About right: references in Open Access EGU journals

Andrea Pozzer¹

¹Max Planck Institute for Chemistry, 55128, Mainz, Germany

Correspondence: Andrea Pozzer (andrea.pozzer@mpic.de)

Abstract. We investigated the number of references per page for different European Geophysical Union journals, which share
the same text formatting. Although the journals formally all focus on geoscience, different disciplines are accommodated, from
ocean science and biogeosciences to the technical description of numerical model development. In this study, we show that the
number of references per page is remarkably constant across these journals. In addition, this value has remained constant in
the last decade, despite the consistent increase in the number of pages and in the number of references in almost all journals
considered. Independently of the quality of the references used in a article, we show that for the EGU journals the average
number of references per page is 3.82 (1.87 – 6.18 at 90% confidence level). This reveals that there is a consensus regarding
optimum reference density, which depends on the journal’s layout and not on the journal’s covered discipline.

1 Introduction

The number of references in a scientific paper can influence reader judgement of the paper’s quality (Lovaglia, 1991), and is
thus an important factor in defining its success, i.e. its number of citations (Fox et al., 2016). Therefore it is important that
authors include an optimal (and balanced) quantity and quality of references in their articles.

It has been shown (Abt and Garfield, 2002) that the number of references per unit page is remarkably constant across a large
number of disciplines, and that longer papers are, on average, more cited than shorter papers (Leimu and Koricheva, 2005).

Nevertheless, the creation of homogeneous and standardised text length is a challenging task, with each journal having different
formatting layouts, which could influence the perception of reference quantity and, indirectly, result in pressure for an increase
or decrease in their numbers.

In the last decades, the length of scientific papers undergone a significant increase. Ucar et al. (2014) showed not only a
clear trend towards an increase in the number of pages in papers in engineering journals, but also showed that this increase
has not yet begun to level off. This increase in paper length is mirrored by a constant increase in the number of references
over time (Biglu, 2008; Jaunich, 2018). Bornmann and Mutz (2015) revealed a large increase in the number of references
from the middle of the twentieth century onward. The temporal increase in the number of references per papers varies among
different disciplines (Sánchez-Gil et al., 2018). Furthermore, Nicolaisen and Frandsen (2021) showed that ”there is a drop in
short reference lists and a corresponding increase in a bit longer and medium size reference lists. Long and very long reference
lists remain much more stable in shares over time, and does therefore not contribute much to the observed growth.” A steady
state in reference numbers has until now only been artificially reached in a few journals and/or manuscript types, through
the enforcement of limits in the number of references (Anger, 1999). Nevertheless, most of these studies only focused on
the number of references per article, without analysing this parameter with respect to the paper length, or, similarly, without
investigating reference density. A notable exception is the work of Milojević (2012), which found different temporal trends in
reference per page, depending on the field of study.

The European Geophysical Society (one of the predecessors of the European Geophysical Union), started its first open access
(OA) journal in 2001, with the launch of the Atmospheric-Chemistry and Physics journal (Pöschl, 2004, 2012). The success of
this first journal prompted The European Geophysical Union (EGU), through Copernicus Publications, to establish additional
OA journals; 19 journals are currently published by Copernicus (for EGU), covering various topics of the Earth, planetary and
space sciences.

In this work we took advantage of the OA EGU journals, which have identical layouts, and therefore allow for a direct
comparison between the different journals. In addition, all the metadata of the papers have been published online in a search-
able XML format, which allows automatic computer scripting for information gathering. It must be stressed that Copernicus
Publications publish other OA journal in addition to the EGU journals studied. However, these journals use diverse layout, and
the comparison between such different formats is not always straightforward.

In this work we analyse the reference density, i.e. the number of references per page, in the OA journals published by the
EGU. The goal is to investigate if the reference density does change among journals which cover different topics but have the
exact same layout. We show that there exists a sound range for the number of references per page, similar for all OA EGU
journals, and that this has remained remarkably constant over time. In the Sect. 2, the methods for data collection are explained,
followed by an analysis of the temporal trends (Sect. 3) is presented. Finally, the main results are derived in Sec. 4, followed
by the conclusions.

2 Methodology

We considered articles accepted and published in XML form in the 2010–2020 period from the EGU OA journals. Therefore,
only EGU journals which started operating in 2010 at the latest were used in this study, which resulted in the inclusion of a
total of 12 journals (see Table 1):

- GMD, Geoscientific Model Development,
- ACP, Atmospheric Chemistry and Physics,
- BG, Biogeosciences,
- CP, Climate of the Past,
- AMT, Atmospheric Measurement Techniques,
- OS, Ocean Science,
− *ESD, Earth System Dynamics,*

− *TC, The Cryosphere,*

− *NHESS, Natural Hazards and Earth System Sciences,*

− *NPG, Nonlinear Processes in Geophysics,*

− *SE, Solid Earth,*

− *HESS, Hydrology and Earth System Sciences.*

An automatic Python script was used to recursively collect all the needed information, such as the number of pages and the number of references, from the XML version of each manuscript.

To avoid counting papers which cited an unrepresentative number of references (such as zero references or pure compilation articles), the outliers, which were defined as (i) papers containing no references or (ii) the papers containing a number of references above the average plus 3 times the reference’s standard deviation, were removed. In total 30,028 papers were downloaded, of which 787 were excluded as outliers, i.e. 29,241 published papers were used in this analysis.

In Table 1 the total number of papers analysed and those excluded from the analysis for each journal are presented. Roughly, \(~ 3\%~\) of the papers published in each journal were excluded as outliers. The outlier fraction is almost constant among all journals, ranging from 1.7\% for TC to 3.6\% for NHESS.

**Table 1.** Summary of journal characteristics. The number of papers analysed in each journal is listed, as well as the number of papers excluded (also expressed as a fraction) as outliers. The papers with the highest number of pages and the highest numbers of references are also listed for each journal.

<table>
<thead>
<tr>
<th>Journal</th>
<th>papers</th>
<th>excluded</th>
<th>% excluded</th>
<th>highest num.pages</th>
<th>highest num. refs</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMD</td>
<td>1931</td>
<td>45</td>
<td>2.3</td>
<td>66 gmd-13-3643-2020</td>
<td>526 gmd-12-3149-2019</td>
</tr>
<tr>
<td>BG</td>
<td>4197</td>
<td>147</td>
<td>3.5</td>
<td>56 bg-11-3547-2014</td>
<td>469 bg-7-2851-2010</td>
</tr>
<tr>
<td>CP</td>
<td>1318</td>
<td>35</td>
<td>2.7</td>
<td>50 cp-13-1851-2017</td>
<td>531 cp-13-1851-2017</td>
</tr>
<tr>
<td>AMT</td>
<td>3210</td>
<td>80</td>
<td>2.5</td>
<td>75 amt-9-4181-2016</td>
<td>346 amt-12-525-2019</td>
</tr>
<tr>
<td>OS</td>
<td>849</td>
<td>26</td>
<td>3.1</td>
<td>51 os-6-91-2010</td>
<td>255 os-14-471-2018</td>
</tr>
<tr>
<td>ESD</td>
<td>468</td>
<td>11</td>
<td>2.4</td>
<td>74 esd-10-379-2019</td>
<td>572 esd-10-379-2019</td>
</tr>
<tr>
<td>TC</td>
<td>1757</td>
<td>30</td>
<td>1.7</td>
<td>47 tc-13-3261-2019</td>
<td>402 tc-12-759-2018</td>
</tr>
<tr>
<td>NHESS</td>
<td>2500</td>
<td>91</td>
<td>3.6</td>
<td>42 nhess-18-2561-2018</td>
<td>308 nhess-11-2617-2011</td>
</tr>
<tr>
<td>SE</td>
<td>791</td>
<td>26</td>
<td>3.3</td>
<td>49 se-7-1417-2016</td>
<td>293 se-5-1243-2014</td>
</tr>
<tr>
<td>HESS</td>
<td>3378</td>
<td>95</td>
<td>2.8</td>
<td>47 hess-23-303-2019</td>
<td>431 hess-20-3799-2016</td>
</tr>
</tbody>
</table>
In addition to the number of analysed and disregarded papers, Table 1 lists the papers with the highest numbers of pages and references in the period 2010–2020 for each journal. The longest articles range in length from 42 pages (nhess-18-2561-2018) to 583 pages (acp-15-4399-2015). The maximum number of references in an article ranges from 255 in os-14-471-2018 to 793 in acp-15-4399-2015. acp-15-4399-2015 stands out among all other EGU articles both with respect to its number of pages as well as the number of references. In this paper, a list of measurements of Henry’s law coefficient for numerous gases of atmospheric relevance are presented. However, it should be noted that not all the papers with higher numbers of references are review articles or compilations of measurements (see for example, amt-12-525-2019 or gmd-12-3149-2019).

3 Temporal trends

We estimated for each analysed journal the trends in number of pages, references and references per page, and our results are presented in Table 2. In EGU publications, the number of pages and references per paper have been increasing in the last decade. The increase in pages per paper ranges from 0.26 pages/yr in ESD to 0.90 pages/yr in SE. Similarly, the number of references also increased in the same period, ranging from 1.06 to 3.91 references/yr in ESD and SE, respectively. Importantly, all these estimated temporal trends (both number of pages and number of references) are statistically significant at 99% confidence level, with the exception of the ESD journal.

Table 2. Linear fit of the temporal trends of pages, references (column refs) and references per page (column refs/page) for different EGU journals for all analysed papers between 2010 and 2020. The numbers inside the parentheses are the standard deviations of the estimated time trends (slope of the linear fit). The units are in yr⁻¹.

<table>
<thead>
<tr>
<th>Journal</th>
<th>pages</th>
<th>refs</th>
<th>refs/pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMD</td>
<td>0.52  (0.06)</td>
<td>2.01  (0.27)</td>
<td>0.02  (0.01)</td>
</tr>
<tr>
<td>ACP</td>
<td>0.34  (0.02)</td>
<td>1.81  (0.09)</td>
<td>0.03  (0.01)</td>
</tr>
<tr>
<td>BG</td>
<td>0.38  (0.02)</td>
<td>1.86  (0.14)</td>
<td>0.01  (0.01)</td>
</tr>
<tr>
<td>CP</td>
<td>0.43  (0.04)</td>
<td>2.43  (0.30)</td>
<td>0.03  (0.01)</td>
</tr>
<tr>
<td>AMT</td>
<td>0.39  (0.03)</td>
<td>1.27  (0.12)</td>
<td>0.01  (0.01)</td>
</tr>
<tr>
<td>OS</td>
<td>0.31  (0.05)</td>
<td>1.61  (0.22)</td>
<td>0.03  (0.01)</td>
</tr>
<tr>
<td>ESD</td>
<td>0.26  (0.10)</td>
<td>1.06  (0.52)</td>
<td>0.01  (0.03)</td>
</tr>
<tr>
<td>TC</td>
<td>0.45  (0.04)</td>
<td>1.99  (0.18)</td>
<td>0.02  (0.01)</td>
</tr>
<tr>
<td>NHESS</td>
<td>0.56  (0.03)</td>
<td>2.34  (0.13)</td>
<td>0.03  (0.01)</td>
</tr>
<tr>
<td>NPG</td>
<td>0.50  (0.06)</td>
<td>1.04  (0.26)</td>
<td>-0.04 (0.02)</td>
</tr>
<tr>
<td>SE</td>
<td>0.90  (0.08)</td>
<td>3.91  (0.39)</td>
<td>-0.00 (0.02)</td>
</tr>
<tr>
<td>HESS</td>
<td>0.51  (0.02)</td>
<td>2.28  (0.13)</td>
<td>0.03  (0.01)</td>
</tr>
<tr>
<td>total</td>
<td>0.45  (0.01)</td>
<td>1.90  (0.05)</td>
<td>0.01  (0.01)</td>
</tr>
</tbody>
</table>

The increase in number of citations may be attributed to the increasing growing of available literature. In fact, by publishing more papers, more manuscripts can (or must) be cited in future work. Analogously, the increase in absolute number of citations
reflects also the maturity that a specific science field has reached, where the large (and increasing) number of citations mirrors the large (and increasing) research performed on the specific topic. Further, accessibility could be a major point for increasing citations over time: OA papers (with the leading role of pure OA journals) allow an easy access to citable material, which can be easily referred to. In addition, technological development (e.g. fast internet connection, searchable and online downloadable journals) allows favourably the search and usage of precedent literature. Finally, Persson et al. (2004) suggested that with the intensification of scientific collaboration an increase of citations of co-published paper must be expected, and therefore this increase is a sign of increasing national and international collaboration between research teams.

In addition to the increase in the number of pages and the number of references in the period 2010-2020, we estimated also the evolution of reference density over this period. As shown in Table 2, these trends are very close to zero. The only journal with a clear statistically trend is ACP, which present an increase of references per page per year equal to 0.032, while none of the other journals present a statistically significant trend. This is in contrast to the findings of Ucar et al. (2014), which found a variable references to pages ratio along the 50 years of study, but is in agreement with the work of Abt and Garfield (2002).

Based on these estimates, we can consider reference density to be constant in OA EGU journals, which will allow us to use all papers published in the considered period to estimate it.

4 Results

The probability density distribution of pages against references is presented in Fig. 1. Both pages and references exhibit a clear log-normal distribution, although for a few journals (e.g. ESD) the number of papers available was quite low, which complicated the estimation of a meaningful statistics. In each plot the linear fit (with no intercept) was also overlaid on the distribution. The linear fits range from 2.8 (AMT) to 4.6 (CP) references/page, showing quite homogeneous behaviour within all the papers, with a coherent and similar reference density in all EGU journals.

For each journal the average number of pages and the number of references have been calculated, and the results are presented in Table 3. The average number of pages and references can exhibit strong variations between the journals, with differences of up to 60%. The longest papers appear on average in GMD, with 19 pages, while the shortest were published in NPG with 12 pages. NPG also exhibit the lowest number of references per paper (i.e. 40 references per paper), while CP has the highest, with 77 references per manuscript on average.

Finally, the average reference densities for each journals (based on the reference density for each manuscript) have been estimated (see Table 3 and Fig. 2). The number of references per page ranges from 3.00 to 4.77, for AMT and CP, respectively. Despite the differences in reference number or page distribution between the journals, the numbers of reference per page are statistically similar for all journals.

We have seen that the reference density for each journal presents a classical log-normal distribution. We then combined all the reference density distributions, obtaining, in the first order approximation, also a log-normal distribution (Mitchell, 1968; Cobb et al., 2012; Dufresne, 2008). We therefore estimated the overall reference density obtaining an average of 3.82 references/page with a confidence level of 90% between 1.87 and 6.18 references/page.
Figure 1. 2-Dimensional histogram (center) with frequency histogram for pages (top) and references (right) for different EGU journals. The journal name and the total number of papers, pages and references are listed on the top right of each plot. The black line depict the linear fit (with no intercept). The axes for the 2-dimensional histograms are the same in all plots.
Table 3. Average numbers of pages, references (column refs) and references per page (column refs/page) for different EGU journals for all analysed papers. The range at 90% confidence level is listed in parentheses.

<table>
<thead>
<tr>
<th>Journal</th>
<th>papers</th>
<th>pages</th>
<th>refs</th>
<th>refs/pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMD</td>
<td>1886</td>
<td>19 (10-32)</td>
<td>65 (22-129)</td>
<td>3.37 (1.55-5.50)</td>
</tr>
<tr>
<td>ACP</td>
<td>8781</td>
<td>17 (10-27)</td>
<td>65 (30-114)</td>
<td>3.90 (2.11-6.00)</td>
</tr>
<tr>
<td>BG</td>
<td>4050</td>
<td>15 (9-24)</td>
<td>70 (35-118)</td>
<td>4.57 (2.61-6.75)</td>
</tr>
<tr>
<td>CP</td>
<td>1283</td>
<td>16 (9-25)</td>
<td>76 (32-141)</td>
<td>4.77 (2.50-7.38)</td>
</tr>
<tr>
<td>AMT</td>
<td>3130</td>
<td>15 (9-26)</td>
<td>46 (19-84)</td>
<td>3.00 (1.45-4.94)</td>
</tr>
<tr>
<td>OS</td>
<td>823</td>
<td>15 (8-24)</td>
<td>49 (21-90)</td>
<td>3.34 (1.67-5.50)</td>
</tr>
<tr>
<td>ESD</td>
<td>457</td>
<td>16 (9-25)</td>
<td>61 (25-117)</td>
<td>3.88 (1.94-6.24)</td>
</tr>
<tr>
<td>TC</td>
<td>1727</td>
<td>15 (8-24)</td>
<td>56 (23-97)</td>
<td>3.68 (2.07-5.56)</td>
</tr>
<tr>
<td>NHESS</td>
<td>2409</td>
<td>13 (6-22)</td>
<td>45 (16-86)</td>
<td>3.48 (1.62-5.80)</td>
</tr>
<tr>
<td>NPG</td>
<td>647</td>
<td>12 (6-21)</td>
<td>40 (15-74)</td>
<td>3.43 (1.59-5.70)</td>
</tr>
<tr>
<td>SE</td>
<td>765</td>
<td>16 (8-28)</td>
<td>68 (27-129)</td>
<td>4.39 (2.14-7.07)</td>
</tr>
<tr>
<td>HESS</td>
<td>3283</td>
<td>16 (9-24)</td>
<td>58 (26-103)</td>
<td>3.70 (1.94-5.86)</td>
</tr>
<tr>
<td>total</td>
<td>29241</td>
<td>16 (9-26)</td>
<td>60 (24-111)</td>
<td>3.82 (1.87-6.18)</td>
</tr>
</tbody>
</table>

It is difficult to establish the cause of the relationship between pages and references, although it is clear that number of pages and number of references in a paper not only influence positively each other but they are influenced directly and indirectly by multiple factors (e.g. number of authors, see Abt and Garfield, 2002). Nevertheless, here we showed that the journal layout plays an essential role in defining this ratio, as this remains constant between all the OA EGU journals, independently on the research field, therefore substantially confirming the findings of Abt and Garfield (2002).

5 Conclusions

The importance of references in scientific journals has been clearly established. In this work we took advantage of the OA EGU journals, which, although they cover different areas in geoscience, share the same layout, thereby allowing for a direct comparison. It is shown that in the period 2010-2020, the number of pages and the number of references has been increasing in a statistically significant way.

Different reason could be behind this growth, such as open access to existing literature together with technological development which allows easier research of relevant citations. Additionally, we suggested this growth to be especially strong in EGU journals, being geophysics still a quite novel science with strong growing research and, consequently, strong growing of published literature, which tends to be more and more referenced in following studies.

Despite the increases in publication length and number of references in all journals since 2010, the reference density (i.e. number of references per page) has remained remarkably constant. In addition, no statistical difference in reference density can be observed in any of the journals. The average number of references per published page has been estimated based on all
Figure 2. Left: Box plot of numbers of references per page. The box represents the distribution quartiles (25% and 75%), the white lines are the medians, and the black dots the averages. The bars represent the 90% confidence levels. The acronym of the respective journal is listed on the bottom. The light blue area represents the overlap of the 25-75% quartiles range for all the journals. Middle: Probability density histograms of numbers of references per page for all the papers from all journals. Right: Box plot of numbers of references per page as on the left, but for all papers from all journals.

The published papers, which show that the optimal reference density is 3.82 references/page (1.87 – 6.18 at 90% confidence level). This work shows that the layout does influence the number of references per page, confirming previous work.

It has been shown that papers with high number of references tends to be cited more (Lovaglia, 1991); here we showed that these correlate with longer papers, suggesting that papers presenting the work in more detail and with greater availability of data or ideas tend to have greater impact on following literature. It is therefore important the papers to be as long as they need to be, with the authors able to publish as much details as needed.

Code availability. The code for analysis is available upon request to the contact author.

Competing interests. No competing interests are present.

Acknowledgements. The author acknowledge the constructive discussion with U.Pöschl
References


Milojević, S.: How are academic age, productivity and collaboration related to citing behavior of researchers?, PloS one, 7, e49176, 2012.


