Spatial Data, Analysis and Modelling Forums: An initiative to broaden the collaborative research potential at DFO

Catalina Gomez, Jessica Nephin, Shelly Lang, Laura Feyrer,
Freya Keyser, Gordana Lazin

Science Branch
Fisheries and Oceans Canada
Maritimes Region
Dartmouth, Nova Scotia, B2Y 4A2, Canada

2021

Canadian Technical Report of
Fisheries and Aquatic Sciences 3416

Fisheries and Oceans Canada Pêches et Océans Canada

Canada

Canadian Technical Report of Fisheries and Aquatic Sciences

Technical reports contain scientific and technical information that contributes to existing knowledge but which is not normally appropriate for primary literature. Technical reports are directed primarily toward a worldwide audience and have an international distribution. No restriction is placed on subject matter and the series reflects the broad interests and policies of Fisheries and Oceans Canada, namely, fisheries and aquatic sciences.

Technical reports may be cited as full publications. The correct citation appears above the abstract of each report. Each report is abstracted in the data base *Aquatic Sciences and Fisheries Abstracts*.

Technical reports are produced regionally but are numbered nationally. Requests for individual reports will be filled by the issuing establishment listed on the front cover and title page.

Numbers 1-456 in this series were issued as Technical Reports of the Fisheries Research Board of Canada. Numbers 457-714 were issued as Department of the Environment, Fisheries and Marine Service, Research and Development Directorate Technical Reports. Numbers 715-924 were issued as Department of Fisheries and Environment, Fisheries and Marine Service Technical Reports. The current series name was changed with report number 925.

Rapport technique canadien des sciences halieutiques et aquatiques

Les rapports techniques contiennent des renseignements scientifiques et techniques qui constituent une contribution aux connaissances actuelles, mais qui ne sont pas normalement appropriés pour la publication dans un journal scientifique. Les rapports techniques sont destinés essentiellement à un public international et ils sont distribués à cet échelon. II n'y a aucune restriction quant au sujet; de fait, la série reflète la vaste gamme des intérêts et des politiques de Pêches et Océans Canada, c'est-à-dire les sciences halieutiques et aquatiques.

Les rapports techniques peuvent être cités comme des publications à part entière. Le titre exact figure au-dessus du résumé de chaque rapport. Les rapports techniques sont résumés dans la base de données *Résumés des sciences aquatiques et halieutiques.*

Les rapports techniques sont produits à l'échelon régional, mais numérotés à l'échelon national. Les demandes de rapports seront satisfaites par l'établissement auteur dont le nom figure sur la couverture et la page du titre.

Les numéros 1 à 456 de cette série ont été publiés à titre de Rapports techniques de l'Office des recherches sur les pêcheries du Canada. Les numéros 457 à 714 sont parus à titre de Rapports techniques de la Direction générale de la recherche et du développement, Service des pêches et de la mer, ministère de l'Environnement. Les numéros 715 à 924 ont été publiés à titre de Rapports techniques du Service des pêches et de la mer, ministère des Pêches et de l'Environnement. Le nom actuel de la série a été établi lors de la parution du numéro 925.

Canadian Technical Report of Fisheries and Aquatic Sciences 3416

2021

SPATIAL DATA, ANALYSIS AND MODELLING FORUMS: AN INITIATIVE TO BROADEN THE COLLABORATIVE RESEARCH POTENTIAL AT DFO

by

Catalina Gomez[1], Jessica Nephin[2], Shelley Lang[1], Laura Feyrer[1], Freya Keyser[1],
Gordana Lazin[1]

[1]Bedford Institute of Oceanography
Fisheries and Oceans Canada, 1 Challenger Drive PO Box 1006
Dartmouth, Nova Scotia, B2Y 4A2, Canada

[2]Institute of Ocean Sciences
Fisheries and Oceans Canada, 9860 West Saanich Road
Sidney, British Columbia, V8L 4B2, Canada

Cat. No. Fs97-6/3416E-PDF ISBN 978-0-660-37261-7 ISSN 1488-5379

Correct citation for this publication:

Gomez, C. Nephin, J., Lang, S., Feyrer, L., Keyser, F, Lazin, G. 2021. Spatial Data, Analysis and Modelling Forums: An initiative to broaden the collaborative research potential at DFO. Can. Tech. Rep. Aquat. Sci. 3416: v + 36 p.

TABLE OF CONTENTS

ABSTRACT

Gomez, C. Nephin, J., Lang, S., Feyrer, L., Keyser, F, Lazin, G. 2021. Spatial Data, Analysis and Modelling Forums: An initiative to broaden the collaborative research potential at DFO. Can. Tech. Rep. Aquat. Sci. 3416: v + 36 p.

Science requires open, reproducible, and collaborative approaches to maximize efficiency and deliver improved outcomes. These proceedings summarize outcomes of the Spatial Data, Analysis and Modelling Forums organized as part of the R Learning and Development series by the Fisheries and Ocean Canada (DFO) Science Sector (Maritimes and Pacific Regions) in 2020. These proceedings provide an overall summary of this learning and development series, describing the materials, presentations, questions, and discussions. The main intent of this initiative was to provide a forum for DFO staff to present their ongoing work and issues in relation to spatial data, analysis and modelling. A secondary goal was to learn how various programs and regions were resolving these issues and to build a common understanding of each other’s perspectives. This initiative was also conceived with the intent to foster collaborations, by helping DFO staff connect with colleagues that have shared challenges or interests. The information gleaned from discussions and participant surveys were used to guide and support future learning opportunities such as statistical modelling and programming language training. It is our hope that the information gathered from these learning and development opportunities can support already established working groups and task forces currently tackling data discovery and management challenges at DFO. Based on the presentations and discussions that followed during these events we present recommendations for increasing reproducibility and institutional efficiency in spatial analyses and modelling efforts.

RÉSUMÉ

Gomez, C. Nephin, J., Lang, S., Feyrer, L., Keyser, F, Lazin, G. 2021. Spatial Data, Analysis and Modelling Forums: An initiative to broaden the collaborative research potential at DFO. Can. Tech. Rep. Aquat. Sci. 3416: v + 36 p.

La science nécessite des approches ouvertes, reproductibles et collaboratives pour maximiser l'efficience et fournir de meilleurs résultats. Le présent compte rendu résume les résultats des forums sur l'analyse et la modélisation des données spatiales organisés dans le cadre de la série d'apprentissage et de perfectionnement sur le langage R offerte par le Secteur des sciences (régions des Maritimes et du Pacifique) de Pêches et Océans Canada (MPO) en 2020. Le présent compte rendu est un résumé général de cette série d'apprentissage et de perfectionnement; on y décrit notamment le matériel, les présentations, les questions et les discussions. L'objectif principal de cette initiative était de fournir une tribune au personnel du MPO afin qu'il puisse faire part de ses travaux et des problèmes liés aux données, aux analyses et à la modélisation spatiales, et aux outils connexes. Un deuxième objectif était d'apprendre comment les différents programmes et les différentes régions résolvaient ces problèmes et d'établir une compréhension commune des perspectives de chacun. L'initiative visait également à favoriser les collaborations en aidant le personnel du MPO à entrer en contact avec des collègues qui ont des défis ou des intérêts communs. Les informations recueillies dans le cadre des discussions et des sondages menés auprès des participants ont été utilisées pour orienter et appuyer les futures occasions d'apprentissage, telles qu'une formation sur le langage de programmation et de modélisation statistique. Nous espérons que les renseignements obtenus grâce à ces occasions d'apprentissage et de perfectionnement pourront aussi soutenir les groupes de travail et les équipes spéciales déjà établis qui s'attaquent actuellement aux défis liés à la découverte et la gestion des données au MPO. À la lumière des présentations et des discussions qui ont eu lieu lors de cet événement, nous présentons des recommandations pour accroître la reproductibilité et l'efficience organisationnelle des activités d'analyse et de modélisation spatiales.

INTRODUCTION

The Science Branch of Fisheries and Oceans Canada (DFO) is diverse and productive; however, a lack of systems that highlight and integrate research activities in the institution limits its broader collaborative research potential. This, in part, is because dissemination is traditionally limited to peer-reviewed publications, coordination between groups is challenging, and there has been a paucity of centralized research/data inventories. This is changing with the adoption of applications like Microsoft Teams that support networking and open collaboration in the Science Branch at DFO and the establishment of a R coding Learning and Development initiative supported by the Science Executive Committee (SEC). This initiative coordinates lunch series, forums, workshops, and task forces to share reproducible tools for a more open, cost-effective, and efficient approach for science execution, communication and advice. These initiatives bring researchers together, with a focus on multi-disciplinary meetings with clear objectives and hands-on coding workshops to support the strategic vision and priorities of DFO as well as individual research interests. These opportunities have fostered growth, community, and training – all qualities supported and encouraged within the organization.

DFO Science staff are spread across seven regions: Newfoundland and Labrador, Gulf, Maritimes, Quebec, Ontario and Prairie, Arctic, and Pacific. Many researchers working within these regional silos rely heavily on the same data types of spatial data and analyses to provide science advice and to support national initiatives like Marine Conservation Targets, Marine Spatial Planning and Planning for Integrated Environmental Response. Building connections among science staff working on common objectives and employing similar methods across and within regions can reduce duplication of effort and increase efficiency via knowledge sharing.

One of the goals of this initiative is to broaden the collaborative research potential at DFO in the field of spatial data, analysis and modelling. Researchers in Pacific and Maritimes Region coordinated two national forums to support this goal:

1. Species Distribution Modelling Forum: Consolidating an East/West Coast Connection at DFO (February 10 2020);
2. Access to Spatial Data to Support Analysis and Modelling (November 23 2020 2020).

The first forum was held at the Bedford Institute of Oceanography (BIO) and WebEx, and included participants from the DFO Pacific, Maritimes, and Newfoundland and Labrador Regions interested in Species Distribution Modelling. The second virtual forum was held in Microsoft Teams and included participants from the DFO Pacific, Maritimes, Quebec, Newfoundland and Labrador, Gulf, National Capital Region (NCR), Ontario and Prairie, and Arctic Regions.

These events enabled Science staff to:

- Exchange information about strategic planning and workflows to effectively organize data, code, and tools to support spatial analysis and modelling,
- Gain insight on how different data products could be, or have already been, used to inform spatial analysis and modelling,
- Share perspectives on the relevant considerations for using spatial data products and datasets, or the predictors in general, and any of the limitations users should be aware of,
- Learn about new spatial data products and existing platforms for sharing spatial data,
- Discuss the limitations of the available data products in terms of extent, resolution, quality and underlying assumptions.

The intent of these proceedings is to document the discussion and material from these series as well as the projects and initiatives related to spatial data, analysis and modelling being led in the department. This document summarizes information primarily for DFO internal use. However, we hope that these efforts will facilitate future research collaboration with researchers outside of DFO.

I: SPECIES DISTRIBUTION MODELLING FORUM: CONSOLIDATING AN EAST/WEST COAST CONNECTION AT DFO

By Jessica Nephin, Jessica Finney, Tana Worcester and Catalina Gomez

A Species Distribution Modelling Forum was organized with the goal of exchanging technical information on the type of approaches and data used to model species distributions in the Atlantic and Pacific coasts, and a discussion of a path-forward to establish and strengthen a network of researchers working on species distribution modelling, Table 1).

The first part of the forum consisted of presentations by Tana Worcester and Jessica Finney to summarize the context and background of this initiative. This introduction was followed by a series of speed talks by DFO staff in Maritimes and Pacific Regions (Tables 2 and 3). Each speed talk presenter was allotted 5 minutes to share their modeling approach, sources and quality of data, methods, and results. The forum presented a diverse mix of presentations – all available at this link. The speed talk presentations highlighted the diversity of approaches used to represent the spatial distribution of species and the many different data sources, spatio-temporal scales, environmental predictors, methods, and tools available. They also highlighted many commonalities, for example, the majority of researchers are using the R programming language to perform spatial modelling.

The objective of the majority of projects presented during the speed talks were prediction (8 projects across space only and 6 projects across both space and time dimensions). The goal of the remaining 4 projects was temporal projection to future climate scenarios. To achieve these objectives, at least 8 different modelling methods were used (Figure 1). Generalized additive models (GAMs) and maximum entropy (MaxEnt) models were the most popular, followed by generalized linear mixed models (GLMMs). Within the GLMMs methods there was a diversity of model building approaches including Template Model Builder (TMB), integrated nested Laplace approximation (INLA), and Markov chain Monte Carlo (MCMC) methods.

The response variables being modelled also varied across projects. The majority of projects were predicting the probability of occurrence (11 projects), however, abundance (3), biomass (2), species richness (1) and substrate (1) were represented as well. With the exception of one project, species responses were modelled individually, not as a community. The mean number of species modelled per project was 20, ranging from 113 species to 1 species.

Table 1. Agenda of the Species Distribution Modelling (SDM) Forum – Monday, February 10 2020.

Time (AST)	Item
12:30 - 12:35 m	Introduction to the SDM Forum
12:35 - 12:45 m	Maritimes Region Context: Tana Worcester
12:45 - 1:00 pm	Pacific Region Context: Jessica Finney
1:00 - 2:00 pm	Speed-talks by participants from Atlantic Region working on different questions and approaches to SDM
2:00 - 3:00 pm	Speed-talks by participants from Pacific Region working on different questions and approaches to SDM
3:00 - 3:10 pm	Break
3:10 - 3:20 pm	Northeast U.S. Regional Marine Fish Habitat Assessment: An integrated approach to understanding fish habitat use
3:20 - 5:00 pm	Discussion on the benefits of using SDM, and general best practices that are applied across the many different groups working in this realm. Discussion on how to strengthen an east/west coast connection of researchers working in this field.

Table 2. Speed talks by participants in Maritimes Region working on spatial approaches to describe species distribution on Monday, February 10 2020. Presentations are all available at this link.

Time (AST)	Items	Presenters
	SDM and stock assessment	
1:05 - 1:10 pm	Using Gaussian Random Fields to model spatiotemporal variability of groundfish on Georges Bank	Dave Keith
	Zooplankton	
1:10 - 1:15 pm	Estimating spatial patterns of *Calanus* abundance in the northwest Atlantic with a coupled bio-physical model	Catherine Brennan

	SDM and Species at Risk	
1:15 - 1:20 pm	SDM for North Atlantic right whales off eastern Canada based on opportunistic sightings and directed surveys of known aggregations	Shelley Lang
	Seabirds	
1:20 - 1:25 pm	Use of at-sea surveys and predictive spatial models to estimate seasonal densities of seabirds in the Atlantic	Sarah Wong (ECCC)
	Climate vulnerability	
1:25 - 1:30 pm	SDMs to predict climate change impacts on the distribution of the habitat-forming glass sponge *Vazella pourtalesii*	Lindsay Beazley
1:30 - 1:35 pm	A Lobster Story: Measuring potential changes in habitat suitability using ocean climate model	Kiyomi Ferguson
	SDM and Benthic Communities	
1:35 - 1:40 pm	Modelling of benthic communities using Joint Species Distribution Models	Javier Murillo
1:40 - 2:00 pm	Q&A	

Table 3. Speed talks by participants in Pacific Region working on spatial approaches to describe species distribution on Monday, February 10 2020. Presentations are all available at this link.

Time (AST)	**Items**	**Presenter**
	Benthic species	
2:00 - 2:05 pm	Habitat suitability index models for data limited species	Candice St. Germain
2:05 - 2:10 pm	Overview of Pacific SDM Framework and its application to several benthic species in BC	Jessica Nephin
2:10 - 2:15 pm	VMEs: Deep-sea coral and sponge SDM	Jessica Nephin (for Jackson Chu)
	Fish	
2:15 - 2:20 pm	Herring spawn SDM	Chris Rooper

2:20 - 2:25 pm	Sand lance SDM	Cliff Robinson
2:25 - 2:30 pm	Geostatistical models predicting biomass from trawl and long line surveys	Sean Anderson
	Climate shifts	
2:30 - 2:35 pm	Spatiotemporal modelling of BC groundfish ranges and their response to climate	Philina English
2:35 - 2:40 pm	Bayesian SDM to estimate climate vulnerability of groundfish	Karen Hunter
2:40 - 2:45 pm	Modelling shifts in invasive species	Devin Lyons
	Marine Mammals	
2:45 - 2:50 pm	Spatial Modelling of Marine Mammals in British Columbia: A Distance Sampling Approach	Brianna Wright
	Modelling environmental data	
2:50 - 2:55 pm	Using Random Forests to model substrate type	Dana Haggarty
2:55 - 3:00 pm	Q&A	

The mean number of predictor variables used was 11. The number of predictors was highly variable between projects ranging from 2 to 39 predictors. The most commonly used predictors were bathymetric (e.g. depth, slope, complexity) and physical (e.g. current speed, temperature, salinity) variables (Figure 2). Models were less likely to include chemical (e.g. oxygen), other species (e.g. copepod distribution) and human use (e.g. fishing pressure) predictor variables. Some presenters indicated a desire to include additional environmental variables but had concerns over data quality, collinearity with other variables and spatial and temporal resolutions that were not appropriate for their objectives.

Following DFO presentations, the Northeast U.S. Regional Habitat Assessment team provided an overview of an integrated approach to understanding fish habitat use. Victoria Kentner presented this information on behalf of Michelle Bachman (New England Fishery Management Council (NEMFC)), Jessica Coakley (Mid-Atlantic Fishery Management Council), Chris Haak (Monmouth University/National Oceanic and Atmospheric Administration (NOAA) National Marine Fisheries Service (NMFS)), and Laurel Smith (NMFS). Four actions have been identified in their work plan as necessary to describe and characterize estuarine, coastal, and offshore fish habitat distribution, abundance, and quality in the Northeast. These actions will address: 1) Abundance and trends in habitat types in the inshore area, 2) Habitat vulnerability, 3) Spatial descriptions of species habitat use in the offshore area and 4) provide a Habitat Data Visualization and Decision Support Tool.

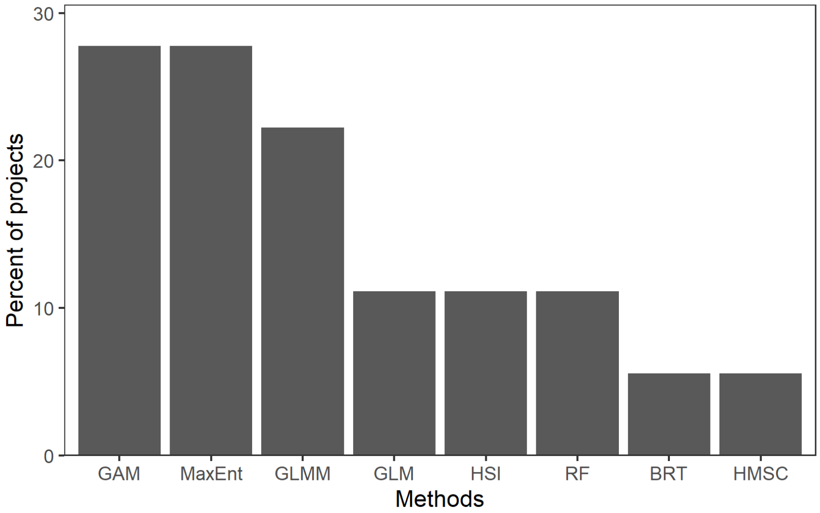


Figure 1. Methods used for species distribution modelling projects presented in the speed talk sessions. Some projects employed more than one method. GAM = generalized additive model, MaxEnt = maximum entropy, GLMM = generalized linear mixed model (built with integrated nested Laplace approximation (INLA), template model builder (TMB) or Markov chain Monte Carlo (MCMC) methods), GLM = generalized linear model, HSI = habitat suitability model (build from expert or literature derived thresholds), RF = random forest, BRT = boosted regression trees, and HMSC = hierarchical modelling of species communities.

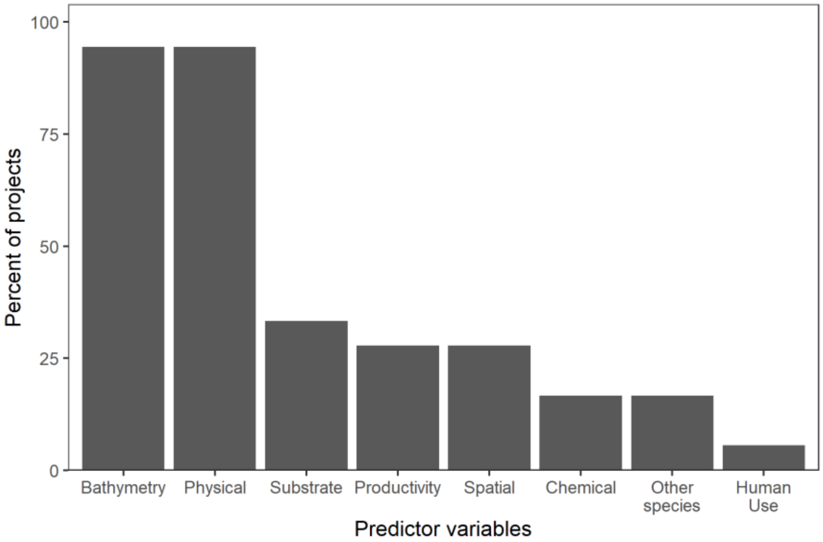


Figure 2. Predictors variable included in species distribution modelling projects presented in the speed talk sessions. Variables were grouped into eight categories: bathymetry (e.g. depth, slope, complexity), physical (e.g. current speed, temperature, exposure), substrate (e.g grain size, rock), productivity (e.g. chlorophyll-a), spatial (e.g. spatial random field), chemical (e.g. oxygen, silica), other species (e.g. distribution of copepod species), human use (e.g. fishing pressure).

Presentations in this forum were followed by a discussion including all participants on the following topics:

- The benefits and drawbacks of using consensus in approaches for spatial modelling, and the feasibility of developing comparable methods to explore shifts in species distribution
- The benefits of streamlining data requests to efficiently access and process environmental predictors (working smarter, not harder)
- Strengthening partnerships with oceanographers, Natural Resources Canada (NRCan) staff, and other stakeholders to improve the quality of environmental predictors
- Opportunities for advancing technical skills to better support species distribution modelling projects in different regions

It was noted that the quality of environmental predictor layers is important and can depend on resolution, extrapolation, and uncertainty of the source data. An identified gap was that predictor layers themselves are not typically being validated (including comparison between difference data sources).

There was a suggestion to include fishing pressure as a predictor layer in species distribution models and the need to coordinate on this approach so that the method for aggregating/smoothing fishing pressure data is consistent and comparable across projects.

There was a question about how to deal with spatial bias in the survey data (e.g. sampling within specific habitats for stock assessment). There is a need to come up with a process/recommendation for how to deal with this bias.

There was a comment about projections under climate change and how it is especially important for these models to incorporate ecological knowledge. For example, if there are range shifts predicted, will the species have time to colonize these new ranges given life history traits such as the time required for reproduction and recruitment.

Participant survey results

Participants in the forum were asked to complete an evaluation survey at the end of the forum. Twenty eight responses were received from the 89 attendees (40% response rate). Responses were received from DFO, NOAA and NEFMC participants based in the Canadian Pacific, Maritimes, and Newfoundland and Labrador Regions, as well as the US Northeast (Figure 3). Participants worked in a wide range of research areas (Figure 4).

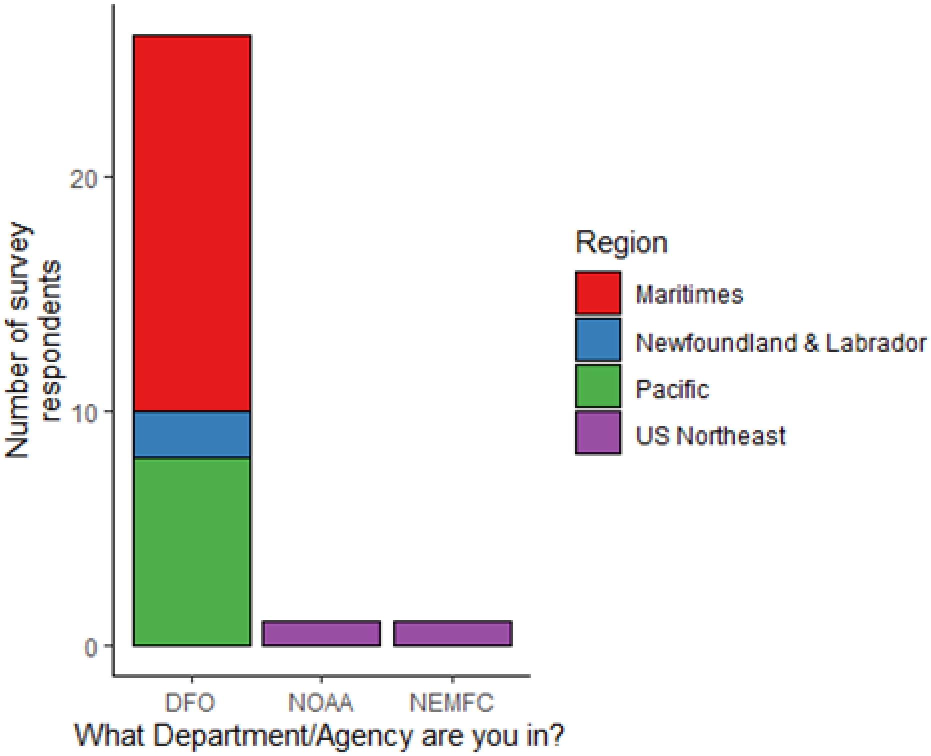


Figure 3. Survey responses by department and region.

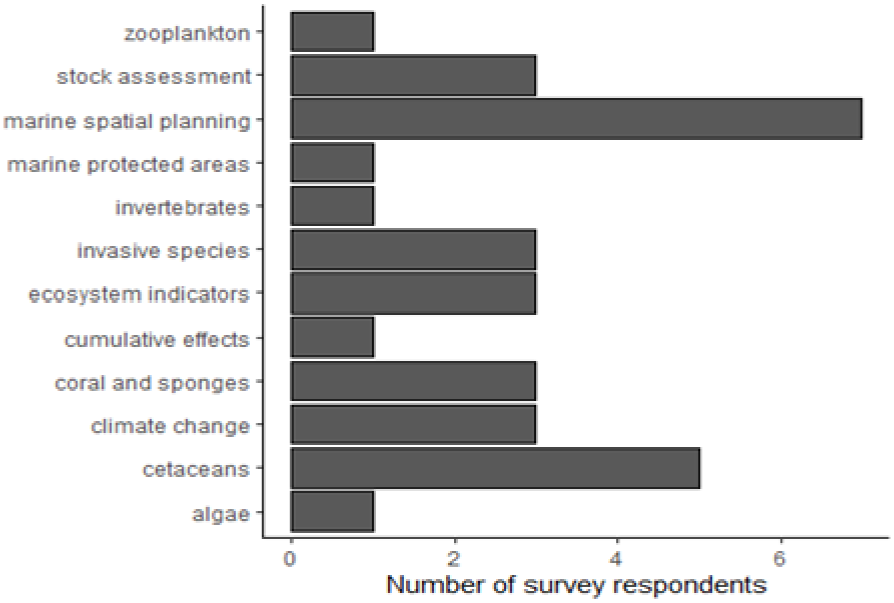


Figure 4. Survey responses by research area.

II: ACCESS TO SPATIAL DATA TO SUPPORT SPATIAL ANALYSIS AND MODELLING

By Shelley Lang, Laura Feyrer, Gordana Lazin, Jessica Nephin, and Catalina Gomez

The species distribution modelling forum identified a diversity of environmental predictors being used by different research groups at DFO and a variety of approaches to access this information. The process of requesting environmental predictor datasets and processing them is drastically different between regions and in some cases it has been conducted in parallel by individual research groups leading to duplication of efforts and inefficiencies. As a result, a national virtual forum was proposed for researchers at DFO interested in assembling spatial data products for spatial analysis and modelling. This forum was proposed to support and foster collaboration, networking, improve efficiency, learn from achievements in different regions, and to help reduce overlapping requests for groups with similar research needs.

The purpose of this workshop was to identify what data products are available, and discuss obstacles and solutions for data access. This workshop allowed us to identify the diversity of tools available at DFO for data discovery and access, as well as gaps in products that are not yet available to better support internal data needs.

Gallery of recordings

Recognizing the adaptation strategies being implemented as the workforce adjusted to the COVID-19 situation, presenters were requested to pre-record their presentations in Microsoft Teams. This allowed DFO participants to watch presentations in their own time and to make the material available for staff in the future. Presenters were asked to limit their presentations to approximately 10 minutes. Presentations enabled participants to gain insight into how specific data products can be used to inform studies and what reproducible tools are available to manipulate and process data for use in spatial analysis and modeling. Presenters provided some perspective and expertise on the general considerations for using the environmental predictors and specific considerations for particular datasets, including any limitations users should be aware of. Presenters also provided information on access and usage considerations of the data for DFO researchers, which we hope will help streamline future data requests for users, data developers and custodians. This resulted in a gallery of recordings (titles are summarized in Table 4) that are available for DFO staff. The recordings could be a valuable resource for training and onboarding new staff in the Department that will be working on spatial modelling and analyses. The intent is that these recordings will become an evergreen gallery of resources to help alleviate some of the burden of data provision for data custodians, support data accessibility to all users across our organization, and share knowledge on available data products and organizational strategies.

Table 4. Presentations prepared by DFO staff in preparation for the forum.

Title	Presenters
Geospatial Layers in Maritimes Region	Gordana Lazin
BNAM: Ocean circulation model output for ecosystem studies	David Brickman and Zeliang Wang
Temperature observations and BNAM	Adam Cook
SC_TEMPERATURE database: a repository for all raw bottom temperature data files	Brent Cameron and Amy Glass
Strategic planning for assembling and managing (via the GIS hub, see APPENDIX) environmental predictors in Pacific Region	Cole Fields and Joanne Lessard
Assembled environmental predictors for species distribution modelling in Pacific Region	Jessica Nephin and Sarah Davies
Ocean Navigator: Ocean Data Visualization and Discovery Tool	Vanessa Sutton-Pande
A geodatabase of historical and contemporary oceanographic datasets	Michelle Greenlaw and Jessica Sameoto
Requesting remote sensing data e.g. Temperature, Chlorophyll-a	Carla Caverhill and Emmanuel Devred
PhytoFit App: Satellite Chlorophyll-a data and phytoplankton blooms	Stephanie Clay
Manipulating remote sensing data	Gordana Lazin
Manipulating Environmental Predictors for use in species distribution modelling	Phil Greyson
The Gulf of St Lawrence Ecosystem Matrix (GSLEA) Data Platform	Daniel Duplisea

Access to spatial data virtual forum

A virtual forum on challenges and solutions to data accessibility took place on November 23, 2020. It included an introductory presentation, two discussion panels led by Science staff, live polls of participants, a presentation from Marine Spatial Data Infrastructure (MSDI) on their internal to DFO Data Viewer application (https://gispi.ent.dfo-mpo.ca/apps/DataViewer/), and a general plenary discussion (Table 5). The forum presented a diverse mix of presentations – all available at this link. One hundred and twelve participants attended the forum, including staff from Gulf, Quebec, Newfoundland and Labrador, Maritimes, Pacific Regions, and national capital region (NCP) (Appendix 1).

Table 5. Agenda of Virtual Forum, November 23 2020.

Time	Item	Presenter/Panelist
9:00-9:15 am PDT 1:00-1:15 pm AST	Introduction and Context	Tana Worcester (Chair) Shelley Lang
9:15-10:15 am PDT 1:15-2:15 pm AST	Panel: Testimonials about challenges and solutions to data accessibility	Gordana Lazin, Frédéric Cyr, Joanne Lessard, Daniel Duplisea
10:15-10:30 am PDT 2:15-2:30 pm AST	Outputs of Live Poll	
10:30-10:45 am PDT 2:30-2:45 pm AST	Data Viewer, and other tools for Data Discovery at DFO as part of the Marine Spatial Data Infrastructure (MSDI)	Bill Goodine
10:45-11:00 am PDT 2:45-3:00 pm AST	Break	
11:00-11:15 am PDT 3:00-3:15 pm AST	Panel: Open Data is not always the solution for internal data needs	Jessica Nephin, Laura Feyrer
11:15-11:45 am PDT 3:15-3:45 pm AST	Discussion: what do we need to maximize efficiency in our everyday internal work to better support work in spatial analysis and modelling. Is there a vision for data sharing/distribution/access for internal use? What type of solutions are required? Pros/cons of national vs regional solutions? Next steps	All
11:45-12:00m PDT 3:45-4:00 pm AST	Adjourn	

The first part of the virtual forum consisted of presentations by Tana Worcester and Shelley Lang to provide background information from the first forum held earlier in the year. Tana Worcester, Science management champion for Marine Spatial Planning (MSP) in the Maritimes Region, provided an introduction for context setting. She highlighted that while it is beneficial to have diverse groups of people working on various projects, programs, and priorities, communication between groups is difficult and a siloed approach can affect the broader collaborative research impact. In this context, this virtual forum offered unique opportunities to:

- Increase networking and participation between various regions and programs
- Learn about the challenges and successes of our colleagues
- Create centralized research and data inventories
- Disseminate information beyond peer-reviewed publications
- Increase the availability and use of online tools

During planning, it was recognised that the success of the virtual forum would depend on contributors and participants working cooperatively together. To that aim, some common operating principles were presented as a guide (Table 6). These principles represent our overarching approach to these Learning & Development initiatives.

Table 6. Operating principles for Learning & Development events. Courtesy of Tana Worcester.

Operating principle	Examples of applications
Think rigorously	● We explore options ● We learn from others ● We challenge each other in a respectful manner (challenge the idea, not the person) ● We value critical thinking and constructive feedback ● We use structured decision-making
Be engaged	● We stay informed ● We stay connected ● We participate
Trust & amplify	● We build trust within the team ● We build trust with partners and “clients” ● We work to people’s strengths
Service	● We serve the Canadian public ● We constantly evaluate whether we are providing good value for money ● We strive to make the planet and the lives of Canadians better
Work with purpose	● We are clear about why we’re doing something, and we communicate this ● We move forward with persistence and focus ● High performers are recognized, enabled and rewarded ● We monitoring and evaluate progress ● We celebrate success
Optimism Prevails	● We believe that change for the better is possible and will happen if we work together
Excellence in Team Work	● We explore and adopt the best ideas, regardless of where they come from ● We celebrate teams and individuals ● Our pride comes from feeling like we’ve contributed productively to the team

The introductory presentations were followed by a panel of Science staff from several regions: Gordana Lazin (Maritimes), Frédéric Cyr (Newfoundland and Labrador), Joanne Lessard (Pacific), and Daniel Duplisea (Quebec), who identified data discovery and access obstacles and shared innovative solutions for accessing and documenting internal data to advance DFO's science priorities and advice.

Gordana Lazin from Maritimes Region shared information about the MSP approach to data discovery and sharing. The approach consists of using a data inventory as a way to organize geospatial data products (maps) available in Maritimes Region, and subsequent publications of the data layers on the open data portals, which enable data viewing, data download, and include the Harmonized North American Profile of ISO 19115:2003 (HNAP) minimum mandatory metadata elements. The information captured in the inventory includes the following fields:

● Group, Subgroups	● Comments
● Parameter	● Data available
● Source data set	● Restrictions
● Method	● Contact
● Data type	● Data Link (open portals)
● Spatial coverage	● Publications/Reports
● Spatial resolution	● Data Assembly
● Temporal coverage	● Authoritative source
● Temporal resolution	● Publication Status

This inventory is now being used in the Maritimes Region by staff in Science and Marine Planning and Conservation and was also adopted by the Newfoundland and Labrador and Pacific Regions as a way to coordinate efforts under the MSP program. The inventory in Maritimes region is being used to prioritize spatial products for publication to the Government of Canada data portals, such as Marine Spatial Data Infrastructure (MSDI, DFO internal), Federal Geospatial Platform (FGP, internal to federal government, includes data from 21 departments), and Open Data (open to public). A similar data inventory/publication approach could be adopted for environmental predictors used for the species distribution modelling in Maritimes Region.

Open data publications to date from Maritimes Region include:

- Monthly temperature, salinity and currents climatology of the North Atlantic Bedford Institute of Oceanography North Atlantic model (BNAM) (3 datasets)
- Coral and Benthic Habitat, hotspots and significant benthic areas (2 datasets)
- Aquatic Invasive Species: DFO Biofouling program

- Invasive Species: Marine Invasion Hotspot, modelling study (present and 2075)
- Priority areas for cetacean monitoring, Scotian Shelf and Newfoundland & Labrador

Frédéric Cyr, research scientist from Newfoundland and Labrador Region, presented his personal point of view in relation to the Atlantic Zone Monitoring Program (AZMP) data stream. His group is very small and they have significant challenges in sharing and disseminating data from the AZMP program. His group receives a large number of incoming data requests, which require the development of solutions such as links that a variety number of staff can access, and a mechanism to permanently identify datasets to make the data products citable (e.g. archiving on the Federated Research Data Repository: FRDR).

Daniel Duplisea, research scientist from Quebec Region, presented the Ecosystem Matrix Approach (Gulf of St Lawrence Ecosystem Approach: GSLEA). GSLEA was designed to facilitate data access to support an Ecosystem Approach to Fisheries Management (EAFM). He noted that an ecosystem approach would not be possible without data. Since an ecosystem approach to fisheries management is a priority within DFO, there is a need to provide staff with easy access to appropriate data. Challenges identified in the Quebec Region included:

- The data supply burden falls overwhelmingly to particular individuals: It has been challenging for particular data stewards as they are inundated with requests for environmental data, or updates of their oceanographic data.
- Loss of traceability and acknowledgement: It is difficult to keep track of data sources, owners, and contributors manually. This can lead to a failure to acknowledge or cite the appropriate people.
- Data requesters do not necessarily know what they want: Initial data requests are often non-specific and people may not actually know what they want until they attempt to use the data provided. This leads to repeated data requests which can place a heavy burden on data providers. Alternatively, it can discourage someone from asking for the most appropriate data for their needs and can lead to misuse of the data provided.
- Data products become stale: Data extracted and stored on a shared network drive may be updated intermittently, irregularly, or not at all.
- Data may be supplied differently between updates: sometimes the format of data changes between updates (variable names, units). It is important to have a tool where existing analyses are not broken by updates and which can be called directly from within analyses.

For Duplisea et al., the solution was to develop an R package and make it available to the public via GitHub (https://github.com/duplisea/gslea). Most Science staff, particularly in stock assessment, are using R and the required data can be integrated into their analysis with a simple library call. A spatial structure (8 regions) for the Gulf of St. Lawrence has been defined and data are provided by region. This may suffice as a first analysis, or even as a final one for many project needs. Basic plotting, data querying, and lagged correlation functions for initial exploration are provided. GSLEA can be seen as the first resource someone might look to for their data needs, and then follow through with the key contacts if something different is required.

For those who do not use R, the data can be downloaded as an Excel file. To further increase the ease of access, a point-and-click R-Shiny interface that could be made available via a cloud platform is under development. To prevent the data from becoming stale, the database can be updated quickly using automated scripts that query the various source dataset. To support data traceability, sources (name, address, email) and main citation for every variable is supplied. Furthermore, GSLEA provides some external data by scraping datasets from partners that make their data available (e.g. NOAA, Can Space Agency, primary publication data archives). It was noted that GSLEA is simply an accessibility tool that is not meant to duplicate the numerous relational databases available regionally or nationally.

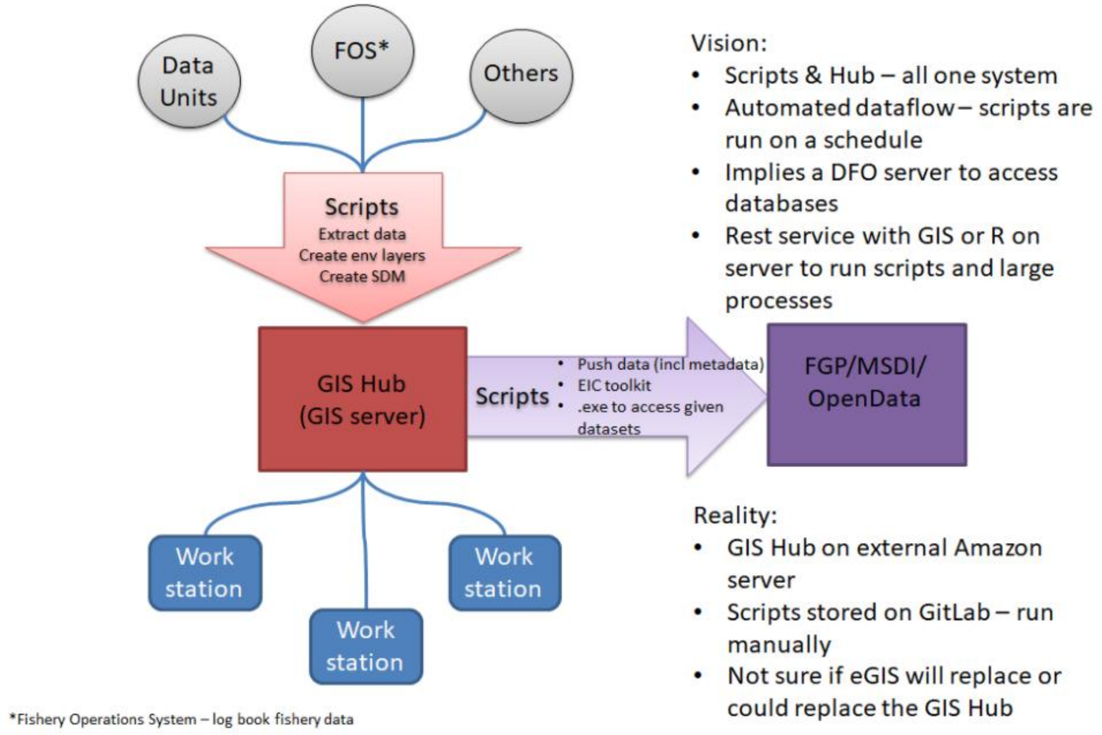
Joanne Lessard, biologist in the Pacific Region, shared a summary of the ongoing challenges with data access and management, their consequences and the emerging solutions (Table 7). She also shared her vision for streamlining data accessibility and distribution (Figure 5) and introduced a new initiative in Pacific Region, the Regional Spatial Data Coordination Working (ReSDaC) that is aiming to provide coordination across Science, Marine Planning and Conservation, Fisheries Management and other DFO sectors, to articulate what the spatial data needs are in terms of management and infrastructure, and communicate that to SEC. Goals of this group include identifying cross-cutting data management challenges and assessing the need for regional data governance.

Table 7. Challenges, consequences and solutions prepared by Joanne Lessard. The Pacific Region GISHub Metadata Standard is included in the appendix. Note that there is a tool on the GIS Hub that exports the metadata from the GIS Hub to HNAP standard.

Challenges	Consequences	Solutions
No central repository for spatial data	Data everywhere, incl. personal computers, external hard drives	Create the GIS Hub only for DFO staff. For more information about the GIS Hub please contact J. Lessard.
Lack of infrastructure	Really hard to find authoritative/current datasets	
	Duplication of efforts & resources	

	Possibly differing science advice	
Datasets not properly documented	Improper use of spatial datasets (unknown limitations, uncertainty, etc.)	Comprehensive metadata required for GIS Hub, including links to GitLab R or Python code (https://gitlab.com/dfo-msea)
	60-80% used data without knowing origin, method, version	Includes scripts to export HNAP/International Organization for Standardization (ISO) compatible
	the HNAP metadata standard is not enough	
Interference by and lack of support from IMTS/SSC	Cannot build the infrastructure we need to manage spatial datasets	Use of external cloud server is possible, but this is not in line with IMTS/MSDI policy
	Server requests denied because it was communicated that enterprise geographic information system (eGIS) may provide those capabilities in the future	
	No interim solutions provided until eGIS & TADAP are functional	
No governance and overarching data management strategy for spatial data	Lack of spatial data management	ReSDaC
	GIS Hub managed by Marine Spatial Ecology and Analysis (MSEA) section in the Pacific Region	
	No accountability across sectors	
	No long-term funding	
Lack of communication	Duplication of efforts & resources	ReSDaC
	Possibly differing science advice	
Limited functionality and access to eGIS	We really cannot do what we need to do in eGIS given the current functionality available	We will continue to use GIS Hub
Interference from MSDI/CHS	Removed all CHS datasets from GIS Hub	Publish metadata of CHS datasets onto GIS Hub with

	Again no central repository to search for available datasets - most CHS datasets are NOT on FGP/OpenData	links to where they can be accessed
Slow network connectivity	Cannot work of web/network files, have to download ALL datasets for analyses (could potentially lead to duplication again)	None identified
Not compliant with Open Data policy	Scattered storage leads to no accountability which leads to not sharing	ReSDaC



Vision:
- Scripts & Hub – all one system
- Automated dataflow – scripts are run on a schedule
- Implies a DFO server to access databases
- Rest service with GIS or R on server to run scripts and large processes

Reality:
- GIS Hub on external Amazon server
- Scripts stored on GitLab – run manually
- Not sure if eGIS will replace or could replace the GIS Hub

*Fishery Operations System – log book fishery data

Figure 5. Contrasting the vision and current reality for data accessibility and distribution to staff by Joanne Lessard.

During the panel, 68 participants responded to our live-poll. Majority of respondents were from Maritimes Region (31), followed by Pacific (24), Gulf (5), Quebec (3), NHQ (3), Central & Arctic (2) and Newfoundland and Labrador (2). The most important tool identified by participants to find out what environmental data is available in each region was emailing/asking colleagues (88%) followed by search of the data holdings (42%), checking the Pacific GIS Hub (see APPENDIX) (32%), and checking regional inventories (19%). 10% responded that they do not know where to start. Other options (<3%) included google searches, literature reviews, Marine Environmental Data Section (MEDS), Oceans Science Branch (DFO - OSB) website, Biological and chemical marine data: BioChem (https://www.dfo-mpo.gc.ca/science/data-donnees/biochem/index-eng.html), NOAA, OGSL (St. Lawrence Global Observatory), Ocean Navigator (https://navigator.oceansdata.ca/), and NASA (https://nasa.github.io/data-nasa-gov-frontpage/). This live poll reflected the variety of options available, and the need for data discovery tools to improve the knowledge of and access to the data available in our organization.

Bill Goodine, from the Marine Spatial Data Infrastructure (MSDI), provided a demo on the Data Viewer, a data discovery tool under development to locate spatial data products: https://gisd.dfo-mpo.gc.ca/portal/apps/sites/#/msdi. Open Data products that are visualized in this tool are from DFO, and there were questions as to whether spatial products from other organizations may be included in the data viewer. In support of the MSP program, an atlas for Atlantic Canada’s three bioregions (Scotian Shelf, Gulf of St. Lawrence, Newfoundland and Labrador) will be developed by March 2022. This Atlantic-wide compilation of data and information will be a web-based, public platform with interactive maps of ocean ecosystems, human uses and management areas. The MSDI data viewer is one of the tools currently under examination as an example of spatial applications to support this deliverable.

There are many different solutions being developed at DFO for data discovery in support of program deliverables such as the MSP atlas. The Directive on Open Government and the Open Data infrastructure (https://open.canada.ca/en/open-data) including the Federal Geospatial Platform (https://gcgeo.gc.ca/en/index.html) are very important steps towards making data and data products more accessible to everyone, including DFO Science staff. However, this is not always the solution for internal data needs. Jessica Nephin and Laura Feyrer provided two examples of some of the challenges to accessing and using datasets available on Open Data for spatial modelling and analysis. These examples highlight the continuing need for an internal spatial data platform to reduce the duplication of effort of data users and the workload of data providers.

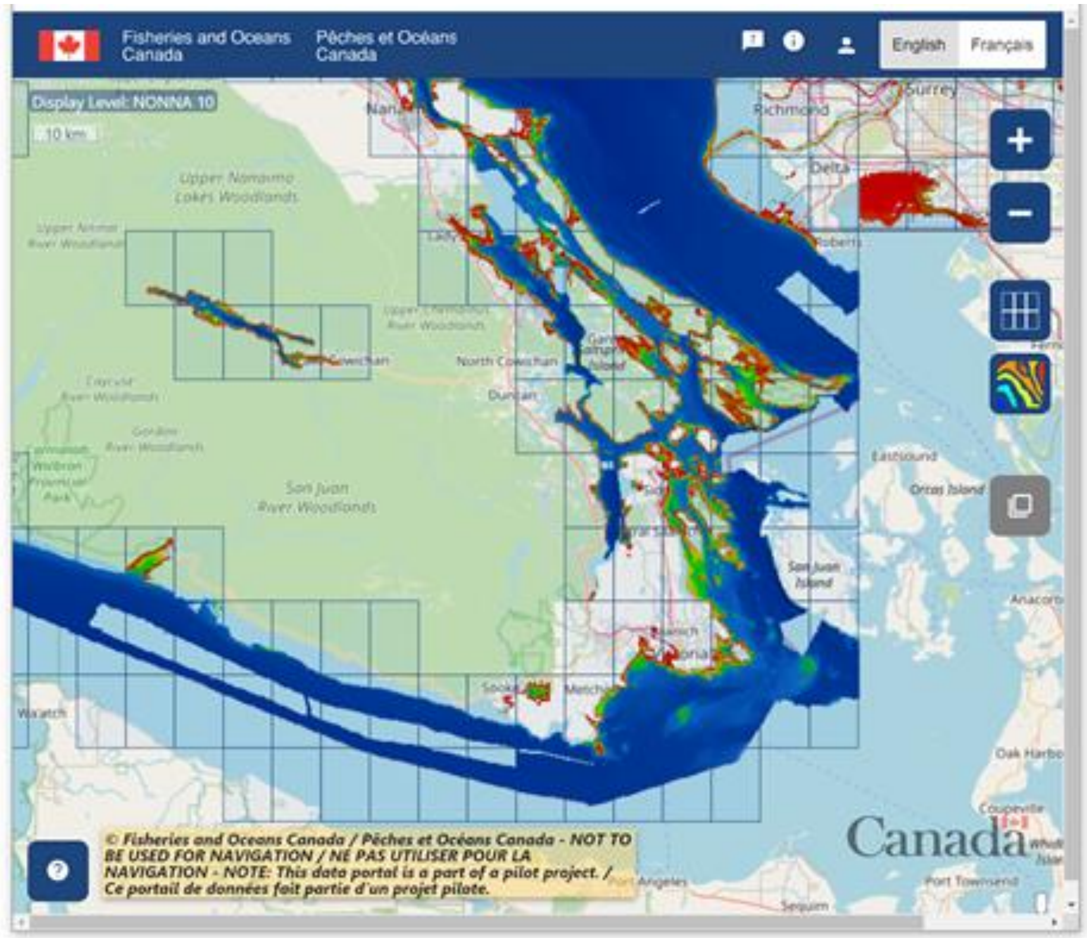
The first Open Data example focused on accessing and processing the Canadian Hydrographic Service (CHS) non-navigational (NONNA) bathymetric data 10 m product within the Pacific Region extent. The objective was to determine the coverage of the bathymetry product to evaluate whether it was appropriate to use for spatial modelling on the Pacific Coast. The bathymetry data was accessed using the CHS NONNA Data Portal (https://data.chs-shc.ca/login). The data portal made it easy to navigate the individual files, or blocks (Figure 6) and multiple blocks could be selected for download at one time. However, only up to 100 files (2% of the dataset) could be downloaded at one time, which made it quite laborious to download the entire Pacific dataset which is comprised of several thousands of files. Thus, this portal approach to data access, while providing an easy way for the public to interact with the data, can act as a barrier to access for DFO analysts who require a way to automate the process so it can be repeatable. Providing users with a simple and consistent way to download data, like an ERDDAP data server, (e.g. https://coastwatch.pfeg.noaa.gov/erddap/download/setup.html) would allow users to automate data access and support reproducible analysis and research.

Once the entire Pacific 10 m bathymetry dataset was downloaded from the data access portal, there was a need to mosaic the raster files into a single or several larger raster files so the bathymetry could be used for a variety of purposes (e.g., spatial modelling) and its coverage could be evaluated. Once the larger mosaicked rasters were created, there were several requests by colleagues for the mosaics so they would not have to repeat the time consuming data access and processing steps. These requests brought several questions to light:

1) Should these sorts of data products, those derived from other data products, be available internally and where should they be stored?

2) What metadata should be included with them?

3) Are there other considerations such as data ownership or user restrictions that should be considered when sharing among colleagues?



© Fisheries and Oceans Canada / Pêches et Océans Canada - NOT TO BE USED FOR NAVIGATION / NE PAS UTILISER POUR LA NAVIGATION - NOTE: This data portal is a part of a pilot project. / Ce portail de données fait partie d'un projet pilote.

Figure 6. CHS NONNA Data Portal https://open.canada.ca/data/en/dataset/d3881c4c-650d-4070-bf9b-1e00aabf0a1d

The second example looked at accessing BNAM modelled climatology data, which is available from Open Data as a set of averaged rasters contained within a ArcGIS geodatabase (GDB). Opening rasters stored in a GDB requires an ArcGIS license to open, which only some users may have access to. The dataset available was also partitioned in ways that differed from initial needs, providing a larger extent with greater spatial (entire western North Atlantic) and depth (8 strata) coverage than required, which encompassed a study area at two depths on the Scotian Shelf. In addition, while data is continuously being updated, the temporal coverage ended in 2015. The alternative was to make a custom request to the model developers. Through this process it became clear that such custom data requests are time consuming on both sides, as understanding the potential options for defining spatial extent, data averaging, and time periods of selected variables took multiple emails to confirm. Data was provided via a temporary link using an external web hosting service. This is currently the easiest option for sharing large custom datasets, but is vulnerable to data loss if the data isn't downloaded in time or original files are lost and required at a later date. The dataset provided was a large number of text files and although other colleagues had written scripts to process BNAM data, they had received it in Network Common Data Form (NetCDF). As a result custom scripts had to be written to read, compile, clip to the study area extent and summarize at various different temporal and spatial resolutions. Since this time other DFO colleagues have since enquired about the script used to process these files, however the data format they have access to is again different (Matlab files) and so requires new code to read and process the data. The experience of requesting custom data products brought several questions to light:

1) Is there a way to simplify the number of steps involved in large custom data requests and processing to ease burden on providers and support multiple DFO users?
2) Is it possible to design a standardized access template or pipeline for large datasets that provides flexibility required for different internal users, without limiting their options?

This forum finished with a general discussion about the challenges and opportunities to continue to connect across all DFO regions, to have a better understanding of the tools available, and to identify synergies between different initiatives and projects. We will continue to organize initiatives to broaden the collaborative research potential at DFO that we hope will continue to create spaces to network and support the exchange of information to ultimately improve our efficiency.

Participant survey results

Forum participants were asked to complete an evaluation survey at the end of the forum. Of the 111 attendees, 28 responses were received (25% response rate). Responses were received from DFO and ECCC participants based in the Pacific, Maritimes, Gulf, NHQ, Ontario and Prairie, Arctic, and Newfoundland and Labrador Regions (Figure 7). Participants worked in a wide range of research areas (Figure 8).

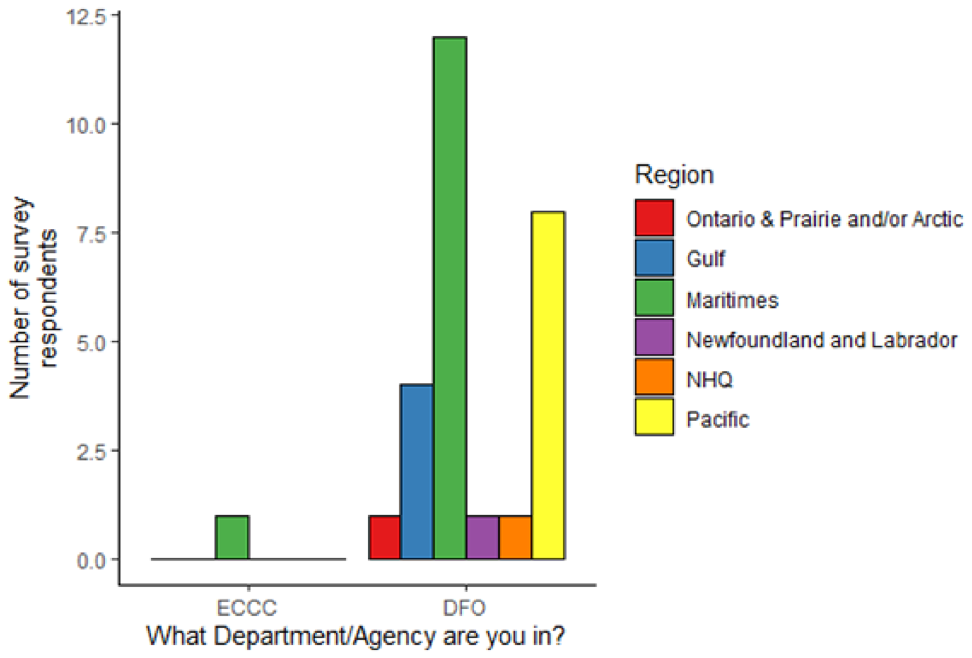


Figure 7. Survey responses by department and region.

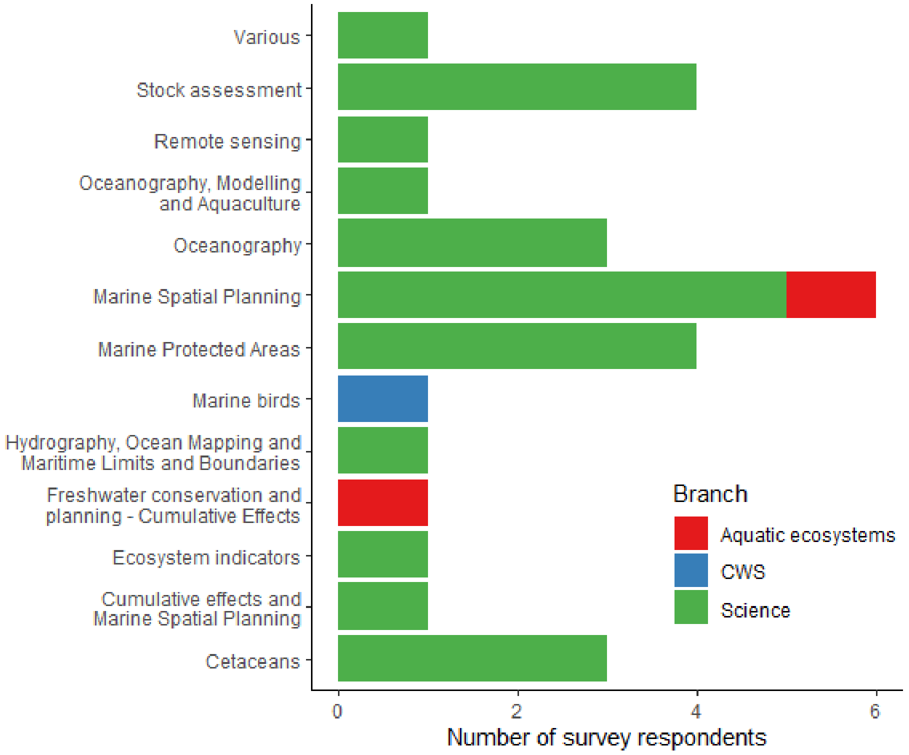


Figure 8. Survey responses by research area.

Participants were asked to rate the forum on various factors on a scale of 0-5 (low to high). The minimum rating for any question was 2, and median ratings were 4 or 5 (Figure 9).

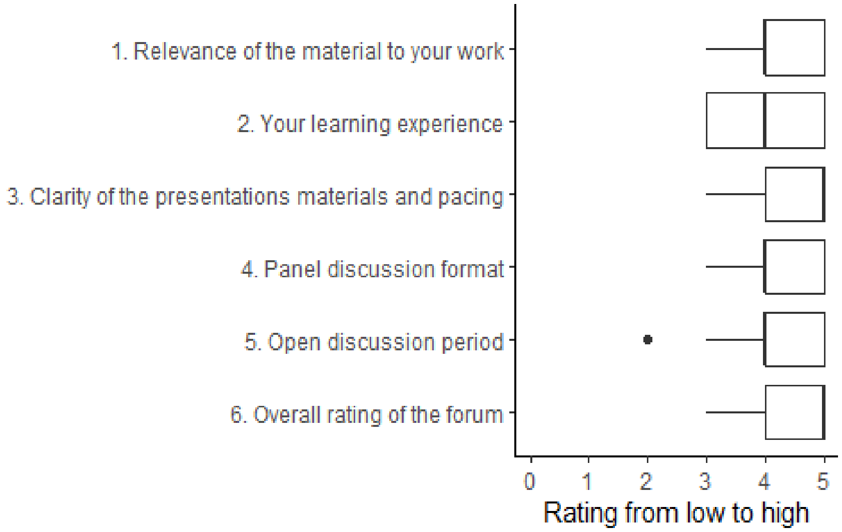


Figure 9. Summary of responses to rating questions. Each rating question was answered by all 28 respondents.

Survey respondents provided examples of other projects or successful initiatives that provide broader access to spatial data beyond those discussed in the forum. The examples were:

- Bilateral (Gulf/Maritimes) collaborative meetings between Integrated Planning and Science sector
- An R Shiny tool under-development for freshwater ecosystems
- External stakeholder outreach project to promote the use of public databases
- Mendeley Data online repository
- Environment and Climate Change Canada’s Emergency Response App

All survey respondents said they would recommend this forum to others. Reasons provided were:

- Inter-regional learning opportunity
- Discussions were well moderated
- Broad overview of information
- Good participant engagement
- Collaborative problem solving
- Opportunity for data managers to receive feedback
- Informative and concise presentations

The majority (89%) of respondents will try to adopt ideas presented in the forum. Some respondents noted that they were already considering using some of the tools discussed (e.g. GitHub, online data platforms), others said that they were previously unaware of the available resources but are keen to use them now, and others felt that the existing resources were not applicable to their work. Respondents also acknowledged a need for creating standards and collaborating with IT to ensure interoperability, and stressed the need for ongoing engagement to develop an effective long term structure.

Survey respondents had remaining questions on:

- Matching data output formats to data analysis workflows (e.g. format should be R-friendly)
- Government and IT-approved processes for data sharing
- Development of standards for sharing data and code
- Limitations and benefits of using code over GIS point-and-click tools
- Potential for artificial intelligence and machine learning
- How to share their data
- Linkages to Marine Spatial Planning

For future training opportunities, respondents suggested the following topics:

- Tutorials on code-writing best practices to facilitate sharing
- Incorporating open-source tools (Geospatial Data Abstraction Library [GDAL], R) into data pipelines
- Developing data discovery/viewing platforms
- Attribution of data products; intellectual property licensing
- Presentations from IM&TS and Science Data Management on infrastructure and plans
- Discussions with Open Government program
- Marine Spatial Planning

III: SCIENCE PODCASTS

In 2020, we initiated a series of science podcasts, a series of one-hour long informal online dialogues that provide an informal online space to discuss different science related topics, a forum to pose questions to colleagues, and to explore ideas informally across all DFO regions in support of innovation. During the first two podcasts we explored the challenges and promise of reproducible research, followed by a podcast about the latest techniques and advancements in image annotation, machine learning and Artificial Intelligence (AI). In preparation for the forum, Peter O'Blenis and Jim Theriault organized a third podcast titled "Digital Transformation Strategy - Exploring Artificial Intelligence technology at DFO" on Friday November 20, 2020. Andrew

Cogswell introduced this podcast and provided some context about the importance of investments in data management and data management infrastructure at DFO as decisions are informed by data, and the strength of those decisions, policies and regulations are a product of the quality of the data that are used to inform them. Our ability to provide accurate and timely science advice is impaired if the necessary data are not readily accessible. The objective of this podcast was to discuss practical internal initiatives that are helping us move towards the ideal of readily accessible data, automated tools and collaborative projects and approaches.

The podcast was hosted by Peter O'Blenis who is working with the Client Portfolio Management (CPM) team acting as a liaison between Science branch and DFO's Information Management & Technology Service (IM&TS), and included presentations by Jim Theriault and Steve Zhang. In the podcast they shared their progress on a project to centralize the processing of Automatic Identification System (AIS) vessel monitoring information at DFO. This initiative leveraged efforts from the CANDEV data challenge (hackathon) that DFO participated in at the University of Ottawa in January 2020 that led to the recruitment of 3 students. The project captured the interest of the TADAP (Target Architecture for Data and Application Platform) team. As a result, the team is working with a database analyst from TADAP with the goal of developing an AIS data pipeline using TADAP tools. This project was able to leverage financial support secured from the SEC to partner with the TADAP team in order to develop a centralized approach to ingesting, processing, decoding, storing and making the AIS data available via an API to DFO researchers. In addition, by moving this process to a centralized cloud-based resource, the team hopes to make available tools to analyze and visualize AIS Data. This podcast highlighted internal efforts to explore solutions to solve issues related to ingestion, processing, decoding, and storing of big data.

FINAL REMARKS AND RECOMMENDATIONS

This series of initiatives offered the opportunity to come together, network, learn, and exchange information in a collaborative setting on data access, spatial analysis and modelling. Although the many groups that participated have different work objectives, and are part of different programs, participants shared common goals in relation to:

- Improving the quality and reproducibility of spatial analysis and modelling
- Developing reproducible tools using version control to maintain institutional knowledge as people move jobs or retire
- Strengthening relationship building amongst regions to maximize opportunities for collaboration
- Supporting network building and programming language and statistical modelling training opportunities that collectively build on the active and open coding community within DFO

The main intent of this initiative was to provide a forum for DFO staff to present their ongoing work and issues in relation to spatial data, analysis and modelling. A secondary goal was to learn how various programs and regions were resolving these issues and to build a common understanding of each other’s perspectives. Based on the presentations and discussions that followed during the two forums we present the following recommendations for increasing reproducibility and institutional efficiency in spatial analyses and modelling efforts:

1) Make data processing and analysis steps available to others by sharing well documented code (e.g. using GitLab or GitHub)
2) Make spatial data products available via internal data repositories (e.g. Pacific GISHub, GSLEA R package, etc.) and on Open Data if appropriate
3) When sharing data products use open data formats that will be accessible to all users and follow well established metadata standards
4) When publishing results, list all the data sources and correctly attribute and acknowledge data providers
5) Participate in collaborative initiatives and training when available, to share and build your knowledge within the DFO spatial analysis and modelling community

It is our hope that the information gathered from these learning and development opportunities can support already established working groups and task forces currently tackling data discovery and management challenges at DFO. As a starting point for data managers, we recommend the development and use of data inventories with consistent metadata fields. These inventories could be used to populate the Target Architecture for Data and Application Platform (TADAP) catalogue and subsequently facilitate data discovery and sharing within the department.

ACKNOWLEDGMENTS

Special thanks to Tana Worcester for encouraging the principles of open science, collaboration and transparency that have inspired the R Learning and Development initiative, and also for chairing and developing material for the initiatives described in this report. Many thanks to the Science Executive Committee for supporting this virtual events and virtual community. Special thanks to participants, advisors, panelists, presenters, and podcast invitees including Jessica Sameoto, Andrew Cogswell, Daniel Duplisea, Joanne Lessard, Bill Goodine, Frédéric Cyr, David Brickman and Zeliang Wang, Brent Cameron, Amy Glass, Cole Fields, Jessica Finney, Sarah Davies, Vanessa Sutton-Pande, David Keith, Catherine Brennan, Sarah Wong, Lindsay Beazley, Ellen Kenchington, Javier Murillo, Kiyomi Fergusson, Jessica Coakley, Phil Greyson, Stephanie Clay, Candice St Germain, Chris Rooper, Cliff Robinson, Sean Anderson, Philina English, Andrew

Mcmillan, Devin Lyons, Brianna Wright, Dana Haggarty, Peter O'Blenis, Tammy Blair, and Jim Theriault. Thank you to Dewey Dunnington and Javier Guijarro-Sabaniel for reviewing and providing feedback on this manuscript. This workshop series would not have been possible without the interest and participation of DFO that provided ideas and support to prepare, deliver, and implement these workshops. The organizing committee was composed of the following DFO staff: Tana Worcester, Jessica Nephin, Shelly Lang, Gordana Lazin, Laura Feyrer and Catalina Gomez.

APPENDIX

The Pacific Region GISHub Metadata Standard

Metadata is a key part of any dataset that is published and shared with others. The GISHub metadata standard is an extension of ISO 19115:2003 (https://www.iso.org/standard/26020.html) that requires several additional metadata fields to be completed.

The GISHub stores metadata at two levels:

1) Dataset metadata that contains the majority of the information such as **basic**, **general**, and **science** metadata (described in the tables below).

2) Resource metadata that contains information like attributes in the layer and a description for the layer as well as spatial information.

The dataset metadata standard is divided into modules that group related metadata fields together. All fields in each metadata module are required unless marked as optional.

Note that there is a tool on the GIS Hub that exports the metadata from the GIS Hub to HNAP standard.

Basic Information

This section is a set of basic information required for all metadata entries. The metadata standard requires contact information for organizations responsible for various aspects of the data. Two sets of contact information are required: data creator and program manager.

Field	**Description**
Title	Short title, no longer than a newspaper headline. This is displayed on the GIS Hub when this dataset appears in search results.
URL (Dataset ID)	The URL for a dataset is a human-friendly unique identifier, which also forms part of the complete URL for a dataset.

Quality Control	Status of quality control for this dataset. By default, it is set to Check Required. Change to the appropriate status when quality control is complete or if problems found.
Summary	Brief narrative summary of the dataset's contents. Please summarize the following: · What data is included · What accessory information (reports, scripts) is included · Objectives · Describe the knowledge gap(s) the dataset is intended to fill · If there are multiple tables/layers in the dataset, how are they related?
Maintainer Email	Email address of a person responsible for the dataset. Notifications about access requests and other updates will be sent to this email.
Organization	Choose from the list - The organization (GC Department or Agency) primarily responsible for publishing the dataset. Departments within government should be specified down to the section level.
Visibility	Set to Public when all required metadata fields have been entered and the data is ready to be published. Note that setting Public means that the dataset is published and will appear in search results on the GIS Hub. All users on the GIS Hub will be able to view the metadata. Access permissions for resources (who is allowed to download / view) are set elsewhere. Set this to Private if you do not yet have all the information you need to complete the required metadata fields. When set to Private, you can save the metadata form, even if it is incomplete, and come back to it later. Your Private (i.e., incomplete) datasets are visible only to you and members of your organization. To come back to one of your private datasets later, click your username at the top right and look for entries marked Draft. You may also wish to bookmark the URL of the dataset, which will not change.

Cite this data as	Describe how this data should be cited. This is generated automatically by the GIS Hub from the metadata you entered, but it can be edited manually.
Start Date	Indicate the earliest date represented in this dataset.
End Date	Indicate the latest date represented in this dataset. If data collection is ongoing, do not leave this blank; choose Ongoing under the Status section.) If this is a model or derived data (as opposed to data collected in the field), enter the last date that the model used to generate this data was updated. In this case, end date can be the same as start date.
Data Creator	The Principal Analyst. The lead person responsible for creating the data. Provide the Name, Role, Department, Address, Phone, email.
Co-Creators *(optional)*	Provide the names of any additional co-creators (secondary authors) for this dataset.
Program Manager	The DFO Program Head responsible for the data. Provide the Name, Role, Department, Address, Phone, email.

General Metadata

This section describes general-purpose metadata required for all spatial data.

Field	**Description**
Topic Category	Main theme of the data. Choose the best match from the list.
Date Completed	The date on which the dataset was substantially complete in its current form.

Date Published	Date of publication of dataset. Default: today's date. This refers to publication of the data on this portal, not necessarily the date of publication of the associated academic research (academic references are described elsewhere).
Status	Development phase of the dataset. Choose the best match from the list.
Update Frequency	Revision cycle of the data. Choose the best match from the list.
Dataset Level	Use Dataset if this is a standalone data product. Series is individual regions in a series of datasets covering a larger area, or some other subset of a larger data package.
Keywords (GoC Thesaurus)	At a minimum, one keyword must be supplied from the Government of Canada Core Subject Thesaurus. Refer to http://canada.multites.net/cst/index.htm. You are most likely to find useful keywords in the "nature and environment" and "science and technology" categories. All keywords should be lowercase.

Science Metadata

This section is required for all scientific data.

Field	**Description**
Science Keywords	Enter additional comma-separated keywords not included in the general-purpose Government of Canada Core Subject Thesaurus. These keywords may include domain-specific vocabulary appropriate for scientific users. All keywords should be lowercase, including acronyms. Acronyms should be in singular form (no trailing s).
Theme	Choose the best match from the list.

Methods	The methods field documents scientific methods used in the collection or derivation of this dataset. It includes information on items such as tools, instrument calibration and software. It can also include a complete description of the lineage of the data. It may refer to sections of an associated academic citation or other published resource. Understanding the pedigree of the data is critical. Ideally, this will link the data set, through its methods, all the way back to the source data set(s), for which the data collection methods are known. Relevant raster methods include: any interpolation or extrapolation, and any masking or numeric manipulation of the source data.
Data Sources	List of source data inputs used to produce this dataset. For each input, provide a link to the source data where possible (to another dataset on the GIS Hub, or an external link). Include the date that the input was extracted, if known. If the source data changes over time, indicate the date on which the data was retrieved. Where possible, provide a citation for each input. Example input: Originator: Canadian Hydrographic Service, Fisheries and Oceans Canada (DFO) Publication_Date: 2015 Title: Geodatabase of Bathymetric point data for entire BC coast Other_Citation_Details: Geodatabase: GEOBASE_Entire_Coast.gdb (link to dataset if available)

Scripts or Software Routines	Link to script(s) used to process the data, for example on the GCCode GitLab site. Alternatively, include the scripts or queries in a Scripts folder inside in the zip file that you will upload. Document the purpose of each script, and anything required to run the script again. This field is now displayed using Markdown formatting, so links pasted here will be clickable. If no scripts were used to create the data, use this field to indicate how it was created, for example “manually created using analysis tools in ArcGIS.”
Spatial Data Quality	A measure of the quality of the spatial data representing some real-world entity. Describe the spatial data quality with reference to the data processing steps. Was location data captured with survey-grade GPS, consumer GPS, paper charts, ship’s log? Was data digitized from a scanned and georeferenced paper map? Is this dataset the result of intentional degradation of the input data to preserve privacy (e.g. 3 boat rule) and therefore has a known spatial resolution? Were multiple datasets captured at different spatial resolutions used in this analysis? Were point data extrapolated to polygons, or polygons/lines converted to centroids?
Positional Accuracy	May include horizontal and vertical. Express in metric units where possible.
Attribute Accuracy	Describe in general the degree to which attribute values reflect real-world conditions. For details on specific attributes, see the Attributes Metadata section.
Logical Consistency	Report on fidelity of relationships in the data set, the validity of its relationships to other data sets, and the degree to which the entities represent the real-world objects or concepts they are intended to capture.
Completeness	Report on the completeness of the dataset relative to its areal coverage, and the overall completeness of attributes.
Absence Data	Does the dataset include recorded (observed) absence of specific features, and if so, for which attributes?

Uncertainties	Describe uncertainties in measurement and methodology. Include any known biases. For example, was data collection opportunistic, or part of the survey design? In there any known bias from individual observers?
Use Restrictions	Restrictions on appropriate use of the resource. Examples: Not suitable for navigation. Do not use for fisheries assessment.
Change History	A description of changes made to the data since its release. Create an entry for each change, indicate the date of change and describe what was changed and why. When a dataset is first published, there is one change history entry for the initial creation of the dataset. Add additional change history entries as needed. If scripts have been used to generate a product, it is recommended to create a 'Release' of the code through GitLab or GitHub and reference that version of the release to track spatial products built using certains versions of code.
Change Date	Date on which the change was made to the data, or the metadata entry.
Temporal Coverage (optional)	If the temporal coverage of the dataset is more complex than a simple start and end date, describe it here. You may describe temporal coverage as a single point in time, multiple points in time, and one or more range of dates. Seasonality can also be described here. For derived datasets, please specify the temporal coverage of all input data here. Specific start and end dates are preferable, but year ranges are acceptable if the exact dates are uncertain or not applicable.
References	Academic references that are closely associated with the data. Create one entry for each reference. Indicate whether the reference is the place where this data was published (e.g. as a map).
Collaboration	Describe collaboration with academia, other government agencies, and external contributors. If produced internally by DFO, enter 'No collaboration outside of DFO.' Indicate collaboration with organizations rather than individuals.

Other Information (optional)	Relevant information that does not fit in any other category.
Confidentiality	Level of confidentiality or sensitivity of the dataset. For a description of the categories, refer to: https://www.tpsgc-pwgsc.gc.ca/esc-src/protection-safeguarding/niveaux-levels-eng.html. · Not Protected means that no harm would occur if the data were released to the public. · Protected A applies to data that could harm individuals or corporations if released, such as fishing events and bathymetry. Note that fishing data received 'in confidence' should be Protected A, not Confidential. Confidential applies to data that could harm the national interest if released. Data classified as Protected B, Secret, or any higher classification cannot be uploaded on this system. For these security classification, you may still create a dataset record using this system, but do not upload the actual data.