. The ratio of missing monthly values from 1920 to now was 9%, 975 out of 11222 values.

Comments from referees:

L219: It is recommended to add some explanations on the shortcomings and improvements in this work in “section 5 discussion”.

Author response:

Yes we will add some discussion on the shortcomings and improvements in this work in “section 5 discussion”.

Author's changes in manuscript:

5 Discussion

The scanning of the original journals enables an additional electronic backup of the data and improve the accessibility when the electronic files can be sent to external users while the original journals have to be kept in the archive. The electronic copies also made it possible to perform the digitization at any place suitable for the staff doing the digitization. It would be beneficial if the work can continue with digitization of more stations further back in time. This work provides a significant addition to the available wind observations for Sweden, but thirteen stations is far from enough to be representative for the whole country of Sweden. One possibility could be to digitize wind

speed observations recorded with the different optical wind scales previously used, which would demand a profound quality evaluation to determine their usability.

This work contributes to satisfy the need as stated by (WMO, 2016) to increase the amount of digitized wind observations available for the international research community and enable additional insight into the historical variability and trends of the wind climate in Sweden and globally as highlighted by (IPCC, 2021). There is still a vast amount of meteorological data that are only stored in paper journals that need to be digitized to be rescued for future generations and made practically accessible and usable for fellow researchers. This urges the national meteorological services to continue their efforts to rescue and digitize meteorological data. In this project we acknowledge the benefit of having experienced digitization staff that work in close collaboration with climate experts to identify and correct errors in the digitization. The International Data Rescue (I-DARE; idare-portal.org; last accessed 26 December 2022) portal collects and provides information on a large number of data rescue activities and facilitates sharing of experience and methods in this field. Some projects have successfully tried to engage the public in citizen science contributions where people could volunteer to digitize data from scanned historical records (Hawkins et al. (2019); Craig and Hawkins (2020)) using the web-service of weatherrescue.org: last accessed 26 December 2022, where the quality control was conducted by triple parallel digitization of three independent persons. To further accelerate the pace in rescuing and digitization of historical data the techniques of artificial intelligence (AI), machine learning (ML), and optical character recognition (OCR) could be developed and applied in this field as discussed by Campbell et al. (2021). These techniques have so far been of limited value due to the low ratio of accuracy but the methods are improving fast. In the near future, there will probably be a growing number of successful projects rescuing data by these and similar methods.

Comments from referees:

L236: It is recommended to incorporate “section 6 outlook” into “section 5 discussion”.

Author response:

Yes we will incorporate section 6 Outlook into section 5 Discussion.

Author's changes in manuscript:

5 Discussion

The scanning of the original journals enables an additional electronic backup of the data and improve the accessibility when the electronic files can be sent to external users while the original journals have to be kept in the archive. The electronic copies also made it possible to perform the digitization at any place suitable for the staff doing the digitization. It would be beneficial if the work can continue with digitization of more stations further back in time. This work provides a significant addition to the available wind observations for Sweden, but thirteen stations is far from enough to be representative for the whole country of Sweden. One possibility could be to digitize wind speed observations recorded with the different optical wind scales previously used, which would demand a profound quality evaluation to determine their usability.

This work contributes to satisfy the need as stated by (WMO, 2016) to increase the amount of digitized wind observations available for the international research community and enable additional insight into the historical variability and trends of the wind climate in Sweden and globally as highlighted by (IPCC, 2021). There is still a vast amount of meteorological data that are only stored in paper journals that need to be digitized to be rescued for future generations and made practically accessible and usable for fellow

researchers. This urges the national meteorological services to continue their efforts to rescue and digitize meteorological data. In this project we acknowledge the benefit of having experienced digitization staff that work in close collaboration with climate experts to identify and correct errors in the digitization. The International Data Rescue (I-DARE; idare-portal.org; last accessed 26 December 2022) portal collects and provides information on a large number of data rescue activities and facilitates sharing of experience and methods in this field. Some projects have successfully tried to engage the public in citizen science contributions where people could volunteer to digitize data from scanned historical records (Hawkins et al. (2019); Craig and Hawkins (2020)) using the web-service of weatherrescue.org: last accessed 26 December 2022, where the quality control was conducted by triple parallel digitization of three independent persons. To further accelerate the pace in rescuing and digitization of historical data the techniques of artificial intelligence (AI), machine learning (ML), and optical character recognition (OCR) could be developed and applied in this field as discussed by Campbell et al. (2021). These techniques have so far been of limited value due to the low ratio of accuracy but the methods are improving fast. In the near future, there will probably be a growing number of successful projects rescuing data by these and similar methods.

At SMHI the work continues to rescue and digitize historical meteorological, hydrological and oceanographic observations from the archives of paper journals spanning over more than 150 years and the methods discussed above are developed to facilitate and speed up this work further.

~~6 Outlook~~

~~At SMHI the work to rescue and digitize historical meteorological, hydrological and oceanographic observations from the archives of paper journals spanning over more than 150 years continues and the methods discussed above are developed to facilitate and speed up this work further.~~

Comments from referees:

Minor comments:

Please adjust or correct the grammar or format in the following sentences.

1. L46: "stilling" and "reversal" periods is due is the low availability of centennial wind series.
   L46: "stilling" and "reversal" periods is due to the low availability of centennial wind series.

1. L82: This created a pressure drop in the tube moving a pointer that showed the wind speed on a scale.
   The wind created a pressure drop in the tube that moved a pointer which showed the wind speed on a scale.

1. L160-L162: Observations have been continuously registered since at Malmslätt since 1945. Before that storing of journals were more random which caused the data gap from July 1936 to January 1939.
   Observations have been continuously registered ~~since~~ at Malmslätt since 1945.

Before that storing of journals were more random which caused the data gap from July 1936 to January 1939.

1. L165-L166: The data gap in the 1940s is due to observations that by mistake were not digitized in the project.
   The data gap in the 1940s was due to observations that by mistake were not digitized in the project.

1. L220: This work contributes to satisfy the need as stated by (WMO, 2016) to increase.
   This work contributes to satisfy the need to increase the amount of digitized wind observations available for the international research community, as stated by (WMO, 2016).

1. L222: the wind climate in Sweden and globally as highlighted by (IPCC, 2021).
   The digitized observations enable additional insight into the historical variability and trends of the wind climate in Sweden.

1. L237-L238: At SMHI the work to rescue and digitize historical meteorological, hydrological and oceanographic observations from the archives of paper journals spanning over more than 150 years continues and.
   At SMHI the work continues to rescue and digitize historical meteorological, hydrological and oceanographic observations from the archives of paper journals spanning over more than 150 years ~~continues~~ and the methods discussed above are developed to facilitate and speed up this work further.

Author response:

Also thank you for the comments on grammar and format.