Seven decades of neutron monitors (1951–2019): Overview and evaluation of data sources

Pauli Kalervo Väisänen[1], Ilya Usoskin[1], and Kalevi Mursula[1]

[1]University of Oulu

November 24, 2022

Abstract

The worldwide network of neutron monitors (NMs) is the primary instrument to study cosmic-ray variability on time scales of up to 70 years. Since the 1950s, 147 NMs with publicly available data have been in operation, and their records are archived in and distributed through different repositories and data sources. A comprehensive analysis of all available NM datasets (300 datasets from 147 NMs) is performed here to check the quality and consistency of the data. The data sources include World Data Center for Cosmic Rays (WDCCR), the Neutron Monitor Database (NMDB), the Pushkov Institute of Terrestrial Magnetism, Ionosphere and Radiowave Propagation (IZMIRAN) and individual station/institution databases. It was found that The data from the same NM can be non-identical and of different quality in different sources. We give and tabulate here a recommendation for the optimal data source of each NM. We also present here a list of 29 'prime' stations with the longest and most reliable data. Verified datasets for these prime stations are provided as supplementary information.

Seven decades of neutron monitors (1951 – 2019): Overview and evaluation of data sources

Pauli Väisänen[1], Ilya Usoskin[1,2], Kalevi Mursula[1]

[1]Space Physics and Astronomy Research Unit, University of Oulu, Finland
[2]Sodankylä Geophysical Observatory, University of Oulu, Finland

Key Points:

- The quality of hourly data of almost 300 datasets for 147 neutron monitors was assessed.
- Individual neutron monitor datasets across multiple sources were cross-compared and the best source(s) for each monitor were determined.
- An up-to-date assessment of all available neutron monitor data was conducted.

Corresponding author: Pauli Väisänen, `pauli.vaisanen@oulu.fi`

Abstract

The worldwide network of neutron monitors (NMs) is the primary instrument to study cosmic-ray variability on time scales of up to 70 years. Since the 1950s, 147 NMs with publicly available data have been in operation, and their records are archived in and distributed through different repositories and data sources. A comprehensive analysis of all available NM datasets (300 datasets from 147 NMs) is performed here to check the quality and consistency of the data. The data sources include World Data Center for Cosmic Rays (WDCCR), the Neutron Monitor Database (NMDB), the Pushkov Institute of Terrestrial Magnetism, Ionosphere and Radiowave Propagation (IZMIRAN) and individual station/institution databases. It was found that The data from the same NM can be non-identical and of different quality in different sources. We give and tabulate here a recommendation for the optimal data source of each NM. We also present here a list of 29 'prime' stations with the longest and most reliable data. Verified datasets for these prime stations are provided as supplementary information.

1 Introduction

Neutron monitors (NMs) are ground-based particle detectors, which detect secondary nucleons produced locally in the atmosphere as a product of cascades initiated by primary cosmic-ray particles (Simpson, 2000; Belov, 2000). The flux of cosmic rays varies as modulated by solar magnetic activity, and this variability is continuously monitored by NM count rates. Natural sources for changes in NM count rates include the varying cosmic-ray flux in near-Earth space (heliospheric modulation by the solar wind and heliospheric magnetic field; solar particle events), geomagnetic shielding (geomagnetic rigidity cutoff at the NM location), atmospheric parameters affecting the development of the cascade (altitude or barometric pressure; weather conditions, e.g. snow), and instrumental changes (technical characteristics of the detector, e.g., electronic setup, number of counters, registration efficiency, local surroundings). In order to study cosmic-ray modulation in solar variability, NM data are corrected for the terrestrial (geomagnetic, atmospheric and instrumental) effects as a standard procedure. Here, we will analyze pressure and efficiency (whenever possible) corrected data unless specified differently.

The NM measurements started in 1951 with the Climax NM (USA) and later developed to a global network (Moraal et al., 2000), thus covering nearly 70 years and producing a unique long dataset in the field of solar-terrestrial physics.

Data from the global NM network have been collected in different repositories and databases that offer the data freely online. However, these repositories often employ different data practices and may contain different versions or only a fraction of the full data. Effectively, this means that data from different repositories may not be congruent with each other, leading to differences when comparing or reproducing the results. This in turn makes the results of analyses of such data-dependent on the exact source. A special question is related to the instrumental stability of long-operating NMs with multi-decadal lifetimes. This issue was studied by Gil et al. (2015) and Usoskin et al. (2011, 2017), but the dependence on the exact data source was not evaluated there.

In this paper, we analyze the history and the current global status of publicly available NM data. Using an automated data collection and analysis system, we obtain, study and cross-compare datasets from different NMs and sources to produce an up-to-date assessment of the NM datasets and reliable recommendations for their usage, with the aim to assist NM data users to produce more reliable and reproducible results.

This paper is organized as follows. In Section 2, we present a brief history of the NM network and NM data practices. Section 3 gives an overview of the NM data repositories, common practices, problems and limitations. Selection of the prime stations and their assessment are presented in Section 4. Section 5 gives our recommendations for future improvements of the NM data archiving. Conclusions are summarized in Section 6.

2 Brief history of neutron monitors as space-physics instruments

NMs were invented by Simpson (1948) as a detector to register and study the secondary neutron particles generated by cosmic rays. The Climax NM started operating in 1951, whereas many other NM stations were launched during the International Geophysical Year (IGY) in 1957. These early NMs are therefore referred to as "IGY" type. Based on the collected experience, the design was improved, and a new type of detector, called NM64 or "super-monitor", was introduced during the International Quiet Sun Year (IQSY) of 1964. This design was so good (Hatton & Carimichael, 1964) with stable operation and robust data production, that it remains a standard design since then, and the number of NM64's operating around the globe reached many dozens. It should be noticed that the standard NM64 design (Hatton, 1971) was initially based on the BF_3-filled proportional counters BP-28 produced by the Chalk River laboratory in Canada and their Soviet analog SNM-15 used in USSR and Eastern Europe. The latter are about

15% less effective than NM64 (Abunin et al., 2011; Gil et al., 2015) because of the less pure filling gas. Later, there was a tendency to use ^{3}He-filled counters but, because of high pressure and leaking ability of helium, they appeared unstable in the long run. At present, BF_3-filled proportional counters of slightly improved design (higher gas pressure) are used again (Strauss et al., 2020).

The data obtained from individual NMs are traditionally collected by the World Data Center for Cosmic Rays (WDCCR) which was established during the International Geophysical Year of 1957 at IZMIRAN (Pushkov Institute of Terrestrial Magnetism, Ionosphere and Radiowave Propagation), USSR and RIKEN, Japan. Through WDCCR, data were exchanged between the Soviet Union, the USA and Japan. WDCCR is currently maintained by Nagoya University, Japan, and is mirrored at IZMIRAN. It offers historical datasets, provided as a set of ASCII data-files in several formats, through an online FTP-service that is updated on a monthly basis. WDCCR stores data from many old and short-lived stations that cannot be found anywhere else. IZMIRAN not only maintains a mirror of the WDCCR dataset but also continuously develops its own database by collecting data and implementing apparent corrections to the raw data.

The first real-time data available service online was provided by the Moscow NM station in 1997. In 2000, Oulu NM launched an online database, the first in Western countries. Since then, several NM stations started their own data service, each in its own style. A decade later, in 2008, the Neutron Monitor Database (NMDB) project started under the EU FP7 program, providing an accessible database of archival and real-time verified data from about 50 monitors. It started as a European project but currently includes NMs from around the globe.

Many active NM stations also offer data through their own web services or other systems. These also include stations and research institutes that manage and distribute data from multiple stations, as will be discussed below.

3 Data and methods

In this work, we collected all available NM count-rate data from all the repositories, databases and individual NM homepages. We have identified 147 NM stations whose data are available in any of the main sources of data listed in Table 1. The station list is provided in Table 2 and in the Supplementary Information.

We developed an automated system for fetching online NM 1-hour resolution data from all the sources of Table 1 up to the end of the year 2019. Each dataset was then parsed and transformed into the Matlab data format. Thus, a dataset of hourly NM count rates was created for further analysis. All data were downloaded during 20 − 23 June 2020.

A brief description of the data repositories is provided in the following subsections.

3.1 WDCCR

The World Data Center for Cosmic Rays (WDCCR) started its operation in 1957 (Lincoln & Shea, 1973). It collects pressure-corrected data from NM stations and makes them available online as ASCII files of 1-hour time resolution, through an FTP service. There are 140 sub-folders for NM data in the FTP folder, but two of them (Bergen & Cape_H) are empty. Metadata is provided in each file, and changes, e.g., the number of counters, can be traced in the metadata. The data in WDCCR are typically from the time of their recording, while revisions/corrections/updates of the already written data are not foreseen.

The data for this study was collected from the WDCCR repository `http://cidas.isee.nagoya-u.ac.jp/WDCCR/`

Table 1. Summary of data repositories and number of recommended data sources.

Data repository	Available stations	# of recommended sources	# of secondary sources
NMDB (1h)	53	29	10
NMDB (revori)	51	3	2
WDCCR	138	59	24
IZMIRAN	81	50	18
Polar Geophys. Inst.	1	1	
Bartol Inst.	8	5	3
Jungfraujoch NM	2	0	2
Lomnicky Stit NM	1	1	
Mexico NM	1	0	1
Oulu NM	3	3	
South African stations	5	2	2
Yakutsk+Tixie Bay	2	0	0

3.1.1 Data format

WDCCR offers data in three formats: LONGFORMAT, SHORTFORMAT and CARDFORMAT, described in the WDCCR homepage under "Data Formats". All the formats contain the same data in yearly ASCII files, which are different only in presentation. The long format displays monthly values in 12 lines, with relevant metadata at the start of each line. The short format displays the same data, but the monthly metadata is more thoroughly described, and the count rates are displayed with 12 hourly values on each line. The card format is similar to the short format in the form of displaying data but does not contain metadata beyond the basic station descriptors (NM name, type, pressure corrections etc.) at the start of each line. For this study, we use data in the LONGFORMAT.

3.1.2 Scaling factors

Count rates in WDCCR are provided as unscaled values (DATA), with a Scaling Factor (SF) and a Constant (CONST) provided in the metadata. The real count rates are defined as:

$$\mathrm{Real\,Counts} = (\mathrm{DATA} + \mathrm{CONST}) \cdot \mathrm{SF}.$$

However, these scaling factors do not always correct such apparent problems as jumps related to the changing number of counters, their malfunctioning, change of type, etc. The scaling factors and their source or methodology are not described in any way. Such apparent jumps need to be analyzed and corrected separately.

3.2 NMDB

The Neutron Monitor Database (NMDB) was established in 2008 as a part of a European Union funded project (FP7 Programme) to create a modern database of NM data, including real-time updates (Mavromichalaki et al., 2011). Originally, it was built on mostly European NMs, but data from several non-European stations have been added later. In total, NMDB hosts data from 66 NMs, 8 of which contain no data, leaving 58 stations with data available. Except for Leadville and Polarstern, all NMs listed there have data available from other sources as well.

3.2.1 Data format

NMDB provides data for uncorrected (raw) counts, pressure- and efficiency-corrected count rates and barometric pressure. Here we always use the ‘corrected’ data.

The NMDB contains three data table options for each station: ”ori” ”revori” and ”1hour”, which contain originally loaded data, the revised data in the best time resolution (usually 1 minute), and the 1-hour validated data, respectively. Short descriptions are available at `http://www01.nmdb.eu/nest/help.php#helptable` and `http://www01.nmdb.eu/nest/statements.html`. Status of the currently available data and their version date for different tables can be found at `http://www01.nmdb.eu/status/status.php`. The NDMB-ori dataset cannot be changed after the first load, while all later corrections/modifications are reflected in NMDB-revori (and NMDB-1hr) datasets. Accordingly, the NMDB-revori table supersedes the ori table (i.e. the NMDB-ori table is just the first version of NMDB-revori table). In this analysis we will not discuss -ori and -revori tables separately, and will only analyze the -revori and -1hour tables.

For the NMDB data retrieval, we employed an automated web query method, which downloads and parses the data at 1-hour resolution for each station from both the revori and 1-hour tables in 1-year increments. The queries were split into 1-year increments since the NMDB system automatically decreases the resolution (e.g. from 1-hour to 1-month time resolution) for too long queries. Finally, the data subsets were compiled into a single matrix for the subsequent analysis.

The web query method utilizes the following url when fetching the data: `http://www01.nmdb.eu/nest/draw_graph.php?formchk=1&stations[]=',*StationAcronym*,tabchoice=',*NMDBtable*,'&dtype=corr_for_efficiency&tresolution=60&force=1&yunits=0&date_choice=bydate&start_day=1&start_month=1&start_year=',*StartYear*,'&start_hour=0&start_min=0&end_day=31&end_month=12&end_year=',*EndYear*,end_hour=24&end_min=00&output=ascii`, where **StationAcronym** is the acronym associated with the specific station, **NMDBtable** is the selected data table, **StartYear** is year for which to collect data and **EndYear** is **StartYear+1**.

3.3 IZMIRAN

The Pushkov Institute of Terrestrial Magnetism, Ionosphere and Radiowave Propagation (IZMIRAN) of the Russian Academy of Sciences was established in 1939 by Nikolay Pushkov. The IZMIRAN database offers data for most Russian (former Soviet) NM stations, but it also offers data from other NM stations. Altogether, IZMIRAN provides data from 82 NMs (Belov et al., 1998). Only one of these does not contain any data (Putre), leaving 81 stations with available data.

The database does not simply copy data from original sources, but apparently applies an automated procedure of validation and correction of the raw data. However, the procedure is not documented nor traceable and may distort the data. We have found, e.g., that outliers of unknown origin occasionally appear in otherwise good data.

3.3.1 Data format

The IZMIRAN database is located at `http://cr0.IZMIRAN.ru/common/links.htm`. The IZMIRAN data is available through the ”iDB”-button next to each station. There are options for pressure-corrected data, barometric pressure data and non-pressure-corrected data. The queried data only includes timestamps and the data values. Empty values are denoted by 0.

The pressure-corrected data for the full analysis period were downloaded on 22-Jun-2020 using the following web query: `http://cr0.IZMIRAN.ru/scripts/nm64queryD`

`.dll/',*StationAcronym*,'?y1=1951&m1=1&d1=1&h1=0&mn1=0&y2=2019&m2=12&d2=31&h2=0&mn2=0&res=1_hour`, where **StationAcronym** is the acronym associated with the specific station.

3.4 NM station homepages

Many NM stations also publicly distribute data through dedicated web-pages, either individual for that NM or institutional, providing a data portal for several NMs operated by the same institution, as briefly described below.

3.4.1 Polar Geophysical Institute

The Polar Geophysical Institute (Murmansk region, Russia) distributes data of Apatity http://pgia.ru/data/nm. There is also an option for Barentsburg NM data, but data retrieval for it did not work for the present analysis.

3.4.2 Bartol

The Bartol Research Institute of the University of Delaware (Newark, USA) operates eight NM stations: McMurdo, Swarthmore/Newark, South Pole, Thule, Fort Smith, Peawanuck, Nain, and Inuvik. Data are distributed through the web-page and FTP at `http://neutronm.bartol.udel.edu/~pyle/bri_table.html`, but the datasets are not updated after 2017.

3.4.3 Jungfraujoch

The Physikalisches Institut of the University of Bern (Switzerland) operates two NMs (one of NM64 and one of IGY type), both located at the Jungfraujoch high-mountain station, for which they distribute data via FTP access at `http://cosray.unibe.ch/`.

3.4.4 Lomnický Štit

The Institute of Experimental Physics of the Slovak Academy of Sciences in Košice (Slovakia) operates the Lomnický Štit NM station and distributes its data through the web-page at `http://neutronmonitor.ta3.sk/`.

3.4.5 Mexico

Data for the Mexico City Cosmic Ray Observatory is available distributes its data through the web-page at `http://www.cosmicrays.unam.mx/`.

3.4.6 Oulu

The Oulu NM started operation in 1964 in the Kontinkangas district and was moved to the Linnanmaa campus where it is still located. The University of Oulu also operates two mini-NMs (a standard DOMC and a bare (lead-free) DOMB) at the Concordia station on the Central Antarctic plateau. The dataset of these stations, which are continuously updated, can be directly accessed through the Oulu NM web-page `http://cosmicrays.oulu.fi` (Usoskin et al., 2001; Poluianov et al., 2015)

3.4.7 South African stations

The Centre for Space Research in the North-West University (NWU) in Potchefstroom (South Africa) operates NMs at five locations: Hermanus, Potchefstroom, Sanae64, Sanae80 and Tsumeb. The data are available as ASCII files at the web-page in `http://natural-sciences.nwu.ac.za/neutron-monitor-data`.

3.4.8 Yakutsk/Tixie Bay

Yu.G. Shafer Institute for Cosmophysical Research and Aeronomy of Russian Academy of Sciences (Yakutsk, Russia) operates two NMs, viz. Yakutsk and Tixie Bay stations, and distributes their data at `https://www.ysn.ru/ipm/`.

3.4.9 Other sources

We also list here a few other possible data sources which we did not use because of some problems reported below.

The data for the Australian NMs at Mawson and Kingston are available through their web page at `http://www.sws.bom.gov.au/World_Data_Centre/1/7` and FTP at `ftp://ftp-out.sws.bom.gov.au/wdc/wdc_cosray/`. However, the website offers only daily files. Moreover, because of a very slow and unstable connection, we were unable to download the entire dataset. Since data from these NMs are available from other sources even at the 1-hour resolution used here, we did not analyze this dataset.

The Tibet/Yang Ba Jing NM has a data distribution web-page at `http://ybjnm.ihep.ac.cn/nm/`, which however, was not working during the preparation of this paper.

3.5 Metadata

Data for each NM station are usually accompanied by metadata either in a station information page or at the header of a data file, which typically includes the following parameters:

Name, typically denoting the geographical name of the location. Historically, because of the limited length for the filename in old data formats, each NM station also has a 4-letter or 6-letter acronyms, which are usually the same for the same station across databases, but can also be different (e.g., McMurdo station is called MCMU and MCMD in NMDB and IZMIRAN databases, respectively). Also, sometimes the same station may have different names through times, as e.g., Swarthmore and Newark NM. We have performed a careful check to make sure that we always refer to the same station even if the names/acronyms are not identical across the databases.

Location includes the geographical latitude, longitude and altitude above sea level.

Geomagnetic cutoff rigidity provides an estimate of the sensitivity of a NM to the energy/rigidity of cosmic rays. It is roughly interpreted so that the primary cosmic-ray particles must possess rigidity higher than the cutoff (Cooke et al., 1991). The cutoff rigidity may slowly change for a fixed geographical location, because of the migration and current weakening of the geomagnetic dipole, but this is not always taken into account in the NM metadata. Sometimes metadata (e.g., the IZMIRAN "see info" page) mentions the rigidity computation year but does not provide the exact model. This information can be used as a rough estimate, but for a detailed long-term analysis, the cutoff rigidity is recommended to be calculated for each location and each given time, rather than being blindly copied from the metadata.

The metadata, including also years of operation are available from the following locations:

NMDB: Station list at `http://www01.nmdb.eu/station/` but it does not reflect possible temporal changes (e.g., changes in rigidity cut-off).

WDCCR: Station Information table at `http://cidas.isee.nagoya-u.ac.jp/WDCCR/station_list.php`, and also in the headers of data files

IZMIRAN: Station info is available under the "see info" button under the specific station "idB" page, or under http://cr0.IZMIRAN.ru/*station*/baseinfo.htm, where **station** is the short acronym of the station.

Station homepages usually also provide metadata.

4 Prime stations

With so many stations, it is difficult to check the stability of any individual NM. In order to have a reliable baseline for data comparison and validation, we have constructed an aggregate based on data from stable long-lived NMs that we here call 'prime' (or 'reference') stations. The selection of the prime stations was based solely on the quality of data, not involving any a-priori or subjective knowledge or preferences, using the following criteria:

1. Times of ground-level enhancements (GLEs) were removed from each dataset of hourly pressure- and efficiency-corrected count rates using the list of the International GLE Database (https://gle.oulu.fi).
2. The data was normalized by the median over two-year interval of years 1995 – 1996 (or 1975 – 1976 if the data for 1995 – 1996 was not available).
3. Outliers were excluded using a 5-point moving median filter which removes points that are more than three median absolute deviations (MAD) from the 5-point median.
4. After the previous steps, stations with less than 20 years of total data coverage were excluded.
5. All datasets were visually checked for apparent steps, drifts or other obvious errors in the data. Some of the errors could be corrected using metadata (e.g., change of the number of counters, or incorrect scaling factor) or using information from other data sources.
6. Datasets, which could not be corrected above, were excluded. To automatically exclude datasets with too large steps or unphysical variation, the following method was applied. Using the knowledge that the natural variability of hourly cosmic-ray data does not exceed $\pm 30\%$ even for polar NMs and is much smaller for lower-latitude stations, we excluded datasets with large steps or drifts by requiring that the max-to-min hourly-value ratio does not exceed two (i.e. the variations from the mean in the dataset do not exceed $\pm 33\%$).
7. In cases with several data sources available for a prime station candidate, the source with the longest data coverage was used.

Using this procedure, we selected 29 prime stations, listed in bold in Table 2. For further analysis we divided them in three groups according to their nominal geomagnetic cutoff rigidity R_c: low- ($R_c \leq 1.75$ GV, 12 NMs), mid- ($1.75 < R_c \leq 2.75$ GV, 5 NMs) and high-rigidity ($R_c > 2.75$ GV, 12 NMs) stations. The temporal variability of these prime stations is shown in the Supplementary Information Figure S4. For the low-rigidity prime NMs, we computed a reference dataset $\mathrm{NM_{low}}$ as the mean of the normalized prime stations with $R_c \leq 1.75$GV, shown by the black curve in the upper panel of Figure S4. The reference dataset for the medium-rigidity stations $\mathrm{NM_{med}}$ was composed in a similar way (Figure S4 middle panel). For the high-rigidity group of NMs, averaging was not done, because of the too wide range of the R_c values, from 2.9 to 11 GV, so that the modulation effects would make the averages to be solar-cycle dependent. This would cause variation around the mean when comparing station data to prime data.

The prime datasets were used to check the data quality of all stations and their different sources. For low- and medium rigidity NMs we compared the data of each individual NM with the corresponding reference datasets NM_{low} and NM_{med}. For the high-rigidity range we compared the individual NM data with the prime station with the nearest rigidity cut-off, or in case of no time overlap, to the second or the third nearest ones. For the comparison, we computed the ratio of the normalized count rates of the analyzed NM to the prime reference dataset.

As an example, we provide a detailed analysis of the mid-rigidity Newark (before 1978 known also as Swarthmore) NM in the supplementary information S1. Newark/Swarthmore has data represented in all the analyzed sources for a long time period and also nicely depicts common characteristics related to the different sources. Similar analyses were made for all stations and all data sources. Basing on the fraction of the good data (and manual inspection of the comparisons), we constructed a list of recommended data sources as described below.

5 Recommendations

The following information on all available NM datasets is given and described in Supplement Table S5 as an Excel-file. This table contains a large amount of information that can be useful for NM data users. The acronyms are helpful when accessing data, since the data retrieval methods usually employ the acronym specific for the database. The number of all hourly data points from each source gives a rough estimate of data coverage. The overall usability of the whole length of data depends on the data quality and potential corrections that can be applied to the data. Latitude, longitude, altitude and geomagnetic cutoff of the stations were collected from the metadata sources, as described in Section 3.6. These values might be not correct in cases where the station has been moved during its operation.

Based on the analysis described in Section 4, we have summarized our recommendations on the data sources for each station in Table 2. More detailed information on the recommended data sources is collected in Supplement Data Set S6, which includes station name, recommended source, secondary source(s) and notes about the data. The 'secondary' (or alternative) sources are nearly equivalent to the primary ones and may contain additional data. Summary statistics of the primary and secondary data source recommendations are presented in Table 1.

The following caveats should be noted. First, the 'data quality' used here as a means for data source selection is only examined relative to individual station: even if a specific source is recommended for the station, it may not correctly describe the general data quality. It only indicates which of the sources is the best according to our criteria. Moreover, the data quality was assessed in late June 2020 and may change later.

Table 2. List of recommended data sources, given as: 1 – Station's website; 2 – IZMIRAN; 3 – WDCCR ; 4 – NMDB1h ; 5 – NMDB1hrevori. Prime stations are in bold.

Ahmedabad	4	Herstmonceux	3	**Newark**	**4**
Albuquerque	3	Hobart	3	**Nobosibirsk**	**2**
Alert	**2**	**Huancayo**	**4**	Nor-Amberd	4
Alma-Ata A	2	**Inuvik**	**2**	Norilsk	2
Alma-Ata B	**4**	Invercargill	3	Northfield	3
Alma-Ata C	2	Irkutsk	2	Ottawa	2
Apatity	1	Irkutsk 2	2	**Oulu**	**1**
Aragats	4	Irkutsk 3	2	Peawanuck	1
Athens	4	Jang Bogo	5	Pic du Midi	2
Bagneres	3	**Jungfraujoch IGY**	**4**	**Potchefstroom**	**1**
Baksan	2	Jungfraujoch NM64	4	Prague	3
Barentsburg	2	Kampala	3	Predigtsthul	3
Beijin	2	**Kerguelen**	**4**	Resolute Bay	3
Beirut	3	Khabarovsk	3	Rio De Janeiro	3
Berkeley	3	**Kiel**	**4**	Rome	2
Brisbane	3	Kiel 2	4	**Sanae64**	**2**
Buenos Aires	3	**Kiev**	**3**	Sanae80	4
Bure	2	Kingston	2	Santiago	2
Calgary	2	Kiruna	3	Seoul	3
CALM	5	Kodaikanal	3	Simferopol	3
Cape Schmidt	2	Kuhlungsborn	3	South Pole	1
Casey	3	Kula	3	South Pole Bare	4
Chacaltaya	3	Lae	3	Sulphur Mt IGY	3
Chicago	2	Larc	2	Sulphur Mt NM64	2
Churchill	2	**Leeds**	**2**	Swarthmore	2
Climax	**4**	Lincoln	3	Sverdlovsk	2
College	3	Lindau_IGY	3	Sydney	3
Cordoba	3	Lindau_NM64	3	Syowa	3
Daejeon	4	**Lomnicky Stit**	**1**	Tashkent	2
Dallas	3	London	3	Tbilisi	2
Darwin	3	Magadan	2	**Terre Adelie**	**4**
Deep River	**2**	Makapuu_Pt	3	Thailand	4
Denver	3	Mawson	2	**Thule**	**4**
Dome B	1	**McMurdo**	**1**	Tibet	4
Dome C	1	Mexico	3	Tixie Bay	2
Dourbes	4	Mina Aguilar	3	**Tokyo**	**2**
Durham	**2**	Mirny	4	**Tsumeb**	**4**
Ellsworth	3	Mobile CR Laboratory	2	Uppsala	3
ESOISR	2	Morioka	3	Ushuaia	3
Fort Smith	5	**Moscow**	**2**	Utrecht	3
Freiburg	3	Moscow experimental	2	Weissenau	3
Fukushima	3	**Mt Norikura**	**2**	Wellington	3
Goettingen	3	**Mt Washington**	**2**	Victoria	3
Goose Bay	**2**	**Mt Wellington**	**2**	Wilkes	3
Hafelekar	2	Munchen	3	Vostok	2
Haleakala_IGY	2	Murchison Bay	3	Yakutsk	2
Haleakala_SM	2	Murmansk	3	Zugspitze	4
Halle	3	Nain	1		
Heiss Is	3	Nederhorst	3		
Hermanus	**1**	Neumayer 3	4		

6 Discussion and conclusions

We have performed a survey of all available NM records in a number of publicly available datasets and assessed their quality. We present a comprehensive table containing detailed information about the available datasets and also a list of recommended data sources for each station. This information is collected based on the state of affairs as of writing; the datasets are subject to change and therefore users of this information need to keep this in mind. Nevertheless, these results form the most extensive and up-to-date analysis of the NM datasets and provide useful basic information for users and developers of the related services.

It appears that datasets for the same NMs are not identical between different sources, making it difficult to control the reliability and reproducibility of studies based on NM data. While the WDCCR provides a simple repository for the data without corrections and updates of the data, other data sources try to resolve this problem. However, even for the NMDB project, there are discrepancies between different data tables, in particular the 1-hour and revori ones.

Somewhat surprisingly, station homepages are not the recommended sources for multiple stations. It seems that through the advent of NMDB, many neutron monitor stations have switched to preferring to use NMDB to distribute their station data. This often leads to a situation where NMDB has more up-to-date and reliable (corrected) data available. Nevertheless, nearly all station homepages are at least a secondary recommended source, so using station homepages is mostly reliable.

IZMIRAN implements corrections in many datasets that are not available elsewhere. One such example is the Rome station, where IZMIRAN has corrected a large number of steps. This is useful, but a proper description of corrections is not readily available.

The results seem to indicate that a rule-of-thumb for selecting which data source to use is as follows:

1. Station homepages are often a good choice, but might not always have the most up-to-date data
2. NMDB is usually a good choice for long-lived European NMs but also houses reliable data from many NMs from around the world.
3. IZMIRAN is a good choice for most Russian and East European NMs but also has good and/or corrected data from other areas. IZMIRAN often has a corrected version of WDCCR data.
4. WDCCR has data from many (short-lived) stations that are not available elsewhere, but usually other sources have more reliable data.

A summary geographic map of these recommendations is shown in Figure 1. Because of the large number of stations, names are not shown. For more detailed information and station names, the reader should refer to the supplementary table.

The metadata of the stations are sometimes not identical across different data sources. In particular, the location information is not always exact and might have changed throughout records. The geomagnetic cut-off rigidity is typically given as a single value without details on how it was calculated and to what time refers, while it may change significantly, especially for mid-latitude stations operated for decades (Smart & Shea, 2009). The naming of some stations can also cause confusion for data users which are not aware of the histories of specific stations. Such examples involve the Swarthmore/Newark station which moved from one location to another nearby one in 1978, and can be referred to as "Newark", "Swarthmore" and "Swarthmore/Newark" in different data sources. The "Newark" dataset can either have data for the whole Swarthmore+Newark period (Station, IZMIRAN, Station) or only for non-Swarthmore-period (WDCCR). Separate datasets

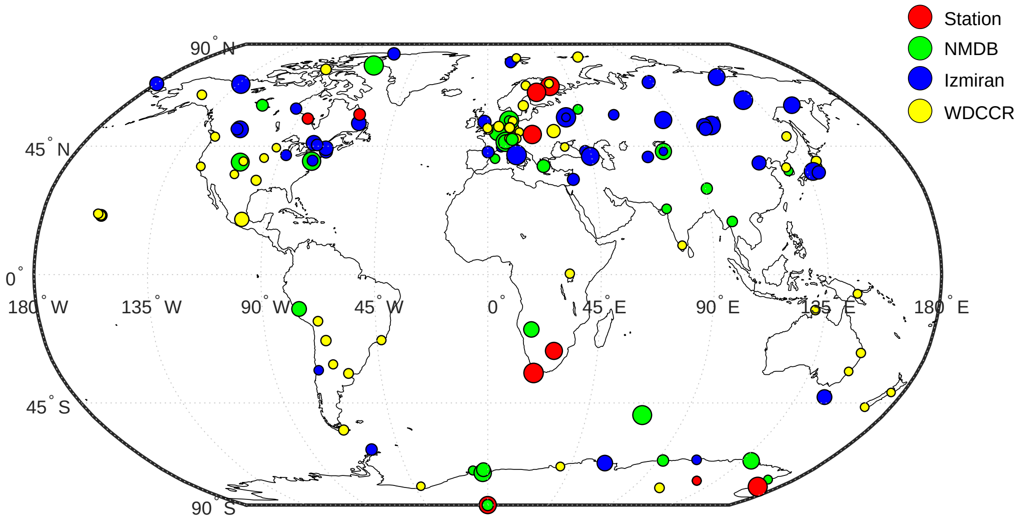

Figure 1. Geographical distribution of NM stations with recommended sources shown as marker colors. Size of markers indicate the amount of available data in the station.

only for Swarthmore data are available in WDCCR and in IZMIRAN, called Swarthmore and Swarthmore/Newark, respectively. This is confusing since the Bartol institute uses Swarthmore/Newark as the name for the dataset containing the full dataset, whereas IZMIRAN only contains Swarthmore data. Also, the Aragats and Nor-Amberd stations (in NMDB) have differing names, which are also called "Yerevan3000" and "Yerevan2000" in IZMIRAN or "Erevan3" and "Erevan" in WDCCR, respectively. The acronyms of the stations may also differ accordingly in the data-sources.

These inconsistencies make the use of data difficult for a non-expert, who is not familiar with datasets and the history of ground-based observations. Here we made an effort to systematize the available and partly controversial information and to provide a user with a verified set of ground-based cosmic-ray measurements. A detailed analysis of the stability of the data from different stations is planned for forthcoming work.

It should be noted that this survey presents only a momentary snapshot (as for June 2020) of the situation with data sources. The analysis has only been conducted for the 1-hour data resolution, and results with other resolutions may differ. Due to the nature of online data services, the presented results may change when data are changed, corrected, removed or combined in the analyzed data-sources. The selection of data-source recommendations includes a visual inspection of the data to account for the incompleteness of the prime station validation, which can introduce a subjective bias in the results. This analysis also does not take into account possible corrections that might easily render the source in question to have reliable and comparable measurements to other sources. When selecting the data source to use, one should refer to the data coverage (number of data points) in the information table to check out if an "non-recommended" source could possibly have more data coverage after corrections. The prime-station method utilized here only roughly validates the data quality in relation to other stations, and may not be accurate for high-rigidity stations, because of their low statistic. For example, the $< 10\%$ limit for good data did not catch the clear 4% step in many Newark datasets (See Supplementary Information S1). A more sophisticated method, based on theoretical modeling of cosmic-ray modulation derived from the entire NM network would provide a more robust assessment, and it is planned for the subsequent work.

Acknowledgments

The prime dataset used in this paper is archived at `http://cosmicrays.oulu.fi/primedata/` according to FAIR guidelines. We acknowledge all the databases and station teams for their datasets that are available in the sources/links described in Section 3. These include the following: NMDB database (`http://nest.nmdb.eu/`), founded under the European Union's FP7 programme (contract no. 213007); IZMIRAN database (`http://cr0.IZMIRAN.ru/common/links.htm`) operated by the Pushkov Institute of Terrestrial Magnetism, Ionosphere and Radiowave Propagation; WDCCR database (`ftp://ftp.isee.nagoya-u.ac.jp/pub/WDCCR/STATIONS/`) operated by the Solar-Terrestrial Environment Laboratory, Nagoya University; The Bartol Research Institute neutron monitor program (`http://neutronm.bartol.udel.edu/~pyle/bri_table.html`) which is supported by National Science Foundation grant ATM-0000315; Apatity station (`http://pgia.ru/data/nm`) operated by The Polar Geophysical Institute (Murmansk region, Russia); Jungfraujoch station (`http://cosray.unibe.ch/data/data.html`) operated by the Physikalisches Institut, University of Bern, Switzerland; Lomnický Štit station (`http://neutronmonitor.ta3.sk/archive.php`), which is supported by the IEP SAS in Košice, by the APVV grant agency project APVV-15-0194 and by VEGA, project no. 2/0026/16; Mexico City Cosmic Ray Observatory (`http://www.cosmicrays.unam.mx/`) operated by the Universidad National Autónoma de Mexico (UNAM); Stations (`http://natural-sciences.nwu.ac.za/neutron-monitor-data`) operated by the Centre for Space Research at the Potchefstroom Campus of the North-West University; Yakutsk and Tixie bay stations (`https://www.ysn.ru/ipm/`) operated by the Yakutsk Scientific Center of SB RAS, funded by RFBR grant 04-07-90054; Oulu cosmic rays station (`http://cosmicrays.oulu.fi/`) is operated by Sodankylä Geophysical Observatory of the University of Oulu, Finland; as well as all other individual stations mentioned in this work that are included in the databases (NMDB, IZMIRAN or WDCCR). This work was partially supported by the Academy of Finland (project No. 321882 ESPERA) and by the Finnish Cultural Foundation (grant number 00191177).

References

Abunin, A. A., Pletnikov, E. V., Shchepetov, A. L., & Yanke, V. G. (2011). Efficiency of detection for neutron detectors with different geometries. *Bull. Russian Acad. Science, Phys.*, *75*, 866-868. doi: 10.3103/S1062873811060037

Belov, A. (2000). Large Scale Modulation: View From the Earth. *Space Sci. Rev.*, *93*, 79-105. doi: 10.1023/A:1026584109817

Belov, A., Belov, M., Gushchina, R., Eroshenko, E., Kartyshev, V., Struminsky, A., & Yanke, V. (1998). Monitoring of cosmic rays in real time and information system of the Moscow cosmic ray station. In *Esa workshop on Space Weather Proceedings.*

Cooke, D., Humble, J., Shea, M., Smart, D., Lund, N., Rasmussen, I., … Petrou, N. (1991). On cosmic-ray cut-off terminology. *Nuovo Cimento C*, *14*, 213–234. doi: 10.1007/BF02509357

Gil, A., Usoskin, I. G., Kovaltsov, G. A., Mishev, A. L., Corti, C., & Bindi, V. (2015). Can we properly model the neutron monitor count rate? *J. Geophys. Res.*, *120*, 7172-7178. doi: 10.1002/2015JA021654

Hatton, C. J. (1971). The Neutron Monitor. In Wilson, J.G. and Wouthuysen, S.A (Ed.), *Processes in elementary particle and cosmic ray physics* (pp. 3–100). Amsterdam-London: North-Holland Publ. Comp.

Hatton, C. J., & Carimichael, H. (1964). Experimental Investigation of the NM-64 Neutron Monitor. *Canad. J. Phys.*, *42*, 2443-2472. doi: 10.1139/p64-222

Lincoln, J. V., & Shea, M. A. (1973). Cosmic Ray Data in World Data Centers. In *International Cosmic Ray Conference* (Vol. 2, p. 1064).

Mavromichalaki, H., Papaioannou, A., Plainaki, C., Sarlanis, C., Souvatzoglou, G., Gerontidou, M., … Pustil'nik, L. (2011). Applications and usage of the real-time neutron monitor database. *Advances in Space Research*, *47*(12), 2210 - 2222. doi: https://doi.org/10.1016/j.asr.2010.02.019

Moraal, H., Belov, A., & Clem, J. M. (2000). Design and co-Ordination of Multi-Station International Neutron Monitor Networks. *Space Sci. Rev.*, *93*, 285-303. doi: 10.1023/A:1026504814360

Poluianov, S., Usoskin, I., Mishev, A., Moraal, H., Kruger, H., Casasanta, G., … Udisti, R. (2015). Mini Neutron Monitors at Concordia Research Station, Central Antarctica. *Journal of Astronomy and Space Sciences*, *32*(4), 281-287. doi: 10.5140/JASS.2015.32.4.281

Simpson, J. A. (1948). The Latitude Dependence of Neutron Densities in the Atmosphere as a Function of Altitude. *Phys. Rev.*, *73*, 1389-1391. doi: 10.1103/PhysRev.73.1389

Simpson, J. A. (2000). The cosmic ray nucleonic component: The invention and scientific uses of the neutron monitor – (keynote lecture). *Space Science Reviews*, *93*(1), 11–32. doi: 10.1023/A:1026567706183

Smart, D. F., & Shea, M. A. (2009). Fifty years of progress in geomagnetic cutoff rigidity determinations. *Adv. Space Res.*, *44*, 1107-1123. doi: 10.1016/j.asr.2009.07.005

Strauss, D., Poluianov, S., van der Merwe, C., Krüger, H., Diedericks, C., Krüger, H., … Traversi, R. (2020). The mini-neutron monitor: a new approach in neutron monitor design. *J. Space Weather Space Clim.*, *10*, 39. doi: 10.1051/swsc/2020038

Usoskin, I. G., Bazilevskaya, G. A., & Kovaltsov, G. A. (2011). Solar modulation parameter for cosmic rays since 1936 reconstructed from ground-based neutron monitors and ionization chambers. *J. Geophys. Res.*, *116*, A02104. doi: 10.1029/2010JA016105

Usoskin, I. G., Gil, A., Kovaltsov, G. A., Mishev, A. L., & Mikhailov, V. V. (2017). Heliospheric modulation of cosmic rays during the neutron monitor era: Calibration using PAMELA data for 2006-2010. *J. Geophys. Res. (Space Phys.)*, *122*, 3875-3887. doi: 10.1002/2016JA023819

Usoskin, I. G., Mursula, K., & Kangas, J. (2001). On-Line Database of Cosmic Ray Intensities*. In *International Cosmic Ray Conference* (Vol. 9, p. 3842).

Supporting Information for "Seven decades of neutron monitors (1951 – 2019): Overview and evaluation of data sources"

Pauli Väisänen[1], Ilya Usoskin[1,2], Kalevi Mursula[1]

[1]Space Physics and Astronomy Research Unit, University of Oulu, Finland

[2]Sodankylä Geophysical Observatory, University of Oulu, Finland

Contents of this file

1. Text S1
2. Figures S2 to S4

Additional Supporting Information (Files uploaded separately)

1. Neutron monitor information table S5 (NMStationInfoList.xlsx)
2. Data source recommendation list S6 (NMSourceRecommendations.xlsx)

Introduction This supporting information contains information and examples about the data analysis process of the main article. The station recommendations and station information which cannot be incorporated in the main article text are also included here. A description for all supplementary information is given.

Text S1. Here we give an overview of the analysis of the quality of NMs on an example of the Newark station, referred henceforth as NWRK.

Figure S2 shows the original data, the normalized data and the ratios between them, corresponding to different sources. One can see from the upper panel that the datasets are almost equivalent in terms of shape, but the absolute level of datasets, retrieved from the NMDB source, is higher than the others (WDCCR line is at the same level as IZMIRAN and the Station's web-site), by a constant factor of 5/3. After the data normalization to the median of years $1995-1996$ (middle panel), the datasets become nearly identical, except for some differences related to the data coverage after 2017 and a small step-like difference between NMDB and the others in $2012-2015$. The ratios (low panel) between the normalized datasets are very close to unity during most of the time, except for small discrepancies across the dataset. The offset-type difference in $2012-2015$ is about 4 %.

Figure S3 depicts ratios between the hourly values of the normalized count rates of the NWRK NM, obtained from the different sources and those of the reference dataset $\mathrm{NM}_{\mathrm{med}}$. For WDCCR, the data for Swarthmore (the previous name of the NWRK station) is also shown. First, one can see that the length of the NWRK dataset is different in different sources. It is the longest (since 1964) in WDCCR, IZMIRAN, station's webpage and NMDB *1hr*, but shorter (since 2000) in the NMDB-revori data tables. There are also some outliers (red points in the Figure), defined as hourly values which deviate by $> 10\%$ from the normalized reference $\mathrm{NM}'_{\mathrm{med}}$ values. Such outliers are relatively frequent after 2010 in the NMDB-revori table but absent in the WDCCR and the station web-page. However, the latter two do not contain the data since 2017. NMDB and IZMIRAN have data also after 2017 and contain several outliers. Another important aspect is the long-term stability of the data. While NMDB-1hr dataset is consistently tied to unity, other datasets exhibit some apparent features, which are not outliers: a 4 % drop during 2012 – 2015 in WDCCR, IZMIRAN and station's tables, or a long-term trends the NMDB-revori table. This illustrates that datasets from different sources are not identical, and need to be carefully checked. For the case of NWRK NM we recommend that the NMDB-1hr dataset is used as the data source. This recommendation is given even in spite of the 5/3 ratio difference, since NM data are mostly used for analyzing relative and not absolute changes, and the difference is very easily corrected if needed.

A similar analysis has been performed for all the 147 stations in the study, and the resulting list of recommendations S5 is based on all of them.

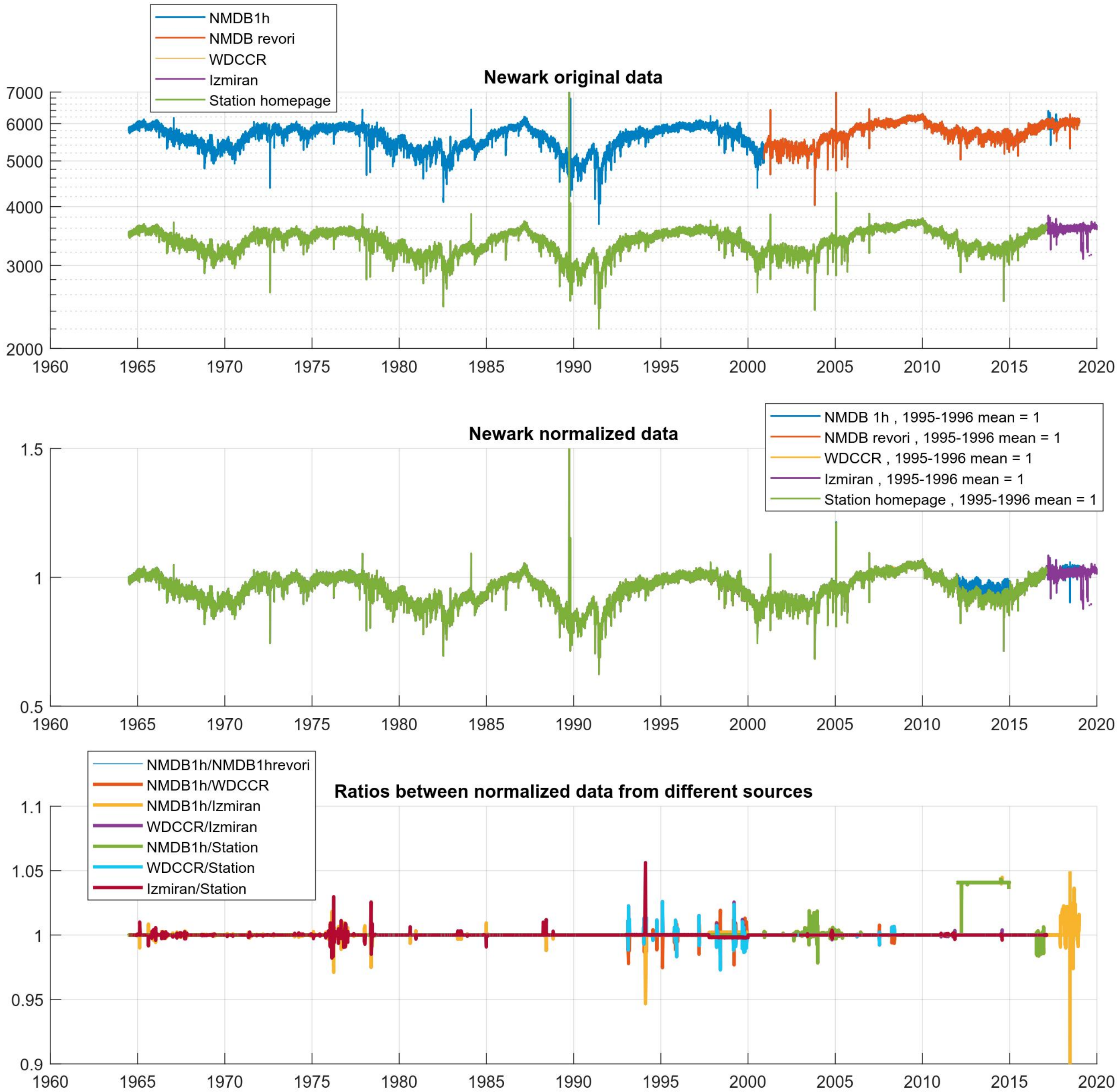

Figure S2. Top: Original Newark/Swarthmore (pressure-corrected) data from different sources. Middle: The same data normalized to the median of years 1995 – 1996. Bottom: Ratios between the normalized datasets, shown in the middle panel.

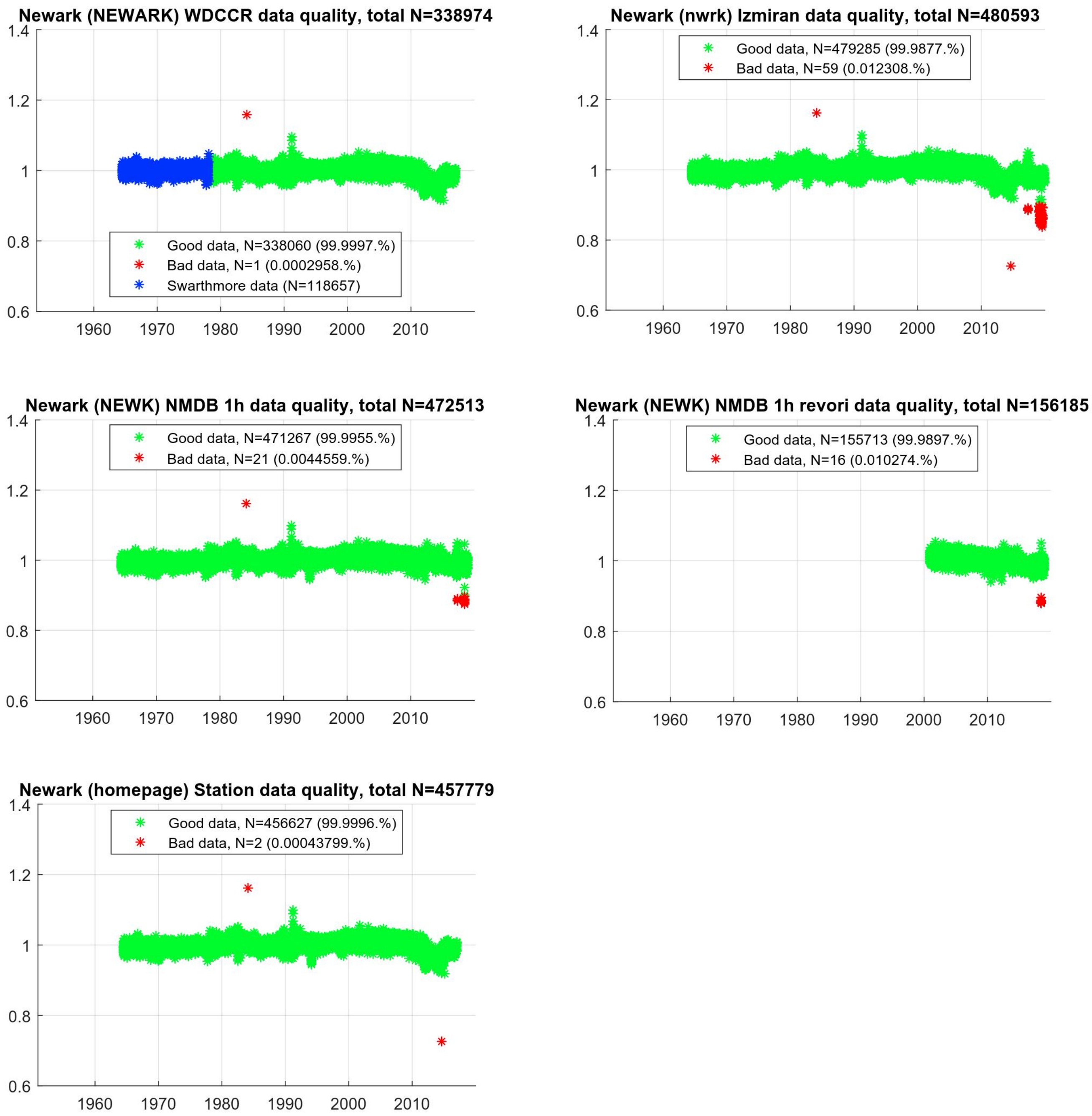

Figure S3. The ratio of the normalized Newark (NWRK) NM hourly count rates, as obtained from different data sources, to those of the reference $\mathrm{NM}'_{\mathrm{med}}$ dataset (Figure S4b). The NWRK data sources are WDCCR, NMDB-1hr, NMDB-revori, IZMIRAN, and station's site, respectively. The color indicates the data quality: *green* indicates good data within ±10% of the $\mathrm{NM}'_{\mathrm{med}}$ values. The total number and the percentage of hourly data points of different quality are given in the legends.

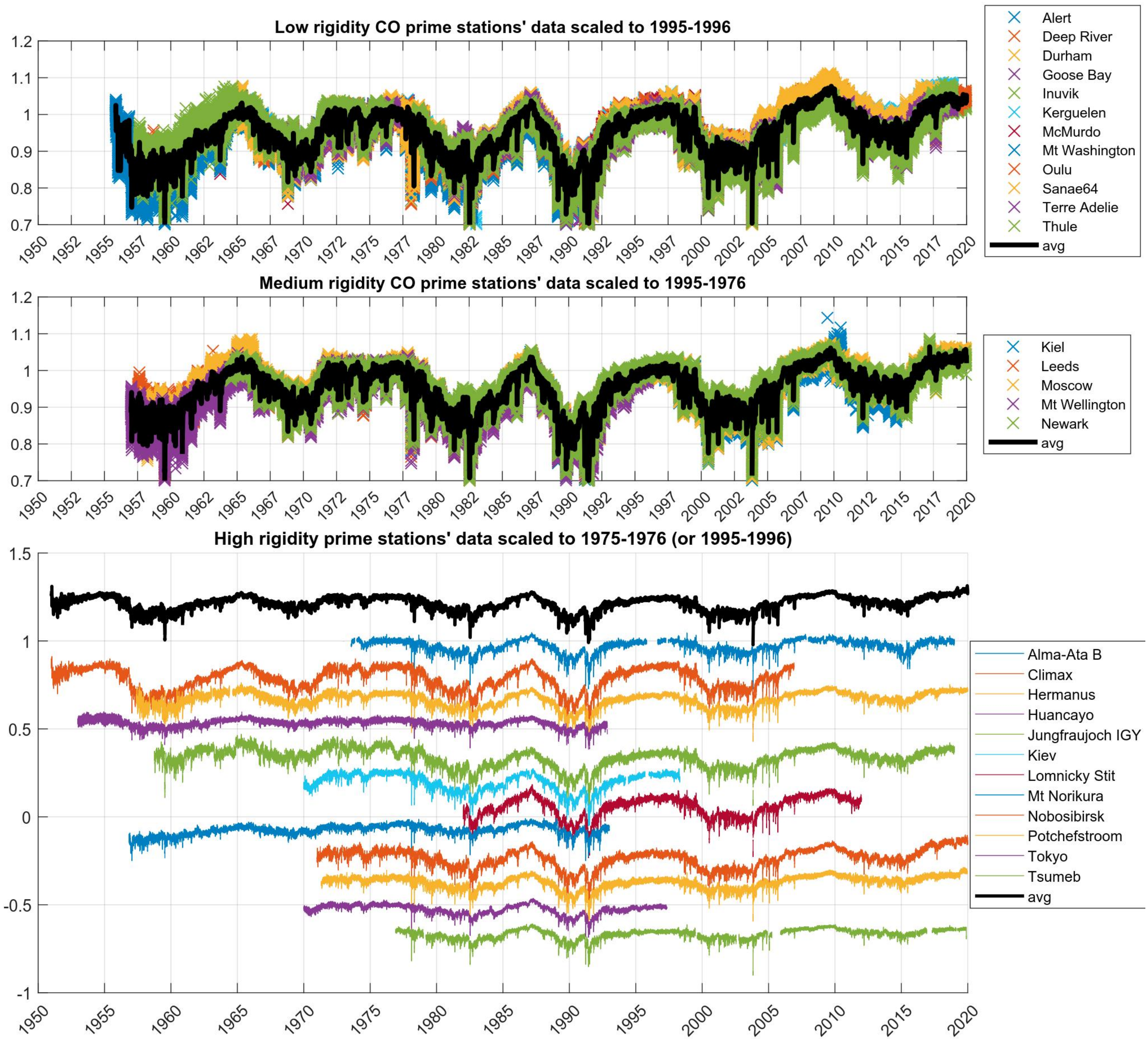

Figure S4. Normalized (see text) time profiles of hourly datasets for the primary NM stations for low ($R_c \leq 1.75$ GV, upper panel) and medium ($1.75 < R_c \leq 2.75$ GV, middle panel) rigidity groups. Colored curves depict individual datasets, while black ones represent the group-averaged reference dataset. High rigidity stations (below) are each offset by 0.15 from each other. Full description is given in the main article Section 4.

Table S5. The data table named NMStationInfoList.xlsx contains metadata, coverage and data quality information. It includes the following 29 columns:

1. Station name	16. Geographical longitude
2. Other name(s)	17. Altitude of the NM location
3. Start year - End year	18. Geomagnetic Cut-off rigidity
4. WDCCR acronym	19. Number of "Good" data points in WDCCR
5. NMDB acronym	20. Fraction of "Good" data points in WDCCR
6. IZMIRAN acronym	21. Number of "Good" data points in NMDB-1h
7. URL of Station homepage	22. Fraction of "Good" data points in NMDB-1h
8. Number of available data sources	23. Number of "Good" data points in NMDB-revori
9. Data points (1h) in WDCCR data	24. Fraction of "Good" data points in NMDB-revori
10. Data points (1h) in NMDB-1h data	25. Number of "Good" data points in IZMIRAN
11. Data points (1h) in NMDB-revori data	26. Fraction of "Good" data points in IZMIRAN
12. Data points (1h) in IZMIRAN data	27. Number of "Good" data points in Station
13. Data points (1h) in Station data	28. Fraction of "Good" data points in Station
14. Maximum coverage in years	29. Maximum coverage by "Good" data in years
15. Geographical latitude	

Table S6. The table named NMSourceRecommendations.xlsx (uploaded separately) contains the recommended and secondary source of neutron monitor data alongside short notes regarding the dataset.