AGU ADVANCING EARTH AND SPACE SCIENCES

The LILY Database: Linking Lithology to IODP Physical, Chemical, and Magnetic Properties Data

Laurel B. Childress[1] , **Gary D. Acton[1]** , **Vincent P. Percuoco[1]**, and **Margaret Hastedt[1]**

[1]Texas A&M University, International Ocean Discovery Program, College Station, TX, USA

Abstract During each expedition of the International Ocean Discovery Program and its precursor, the Integrated Ocean Drilling Program (jointly referred to as IODP), vast arrays of data are collected from drill cores. These data, which are accessible from the IODP LIMS (Laboratory Information Management System) database, include physical, chemical, and magnetic properties collected semi-continuously along cores using automated track systems, as well as a variety of analyses conducted on discrete subsamples taken from the cores. In addition, the lithology of all cores is described based on visual characteristics of the surface of split cores, visual examination of smear slides and thin sections, and compositional or mineralogical information derived from geochemical analyses. We extract basic lithologic information from this complex array of descriptive information and then tie that information to all other measurements. This new database is referred to as **L**IMS with **L**itholog**y** (LILY). LILY currently contains over 34 million data from 89 km of core recovered on 42 expeditions conducted 2009–2019. Some uses of LILY include identifying the abundance of different lithologies, finding data from core intervals with a specific lithology, assessing the efficacy of coring systems in different lithologies, or characterizing and analyzing physical, chemical, and magnetic properties based on lithology. We illustrate the use of LILY by computing the grain density by lithology from over 24,000 moisture and density measurements and then use those grain densities, along with the large IODP bulk density dataset, to compute a new high-resolution porosity dataset with over 3.7 million new porosity estimates.

Plain Language Summary During expeditions on the drillship *JOIDES Resolution*, large amounts of data are collected along drill cores. While stored in a singular database (LIMS), these data are difficult to associate with the lithology (type of rock or sediment) from which they were collected because of the different types of data that have been stored and formatted. For example, density is measured at high resolution using a semi-automated track system, while lithologic description intervals are entered by hand at resolutions varying from 1 to 150 cm and the terms applied to these descriptive data vary considerably for each expedition. We organize the descriptive information so that it can be easily paired with all the types of data in LIMS. This LIMS with Lithology (LILY) database brings the expedition data and primary lithologic description information into a singular and easily accessible format. We illustrate how LILY can be used by determining the most common lithology cored, by estimating the density for each lithology, and by computing a new high-resolution porosity dataset from the density data.

1. Introduction

1.1. General Overview

The standardized collection of data during each Integrated Ocean Drilling Program or International Ocean Discovery Program (IODP) expedition includes a wide variety of physical, chemical, and magnetic properties data, digital images, and descriptive information. Most of these data are collected every few centimeters along the cores using automated track systems. Additional datasets are constructed from a variety of analyses on discrete subsamples taken from the cores. Some of the properties typically measured include P-wave velocity, bulk density, grain density, porosity, magnetic susceptibility, natural remanent magnetization, natural gamma radiation, thermal conductivity, and visible spectral reflectance as well as a wide range of geochemical properties (Table S1 provides a full list of data types; Table S2 provides data quantity). In addition to these measurements, descriptive information is recorded continuously. Ten years of digital tabular descriptive information now exists for 42 expeditions, representing data from multiple ocean basins and geological time.

Since 2009, the DESClogik software system has been used onboard the *JOIDES Resolution* to collect descriptive data. DESClogik is a software package developed internally by IODP that includes a tabular interface to enter data

Project administration: Laurel B. Childress, Gary D. Acton
Resources: Laurel B. Childress, Gary D. Acton, Vincent P. Percuoco, Margaret Hastedt
Software: Laurel B. Childress, Gary D. Acton, Vincent P. Percuoco, Margaret Hastedt
Supervision: Laurel B. Childress, Gary D. Acton
Validation: Laurel B. Childress, Gary D. Acton, Vincent P. Percuoco, Margaret Hastedt
Visualization: Laurel B. Childress, Gary D. Acton, Vincent P. Percuoco, Margaret Hastedt
Writing – original draft: Laurel B. Childress, Gary D. Acton
Writing – review & editing: Laurel B. Childress, Gary D. Acton, Vincent P. Percuoco, Margaret Hastedt

into separate spreadsheets that are used to capture macroscopic (e.g., texture, structure) and microscopic (e.g., thin section and smear slide) observations (Expedition 320T Scientists, 2009; additional information about DESClogik is available at https://wiki.iodp.tamu.edu/display/LMUG/Core+Description). Within each spreadsheet, each column is used for a different type of observation (e.g., lithology, color, drilling disturbance, structure, and bioturbation). Pre-defined terms available in customizable "value lists" can be used to further qualify those observations. From this extensive dataset, subsets of data are presented in visual core descriptions (VCDs, a.k.a. "barrel sheets"), such as Hole 341-U1417C, Core 17H VCD of the Expedition 341 Proceedings Volume (Jaeger, Gulick, Asahi, et al., 2014; http://publications.iodp.org/proceedings/341/EXP_REPT/CORES/CORU1417.PDF#page=83).

There is currently no convenient method to compare, combine, or analyze DESClogik data, particularly lithology data, across multiple expeditions or to integrate descriptive data with the other track and discrete measurements collected onboard. Most shipboard data collected onboard the *JOIDES Resolution* are tabularly stored in the Laboratory Information Management System (LIMS) database. Descriptive information is an exception; it is available within the LIMS database in multi-tab Excel files, which complicates directly pairing the lithologic descriptions with other data.

The subjective nature of lithologic descriptions further complicates the pairing process, as the same lithology may not be described consistently from one expedition to the next. Inconsistencies in descriptions may even exist within the same expedition. Furthermore, lithologic terminology evolves over time. An additional complication is that lithologic descriptions are not made on the same scale as other observations. For example, each measurement made along a core section by a track system is representative of a distinct spot or narrow interval and data from each discrete sample are representative of a small volume of material, while descriptive information for a core section frequently includes large intervals (up to an entire ~1.5 m section).

In this study, we extract the lithologic information from the complex array of descriptive information and then tie that information to other observations made on the cores in order to characterize the physical, chemical, and magnetic properties of a myriad of lithologies. The primary products are consistently cleaned, vetted, and organized tables of lithologies for each expedition and tables of all the other observations with a lithology attached. In these tables, each observation also contains the expedition, site, hole, core, core type, section, offset (interval), and depth in meters below seafloor (mbsf; referred to as CSF-A[m] in LIMS) and has associated metadata derived from the Hole and Core Summary files, including geographic coordinates, water depth, and additional coring type information (coring types are listed in Table S3). This database, which we refer to as LIMS with LithologY (LILY), contains millions of observations, some that are made on the scale of tens of microns to millimeters and others that are on the scale of centimeters to meters.

LILY may be used in a variety of investigations including simply determining the characteristics of different lithologies. For instance, what is the magnetic susceptibility of nannofossil ooze and how does it differ from clay or diatom ooze? While one might be able to find some ranges of magnetic susceptibilities for common rock types with an internet search or in geophysical textbooks, those ranges are generally based upon a few tens to a few hundreds of measurements for a limited number of general lithologies. There are over 6.8 million magnetic susceptibility measurements in LILY and over 98 principal lithologies, each with more than 1,000 susceptibility observations. Other studies might seek to investigate relationships between physical properties as a function of lithology, depth, latitude, tectonic province, or other criteria. For example, how does porosity vary with density for different lithologies? How does overburden affect seismic P-wave velocity for different lithologies? In what lithologies is there the largest P-wave velocity anisotropy? How does the lightness (L^*) vary with carbonate content as a function of lithology and are the relationships dependent on location and paleoclimate, for example, intervals that are lighter and have higher carbonate content are associated with warmer interglacial intervals in the north Atlantic over the past 3 million years (Stein et al., 2006, 2009).

With lithologies paired with over 34 million track/discrete data—including digital image red/green/blue (RGB) color, visible spectral reflectance, magnetic susceptibility, natural remanent magnetization (NRM) before and after demagnetization treatment, P-wave velocity, natural gamma radiation (NGR), density (wet and dry), porosity, thermal conductivity, carbonate content, major element oxides and minor element content from ICP (inductively coupled plasma) measurements, and other chemical observations—LILY is a substantial database that can be probed with big data analytics, business intelligence, and machine learning methods to allow users to extract the information or attributes they desire. Below, we provide an example of this using grain and bulk

densities to derive over 3.7 million new porosity estimates and characterize the densities and porosities by lithology. We chose this example because calculating porosities from bulk densities requires knowledge of the grain density, which in turn depends on the composition and hence the lithology of each sample.

1.2. Background on Lithologic Naming Conventions

From the Deep Sea Drilling Project (DSDP) to current IODP expeditions, lithologic naming conventions have evolved. In addition, the detail to which one science party describes the lithology of the recovered material varies depending on the goals of the expedition, the size and expertise of the science party, and the volume of material recovered. Since DSDP Leg 2, most legs or expeditions have provided an explanation of their lithologic naming conventions in an "Explanatory Notes" or "Methods" section of their "Initial Reports" or "Proceedings Volume" (e.g., Huber et al., 2019; The Shipboard Scientific Party, 1970), all of which are available online at http://publications.iodp.org. The understanding of lithologic naming conventions as well as other core description practicalities is important for the appropriate use of IODP descriptive data (e.g., Marsaglia & Milliken, 2023).

Although not ubiquitously or consistently used from one expedition to the next, a common theme for sedimentary rock classification for IODP expeditions has been to follow the guidelines in Mazzullo et al. (1988), with exceptions that generally entailed devising alternative naming conventions for the "mixed sediment" classification of Mazzullo et al. (1988). While the basic descriptive terminology varies amongst the expeditions, lithologies for granular sedimentary rocks generally have been described with a principal name that may have major modifiers given as prefixes and minor modifiers as suffixes. For sedimentary rocks, the prefix (major modifier) generally infers a composition that is >25% and suffixes (minor modifiers) are for compositions between 10% and 25% (Mazzullo et al., 1988). The principal lithology is representative of the dominant composition (>60%) derived from one of four types of granular sedimentary rocks (pelagic, neritic, siliciclastic, or volcaniclastic). For example, an unconsolidated sediment that is 40% silt + 35% clay (total siliciclastic = 65%) + 25% nannofossil grains would have a principal name = silt, because silt is the dominant component of the siliciclastics, which comprise >60% of the total composition. The prefix or major modifier would be "clayey" and the suffix or minor modifier would be "with nannofossils", resulting in a full lithologic name of "clayey silt with nannofossils". When none of the four granular types comprised >60% of the total, the term mixed sediment was used by Mazzullo et al. (1988). The "mixed sediment" principal lithology is not, however, used by any of the expeditions covered in this study. Instead, most expeditions have derived principal names from the granular sedimentary rock type that represents >50% of the composition. Given the difference in usage and some inherent ambiguities in the nomenclature, full lithologic names at best provide semiquantitative estimates of the amount of each component in a sample. For example, a lithology such as a clayey sand with foraminifera is dominantly sand with >25% clay and 10%–25% foraminifera. Given the percentages for the major and minor modifiers, sand comprises 40%–65% of the composition and clay must also be <45%.

Given this terminology, lithologic descriptions have been stored in the DESClogik "macroscopic" spreadsheets in three main columns or categories, which we refer to as "prefix", "principal", and "suffix", with a fourth column that gives the "full lithology" that is a combination of the prefix, principal, and suffix. Not every expedition uses all three columns, nor do they all use them in the same manner.

Some expeditions (e.g., Expeditions 318, 341, 352, 363, 366, 374) follow or partly follow the classification system of Dean et al. (1985) or some modified version of that system, in which they include major modifiers (25%–50% of the composition) and minor modifiers (10%–25% of the composition) in the prefix, and differentiate them by adding "-rich" or nothing for major modifiers and by adding "-bearing" for minor modifiers. The principal lithology name is determined from the dominant (>50%) non-biogenic or biogenic component. This system does not use the "suffix" category. A lithology with the prefix = "radiolarian-bearing nannofossil-rich" and principal = "diatom ooze", full lithology = "radiolarian-bearing nannofossil-rich diatom ooze" would indicate the ooze was between 10% and 25% radiolaria, 25%–40% nannofossils, and >50% diatoms (Note that nannofossils cannot be 25%–50% in this example because of the relative abundance of radiolaria and diatoms). In the unmodified version of the Dean et al. (1985) classification, the "-rich" was not added to a prefix of a component that comprises 25%–50% of the composition, and if more than two major modifiers were present, they were listed in order of increasing abundance. These practices result in ambiguities, for example, if the full lithology is given as "radiolarian-bearing nannofossil diatom ooze", diatoms may comprise 25%–50% if they are a major modifier of "ooze" or >50% if they are part of the principal lithology "diatom ooze".

Confusion can be caused when some descriptions include a combination of methods that do not agree with what is stated in the "Methods" section of the associated Proceeding Volume. Several expeditions have minor modifiers in both the prefix and suffix columns even though their methods state that all minor modifiers will be placed in the prefix with "-bearing" added. For example, a lithology description that violates the classification systems of both Dean et al. (1985) and Mazzullo et al., 1988) is given as a prefix = "foraminifer bearing", principal = "diatom ooze", and suffix = "with radiolarians". In this example, a hyphen is missing in the prefix of the minor modifier and a second minor modifier is listed in the suffix, which presumably should have been in the prefix as "radiolaria-bearing." These types of inconsistencies or ambiguities are not uncommon and so we have developed some conventions of our own that hopefully add some consistency without altering the intent of the original lithologic descriptions, the details of which are given in Text S2 in Supporting Information S1 and Table S4.

Igneous rock classification, which is based on composition and texture, can also vary considerably between expeditions. As might be expected, basalt is by far the most common principal lithology for igneous rocks, but there are many compositional and textural terms applied to this base lithology. Prefixes are commonly used to define phenocryst composition, abundance, and/or size, for example, highly olivine-phyric basalt, fine-grained basalt, and aphyric cryptocrystalline basalt. Again the "Explanatory Notes" or "Methods" sections of the expedition publications provide the specifications for the nomenclature. For example, Expedition 335 defines aphyric cryptocrystalline basalt as basalt with <1% phenocrysts and with phenocryst grain size <0.1 mm (Expedition 335 Scientists, 2012). Suffixes are applied more rarely and have commonly been used for flow morphology, for example, "sparsely plagioclase phyric basalt pillow lava flow".

2. Materials and Methods

2.1. Methods Summary

Multiple script-based transformations are applied to data downloaded from IODP data repositories to generate the LILY database (Figure 1). All data used in this analysis are available from LORE (LIMS Online Reports; https://web.iodp.tamu.edu/LORE/). An exception to this is lithologic data for Expeditions 318 and 321, which are incomplete in LORE and are instead derived from IODP Publications archives and are provided in Zenodo as part of the full LILY database in Childress et al. (2023); Data Set S1, https://doi.org/10.5281/zenodo.8408297.

2.2. Data Gathering

DESClogik, a core and fossil description application, was first tested onboard the *JOIDES Resolution* in 2009 during Expedition 320T, although no data were archived, and then first implemented during Expedition 320; it was later updated in 2011 during Expedition 335 (sometimes generically known as DESClogik 2.0). DESClogik data from IODP expeditions that are out of moratorium are available at http://web.iodp.tamu.edu/DESCReport/. We use data from the 42 expeditions listed in Table 1, all of which were conducted between March 2009 and March 2019 and the last of which was out of moratorium as of March 2021 (Figure 2). Descriptive information is organized at the expedition-site-hole level, resulting in a separate final product Excel document for each IODP hole drilled. For some expeditions, multiple versions of the Excel document of descriptive information are available for a single hole, usually as the result of corrections made to descriptive information during post-expedition editorial meetings. In the case of multiple versions of the same descriptive file, the most recently dated file is used.

The descriptive information is available as Microsoft Office Excel format files (including .xls and .xlsx) containing multiple sheets within each file, which include descriptive information such as drilling disturbance, lithology, structure, etc. While multi-sheet Excel file formats are a convenient method for storing tabular information, this format can present obstacles to data mining, and the naming of these files is inconsistent across the downloadable LORE dataset. For example, an export of lithologic data from Expedition 320 is given the filename "320 Core Description_U1332B". while an export from Expedition 363 is given the filename "X363-U1490B-macroscopic". Although not detrimental to the integrity of the data contained within, these inconsistencies create data mining hurdles. To more easily access individual data and remove unnecessary formatting, the R software environment (R Core Team, 2023) is used to extract descriptive data and, as described later, to pair descriptive and measured data.
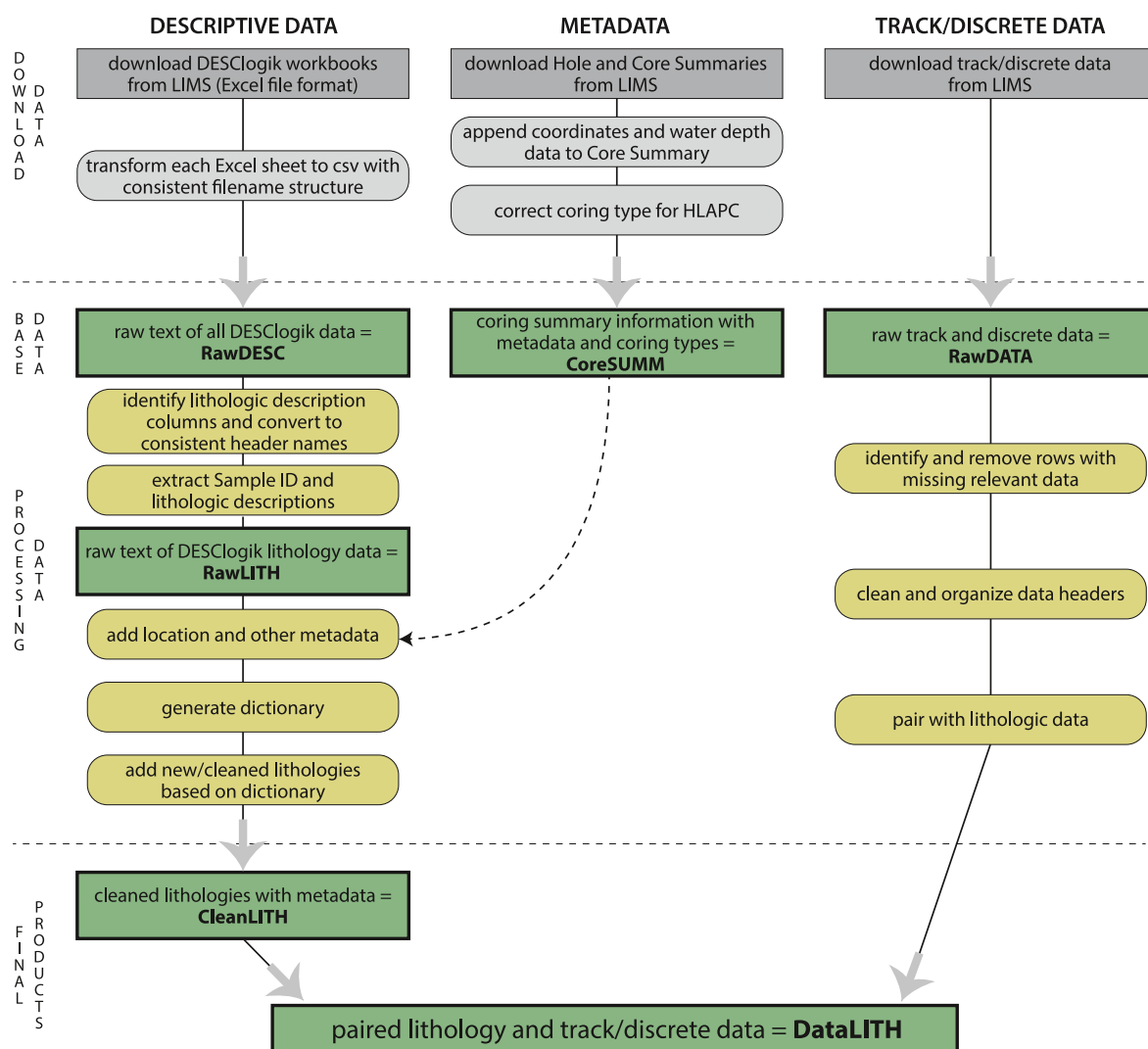
Figure 1. Schematic of data management.

To create independently accessible descriptive textual files that contain data records in a tabular format, each sheet of each Excel file was extracted, converted to comma separated value (CSV) format, and a consistent naming convention was applied (Table S5). These exported data are available as LILY-RawDESC. LILY-RawDESC files contain the raw text content of the Excel worksheet without applying any corrections or conversions. They thus have all the blemishes of the original observations but provide a raw, searchable dataset for all DESClogik data. However, because this is the raw data, errors can include misspelled words, undefined acronyms, extra carriage returns, extra spaces, improper punctuation, and erroneous sample identification information (also referred to as "scope of description" in DESClogik terminology), and are sometimes derived from worksheets with odd naming conventions. Because DESClogik templates for data capture are often reused across expeditions, many Excel sheets (and thus resulting CSV products) contain no data because the descriptive category was irrelevant to a later expedition and ignored rather than removed. These inclusions are less noticeable in the Excel-sheet format of the original data but, in the CSV format, many files contain only header information. These empty datasets are left in the LILY-RawDESC collection for completeness and transparency to the user that data have not been removed.

We devise a naming convention for the LILY-RawDESC files that allows them to be related to the original worksheet from which they were derived, while also containing a consistent convention that aids users in identifying the type of data in the files. The naming convention for the CSV dataset is formatted as

Table 1
Expeditions Included in LILY

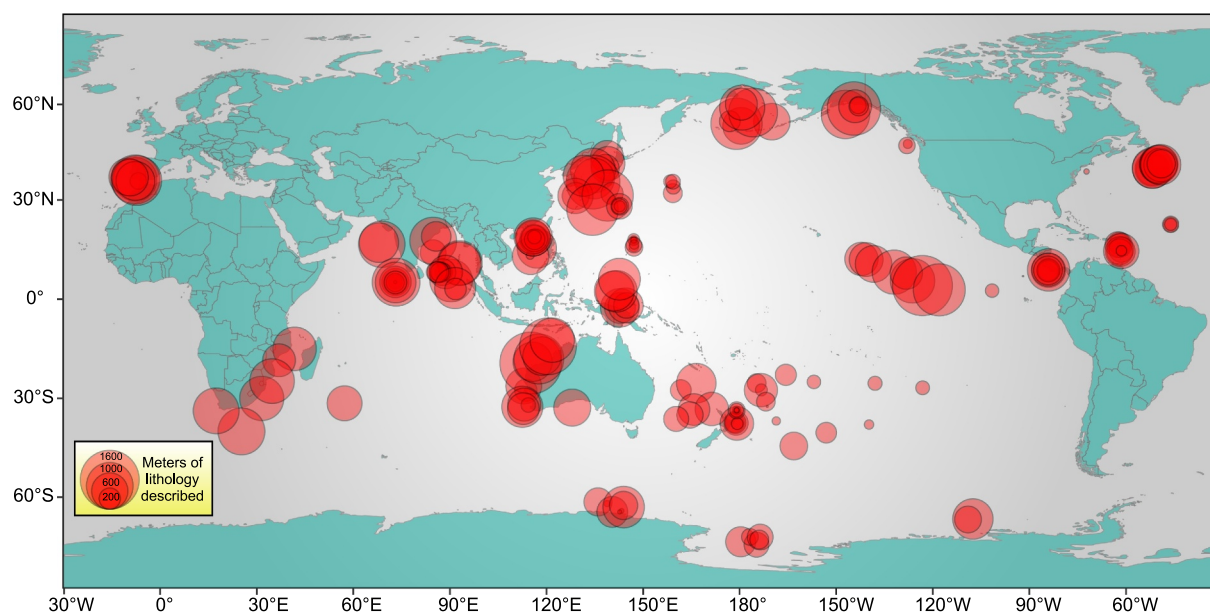| Expedition number	Expedition name	Dates	Reference
318	Wilkes Land Glacial History	January–March 2010	Escutia et al. (2011)
320	Pacific Equatorial Age Transect	March–May 2009	Pälike et al. (2010)
321	Pacific Equatorial Age Transect	May–July 2009	Pälike et al. (2010)
323	Bering Sea Paleoceanography	July–September 2009	Takahashi et al. (2011)
324	Shatsky Rise Formation	September–November 2009	Sager et al. (2010)
327	Juan de Fuca Ridge-Flank Hydrogeology	July–September 2010	Fisher et al. (2011)
329	South Pacific Gyre Subseafloor Life	October–December 2010	D'Hondt et al. (2011)
330	Louisville Seamount Trail	December 2010–February 2011	Koppers et al. (2012)
334	Costa Rica Seismogenesis Project (CRISP-A1)	March–April 2011	Vannucchi et al. (2012)
335	Superfast Spreading Rate Crust 4	April–June 2011	Teagle et al. (2012)
336	Mid-Atlantic Ridge Microbiology	September–November 2011	Edwards et al. (2012)
339	Mediterranean Outflow	November 2011–January 2012	Stow, Hernández-Molina, Alvarez Zarikian, and the Expedition 339 Scientists (2013)
340	Lesser Antilles Volcanism and Landslides	March–April 2012	Le Friant et al. (2013)
341	Southern Alaska Margin	May–July 2013	Jaeger, Gulick, LeVay, and the Expedition 341 Scientists (2014)
342	Paleogene Newfoundland Sediment Drifts	June–July 2012	Norris et al. (2014)
344	Costa Rica Seismogenesis Project (CRISP-A2)	October–December 2012	Harris et al. (2013)
345	Hess Deep Plutonic Crust	December 2012–February 2013	Gillis et al. (2014)
346	Asian Monsoon	July–September 2013	Tada et al. (2015)
349	South China Sea Tectonics	January–March 2014	Li et al. (2015)
350	Izu-Bonin-Mariana Rear Arc	March–May 2014	Tamura et al. (2015)
351	Izu-Bonin-Mariana Arc Origins	May–July 2014	Arculus et al. (2015)
352	Izu-Bonin-Mariana Forearc	July–September 2014	Reagan et al. (2015)
353	Indian Monsoon Rainfall	November 2014–January 2015	Clemens et al. (2016)
354	Bengal Fan	January–March 2015	France-Lanord et al. (2016)
355	Arabian Sea Monsoon	March–May 2015	Pandey et al. (2016)
356	Indonesian Throughflow	July–September 2015	Gallagher et al. (2017)
359	Maldives Monsoon and Sea Level	September–November 2015	Betzler et al. (2017)
360	SW Indian Ridge Lower Crust and Moho	November 2015–January 2016	MacLeod et al. (2017)
361	South African Climates (Agulhas LGM Density Profile)	January–March 2016	Hall et al. (2017)
362	Sumatra Seismogenic Zone	August–October 2016	McNeill et al. (2017)
363	Western Pacific Warm Pool	October–December 2016	Rosenthal et al. (2018)
366	Mariana Convergent Margin & S Chamorro Seamount	December 2016–February 2017	Fryer et al. (2018)
367	South China Sea Rifted Margin	February–April 2017	Sun et al. (2018)
368	South China Sea Rifted Margin	April–June 2017	Sun et al. (2018)
368X	Return to Hole U1503A (South China Sea)	November–December 2018	Childress et al. (2020)
369	Australia Cretaceous Climate and Tectonics	September–November 2017	Hobbs et al. (2019)
371	Tasman Frontier Subduction Initiation and Paleogene Climate	August–September 2017	Sutherland et al. (2019)
372	Creeping Gas Hydrate Slides and Hikurangi LWD	November–December 2017	Wallace et al. (2019)
374	Ross Sea West Antarctic Ice Sheet History	January–March 2018	McKay et al. (2019)
375	Hikurangi Subduction Margin	March–May 2018	Wallace et al. (2019)
376	Brothers Arc Flux	May–July 2018	de Rhonde et al. (2019)
379	Amundsen Sea West Antarctic Ice Sheet History	January–March 2019	Gohl et al. (2021)

Figure 2. Geographic location of sites for the 42 expeditions. Symbol size is proportional to the meters of core described for each site.

"XEXP_SHEETNAME_type_type2_SITEHOLE" where: X is an acronym for Expedition; EXP = expedition number; SHEETNAME = the original Excel file sheet name; type = a low-order data designation (ex.—radiolaria, sediment, diatom) based on the original filename; type2 = a high-order data designation of "microscopic", "macroscopic", or "paleo"; and SITEHOLE = the conventional IODP Site-Hole concatenation (ex.—U1417A for Site U1417, Hole A). If no low-order designation is provided by the filename, the high-order designation is repeated. For example, X374_General_macroscopic_macroscopic_U1524A.csv contains the macroscopic lithostratigraphy for Hole U1524A cored during Expedition 374. All CSV files can be traced to the Excel file of origin using Table S5. For example, the file X374_General_macroscopic_macroscopic_U1524A.csv was derived from the Excel file 374_U1524A_macroscopic.xlsx downloaded from LIMS.

2.3. Data Transformation and Management—Lithology

One of the most consistently collected and useful descriptive parameters in the DESClogik data is lithology. While lithology is recorded for cores from all expeditions, it can be stored across multiple sheets of the Excel files. For example, an expedition that recovers sedimentary, igneous, and metamorphic rocks may record the lithology of these materials across three separate sheets within the description file(s).

Additionally, DESClogik has not required a consistent header name for data entry of lithology. For example, the principal lithology column across expeditions may be written as "Principal lithology", "Lithology principal name", "PRINCIPALLITHOLOGY", "LithologyPRINCIPAL NAME", "Lithology", "Lithology 1 principal", "MAJ Lith. Principal name", "MAJ Principal lithology B", and a variety of variations of these combinations. In older data (prior to Expedition 339), lithology headers can vary hole-to-hole within a single expedition. Lithology prefix and suffix column headers vary similarly in the wide variety of styles and formats used for header column names. Some expeditions do not include a prefix and/or suffix column in the lithologic description and even within a single expedition, this may not be consistent across multiple description type sheets (ex.—sediment, metamorphic). Attempts to identify lithology columns by searching for a string or phrase such as "Lith" in the header are hindered by the presence of many other non-lithology column headers that include variations of the term "lithology," such as "Lith. Color" or "Lithologic accessories". The formatting of the column headers and column order for sample identification information (Expedition, Site, Hole, Core, etc.) is equally diverse across expeditions, and occasionally within a single expedition. Therefore, each sheet (as CSV) in LILY-RawDESC is processed separately to identify lithology columns (a list of these is given in Table S6), extract the principal lithologies along with major and minor modifiers, transform these observations into a consistent format, and add associated identification information and metadata to generate the LILY-RawLITH dataset (Figure 1). In order to

generate a consistently formatted dataset, decisions had to be made on how to handle cases in which there was no clear principal lithology (either with or without major and minor modifiers) in the DESClogik data. For example, problems arise due to inconsistent use of the term "principal lithology". In some expeditions, "principal lithology" is referred to as "major lithology", although without the presence of a "minor lithology". Note that this "major lithology" is not the "major modifier" of a principal lithology; it is instead used in some expeditions in place of "principal lithology". Other expeditions employ a scheme of "major lithology" with some records of "minor lithology". For example, a lithologic description for a section of core may be given as having a major lithology that is "nannofossil ooze with diatoms" and a minor lithology as "diatom ooze with nannofossils". Here, the term "major lithology" indicates the most common lithology in an interval that has varying amounts of nannofossils and diatoms, with nannofossils being more common than diatoms for most of the described interval. In this case, we would use only the "major lithology" and record the "principal lithology" as "nannofossil ooze" and the "minor suffix" as "with diatoms" in the LILY-RawLITH dataset. Minor lithologies, such as those described above, are uncommonly recorded; they are found in only nine expeditions. These minor lithologies are extracted and included in LILY-RawLITH for availability to other researchers but are not used in the lithology analysis presented here.

Another issue arises because some DESClogik datasets (Table S6) contain multiple principal lithologies for the same depth interval and an additional column indicating the percentage of material of each lithology is included. In instances of multiple principal lithologies for the same interval, the dominant (largest percentage value) lithology is retained. Fundamentally, only lithologies ≥50% should be retained, as this would be equivalent to a "principal lithology". Multiple expeditions, however, include three or more principal lithologies for the same description interval, and all percentages provided are <50%. In cases where multiple principal lithologies for the identical description interval are present with equivalent greatest percentage values (ex. 50%–50% or 40%–40%), each entry is kept because no clear principal lithology can be determined. For example, consider an interval within a section that, rather than having a single row description in the DESClogik table with a principal lithology and with major and minor modifiers, is instead described in three rows by three different principal lithologies and each lithology has been assigned a percentage: "nannofossil ooze" = 40%, "clay" = 40%, and "mud" = 20%. In this case, there is no clear singular principal lithology, and both the "nannofossil ooze" and "clay" are retained as the principal lithology for the interval. This example is even more insidious because mud is a mixture of clay and silt; clay is mud, silt is mud, and mud would preferably never occur in lithologic descriptions that include silt or clay. Such terminology, however, does occur in the original descriptions, including principal lithologies given as "silty mud", "muddy clay", and "muddy silt".

2.4. Data Cleaning and Metadata Additions—Lithology

Description intervals with erroneous, missing, or corrupted identification information are common. Some errors in identification information, such as a randomly missing "A" for archive half in a single row of description, are reasonably obvious. Other errors are more complicated, such as several expeditions where pieces are described by the correct mbsf (CSF-A m) values but incorrect or ambiguous cm offset values. The most common of these are from the ambiguous use of "offset", which is the distance measured from the top of a sample, where the sample can be a section, piece, or subsample of a sample. Failure to identify the sample clearly means the distance given is measured from an unknown reference point. Generally, "offset" is used in IODP observations like the term "interval" was used for DSDP and ODP, in which "interval" is always a measure of the distance from the top of a section. To ensure consistent use of "offset" across all the expeditions, we only use offset as the distance from the top of a section and correct any of the originally ambiguous entries. Other erroneous description intervals include missing section values, descriptions at the core level that need to be expanded to the section level, missing cm offset values, missing mbsf (CSF-A m) values, piece identification information located in the wrong column, section identification information located in the wrong column, and rows that are otherwise flawed. Some of these errors are generated by the methods used to extract data from the CSV files. For example, many expeditions do not explicitly include identification information by column and so we parse it from the Sample ID during the lithology extraction process. The wide variety and format of Sample IDs further complicates the automation of the extraction process, requiring manual intervention. All noted identification errors are corrected by script.

To assess the completeness of lithologic descriptions, overlaps and omissions (gaps) in description are determined for LILY-CleanLITH. Overlaps in lithologic description are defined in this context as intervals that are associated with two or more lithologic descriptions. An exception to this is made for identical description intervals with

different descriptions, as these are primarily derived from the early use of percentages in DESClogik (see above). An example of an overlap counted here is 369-U1513E-6R-2A, 0–29 cm (dolerite) and 21–29 cm (basalt). In total, 350 m of overlapping descriptions are found in LILY-CleanLITH. To determine omissions (gaps) in the lithologic description, we consider the total meters of curated material minus whole-round shipboard samples. Whole round (WRND) samples collected on the catwalk are typically not visually examined by core describers and therefore should lack a lithologic description. With WRND samples removed from the total curated length, 87,227 m of curated material are available for description, and of this, only 1,867 m are undescribed (~2%). The largest discrepancies come from expeditions with high recovery where multiple holes or large portions of holes were not described owing to time constraints during the expeditions, for example, Expeditions 323, 339, and 363. This results in 85,360 m of lithology described in LILY-CleanLITH, regardless of the potential overlapping descriptions available.

Core Summary information includes the expedition, site, hole, core, coring type, top and bottom depths drilled, advances and recoveries, time and date of recovery, and the number of sections. Coordinates and water depth for each hole are derived from LIMS (and the Janus database at http://www-odp.tamu.edu/database/ for older expeditions) and appended to Core Summary information [LILY-CoreSUMM]. During Expeditions 341 and 346, the half-length advanced piston corer (HLAPC) system was used, but the core type for the advanced piston corer (APC) and HLAPC were both recorded in the IODP database as "H." A distinction of "F" was later used for the HLAPC core type. Because the original "core type" is part of the Label and Sample ID (ex.—tracking information attached to real sections and samples that exist) this information is not corrected in the core type column ("Type"). Rather than changing the sample ID information, which would generate problems with other datasets, we provide a new column ("Expanded Core Type") to indicate the true coring system used. Coring types in column "Expanded Core Type" include a wide variety of terminology (Table S3; Storms, 1990), however the most common types of interest are generally: APC = advanced piston corer, HLAPC = half-length advanced piston corer, XCB = extended core barrel, RCB = rotary core barrel, WASH = washed interval, DRILLED = drilled interval, and GHOST = ghost core. Because basic coring summaries are available for all DSDP, ODP, and IODP expeditions in the Janus and LORE databases, we have extracted that data, paired them with additional metadata (expanded core type, latitude, longitude, and water depth) and included them in LILY-CoreSUMM along with the expeditions included in the rest of this study.

The final cleaning of lithologic data is to standardize the terminology used for the prefix, principal, and suffix lithology terms. This standardization and linkage to the original DESClogik data is contained in the LILY dictionary (Table S4), which was created following the guidelines described in Text S2 in Supporting Information S1. Briefly, standardization includes the correction of variable capitalization (ex.—sand, Sand, SAND), misspellings (ex.—"bbasalt"), and redundant terminology (ex.—removal of "with nannofossils" from "nannofossil ooze with nannofossils"). Other normalizations include ordering of multiple terms (rich, bearing, poor). Alphabetization of mineral nomenclature is also included. While in some instances the ordering has meaning and purpose, this is not always guaranteed and without standardization of these terms over the 42-expedition dataset, nearly identical lithologies are less readily compared. Some choices made in this dictionary are for the purpose of exploring the entirety of the 42-expedition dataset and may not be suitable/preferred for other types of analysis. This includes the removal of 27.75 m worth of descriptions with principal lithologies deemed non-lithological ("drilling disturbed zone of rock fragments", "Void", "alteration vein", "void", "???", "Vein fill", "fall-in"). The final LILY dictionary (Table S4) contains unique combinations of prefix, principal, and suffix linking all the original (DESClogik) full lithologies to updated (LILY) full lithologies.

2.5. Data Transformation and Management—Track and Discrete Data

All non-descriptive (physical, chemical, magnetic, etc.) data are available at http://web.iodp.tamu.edu/LORE/. Datasets paired with descriptive information are listed in Table S1 in Supporting Information S1 and information about the quantity of data in LILY by data type and expedition is given in Table S2 in Supporting Information S1.

Each track/discrete dataset is downloaded by expedition from LORE [LILY-RawDATA]. An exception to this is when data are collected at a density (and therefore total volume) too large for LORE to download at the expedition level. In these instances, data are downloaded at the Site level and combined locally. Lithologic descriptive data (LILY-CleanLITH) are paired to the track/discrete data by matching the Expedition, Site, Hole, Core, Type, Section, and "Offset/Top-Bottom Offset cm". For the special case of IW (interstitial water) discrete samples,

which are collected on the catwalk prior to core description, the lithology from immediately above the IW sample interval is presumed to continue through the sample. Where multiple principal lithologies are provided coincident with a track or discrete measurement, the track/discrete dataset row is duplicated. The track/discrete data paired with lithology are provided for each expedition in LILY-DataLITH.

3. Data

3.1. The Lithologies of LILY

The lithologic dataset (LILY-CleanLITH) contains descriptions of over 85 km of core collected on the 42 expeditions included in this study. Coring locations span the globe, but the total amount of core recovered is unevenly distributed with >68% from the northern hemisphere (Figure 2). A quarter of Earth's surface, consisting of the western hemisphere south of the equator, contains only 4.8% of the core, although this lack of sampling will be partially alleviated by several recent expeditions that were conducted in this region (i.e., Expeditions 382, 383, and 378) but were still in moratorium at the start of this project. Nearly half (47%) of the core was collected in the tropics (between 23.4336°S and 23.4336°N), while only 5.4% was collected in polar regions (south of 60°S and north of 60°N). The cores were collected in water depths that vary from 87.15 to 5,696.66 m.

Principal lithology is a required observation for the dataset and is present for all >85,000 m of core described. Within this dataset, 209 unique principal lithologies are identified and the most common (by meters of core described) is "nannofossil ooze" (>20% of all descriptions). The 20 most common principal lithologies account for >80% of the meters described (Table 2), whereas the four least common principal lithologies comprise only 0.04 m of the core described.

Prefixes are used for 37,958 m of the dataset, with over half of the descriptions (>49,200 m) containing no prefix. Within this dataset, 431 unique prefix values are identified and the most common (by meters of core described) is "nannofossil-rich" (~9% of all prefixes by meters). The 20 most common prefixes account for almost 75% of the meters described that include a prefix (Table 2), whereas the five least common comprise only 0.06 m of all core described.

Suffixes are used for 24,603 m of the dataset, with the bulk of descriptions (>62,600 m) containing no suffix. Within this dataset, 185 unique suffix values are identified and the most common (by meters of description) is "with foraminifera" (~25% of all suffixes by meters). The 20 most common suffixes account for >90% of the meters described that include a suffix (Table 2). The least common suffix is "with algae," which comprises 0.01 m of the core described.

Thirty-nine percent of principal lithologies (by meter of core described) have no prefix or suffix, 34% have only a prefix, ~18% have only a suffix, and ~10% have both. The most common lithology with both a prefix and suffix is "nannofossil-rich clay with foraminifera", which comprises ~1.0% of the core described.

Rather than look at specific lithologies, one can group lithologies by types. For example, the total amount of unconsolidated core collected accounts for ~71% of the core described. Within the unconsolidated group, all clastic sediments (ex.—clay, silt, mud, and sand) can be grouped and these account for 58% of the unconsolidated core described. Biogenic oozes account for another 41% of unconsolidated sediments, leaving only ~1% for all other unconsolidated lithologies.

3.2. Coring Type and Core Recovery Observations With and Without Lithologies

As noted above, LILY-CoreSUMM contains coring summaries that include coring type and recovery percentage information as well as associated metadata for all cores, drilled and washed intervals, and any other coring records from DSDP (21,725 records), ODP (36,388 records), and IODP (23,001 records; through Expedition 398). These cores were primarily collected (94% of LILY-CoreSUMM records) with one of four main coring systems: APC, HLAPC, XCB, and RCB. The data in the "Expanded Core Type" column will therefore generally be designated as being one of these four systems, although other core types, such as GHOST, NCB, PDCM, and MISC, occur much more rarely (Table S3). Of the 81,114 records in LILY-CoreSUMM, 35,344 are RCB, 24,055 are APC, 14,896 are XCB, and 2,693 are HLAPC cores.

The overall LILY-CoreSUMM dataset can be used to examine attributes of coring type, regardless of whether an expedition is included in the set of 42 expeditions with detailed lithology and track/discrete data primarily

Table 2
Twenty Most Common Prefixes, Principal Lithologies, Suffixes, and Full Lithologies by Meters of Lithology Described

Prefix	meters	Principal	meters	Suffix	meters	Full lithology	meters
nannofossil-rich	3526.12	nannofossil ooze	17494.84	with foraminifera	6259.35	nannofossil ooze	4132.04
calcareous	2794.08	clay	10244.79	with nannofossils	2802.67	silty clay	3629.64
diatom-rich	2746.10	silty clay	6882.42	with clay	2161.72	clay	2768.02
clayey	2615.74	mud	5075.49	with silt	1663.83	diatom ooze	2750.82
planktic foraminifera-rich	2224.34	diatom ooze	4201.24	with dispersed clasts	1083.65	nannofossil ooze with foraminifera	2352.67
diatom-bearing	2215.02	nannofossil chalk	3347.54	with radiolarians	1083.62	calcareous mud	1692.34
foraminifera-rich	1802.82	wackestone	3172.24	with glauconite	1065.93	wackestone	1572.63
tuffaceous	1774.14	packstone	3010.72	with biogenic carbonate	888.01	mud	1495.12
clay-rich	1061.71	clayey silt	2605.98	with diatoms	831.54	mudstone	1452.17
volcaniclastic	1010.66	mudstone	2294.20	with fine sand	783.61	nannofossil mud	1422.86
biosilica-rich	946.64	sand	2253.92	with biogenic silica	736.74	packstone	1404.03
foraminifera-bearing	795.38	claystone	2120.98	with ash	540.32	claystone	1280.35
diatom	768.39	nannofossil clay	1820.15	with sponge spicules	475.77	nannofossil chalk	1210.06
muddy	759.22	basalt	1453.00	with biogenic components	459.52	planktic foraminifera-rich wackestone	1209.94
foraminifera-bearing clay-rich	604.51	nannofossil mud	1423.50	with diatoms and radiolarians	438.12	nannofossil-rich clay	1079.80
clay-bearing	578.66	carbonate ooze	1145.59	with sand	354.40	carbonate ooze	1031.10
aphyric	469.02	silty mud	965.74	with gravel	303.61	planktic foraminifera-rich packstone	900.38
nannofossil	431.65	silt	930.74	with mica	268.15	nannofossil-rich clay with foraminifera	872.30
olivine	423.39	silty claystone	899.32	with lithic clasts	242.55	nannofossil clay	862.65
nannofossil-bearing	409.57	breccia	779.85	pillow lava	146.79	nannofossil ooze with clay	794.86

included in this study. Hence, one could examine the success of each coring system in terms of recovery percentage over the entire period that the coring systems have been in use for scientific ocean drilling, as summarized in Figure S1.

An analysis of how the coring systems work in different lithologies is possible using the 42 expeditions where coring type is paired with lithology, which contains over 14,900 records of which 5,701 are APC, 4,318 are RCB, 2,534 are XCB, and 1,890 are HLAPC cores. The coring systems were designed to collect certain rock types and thus their use naturally correlates with lithology. In general, piston coring (APC, HLAPC) is used for unconsolidated materials (i.e., sediments), XCB for semi-lithified materials (e.g., chalk), and RCB for rock. The XCB system is also used to collect short intervals of hard rock at the base of an APC/XCB cored hole, for example, to collect a short interval of igneous basement after penetrating the sedimentary overburden. Its primary purpose is to extend coring in a hole started with the APC system, without needing to trip the pipe back to the surface to change bits, which is required for using the RCB system. The RCB system may be used to collect soft or semi-lithified materials as it passes through the unconsolidated sedimentary overburden on its way to deeper, targeted lithified or igneous intervals.

A natural question to ask is how successful each system is in collecting different lithologies. A study by Evans (2020) first addressed this question for IODP Expeditions 317–375. In that study, all cores from a hole were assigned the lithology of the most common rock type from the hole and then placed into one of six lithologic groups (Ooze/chalk; Siliciclastics [silt, clay, sand]; Shallow-water carbonate; Siliceous ooze; Volcaniclastic; and Glacial), which allows for studying only first-order differences in coring system. Nonetheless, even those basic lithologic designations revealed interesting observations, like recovery for the APC and HLAPC system is nearly
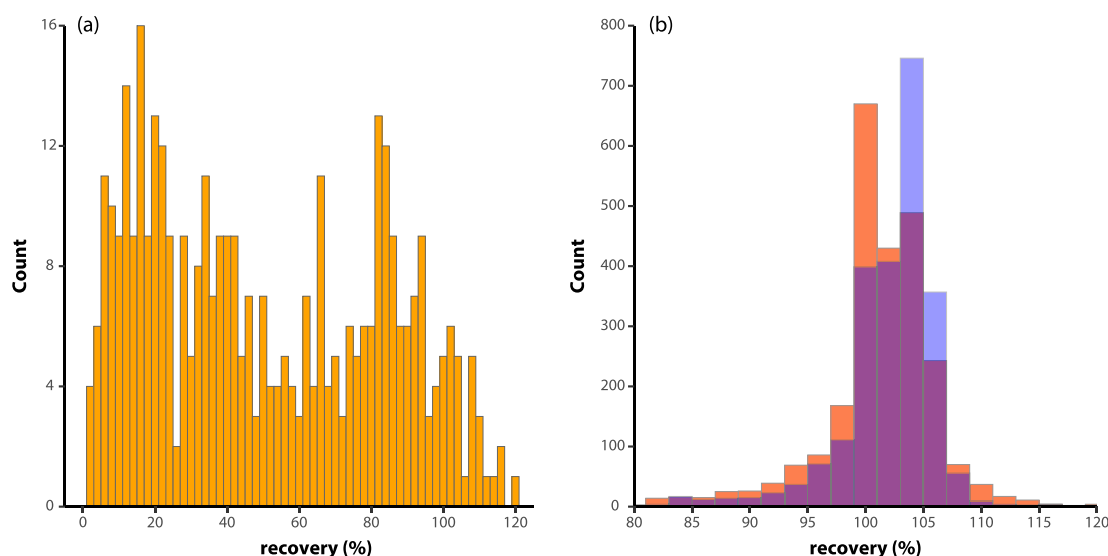
Figure 3. (a) RCB recovery of cores that are at least 50% basalt. (b) Histograms of percent recovery of the APC coring system in cores that are at least 50% clastic sediments (orange), biogenic lithologies (light blue), and their overlap (purple). Only the interval from 80% to 120% is shown because few cores have recoveries outside that range.

100% in all six groups, whereas the XCB system had 63%–76% recovery in ooze/chalk and siliciclastics and only 27%–37% recovery in glacial, volcaniclastic, and carbonate sediments/rock.

Our results agree with these generalizations, but with LILY-CleanLITH one can refine the relative success of each coring system in different lithologies. Because core recovery is known only for each core, all lithologies within a core are assumed to be recovered at the percent recovery for that core, which should be a fairly accurate indicator for each lithology. As an example of how the database can be used, we consider some questions that might commonly arise, such as what recovery percentage can be expected when using the RCB system to core in basalt? We consider only cores that have basalt as the principal lithology for 50% or more of the core recovered, for which there are 392 cores (Figure 3a). The mean recovery for these is 51.7%. The distribution of recovery percentages has a somewhat bimodal distribution with modes at about 16% and 82%. Thus, someone planning an expedition that targets basalt might expect about 50% recovery, with relatively poor recovery (~16%) in about half of the cores recovered and good recovery (~82%) in the other half.

As another example, LILY might be used to assess if the APC coring system has higher success recovering unconsolidated clastic sediments or biogenic oozes. LILY-CleanLITH has 2,560 APC cores that contain 50% or more clastic sediments (i.e., those with principal lithologies such as clay, silt, mud, or sand) and 2,365 APC cores that contain 50% or more biogenic oozes. Histograms of the recovery percentages for these two lithologies (Figure 3b) indicate highly unimodal distributions with modes for clastic sediments and biogenic oozes at 101% and 105% recovery and means of 101.7% and 103.1%, respectively [Note, recovery percentages greater than 100% generally result from core expansion (e.g., gas expansions in voids) when cores are brought from below the seafloor to the ship and because curation practices aim to preserve all material recovered, including soupy sediment at the tops of cores, some of which may have fallen downhole from above]. The mean difference in the recovery of clastic sediments versus biogenic ooze with the APC coring system is small (1.4%) but is actually significant at the 95% confidence level because of the large number of cores. There are several explanations for this difference of which the two most likely are that the APC system either cores biogenic sediments slightly better than it cores siliciclastic sediments or that biogenic sediments expand more than siliciclastic sediments when they are retrieved from below the seafloor. With recovery rates over 100% for both lithologic groups, the main conclusion is that the APC system is very effective at recovering both types of sediments.

One might also investigate the efficacy of the RCB system relative to other systems in collecting the most common lithology, nannofossil ooze, and its more indurated version, nannofossil chalk. Figure 4 shows that nannofossil chalk is cored with moderate to good success with RCB, whereas nannofossil ooze recovery is poor to moderate. In contrast, nannofossil chalk can be reasonably well recovered with XCB, and in the few occasions it is
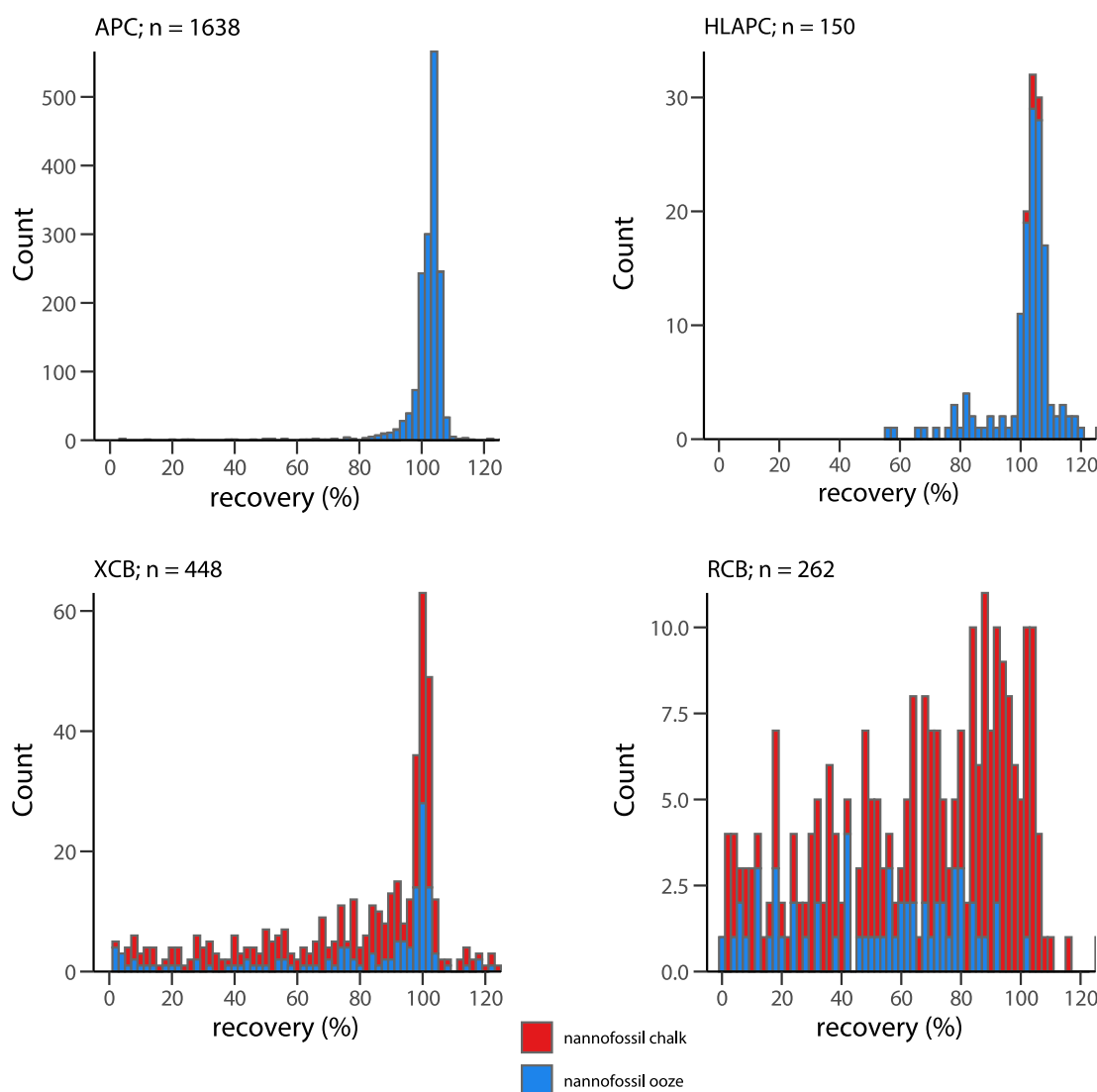
Figure 4. Recovery by coring type for cores with at least 50% nannofossil chalk or nannofossil ooze.

cored by HLAPC, the recovery is quite good. Almost anything except RCB is a good bet for recovering nannofossil ooze.

3.3. Track and Discrete Data

More than 34 million unique individual track and discrete measurements from 23 instruments/tracks were paired with lithologic data (Figure 5; Table S2). The largest track data source, RGB, includes 10,789,797 measurements comprising ∼31% of the paired data. The smallest data source, the discrete measurement SRA, includes 121 measurements (<0.01% of the paired data). Some measurements (e.g., NGR) are collected consistently across all expeditions. However, several measurements are collected ad hoc (e.g., IW) or are collected only after a particular instrument or capability was acquired (e.g., RGB, SRA).

3.4. Data Quality

Given the large number of measurements, most made continuously along a core section, it is not surprising that anomalous or biased values are recorded, such as those that might occur for measurements made with any instrument that was poorly calibrated. The advantage of analyzing big data sets is that even though poor-quality measurements may occur, their numbers should be a relatively small percentage of the overall data, assuming
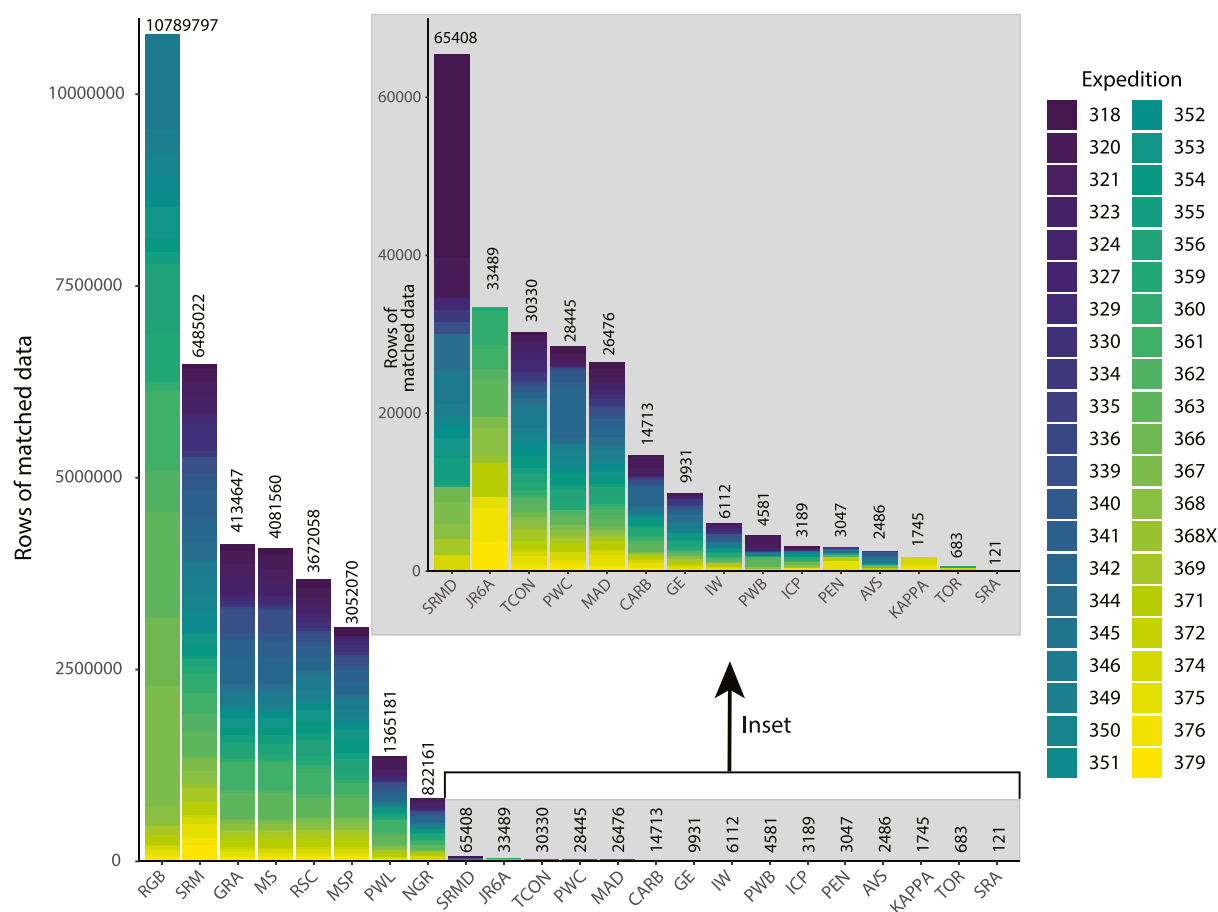
Figure 5. Extent of physical, chemical, and magnetic datasets paired with lithologic descriptions for this study. All data were retrieved from the IODP Laboratory Information Management System (LIMS) [http://web.iodp.tamu.edu/LORE/]. See Table S1 in Supporting Information S1 for track/discrete data abbreviations and Table S2 in Supporting Information S1 for data quantity.

good laboratory practices are generally followed when collecting data. Thus, meaningful estimates of physical, chemical, and magnetic properties can be extracted from the big datasets, with relatively little or even no filtering of "bad" data. Ultimately, some values recorded are so egregious that they are easy to filter out when analyzing the data. Other "bad" data may not be possible to identify or remove.

In the following analyses, it is important to understand some common reasons that anomalous or biased data occur, beyond typical uncertainties related to instrument sensitivity, which are covered in the IODP lab manuals (available at wiki.iodp.tamu.edu) and in the "Methods" section of the IODP "Proceedings Volumes". As an example, consider track measurements, most of which have measurement-dependent parameters in their software, such as the expected core diameter, that are set and remain fixed for all cores measured, rather than being adjusted for real differences that occur along a core or between cores. For example, the inner diameter of the plastic core liner is 6.6 cm. For APC sediment cores, the liner is typically filled to the inner periphery. Thus, for measurements that vary with core diameter or volume of material being measured, such as density, P-wave velocity, susceptibility, and intensity of magnetization, 6.6 cm is a reasonably accurate estimate for the core diameter for intervals along a core that are filled with sediment. Of course, some intervals are not filled owing to cracks, gas voids, poor recovery, and various forms of disturbance related to drilling or processing the core. Such incompletely filled areas will give anomalous observations for those measurements depending on diameter or volume. For example, gamma ray attenuation (GRA) density measurements for some intervals will be lower than the density of water and may even be slightly negative; the former occurs over intervals that are only partially filled with material and the latter may occur over voids. Such intervals are not representative of any sediment or rock type and, if possible, need to be filtered out of the data when

analyzing the properties of the different lithologies. The best use of IODP data also includes the review of expedition volume materials, including "Methods" chapters, which may identify or explain certain anomalous data.

Systematic biases may also occur. For example, the diameter of RCB cores is never 6.6 cm because the RCB bit cuts a core narrower than the core liner. The throat of the RCB bit is 5.87 cm, and the typical core diameter for an RCB core is between 5.5 and 5.8 cm and it commonly gets smaller the more hours a bit is used, because the roller cones tend to wobble more with wear. Track systems that have the diameter fixed at 6.6 cm will produce biased values for RCB cores for those properties that depend on the diameter, such as density, and for those that are normalized by volume, such as susceptibility and intensity of magnetization. Rarely, the operator of the track system may adjust the diameter value being used, but this is not done consistently and thus biases lurk in the data. Removing all anomalous data and correcting data for biases are beyond the scope of this study, although below we provide an example of how both can be done for density data. Future studies may use the joint observations made along each core by the different instruments to devise ways to detect and filter out more of the unrepresentative data and to correct for biases.

4. Results

4.1. LILY Example Studies of Density and Porosity

LILY data can be probed for a myriad of studies. We illustrate this using the moisture and density (MAD) data along with the GRA bulk densities in the following examples:

1. MAD data are used to estimate the crystallographic (grain) densities, bulk densities, dry densities, and porosities for the different principal lithologies, simplified lithologies, lithology types and subtypes, degree of consolidation, and core type.
2. MAD and GRA bulk densities are compared to assess consistency and to investigate and correct for biases, improving the overall accuracy of the GRA densities.
3. Porosity is estimated for 3.7 million GRA density data using the grain densities determined from the MAD data.

4.2. Moisture and Density (MAD) Data

For background, MAD data are collected from samples, typically 6–10 cm^3 in volume, that are taken from representative intervals of the core typically at a frequency of one to two samples per core section. The wet and dry masses of the samples are weighed with an accuracy of 0.1%, the dry volume of each sample is carefully determined in a pycnometer with an accuracy of 1%, and the resulting bulk densities, grain densities, porosities, void ratios, and moisture contents are estimated generally with an accuracy of few percent (Blum, 1997).

The densities and porosities in the MAD data are expected to follow relationships based on Archimedes principle (Hall & Hamilton, 2015), which is given in air by

$$v = 1 - \rho_B/\rho_G \tag{1a}$$

$$\rho_B = \rho_G - \rho_G v \tag{1b}$$

and in a fluid by

$$v = (\rho_B - \rho_G)/(\rho_F - \rho_G) = \frac{1}{(\rho_F - \rho_G)}\rho_B + \frac{\rho_G}{(\rho_F - \rho_G)} \tag{2a}$$

$$\rho_B = (\rho_F - \rho_G)v + \rho_G \tag{2b}$$

where ρ_B is the bulk density, ρ_G is the grain density (i.e., the density of the solid components, also referred to as the crystallographic density), ρ_F is the fluid density, and v is the porosity given as a ratio of densities (we plot and express these values as percentages). The MAD densities are determined wet, where the pore-water fluid is considered to be seawater with a density of 1.024 g/cm^3 (Blum, 1997). As apparent in the equations, porosity
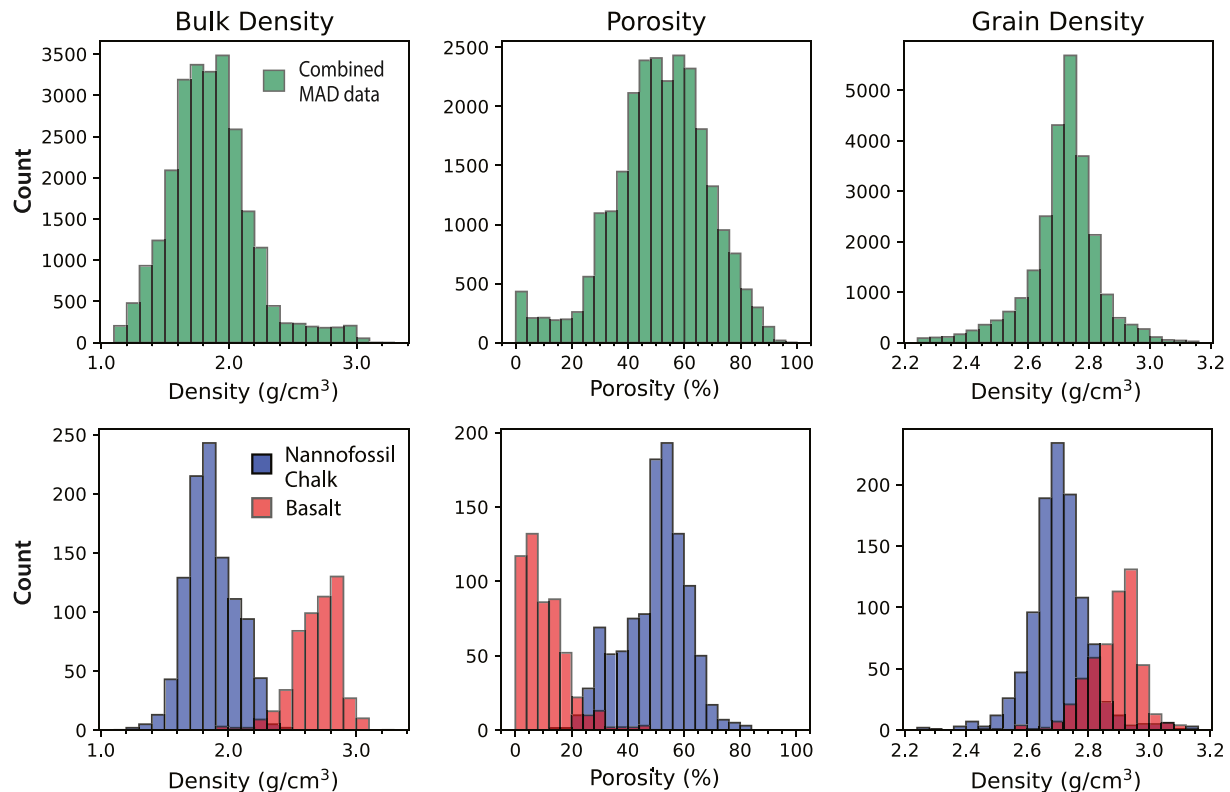
Figure 6. Histograms of MAD data showing distributions of the combined data and differences in the distributions for two lithologies, nannofossil chalk and basalt.

varies linearly with bulk density. In the absence of any solid material, porosity is 100% and the bulk density is that of seawater ($\rho_B = \rho_F$). As porosity decreases, bulk density increases linearly toward the grain density, and $\rho_B = \rho_G$ when the porosity is zero. These relationships are valid for porous materials, where liquid fills the open pores and any closed pores are treated as part of the solid material.

From the above, it is apparent that if the grain density is accurately known, the expected porosity can be accurately calculated for a sample by measuring only the bulk density. This can prove useful because there are about 4 million GRA bulk density measurements without associated porosity determination. The only catch is that the grain density is dependent on the lithology. The value of LILY becomes apparent as it gives us the ability to compute the grain density for any lithology represented in the MAD dataset and to apply that information to the GRA bulk density dataset to get porosities. Likewise, we can compute the expected MAD porosities using the MAD grain densities and MAD bulk densities and then compare them with the observed MAD porosities as a check on the quality of the data.

We first examine some general properties of the MAD dataset using data visualization and basic statistics. The dataset contains 26,467 rows of data. After discarding duplicates, such as samples that are at lithologic contacts with two different principal lithologies or sections that were measured twice, and discarding two rows that are missing either a value in the bulk density, grain density, or porosity columns, 25,594 rows of data (independent samples with unique principal lithologies) remain. We do some minor cleaning of the data to remove extreme outliers, such as samples with density or porosity that is beyond what is possible (e.g., negative porosity) or probable for marine sediments and rocks. For example, marine sediments and rocks are predominantly composed of minerals that have grain densities well between 2.0 and 3.2 g/cm^3, with only very few and scattered values plotting outside that range. Applying the following cutoffs for extreme outliers:

- 1.024 g/cm^3 < bulk density < 3.2 g/cm^3
- 2.0 g/cm^3 < grain density < 3.2 g/cm^3
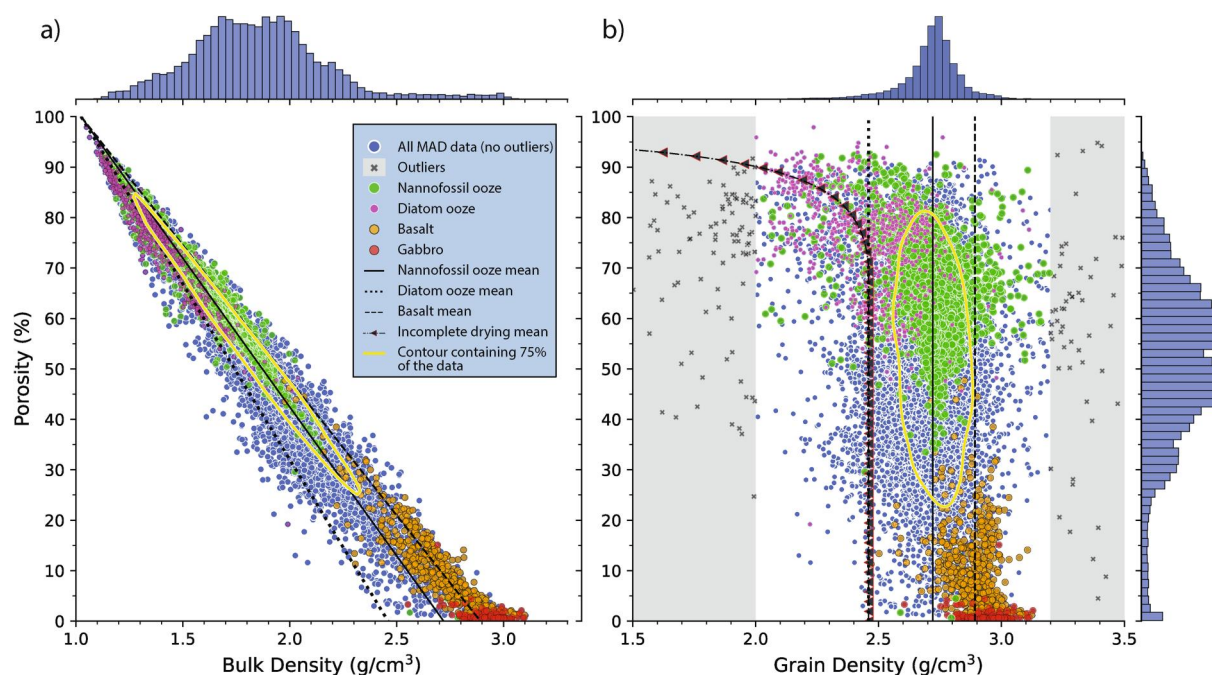- 0% ≤ porosity ≤ 100%

Figure 7. (a) MAD bulk densities versus porosity data and (b) MAD grain densities versus porosity are plotted for all the data and for four representative principal lithologies along with histograms for all the data. The solid, dotted, and dashed lines in (a) show the expected linear relationships between bulk density and porosity (see Equation 2) for nannofossil ooze, diatom ooze, and basalt, respectively, which have grain densities of 2.72, 2.46, and 2.89 g/cm³, respectively (see Section 4.2.1 Estimating MAD Grain, Bulk, and Dry Densities and Porosities). The same grain densities are vertical lines in (b) as grain density is independent of porosity. Diatom ooze, however, shows a decrease in grain density when the porosity is higher than about 80%, which we attribute to incomplete drying of samples with very high porosity. The dotted-dashed line with triangles is a model that shows how incomplete drying of only a few percentage of the total water can affect the grain density. The solid yellow line is a contour that contains 75% of the data. The contour and the histograms convey that the data are highly concentrated in a restricted range in both plots.

reduces the number of MAD data to 25,335. There are 1,336 unique full lithologies and 148 unique principal lithologies represented in these measurements.

The distribution of the grain densities, bulk densities, and porosities for the full MAD data are Gaussian in shape, but these are in turn composed of distinct lithologic-dependent distributions, as illustrated for nannofossil chalk and basalt in Figure 6. Lithological differences in linear trends in bulk density versus porosity data are also apparent (Figure 7a). While the overall trend of all the data is linear, some non-linearity occurs in the combined data because bulk densities and porosities depend on lithology. This is particularly evident when the porosity is small and the bulk density approaches that of the crystallographic (grain) density of the constituent mineral. Hence, lithologies that are largely composed of calcite, like nannofossil ooze, will have grain densities near that of calcite (2.711 g/cm³; Smyth and McCormick (1995)), and lithologies, like basalt and gabbro, composed of mafic minerals will have grain densities approaching that of mixtures of mafic minerals, e.g., pyroxenes (~3.3 g/cm³), olivine (~3.2 g/cm³), plagioclase (~2.7 g/cm³), and hornblende amphibole (~3.0 g/cm³).

Scatter plots of grain density versus porosity, when combined with histograms and cumulative data density contours, proved to be useful for identifying outliers or anomalous observations that exist in addition to the "extreme outliers" (Figure 7b). The expectation is that grain density should vary with lithology and be independent of porosity for each separate lithology. Hence, we would expect to see the carbonate lithologies, such as nannofossil ooze, cluster around 2.71 g/cm³ and mafic igneous rocks, such as basalt, cluster around 2.9 g/cm³ regardless of porosity (see the vertical solid and the vertical dashed lines, respectively, in Figure 7b). While the grain densities for nearly all the lithologies do exactly that, we also observe a small population (<400 samples) with grain densities that trend toward lower values as the porosity increases above about 80%. Most of these samples are biogenic sediments, with 56% being diatom and radiolarian ooze. The mean grain density for diatom ooze is only 2.34 g/cm³ for samples with porosity >80% and 2.49 g/cm³ for samples with porosity <80%. We attribute this behavior to the incomplete drying of very wet and mostly

Table 3
Mean Grain Densities, Bulk Densities, Dry Densities, and Porosities and Their Uncertainties Computed From the MAD Physical Properties Data Set

MAD density and porosity data		Grain density			Bulk density		Dry density			Porosity	
Principal lithology	N	Mean	Standard deviation	95% confidence limits	Mean	Standard deviation	Mean	Standard deviation	95% confidence limits	Mean	Standard deviation
basalt	514	2.893	0.076	0.018	2.708	0.159	2.605	0.225	0.056	10.0	6.9
boninite	71	2.809	0.078	0.023	2.422	0.144	2.201	0.222	0.069	21.6	8.1
breccia	320	2.757	0.133	0.034	2.315	0.291	2.048	0.429	0.058	26.0	13.8
calcareous ooze	131	2.705	0.071	0.012	1.675	0.162	1.046	0.252	0.122	61.4	8.9
carbonate ooze	285	2.693	0.078	0.047	1.610	0.125	0.944	0.194	0.040	65.0	6.8
chalk	193	2.703	0.049	0.008	1.933	0.185	1.463	0.290	0.042	45.9	10.4
clast-poor diamict	52	2.830	0.075	0.021	2.145	0.132	1.757	0.193	0.057	38.0	6.3
clast-poor sandy diamict	118	2.604	0.070	0.027	2.009	0.145	1.621	0.219	0.094	37.8	7.3
clast-rich diamict	98	2.855	0.094	0.047	2.270	0.101	1.943	0.148	0.072	31.9	5.3
clast-rich sandy diamict	80	2.630	0.071	0.023	2.091	0.148	1.746	0.228	0.066	33.7	8.0
clay	2912	2.738	0.092	0.025	1.720	0.247	1.109	0.386	0.025	59.6	13.7
clayey silt	626	2.696	0.078	0.006	1.802	0.153	1.256	0.239	0.043	53.5	8.7
clayey siltstone	197	2.721	0.044	0.008	2.004	0.152	1.572	0.249	0.149	42.2	9.5
claystone	834	2.718	0.133	0.064	1.973	0.203	1.521	0.302	0.077	44.2	10.0
conglomerate	155	2.802	0.070	0.017	2.146	0.262	1.766	0.404	0.309	37.0	13.9
diatom ooze	705	2.456	0.188	0.049	1.372	0.154	0.594	0.246	0.064	76.2	9.2
diatom silt	77	2.555	0.168	0.078	1.468	0.118	0.744	0.177	0.044	71.2	6.1
diatomite	63	2.427	0.089	0.022	1.318	0.064	0.509	0.111	0.086	79.0	4.7
fine sand	109	2.719	0.028	0.006	2.073	0.120	1.682	0.190	0.052	38.1	6.9
fine sandstone	65	2.572	0.144	0.041	1.979	0.169	1.589	0.268	0.090	38.0	10.5
foraminifera ooze	100	2.723	0.100	0.048	1.570	0.088	0.875	0.134	0.040	67.9	4.7
gabbro	243	2.965	0.058	0.025	2.952	0.061	2.945	0.064	0.034	0.7	0.6
grainstone	211	2.781	0.048	0.014	1.796	0.262	1.223	0.416	0.081	56.0	15.1
interbedded sand and mud	125	2.727	0.041	0.021	1.948	0.155	1.479	0.243	0.052	45.8	8.7
interbedded sandstone and mudstone	85	2.666	0.021	0.007	1.994	0.044	1.574	0.071	0.018	41.0	2.7
interbedded silt and clay	731	2.721	0.044	0.012	1.986	0.140	1.542	0.215	0.039	43.4	7.4
interbedded siltstone and mudstone	211	2.712	0.035	0.007	1.948	0.071	1.485	0.112	0.022	45.2	4.1
lapilli tuff	103	2.711	0.092	0.021	2.209	0.111	1.905	0.171	0.046	29.7	6.5
lapillistone	50	2.714	0.085	0.024	2.183	0.111	1.862	0.183	0.054	31.3	7.5
limestone	63	2.665	0.139	0.041	2.177	0.245	1.872	0.370	0.103	29.8	12.9
marl	158	2.699	0.050	0.017	2.024	0.141	1.611	0.213	0.134	40.4	7.2
medium to coarse sandstone	153	2.661	0.169	0.029	2.007	0.112	1.604	0.190	0.045	39.3	9.0
mud	1304	2.747	0.086	0.005	1.836	0.196	1.294	0.303	0.122	53.0	10.6
mudstone	727	2.688	0.102	0.052	1.932	0.231	1.467	0.368	0.052	45.4	13.6
nannofossil chalk	1025	2.705	0.083	0.005	1.871	0.186	1.363	0.297	0.077	49.6	11.0
nannofossil clay	372	2.738	0.046	0.013	1.635	0.135	0.976	0.213	0.048	64.4	7.6
nannofossil claystone	71	2.752	0.029	0.008	1.733	0.086	1.129	0.136	0.118	59.0	4.9
nannofossil mud	340	2.766	0.036	0.013	1.968	0.082	1.499	0.127	0.017	45.8	4.5
nannofossil ooze	3228	2.721	0.085	0.007	1.633	0.149	0.975	0.233	0.048	64.2	8.3

Table 3
Continued

MAD density and porosity data		Grain density			Bulk density		Dry density			Porosity	
Principal lithology	*N*	Mean	Standard deviation	95% confidence limits	Mean	Standard deviation	Mean	Standard deviation	95% confidence limits	Mean	Standard deviation
packstone	854	2.762	0.042	0.017	1.945	0.211	1.465	0.335	0.063	47.0	12.1
pebbly mud	216	2.602	0.058	0.009	1.906	0.172	1.455	0.278	0.057	44.1	10.5
radiolarian ooze	105	2.437	0.215	0.049	1.227	0.056	0.351	0.087	0.037	85.6	3.4
sand	360	2.746	0.068	0.022	1.939	0.147	1.458	0.227	0.097	46.9	8.1
sandstone	417	2.659	0.148	0.057	2.047	0.234	1.663	0.350	0.110	37.6	12.0
sandy clay	53	2.680	0.062	0.021	1.700	0.156	1.094	0.251	0.070	59.2	9.3
sandy silt	100	2.709	0.070	0.019	1.899	0.145	1.407	0.226	0.079	48.1	8.2
sandy siltstone	71	2.699	0.061	0.014	2.008	0.094	1.586	0.149	0.075	41.2	5.6
silt	241	2.718	0.077	0.010	1.829	0.209	1.291	0.323	0.046	52.7	11.3
siltstone	305	2.618	0.153	0.071	1.942	0.150	1.510	0.236	0.041	42.1	9.5
silty clay	2055	2.700	0.090	0.021	1.767	0.227	1.195	0.352	0.044	55.9	12.4
silty claystone	330	2.709	0.087	0.010	2.002	0.124	1.571	0.186	0.033	42.0	6.4
silty mud	210	2.761	0.051	0.012	1.987	0.157	1.531	0.245	0.066	44.6	8.7
silty sand	182	2.747	0.058	0.018	1.958	0.127	1.488	0.198	0.046	45.8	7.1
silty sandstone	151	2.749	0.061	0.011	2.132	0.138	1.766	0.214	0.040	35.8	7.6
tuff	85	2.649	0.126	0.062	2.012	0.241	1.609	0.369	0.165	39.4	12.9
very fine sand	119	2.733	0.024	0.006	2.111	0.120	1.739	0.191	0.051	36.4	7.0
volcanic rock	51	2.786	0.091	0.026	2.422	0.123	2.210	0.171	0.047	20.7	5.4
wackestone	714	2.744	0.032	0.007	1.908	0.181	1.410	0.288	0.051	48.6	10.4

Note. A more complete list of grain densities, bulk densities, dry densities, and porosities along with their statistics is given in Table S7. Results in that table also include values for data grouped by simplified lithology, lithology type, lithology subtype, degree of consolidation, and expanded core type. N is the number of observations. 95% Confidence Limits include random and systematic errors as described in the text.

porous biogenic samples. A simple model (line with triangles in Figure 7b) in which a small percentage of the pore water is not removed by drying samples with high porosity can explain the anomalous values. In the model, the percentage of pore water not removed increases nonlinearly from 0% for porosities <70% to about 2% for a porosity of 90%. Possibly some of the pore water is locked in the tests of microfossils in the high porosity samples, which will be those near the mudline (top of the seafloor) where microfossil shells are generally well preserved. Alternatively, samples with very high porosities may merely need to be dried longer than those without so much water in their pore spaces.

4.2.1. Estimating MAD Grain, Bulk, and Dry Densities and Porosities

To compute the best estimate of the grain, bulk, and dry densities and the porosity for each lithology, we first remove the extreme outliers as described above. Next, we group the data, where grouping is done by principal lithology, along with subsequent groupings by simplified lithology, lithology type, lithology subtype, degree of consolidation, and expanded core type. Each group is cleaned by excluding values that are more than three standard deviations from the group mean values, and then the means, medians, and their attendant statistics are recomputed (Table S7, with a subset of the data in Table 3). This cleaning reduces the number of MAD data from 25,335 to 24,748 samples, which is a reduction of <1%. We refer to this as the "ultra-clean" dataset and we use the mean values from this dataset as the most representative estimate of the true values.

Representative grain and dry densities for the principal lithologies with more than 50 observations are shown along with their standard deviation and 95% confidence limits in Figure 8. The 95% confidence limits listed in Table 3 and Table S7 are derived by first adding in quadrature the random standard error and a systematic standard error equivalent to the difference between the mean and median, and then multiplying this sum by
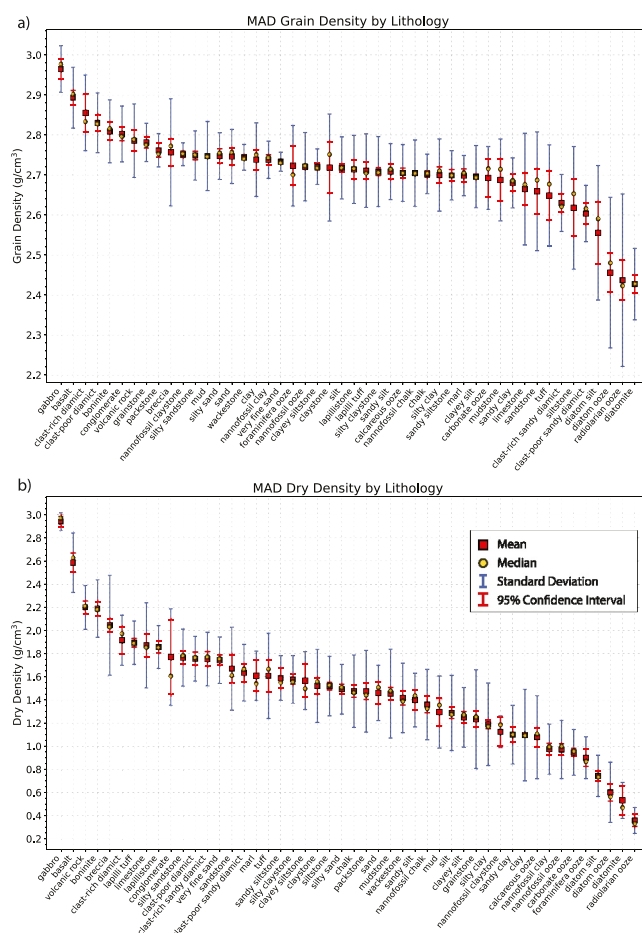
Figure 8. Mean (red squares) and median (yellow dots) for (a) grain densities and (b) dry densities for some of the principal lithologies with more than 50 observations each, along with their standard deviations (blue bar) and 95% confidence limits (red bar).

1.96. This systematic error is included to account for the non-Gaussian nature of some of the distributions. The most extreme example of this is for the claystone grain density, where the mean and median differ by 0.03 g/cm^3, which significantly exceeds the 0.01 g/cm^3 95% confidence limits for random errors alone.

The grain densities computed from the large number of observations for some lithologies give mean values that have very small 95% confidence limits, averaging ±0.024 g/cm^3 for the 60 lithologies with 50 or more observations. This leads to some interesting observations such as nannofossil ooze (2.721 ± 0.007 g/cm^3) has a small but statistically significantly higher grain density than nannofossil chalk (2.705 ± 0.005 g/cm^3), which in turn has a higher grain density than limestone (2.665 ± 0.041 g/cm^3). Likewise, sand (2.746 ± 0.022 g/cm^3) has a significantly higher grain density than sandstone (2.659 ± 0.057 g/cm^3), silt (2.718 ± 0.010 g/cm^3) has a significantly higher grain density than siltstone (2.618 ± 0.071 g/cm^3), and clay (2.738 ± 0.025 g/cm^3) has a higher, although not statistically significant, grain density than claystone (2.718 ± 0.064 g/cm^3). These seemingly small differences apparently have a physical origin. Possibly, the more lithified versions of these lithologies may have some closed pore spaces that retain water, which would result in lower grain densities and higher dry densities than the unconsolidated analogs. This explanation is supported by the dry density data, where the more lithified lithologies (limestone, claystone, siltstone, and sandstone) have higher values than the unconsolidated analogs (nannofossil ooze, clay, silt, and sand).

The grain densities are generally consistent with the expected values based on the grain densities of the samples' constituent minerals, for example, the expected (calculated) grain density of pure calcite is 2.711 g/cm^3 and the observed grain densities of the lithologies mainly composed of calcite have comparable values: nannofossil ooze (2.721 g/cm^3), nannofossil chalk (2.705 g/cm^3), chalk (2.703 g/cm^3), and foraminifera ooze (2.723 g/cm^3). Although none of the calcite-rich lithologies are pure calcite, they differ by no more than 0.012 g/cm^3 from the expected value for calcite. An exception to this is limestone (2.665 ± 0.041 g/cm^3), which differs only marginally given its uncertainty.

MAD bulk densities and porosities show the expected inverse relationship (Figure 9). Lithologies with large porosities (>60%), such as oozes and other unconsolidated clastic sediments, have low bulk densities (<1.7 g/cm^3), and rocks with very little porosity (<25%), such as igneous rocks, have high bulk densities (>2.4 g/cm^3). Lithologies composed of microfossils with silica tests, like diatomite, diatom ooze, diatom silt, and radiolarian ooze have the lowest density (<1.5 g/cm^3) and highest porosity (>70%), which can be attributed to how the silica tests of diatoms provide support for a porous skeletal structure.

Grouping data using categories other than principal lithology, such as the degree of consolidation or coring type, provide additional insights (Table S7): The average porosity of all the consolidated lithologies is 38% compared with 58% for unconsolidated lithologies. RCB cores are collected in material with an average porosity of 37% compared with 64% for APC cores and 50% for HLAPC.

4.3. MAD and GRA Densities Compared

Besides the MAD measurements on discrete samples, bulk density is measured indirectly by gamma ray attenuation (GRA) along whole-round core sections. While generally more accurate, the MAD measurements are time consuming and require destructive sampling and therefore cannot be conducted at the same frequency as indirect, track-based, non-destructive methods. As a result, the GRA density data set in LILY includes over 4 million measurements, more than two orders of magnitude more data than MAD measurements.
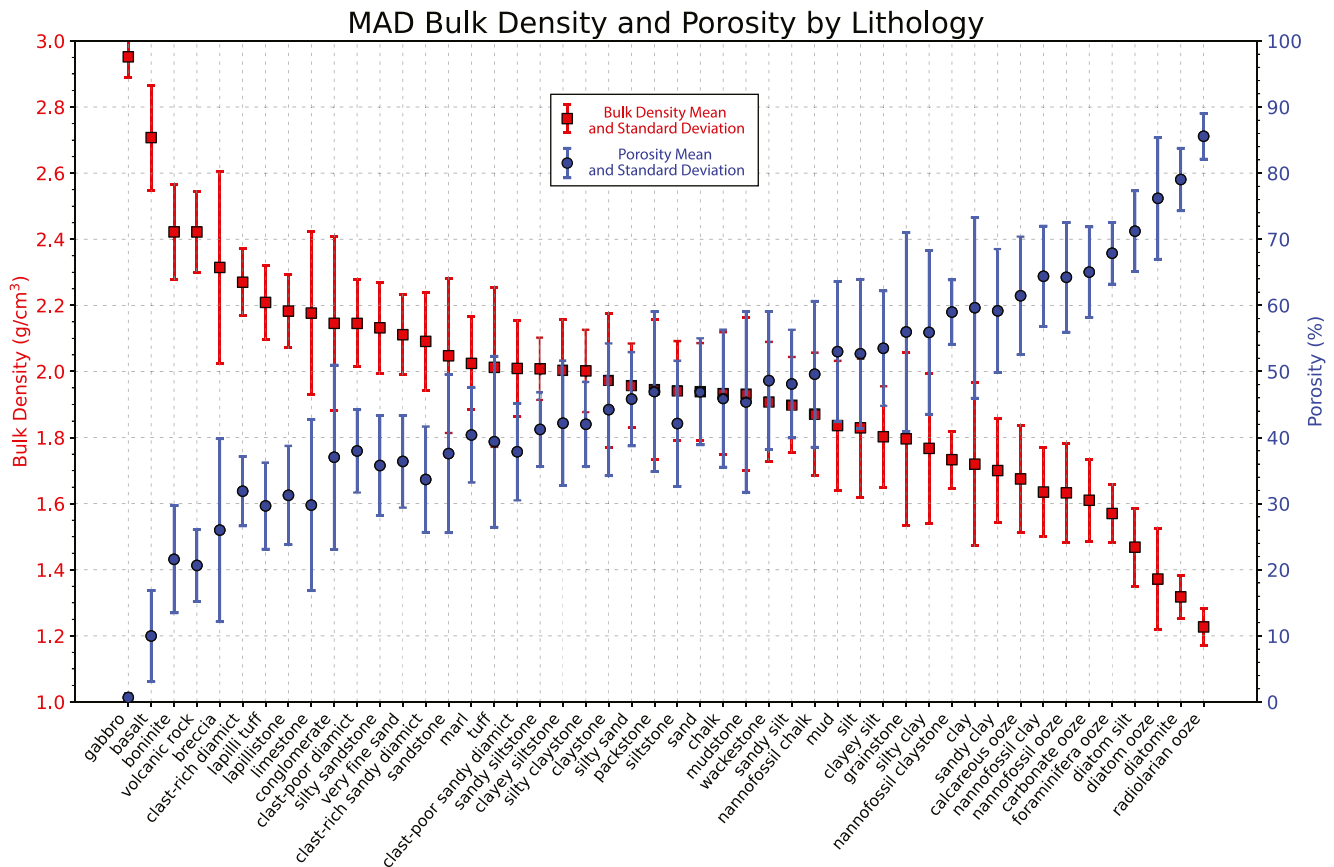
Figure 9. Mean bulk densities and porosities and their standard deviations for some of the principal lithologies with more than 50 observations each.

To compare the GRA bulk densities to the MAD bulk densities, we find the nearest GRA datum for each MAD datum, where the two data must come from the same core section and have the same principal lithology. We required that the paired data be no further than 5 cm from each other. This leaves 24,178 paired data for comparison, with over 95% of them within 2 cm of each other. The results compare remarkably well for APC, HLAPC, and XCB cores (Figure 10). The GRA densities are on average 0.036 g/cm^3 and 0.016 g/cm^3 higher than the MAD densities for APC and HLAPC, respectively, and they are -0.048 g/cm^3 lower for XCB. For APC and HLAPC, we attribute the higher GRA densities relative to MAD densities to the sediments being under compression in the core liner for GRA measurements. Once the cores are split, the subsamples taken for MAD measurements are no longer under compression and hence their densities are marginally lower. The slightly lower GRA values relative to MAD values for XCB cores would be expected because XCB cores often do not fill the core liner completely. The expectation is that there would be more outliers with GRA densities significantly lower than MAD densities due to GRA densities being measured over intervals where the core liner is only partly filled, and this tendency is observed. This tendency is much more apparent for RCB cores, where voids and intervals with partially filled core liner are much more abundant (see the "Anomalous" regions in Figure 10d).

RCB cores have significantly lower GRA densities than MAD densities (on average by 0.210 g/cm^3) and the difference depends on lithology and degree of lithification. The harder rock types, for example, basalt, gabbro, diorite, breccia, conglomerate, tuff, limestone, siltstone, and sandstone, typically have GRA densities 0.2–0.4 g/cm^3 less than the MAD densities. Softer, less consolidated sediments, such as mud, clay, silt, and ash, have GRA densities 0.1–0.2 g/cm^3 less than the MAD densities. The softest sediments, like nannofossil ooze, diatom ooze and diatomite, have GRA densities that are smaller than MAD densities on average by no more than 0.02 g/cm^3.

These differences predominantly result from RCB cores having a smaller diameter than the standard core diameter (6.6 cm), which is a set parameter in the software for GRA measurements. This interpretation is
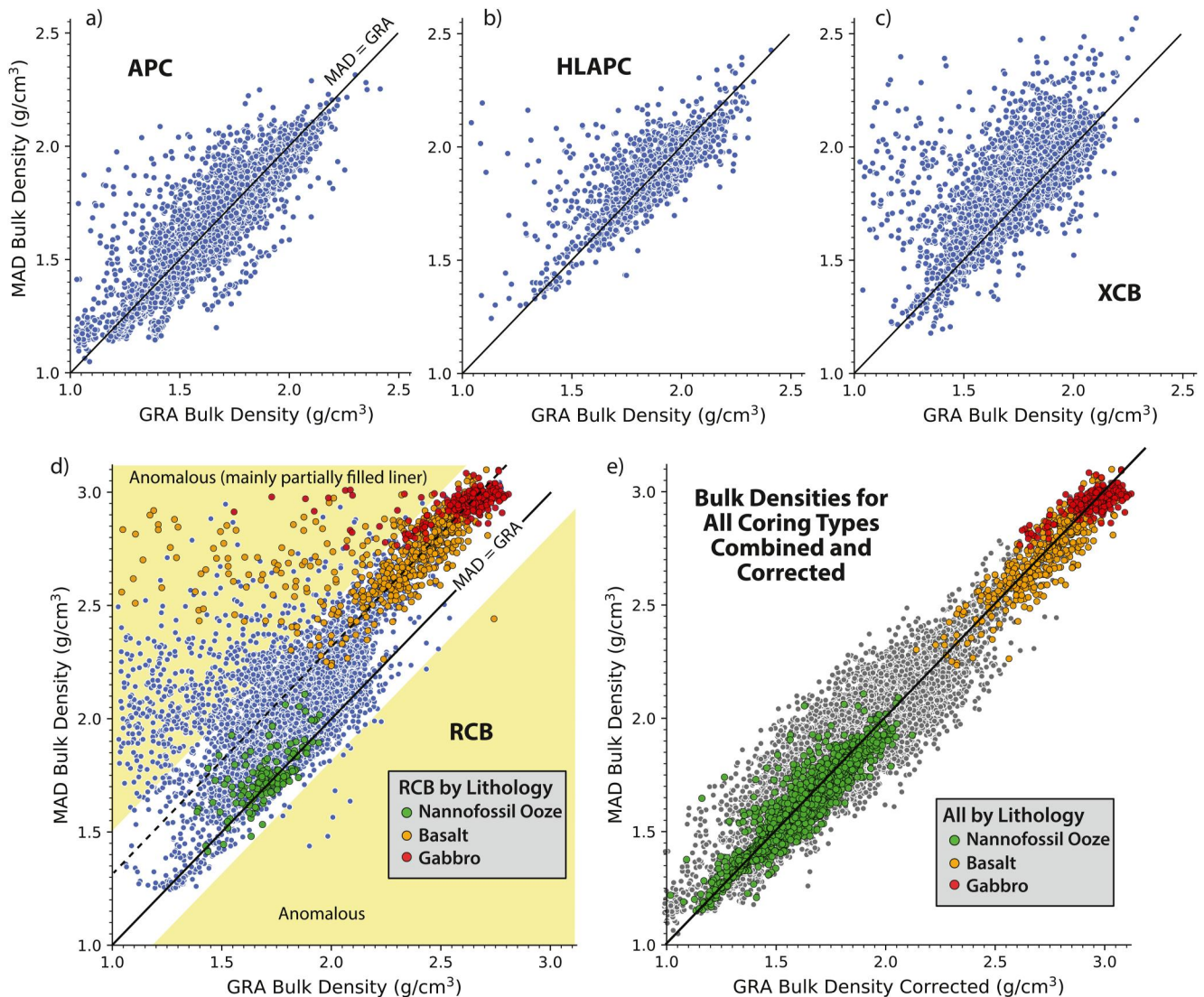
Figure 10. Comparison of GRA and MAD bulk densities measured from the same interval for (a) APC cores, (b) HLAPC cores, (c) XCB cores, and (d) RCB cores. In panels (a–d) the blue dots include all the data for the specific coring system. In panel d, besides plotting all the RCB data, data for three lithologies are superimposed to illustrate how GRA densities are more biased for some lithologies than for others. As discussed in the text, the difference between GRA and MAD densities is a bias in the GRA data that is dependent on the coring system and lithology. (e) We have determined the size of this bias as a function of those dependent variables and corrected for it. This panel shows data from all the coring systems combined (gray dots), with data from three specific lithologies superimposed and with outliers [from the "Anomalous" region shaded yellow in panel (d)] removed (see text).

supported not only by visibly examining cores but also by computing what the RCB core diameter is for different lithologies using the relationship:

$$d_{\text{true}} = (6.6 \text{ cm})(\rho_{\text{obs}}/\rho_{\text{true}}) \qquad (3)$$

where d_{true} is the actual core diameter, 6.6 cm is the assumed diameter, ρ_{obs} is the observed GRA bulk density, and ρ_{true} is the true bulk density, which can be estimated from the average of all the MAD bulk densities for each lithology. For example, the average MAD bulk density for basalt is 2.708 g/cm^3 and the average paired GRA bulk density is 2.267 g/cm^3, which indicates that the average diameter of the core is 5.53 cm (or about 5.81 cm if we exclude GRA data that differ more than 0.5 g/cm^3 from their paired MAD data, as discussed below). As noted above, even though the liner in the core barrel is 6.6 cm, the throat of the RCB bit is 5.87 cm. Unlike lithified rock, soft sediments collected with the RCB tend to expand and fill the liner. For example, the MAD bulk density for
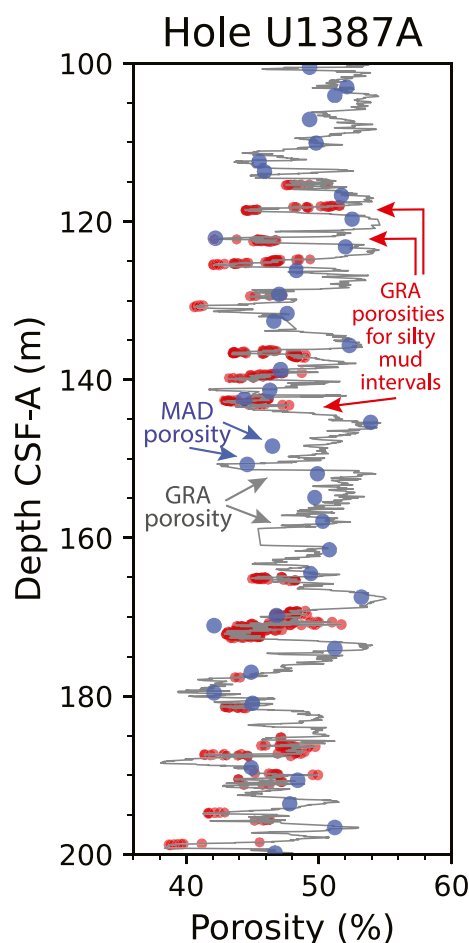
Hole U1387A

Figure 11. The new high-resolution GRA porosities (gray line) are plotted along with MAD porosities (blue dots). The GRA porosities for intervals with silty mud are plotted as red dots. The principal lithology for all other intervals is nannofossil mud.

nannofossil is 2.721 g/cm^3 and the averaged paired GRA bulk density (collected with the RCB system) is 2.701 g/cm^3, which indicates that the actual average diameter of a RCB core with nannofossil ooze is 6.55 cm.

It is apparent that the bias or difference between the GRA and MAD densities is dependent on the diameter of core recovered and the diameter is dependent on the coring system and lithology, with the coring systems themselves being lithology dependent. For APC, HLAPC, and XCB systems, little bias is evident and what bias there is does not have a clear dependence on lithology. Basically, the lithologies collected by these coring systems typically fill the core liner. For the RCB system, which collects a range of lithologies from hard rock to soft sediment, the lithology dependence is clear (Figure 10d).

Using LILY, we can compute the median differences between the paired GRA and MAD densities for each lithology collected with the RCB system and then use those values to correct all the GRA data (not just the paired data) from RCB cores (Table S8; Figure 10e). In computing the differences, we minimize the influence of data collected over intervals with partially filled core liners by using the median rather than mean and by excluding data where the GRA minus MAD results are <0.5 g/cm^3 or > −0.2 g/cm^3 (labeled as the "Anomalous" region in Figure 10d). Given the small differences between GRA and MAD densities for data from APC, HLAPC, and XCB cores, we use the median difference for each coring system rather than for each lithology as a robust correction factor. After excluding the anomalous values, the remaining 23,155 paired GRA and MAD data were used to compute correction factors (Table S8) and those factors were used to correct all the GRA bulk density data.

4.4. Porosities From GRA Densities

Before correcting the GRA bulk densities, we clean the GRA data similarly to how the MAD data were cleaned: duplicates and extreme outliers (densities <1.024 g/cm^3 or >3.2 g/cm^3) are removed. This leaves 3,752,145 GRA bulk densities, from which we compute corrected GRA bulk densities using the correction factors and compute porosities using the corrected GRA bulk densities, the MAD grain densities for the corresponding principal lithology, and Equation 2a (Data Set S2 in Acton et al., 2023; https://doi.org/10.5281/zenodo.10001854).

The new GRA porosity dataset has more than two orders of magnitude more data than in the MAD dataset and is spaced every few centimeters giving it a much higher resolution, as illustrated for an interval from Hole U1387A off the southern margin of Portugal (Figure 11). The sediment in this interval is predominantly nannofossil mud with thin interbeds of silty mud (Stow, Hernández-Molina, Alvarez Zarikian, Lofi, et al., 2013). GRA porosity shows a clear decrease associated with these silty mud layers (red dots along the gray line in Figure 11). This association was missed by the more widely spaced MAD porosity data. It is also apparent that had the sedimentologists had the high-resolution GRA porosity dataset, they may have identified a few other narrow intervals that are likely silty mud.

4.5. Comparison of IODP and DSDP Density Compilations

Tenzer and Gladkikh (2014) compiled bulk density measurements from 716 DSDP drill sites (21,937 samples). These data are similar to a mix of IODP MAD and GRA bulk density data because the DSDP data include results from discrete samples and from measurements of whole-round core samples by gamma ray attenuation (e.g., Boyce, 1976; Mills, 1985; Scientific Party, 1970). In addition, DSDP discrete sample volumes were not estimated by a pycnometer, but were instead estimated from the shape of syringe and mini-core cylinder samples or by water displacement for irregular shaped chunks of rock (e.g., Lee, 1973). As noted by Boyce (1976), the DSDP GRA densities are biased toward lower values when the core does not fill the core liner. This is exactly what we noted
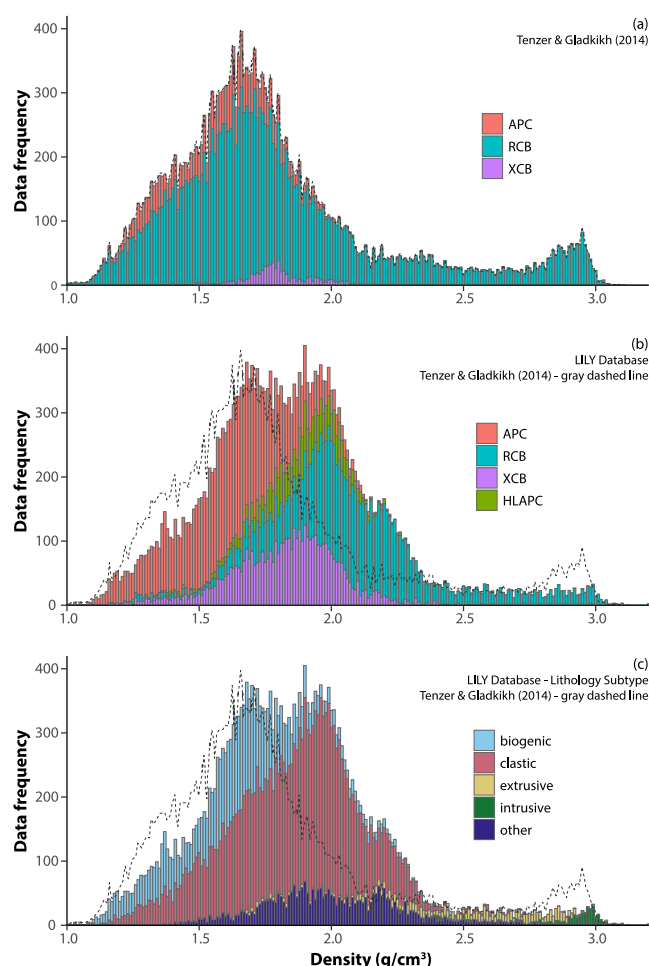
Figure 12. IODP MAD bulk density compared with the DSDP bulk density compilation of Tenzer and Gladkikh (2014).

for the IODP GRA densities before we corrected them. The bias remains in the DSDP data, whereas it is absent from the IODP MAD data and has been removed from the IODP GRA data via the correction we applied (see Section 4.3 MAD and GRA Densities Compared).

This bias is apparent in the comparison of the DSDP densities with the IODP MAD densities (Figure 12). The DSDP data are on average about 0.2 g/cm^3 lower than the IODP MAD data, which is an amount virtually identical to what we noted for the GRA bulk densities for RCB cores. Note that the DSDP data are mostly from RCB cores (Figure 12a).

The distribution of DSDP versus IODP density values differs in at least two other notable ways. The DSDP distribution contains a small secondary mode between 2.8 and 3.1 g/cm^3 that is absent in the IODP data. About 45% of these data come from just three holes (395A, 462A, and 504B), which targeted mostly igneous basement. The GRA bias is not evident in these studies because these data are predominantly from discrete sample studies in which the bulk density was determined from weighing the sample and determining the volume from water displacement (e.g., Karato, 1983). The larger number of samples from igneous rocks reflects both the greater emphasis in coring oceanic basement by DSDP and the way cores were collected. During DSDP, spot coring was frequently employed to target intervals of interest, particularly igneous rock, while drilling through much of the sedimentary overburden.

Another noticeable difference is the bimodality of the primary maxima of IODP data for densities <2.5 g/cm^3, whereas the DSDP data have a single maximum. The two peaks in the IODP data are associated with differences in coring systems used and types of lithologies recovered. RCB coring was the most common method used by DSDP prior to 1979, when the APC coring system was introduced. Thus, during DSDP much of the soft sediment that would be cored currently with an APC coring system was either drilled through or cored by RCB, resulting in much less recovery of soft sediment, which is quite evident in Figures 12a and 12b. Furthermore, when the bias in unfilled core liners is considered, the mode at about 1.65 g/cm^3 in the DSDP distribution actually corresponds to the mode at about 1.85 g/cm^3 in the IODP distribution. Thus, the mode at 1.65 g/cm3 in the IODP data is absent in the DSDP data, mainly because DSDP collected so few cores with the APC system and none with the XCB or HLAPC systems.

5. Conclusions

The continuous collection of consistent types of data on IODP expeditions over a long time makes them a natural target for broad studies that span beyond a single site or expedition. The scientific ocean drilling database is big data and an endless array of studies can be conducted with it. Undoubtedly, many new discoveries will be made with those data. That was true of the IODP data even before we added descriptive lithologic information to each datum to make LILY.

The many datasets that make up the IODP LIMS database are, however, from measurements collected along geologic units or from samples taken from those geologic units and it is only natural that the type of geologic unit be part of the metadata for each measurement. LILY ensures that information is part of the IODP data and enables users to investigate the properties and characteristics of the data that may be connected by or depend on lithology.

As we have shown in an example using MAD and GRA data, the addition of descriptive lithologic information allows new attributes of datasets to be resolved, features to be characterized, and hypotheses to be explored. In our example, we have used LILY to estimate grain densities from lithology from MAD data, to correct a bias that existed in the GRA bulk densities that was dependent on lithology and coring system, and to compute a high-resolution porosity dataset using the lithologic-dependent grain densities and the corrected GRA bulk

densities. With over 3.7 million new porosity estimates, this dataset should be of use in identifying fluid pathways, estimating reservoir capacity, determining the CO_2 sequestration potential for geologic units, investigating how changes in porosity with the depth might be related to landslide or earthquake nucleation or facilitation, and much more. We hope future investigators will uncover endless possibilities for the use of the LILY database and derive new observations and insights from it.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

The LILY database is available on Zenodo in the International Ocean Discovery Program community (Childress et al., 2023; Data Set S1, https://doi.org/10.5281/zenodo.8408297). The corrected IODP densities and calculated porosities derived from the LILY Database are also available on Zenodo (Acton et al., 2023; Data Set S2, https://doi.org/10.5281/zenodo.10001854). Data from the LIMS and Janus databases (https://web.iodp.tamu.edu/LORE/; http://www-odp.tamu.edu/database/) were used in this manuscript. Data (e.g., DESClogik workbooks from early expeditions that were improperly preserved) not available through LIMS or Janus are provided in the Supporting Information S1 or in Zenodo as part of the LILY database. Figures were made with R and Python/Jupyter. R version 4.2.3 (R Core Team, 2023) is available at https://www.R-project.org/. Python programming was done with version 3.9.7 in Visual Studio Code Jupyter Notebooks. Example programs for reading LILY data, cleaning data, plotting, and merging different data types are available in Childress et al. (2024).

References

Acton, G., Childress, L., Percuoco, V., & Hastedt, M. (2023). Corrected IODP gamma ray attenuation (GRA) densities and calculated porosities derived from the LILY database [Dataset]. Geochemistry, Geophysics, Geosystems. Zenodo. Vol. TBD (2023-12-26, Number TBD). https://doi.org/10.5281/zenodo.10001855

Arculus, R. J., Ishizuka, O., Bogus, K., & the Expedition 351 Scientists. (2015). In *Proceedings of the international ocean discovery program, expedition 351: Izu-Bonin-Mariana arc origins: College Station, TX*. International Ocean Discovery Program. https://doi.org/10.14379/iodp.proc.351.2015

Betzler, C., Eberli, G. P., Alvarez Zarikian, C. A., & the Expedition 359 Scientists. (2017). In *Proceedings of the international ocean discovery program, 359, Maldives monsoon and sea level: College Station, TX*. International Ocean Discovery Program. https://doi.org/10.14379/iodp.proc.359.2017

Blum, P. (1997). Physical properties handbook: A guide to the shipboard measurement of physical properties of deep-sea cores. ODP tech. Note, 26. https://doi.org/10.2973/odp.tn.26.1997

Boyce, R. E. (1976). Definitions and laboratory techniques of compressional sound velocity parameters and wet-water content, wet-bulk density, and porosity parameters by gravimetric and gamma ray attenuation techniques. https://doi.org/10.2973/dsdp.proc.33.app1.1976

Childress, L. B., Acton, G. D., Percuoco, V. P., & Hastedt, M. (2023). The LILY database: Linking lithology to IODP physical, chemical, and magnetic properties data [Dataset]. Geochemistry, Geophysics, Geosystems. Zenodo. Vol. TBD (2023-12-26, Number TBD). https://doi.org/10.5281/zenodo.8408297

Childress, L. B., Acton, G. D., Percuoco, V. P., & Hastedt, M. (2024). IODP/LILY: LILY supplement [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.10463680

Childress, L. B., Alvarez Zarikian, C. A., Briais, A., Dadd, K. A., Deng, J.-M., et al. (2020). Return to site U1503. In Z. Sun, Z. Jian, J. M. Stock, H. C. Larsen, A. Klaus, C. A. Alvarez Zarikian, & the Expedition 367/368 Scientists. (Eds.), *Proceedings of the international ocean discovery program, 367/368: College Station, TX*. International Ocean Discovery Program. https://doi.org/10.14379/iodp.proc.368X.103.2020

Clemens, S. C., Kuhnt, W., LeVay, L. J., & the Expedition 353 Scientists. (2016). In *Proceedings of the international ocean discovery program, 353, Indian monsoon rainfall: College Station, TX*. International Ocean Discovery Program. https://doi.org/10.14379/iodp.proc.353.2016

Dean, W. E., Leinen, M., & Stow, D. A. V. (1985). Classification of deep-sea, fine-grained sediments. *Journal of Sedimentary Research*, *55*(2), 250–256. https://doi.org/10.1306/212F868E-2B24-11D7-8648000102C1865D

de Ronde, C. E. J., Humphris, S. E., Höfig, T. W., & the Expedition 376 Scientists. (2019). In *Proceedings of the international ocean discovery program, 376, Brothers arc flux: College Station, TX*. International Ocean Discovery Program. https://doi.org/10.14379/iodp.proc.376.2019

D'Hondt, S., Inagaki, F., Alvarez Zarikian, C. A., & the Expedition 329 Scientists. (2011). In *Proceedings of the international ocean drilling program, 329*. Integrated Ocean Drilling Program Management International, Inc. https://doi.org/10.2204/iodp.proc.329.2011

Edwards, K. J., Bach, W., Klaus, A., & the Expedition 336 Scientists. (2012). In *Proceedings of the international ocean drilling program, 336: Tokyo*. Integrated Ocean Drilling Program Management International, Inc. https://doi.org/10.2204/iodp.proc.336.2012

Escutia, C., Brinkhuis, H., Klaus, A., & the Expedition 318 Scientists. (2011). In *Proceedings of the international ocean drilling program, 318: Tokyo*. Integrated Ocean Drilling Program Management International, Inc. https://doi.org/10.2204/iodp.proc.318.2011

Evans, H. F. (2020). Relationships between core recovery, coring systems, and sedimentary lithology for scientific ocean drilling. Retrieved from http://iodp.tamu.edu/publications/JRSO/TR_1/TechReport_TR_1.pdf

Expedition 320T Scientists (2009). USIO sea trials and assessment of readiness transit (START): Ontong Java Plateau. IODP Prel. Rept., 320T. https://doi.org/10.2204/iodp.pr.320T.2009

Expedition 335 Scientists (2012). Methods. In D. A. H. Teagle, B. Ildefonse, P. Blum, & Expedition 335 Scientists (Eds.), *Proceedings of the international ocean drilling program, 335: Tokyo*. Integrated Ocean Drilling Program. https://doi.org/10.2204/iodp.proc.335.102.2012

Fisher, A. T., Tsuji, T., Petronotis, K., & the Expedition 327 Scientists. (2011). In *Proceedings of the international ocean drilling program, 327: Toky*. Integrated Ocean Drilling Program Management International, Inc. https://doi.org/10.2204/iodp.proc.327.2011

France-Lanord, C., Spiess, V., Klaus, A., Schwenk, T., & the Expedition 354 Scientists. (2016). In *Proceedings of the international ocean discovery program, 354, Bengal fan: College Station, TX*. International Ocean Discovery Program. https://doi.org/10.14379/iodp.proc.354.2016

Fryer, P., Wheat, C. G., Williams, T., & the Expedition 366 Scientists. (2018). In *Proceedings of the international ocean discovery program, 366, Mariana convergent margin and South Chamorro seamount: College Station, TX*. International Ocean Discovery Program. https://doi.org/10.14379/iodp.proc.366.2018

Gallagher, S. J., Fulthorpe, C. S., Bogus, K., & the Expedition 356 Scientists. (2017). In *Proceedings of the international ocean discovery program, 356, Indonesian throughflow: College Station, TX*. International Ocean Discovery Program. https://doi.org/10.14379/iodp.proc.356.2017

Gillis, K. M., Snow, J. E., Klaus, A., & the Expedition 345 Scientists. (2014). In *Proceedings of the International ocean drilling program, 345: College Station, TX*. Integrated Ocean Drilling Program. https://doi.org/10.2204/iodp.proc.345.2014

Gohl, K., Wellner, J. S., Klaus, A., & the Expedition 379 Scientists. (2021). In *Proceedings of the international ocean drilling program, 379, Amundsen Sea West Antarctic ice sheet history: College Station, TX*. International Ocean Discovery Program. https://doi.org/10.14379/iodp.proc.379.2021

Hall, C., & Hamilton, A. (2015). Porosity–density relations in stone and brick materials. *Materials and Structures/Materiaux et Constructions*, *48*(5), 1265–1271. https://doi.org/10.1617/s11527-013-0231-1

Hall, I. R., Hemming, S. R., LeVay, L. J., & the Expedition 361 Scientists. (2017). In *Proceedings of the international ocean discovery program, 361, South African climates (Agulhas LGM density profile): College Station, TX*. International Ocean Discovery Program. https://doi.org/10.14379/iodp.proc.361.2017

Harris, R. N., Sakaguchi, A., Petronotis, K., & the Expedition 344 Scientists. (2013). In *Proceedings of the international ocean drilling program, 344: College Station, TX*. Integrated Ocean Drilling Program. https://doi.org/10.2204/iodp.proc.344.2013

Hobbs, R. W., Huber, B. T., Bogus, K. A., & the Expedition 369 Scientists. (2019). In *Proceedings of the international ocean discovery program, 369, Australia cretaceous climate and tectonics: College Station, TX*. International Ocean Discovery Program. https://doi.org/10.14379/iodp.proc.369.2019

Huber, B. T., Hobbs, R. W., Bogus, K. A., Batenburg, S. J., Brumsack, H.-J., do Monte Guerra, R., & the Expedition 369 methods. (2019). In R. W. Hobbs, B. T. Huber, K. A. Bogus, & Expedition 369 Scientists (Eds.), *Australia cretaceous climate and tectonics. Proceedings of the international ocean discovery program, 369: College Station, TX*. International Ocean Discovery Program.

Jaeger, J. M., Gulick, S. P. S., Asahi, H., Belanger, C. L., Berbel, G., Childress, L. B., et al. (2014). In *Proceedings of the integrated ocean drilling program*. In *Methods* (Vol. 341). https://doi.org/10.2204/iodp.proc.341.102.2014

Jaeger, J. M., Gulick, S. P. S., LeVay, L. J., & the Expedition 341 Scientists. (2014). In *Proceedings of the international ocean drilling program, 341: College Station, TX*. Integrated Ocean Drilling Program. https://doi.org/10.2204/iodp.proc.341.2014

Karato, S. (1983). Shipboard physical properties measurements of basalts from the Costa Rica rift deep sea drilling project legs 69 and 70. In *Initial reports of the deep sea drilling project* (Vol. 69, p. 675). https://doi.org/10.2973/dsdp.proc.69.141.1983

Koppers, A. A. P., Yamazaki, T., & Geldmacher, J., & the Expedition 330 Scientists. (2012). In *Proceedings of the international ocean drilling program, 330: Tokyo*. Integrated Ocean Drilling Program Management International, Inc. https://doi.org/10.2204/iodp.proc.330.2012

Lee, H. J. (1973). Measurements and estimates of engineering and other physical properties, leg 19. https://doi.org/10.2973/dsdp.proc.19.122.1973

Le Friant, A., Ishizuka, O., Stroncik, N. A., & the Expedition 340 Scientists. (2013). In *Proceedings of the international ocean drilling program, 340: Tokyo*. Integrated Ocean Drilling Program Management International, Inc. https://doi.org/10.2204/iodp.proc.340.2013

Li, C.-F., Lin, J., Kulhanek, D. K., & the Expedition 349 Scientists. (2015). In *Proceedings of the international ocean discovery program, 349: South China Sea tectonics: College Station, TX*. International Ocean Discovery Program. https://doi.org/10.14379/iodp.proc.349.2015

MacLeod, C. J., Dick, H. J. B., Blum, P., & the Expedition 360 Scientists. (2017). In *Proceedings of the international ocean discovery program, 360, southwest Indian ridge lower crust and Moho: College Station, TX*. International Ocean Discovery Program. https://doi.org/10.14379/iodp.proc.360.2017

Marsaglia, K. M., & Milliken, K. L. (2023). Technical note 5: Handbook for sedimentological core description: Shipboard and shorebased. In *International ocean discovery program, 83*. https://doi.org/10.14379/iodp.tn.5.2023

Mazzullo, J., Meyer, A., & Kidd, R. (1988). Appendix I: New sediment classification scheme for the ocean drilling program. In J. Mazzullo & A. G. Graham (Eds.), *Handbook for shipboard sedimentologists. Ocean drilling program technical note: College Station, TX* (Vol. 8, pp. 44–67). Ocean Drilling Program. https://doi.org/10.2973/odp.tn.8.1988

McKay, R. M., De Santis, L., Kulhanek, D. K., & the Expedition 374 Scientists. (2019). In *Proceedings of the international ocean discovery program, 374, Ross Sea West Antarctic ice sheet history: College Station, TX*. International Ocean Discovery Program. https://doi.org/10.14379/iodp.proc.374.2019

McNeill, L. C., Dugan, B., Petronotis, K. E., & the Expedition 362 Scientists. (2017). In *Proceedings of the international ocean discovery program, 362, Sumatra subduction zone: College Station, TX*. International Ocean Discovery Program. https://doi.org/10.14379/iodp.proc.362.2017

Mills, W. G. (1985). Explanatory notes, deep sea drilling project leg 82. In *Initial reports of the deep sea drilling project* (Vol. 82, p. 45). https://doi.org/10.2973/dsdp.proc.82.104.1985

Norris, R. D., Wilson, P. A., Blum, P., & the Expedition 342 Scientists. (2014). In *Proceedings of the international ocean drilling program, 342: College Station, TX*. Integrated Ocean Drilling Program. https://doi.org/10.2204/iodp.proc.342.2014

Pälike, H., Lyle, M., Nishi, H., Raffi, I., Gamage, K., Klaus, A., & the Expedition 320/321 Scientists. (2010). In *Proceedings of the international ocean drilling program, 320/321: Tokyo*. Integrated Ocean Drilling Program Management International, Inc. https://doi.org/10.2204/iodp.proc.320321.2010

Pandey, D. K., Clift, P. D., Kulhanek, D. K., & the Expedition 355 Scientists. (2016). In *Proceedings of the international ocean discovery program, 355, Arabian Sea monsoon: College Station, TX*. International Ocean Discovery Program. https://doi.org/10.14379/iodp.proc.355.2016

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

Reagan, M. K., Pearce, J. A., Petronotis, K., & the Expedition 352 Scientists. (2015). In *Proceedings of the international ocean discovery program, 352, Izu-Bonin-Mariana fore arc: College Station, TX*. International Ocean Discovery Program. https://doi.org/10.14379/iodp.proc.352.2015

Rosenthal, Y., Holbourn, A. E., Kulhanek, D. K., & the Expedition 363 Scientists. (2018). In *Proceedings of the international ocean discovery program, 363, western Pacific warm pool: College Station, TX*. International Ocean Discovery Program. https://doi.org/10.14379/iodp.proc.363.2018

Sager, W. W., Sano, T., Geldmacher, J., & the Expedition 324 Scientists. (2010). In *Proceedings of the international ocean drilling program, 324: Tokyo*. Integrated Ocean Drilling Program Management International, Inc. https://doi.org/10.2204/iodp.proc.324.2010

Scientific Party. (1970). Explanatory notes. In *Initial reports of the deep sea drilling project* (Vol. 3, p. 11). https://doi.org/10.2973/dsdp.proc.3.102.1970

Smyth, J. R., & McCormick, T. S. (1995). Crystallographic data for minerals. In T. J. Ahrens (Ed.), *Mineral physics and crystallography: A handbook of physical constants* (Vol. 2). American Geophysical Union. https://doi.org/10.1029/RF002p0001

Stein, R., Hefter, J., Grützner, J., Voelker, A., & Naafs, B. D. A. (2009). Variability of surface water characteristics and Heinrich-like events in the Pleistocene midlatitude North Atlantic Ocean; biomarker and XRD records from IODP Site U1313 (MIS 16-9). *Paleoceanography*, *24*(2), PA2203. https://doi.org/10.1029/2008PA001639

Stein, R., Kanamatsu, T., Alvarez Zarikian, C., Higgins, S. M., Channell, J. E. T., Aboud, E., et al. (2006). North Atlantic paleoceanography; the last five million years. *Eos, Transactions of the American Geophysical Union*, *87*(13), 129–133. https://doi.org/10.1029/2006EO130002

Storms, M. A. (1990). Ocean Drilling Program (ODP) deep sea coring techniques. *Marine Geophysical Researches*, *12*(1–2), 109–130. https://doi.org/10.1007/BF00310567

Stow, D. A. V., Hernández-Molina, F. J., Alvarez Zarikian, C., Lofi, J., Williams, T., Acton, G. D., et al. (2013). In *Proceedings of the integrated ocean drilling program; Mediterranean outflow; expedition 339 of the riserless drilling platform; Ponta Delgada, Azores (Portugal), to Lisbon, Portugal, sites U1385-U1391, 16 November 2011–2016 January 2012*. IODP Management International. https://doi.org/10.2204/iodp.proc.339.2013

Stow, D. A. V., Hernández-Molina, F. J., Alvarez Zarikian, C. A., & the Expedition 339 Scientists. (2013). In *Proceedings of the international ocean drilling program, 339: Tokyo*. Integrated Ocean Drilling Program Management International, Inc. https://doi.org/10.2204/iodp.proc.339.2013

Sun, Z., Jian, Z., Stock, J. M., Larsen, H. C., Klaus, A., Alvarez Zarikian, C. A., & the Expedition 367/368 Scientists. (2018). In *Proceedings of the international ocean discovery program, 367/368, South China Sea rifted margin: College Station, TX*. International Ocean Discovery Program. https://doi.org/10.14379/iodp.proc.367368.2018

Sutherland, R., Dickens, G. R., Blum, P., & the Expedition 371 Scientists. (2019). In *Proceedings of the international ocean discovery program, 371, Tasman Frontier subduction initiation and Paleogene climate: College Station, TX*. International Ocean Discovery Program. https://doi.org/10.14379/iodp.proc.371.2019

Tada, R., Murray, R. W., Alvarez Zarikian, C. A., & the Expedition 346 Scientists. (2015). In *Proceedings of the international ocean drilling program, 346: College Station, TX*. Integrated Ocean Drilling Program. https://doi.org/10.2204/iodp.proc.346.2015

Takahashi, K., Ravelo, A. C., Alvarez Zarikian, C. A., & the Expedition 323 Scientists. (2011). In *Proceedings of the international ocean drilling program, 323*. Integrated Ocean Drilling Program Management International, Inc. https://doi.org/10.2204/iodp.proc.323.2011

Tamura, Y., Busby, C. J., Blum, P., & the Expedition 350 Scientists. (2015). In *Proceedings of the International Ocean Discovery program, expedition 350: Izu-Bonin-Mariana rear arc: College Station, TX*. International Ocean Discovery Program. https://doi.org/10.14379/iodp.proc.350.2015

Teagle, D. A. H., Ildefonse, B., Blum, P., & the Expedition 335 Scientists. (2012). In *Proceedings of the international ocean drilling program, 335: Tokyo*. Integrated Ocean Drilling Program Management International, Inc. https://doi.org/10.2204/iodp.proc.335.2012

Tenzer, R., & Gladkikh, V. (2014). Assessment of density variations of marine sediments with ocean and sediment depths. *The Scientific World Journal*, *9*, 1–9. https://doi.org/10.1155/2014/823296

The Shipboard Scientific Party. (1970). Explanatory notes—Leg 2. https://doi.org/10.2973/dsdp.proc.2.102.1970

Vannucchi, P., Ujiie, K., Stroncik, N., & the Expedition 334 Scientists. (2012). In *Proceedings of the international ocean drilling program, 334: Tokyo*. Integrated Ocean Drilling Program Management International, Inc. https://doi.org/10.2204/iodp.proc.334.2012

Wallace, L. M., Saffer, D. M., Barnes, P. M., Pecher, I. A., Petronotis, K. E., LeVay, L. J., & the Expedition 372/375 Scientists. (2019). In *Proceedings of the international ocean discovery program, 372B/375: College Station, TX*. Hikurangi Subduction Margin Coring, Logging, and Observatories. International Ocean Discovery Program. https://doi.org/10.14379/iodp.proc.372B375.2019