AGU PUBLICATIONS

JGR

Journal of Geophysical Research: Biogeosciences

RESEARCH ARTICLE

10.1002/2017JG004277

Key Points:
- Data sharers in impact-ranked ecologists follow a long-tail distribution
- Influential ecologists with higher scientific impact are more likely to have had publish their data in public data sets
- Regional imbalances exist in flux data sharing, publications, and observation networks

Supporting Information:
- Supporting Information S1
- Table S1
- Table S2
- Table S3
- Data Set S1

Correspondence to:
B. Zhao,
zhaobin@fudan.edu.cn

Citation:
Dai, S.-Q., Li, H., Xiong, J., Ma, J., Guo, H.-Q., Xiao, X., & Zhao, B. (2018). Assessing the extent and impact of online data sharing in eddy covariance flux research. *Journal of Geophysical Research: Biogeosciences*, *123*. https://doi.org/10.1002/2017JG004277

Received 2 NOV 2017
Accepted 29 DEC 2017
Accepted article online 8 JAN 2018

Assessing the Extent and Impact of Online Data Sharing in Eddy Covariance Flux Research

Sheng-Qi Dai[1], Hong Li[1], Jun Xiong[1], Jun Ma[1], Hai-Qiang Guo[1], Xiangming Xiao[1,2], and Bin Zhao[1]

[1]Ministry of Education Key Laboratory for Biodiversity Science and Ecological Engineering, and Coastal Ecosystems Research Station of the Yangtze River Estuary, Fudan University, Shanghai, China, [2]Department of Microbiology and Plant, Biology, College of Arts and Sciences, University of Oklahoma, Norman, OK, USA

Abstract Research data sharing is appealing for its potential benefits on sharers' scientific impact and is also advocated by various policies. How do scientific benefits and policies correlate with practical ecological data sharing? In this study, we investigated data-sharing practices in eddy covariance flux research as a typical case. First, we collected researchers' data-sharing information from major observation networks. Then, we downloaded bibliometric data from the Web of Science and evaluated scientific impact using LeaderRank, a synthetic algorithm that takes both citation and cooperation impacts into consideration. Our results demonstrated the following: (1) specific to eddy covariance flux research, 8% of researchers published information in public data portals, whereas 64% of researchers provided their available data online in a downloadable form; (2) regional differences in data sharing, publications, and observation networks existed; and (3) the data sharers in impact-ranked ecologists followed a long-tail distribution, which suggested that, although sharing data is not necessary for researchers to be influential, data sharers are more likely to be high-impact researchers. Differentiated policies should be proposed to encourage ecologists in the long tail of data sharers, and from regions with little tradition of data sharing, to embrace a more open model of science.

1. Introduction

With multiple interdisciplinary, long-term, highly instrumented observation projects such as the Long Term Ecological Research Network, National Ecological Observatory Network, and FLUXNET, ecology has already become a data-intensive science (Goring et al., 2014; Hampton, Strasser, & Tewksbury, 2013; Heffernan et al., 2014). Large amounts of observation data are collected in near real time by field investigation, sensor networks, and space satellites (Porter, Hanson, & Lin, 2012). Massive observation data require professional analyses from ecologists to extract its scientific value. Thus, to make data easily accessible for each researcher, ecology is calling for a widely accepted data-sharing culture (Duke & Porter, 2013; Fecher, Friesike, & Hebing, 2015).

Based on the premise that data sharing contributes significantly to the advancement of ecology, many studies have been conducted to build a data-sharing culture (Peng et al., 2016; Reichman, Jones, & Schildhauer, 2011; Soranno et al., 2014). Funding agencies, journal publishers, policy makers, and research institutions are encouraging researchers to publish their observation data (Costello, 2009; Wallis, Rolando, & Borgman, 2013; Whitlock, 2011). In addition, more ecologists recognize that data sharing provides multiple benefits such as discouraging fraud, validating original research, building new hypotheses, educating students, accelerating scientific collaboration, and avoiding duplicate data collection (Fecher et al., 2015; Kim & Zhang, 2015).

However, the current levels of data sharing in ecology are not encouraging. The majority of ecologists interviewed said they were willing to share data, but the actual situation is not as optimistic (Huang et al., 2012). It was reported that only 8% of ecological projects that produced nongenetic data could be found online, and only an estimated 1% of ecological data are accessible after paper publication (Hampton et al., 2013; Reichman et al., 2011; Wolkovich, Regetz, & O'connor, 2012). As ecological data sets are typically collected by individuals or small groups, they are difficult to share for multiple reasons: First, the data sets tend to be small in volume, local in character, gathered for specific analyses, and obscure to unfamiliar researchers. Some of these data sets are unstructured data and even incompatible in long-term structuralized data management (Ferguson et al., 2014; Jarnevich et al., 2007; Tenopir et al., 2011). Second, although the majority of

scientific journal publishers expect data availability underlying published articles for future verification, there are few easy-to-use and widely accepted data repositories available (Enke et al., 2012; Michener, 2015). Third, similar to other experimental data sets, ecological data sets are gathered through years of fieldwork and expected to support multiple innovative papers. To maintain the privilege of first publication, data sets are viewed as being too previous to be shared immediately after collection, but this risks them never being published or shared and thus lost (Kenall, Harold, & Foote, 2014). Fourth, due to the standard deficiencies of data citation, the contributions of external data providers cannot be recognized. Without the guarantee of coauthorship, the scientific benefits of data sharing seem farfetched to potential sharers (Belter, 2014; Duke & Porter, 2013).

Career benefits such as scientific impact are one of the most important motivations in data sharing (Kim & Stanton, 2016; Michener, 2015; Cynthia S Parr & Cummings, 2005). However, ecologists need ethical principles to guarantee scientific impact returns after sharing their data (Duke & Porter, 2013). Recent discussions on data sharing and scientific impact in ecology were mostly based on questionnaire surveys and literature review studies. Kim and Zhang (2015) proposed a quantitative framework to study the drivers in data-sharing behaviors of science, technology, engineering, and mathematics researchers on the theory of planned behavior using a questionnaire survey. Their results emphasized that perceived career benefits such as scientific impact played a positive role in the attitudes toward data sharing. Similarly, Fecher et al. (2015) built another framework to explain drivers in academic data sharing through both literature review and surveys, which also showed that data sharers expected scientific impact returns if they choose to share data.

However, results from questionnaires, interviews, and surveys might shift due to the subjectivity of interviewees. To make this conclusion more objective and convincing, quantitative research is required. Open-access articles with data available have been widely shown to have greater citation impact (Craig et al., 2007; Eysenbach, 2006; Norris, Oppenheim, & Rowland, 2008). In addition, Heather A. Piwowar, Day, and Fridsma (2007) found that publicly available data were significantly associated with a 69% increase in citations. Furthermore, Heather A. Piwowar and Vision (2013) produced another analysis in microarray research data. They found that papers with data available in a public repository received 9% more citations than those with no data available, which would be a powerful incentive for data sharing. Similarly, in ecology, a quantitative survey on actual data activity is needed to discover who is sharing data and to determine the relationship between scientific impact and data sharing.

Investigating the status of data sharing in a specific research field was once difficult due to the scarcity of data and information retrieval techniques. Currently, parser-based Web crawlers can retrieve data from the Internet automatically. Regarding scientific impact evaluation, evaluating researchers using a single bibliometric factor like total citations has been shown to be inappropriate (Deng & Wang, 2015). Börner et al. (2005) redefined the edge weights in cooperation networks by citation rates and evaluated scientific impact using centrality, and Yan and Ding (2011) introduced a citation rate-weighted PageRank algorithm into cooperation networks to identify influential authors. A state-of-the-art bibliometric method known as the weighted LeaderRank algorithm was published to evaluate researchers quantitatively in a specific research field using both cooperation and citation impacts. This algorithm converged faster and was more robust than the classic PageRank algorithm (Li et al., 2014; Lü et al., 2011).

For the selection of typical research fields in ecology, eddy flux research was chosen. Together with multiple high-frequency environmental sensors, the eddy flux tower sites are equipped with eddy covariance methods to measure the exchange of carbon dioxide (CO_2), water vapor, and energy between terrestrial ecosystems and the atmosphere (Baldocchi, 2003). Data sets in eddy flux research are of large volume, great value, complex variety, and high-generating velocity, and these data sets are shared through observation networks. The observation networks, which include multiple observation sites as affiliates to reinforce communication and management, are important organizations in flux data sharing, and most data set and data-sharing policies are published by these networks. After over 25 years of development, eddy covariance flux research is the pioneer of data-sharing in modern ecology and deserves special attention.

In this article, we explored the data-sharing activities among ecologists in the eddy covariance flux research field and proposed a comprehensive framework to assess their scientific impact and detect data-sharing activity. The results were based on (a) a systematic bibliometric analysis of the full records of 5,654 research papers in eddy flux research during 1985–2016 from Thomson Reuters Web of Science and (b)

comprehensive flux site information retrieved from major databases using a Web crawler. We hypothesized a close correlation between scientific impact and data-sharing activities, aiming to discuss the regional differences in flux data-sharing polices to support future policy making in ecology.

2. Data and Methods

2.1. Data Sharer and Flux Researcher Definition

Eddy flux sites are tower-based infrastructures where the sites collect data sets covering a certain time period that can be regarded as typical units of data observation in ecology. Researchers who want to make their data sets public will join observation networks and publish their sites online, and thus, the information of flux sites and their investigators can be found online in major databases. Therefore, we defined all investigators and their teams whose names are listed online as participants in data sharing and tagged them as data sharers. Similarly, the authors involved in eddy flux publications together with all investigator teams were both defined as flux researchers.

Flux researchers were further classified into three categories: (1) for flux sites that published their observation data in public data sets or directly downloadable online forms (data set site year >0), their investigator teams were tagged as direct data sharers (DDS). (2) For sites that published metadata but no available observation data (data set site year = 0), corresponding investigator teams were tagged as metadata publishers (MDP). (3) For other researchers with no online information, they were tagged as no data available (NDA).

2.2. Flux Information Collection

Major observation networks including FLUXNET (FLUXNET, 2017), AmeriFlux (AmeriFlux, 2017b), AsiaFlux (AsiaFlux, 2017), European Fluxes Database Cluster (EFDC, 2017), and OzFlux (OzFlux, 2017) served as reliable data sources for flux site information. We deployed a Scrapy Web crawler (https://scrapy.org/, open source, Version 1.4) on 15 August 2017 to fetch information from the websites listed above, and URLs of the observation networks can be found in the supporting information. Then, BeautifulSoup (https://www.crummy.com/software/BeautifulSoup/, open source, Version 4.2.0) was used to transform webpage table data into structured text records. All information retrieval work was processed using Python (https://www.python.org/, open source, version 2.7.13).

2.3. Bibliometric Data

To evaluate the scientific impact of flux researchers, bibliometric information was required. We built the bibliometric database on 10 August 2017 using Clarivate Analytics' Web of Science. When compared to Google Scholar, Web of Science data may suffer from data incompleteness. However, when taking data robustness and feasibility into consideration, we still chose the Web of Science data set as a nonbiased sampling from all eddy flux-related papers. The search formula was as follows:

TS = ((eddy covariance OR (flux tower AND ecology)) OR (flux tower AND carbon cycling) OR (flux tower AND land atmosphere interaction) OR fluxnet OR ameriflux OR mexflux OR asiaflux OR japanflux or chinaflux or OzFlux) AND (PY = 1985–2016).

Using the formula above, 5,654 publications were identified in the Web of Science core database.

2.4. Data Analysis and Visualization

Duplicate researcher names existed in both the bibliometric list and the website investigator list. We unified these name expressions using an algorithm based on an automatic similarity check and a manual recheck.

Then, we deployed the researchers' scientific impact evaluation based on the weighted LeaderRank algorithm. Taking both citation and cooperation impact into account, this algorithm was shown to be more efficient in identifying influential authors in research networks than classical publication and citation count methods (Li et al., 2014). Detailed algorithm information is provided in the supporting information. Based on researchers' LeaderRank scores, statistical analysis and data visualization were implemented on 25 August 2017 using Python (ver. 2.7.13) and Microsoft Excel 2013. Original data, source codes and a detailed list of flux sites, researchers, and papers are included in the supporting information and can also be found at https://github.com/daishengqi/Data_Sharing_Impacts.

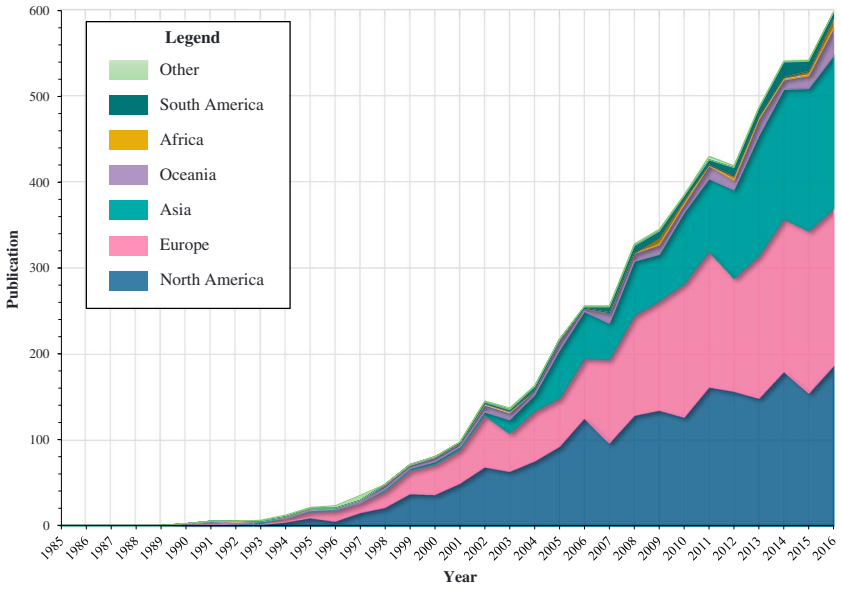

Figure 1. Annual publications by different regions from 1985 to 2016.

3. Results

3.1. Flux Sites and Bibliometric Records

From the Web pages, we detected a total of 930 public flux observation sites. In total, 407 sites (44%) had observation data directly available online, whereas 523 sites (56%) had no available data online. For staff, 680 investigators who managed these flux sites were detected, with 436 DDS (64%) and 244 MDP (36%).

A total of 5,654 research papers on flux research were collected, and a steady growth over the past 31 years of flux research papers is shown in Figure 1. With an annual publication increase of 2 to 598 since 1990, this indicates that more attention has been focused on flux research. The publications were dominated by North American and European researchers, with a total of 2,097 and 1,940 publications, respectively. Notably, although the first Asian paper was published in 1998, Asian flux research grew rapidly to an annual total of 178 in 2016.

3.2. Data Sharing and Scientific Impact Evaluation

Among the 9,628 unique researchers involved in both bibliometric analysis and website data sharing, there were 436 DDS (5%), 244 MDP (3%), and 8,948 researchers with NDA (92%). By directly defining DDS and MDP together as online data sharers, only 8% of authors published information on public data portals. It may be an underestimate of data sharers, because some researchers in NDA may represent modelers as data users; however, they did not collect the data and should not be regarded as potential data sharers. Therefore, 64%, which was the proportion of DDS in online data sharers (DDS/(DDS + MDP)), is the maximum proportion. It may be an overestimate but a fairer representation of data-sharing willingness in flux research.

For the scientific impact evaluations, multiple indicators were calculated, including total publication, total citation, citation score, and LeaderRank score. We chose LeaderRank score as the final indicator of impact evaluation, as researchers with higher LeaderRank scores gained higher positions in science impact ranking.

Notably, we found 21 data sharers (including both DDS and MDP) in the top 30 researchers and 69 in the top 100; only 2 MDP were in the bottom 100, which uncovered a probable relationship between high scientific impact and data-sharing actions. Further statistical analyses and visualization were implemented to test our hypothesis. First, the distribution of the different data-sharing groups in impact-ranked researchers was visualized, as shown in Figure 2.

In Figure 2a, the top 900 researchers were divided into 9 groups by descending order of LeaderRank score. Each group was arranged into a chromatic matrix with each pixel colored as the data-sharing type of the

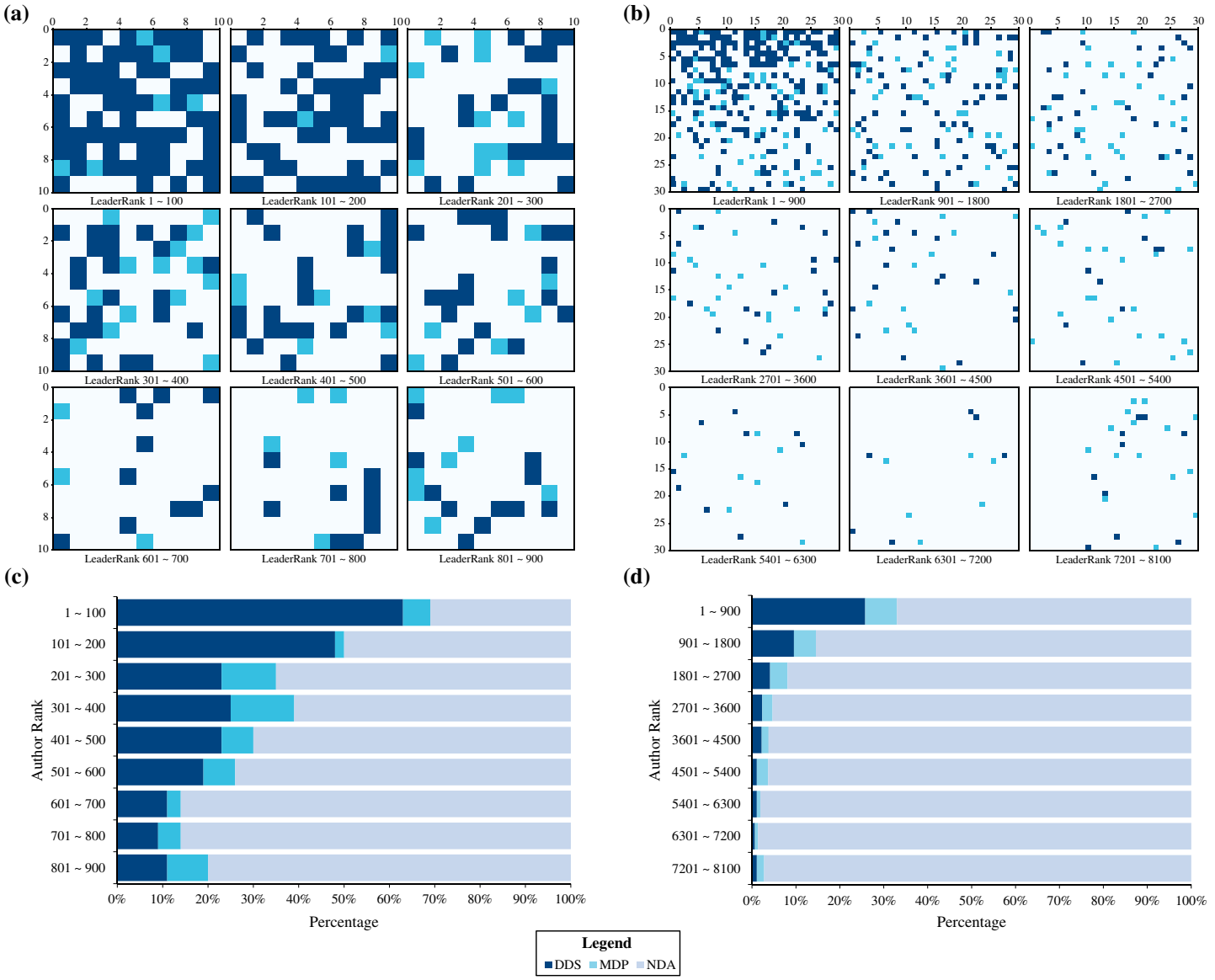

Figure 2. Distribution and proportion of different data-sharing groups of researchers. (a) Distribution of data-sharing groups in the top 900. (b) Distribution of data-sharing groups in the top 8,100. (c) Proportion of data-sharing groups in the top 900. (d) Proportion of data-sharing groups in the top 8,100.

corresponding author. The color brightness of each pixel represented the data-sharing group of the researcher, as shown in the legend. These nine matrices were mosaicked into a 3 * 3 figure in rank order. The same visualization was also deployed to the top 8,100 to further demonstrate the data-sharing distribution of additional researchers in Figure 2b. To better describe the distribution of data sharers, the proportion of different data-sharing groups in the top 900 and top 8,100 was also visualized by stacked percentage bar charts, as shown in Figure 1b.

It is clear that data sharers had a denser distribution in the top ranks than lower ranks. According to Figure 2d, the occupancy of data sharers was 33% in the top 900, and it dropped continuously to only 3% in the bottom 900. By taking a closer look at the top 900 researchers (Figure 2c), the same pattern appeared, with 69% of data sharers in the top 100; the occupancy dropped to only 20% in ranks 801 to 900. The majority of data sharers represented high-ranking researchers. Among the total of 664 data sharers in the top 8,100, 297 (45%) represented the top 900 researchers, which directly indicated a close correlation between scientific impact and data sharing.

4. Discussion

4.1. Policy-Driven Online Data Sharing in Flux Networks

Though FLUXNET collects global data and metadata, most flux sites and data sets are also managed by regional observation networks like AmeriFlux, European Fluxes Database Cluster, OzFlux, and AsiaFlux. Different regional networks prefer various affiliates, policies, and infrastructures, which creates regional differences in flux data sharing.

After over 20 years of development, FLUXNET has already constructed a mature data portal and established three data sets to serve the flux research community. The FLUXNET Marconi data set was the first data set released by FLUXNET, which compiled 97 site-years of gap-filled data and was completely open online. Without further limitations, only data citations were required when using this data set. The second data set released by FLUXNET was the La Thuile Synthesis data set, which consisted of 965 site-years of data from over 252 flux sites. As the successor of the Marconi data set, a detailed data policy was proposed to release different versions of La Thuile to data users, as follows: (a) complete data sets were only available to data contributors, (b) users could download "open" data sets designated by data contributors after submitting a proposal, and (c) free fair-use data sets were open to all users. Acknowledgements and data citation were always required when using any version of the La Thuile data set. Furthermore, the latest FLUXNET2015 data set included several improvements in data quality control and the processing pipeline. Based on its data policy, data sets were classified into two tiers: (a) Tier one data, complete open and free data, and (b) Tier two data, open but not free data, which meant that data producers must have opportunities to collaborate with data users. When using both tiers of data, submitting a download form, an intended-use statement, proper acknowledgements, and citations were all required. Apart from the improvement in data quality and data portals, the evolution of FLUXNET data policies represents a better attempt to benefit data producers and push flux data sets to the open source era.

AmeriFlux was the pioneer in flux data management and sharing for several reasons. First, all the flux data sets are downloadable through a user-friendly search interface after registration. Second, DOI citation is required when using data and easily fetched when citing data, while data use logs also serve as a good supervision in data downloading (AmeriFlux, 2017a). Third, an up-to-date publication list of papers that cite the corresponding data sets provides a great return for data sharers. Driven by strict AmeriFlux data policies and its mature management platform, researchers in North America have already built a virtuous continual data-sharing culture.

Similarly, European researchers constructed the European Fluxes Database Cluster (EFDC) to store and manage their flux data sets, which are yielded from multiple observation projects and networks. The EFDC is more like a database than a data-sharing portal. First, only data sets with public access policies are available to all users, and flux data sets in the EFDC are strictly archived and distributed to different user groups according to their authorization. Second, detailed documentation requirements for data formation and submission were carefully created and proposed. Third, the data availability is clear on each flux site, but no data use logs or publication lists are available on the website.

An additional network is OzFlux, which was supported by the Terrestrial Ecosystem Research Network and was once part of FLUXNET; thus, most information on its flux sites can be found in FLUXNET. Metadata on OzFlux websites are more descriptive by using text rather than the table-like information in AmeriFlux or EFDC, which functionally serve as data-publishing platforms rather than databases. In particular, a restricted access period of no more than 18 months after submission of the data to OzFlux can be used by collection owners to allow postgraduate researchers working with the data provider to conduct their research.

AsiaFlux mostly consists of ChinaFlux, JapanFlux, and KoFlux. With strict data fair use polices and detailed data uploading instructions, metadata of AsiaFlux sites are well arranged at the corresponding Web pages, but no data are connected by hyperlinks on site pages. AsiaFlux data sets are currently archived in the Japan global environmental database (https://db.cger.nies.go.jp/asiafluxdb), and a total of 36 flux data sets were published in a list. Users must submit a detailed data request form before registration and data download. JapanFlux and KoFlux provide similar metadata as AsiaFlux, and their public data sets are archived in the AsiaFlux database. However, the ChinaFlux website is different in that metadata of flux sites are described by text and grouped by observed ecosystem type instead of site code. Its data portal is only available in Chinese, and applications are required before data download. Additional efforts with the ChinaFlux website are

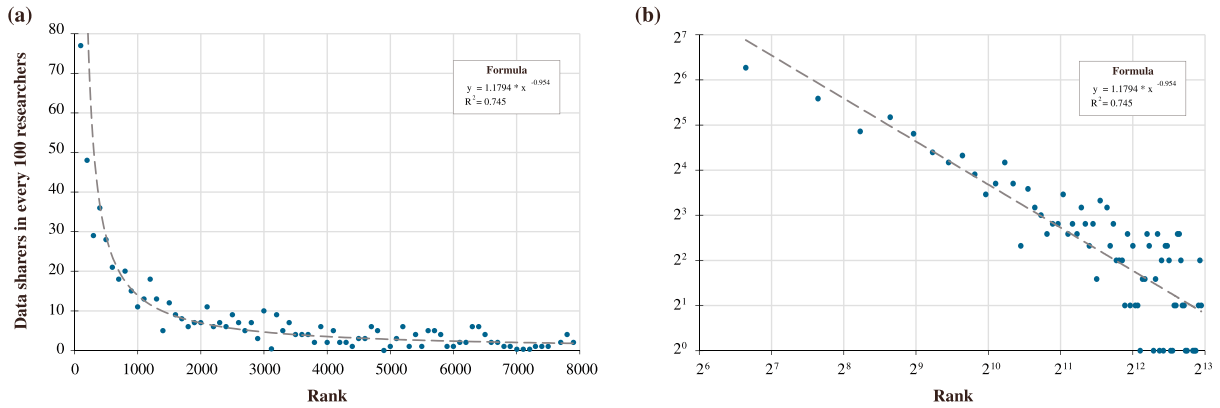

Figure 3. Distribution of data sharers in ecology researchers. (a) The power distribution of data sharers in the researcher rank list. (b) Log-log verification of the previous power distribution.

required to build a global and open data-sharing platform. The Chinese research community and individual researchers should think globally and act personally to promote a paradigm of open, free, and timely data sharing (Peng et al., 2016).

4.2. "Long-Tail" Distribution of Flux Data Sharers

The distributions of a wide variety of physical, biological, and man-made phenomena approximately follow a power law over a wide range of magnitudes, such as the Pareto distribution and the Zipfian distribution. Similarly, in this study, we found that the distribution of data sharers in flux researchers also followed this pattern, as shown in Figure 3a. To verify this power distribution, a log-log linear regression was deployed and is shown in Figure 3b.

According to Figure 2, 680 researchers chose to share flux data sets or metadata online, and the majority of these data sets were of high impact. Further, the result of Figure 3 indicates a strong correlation between data sharing and individual scientific impact, while the causality requires further exploration. Thus, we suggest that, although sharing data is not necessary for researchers to be influential, data sharers are more likely to be high-impact researchers. This conclusion can be interpreted in multiple ways: First, as a top-down action, flux data sharing is mainly advocated by well-known scientists, publishers, and governments. In order to convince other data holders to share their data, leading scientists must act as data-sharing pioneers. Second, data sharing has been shown to aid citation rates (Heather A. Piwowar & Vision, 2013). Similarly, active data sharers in flux research might attract more attention by publishing their data sets online. Third, multiple studies have shown that authors were most likely to share data if they had prior experience sharing or reusing data (Fecher et al., 2015; Heather A Piwowar, 2011; Wolkovich et al., 2012). Equipped with mature data infrastructures, sufficient data quantity, and good data quality, it is more feasible for leading ecologists to publish their data. Fourth, Kim and Stanton (2016) indicated that perceived career risk at an individual level was not found to have any significant relationship with data-sharing behaviors, but our results may demonstrate that exclusive data mining privilege was less crucial to leading ecologists than average ecologists. For leading ecologists, multiple papers have been published before sharing their data, which minimizes the career risk when they publish data.

4.3. Moving Toward a Better Data-Sharing Future

A sustainable data-sharing culture will likely be built as ecology enters the age of big data. As a human enterprise, ecological data sharing should be guided by effective policies (Cynthia Sims Parr, 2007). These policies are based on reliable and comprehensive investigation results, intend to resolve critical problems in actual data-sharing practices and call for the efforts from not only researchers but also institutes, publishers, and governments.

The results presented in this paper can help us propose the following suggestions: First, regional differences in flux data sharing are prominent. Thus, the strategies in different regions should differ. North American flux

researchers have already built an integrated data platform; thus, future work should focus on refining the data sets and improving the usability of this system, as data management and organization support greatly affect the willingness of data sharing and data reusability (Sayogo & Pardo, 2013). For European flux researchers, the EFDC has an excellent infrastructure but needs some improvements. Attaching data use logs and corresponding publications to data sets or flux sites and publishing them as metadata will directly benefit data sharing. For Asian researchers, the priority is refining Asiaflux to be an easy-to-use data-sharing infrastructure, as this integrated data platform calls for closer cooperation among Chinaflux, Japanflux, and KoFlux. Mandatory data policies established by both publishers and governments are required to persuade more researchers in Asia to publish their data sets.

Second, there was an imbalanced distribution of data sharers in researchers. The average flux researchers were mostly composed of small research groups, individual researchers, or interdisciplinary researchers who were located far from core scientific cooperation networks. They represented the majority of researchers, but few of these individuals shared data, which indicated that the efforts or perceived risks of data sharing still overweigh the potential benefits for average researchers. Both easy-to-use data platforms and standardized data policies serve as stimuli to involve these groups in data sharing. As long as academia remains publication centered, first publication privilege, the promise of authority assignment, and the publicity of data use that reduce career risks are crucial for small groups and individuals to promote data sharing. Peer-reviewed data will be a good choice in publishing primary observation data (Huang, Hawkins, & Qiao, 2013).

5. Conclusion

In this study, we investigated flux research by evaluating scientific impact and public data publishing to discover the underlying patterns in data sharing. First, 8% of flux researchers published information on public portals, and 64% of these researchers made flux data downloadable; the willingness of sharing in flux data sharers was appreciated. Second, large regional imbalances in data sharing still existed in flux research. North American and European researchers did better in data sharing, while Asian researchers still had improvements to make regarding refining infrastructures. Last but not least, higher-impact scientists tended to make their data sets more easily accessible. Although sharing data is not necessary for researchers to be influential, data sharers are more likely to be high-impact researchers. Moreover, the distribution of data sharers in ranked researchers followed a classical "long-tail" pattern. We suggest that more advantageous polices from governments and scientific organizations should be proposed to encourage average ecologists with relative low scientific impact to share their data.

More research is needed to better understand data sharing in the future. The long-tail distribution of data sharers we introduced is based on a typical case study in flux research. The causality between data sharing and scientific impacts remains unclear, and the results will be more robust if survey data that cover data sharers and top researchers are acquired. Particularly, offline data communication and private data holders are not included in this study. Together with accurate data reuse data, the map of data flow can be determined to best evaluate the contribution of data sharing in ecology.

Acknowledgments

Detail source codes of the Web crawler, name unification, bibliometric analysis, data visualization, and improved weighted LeaderRank algorithm are available on Github at https://github.com/daishengqi/Data_Sharing_Impacts. Original data including the related Web pages and bibliometric data can also be found in the supporting information or at http://www.mediafire.com/file/7e4nlshpsp3z4qa/Original_Data_Upload.rar. This research was financially supported by the National Natural Science Foundation of China (31170450) and Shanghai Science and Technology Innovation Action Plan (13JC1400400 and 13231203503).

References

AmeriFlux (2017a). AmeriFlux data policy—For users and contributors. Retrieved from http://ameriflux.lbl.gov/data/data-policy/

AmeriFlux (2017b). AmeriFlux: Measuring carbon, water and energy flux across the Americas. Retrieved from http://ameriflux.lbl.gov/

AsiaFlux (2017). AsiaFlux. Retrieved from http://asiaflux.net/

Baldocchi, D. D. (2003). Assessing the eddy covariance technique for evaluating carbon dioxide exchange rates of ecosystems: Past, present and future. *Global Change Biology*, *9*(4), 479–492. https://doi.org/10.1046/j.1365-2486.2003.00629.x

Belter, C. W. (2014). Measuring the value of research data: A citation analysis of oceanographic data sets. *PLoS One*, *9*(3), e92590. https://doi.org/10.1371/journal.pone.0092590

Börner, K., Dall'Asta, L., Ke, W., & Vespignani, A. (2005). Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams. *Complexity*, *10*(4), 57–67. https://doi.org/10.1002/cplx.20078

Costello, M. J. (2009). Motivating online publication of data. *Bioscience*, *59*(5), 418–427. https://doi.org/10.1525/bio.2009.59.5.9

Craig, I. D., Plume, A. M., McVeigh, M. E., Pringle, J., & Amin, M. (2007). Do open access articles have greater citation impact?: A critical review of the literature. *Journal of Informetrics*, *1*(3), 239–248. https://doi.org/10.1016/j.joi.2007.04.001

Deng, Q., & Wang, X. (2015). Identifying influential authors based on LeaderRank. New Technology of Library and Information Service(09), pp. 60–67. https://doi.org/10.11925/infotech.1003-3513.2015.09.09

Duke, C. S., & Porter, J. H. (2013). The ethics of data sharing and reuse in biology. *Bioscience*, *63*(6), 483–489. https://doi.org/10.1525/bio.2013.63.6.10

EFDC (2017). European fluxes database cluster. Retrieved from http://gaia.agraria.unitus.it/
Enke, N., Thessen, A., Bach, K., Bendix, J., Seeger, B., & Gemeinholzer, B. (2012). The user's view on biodiversity data sharing—Investigating facts of acceptance and requirements to realize a sustainable use of research data. *Ecological Informatics*, *11*, 25–33. https://doi.org/10.1016/j.ecoinf.2012.03.004
Eysenbach, G. (2006). Citation advantage of open access articles. *PLoS Biology*, *4*(5), e157. https://doi.org/10.1371/journal.pbio.0040157
Fecher, B., Friesike, S., & Hebing, M. (2015). What drives academic data sharing? *PLoS One*, *10*(2), e0118053. https://doi.org/10.1371/journal.pone.0118053
Ferguson, A. R., Nielson, J. L., Cragin, M. H., Bandrowski, A. E., & Martone, M. E. (2014). Big data from small data: Data-sharing in the 'long tail' of neuroscience. *Nature Neuroscience*, *17*(11), 1442–1447. https://doi.org/10.1038/nn.3838
FLUXNET (2017). Fluxdata. Retrieved from http://fluxnet.fluxdata.org/
Goring, S. J., Weathers, K. C., Dodds, W. K., Soranno, P. A., Sweet, L. C., Cheruvelil, K. S., … Utz, R. M. (2014). Improving the culture of interdisciplinary collaboration in ecology by expanding measures of success. *Frontiers in Ecology and the Environment*, *12*(1), 39–47. https://doi.org/10.1890/120370
Hampton, S. E., Strasser, C. A., & Tewksbury, J. J. (2013). Growing pains for ecology in the twenty-first century. *Bioscience*, *63*(2), 69–71. https://doi.org/10.1525/bio.2013.63.2.2
Heffernan, J. B., Soranno, P. A., Angilletta, M. J., Buckley, L. B., Gruner, D. S., Keitt, T. H., … Xiao, J. (2014). Macrosystems ecology: Understanding ecological patterns and processes at continental scales. *Frontiers in Ecology and the Environment*, *12*(1), 5–14. https://doi.org/10.1890/130017
Huang, X., Hawkins, B. A., Lei, F., Miller, G. L., Favret, C., Zhang, R., & Qiao, G. (2012). Willing or unwilling to share primary biodiversity data: Results and implications of an international survey. *Conservation Letters*, *5*(5), 399–406. https://doi.org/10.1111/j.1755-263X.2012.00259.x
Huang, X., Hawkins, B. A., & Qiao, G. (2013). Biodiversity data sharing: Will peer-reviewed data papers work? *Bioscience*, *63*(1), 5–6. https://doi.org/10.1525/bio.2013.63.1.2
Jarnevich, C. S., Graham, J. J., Newman, G. J., Crall, A. W., & Stohlgren, T. J. (2007). Balancing data sharing requirements for analyses with data sensitivity. *Biological Invasions*, *9*(5), 597–599. https://doi.org/10.1007/s10530-006-9042-4
Kenall, A., Harold, S., & Foote, C. (2014). An open future for ecological and evolutionary data? *BMC Ecology*, *14*(1), 10. https://doi.org/10.1186/1472-6785-14-10
Kim, Y., & Stanton, J. M. (2016). Institutional and individual factors affecting scientists' data-sharing behaviors: A multilevel analysis. *Journal of the Association for Information Science and Technology*, *67*(4), 776–799. https://doi.org/10.1002/asi.23424
Kim, Y., & Zhang, P. (2015). Understanding data sharing behaviors of STEM researchers: The roles of attitudes, norms, and data repositories. *Library and Information Science Research*, *37*(3), 189–200. https://doi.org/10.1016/j.lisr.2015.04.006
Li, Q., Zhou, T., Lü, L., & Chen, D. (2014). Identifying influential spreaders by weighted LeaderRank. *Physica A: Statistical Mechanics and its Applications*, *404*, 47–55. https://doi.org/10.1016/j.physa.2014.02.041
Lü, L., Zhang, Y.-C., Yeung, C. H., & Zhou, T. (2011). Leaders in social networks, the delicious case. *PLoS One*, *6*(6), e21202. https://doi.org/10.1371/journal.pone.0021202
Michener, W. K. (2015). Ecological data sharing. *Ecological Informatics*, *29*, 33–44. https://doi.org/10.1016/j.ecoinf.2015.06.010
Norris, M., Oppenheim, C., & Rowland, F. (2008). The citation advantage of open-access articles. *Journal of the Association for Information Science and Technology*, *59*(12), 1963–1972. https://doi.org/10.1002/asi.20898
OzFlux (2017). OzFlux - Home. Retrieved from http://www.ozflux.org.au/
Parr, C. S. (2007). Open sourcing ecological data. *Bioscience*, *57*(4), 309–310. https://doi.org/10.1641/B570402
Parr, C. S., & Cummings, M. P. (2005). Data sharing in ecology and evolution. *Trends in Ecology & Evolution*, *20*(7), 362–363. https://doi.org/10.1016/j.tree.2005.04.023
Peng, C., Song, X., Jiang, H., Zhu, Q., Chen, H., Kim, Y., & Stanton, J. M. (2016). Towards a paradigm for open and free sharing of scientific data on global change science in China. *Ecosystem Health and Sustainability*, *67*(4), 776. https://doi.org/10.1002/ehs2.1225
Piwowar, H. A. (2011). Who shares? Who doesn't? Factors associated with openly archiving raw research data. *PLoS One*, *6*(7), e18657. https://doi.org/10.1371/journal.pone.0018657
Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. *PLoS One*, *2*(3), e308. https://doi.org/10.1371/journal.pone.0000308
Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, *1*, e175. https://doi.org/10.7717/peerj.175
Porter, J. H., Hanson, P. C., & Lin, C.-C. (2012). Staying afloat in the sensor data deluge. *Trends in Ecology & Evolution*, *27*(2), 121–129. https://doi.org/10.1016/j.tree.2011.11.009
Reichman, O. J., Jones, M. B., & Schildhauer, M. P. (2011). Challenges and opportunities of open data in ecology. *Science*, *331*(6018), 703–705. https://doi.org/10.1126/science.1197962
Sayogo, D. S., & Pardo, T. A. (2013). Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data. *Government Information Quarterly*, *30*, S19–S31. https://doi.org/10.1016/j.giq.2012.06.011
Soranno, P. A., Cheruvelil, K. S., Elliott, K. C., & Montgomery, G. M. (2014). It's good to share: Why environmental scientists' ethics are out of date. *Bioscience*, *65*(1), 69–73. https://doi.org/10.1093/biosci/biu169
Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., … Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLoS One*, *6*(6), e21101. https://doi.org/10.1371/journal.pone.0021101
Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLoS One*, *8*(7), e67332. https://doi.org/10.1371/journal.pone.0067332
Whitlock, M. C. (2011). Data archiving in ecology and evolution: Best practices. *Trends in Ecology & Evolution*, *26*(2), 61–65. https://doi.org/10.1016/j.tree.2010.11.006
Wolkovich, E. M., Regetz, J., & O'connor, M. I. (2012). Advances in global change research require open science by individual researchers. *Global Change Biology*, *18*(7), 2102–2110. https://doi.org/10.1111/j.1365-2486.2012.02693.x
Yan, E., & Ding, Y. (2011). Discovering author impact: A PageRank perspective. *Information Processing and Management*, *47*(1), 125–134. https://doi.org/10.1016/j.ipm.2010.05.002