PROCEEDINGS OF SCIENCE

ONLINE ICRC 2021
THE ASTROPARTICLE PHYSICS CONFERENCE
Berlin | Germany
37th International Cosmic Ray Conference
12–23 July 2021

Quality survey of Neutron Monitor data sources for 1951-2019

Pauli Väisänen,[a,*] **Ilya Usoskin**[a,b] **and Kalevi Mursula**[a]

[a] *University of Oulu,*
Pentti Kaiteran katu 1, Oulu, Finland
[b] *Sodankylä Geophysical Observatory,*
Tähteläntie 62, Sodankylä, Finland
E-mail: pauli.vaisanen@oulu.fi, ilya.usoskin@oulu.fi, kalevi.mursula@oulu.fi

Long-term measurements from the global neutron monitor (NM) network allow us to study galactic cosmic ray (GCR) variations for the last seven decades. However, the network offers data of quite different quality from several sources. Historically, NM data is distributed through different data repositories, which include the Neutron Monitor Database (NMDB), World Data Center for Cosmic Rays (WDCCR), The Pushkov Institute of Terrestrial Magnetism, Ionosphere and Radiowave Propagation (IZMIRAN) repositories and individual homepages of stations/teams. Here we present a detailed quality survey by comparing the consistency of hourly resolution NM datasets of different origin. The analysis includes 300 datasets from 147 NMs in 1951-2019. As the main result of the survey, we found that the data of individual stations are not often uniform across the different sources. This results in problems with the reliability and reproducibility of scientific results. Our survey also underlines that special efforts should be given to a proper documentation of the datasets. This is particularly true for the oldest data that are in danger of getting lost to time. We also offer a list of currently recommended data sources for each station, based on their comparison with a "prime" dataset composed from long-lived NM stations that fulfil specific quality criteria.

37th International Cosmic Ray Conference (ICRC 2021)
July 12th – 23rd, 2021
Online – Berlin, Germany

*Presenter

PoS(ICRC2021)1244

1. Introduction

Neutron monitors (NMs) measure the variation of galactic cosmic rays (GCR) on Earth. The global neutron monitor network consists of over 100 NMs on all continents. These measurements are distributed via different repositories or through the homepages of NMs. There is an overlap between different repositories, and many stations have data available from multiple sources. This wide availability is beneficial for the accessibility of data, but can lead to confusing situations, where different repositories are hosting different version of the data.

In [1], all available NM datasets at 1-hour resolution from all sources were analysed using a set of good quality "prime stations" to produce a list of recommended data sources for each possible station. This recent list and work is an unique survey of the extent of the global NM network and its datasets.

Here we will present the data from the recommended sources to present and visualize, what does the NM data from all recommended sources look like after collecting and corrections. We show the data collection and correction methods in Section 2, and present and analyse the resulting "recommended dataset" in Section 3. We conclude the findings in Section 4.

2. Data and methods

The data from the recommended stations shown in Table 1 was collected using an automated process in in the same way as described in [1]:

1. Data from NMDB and IZMIRAN was downloaded using the relevant formulation of the http address with parameters.
2. Data from WDCCR was downloaded via the FTP
3. Data from individual station homepages were downloaded via the relevant web page. The download locations are available in [1]

2.1 Corrections

The analysis of data sources which was used to compile the recommendation list in [1] did not consider the general data quality, but instead focused on the station-specific quality. This means that even the "best" source of the station can have problems like outliers, jump and other inconsistencies. Therefore, we still need to do corrections to some of the datasets in order to use the data.

Outliers were removed using a median (Hampel) filter by removing elements with more than three local scaled MADs (median absolute deviation) from the median of 5 points. Times of GLE's from gle.oulu.fi were not included in the outlier removal process.

In addition to outliers, NM datasets often have uncorrected jumps of the data, which we removed by finding the peak of a kernel density estimate and removing values that were ±30% from the peak. This method means that the longest section of the dataset is preserved.

We then scale data from all the stations to the median count-rate of 1975-1976 (or 1995-1996 if no data from 1975-1976). If those years are not available, we scale the station to the median of other stations in the same rigidity bin that have available data in 1975-1976 or 1995-1996. After

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

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	Predigtstuhl	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	**Lomnický Štit**	**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		

the following rigidity binning, we also require the data to not be over ±10% off from the median values of the bin.

2.2 Rigidity cutoff grouping

Since measurements from stations with different rigidity cutoffs is hard to combine, we group the stations to specific rigidity bins. We use the following bins:

- Low rigidity stations: $R < 1.75\,GV$
- Medium rigidity stations: $1.75\,GV \leq R \leq 2.75\,GV$
- High rigidity stations: $R > 2.75\,GV$

The high rigidity bin is still very wide and mixed, so that needs to be taken into consideration when viewing the results. But for illustrative purposes this binning is sufficient.

3. Results and discussion

An overview of the raw data and corrected+scaled version of all datasets is shown in Figure 1. The raw data shows the extent of the data, but also indicates some very clear problems, outliers and steps in the data. Using the corrections

In order to better quantify the full dataset, in Figure 2 we present the 27-day moving average count rates of all the stations and medians separated by rigidity cutoff bins (low, medium or high). We also show the median absolute deviations of the specific rigidity bins, which is a descriptor of variance around the median, similar to the standard deviation. In the bottom panel we show the number of available stations.

In the upper panel we see that when using this scaling, count rates are quite similar during solar minima, whereas the differences of rigidity cutoff are evident during solar maximum times. One exception is the deep solar minimum of 2008-2010, where the low rigidity stations show clearly higher count rates than medium or high cutoff stations, that is related to the flattening of the heliospheric current sheet to which lower-energy cosmic rays are more sensitive.

The median absolute deviations also underline the activity during solar maximum which causes differences based on rigidity cutoff. We can see that during minimum years in 1976, 1986 and 1996, deviations across stations of all cutoffs is small, although the scaling can affect these. Interestingly, deviations during and after solar cycle 24 (2008-2014) are slightly higher in all rigidity bins. This could be due to change in the observing stations, temporal distance to the scaling years (which might strengthen any possible drifts in NM count rates) or maybe a change in the modulation of GCR in the heliosphere and near-Earth space.

We can see that after the gradual start of operated NMs from 1951, the number of stations rose sharply in 1964, which was the IQSY campaign. There was also a small decline in 1960, when many short-lived IGY stations were shut down. We also see that in recent years after 2017, the number of available stations has started declining. This is troubling, since a good coverage of long-lived NMs is important for the monitoring of the Earth's radiation environment. The total coverage of the raw dataset gives the theoretical maximum coverage, if the outliers and errors removed could be corrected.

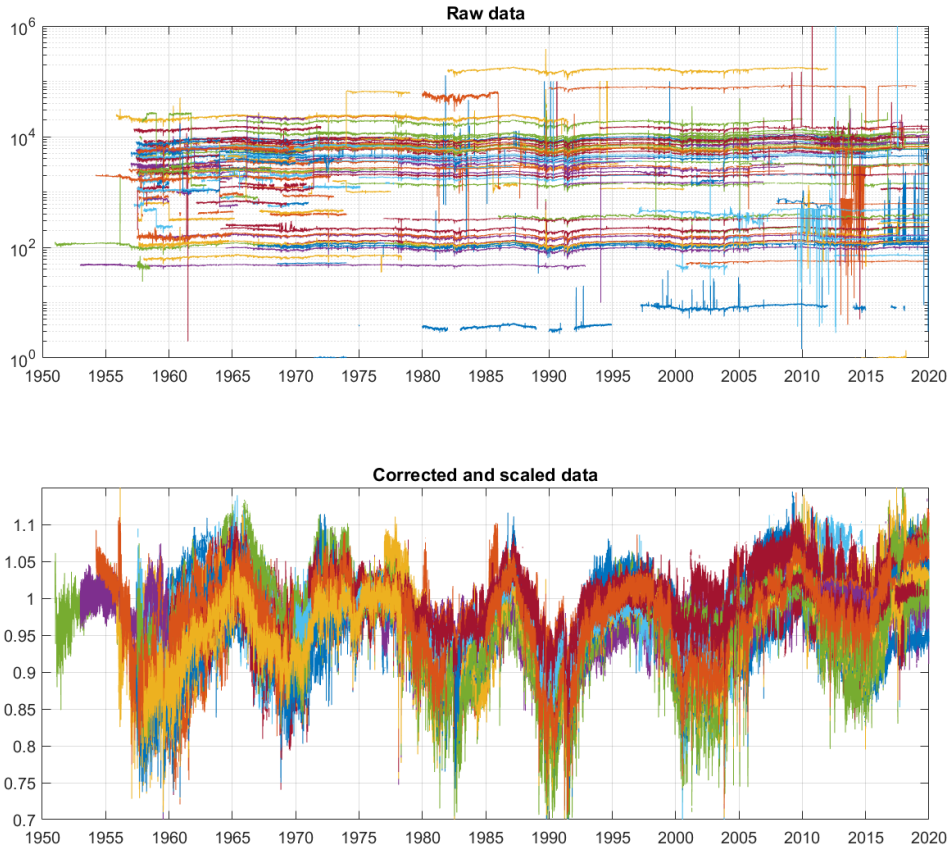

Figure 1: Top: Raw hourly count rates from all recommended stations (147). Bottom: Corrected and scaled data of all stations

4. Conclusion

This visualisation of the 147 recommended NM stations demonstrates the current extent of the measurements of the global NM network [1]. The recommended datasets can be used for high and accurate studies, but they still require a number of (automated) corrections for the data to be usable or reliable. Corrections at the data sources could be the most efficient method, since then users do not need to implement their own corrections to each dataset every time they use the data.

This analysis of the measurements from the six last decades also raises questions about the future of the global NM network. NMs requires long-term resourcing and knowledgeable, trained staff for reliable operations and usable datasets. The current situation with scattered and differing datasets across different repositories might not be sustainable in the long-term, and some kind of action or project might be needed for the preservation, collection and correction of the historical NM datasets. Nevertheless, the current total extent of the NM datasets is still sufficient and enables high-end research on solar activity, cosmic ray modulation and the radiation environment around Earth. Users just need to be more aware of all the different caveats related to NM data.

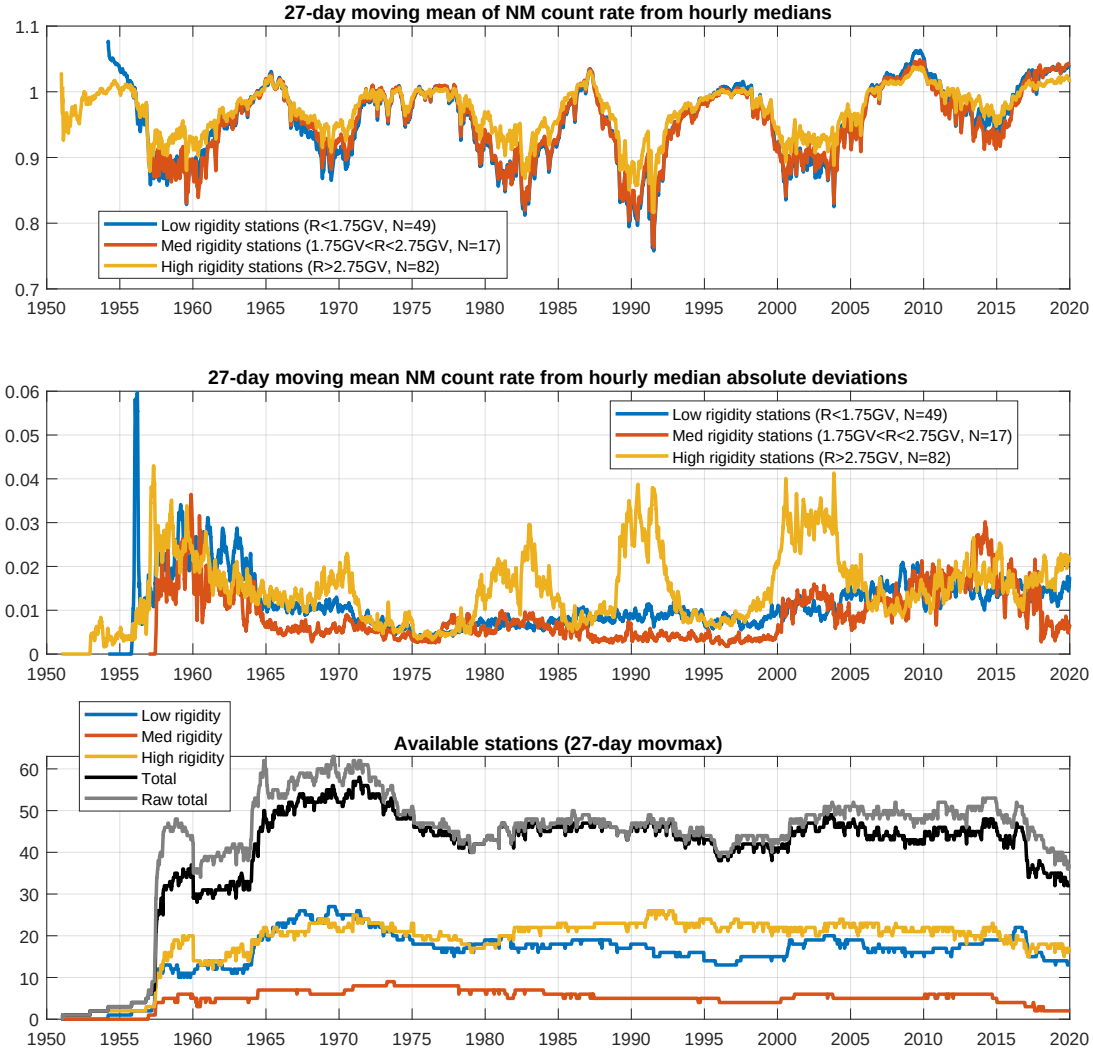

Figure 2: Moving 27-day averages of median count-rates (top), median absolute deviations (middle) and coverages (bottom) of NM stations of the different rigidity bins.

Acknowledgements

This work was supported by the Academy of Finland (project No. 321882 ESPERA). We acknowledge all the stations teams and the three main repositories (NMDB, IZMIRAN and WDCCR) for their datasets.

References

[1] P. Väisänen, I. Usoskin and K. Mursula, *Seven Decades of Neutron Monitors (1951–2019): Overview and Evaluation of Data Sources*, *Journal of Geophysical Research (Space Physics)* **126(5)** (2021).