Peer comparison of XSEDE and NCAR publication data

Gregor von Laszewski, Fugang Wang, Geoffrey C. Fox
School of Informatics and Computing
Indiana University
Bloomington, Indiana, U.S.A.
laszewski@gmail.com

David L. Hart
Computational and Information Systems Laboratory
NCAR
Boulder CO 80307-3000

Thomas R. Furlani, Robert L. DeLeon, Steven M. Gallo
Center for Computational Research
University at Buffalo, SUNY
Buffalo, New York, 14203

Abstract—We present a framework that compares the publication impact based on a comprehensive peer analysis of papers produced by scientists using XSEDE and NCAR resources. The analysis is introducing a percentile ranking of citations of the XSEDE and NCAR papers compared to peer publications in the same journal that do not use these resources. This analysis is unique in that it is a comprehensive study in which all reported published papers are compered to peer publications selected from within the same issue of the same journal. From this analysis, we can see that papers that utilize XSEDE and NCAR resources are cited statistically significantly more often. Hence we find that reported publications indicate that XSEDE and NCAR resources exert a strong positive impact on scientific research.

I. INTRODUCTION

To identify the impact on *scientific advancements enabled by enhanced cyberinfrastructure*, it is important to conduct a comprehensive analysis of achievements that can be attributed to the use of advanced infrastructure, such as provided by the Extreme Science and Discovery Environment (XSEDE) [1], [2]. Many recent science and engineering innovations and discoveries are increasingly dependent on access to high-end computing resources [?]. The demand for high-end resources is met by large-scale compute resources located in geographically dispersed locations that cannot typically be supported by any single research group. Accordingly, dedicated large-scale computing facilities play an important role in scientific research, in which resources are shared among groups of researchers, while the facilities themselves are managed by dedicated staff. The National Science Foundation (NSF) and the Department of Energy have supported such facilities for many years. XSEDE allocates resources to approved projects, which represent a substantial financial investment by NSF. Thus, justification for their use is warranted and questions regarding the scientific impact of these resources naturally arise. In previous work [?] we focused on the creation of a framework that collects bibliometric data and analyzes them with respect to a number of metrics. However, this work did not yet include a mechanism to compare publications with peers not using such resources.

In this paper, we significantly enhance our previous work by comparing publication impact based on a comprehensive peer analysis carried out on papers produced by scientists using XSEDE and National Center for Atmospheric Research (NCAR) resources. The analysis is based on a percentile ranking of citations of the papers derived from a comparison to peer publications in the same journal not using the resources. This analysis is unique in that it is a comprehensive study in which thousands of published papers are are compered to peer publications selected from within the same issue of the same journal. From this analysis we can see that papers that utilize XSEDE and NCAR resources are cited statistically significantly more often.

The paper is structured as follows. First, we review some portions of our previous work and relate it to the work reported in this paper (Section II). We review our design and the architecture of our framework supporting this effort (Section III). Next we introduce a journal-based peer metric that allows us to compare any resource provider's related publications with publications not using the resources (Section IV). To demonstrate general applicability of this method, we introduce in the next section a peer analysis of XSEDE (Section V) and NCAR data (Section VI). We present important statistics about this metric for both XSEDE and NCAR. Finally, we present our conclusions (Section VII).

II. RELATED WORK

Although a number of related studies have been conducted [3], [4], [5], [6], [7] our work is unique in that it provides a *comprehensive* analysis superior in data volume to other studies and is focused on the analysis of XSEDE and NCAR data. More information about related work can be found in [?]. We are happy to engage in collaborative efforts to enhance this work or to integrate other related work by contacting us.

III. Bibliometric Data Analysis Framework for Resource Providers

The work described in this paper has been motivated by analyzing data related to XSEDE[1]. However, as many other resource providers may have similar needs to analyze their data, this framework is general enough that it can be adapted for use by other resource providers. In case this is desired, a custom integration can be performed and the analysis can be adapted by the team from Indiana University.

A. Requirements

The motivation for the framework includes three main objectives. First, our framework must be able to compare the impact of the use of resources on their users. Second, we have to define suitable metrics that allow such a comparison. Third, we need to provide a framework that provides an integrated set of data services on which these metrics can be applied to conduct a comparative impact analysis. This is depicted in Figure 1 on the left-hand side.

B. Design

For XSEDE we have identified specific design criteria that we are integrating into our architecture. Many of these design decisions will overlap with other resource providers. Certainly our design needs to be able to provide a comparative impact study for peers and also the funding agencies of the peers in conjunction with meaningful metrics. The metrics must be not only targeting the resource provider but also the community. Specifically for XSEDE interested parties in this analysis are not only the users, but also the Resource Allocation Committee (RAC) and the XSEDE leadership. The bibliometric data services and the data mashup are largely hidden from these groups. The groups benefit form a number of preconfigured analysis that may be further customized or enhanced. To show generality of this approach we have taken an analysis and applied it not only to XSEDE as a resource provider, but also to National Center for Atmospheric Research (NCAR) as a resource provider [8].

C. Architecture

The requirements for XSEDE and our user community have resulted in a layered architecture as depicted in 2. The framework is based on a distributed set of services. The service-oriented system consists of components for (a) publication and citation data retrieval (e.g., from the NSF award database, Google Scholar, and ISI Web of Science); (b) parsing and processing while correlating data from various databases and services, such as the XSEDE central database (XDCDB), which stores all usage data

[1]The XSEDE data set includes publications from both XSEDE and its predecessor program, TeraGrid. For simplicity, we refer to *XSEDE* throughout this paper.
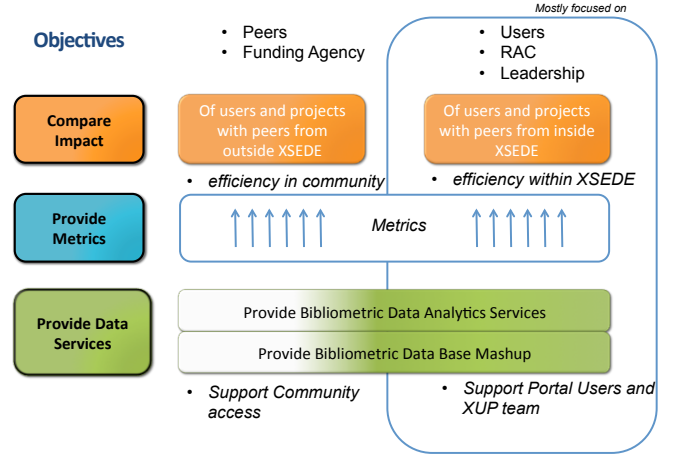

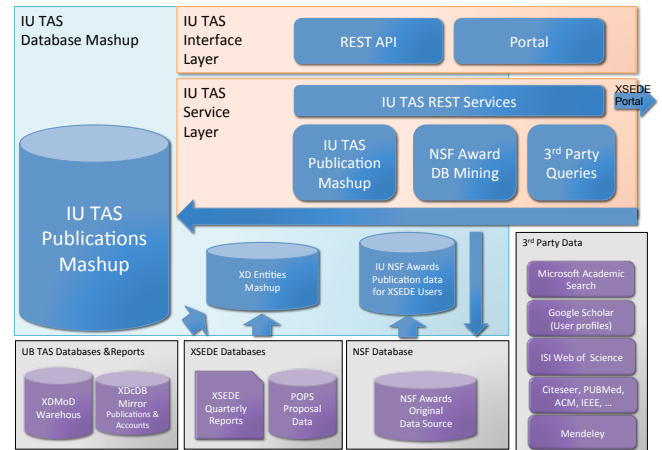Figure 1: High level Objectives impacting the design of our framework

Figure 2: Architecture with the Interface, Service, Database Mashup, and Database Resources Layers

for jobs run on XSEDE resources; and (c) the XSEDE allocations database, which stores publication and grant funding information for PIs applying for XSEDE allocations. The system also includes components for metrics generation and an analysis system for different aggregation levels – users, projects, organization, field of science (FOS) – as well as a presentation layer using a lightweight portal in addition to exposing some data via a RESTful API [**?**]. The main layers include (a) *Interface Layer* – easy to use interfaces for various communities, including an API, REST, and a Web GUI interface; (b) *Service Layer* – advanced services to bridge to a database mashup via sophisticated services and queries to the underlying database layer; (c) *Database Mashup Service Layer* – a sophisticated database mashup that contains the integration of data from a variety of data resources; and (d) the *Database Resources* that provide the underlying information for our service.

Due to this approach, our framework is expandable as we are able to integrate not only resources relevant for XSEDE,

but also other resource providers such as NCAR. However, we need to customize this integration and allow relevant services to be exposed to the framework. In some cases it could be as simple as replacing the database with a minor adaptation in the service layer.

Our current framework for XSEDE includes specialized services focusing on XSEDE user specific publication data as well as user, project, and FOS views. The mashup component aggregates the publication data mined from the previous component, in addition to those from XDCDB, and from other available external services. It also retrieves citation data for each publication from external services. Another essential task of this framework is to generate metrics for users, projects, and FOS, in which the XSEDE allocations database is involved to get proposal and project data. These data will be stored in the mashup database, which can then be integrated into the XDMoD system [9] by our partners at the University of Buffalo.

To conduct the analysis, the general workflow includes obtaining the publication data for each XSEDE user, and then retrieving the citation data for each publication. Hence, the data is originally collected on a per-user and per-publication basis. As part of processing, the data are aggregated based on organization, XSEDE project/account, and FOS. By correlating the publication data with XSEDE usage and allocation data (for example the allocation amounts awarded by XSEDE), our intention is to determine if the analysis reveals patterns and trends in how XSEDE impacts the sciences and possibly to help better measure the return on investment (ROI) for NSF. Naturally, the same is applied to the NCAR data to demonstrate the generality of our methods.

IV. METRIC FOR JOURNAL PUBLICATION-BASED PEER COMPARISON

While we focused in our previous work more on the internal analysis of data within XSEDE in regards to FOS, h-index [10], g-index [11] and i-index [12], in this work we extend the scope to a comparison with external peer publications and hence significantly expand our original analysis. To conduct this analysis, we introduce the definition of a metric that allows us to analyze and compare journal publications amongst peers, as well as identify a suitable peer group for the comparison.

Hence we considered for this work **all** publications that are uploaded through self-identification by users into the XSEDE portal, as well as XSEDE reports. We then identified from this set all *all* journal publications and identified from that subset all journals that had at least ten self-identified publications from XSEDE. Although the number to meet this threshold is small, we found the restriction useful as it defines a peer group of scientists that publish repeatedly in these journals, thus making the comparison more meaningful. Once we identified such journals, we compared the

citation count for each publication located in a journal issue that had at least one XSEDE publication. Our comparison is between publications in such journals that we identified as XSEDE papers and those that were not.

Next we introduce the percentile ranking of citations as the metric for the comparison of XSEDE publications with their non-XSEDE peers. The percentile ranking is based on the sum of all citations for a paper while at the same time ranking that number in four uniform percentile categories. Thus the papers in the first percentile have the most citations, the ones in the second have fewer, and so forth. We add the weighted sums of these counts. Thus the performance score is defined as:

$$S = 1 * P_{Q_1} + 0.5 * P_{Q_2} + (-0.5) * P_{Q_3} + (-1) * P_{Q_4}$$

In which, P_{Q_i} is the percentage of pubs falling into the top i quarter. P_{Q_i} gets the value in $[0, 1]$ and $\sum_i P_{Q_i} = 1$ for one FOS.

It is trivial to see that S has its value from $[-1, 1]$. A positive value implies more publications appear in the upper half in ranking and negative means more in the lower half.

V. XSEDE PEER DATA ANALYSIS

To apply the percentile ranking to the field of science of XSEDE/TG publications among the journal issues where each publication was published, we aggregate them based on FOS, according to the self-reported categories defined in the XDCDB, and calculate the average and median percentile ranking for each field of science, as well as the resulting *performance score*. We include only those with at least ten[2] publications so the results have a higher statistical power.

To identify the FOS for each publication, we followed this process:

1) Find the FOS information out of the past XSEDE/TG quarterly reports as this information may have explicitly been associated with them.
2) Find the FOS information from the project data in the XDCDB. Unfortunately, it is possible that one project is associated with more than one FOS. In such cases we counted the publications of the project for all involved FOSs.
   a) Some publications from XSEDE/TG quarterly reports were identified only by the project proposal number. We mapped them to the project charge number and account id used internally within the XSEDE central database.
   b) For user uploaded publications data via the XSEDE user portal, a project charge number was associated with each publication.

Through this data mashup we obtained enough data to conduct our analysis. We present our results in a number of

---

[2]For NCAR data, we used a value of five due to the smaller number of overall publications

graphs and tables. We first present some overall comparisons of the XSEDE and non-XSEDE citation comparisons. Figure 3 shows the kernel density of the distributions of XSEDE publications' percentile ranking and that of peers'. As expected, the non-XSEDE peer publications are symmetrically distributed by percentile ranking. The XSEDE publications are weighted to the higher percentile ranking. *This shows that XSEDE publications tend to be more highly cited.* Table I lists the average and median rankings and citations received of the two groups to evaluate and compare. We performed a T-test to test if the citation differences were statistically significant. The results show that the XSEDE group has a statistically higher citation ranking and a statistically higher mean citation rate than the non-XSEDE peer group.

1) T-test for ranking (Welch Two sample t-test)
   a) T=21.4134, df=2412.99, p-value¡2.2e-16
   b) 95% confidence interval: $[10.80, 12.98]$
2) T-test for citation count:
   a) T=7.057, df=2358.929, p-value=2.228e-12
   b) 95% confidence interval: $[9.40, 16.63]$

Table I: Basic statistics of XSEDE publications group and peers group

	Number of Publications	Rank Average	Rank Median	Citations Average	Citations Median
XD	2349	61	65	26	11
Peers	168422	49	48	13	5

The results are depicted in Figures 6, 7, 8, 4, and 5.

NSF and XSEDE define a hierarchy of FOSs. In Figure 5, we show the top level FOS as defined by NSF. When we look to expand the FOS to the next level in the hierarchy, we find the results as depicted in Figure 8. The next level is shown in Figure 7. Each of these figures shows the list of FOSs in decreasing order by the performance score S.

From Figure 7 we can identify that for most fields of science the XSEDE publications performed better than their peers. The average and median scores were higher than 50 and the score is positive. When looking at individual results, we see that astronomy and physics benefit most from using XSEDE/TG. When looking at the fields that perform worst, we find fields such as Experimental and Theoretical Geochemistry, Geometric Analysis and Mechanical and Structural systems. Such fields are typically not dominated by simulation science and are less dependent on computational resources. Other fields such as Training include many other areas of training outside of supercomputing usage. We even find fields such as Computer and Computation Research to be less impacted. We certainly have to acknowledge in this case that many theoretical papers and papers not using supercomputers are published.

To show the percentile distribution in more detail for each journal, we present in Figure 6 a stacked barchart. Also here, as expected from our previous results, we see

a positive impact in the percentile citation count for most of the journals. This is made obvious by looking at Figure 4. Here we also have included the average of the top 66 journals in which we found 10 or more XSEDE publications and find that the score metric is positive with a value of 0.284. Thus we can conclude that papers benefit from use of XSEDE resources, based on self-identification in reports and bibliographic upload to the XSEDE portal, resulting in their being more cited on average than their peers from the same journal.

VI. NCAR Peer Data Analysis

We have replicated the publication citation analysis with a text file of 880 publications obtained from NCAR. Because the NCAR data set was smaller than the one from XSEDE, instead of looking at all journals that have at least ten publications, we reduced the value to five. This will give a large enough journal number to conduct our analysis in this case. To conduct the analysis, we took the following steps.

1) We parsed the text file containing 880 publications into a structured database while identifying titles, and DOIs.
2) With our framework, we queried ISI Web of Knowledge to get detailed information on the publications – journal name, issue, citation data and other identifying information. We were able to verify and obtain data for 813 publications.
3) Through our framework, we identified and obtained peer data based on journal issue information. In total, 130 different publication venues have been identified. To ensure the results are statistically meaningful, we eliminated those with less than five publications appearing in them. This leads to 35 different journals that cover 653 publications. Out of the 35 journals, we had peers data for 12 of them (from our XSEDE peers comparison work). For the other 23 journals, we obtained peer publication data for an additional 39,000 publications.
4) For each NCAR publication, we computed the percentile ranking among peers within the same journal issue against all publication entries as obtained from ISI Web of Knowledge.
5) We grouped the percentile ranking scores based on each journal, divided the scores into each ranking quarters (Q_1: top 25%; top Q_2: 25% 50%; bottom Q_3: 50%-75%, bottom Q_4: 75%-100%), and computed the percentage for those that fall into each ranking quarter.

We depict the result in a stacked column chart similar to that used in Figure 6. In Figure 9, we show the distribution of the publications in the top 10 journals sorted by the number of publications. The top 10 journals total 484 publications. They account for about three quarters of the 653 publications from the 35 journals with at least five publications appearing
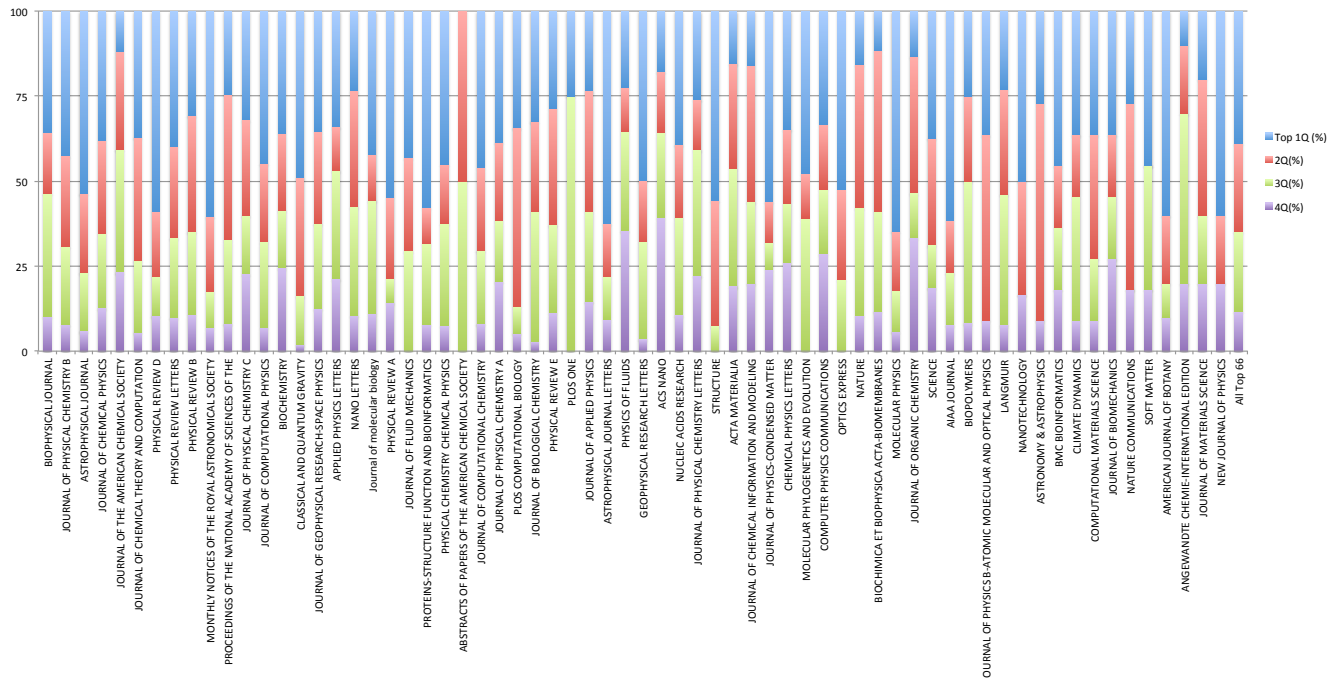
Figure 6: Percentile ranging by FOS in a stacked barchart of XSEDE publications.

in them; or 60% of the 813 publications from the 130 journals we identified via the ISI source (see Table II).

Table II: Top ten journals with the most publicationsns of NCAR publication data

Publication Venue	Publication Count
Journal of Climate	110
Journal of Geophysical Research-Atmospheres	72
Geophysical Research Letters	66
Atmospheric Chemistry and Physics	56
Monthly Weather Review	53
Journal of the Atmospheric Sciences	42
Climate Dynamics	35
Journal of Geophysical Research-Space Physics	19
Journal of Geophysical Research-Oceans	16
Journal of Hydrometeorology	15

We also computed and compared the *performance score* as defined earlier. The result is shown in Figure 11. Here we see that the average over all 35 entries is a positive value of about 0.35.

VII. CONCLUSION

We observe that the NCAR score is slightly higher than that of XSEDE. However, we need to take into account that XSEDE publications have a wider range of FOS; NCAR focusses on a much narrower FOS. Furthermore, as we can see from the XSEDE results, research in the atmospheric sciences resulted in the highest score values using XSEDE resources. Thus we can understand that the value is higher for NCAR as they do not include the other disciplines that have lower values.

However, the most important conclusion we make for our metric is that, for both XSEDE and NCAR publications, the impact measured by a percentile score is positive and higher than their peers that have not used such resources.

We will further expand upon our metric and can define such values also for other groups defined in XSEDE or for other resource providers. The important information we need is a reliable source that identifies publications that can be associated with the resource. While exploring not only XSEDE, but also NCAR data we have in principle shown that our approach is applicable to other resource providers.

VIII. ACKNOWLEDGMENTS

REFERENCES

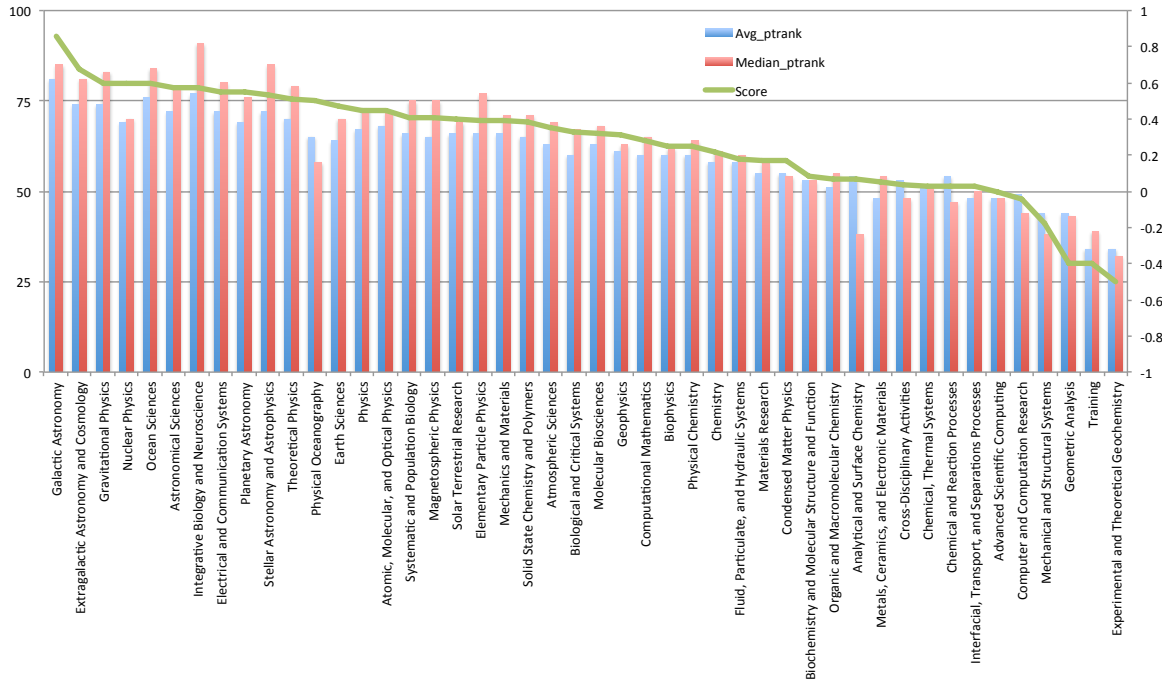[1] "XSEDE," Web Page. [Online]. Available: https://www.xsede.org/

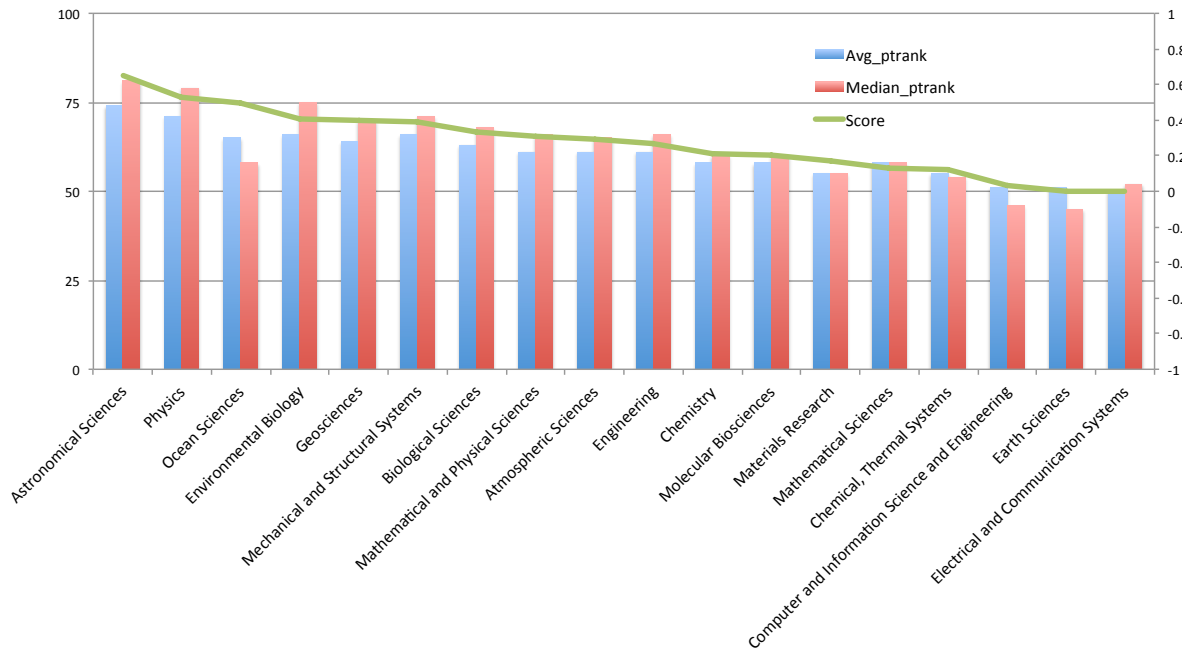Figure 7: Peer comparison based on Field of Science of XSEDE publications.

Figure 8: Peer comparison based on Parent Field of Science from the original analysis of XSEDE data.

[2] J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. Peterson, R. Roskies, J. Scott, and N. Wilkins-Diehr, "XSEDE: Accelerating Scientific Discovery," *Computing in Science Engineering*, vol. 16, no. 5, pp. 62–74, Sept 2014.

[3] P. Thomas and D. Watkins, "Institutional research rankings via bibliometric analysis and direct peer review: A compar-ative case study with policy implications," *Scientometrics*, vol. 41, no. 3, pp. 335–355, 1998.

[4] J. Bollen, M. A. Rodriguez, and H. Van de Sompel, "MESUR: Usage-based Metrics of Scholarly Impact," in *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, ser. JCDL '07. New York, NY,

Table III: The top level Field of Sciences in XSEDE as defined by NSF

FOS	Avarage ptrank	Median ptrank	Score
Mathematical and Physical Sciences	63	68	0.35
Geosciences	62	67	0.33
Biological Sciences	61	64	0.27
Engineering	58	62	0.2
Computer and Information Science and Engineering	50	46	0.01

Table IV: Percentile ranking and score of XSEDE publicatiopn data grouped by parent field of science.

FOS	Number of Pubs	average ptrank	Score
Astronomical Sciences	74	81	0.65
Physics	71	79	0.53
Ocean Sciences	65	58	0.5
Environmental Biology	66	75	0.41
Geosciences	64	70	0.4
Mechanical and Structural Systems	66	71	0.39
Biological Sciences	63	68	0.33
Mathematical and Physical Sciences	61	66	0.31
Atmospheric Sciences	61	65	0.29
Engineering	61	66	0.27
Chemistry	58	60	0.21
Molecular Biosciences	58	60	0.2
Materials Research	55	55	0.17
Mathematical Sciences	58	58	0.13
Chemical, Thermal Systems	55	54	0.12
Computer and Information Science and Engineering	51	46	0.03
Earth Sciences	51	45	0
Electrical and Communication Systems	50	52	0

Table V: Overview of the number of publications, average percentile ranking and performance score for XSEDE publication data based on Field of science.

FOS	Number of Pubs	average ptrank	Score
Galactic Astronomy	14	81	0.86
Extragalactic Astronomy and Cosmology	57	74	0.68
Gravitational Physics	104	74	0.60
Nuclear Physics	5	69	0.60
Ocean Sciences	5	76	0.60
Astronomical Sciences	187	72	0.57
Integrative Biology and Neuroscience	7	77	0.57
Electrical and Communication Systems	54	72	0.55
Planetary Astronomy	10	69	0.55
Stellar Astronomy and Astrophysics	31	72	0.53
Theoretical Physics	49	70	0.51
Physical Oceanography	5	65	0.50
Earth Sciences	19	64	0.47
Physics	147	67	0.45
Atomic, Molecular, and Optical Physics	55	68	0.45
Systematic and Population Biology	37	66	0.41
Magnetospheric Physics	11	65	0.41
Solar Terrestrial Research	10	66	0.40
Elementary Particle Physics	27	66	0.39
Mechanics and Materials	14	66	0.39
Solid State Chemistry and Polymers	8	65	0.38
Atmospheric Sciences	44	63	0.35
Biological and Critical Systems	6	60	0.33
Molecular Biosciences	494	63	0.32
Geophysics	8	61	0.31
Computational Mathematics	9	60	0.28
Biophysics	323	60	0.25
Physical Chemistry	168	60	0.25
Chemistry	259	58	0.22
Fluid, Particulate, and Hydraulic Systems	51	58	0.18
Materials Research	305	55	0.17
Condensed Matter Physics	61	55	0.17
Biochemistry and Molecular Structure and Function	141	53	0.08
Organic and Macromolecular Chemistry	44	51	0.07
Analytical and Surface Chemistry	7	54	0.07
Metals, Ceramics, and Electronic Materials	11	48	0.05
Cross-Disciplinary Activities	26	53	0.04
Chemical, Thermal Systems	38	51	0.03
Chemical and Reaction Processes	31	54	0.03
Interfacial, Transport, and Separations Processes	17	48	0.03
Advanced Scientific Computing	18	48	0.00
Computer and Computation Research	14	49	-0.04
Mechanical and Structural Systems	14	44	-0.18
Geometric Analysis	5	44	-0.40
Training	5	34	-0.40
Experimental and Theoretical Geochemistry	5	34	-0.50

USA: ACM, 2007, pp. 474–474. [Online]. Available: http://doi.acm.org/10.1145/1255175.1255273

[5] J. Bollen, H. Van de Sompel, and M. A. Rodriguez, "Towards Usage-based Impact Metrics: First Results from the Mesur Project." in *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries*, ser. JCDL '08. New York, NY, USA: ACM, 2008, pp. 231–240. [Online]. Available: http://doi.acm.org/10.1145/1378889.1378928

[6] J. Bollen, H. Van de Sompel, A. Hagberg, and R. Chute, "A principal component analysis of 39 scientific impact measures," *PloS one*, vol. 4, no. 6, p. e6022, 2009.

[7] J. Bollen, G. Fox, and P. R. Singhal, "How and where the TeraGrid supercomputing infrastructure benefits science," *Journal of Informetrics*, vol. 5, no. 1, pp. 114–121, 2011.

[8] "NCAR," Web Page. [Online]. Available: http://ncar.ucar.edu

[9] T. R. Furlani, B. L. Schneider, M. D. Jones, J. Towns, D. L. Hart, S. M. Gallo, R. L. DeLeon, C.-D. Lu, A. Ghadersohi, R. J. Gentner, A. K. Patra, G. von Laszewski, F. Wang, J. T. Palmer, and N. Simakov, "Using xdmod to facilitate xsede operations, planning and analysis," in *Proceedings of the Conference on Extreme Science and Engineering Discovery Environment: Gateway to Discovery*, ser. XSEDE '13. New York, NY, USA: ACM, 2013, pp. 46:1–46:8. [Online]. Available: http://doi.acm.org/10.1145/2484762.2484763

[10] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proceedings of the National academy of Sciences of the United States of America*, vol. 102, no. 46, pp. 16 569–16 572, 2005.

[11] "i-10 index — google scholar citations open to all." [Online]. Available: http://googlescholar.blogspot.com/2011/11/google-scholar-citations-open-to-all.html

[12] L. Egghe, "Theory and practise of the g-index," *Scientometrics*, vol. 69, no. 1, pp. 131–152, 2006.
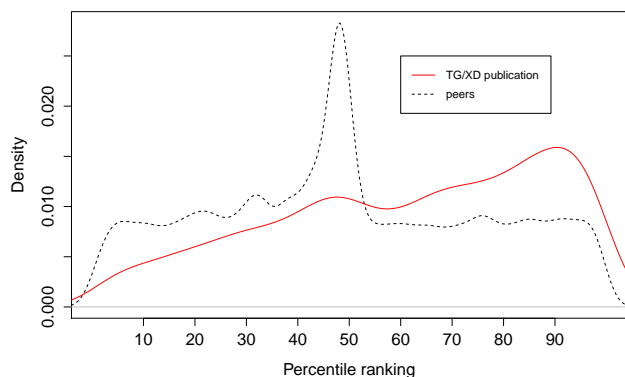
Figure 3: Kernel density of distributions for XSEDE publication percentile ranking versus peers.
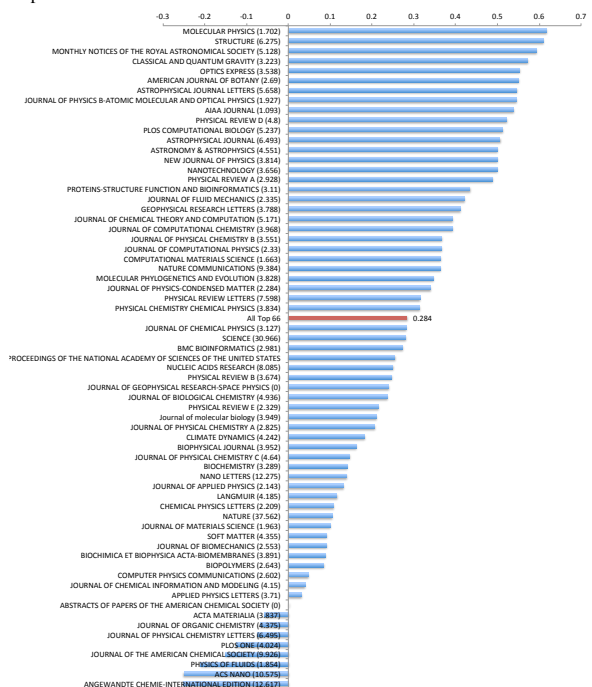

Figure 4: The score of our peer comparison metric for XSEDE publications by journal.
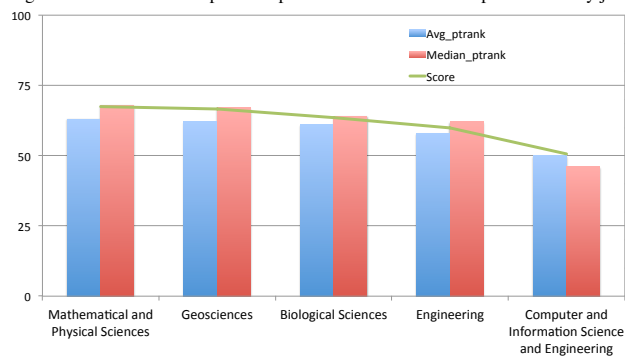

Figure 5: Peers comparison based on the topmost Field of Science category as defined by NSF.
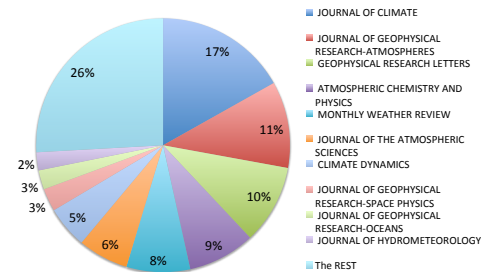

Figure 9: Distribution of the top most journals by publication count.
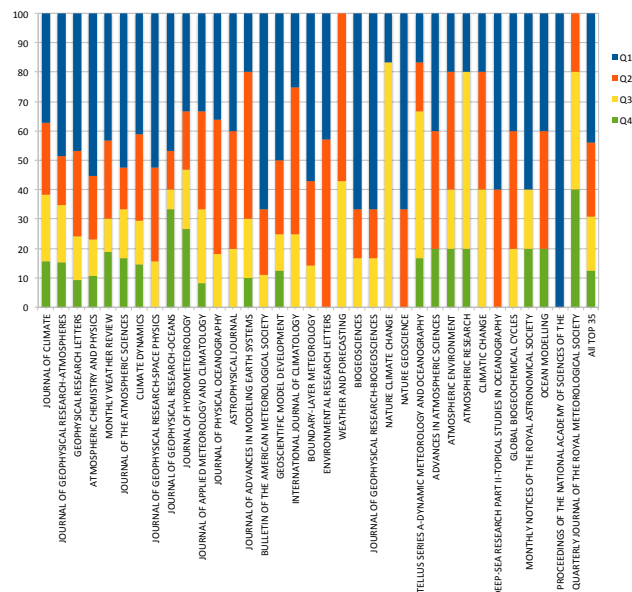

Figure 10: Percentile ranging by FOS in a stacked barchart of NCAR publications.
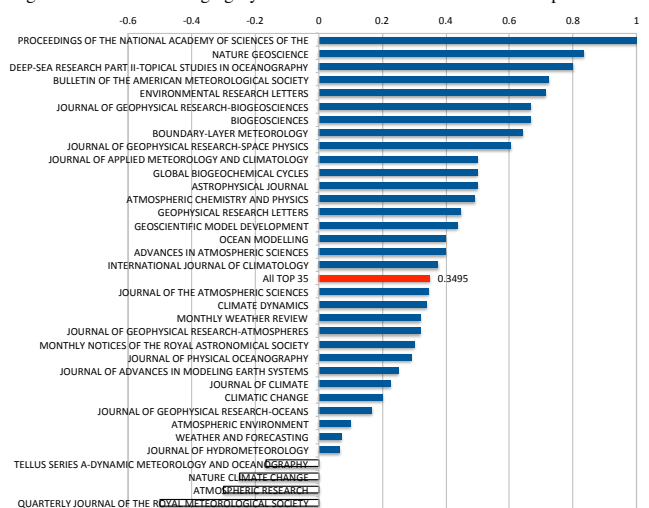

Figure 11: Peer comparison based on Parent Field of Science from the original analysis of XSEDE data.