Data Article

A synthesized dataset of Pacific Arctic marine mammal occurrences (1860-2024)

Check for
updates

Sydney Waloven*, Natalia Portales, Kelly Kapsar, Jianguo Liu*

Manly Miles Rm 115, 1405 S Harrison Rd East Lansing MI 48823, Center for Systems Integration and Sustainability, Department of Fisheries & Wildlife, Michigan State University, USA

A R T I C L E   I N F O

A B S T R A C T

Data on marine mammals in the Pacific Arctic region are limited, variably collected over time and by multiple sources. Therefore, collation and synthesis of previously collected data, such as the dataset presented here, make these data more useful to researchers, managers, policy makers, and residents. This paper presents a synthesized dataset of marine mammal occurrences in the Pacific Arctic from 1860 to 2024. Specifically, the dataset focuses on occurrence records of three species – Pacific walrus (*Odobenus rosmarus*), spotted seal (*Phoca largha*), and bearded seal (*Erignathus barbatus*). These species were chosen because of their ecological importance to the Pacific Arctic ecosystem as well as their cultural and subsistence importance to local Indigenous communities. Data were aggregated from open-access, public online data repositories. Datasets were tidied into a standardized format and then integrated to create a singular dataset of occurrence data for these three species. The final dataset contains 36,438 presence-only occurrence points over 164 years throughout the Bering, Chukchi, and Beaufort Seas. This dataset provides stakeholders with quantitative data that can be used to evaluate spatial patterns of marine mammal species over time in the Pacific Arctic region, which can generate insights into the effects of human disturbances (e.g., shipping, resource

* Corresponding authors.
  E-mail addresses: walovens@msu.edu (S. Waloven), liuji@msu.edu (J. Liu).

extraction) and climate change when combined with other environmental variables.

Specifications Table

Subject	Environmental Science
Specific subject area	Ecology, marine wildlife, marine spatial planning
Type of data	Table
	Filtered, Processed
Data collection	Data were obtained from open-access online databases. These data were either in a raw format or had been tidied by the data collector previously. Each individual dataset was further manipulated to all be in a consistent format
	For a dataset to be included, it had to contain the following information:
	- Species name (either scientific or common name)
	- Date of occurrence (including year, month, day)
	- Coordinates of occurrence
	- Method of observation or instrument used for data collection
	Data were excluded if they did not meet the above criteria. All data were processed and standardized in the statistical software R.
Data source location	City/Town/Region: North Pacific and Arctic Oceans
	Country: USA, Russia, International
	Latitude and longitude (GPS coordinates, if possible) for collected samples/data:
	Approximate bounding box: 172E, 46 N, to 146 W, 73N
	Primary Data Sources (see full reference list in Table 2):
	• GBIF.org (03 July 2024) **GBIF Walrus Occurrence Download**
	• GBIF.org (03 July 2024) **GBIF Spotted Seal Occurrence Download**
	• GBIF.org (03 July 2024) **GBIF Bearded Seal Occurrence Download**
	• Nelson, T., Blees, M. **TGS Seismic Survey Visual Sightings in Chukchi Sea 2013**
	• Kuletz, K., Cushing, D., Labunski, E., Iken, K. **Marine bird sighting data, Chukchi Sea research on the vessel Norseman II from 2015–08–09 to 2015–09–03**
	• Moore, S. **Marine Mammal Watch, Northern Bering Sea and Chukchi Sea, July 2016**
	• Moore, S. **Marine Mammal Watch, Northern Bering Sea and Chukchi Sea, August-September 2017**
	• Stafford, K. **Marine mammal sightings during Sir Wilfrid Laurier cruise to the Pacific Arctic 11–22 July 2019**
	• Stafford, K. **Marine mammal sighting data from cruises in the Pacific Arctic, 2018, from Distributed Biological Observatory (DBO) regions**
	• Ferguson, M. **Marine mammal aerial surveys in the Bering, Chukchi and Beaufort Seas 1979–2010**
	• Jay, C.V., Quakenbush, L., Citta, J., Fischbach, A.S., Battaile, B. **Sex and age composition of walrus groups hauled out on ice floes in the Bering and Chukchi Seas, 2013–2015**
	• Battaile, B., Jay, C.V., Fischbach, A.S. **Behavior of Pacific walruses (*Odobenus rosmarus divergens*) hauled out on sea ice during UAS overflights, Eastern Chukchi Sea, 2015**
	• Moudrý, V., Devillers, R. **Quality and usability challenges of global marine biodiversity databases: An example for marine mammal data**
	• BOEM, NOAA, North Slope Borough. **Aerial Surveys of Arctic Marine Mammals (ASAMM) collected by Marine Mammal Laboratory, Bureau of Ocean Energy Management, and other agencies in the Bering, Chukchi, Beaufort Seas, and Amundsen Gulf from 1979 to 2019 and North Slope Borough, Alaska from 2020 to 2021 (NCEI Accession 0039,614)**
	• Hobbs, R. **NMML Small Cetacean Aerial Survey 1999**
	• Barry, T., Helgason, H., McRae, L., Gill, M., Svoboda, M., Eamer, J. **Conservation of Arctic Flora and Fauna: Arctic Species Trend Index (ASTI): Marine**

- Link, O., Russell, L., Bloom, D., Wieczorek, J., McDonald, D., Gunderson, A., Kyndall, H. **UAM Mammal Collections (Arctos)**
- Waite, J. **NMML Bering Sea Cetacean Survey 2000**
- Ray, G., Hufford, G., Krupnik, I., Labunski, E. **HLY-08–01 and HLY-08–02 Observations of marine mammals of Beringia [Ray, G.C.]**
- Ray, G., Hufford, G., Krupnik, I., Labunski, E. **HLY-07–01 and HLY-07–02 Observations of Marine Mammals of Beringia [Ray, G.C.]**
- Holst, M., Lee, O., Smith, H. **Lamont-Doherty/LGL/NSF cruises**
- Stafford, K. **Acoustic detections of Arctic mammals in the western Beaufort Sea 2010–2011**
- Stafford, K., Moore, S., Iken, K., Turner, C., Benson, A., Biddle, M. **Vessel line-transect surveys of Arctic pinnipeds in the Chukchi Sea, Arctic Marine Biodiversity Observing Network (AMBON) research cruise on the vessel Norseman II, 2015–08–09 to 2015–09–03**
- Stafford, K., Iken, K., Moore, S., Turner, C., Benson, A., Biddle, M. **Vessel line-transect surveys of Arctic pinnipeds in the Chukchi Sea, Arctic Marine Biodiversity Observing Network (AMBON) research cruise on the vessel Norseman II, 2017–08–05 to 2017–08–25**

Data accessibility	Repository name: NSF Arctic Data Center
	Data identification number: doi: 10.18739/A2WH2DH0V
	Direct URL to data: https://doi.org/10.18739/A2WH2DH0V
Related research article	None

1. Value of the Data

- This unique dataset synthesizes occurrences for Pacific walrus (*Odobenus rosmarus*), spotted seal (*Phoca largha*), and bearded seal (*Erignathus barbatus*) from 1860 to 2024, combining public data into one standardized resource to improve the data's usability for analysis.
- Data are prepared and readily available to be applied to maps, statistical models, and other spatial representations. Occurrences with missing or inaccurate information (e.g., coordinates, species name, date) have been eliminated from this dataset.
- These data can support a variety of applications, including studies of marine mammal distributions over time and the impacts of environmental change in the Pacific Arctic. They are valuable to local stakeholders, Indigenous communities, ecologists, marine biologists, resource managers, policymakers, government agencies, commercial fishers, and marine environmental organizations. These data can be useful for wildlife and marine ecosystem management.
- While this dataset is the first of its kind, it will require further refinement to allow species occurrences from new data sources and formats to be integrated in the future. However, the structure of this dataset can be replicated or expanded to focus on different species of interest (e.g., whales, porpoises, ringed seals, harbor seals, polar bears, etc.).

2. Background

The Pacific Arctic provides critical habitat to many Arctic species and supports the lives of local Indigenous communities and those around the globe that depend on the region for its resources. Ensuring the sustainability and protection of the Pacific Arctic is important to prevent repercussions on the well-being and security of the populations, both humans and wildlife, that are interconnected. Marine mammal species that live here—such as the walrus, spotted seal, and bearded seal—play vital roles in marine ecosystems, trophic food webs, and as substantial subsistence resources for Native coastal communities [1,2]. However, these species are at risk of various human-induced threats and are ranked under the U.S. Marine Mammal Protection Act (MMPA) or Endangered Species Act. Despite the range of marine mammals that inhabit the region, data on marine mammal occurrences are archived in many different locations and in inconsistent formats that make re-use challenging. This dataset provides a valuable resource for those
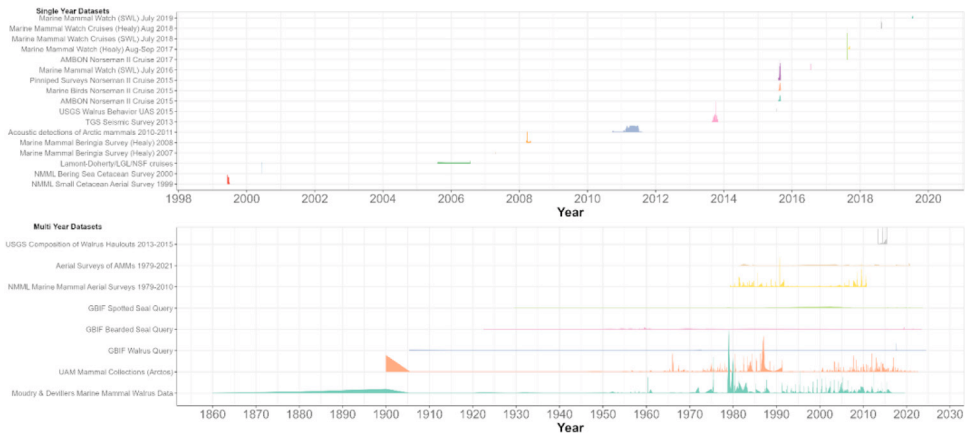
Fig. 1. Ridgeline plot representing data collection periods for each dataset. Height of peaks correspond to the density of occurrences at that time. Single year datasets represent datasets that were collected within one calendar year. Multi year datasets represent datasets that span multiple years.

focused on addressing knowledge gaps in the spatial and temporal distribution of the species of interest. The dataset format is adaptable and can be manipulated to incorporate different species, enabling exploration of various research questions. There are multiple challenges that arise when aggregating data from several sources, related to wide-ranging sampling methodologies and designs, error rates surrounding observations, seasonal availability of data, and more. Thus, a standardized procedure to develop the dataset helps reduce variance and allow future reproducibility and use by others. The creation of this dataset was driven by the goal of providing a unified and reproducible dataset that can fulfill diverse research purposes and supplement existing knowledge on Arctic marine mammal distributions.

3. Data Description

This article describes the synthesized marine mammal dataset of the linked repository containing occurrences for the Pacific walrus, bearded seal, and spotted seal from 1860–2024 (Fig. 1). Our integrated dataset is located within the NSF Arctic Data Center (https://doi.org/10.18739/A2WH2DH0V) and all related code is housed on GitHub (https://github.com/ArcticTelecouplingProject/ATP-MarineMammalSubModel). The dataset consists of data across the Bering, Chukchi, and Beaufort Seas and totals 36,438 occurrences. An occurrence represents an individual row in our dataset where at least one individual of a focal species was observed at a specific location and time. In total, there were 22,444 occurrences for walrus, 8586 for bearded seals, and 5408 for spotted seals. The dataset's columns and contents are detailed in Table 1.

We collected occurrence records from multiple public databases containing studies on the species of interest in this region (Table 2). First, we prioritized searching among the most well-known open access database repositories for biodiversity data such as the Global Biodiversity Information Facility (GBIF) and the Ocean Biodiversity Information System (OBIS). These databases also include international data, when available, which was essential to consider given the international nature of the Bering Strait's waters. No relevant publicly available data were found specifically regarding Russian waters.

We expanded our search to include U.S. government agencies such as the United States Geological Survey (USGS) and National Science Foundation (NSF). USGS has conducted several federal studies regarding the abundance and distribution of Pacific walruses and has made multiple data sources from these studies publicly available. NSF's Arctic Data Center is another primary

Table 1
Column headers and a description of their contents of the integrated dataset.

Column Name	Column Description
uniqueID	Unique identifier for an occurrence record on GBIF.org
scientificName	The taxonomic scientific name of the species
date	Date of the observation (yyyy-mm-dd)
decimalLat	The latitude coordinate of the observation location in decimal format
decimalLon	The longitude coordinate of the observation location in decimal format
count	Number of individuals of a species observed for an occurrence
methodObs	The method of data collection: - Acoustic monitoring survey - Aerial survey - Line transect visual survey - Preserved specimen - Research cruise visual observation - Seismic survey visual observation - Other
source	Unique identifier that abbreviates the title of the study or survey expedition name and year(s) of data collection
dataOrigin	The data repository where each source was downloaded from
sourceFile	Individual dataset file abbreviation

data and software repository for Arctic research whose main goal is housing data and information to promote reproducible research.

Scientific or common names were first used to identify datasets with relevant information for our three species of interest. A dataset was initially selected if it included a species of interest. Then, it was evaluated to determine whether the species data were within our study area. We added additional search terms regarding location (e.g., "Pacific Arctic", "Bering Sea", "Chukchi Sea", "Beaufort Sea", "Bering Strait") when necessary to narrow our search queries.

Regarding our searches on GBIF, we executed a search query on July 3, 2024, for each species with the following criteria:

- Basis of Record: Human observation, Machine observation, Preserved specimen
- Country or Area: Russian Federation, United States of America, Canada
- Scientific Name: *Odobenus rosmarus, Phoca largha*, or *Erignathus barbatus*

After determining that a dataset satisfied the previous criteria, it was downloaded and tidied using individual scripts respectively in the statistical software R, version 4.3.1. We created a standardized tidying procedure, defined in the methods, to restructure each dataset into a consistent format before combining. For example, we standardized the method used for data collection (e.g., methodObs) into seven categories based on the most common methods seen in each study's metadata (Fig. 2).

Many records originating from four datasets found on GBIF had their method of data collection categorized as 'Preserved Specimen'. These four datasets were the University of Alaska Museum's mammal collection (Arctos), the three GBIF species queries, and Moudrý & Devillers (2020). As defined by GBIF, when a record is classified as a 'Preserved Specimen' this indicates that the specimen has been preserved and catalogued and typically represents a specimen sample currently housed in a museum. Depending on the users' intent with the data, this type of presence record can be beneficial; therefore, these records can be kept or removed based on preferences for specimen versus observation records in the data. Caution should be taken when interpreting this categorization because of differences in publisher decisions when completing this field on GBIF. Other potential sources of errors in the records may include missing information or inaccuracies in dates (e.g., dates defaulting to the first of a month rather than on a specific day).

Table 2

Primary dataset sources and corresponding CSV root name of each tidied dataset included in the master dataset. https://doi.org/10.18739/A2WH2DH0V.

Author name(s)	Dataset title	Data repository	Dates of collection	DOI	CSV root name
Misc.	GBIF.org (03 July 2024) GBIF Occurrence Download [3]	GBIF	1860–2024	https://doi.org/10.15468/dl.9mq3hx	*gbif_queries_walrus*
Misc.	GBIF.org (03 July 2024) GBIF Occurrence Download [4]	GBIF	1900–2023	https://doi.org/10.15468/dl.87zdrf	*gbif_queries_spotted_seal*
Misc.	GBIF.org (03 July 2024) GBIF Occurrence Download [5]	GBIF	1900–2023	https://doi.org/10.15468/dl.vwx9bm	*gbif_queries_bearded_seal*
Nelson, T., Blees, M.	TGS Seismic Survey Visual Sightings in Chukchi Sea 2013 [6]	GBIF OBIS	2013–08–25 to 2013–10–30	10.15468/NNBSE9	*gbif_tgs_seismic_survey* *obis_tgs_seismic_survey*
Kuletz, K., Cushing, D., Labunski, E., Iken, K.	Marine bird sighting data, Chukchi Sea research on the vessel Norseman II from 2015- 08–09 to 2015–09–03 [7]	GBIF OBIS	2015–08–09 to 2015–09–03	10.15468/3IFPO0	*gbif_ambon_norseman_2015* *obis_ambon_norseman_2015*
Moore, S.	Marine Mammal Watch, Northern Bering Sea and Chukchi Sea, July 2016 [8]	NSF Arctic Data Center	2016–07–10 to 2016–07–20	10.18739/A27P8TD2J	*nsf_moore_mmw_2016*
Moore, S.	Marine Mammal Watch, Northern Bering Sea and Chukchi Sea, August-September 2017 [9]	NSF Arctic Data Center	2017–08–26 to 2017–09–14	10.18739/A25Q4RM2M	*nsf_moore_mmw_2017*
Stafford, K.	Marine mammal sightings during Sir Wilfrid Laurier cruise to the Pacific Arctic 11–22 July 2019 [10]	NSF Arctic Data Center	2019–07–11 to 2019–07–22	10.18739/A2DZ0320P	*nsf_dbo_laurier_2019*
Stafford, K.	Marine mammal sighting data from cruises in the Pacific Arctic, 2018, from Distributed Biological Observatory (DBO) regions [11]	NSF Arctic Data Center	2018–08–08 to 2018–08–23	10.18739/A2NV99B09	*nsf_dbo_laurier_2018*

Table 2 (*continued*)

Author name(s)	Dataset title	Data repository	Dates of collection	DOI	CSV root name
Ferguson, M.	Marine mammal aerial surveys in the Bering, Chukchi and Beaufort Seas 1979–2010 [12]	GBIF	1979–04–23 to 2010–10–25	*10.15468/UDS72J*	*gbif_nmml_aerial_1979_2010*
Jay, C.V., Quakenbush, L., Citta, J., Fischbach, A.S., Battaile, B.	Sex and age composition of walrus groups hauled out on ice floes in the Bering and Chukchi Seas, 2013–2015 [13]	USGS Alaska Science Center	2013–06–06 to 2015–06–22	*10.5066/F79K4894*	*usgs_sex_age_comp_2013_2015*
Battaile, B., Jay, C.V., Fischbach, A.S.	Behavior of Pacific walruses (*O. rosmarus divergens*) hauled out on sea ice during UAS overflights, Eastern Chukchi Sea, 2015 [14]	USGS Alaska Science Center	2015	*10.5066/F7DB7ZWP*	*usgs_walrus_behavior_uas_2015*
Moudrý, V., Devillers, R.	Quality and usability challenges of global marine biodiversity databases: An example for marine mammal data [15]	GBIF	2015, 2019	*10.1016/j.ecoinf.2020.101051*	*gbif_moudry_devillers*
BOEM, NOAA, North Slope Borough	Aerial Surveys of Arctic Marine Mammals (ASAMM) collected by Marine Mammal Laboratory, Bureau of Ocean Energy Management, and other agencies in the Bering, Chukchi, Beaufort Seas, and Amundsen Gulf from 1979 to 2019 and North Slope Borough, Alaska from 2020 to 2021 (NCEI Accession 0039,614) [16]	NOAA National Centers for Environmental Information	1979–04–21 to 2021–10–10	*10.7289/V51V5BZM*	*aerial_surveys_asamm*
Hobbs, R.	NMML Small Cetacean Aerial Survey 1999 [17]	GBIF OBIS-SEAMAP	1999–06–05 to 1999–07–04	*10.15468/N6QF5X*	*gbif_nmml_cetacean_survey_1999* *nmml_cetacean_survey_1999*
Barry, T., Helgason, H., McRae, L., Gill, M., Svoboda, M., Eamer, J.	Conservation of Arctic Flora and Fauna: Arctic Species Trend Index (ASTI): Marine [18]	GBIF	1950 to 2012	*10.15468/IL2VR5*	*gbif_asti_marine_walrus*

Table 2 (*continued*)

| Author name(s)	Dataset title	Data repository	Dates of collection	DOI	CSV root name
Link, O., Russell, L., Bloom, D., Wieczorek, J., McDonald, D., Gunderson, A., Kyndall, H. Waite, J.	UAM Mammal Collections (Arctos) [19]	GBIF	1900 to 2022	10.15468/LBIXOB	gbif_uam_collections_arctos
	NMML Bering Sea Cetacean Survey 2000 [20]	GBIF OBIS-SEAMAP	2000–06–10 to 2000–07–03	10.15468/4CJQ7F	gbif_nmml_bering_survey_2000 nmml_bering_survey_2000
Ray, G., Hufford, G., Krupnik, I., Labunski, E.	HLY-08–01 and HLY-08–02 Observations of marine mammals of Beringia [Ray, G.C.] [21]	NSF Arctic Data Center	2008–03–15 to 2008–05–04	10.5065/D6DV1GVN	nsf_mm_beringia_2008
Ray, G., Hufford, G., Krupnik, I., Labunski, E.	HLY-07–01 and HLY-07–02 Observations of Marine Mammals of Beringia [Ray, G.C.] [22]	NSF Arctic Data Center	2007–04–11 to 2007–06–18	10.5065/D67S7KS1	nsf_mm_beringia_2007
Holst, M., Lee, O., Smith, H.	Lamont-Doherty/LGL/NSF cruises [23]	GBIF OBIS-SEAMAP	2003–05–28 to 2009–03–07	10.15468/G27YV2	gbif_lamont_doherty obis_lamont_doherty
Stafford, K.	Acoustic detections of Arctic mammals in the western Beaufort Sea 2010–2011 [24]	GBIF OBIS-SEAMAP	2010–09–25 to 2011–08–29	10.15468/JRMEJE	gbif_acoustic_mams obis_acoustic_mams
Stafford, K., Moore, S., Iken, K., Turner, C., Benson, A., Biddle, M.	Vessel line-transect surveys of Arctic pinnipeds in the Chukchi Sea, Arctic Marine Biodiversity Observing Network (AMBON) research cruise on the vessel Norseman II, 2015-08–09 - 2015–09–03 [25]	GBIF USGS	2015–08–09 to 2015–09–03	10.15468/HCZOXZ	gbif_ambon_norseman_2015 gbif_ambon_norseman_2015
Stafford, K., Iken, K., Moore, S., Turner, C., Benson, A., Biddle, M.	Vessel line-transect surveys of Arctic pinnipeds in the Chukchi Sea, Arctic Marine Biodiversity Observing Network (AMBON) research cruise on the vessel Norseman II, 2017-08–05 to 2017–08–25 [26]	GBIF	2017–08–05 to 2017–08–25	10.15468/RVLBKL	gbif_ambon_norseman_2017
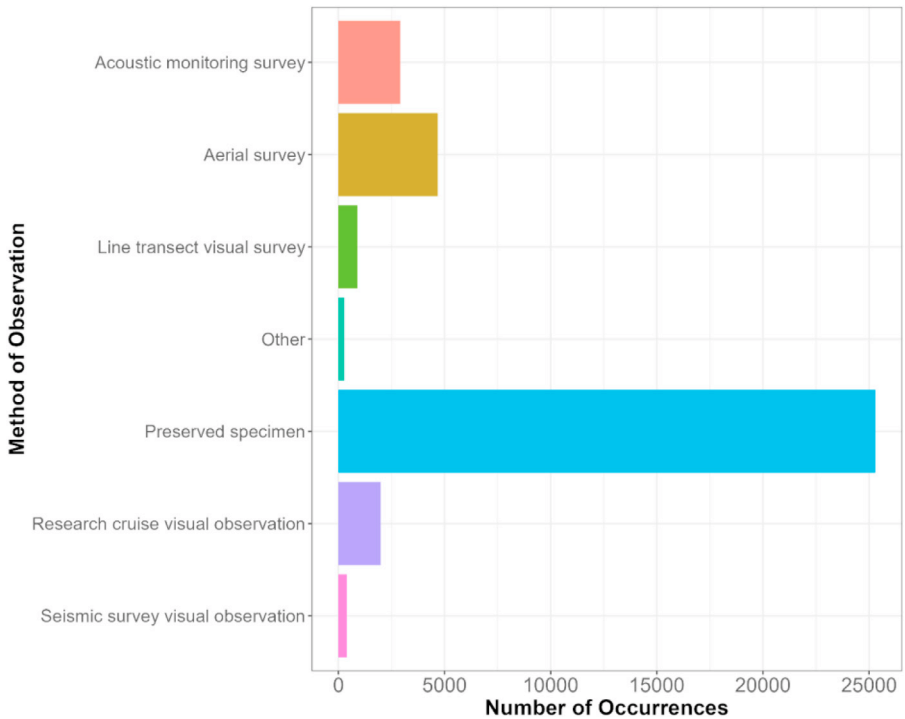
Fig. 2. Total occurrences collected within each method of observation category.

We chose to include the preserved specimens within this dataset because of the abundance of records ($n = 25,300$) and their temporal depth (i.e., since late 1800s). This could potentially provide additional insight into historical ranges of where these species were found and catalogued.

A spatial representation of Fig. 2 can be seen in Fig. 3, where the data are faceted by season and color coded by method of observation. This helps delineate the type of sampling performed during each season and show how the sampling effort is higher during the spring and summer months when there are less adverse weather conditions and sea ice that inhibit recording observations and executing surveys.

The 24 datasets that were included in our synthesized dataset had a large range of occurrences. The datasets that fit within our criteria with the lowest number of occurrences ($n = 2$) were the Marine Mammal Watch (SWL) 2018 and the Marine Mammal Watch Beringia Survey (Healy) 2007 datasets. The highest number of occurrences ($n = 12,534$) was contributed by the UAM Mammal Collections (Arctos) dataset. The breakdown of the number of occurrences by species of interest is represented in Fig. 4.

The number of studies documenting the occurrences of our three species varied considerably because of the wide array of purposes for collecting these data. Certain studies were species focused, collecting data only on their species of interest—for example, the *USGS Composition of Walrus Haulouts 2013–2015* focused on walrus, and the *Acoustic detections of Arctic mammals 2010–2011* focused on bearded seal. Pacific walruses had the highest amount of data available, making up 22,444 records in the synthesized dataset. Bearded seals were second with 8586 occurrences and spotted seals had the lowest amount with 5408 occurrences. The large difference in data available for walruses compared to the two seal species is also seen in Fig. 5, which totals the combined occurrences per month ($n =$). Similar to Fig. 3, there is less data available during the winter months when surveying is difficult due to hazardous weather and presence of
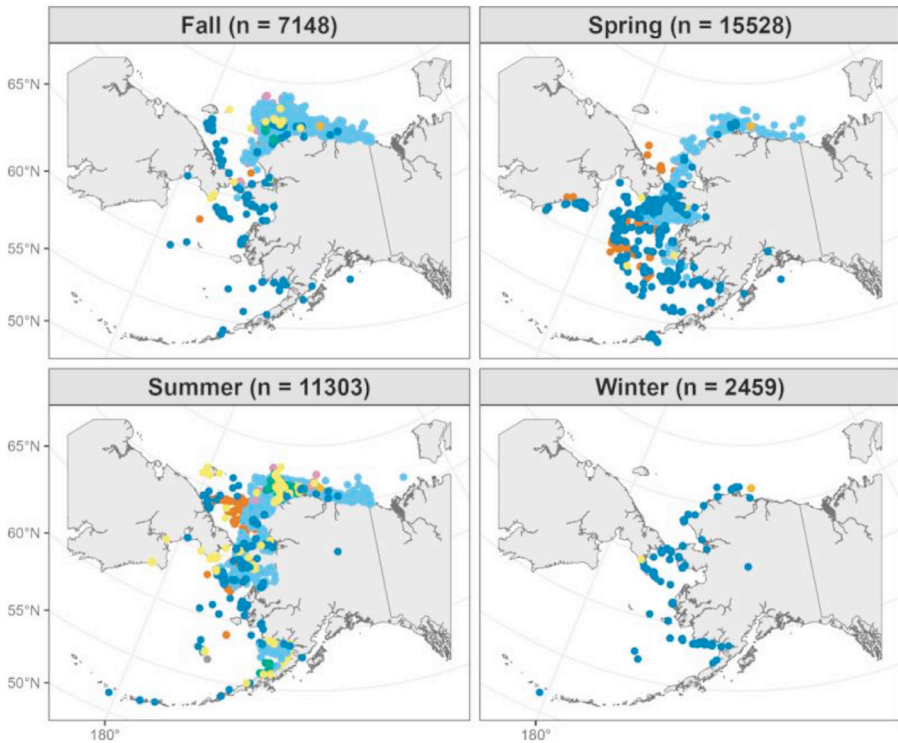
Fig. 3. Total occurrences for all three species faceted by season and color coded by method of data collection. Seasons: Fall (September-November), Winter (December-February), Spring (March-May), Summer (June-August).

sea ice. Most of the data observation methods relied on ship-board surveys or technology, therefore when sea ice impedes the flow of vessel traffic it also impedes the ability for surveying and data collection.

4. Experimental Design, Materials and Methods

4.1. Data collection and sources

Our comprehensive list of data sources that were used in the final dataset are detailed in Table 2.

4.2. Data preparation and manipulation

Datasets satisfying our criteria were downloaded and tidied using individual scripts in the statistical software R, version 4.3.1. All R scripts created and used in this study are publicly available through https://github.com/ArcticTelecouplingProject/ATP-MarineMammalSubModel.

An occurrence is a single recorded instance of at least one animal of interest observed at a specific time and location within the study area. If there were more than one animal observed at a specific time and location, this value is reported by the 'count' column. However, researchers chose to represent and record their data in different ways. In some studies, each individual animal—even if observed at the same time and location—was recorded as a separate
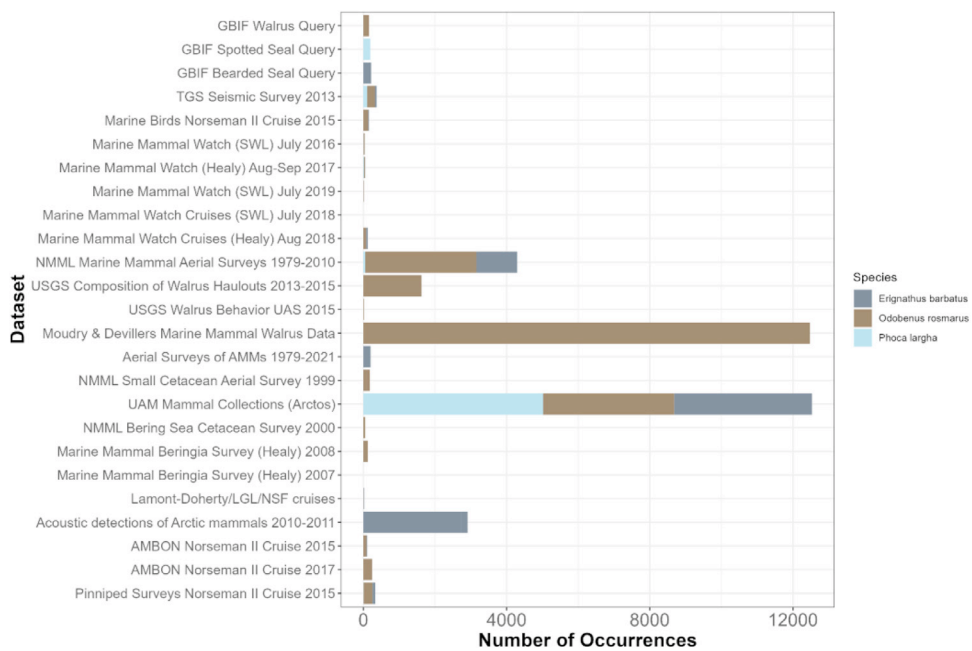
Fig. 4. Number of occurrences for each species of interest within each individual dataset.
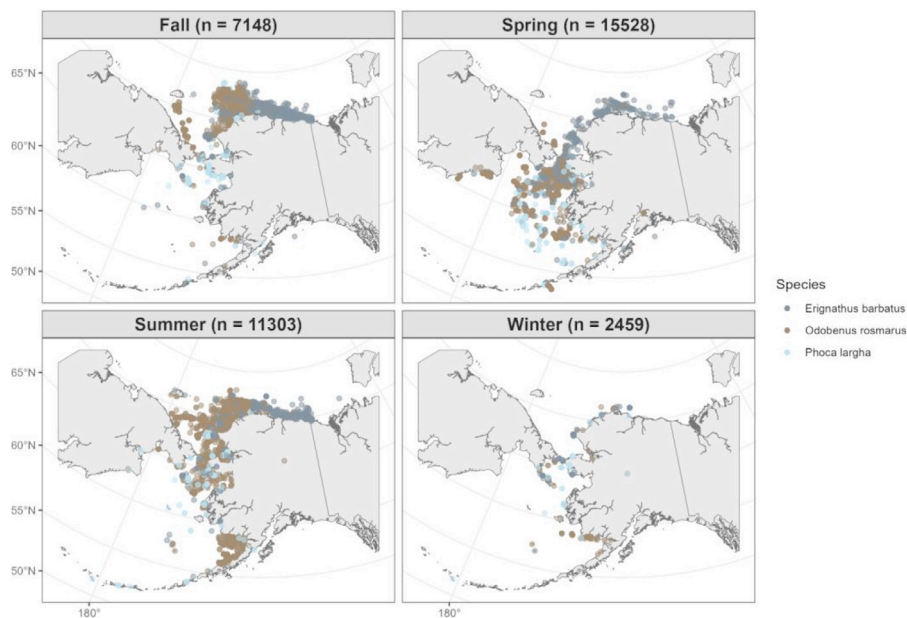

Fig. 5. Total occurrences (*n =*) mapped for each species of interest faceted by month and color coded by species. Seasons: Fall (September-November), Winter (December-February), Spring (March-May), Summer (June-August).

occurrence and no group size estimate or count was included. In other cases, groups of animals at a specific time and location were recorded under a group size or count column; therefore, by our "occurrence" definition this indicates one occurrence.

The following general tidying steps were applied to each dataset to ensure format consistency.

4.3. Remove occurrences with NA values in coordinates, species name, or date columns

- Filter to select species of interest (if other species exist in the data) by scientific or common name.
- Remove occurrences outside the study area bounds:
  - Study area bounding box: 172.18938 E, 46.08836 N, to 146.30217 W, 73.92759 N.
- Add additional columns and standardize contents as needed:
  - count: If there was no group size estimate recorded for an occurrence, it was marked NA.
  - methodObs: Assign a method that is most appropriate based on the process of data collection described in the study's metadata.
  - source: Brief title abbreviating the study name and year(s) of observation or data collection.
  - dataOrigin: Database or repository where the dataset was downloaded from.
- Reformat the date of observation columns into year, month, and day columns.
- Standardize naming and order for columns of interest: gbifID, scientificName, year, month, day, decimalLat, decimalLon, count, methodObs, source, dataOrigin, sourceFile.
- Eliminate duplicate occurrences.
- Rearrange the data chronologically by earliest date.
- Save as CSV.

Once each individual dataset was tidied, they were combined to build the overall synthesized dataset. Additional tidying steps were taken during the joining process:

- Create sourceFile column and standardize cell contents using the same naming convention as the individual tidy scripts (e.g., 04_nsf_moore_mmw_2016.R becomes nsf_moore_mmw_2016).
- Ensure records with missing information in the year, month, day, decimalLat, or decimalLon fields are eliminated.
- Remove duplicate records using unique IDs (either gbifID or a custom ID created with the uuid package in R).

4.4. Standardization decisions

Discrepancies in the year, month, and day format amongst datasets were resolved by formatting into columns for a 4-digit year, 2-digit month, and 2-digit day and creating a date column (i.e., yyyy, mm, dd) with these contents in the final dataset.

We discovered that there were several datasets cross-listed on multiple databases (e.g., the GBIF search queries and other datasets found on GBIF, OBIS, etc.). Each GBIF dataset occurrence is associated with a unique gbifID that corresponds with the original record and study on GBIF for traceability. Additionally, each GBIF search query returns multiple datasets that meet the criteria for the query and provides links to the original dataset pages on GBIF. These search query studies were cross-referenced with the datasets that we pulled from other sources. If there was a dataset found on another database and part of the GBIF query's list of included datasets, then the dataset was downloaded from its original GBIF page to correspond the gbifID with its original study name for traceability and remove duplicates using these unique identifiers. If two datasets from GBIF were found to be duplicated, the gbifID was used to keep one set of these

occurrences. If a dataset was downloaded from a different database and not cross-listed on GBIF or another database, then a custom ID was created using the uuid package in R to maintain consistent column formatting (e.g., custom_id_abc123).

Limitations

Despite its comprehensiveness, there are several limitations to consider when using these data to answer questions related to species distributions, or other relevant topics. The methods of collection were frequently opportunistic observations and therefore are classified as presence-only data. For the purposes of this synthesis, we excluded information on absence data. However, many aerial and ship-based surveys do contain absence data and original sources can be referred to if needed for specific research goals. We acknowledge the limitations that excluding absence data can preclude certain statistical methods, such as abundance estimates, from being applied to these data. Nevertheless, there are many alternative methods that have proven to be similarly robust with comparative predictive power as presence-absence modelling methods, such as maximum entropy, regression-based models, and ecological niche factor analysis [27]. Lastly, our decision for standardizing the count column may lead to potential bias.

Ethics Statement

All authors have read and follow the ethical requirements for publication in Data in Brief. This work did not involve any human subjects, animal experiments, or data collected from social media.

CRediT author statement

Sydney Waloven: Data curation, Writing, Original draft preparation, Reviewing and Editing. **Natalia Portales:** Data curation, Writing, Original draft preparation. Reviewing and Editing. **Kelly Kapsar:** Conceptualization, Supervision, Writing – Reviewing and Editing. **Jianguo Liu:** Supervision, Writing – Reviewing and Editing.

Data Availability

A synthesis of Pacific Arctic marine mammal occurrences (1860–2024) (Original data) (NSF Arctic Data Center).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] R. Gryba, H.P. Huntington, A.L. Von Duyke, B. Adams, B. Frantz, J. Gatten, Q. Harcharek, H. Olemaun, R. Sarren, J. Skin, G. Henry, M. Auger-Méthé, Indigenous knowledge of bearded seal (*Erignathus barbatus*), ringed seal (*Pusa hispida*), and spotted seal (*Phoca largha*) behaviour and habitat use near Utqiaġvik, Alaska, USA, Arct. Sci. 7 (2021) 832–858, doi:10.1139/as-2020-0052.

[2] B.C. Battaile, C.V. Jay, M.S. Udevitz, A.S. Fischbach, Evaluation of a method using survey counts and tag data to estimate the number of Pacific walruses (*Odobenus rosmarus divergens*) using a coastal haulout in northwestern Alaska, Polar Biol. 40 (2017) 1359–1369, doi:10.1007/s00300-016-2060-5.

[3] GBIF.Org User, Pacific walrus occurrence download, (2024) 636487. https://doi.org/10.15468/DL.9MQ3HX.

[4] GBIF.Org User, Spotted seal occurrence download, (2024) 179035. https://doi.org/10.15468/DL.87ZDRF.

[5] GBIF.Org User, Bearded seal occurrence download, (2024) 296252. https://doi.org/10.15468/DL.VWX9BM.

[6] T. Nelson, M. Blees, TGS seismic survey visual sightings in Chukchi sea 2013, (2021). https://doi.org/10.15468/NNBSE9.

[7] K. Kuletz, D. Cushing, E. Labunski, K. Iken, Marine bird sighting data, Arctic marine biodiversity observing network (AMBON) Chukchi sea research cruise on the vessel Norseman II from 2015 to 08-09 to 2015-09-03, (2023). https://doi.org/10.15468/3IFPO0.

[8] S. Moore, Marine mammal watch, northern Bering sea and Chukchi sea, July 2016, (2019). https://doi.org/10.18739/A27P8TD2J.

[9] S. Moore, Marine mammal watch, northern Bering sea and Chukchi sea, August-September 2017, (2019). https://doi.org/10.18739/A25Q4RM2M.

[10] K. Stafford, Marine mammal sightings during Sir Wilfrid Laurier cruise to the Pacific Arctic 11-22 July 2019, (2019). https://doi.org/10.18739/A2DZ0320P.

[11] K. Stafford, Marine mammal sighting data from cruises in the Pacific Arctic, 2018, from distributed biological observatory (DBO) regions, (2019). https://doi.org/10.18739/A2NV99B09.

[12] M. Ferguson, Marine mammal aerial surveys in the Bering, Chukchi and Beaufort seas 1979-2010, (2021). https://doi.org/10.15468/UDS72J.

[13] C.V. Jay, L. Quakenbush, J. Citta, A.S. Fischbach, B. Battaile, Sex and age composition of walrus groups hauled out on ice floes in the Bering and Chukchi seas, 2013-2015, (2020). https://doi.org/10.5066/F79K4894.

[14] B. Battaile, C.V. Jay, A.S. Fischbach, Behavior of Pacific walruses (*Odobenus rosmarus divergens*) hauled out on sea ice during UAS overflights, eastern Chukchi sea, 2015, (2022). https://doi.org/10.5066/F7DB7ZWP.

[15] V. Moudrý, R. Devillers, Quality and usability challenges of global marine biodiversity databases: an example for marine mammal data, Ecol. Inform. 56 (2020) 101051, doi:10.1016/j.ecoinf.2020.101051.

[16] A.O.C.S. (OCS) R. BOEM, N.M.F.S. NOAA, North slope borough, aerial surveys of Arctic marine mammals (ASAMM) collected by marine mammal laboratory, bureau of ocean energy management, and other agencies in the Bering, Chukchi, Beaufort seas, and Amundsen gulf from 1979 to 2019 and north slope borough, Alaska from 2020 to 2021 (NCEI Accession 0039614), (2008). https://doi.org/10.7289/V51V5BZM.

[17] R. Hobbs, NMML small cetacean aerial survey 1999, (2021). https://doi.org/10.15468/N6QF5X.

[18] T. Barry, Arctic species trend index (ASTI) : marine, (2019). https://doi.org/10.15468/IL2VR5.

[19] Link Olson, UAM mammal collection (Arctos), (2024). https://doi.org/10.15468/LBIXOB.

[20] J. Waite, NMML Bering sea cetacean survey 2000, (2021). https://doi.org/10.15468/4CJQ7F.

[21] G. Ray, G. Hufford, I. Krupnik, E. Labunski, HLY-08-01 and HLY-08-02 observations of marine mammals of Beringia. Version 1.0, (2009) 7 data files, 1 ancillary/documentation file, 1218 KiB. https://doi.org/10.5065/D6DV1GVN

[22] G. Ray, G. Hufford, I. Krupnik, E. Labunski, HLY-07-01 and HLY-07-02 observations of marine mammals of Beringia. Version 1.0, (2009) 6 data files, 1 ancillary/documentation file, 1676 KiB. https://doi.org/10.5065/D67S7KS1

[23] M. Holst, O. Lee, H. Smith, Lamont-Doherty/LGL/NSF cruises, (2021). https://doi.org/10.15468/G27YV2.

[24] K. Stafford, Acoustic detections of Arctic mammals in the western Beaufort sea 2010-2011, (2021). https://doi.org/10.15468/JRMEJE.

[25] K. Stafford, S. Moore, K. Iken, Vessel line-transect surveys of Arctic pinnipeds in the Chukchi sea, Arctic marine biodiversity observing network (AMBON) research cruise on the vessel Norseman II, 2015-08-09 - 2015-09-03, (2023). https://doi.org/10.15468/HCZOXZ.

[26] K. Stafford, K. Iken, S. Moore, Vessel line-transect surveys of Arctic marine mammals in the Chuckchi sea, Arctic marine biodiversity observing network (AMBON) research cruise on the vessel Norseman II from 2017 to 08-05 to - 2017-08-25, (2023). https://doi.org/10.15468/RVLBKL.

[27] J. Elith, C.H. Graham, R.P. Anderson, M. Dudík, S. Ferrier, A. Guisan, R.J. Hijmans, F. Huettmann, J.R. Leathwick, A. Lehmann, J. Li, L.G. Lohmann, B.A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J. McC. M. Overton, A. Townsend Peterson, S.J. Phillips, K. Richardson, R. Scachetti-Pereira, R.E. Schapire, J. Soberón, S. Williams, M.S. Wisz, N.E. Zimmermann, Novel methods improve prediction of species' distributions from occurrence data, Ecography 29 (2006) 129–151, doi:10.1111/j.2006.0906-7590.04596.x.