## CITATION

## DOI

## PERMISSIONS

# Topics and Trends in NSF Ocean Sciences Awards

By Ivan D. Lima and Jennie E. Rheuban

**ABSTRACT.** The National Science Foundation Ocean Sciences Division (NSF-OCE) provides the majority of the support for ocean research in the United States. Knowledge of the trends in research and funding for NSF-OCE awards is important to investigators, academic institutions, policy analysts, and advocacy organizations. Here, we apply topic modeling to NSF-OCE award abstracts to uncover underlying research topics, examine the interrelationships between awards, and identify research and funding trends. The 20 topics identified by the model capture NSF-OCE's 10 largest programs (~90% of awards) remarkably well and provide better resolution into research subjects. The distribution of awards in topic space shows how the different topics relate to each other based on their similarity and how awards transition from one topic to another. Awards have become more interdisciplinary over time, with increasing trends in 13 of the 20 topics (65%). Seven topics show a growing fraction of the number of awards while six topics have a declining share. Both the annual inflation-adjusted amount of money awarded and the fraction of the annual funding have been increasing over time in four of the 20 topics. Three other topics show a decline in both the annual amount awarded and the fraction of total annual funding. The identified topics can be grouped into three major themes: infrastructure, education, and science. After 2011, increases in the mean annual cost per project result in a relatively constant fraction of annual funding for infrastructure, despite a significant decline in the infrastructure fraction of awards. The information presented on research and funding trends is useful to scientists and academic institutions in planning and decision-making, while the metrics we employed can be used by NSF to quantify the effects of policy decisions.

## INTRODUCTION

The National Science Foundation Ocean Sciences Division (NSF-OCE) plays a fundamental role in supporting basic ocean research, and its policies have significantly influenced the course of ocean sciences in the United States (NRC, 2000). Long-term analysis of NSF-OCE awards reveals a number of research trends and shifts in focus. While some trends are evident, such as the increasing importance of broader impacts, others are more subtle and not easily discernible. For example,

what research topics have declined or become more popular over the years? Are projects becoming more interdisciplinary? What are the funding trends for different research topics? Given NSF-OCE's importance for ocean research, it is necessary to go beyond anecdotal descriptions and provide more quantitative answers to these questions. In doing so, we also build formal metrics that are helpful to investigators, academic institutions, policy analysts, advocacy organizations, and NSF itself for long-range planning.

## RESEARCH TOPICS AND INTERDISCIPLINARITY

Here we apply topic modeling to NSF-OCE award abstracts to uncover underlying research topics, examine how the different topics and awards relate to one another, and identify research and funding trends. Topic modeling provides a set of machine learning techniques for analyzing the contents of large collections of text documents by extracting semantically meaningful topics embedded in the documents (Blei, 2012). Topic models have been used with great success in the humanities and political and social sciences (DiMaggio et al., 2013; McFarland et al., 2013; Mohr and Bogdanov, 2013). Abstracts are missing for most NSF-OCE records prior to 1985, so we compiled 11,238 abstracts from awards for the period 1985 to 2018 and extracted 20 topics using a Bayesian probabilistic model (see Supplementary Materials 1 and 2 for details). The topic model identifies groups of words that appear together frequently in the documents and that can usually be related to a theme or topic. The model output includes the latent topics, which consist of sets of word frequencies that when sorted in descending order give us the most common words in each topic and the distribution of topic probabilities for each award (see Supplementary Materials 2). The 10 most frequent words in the extracted topics (Table 1) show topics related to:

(1) instrument development, (2) impact of climate change on coral reefs, (3) education and outreach activities, (4) sediment cores and paleoclimatology, (5) geochemistry of hydrothermal vents, (6) observation of physical variables, (7) organic matter in the ocean, (8) modeling physical circulation processes, (9) population management and fisheries, (10) ship equipment and operation, (11) ocean biogeochemistry, (12) microbiology and genomics, (13) coastal management and human health, (14) oil spills in the Gulf of Mexico, (15) ecology and population dynamics, (16) plate tectonics, (17) ocean and atmospheric circulation variability in the North Atlantic, (18) coastal circulation processes, (19) seismology and geophysics, and (20) trace elements and isotopes. The model extracts the topics in no particular order; hereafter in the text and figures, we refer to them using their three most common words (Table 1).

The distribution of awards in topic space shows how the awards and topics relate to each other based on their similarity (Figure 1a). We visualize that distribution using a dimensionality reduction algorithm that provides a two-dimensional representation of the high-dimensional data that groups the awards based on their topic similarity (see Supplementary Materials 3). The distribution of awards forms a coherent structure where the different topics are represented as clusters of awards and related topics are located near each other. We can easily identify groups of topics that correspond to major subjects: biological, chemical, physical, and geological oceanography; equipment and instrumentation; and education and outreach (Figure 1a). The topic *record core past* is located next to *circulation atlantic variability*, even though it is part of geological oceanography, because these topics are closely related to each other in the sense that both deal with issues related to climate variability. Examination of the abstracts from the topic *gulf gas oil* reveals that these awards are requests for additional instrumentation to monitor oil spills, and that this small group of awards is closely associated with *sensor instrument technology* (Figure 1a). Comparison of the distribution of awards classified according to topic (Figure 1a) with that of awards from NSF-OCE's 10 largest programs (Figure 1b) shows that the model captures NSF-OCE major programs remarkably well and gives us better resolution into specific subjects within each program. The topic *vessel equipment operate* comprises two programs: "Oceanographic Instrumentation" and "Shipboard Scientific Support Equipment," and the topic *sensor instrument technology* includes awards from the programs "Ocean Technology and Interdisciplinary Coordination" and "Major Research Instrumentation." Nevertheless, the programs within these topics are quite similar to each other and there is a significant amount of overlap between them (Figure 1b).

Interdisciplinary research plays an important role in ocean sciences. Understanding the ocean as a system and its role in global climate change requires knowledge of the synergistic effects of diverse factors and, consequently, research that crosses many disciplines (Pachauri et al., 2014). In the model, the awards are composed of a mix of all 20 topics and therefore are inherently interdisciplinary. We quantify award interdisciplinarity based on the evenness of the distribution of topic probabilities (see Supplementary Materials 4 for details). Evenly distributed topic probabilities imply an even mix of topics and high interdisciplinarity.

**TABLE 1.** Ten most frequent words in each topic listed in declining order of frequency.

| TOPIC 1 | TOPIC 2 | TOPIC 3 | TOPIC 4 | TOPIC 5 | TOPIC 6 | TOPIC 7 | TOPIC 8 | TOPIC 9 | TOPIC 10 |
|---|---|---|---|---|---|---|---|---|---|
| sensor | coral | workshop | record | hydrothermal | temperature | organic | wave | population | vessel |
| instrument | reef | education | core | vent | depth | matter | flow | fish | equipment |
| technology | host | international | past | fluid | series | compound | eddy | larval | operate |
| capability | ecosystem | school | isotope | ridge | mooring | dissolve | numerical | zooplankton | ship |
| design | acidification | national | proxy | melt | cruise | doc | layer | specie | instrumentation |
| observatory | disease | oceanography | carbonate | rock | profile | chemical | mix | larvae | shipboard |
| instrumentation | temperature | experience | temperature | chemical | hydrographic | source | turbulence | recruitment | instrument |
| vehicle | specie | public | glacial | mantle | salinity | dom | internal | bank | operation |
| situ | ph | opportunity | level | temperature | deploy | material | energy | abundance | share |
| available | increase | participate | isotopic | geochemical | station | composition | flux | growth | acquisition |

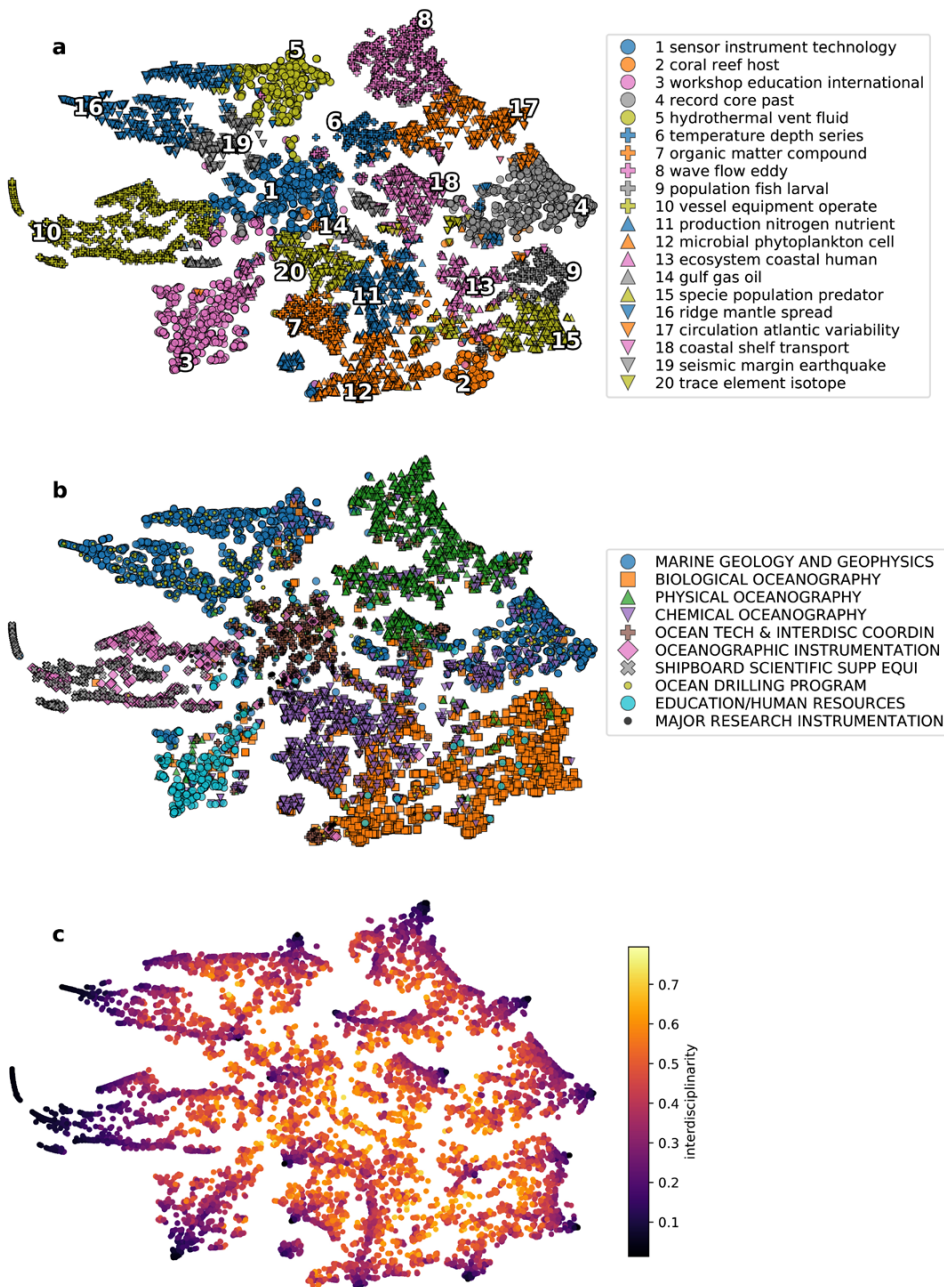| TOPIC 11 | TOPIC 12 | TOPIC 13 | TOPIC 14 | TOPIC 15 | TOPIC 16 | TOPIC 17 | TOPIC 18 | TOPIC 19 | TOPIC 20 |
|---|---|---|---|---|---|---|---|---|---|
| production | microbial | ecosystem | gulf | specie | ridge | circulation | coastal | seismic | trace |
| nitrogen | phytoplankton | coastal | gas | population | mantle | atlantic | shelf | margin | element |
| nutrient | cell | human | oil | predator | spread | variability | transport | earthquake | isotope |
| oxygen | bacteria | environmental | $CO_2$ | prey | seismic | north | river | subduction | particle |
| iron | growth | management | day | genetic | crust | basin | estuary | zone | metal |
| flux | gene | level | mexico | behavior | plate | southern | flow | fault | tracer |
| cycle | molecular | resource | total | animal | mid | atmospheric | wind | event | flux |
| fe | organism | public | proposal | organism | crustal | observation | plume | plate | concentration |
| primary | specific | health | spill | interaction | oceanic | heat | continental | drilling | geotrace |
| biogeochemical | diversity | lake | air | pattern | tectonic | flux | estuarine | deformation | seawater |

**FIGURE 1.** Distribution of National Science Foundation Ocean Sciences Division (NSF-OCE) awards in a t-Distributed Stochastic Neighbor Embedding (t-SNE) two-dimensional representation of 20-dimensional topic space (see Supplementary Materials 3 for details). (a) NSF-OCE awards classified according to topic. Awards are assigned the topic for which the probability is the highest. The white numbers on the plot correspond to the topic numbers on the legend. Topics are named after their three most common words. (b) NSF-OCE awards from the 10 largest programs in number of awards. These programs combined contain about 90% of all NSF-OCE awards. Awards are colored according to NSF-OCE program. (c) NSF-OCE awards colored according to interdisciplinarity. Higher values (closer to one) mean more interdisciplinary.

(65%), NSF-OCE awards become more interdisciplinary in time. The topic *seismic margin earthquake* is the only one where interdisciplinarity decreases. The increasing interdisciplinarity of NSF-OCE awards brings additional benefits to the oceanographic community, as the exchange of ideas and methods between disciplines leads to increased innovation (Balsiger, 2004), a more comprehensive understanding based on combined knowledge (Klein, 1990), and higher impact publications (Hicks et al., 2010).

## RESEARCH AND FUNDING TRENDS

Information on research and funding trends is of great value to investigators and academic institutions for managing resources, identifying opportunities, and planning investments in human capital and infrastructure. We examine topic trends by looking at how the annual fraction of the number of awards for the different topics varies in time (Figure 3). The topic fractions of the number of awards are computed based on the topic probabilities, which represent the proportions of the different topics in each award (see Supplementary Materials 2). The topics *coral reef host*, *workshop education international*, *ecosystem coastal human*, *gulf gas oil*, *coastal shelf transport*, *seismic margin earthquake*, and *trace element isotope* garner a growing fraction of the number of awards, while *temperature depth series*, *organic matter compound*, *population fish larval*, *vessel equipment operate*, *species population predator*, and *ridge mantle spread* decline over the years (Figure 3). The linear regression fit in Figure 3 is a first-order estimate of the trend through time, but for some of the topics, the behavior is more complex. The *coral reef host* fraction of the number of awards is relatively flat initially and starts to increase after 2005 with the emergence of ocean acidification research and concerns with accelerating warming trends and their impact on coral reef communities (Orr et al., 2005; Doney, 2006). The *vessel equipment operate* topic peaked in the 1990s

Conversely, uneven probabilities mean dominant topics and lower interdisciplinarity. The distribution of awards classified according to interdisciplinarity (Figure 1c) shows clusters of awards where each of the identified topics is dominant. Awards become more interdisciplinary as they move away from these regions and gradually transition from one dominant topic to another (Figure 1a,c). We examine temporal trends in interdisciplinarity by averaging the interdisciplinarity of awards for each topic and each year (see Supplementary Materials 4) and by looking at how these annual averages vary in time (Figure 2). In 13 of the 20 topics

and early 2000s with the acquisition (or refit) of several research vessels (*Thomas G. Thompson*, 1991; *Endeavor*, 1993; *Roger Revelle*, 1996; *Atlantis*, 1997; *F.G. Walton Smith*, 2000; and *Kilo Moana*, 2002) and has been declining since then. The upward shifts in *ecosystem coastal human* in 2004 and 2013 are most likely related to the development of the Coupled Natural and Human Systems (CNH) program in the early 2000s and the start of the Coastal Science, Engineering and Education for Sustainability (Coastal SEES) program. The *gulf gas oil* topic is also relatively constant until 2010, when the baseline shifts upward. The two outlier points with high fractions in that topic occurred in 1991 and 2010 and correspond to the *Exxon Valdez* oil spill in 1989, the explosion of the oil tanker *Mega Borg* off the coast of Galveston in 1990, and the Deepwater Horizon disaster in 2010.

In our model, award interdisciplinarity is proportional to the evenness of the distribution of topic probabilities. As awards become more interdisciplinary, the topic probabilities become more evenly distributed and the probability (proportion) of the dominant topic decreases (the sum of topic probabilities for each award always equals one), resulting in a lower fraction of the number or awards for that topic (see Equations 1 and 2 in Supplementary Materials). Therefore, for any given topic, award interdisciplinarity (Figure 2) and the annual fraction of the number of awards (Figure 3) tend to be inversely correlated. As a result, the observed decrease in the annual fraction of awards for the topics *temperature depth series*, *organic matter compound*, *population fish larval*, *vessel equipment operate*, *species population predator*, and *ridge mantle spread* (Figure 3) can be interpreted as a consequence of the increase in interdisciplinarity of the awards in these topics (Figure 2) and not necessarily as a decline in popularity. The annual growth in the topic *coastal shelf transport* is sufficiently large to offset the increase in interdisciplinarity in that topic and to result in an upward trend in its annual fraction of
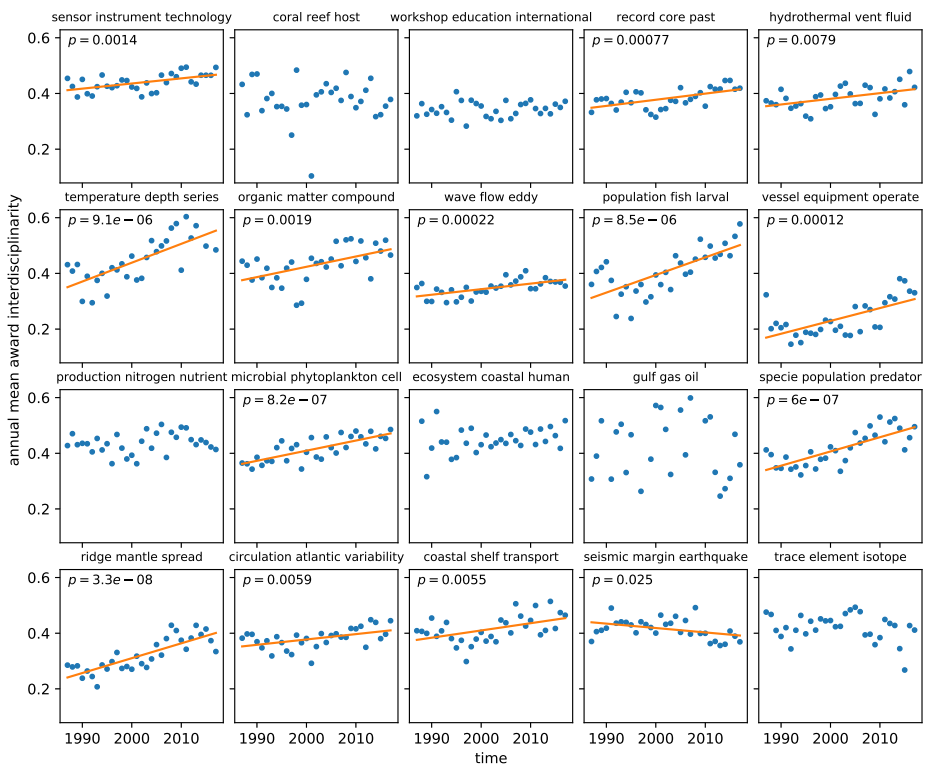


**FIGURE 2.** Time series of annual mean award interdisciplinarity for each topic. Higher values (closer to one) mean more interdisciplinary. Where the *p*-value for the regression slope is lower than 0.05, the regression line (orange) and the *p*-value are included.
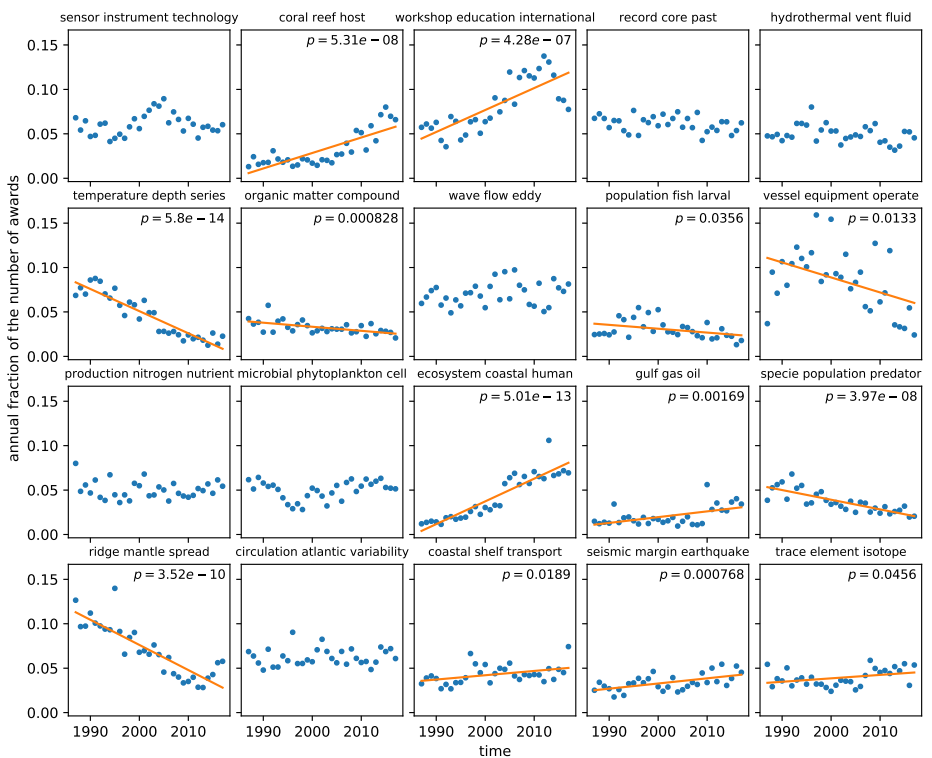


**FIGURE 3.** Time series of annual fraction of the number of awards for each topic. Where the *p*-value for the regression slope is lower than 0.05, the regression line (orange) and the *p*-value are included.

**FIGURE 4.** Time series of the annual amount of money awarded for each topic in millions of US dollars. Amounts are adjusted for inflation to 2017 US dollars. Annual amounts vary by more than an order of magnitude, so a log scale is used in the axis. Where the *p*-value for the regression slope is lower than 0.05, the regression line (orange) and the *p*-value are included.
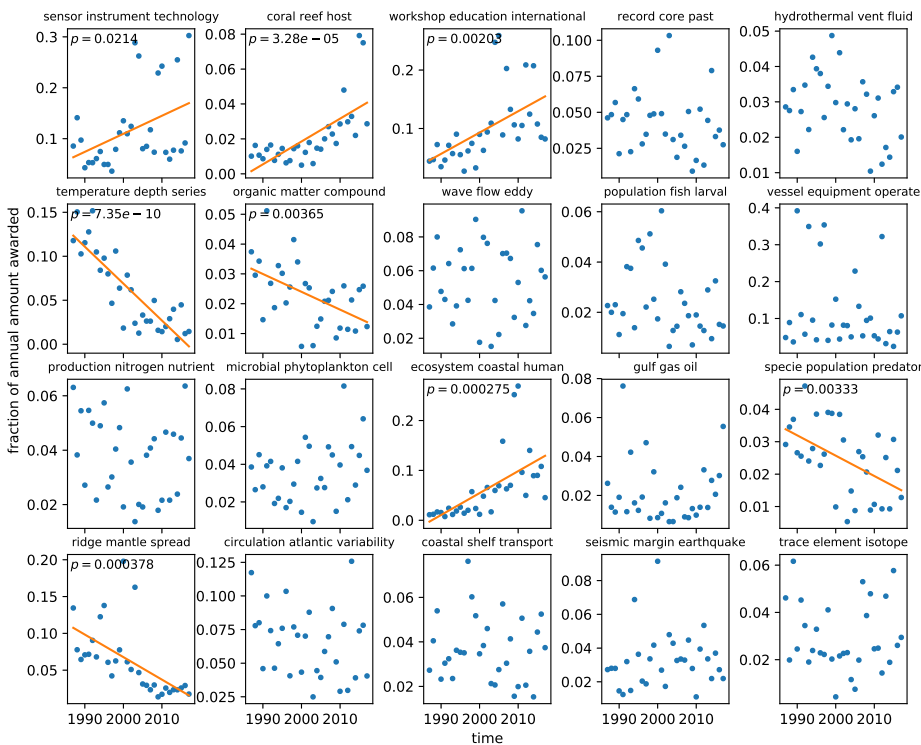


**FIGURE 5.** Time series of the fraction of the total annual amount awarded for each topic. Where the *p*-value for the regression slope is lower than 0.05, the regression line (orange) and the *p*-value are included.
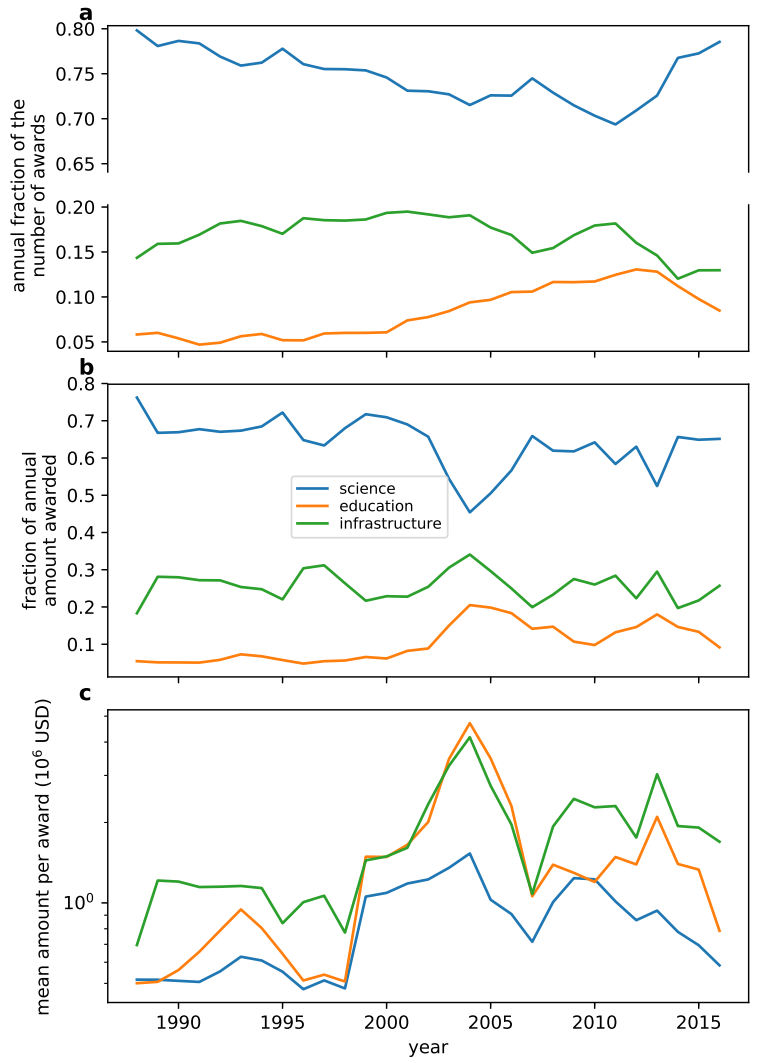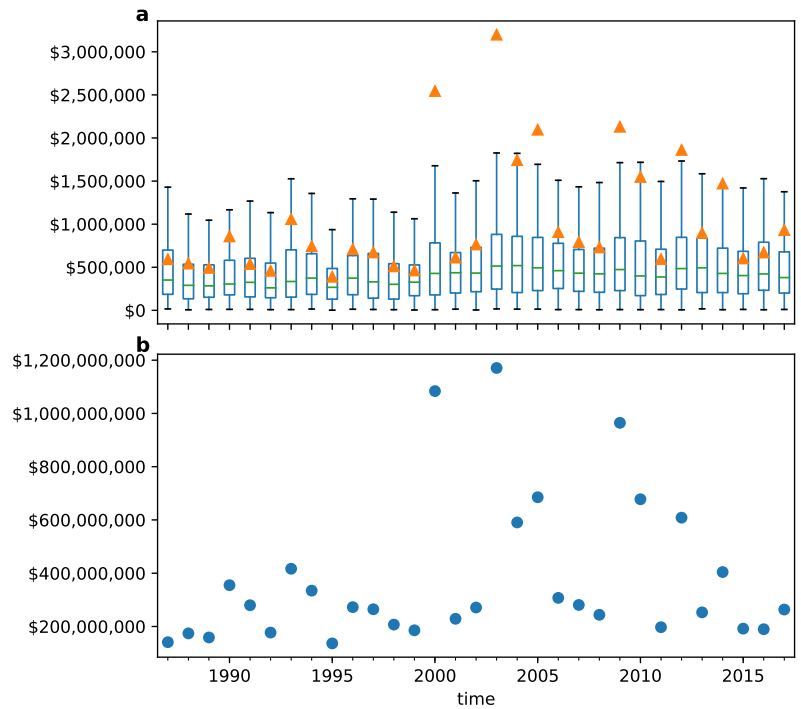
the number of awards (Figure 3).

Awareness of current topic funding trends can also be helpful to academic institutions and research groups for anticipating and preparing for future risks and challenges. We investigate funding trends for the different topics by looking at how the annual amount of money awarded for each topic and the topic fractions of the total annual amount awarded vary in time (Figures 4 and 5). In the calculations, the amounts awarded to individual projects are adjusted for inflation to 2017 US dollars. The estimated topic amounts and fractions, computed based on the topic proportions for each award (see Supplementary Materials 2), are quite noisy due to high variability in the amount awarded to individual projects and in total annual funding (Figure 6). Nevertheless, we can still detect trends. Funding for the topics *sensor instrument technology*, *coral reef host*, *workshop education international*, and *ecosystem coastal human* has been growing in both absolute value and fraction of annual funding, while the topics *temperature depth series*, *species population predator*, and *ridge mantle spread* have seen a decline in both absolute amount and fraction of total annual funding (Figures 4 and 5). The combination of relatively constant annual amounts (Figure 4) and higher frequency of years with high funding levels after 2000 (Figure 6b) could result in a downward trend in the fraction of the annual funding for the topic *organic matter compound* (see Equation 4 in Supplementary Materials 2). The increase in the annual amount awarded for *wave flow eddy* and *microbial phytoplankton cell* is offset by a higher frequency of years with high funding after 2000 (Figure 6b), resulting in no significant trend in topic fraction of annual funding. Award interdisciplinarity also tends to be inversely correlated with the annual amount awarded (Figure 4) and fraction of annual amount awarded (Figure 5), although the relationship is less clear due to variations in the amounts awarded to individual projects and in total annual funding (see Equations 3 and 4 in

**FIGURE 6.** (a) Box plot of the amount awarded to individual projects in US dollars. Orange triangles represent the annual means, and awards outside 1.5 × interquartile range are not plotted. (b) Total annual amount awarded in US dollars. Values represent the annual sums of awards with valid abstracts used in the analysis. The data include awards funded through Major Research Equipment and Facility Construction (MREFC) and the American Recovery and Reinvestment Act (ARRA), so the annual totals will exceed the NSF-OCE annual budget in some years. In both panels, values are adjusted for inflation to 2017 US dollars.

Supplementary Materials 2, Figure 6). The decline in both amount awarded and fraction of annual funding in the topics *temperature depth series*, *species population predator*, and *ridge mantle spread* can be explained by the decrease in the proportion of the dominant topic (Figure 3) due to the increase in interdisciplinarity in the awards in these topics (Figure 2).

Among the 20 topics identified (Table 1), three major themes or metatopics emerge: *infrastructure* (topics 1, 10, 14), *education* (topic 3), and *science* (topics 2, 4, 5, 6, 7, 8, 9, 11, 12, 13, 15, 16, 17, 18, 19, 20). We examine research and funding trends in these metatopics by combining the annual fractions of the number of awards and funding for the topics within each of the metatopics (Figure 7). The steady decline in the *science* fraction of awards up to 2011 is accompanied by increases in *infrastructure* (1987–1997) and *education*, which start to grow after 1997 when broader impacts became a merit review criterion. However, the sharp increase in the *science* content of awards after 2011 (Figure 7a) is not reflected in the *science* fraction of annual funding, which remains relatively constant (Figure 7b), denoting an overall decrease in the mean amount awarded per project in *science* after 2011 (Figure 7c). Conversely, higher costs per project in *infrastructure* result in a relatively steady fraction of annual funding despite a significant decline in the *infrastructure* fraction of the number of awards after 2011 (Figure 7a). The results presented here are consistent with the findings in the *Sea Change* report (NRC, 2015), which shows rising research infrastructure costs approaching and surpassing investments in core research programs after 2011. In our model, funding for *education* is proportional to the *education* content of awards (see Supplementary Materials 2). However,



**FIGURE 7.** Metatopic time series. The metatopic data show large variations from year to year so the annual values were smoothed using a three-year sliding window average to help visualize trends. (a) Annual metatopic fractions of the number of awards. (b) Metatopic fractions of annual amount awarded. (c) Mean amount awarded per project per year for each metatopic. Amounts are adjusted for inflation to 2017 US dollars. A log scale is used in the axis.

that is not necessarily true for *science* awards (particularly after the inclusion of broader impacts as a merit review criteria), as a relatively small fraction of the award funds are allocated to broader impact activities. As a result, our model tends to overestimate funding for *education* (Figure 7b).

## CAVEATS AND LIMITATIONS

In our study, the total amount awarded annually (Figure 6b) will differ substantially from the NSF-OCE budget and other studies (Witze, 2015; Mix, 2017) in some years for three main reasons:

1. This study includes awards funded through Major Research Equipment and Facility Construction (MREFC) and the American Recovery and Reinvestment Act (ARRA). MREFC is an NSF account that supports the acquisition and construction of major research facilities and equipment through special Congressional appropriations, and ARRA was a stimulus package enacted by the 111th US Congress and signed into law by President Barack Obama in February 2009. These funds are generally not part of the NSF-OCE budget. Nevertheless, NSF-OCE makes the awards to other organizations, and in the case of MREFC projects, is responsible for oversight of the development and performance of the facilities. Because we aim to provide a full picture of NSF-OCE science activities, including large investments over limited periods of time, we chose to include awards funded through these programs.

2. This study excludes awards with missing abstracts (see Supplementary Materials 1).

3. Some awards are funded in full in the first year while others are funded one year at a time over the duration of the award. Our data set contains only the total amount awarded, which we assign in full to the year the project starts.

The topic model uses a Bayesian probabilistic approach to approximate the topic structure of the NSF-OCE awards, given the observed collection of abstracts. One of the caveats of this type of model is that it is quite sensitive to the input data. Relatively small changes in the collection of documents or on how the *bag-of-words* is built can change some of the topics extracted. Nevertheless, generative probabilistic models are the simplest and most widely used topic models due to the high quality of their results (Griffiths and Steyvers, 2004; Blei, 2012).

## SUMMARY AND CONCLUSIONS

Here, we uncover the main topics in NSF-OCE awards and identify research and funding trends using formal quantitative metrics. This information on topic and funding trends is valuable to investigators, academic institutions, and program managers for strategic planning and informing policy. Metrics similar to what we employed can be used by NSF and policy analysts to measure progress toward established research goals and to quantify the impacts of policy changes.

### REFERENCES
Balsiger, P.W. 2004. Supradisciplinary research practices: History, objectives and rationale. *Futures* 36(4):407–421, https://doi.org/10.1016/j.futures.2003.10.002.
Blei, D.M. 2012. Probabilistic topic models. *Communications of the ACM* 55(4):77–84, https://doi.org/10.1145/2133806.2133826.
Blei, D.M., and J. Lafferty. 2009. Topic models. Pp. 71–94 in *Text Mining: Classification, Clustering, and Applications*, 1st ed. A.N. Srivastava and M. Sahami, eds, Chapman and Hall/CRC, Boca Raton, FL.
Blei, D.M., A.Y. Ng, and M.I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3(Jan):993–1,022.
DiMaggio, P., M. Nag, and D. Blei. 2013. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics* 41(6):570–606, https://doi.org/10.1016/j.poetic.2013.08.004.
Doney, S.C. 2006. The dangers of ocean acidification. *Scientific American* 294(3):58–65, https://doi.org/10.1038/scientificamerican0306-58.

Griffiths, T.L., and M. Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America* 101(suppl 1):5,228–5,235, https://doi.org/10.1073/pnas.0307752101.
Hicks, C.C., C. Fitzsimmons, and N.V.C. Polunin. 2010. Interdisciplinarity in the environmental sciences: Barriers and frontiers. *Environmental Conservation* 37(4):464–477, https://doi.org/10.1017/S0376892910000822.
Klein, J.T. 1990. *Interdisciplinarity: History, Theory, and Practice*. Wayne State University Press, 340 pp.
Ludwig, J.A., and J.F. Reynolds. 1988. *Statistical Ecology: A Primer in Methods and Computing*. John Wiley & Sons, 362 pp.
McFarland, D.A., D. Ramage, J. Chuang, J. Heer, C.D. Manning, and D. Jurafsky. 2013. Differentiating language usage through topic models. *Poetics* 41(6):607–625, https://doi.org/10.1016/j.poetic.2013.06.004.
Mix, A.C. 2017. Follow the money. *Oceanography* 30(2):7–9, https://doi.org/10.5670/oceanog.2017.202.
Mohr, J.W., and P. Bogdanov. 2013. Introduction—topic models: What they are and why they matter. *Poetics* 41(6):545–569, https://doi.org/10.1016/j.poetic.2013.10.001.
NRC (National Research Council). 2000. *50 Years of Ocean Discovery: National Science Foundation 1950–2000*. The National Academies Press, 276 pp., https://doi.org/10.17226/9702.
NRC. 2015. *Sea Change: 2015–2025 Decadal Survey of Ocean Sciences*. The National Academies Press, 98 pp., https://doi.org/10.17226/21655.
Orr, J.C., V.J. Fabry, O. Aumont, L. Bopp, S.C. Doney, R.A. Feely, A. Gnanadesikan, N. Gruber, A. Ishida, F. Joos, and others. 2005. Anthropogenic ocean acidification over the twenty-first century and its impact on calcifying organisms. *Nature* 437(7059):681–686, https://doi.org/10.1038/nature04095.
Pachauri, R.K., M.R. Allen, V.R. Barros, J. Broome, W. Cramer, R. Christ, J.A. Church, L. Clarke, Q. Dahe, P. Dasgupta, and others. 2014. *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. R.K. Pachauri, and L. Meyer, eds, IPCC, Geneva, Switzerland, 151 pp.
van der Maaten, L., and G. Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9(Nov):2,579–2,605.
Witze, A. 2015. US ocean sciences told to steer a new course. *Nature News* 517(7536):538, https://doi.org/10.1038/nature.2015.16780.

### AUTHORS
**Ivan D. Lima** (ilima@whoi.edu) is Research Specialist, and **Jennie E. Rheuban** is Research Associate, Department of Marine Chemistry and Geochemistry, Woods Hole Oceanographic Institution, Woods Hole, MA, USA.