The necessary optimization of the data lifecycle: Marine geosciences in the big data era

Taylor R. Lee*†, Benjamin J. Phrampus† and Jeffrey Obelcz†

Ocean Sciences Division, US Naval Research Laboratory, Stennis Space Center, MS, United States
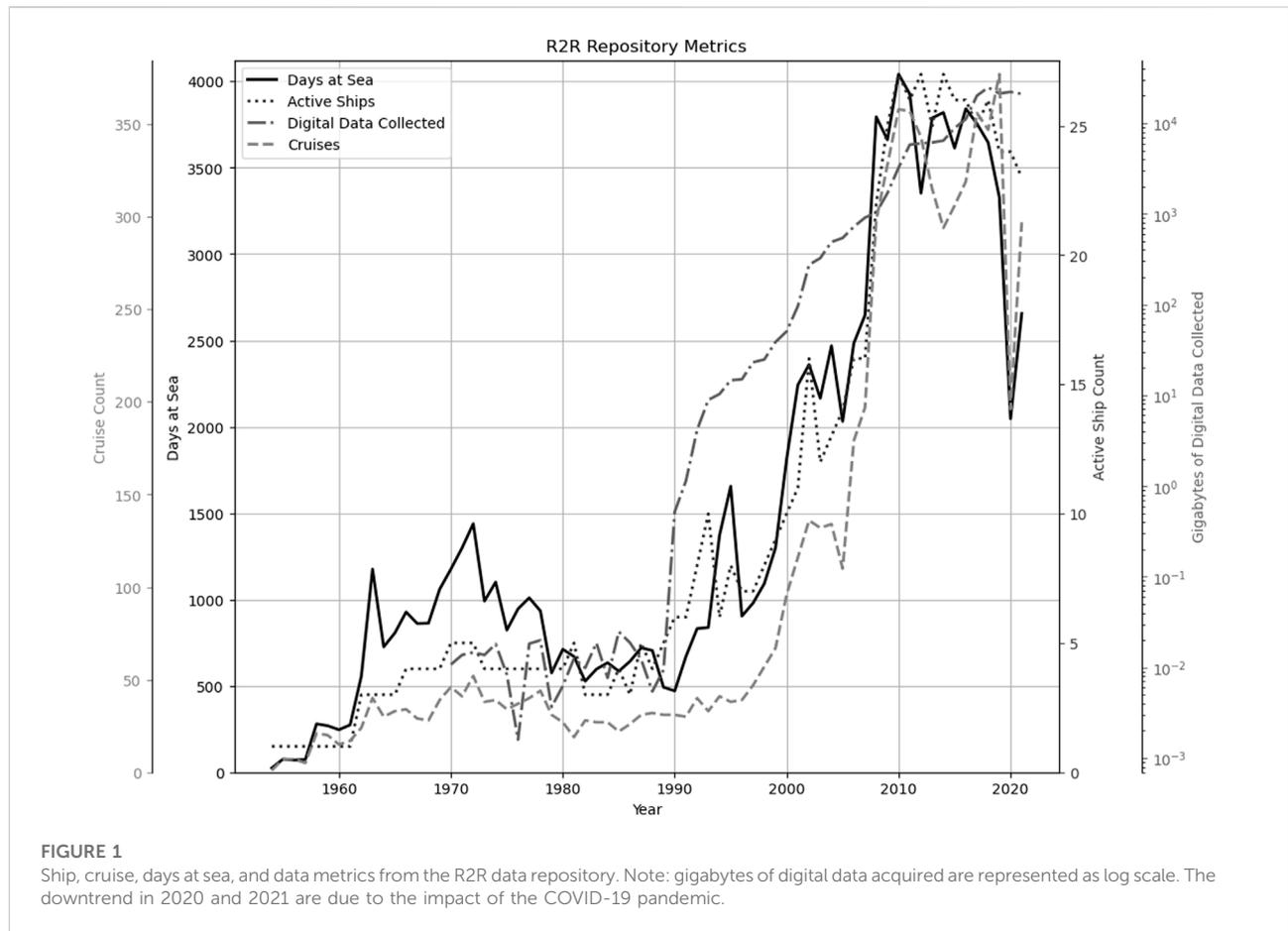
In the marine geosciences, observations are typically acquired using research vessels to understand a given phenomenon or area of interest. Despite the plateauing of ship time and active research vessels in the last decade, the rate of marine geoscience data production has continued to increase. Simultaneously, there exists large quantities of legacy data aggregated within data repositories; however, these data are rarely curated to be both discoverable and machine-readable (i.e., accessible). This results in inefficient use, or even omission, of high-quality data, that is, both increasingly important to utilize and impractical to recollect. The proliferation of newly acquired data, and increasing importance of legacy data, has only been met with incremental evolution in the methods of data integration. This paper describes some improvements at each stage of the data lifecycle (acquisition, curation, and integration) that could align the marine geosciences better with the "big data" paradigm. We have encountered several major issues coordinating these efforts which we outline here: 1) geologic anomalies are the primary focus of data acquisition and pose difficulty in understanding the dominant (i.e., baseline) marine geology, 2) marine geoscience data are rarely curated to be accessible, and 3) aforementioned issues preclude the use of efficient integration tools that can make optimal use of data. In this paper, we discuss challenges and solutions associated with these issues to overcome these concerns in future decades of marine geoscience. The successful execution of these interconnected steps will optimize the lifecycle of marine geoscience data in the "big data" era.

KEYWORDS

big data, data lifecycle, database, data acquisition, data curation, data-driven, data integration

Introduction

The field of open marine geoscience (in contrast to coastal geoscience, which is comparatively more data-rich) has experienced a dramatic increase in the volume and variety of data production since the late 1950s, which corresponds with the initiation of the R2R (Rolling Deck to Repository, https://www.rvdata.us). The R2R houses all UNOLS (University-National Oceanographic Laboratory System) digital data acquired and constitutes a representative, but not comprehensive, repository of both structured (e.g.

FIGURE 1
Ship, cruise, days at sea, and data metrics from the R2R data repository. Note: gigabytes of digital data acquired are represented as log scale. The downtrend in 2020 and 2021 are due to the impact of the COVID-19 pandemic.
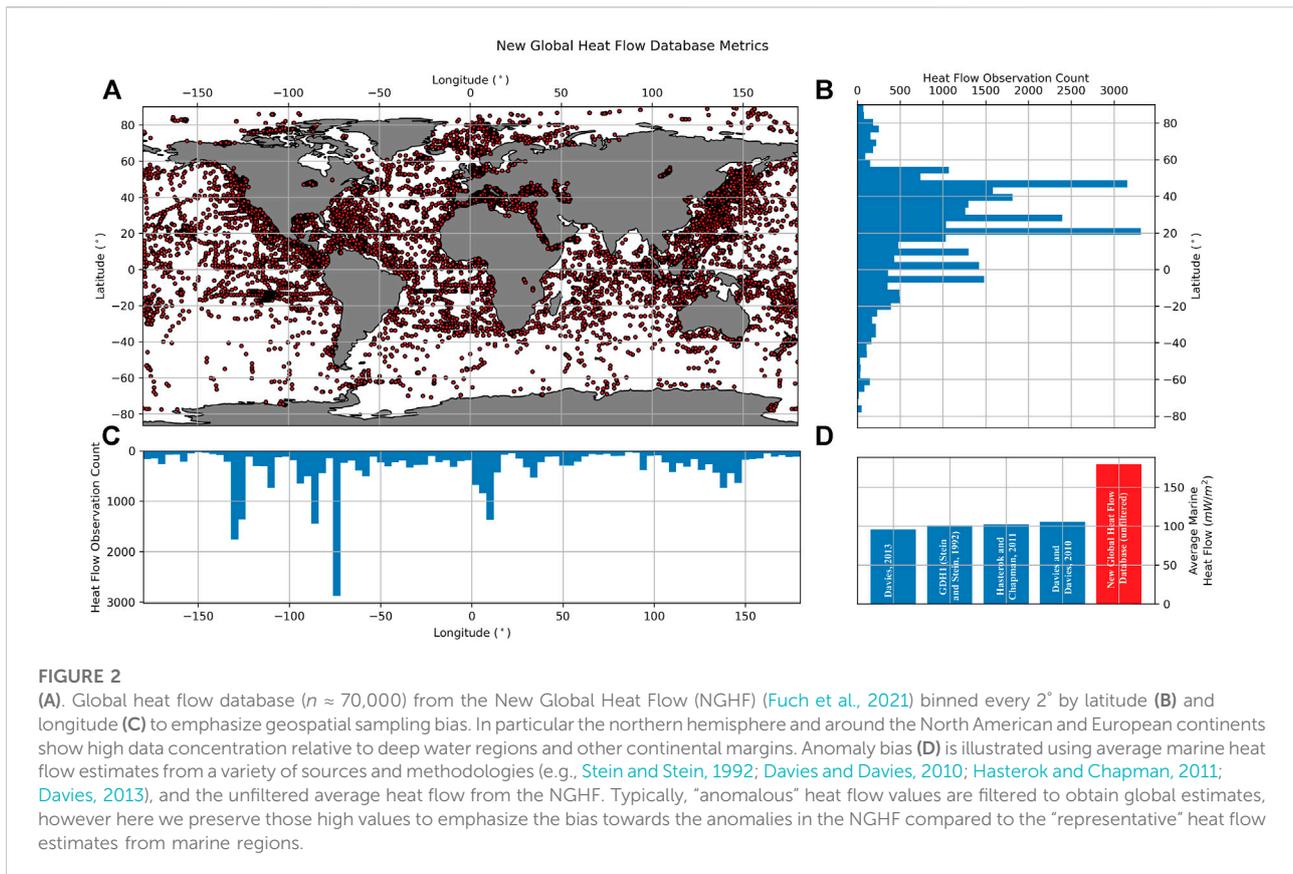
, multibeam, seismic, *etc.*) and unstructured (e.g., core log, grab sample description, *etc.*) marine geoscience data. Using data trends within R2R as a proxy for the broader marine geoscience field shows the total amount of digital data acquired has exponentially increased over the last three decades, while growth of active ships, cruises per year, and days at sea within the last decade (~2010) has stagnated (Figure 1).

These trends in the R2R dataset are consistent with other analyses that observe recent data quantity increases due to new advancements in technologies (e.g., Watts, 2020). Data production velocity stands to further increase as novel technologies mature, particularly the development of autonomous (e.g., autonomous underwater/surface vehicles (AUVs/ASVs)) and stand-alone passive systems (e.g., distributed acoustic sensing). This exponential growth in data volume and the development and utilization of more data-intensive technologies will necessitate 'big data' approaches (informally defined as computational analysis of large, i.e., terabyte-scale, volumes of structured and unstructured data) for the marine geosciences. However, the marine geoscience data

lifecycle is currently not designed to house, or even properly utilize, this data.

The current state of data curation in the marine geosciences is incrementally evolving as data sharing mandates from funding agencies and scientific journals push for open-source data policies (e.g., FAIR principals of findable, accessible, interoperable, and reusable data; Wilkinson et al., 2016). However, this approach still largely resembles the 20th century model, whereby data hosted online in a variety of disparate formats and states of curation, often in a data repository, i.e., a place to store data, which is not a substitute for a structured, discoverable, and machine readable database. Further, the acquisition, curation, and integration of data are often uncoordinated between research campaigns, even if science objectives are heavily linked. Since data-driven workflows require 'big data' volumes, linking the data lifecycle with data acquisition is increasingly important.

This paper is divided into three sections, one for each of the three phases of the data lifecycle: acquisition, curation, and integration. The data lifecycle involves: 1) the acquisition of new data, 2) the curation of data for future use, and 3) the integrating of data by secondary users to test new hypotheses. In

FIGURE 2
(A). Global heat flow database ($n \approx 70{,}000$) from the New Global Heat Flow (NGHF) (Fuch et al., 2021) binned every 2° by latitude **(B)** and longitude **(C)** to emphasize geospatial sampling bias. In particular the northern hemisphere and around the North American and European continents show high data concentration relative to deep water regions and other continental margins. Anomaly bias **(D)** is illustrated using average marine heat flow estimates from a variety of sources and methodologies (e.g., Stein and Stein, 1992; Davies and Davies, 2010; Hasterok and Chapman, 2011; Davies, 2013), and the unfiltered average heat flow from the NGHF. Typically, "anomalous" heat flow values are filtered to obtain global estimates, however here we preserve those high values to emphasize the bias towards the anomalies in the NGHF compared to the "representative" heat flow estimates from marine regions.

each section, we briefly describe the current paradigm using examples from representative, but not comprehensive, marine geoscience datasets and identify the challenges that we, as data-driven marine geoscientists, have dealt with. To address these challenges, we put forth potential improvements as both existing approaches to be encouraged, and new approaches to be implemented and developed. We believe these improvements, while not comprehensive, would represent wholesale steps in moving marine geoscience into the "big data" era.

Marine geoscience data acquisition

Current paradigm

Field data acquisition in the marine geosciences is influenced by a balance of obstacles and incentives. Data acquisition is constrained by barriers which can be: financial (field efforts are expensive), logistical (mobilizing/demobilizing is challenging), and administrative (permitting and regulations are limiting). Therefore, the ability to collect field data on seagoing vessels is limited to those with the proper funding, resources, and permits, resulting in biased datasets (Coperdock et al., 2021). An example of these biases can be found in the New Global Heat

Flow (NGHF) database (Figure 2; Fuch et al., 2021), where data acquisition is biased to the northern hemisphere (e.g., Figure 2B), particularly the United States and Western Europe (e.g., Figure 2C). This also results in a shallow water depth bias (e.g., Figure 2A; Diesing, 2020).

Another acquisition bias stems from field efforts that focus on geologic anomalies. These anomalous and/or societally important phenomena are often quantified more frequently than baseline (i.e., observational majority) regions (Figure 2D). Anomaly-focused observations are implicitly encouraged *via* incentives that drive marine geoscience data acquisition, i.e., appealing to funding agencies and/or high impact peer-reviewed journals. The inherent and sometimes necessary (i.e., marine hazard assessment) bias of funding and subsequent sampling towards anomalies makes for a dataset, that is, not representative of the marine realm as a whole.

Challenges

Despite innovations in marine geoscience data acquisition (e.g., AUVs), the vast majority of the seabed will likely never be surveyed or sampled at high spatial and/or temporal resolution. Under the current paradigm, marine geoscience datasets tend to

be anomaly biased (e.g., Figure 2D). This bias can cause challenges for data-driven modeling of the broader marine realm, since data-driven methods can only learn from what has been previously observed. With anomalies overrepresented in observational datasets, an accurate representation of the marine realm can be more challenging to obtain.

In addition to heat flow, an "anomaly-driven" dataset exists in the study of seafloor fluid expulsion anomalies (SEAFLEAs), such as seafloor seeps (Phrampus et al., 2020). In this example, marine scientists are driven to sites of anomalous fluid flow due to their large chemical gradients, which alter seabed biogeochemistry and host diverse benthic communities (e.g., Skarke et al., 2014). Subsequently, SEAFLEAs observations are most commonly reported as anomalies only (i.e., no absence points), which can heavily influence data-driven analyses. This bias limits data-driven analyses and results in poorly-generalized results due to the limited feature selection capability, and fuzzy delineation between anomaly and no-anomaly locations due to the limited capability in identifying unobserved phenomena (Phrampus et al., 2020). Anomaly bias can also be observed in marine geochronology, which also omits sediment cores with zero net sediment accumulation (e.g., Restreppo et al., 2021). More representative datasets with absence data (e.g., Diesing et al., 2021) would bypass analysis limitations and result in more comprehensive examination of global phenomena.

Without representative datasets, we have limited recourse to deal with anomaly-driven sampling bias, and are restricted to accounting for this bias *post hoc*. Additionally, "anomaly-driven research" tends to be performed under the implicit assumption that "normal" areas will remain "normal" under rapid changes in the thermal, chemical, and biological seabed induced by anthropogenic climate change on region to margin scale (Kopf, 2009; McKenna, 2015; Rillo et al., 2019; Marchese et al., 2022). Any understanding derived from biased marine geoscience datasets will serve as poor baselines for forward modeling efforts.

Improvements

The wholesale adaptation of autonomous research platforms (AUVs/ASVs, see Sahoo et al. (2019) for overview) by the marine geoscience community can serve to combat anomaly bias. These platforms provide obvious appeal to researchers due to reduced operational expense and extensive data collection with days, weeks, or even months of data acquisition between platform recoveries. From our perspective, the limited control of survey patterns may be an overall benefit in making data acquisition less anomaly-focused.

While autonomous research platforms provide promise to deliver systematic data acquisition at reduced cost, seabed sampling (including coring, heat flow measurements, and geotechnical profiling) is unlikely to become an autonomous

activity in the near future. Therefore, we believe public funding agencies, who subsidize the majority of marine geoscience research, should incentivize systematic or exploratory data acquisition designs, as was common in the 1960s and 70s during the early days of marine geoscience exploration (e.g., GeoMapApp archived Analog Seismic Reflection Profiles collected by R/Vs Robert D. Conrad, Eltanin, Vema, *etc.*). One potential method funding agencies could use is adding a prompt such as "do these data contribute to a representative data baseline?" to proposal evaluation rubrics.

Marine geosciences data curation

Current paradigm

Following field collection and data moratoriums, funding agencies often require principal investigators to be good stewards of their funding and publish their data in a data repository. These data can be tremendously useful and even invaluable as certain exploratory datasets will likely never be acquired again. For example, a long offset seismic line acquired continuously from Cape Hatteras to the Mid-Atlantic Ridge is approximately 3,400 km long (Agena et al., 1993). Seismic data along this trackline is unlikely to ever be reacquired due to logistics, expense, and sanctions. Therefore, making this legacy data both discoverable and machine-readable is a high priority.

Currently the largest data holdings are hosted within data repositories operated by public institutions, such as NOAA's National Center for Environmental Information (NCEI) and Germany's PANGAEA (Diepenbroek et al., 2002). These repositories provide some parsing capabilities, including keyword and geographic search parameters. However, these repositories are not discoverable and machine readable databases. For example, the amalgamation of ocean drilling data (e.g., International Ocean Discovery Program JANUS website) is a repository since the data lacks both ease of access and a consistent structure between expeditions. Conversely, Lamont-Doherty Earth Observatory's GeoMapApp application houses a database with a large amount of discoverable marine geoscience data stored within a georeferenced GIS application. The GeoMapApp stands as an exception to the general rule that marine geoscience data are deposited "as-is" in repositories by mandate of funding agencies or publishing journals.

Challenges

The largest challenge in data curation is the lack of incentive for researchers to do more than the bare minimum to curate their data. With the Agena et al. (1993) example, the data have issues such as non-uniform sample rates, missing traces, and

inconsistent or missing deep-water delays. These issues of insufficiently quality-controlled data can result in huge time sinks for data re-users. Issues like these are frequent and persistent throughout public platforms and marine geoscience data types.

Another issue for marine geoscience data curation is the lack of data format uniformity and required metadata. Few marine geoscience data types have an almost universally accepted format, such as SEG-Y for seismic data (SEG Technical Standards Committee, 2017). Point data, such as sediment cores, are inherently less structured than gridded data formats such as seismic and multibeam, and are typically in a delimited text file. However, field names, separators, data units, and other metadata can vary widely between data acquirers. The variety and lack of uniformity of these unstructured data creates a data conditioning time sink before data can be integrated. Finally, data organizations often arrange data based on geographic region and rarely based on data-type, which further inhibits bulk data download capabilities necessary for global analyses.

Improvements

We believe better data curation can be achieved using a "carrot instead of stick" approach, wherein researchers are incentivized to better curate their (or other's) data instead of punishing them for not. Accordingly, instead of withholding funding from researchers if data are not sufficiently curated, funding agencies could include data curation metrics in their proposal evaluation rubric. This would motivate proposal writers to include "data literate" scientists on their teams, such as data scientists and/or personnel from data curating agencies like NCEI, who are funded to curate data (https://www.ncei.noaa.gov). The inclusion of data literate scientists can also lead to a consensus on data formats within a realm of study, resulting in consistent data structures across the marine geoscience community.

Data curation is not always possible for legacy datasets for which data rescue is the only option. Due to the high cost of re-acquiring datasets, rescue efforts can have an outstanding return on investment. For example, Analog Seismic Reflection Profile data housed within the GeoMapApp constitutes ~2.3 M km of continuous single-channel profile data. Assuming a ship speed of ~8.3 km/h (4.5 knots) and 24 h operations, a single ship would cover ~200 km/day. To recollect this data would require 11,769 days at sea. With a conservative day-rate of $50k (USD) for a global class ship, the total cost of re-collection would be ~$588M.

Considering this cost, we encourage funding institutions to treat standalone data rescue proposals with equal priority to data acquisition efforts. Data journals (such as *Scientific Data*) and repositories with standalone DOIs for data (such as Zenodo) are also strong positive reinforcement tools for researchers to properly curate their data after publication. Data journals publish peer-reviewed curated data, creating a product that tangibly counts towards a researchers' productivity.

To address issues of disparate data formats, we believe the marine geoscience community should look to other earth science disciplines for inspiration. SEG-Y is one of the few marine geoscience data formats that has a stringent data/metadata format and is widely adopted. For many other types of data, particularly gridded data, the NetCDF (Network Common Data Format; Rew and Davis, 1990) format provides a self-described and flexible format, that is, compatible with many popular data processing and analysis software packages. Formats such as these, coupled with established metadata (e.g., units) and attribute names, could collate disparate data formats. Finally, adding utilities to bulk download data by type would be useful for analyses of one geologic quantity and would allow for quicker turnaround in utilizing and integrating data.

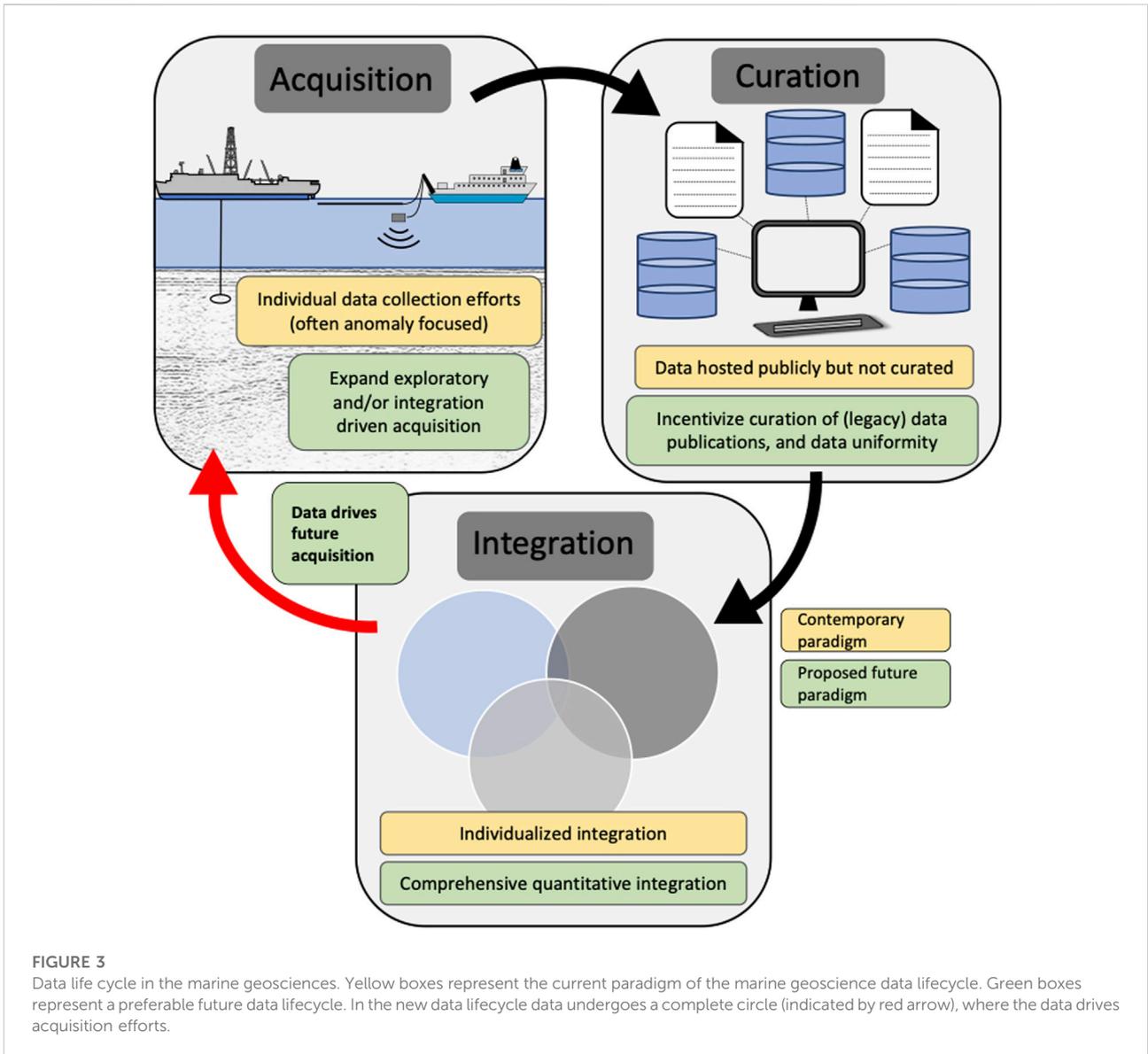## Marine geosciences data integration

Current paradigm

The final, often repetitive, stage of the data lifecycle is the integration and reutilization of data. In this step, datasets are integrated into a holistic analysis tool such as machine learning and/or a GIS-based workflows. Data-driven approaches, like machine learning, have remained relatively novel tools in the marine geosciences, despite the common use in diverse fields such as meteorology and finance (Dixon et al., 2020; Chase et al., 2022). We believe this is, in part, due to the data issues described above which make data mining particularly difficult in the marine geosciences.

In the contemporary paradigm, it is uncommon that integrated data are used quantitatively to guide future research endeavors, including where and what kinds of measurements to collect. Only in specific circumstances has a systematic approach been taken in data collection *via* the filling in geospatial gaps in datasets (e.g., Mayer et al., 2018 initiative). However, this geospatial approach only makes sense when a "complete" dataset is practically attainable, i.e., measurements acquired underway without holding station. Therefore, other methods to identify where and what kinds of measurements should be taken require a more data-driven, instead of geospatial, approach.

Challenges

Within our data integration and rescue efforts, one of the largest issues we've encountered has been finding and training scientists with both marine geoscience and data science expertise.

FIGURE 3
Data life cycle in the marine geosciences. Yellow boxes represent the current paradigm of the marine geoscience data lifecycle. Green boxes represent a preferable future data lifecycle. In the new data lifecycle data undergoes a complete circle (indicated by red arrow), where the data drives acquisition efforts.

The classically trained marine geoscientist uses tools such as geophysical data interpretation and/or geological sediment analysis over relatively small spatial/temporal scales to better understand a region. Such methods require expertise with highly specific types of data (e.g., subbottom profiler data or geochemical isotopes), but generally not expertise in data integration and reuse. This limited perspective requires marine geoscientists to only be data literate within the bounds of their data, which inherently hinders their ability to make their data usable to the community outside their specific expertise.

The challenges discussed above are exacerbated by a lack of collaboration between traditional field-based, observational marine geoscientists and data miners. Without an effective institutional incentive structure in place to aid this collaboration, data miners can only offer authorship (and/or

model outputs) to data acquirers in exchange for data access. This transactional approach is inefficient at best at maximizing the utility of marine geoscience data.

Improvements

Marine geoscience data integration is inherently limited by data acquisition and curation. Accordingly, the suggestions that apply to the previous sections also apply here. Particularly, suggestions to incentivize collaboration between data scientists and marine geoscientists at the proposal stage would help bridge the existing gap in the marine data science lifecycle. Efforts for marine geoscientists to become "data literate" beyond the immediate needs of their own datasets are already underway

through organizations such as Community Surface Dynamics Modeling System (CSDMS) and the Research Data Alliance (Berman et al., 2014).

An example of data-driven marine geoscience can be found in recent machine learning efforts that provide both marine geoscience analyses and identify parametrically unique regions to sample (e.g., Lee et al., 2019; Graw et al., 2020). Analyses such as these pinpoint regions of geologic interest, instead of geographic interest, that are ideal for further data collection. We believe that using the data to inform future data acquisition is the next great Frontier of the marine geosciences, allowing the data to drive future collection and making the data lifecycle come full circle (Figure 3).

Summary

In order to move the marine geosciences into the "big data" era, the three stages of the data lifecycle (acquisition, curation, and integration) need to be deliberately linked. Many challenges discussed herein are due to the recent exponential increase in volume and variety of marine geoscience data (Figure 1), with only incremental changes in how data are handled. Below is a brief summary of our opinion regarding the three largest issues facing the marine geoscience community in the movement towards the "big data" era:

1) Contemporary data acquisition is both geospatially and anomaly focused resulting in biased observational datasets.
2) There are not enough incentives for data acquirers to do more with their data than meet basic funding agency guidelines (i.e., depositing their data "as-is" in a repository).
3) Data integration is currently performed as a largely standalone effort, instead of a coordinated effort between data acquirers and curators.

In this paper, we outline possible steps to address these problems, which can be summarized as:

1) Utilize autonomous research platforms, and fund systematic/ exploratory data efforts to collect data in a less biased manner.
2) Incentivize data curation through research proposal evaluation rubrics and citable/publishable databases and journals.
3) Utilize data-driven sampling methodologies, such as parametric sampling, and "cross-train" marine geoscientists in all three phases of the data lifecycle.

The solutions proposed above will not singlehandedly deliver the marine geoscience community to the 'big data' era. However, we believe these solutions are tangible steps to make the marine geoscience community capable of handling the acquisition, curation, and integration of the data we have today and better face the data challenges of the coming decades. Diepenbroek et al., 2002,Diesing, 2020,Rew and Davis, 1990.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

TL and BP prepared the figures. Authors TL, BP, and JO contributed equally to the discussion and writing of this manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

The views and opinions expressed here are those of the authors only and do not necessarily reflect the views or positions of the entities they represent.

References

Agena, W. F., Hutchison, D. R., Lee, M. W., and Oliver, H. L. (1993). Cape Hatteras to the Mid-Atlantic Ridge – demultiplexing and archiving a unique multichannel seismic reflection data set. *U.S. Geol. Surv. Open-File Rep.* 93-264.

Berman, F., Wilkinson, R., and Wood, J. (2014). Guest editorial: Building global infrastructure for data sharing and exchange through the Research Data Alliance. *D-Lib Mag.* 20 (1/2). doi:10.1045/january2014-berman

Chase, R. J., Harrison, D. R., Burke, A., Lackmann, G. M., and McGovern, A. (2022). A machine learning tutorial for operational meteorology. Part I: Traditional machine learning. *Weather Forecast.* 37, 1509–1529. doi:10.1175/waf-d-22-0070.1

Cooperdock, E. H. G., Chen, C. Y., Guevara, V. E., and Metcalf, J. R. (2021). Counteracting systemic bias in the lab, field, and classroom. *AGU Adv.* 2. doi:10.1029/2020AV000353e2020AV000353

Davies, J., and Davies, D. (2010). Earth's surface heat flux. *Solid earth.* 1, 5–24. doi:10.5194/se-1-5-2010

Davies, J. H. (2013). Global map of solid Earth surface heat flow. *Geochem. Geophys. Geosystems* 14 (10), 4608–4622. doi:10.1002/ggge.20271

Diepenbroek, M., Grobe, H., Reinke, M., Schindler, U., Schlitzer, R., Sieger, R., et al. (2002). PANGAEA—An information system for environmental sciences. *Comput. Geosciences* 28 (10), 1201–1210. doi:10.1016/S0098-3004(02)00039-0

Diesing, M. (2020). Deep-sea sediments of the global ocean. *Earth Syst. Sci. Data* 12, 3367–3381. doi:10.5194/essd-12-3367-2020

Diesing, M., Thorsnes, T., and Bjarnadóttir, L. R. (2021). Organic carbon densities and accumulation rates in surface sediments of the North Sea and Skagerrak. *Biogeosciences* 18, 2139–2160. doi:10.5194/bg-18-2139-2021

Dixon, M. F., Halperin, I., and Bilokon, P. (2020). *Machine learning in finance.* Berlin/Heidelberg, Germany: Springer.

Fuchs, S., and Norden, B. (2021). "International heat flow commission. Data from," in *The global heat flow database: Release 2021* (Potsdam, Germany: GFZ Data Services). doi:10.5880/fidgeo.2021.014

Graw, J. H., Wood, W. T., and Phrampus, B. J. (2020). Predicting global marine sediment density using the random forest regressor machine learning algorithm. *J. Geophys. Res. Solid Earth* 126, 1. doi:10.1029/2020JB020135e2020JB020135

Hasterok, D., and Chapman, D. S. (2011). Heat production and geotherms for the continental lithosphere. *Earth Planet. Sci. Lett.* 307 (1-2), 59–70. doi:10.1016/j.epsl.2011.04.034

Kopf, A. (2009). *The deep-sea and sub-surface frontier initiative – a key link EC research and international scientific ocean drilling.* Vienna, Austria: Copernicus Publications. EGU General Assembly EGU2009-9434. Abstract retrieved from SAO/NASA Astrophysics Data System and the DS3F Steering Committee Team

Lee, T. R., Wood, W. T., and Phrampus, B. J. (2019). A machine learning (kNN) approach to predicting global seafloor total organic carbon. *Glob. Biogeochem. Cycles* 33, 37–46. doi:10.1029/2018GB005992

Marchese, F., Purkis, S., Chimienti, G., Ouhssain, M., Shernisky, H., Terraneo, T., et al. (September 2022). A baseline assessment of seafloor geomorphology and benthic habitat distribution along the Neom Coast (Northern Saudi Arabia, Red Sea). Proceedings of the 10th International Conference on Geomorphology ICG2022-705. Coimbra, Portugal. doi:10.5194/icg2022-705

Mayer, L., Jakobsson, M., Allen, G., Dorschel, B., Falconer, R., Ferrini, V., et al. (2018). The nippon foundation - GEBCO seabed 2030 project: The quest to see the world's oceans completely mapped by 2030. *Geosci.* 8 (2). doi:10.3390/geosciences8020063

McKenna, L., Cantwell, K., Kennedy, B., Elliott, K., Lobecker, E., and Sowers, D. (2015). "Exploring deep sea habitats for baseline characterization using NOAA ship Okeanos Explorer," in *Abstract retrieved from center for coastal and ocean mapping joint hydrographic center* (Chicago Illinois: American Geophysical Fall Meeting).

Phrampus, B. J., Lee, T. R., and Wood, W. T. (2020). A global probabilistic prediction of cold seeps and associated SEAfloor Fluid Expulsion Anomalies (SEAFLEAs). *Geochem. Geophys. Geosystems* 21 (1). doi:10.1029/2019GC008747e2019GC008747

Restreppo, G. A., Wood, W. T., Graw, J. H., and Phrampus, B. J. (2021). A machine-learning derived model of seafloor sediment accumulation. *Mar. Geol.* 440, 106577. doi:10.1016/j.margeo.2021.106577

Rew, R. K., and Davis, G. P. (1990). NetCDF: An interface for scientific data access. *IEEE Comput. Graph. Appl.* 10 (4), 76–82. doi:10.1109/38.56302

Rillo, M. C., Kucera, M., Ezard, T. H. G., Miller, C. G., Johnson, R., and Grear, J. (2019). Surface sediment samples from early age of seafloor exploration can provide a late 19th century baseline of the marine environment. *Front. Mar. Sci.* 5, 1–15. doi:10.3389/fmars.2018.00043

Sahoo, A., Dwivedy, S. K., and Robi, P. S. (2019). Advancements in the field of autonomous underwater vehicle. *Ocean. Eng.* 181, 145–160. doi:10.1016/j.oceaneng.2019.04.011

SEG Technical Standards Committee (2017). *SEG-Y_r2.0: SEG-Y revision 2.0 data exchange format.*

Skarke, A., Ruppel, C., Kodis, M., Brothers, D., and Lobecker, E. (2014). Widespread methane leakage from the sea floor on the northern US Atlantic margin. *Nat. Geosci.* 7 (9), 657–661. doi:10.1038/ngeo2232

Stein, C. A., and Stein, S. (1992). A model for the global variation in oceanic depth and heat flow with lithospheric age. *Nature* 359 (6391), 123–129. doi:10.1038/359123a0

Watts, A. B. (2021). Reflections on a career in marine geoscience. *Perspect. Earth Space Sci.* 2, 1. doi:10.1029/2021CN000144e2021CN000144

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 160018. doi:10.1038/sdata.2016.18