Reusability in Cryospheric Sciences: Fundamental concepts and case studies

Von der Fakultät für Georessourcen und Materialtechnik der Rheinisch-Westfälischen Technischen Hochschule Aachen

zur Erlangung des akademischen Grades einer

Doktorin der Naturwissenschaften

genehmigte Dissertation

vorgelegt von

Anna Lara Simson, M.Sc.

Berichtende: Univ.-Prof. Dr. sc. Julia Kowalski
Univ.-Prof. Florian Wellmann, Ph.D.

Tag der mündlichen Prüfung: 10.02.2025

Abstract

Reusability of research products in form of digital resources such as data sets and modeling software is of utmost importance to science. It makes research transparent, sustainable, and accelerates the scientific endeavor. Over the past few decades, repositories for sharing research products have been established, and standards for the consistency of digital representations of resources have been developed. The aim is to facilitate the reuse of data and software. Since 2016, guidelines for Findable, Accessible, Interoperable and Reusable research products have been formulated, widely known as the FAIR Principles. The FAIR vision describes a future automated state in which humans assign tasks like data analysis and integration to machines.

In light of these FAIR developments, one might think that reuse scenarios ranging from model validation based on existing measurement data, to model coupling using existing modeling software are smoothly executable. In reality, despite FAIR's focus on machines, reuse is still carried out mostly manually, i.e., humans find, integrate, and analyze existing research products, and reuse efforts remain time consuming and ineffective.

Cryospheric sciences, which study all occurrences of frozen water on Earth, would particularly benefit from highly reusable research products. Measurements of a constantly shrinking cryosphere cannot be repeated, in-situ measurements during polar expeditions are costly, and cryospheric modeling software must be easily coupled with models from other geoscience domains. A holistic understanding of the cryosphere is essential to predict and mitigate the impacts of climate change. This requires the combination of cryospheric data and models at different scales, and their seamless, interdisciplinary integration with data and models from other geoscience domains. Therefore, the reusability of cryospheric research products and their current leverage in reuse scenarios needs to be investigated in depth.

In this thesis, I investigate the reusability of cryospheric research products based on two case studies. Cryospheric Case Study I is a physics-based process model for coupled water vapor transport and settling in the snowpack. The model is characterized by modularity and extensibility. The reuse potential of the model's software is highlighted at the example of a real-world reuse scenario in form of a model comparison. Cryospheric Case Study II describes the approach followed to compile sea ice core measurement data into a comprehensive and analysis-ready database. This case study demonstrates the challenges encountered when manually harmonizing and combining distributed

and heterogeneous data sets. Both case studies put a special focus on the transparency of the method and the reusability of the generated research products.

The case studies effectively demonstrate that the reuse of cryospheric data and software is not trivial and not yet executable by machines alone. The sharing of cryospheric research products follows individual preferences, and the products lack standardization of data and metadata as well as quality information, both of which affect their understandability and interoperability. The case studies demonstrate: (1) the need to clearly communicate reuse needs in the form of reuse scenarios; (2) the large discrepancies between the many challenges faced when manually reusing a resource and the FAIRness of the resource, which reflects its machine reusability; and (3) the challenges experienced in manual reuse will be inherited by machines.

In the future, reuse scenarios should be documented more effectively to represent the reuse perspective in a systematic way. A major focus should be on improving reusability with transparent and comprehensive documentation of data and software with metadata, the development of community-agreed standards for terminology and formats, and the prioritization and documentation of data and software quality. Such developments will benefit manual reusers, support the development of technologies that enable autonomous reuse, and it will facilitate the combination of small, heterogeneous, and distributed data sets. To automate reuse scenarios, future research should specifically investigate the application of large language models, including exploring their limitations.

Zusammenfassung

Die Wiederverwendbarkeit von Forschungsprodukten in Form von digitalen Ressourcen wie Datensätzen und Modellierungssoftware ist für die Wissenschaft von größter Bedeutung. Sie macht Forschung transparent, nachhaltig und beschleunigt den wissenschaftlichen Fortschritt. In den letzten Jahrzehnten wurden Repositorien für die Bereitstellung von Forschungsprodukten eingerichtet und Standards für die Konsistenz der digitalen Repräsentationen von Ressourcen entwickelt, um die Wiederverwendung von Daten und Software zu erleichtern und zu verbessern. Seit 2016 wurden Richtlinien für auffindbare, zugängliche, interoperable und wiederverwendbare Forschungsprodukte formuliert, die allgemein als FAIR-Prinzipien bekannt sind. Die FAIR-Vision beschreibt einen zukünftigen, automatisierten Zustand, in dem Menschen die Aufgabe der Datenanalyse und Datenintegration an rechnergestützte Agenten delegieren.

Angesichts dieser FAIR-Entwicklungen könnte angenommen werde, dass Wiederverwendungsszenarien, die von der Modellvalidierung auf der Grundlage vorhandener Messdaten bis hin zur Modellkopplung unter Verwendung vorhandener Modellierungssoftware reichen, problemlos durchführbar sind. In der Realität erfolgt die Wiederverwendung trotz des Schwerpunkts von FAIR auf Computer immer noch weitgehend manuell, und Wiederverwendungsvorhaben sind nach wie vor zeitaufwendig und ineffizient.

Die Kryosphärenwissenschaften, die alle Vorkommen von gefrorenem Wasser auf der Erde untersuchen, würden besonders von hochgradig wiederverwendbaren Forschungsprodukten profitieren. Messungen an einer ständig abnehmenden Kryosphäre können nicht wiederholt werden, In-situ-Messungen während Polarexpeditionen sind kostspielig, und Software zur Modellierung der Kryosphäre muss leicht koppelbar mit Modellen aus anderen geowissenschaftlichen Bereichen sein. Ein ganzheitliches Verständnis der Prozesse und der Entwicklung der Kryosphäre ist für die Entwicklung technologischer Lösungen zur Vorhersage und Minderung der Auswirkungen des Klimawandels unerlässlich. Dies erfordert nicht nur die Kombination von Kryosphären-Daten und -Modellen auf verschiedenen Skalen, sondern auch deren nahtlose interdisziplinäre Integration mit Daten und Modellen aus anderen geowissenschaftlichen Bereichen. Daher ist es notwendig, die Wiederverwendbarkeit von kryosphärischen Forschungsprodukten und ihre derzeitige Nutzung in praktizierten Wiederverwendungsszenarien eingehend zu untersuchen.

In dieser Arbeit untersuche ich die Wiederverwendbarkeit von kryosphärischen Forschungsprodukten anhand von zwei Fallstudien. Fallstudie I ist ein physikbasiertes Prozessmodell für die

Kopplung von Wasserdampftransport und Setzung in der Schneedecke. Das Modell zeichnet sich durch Modularität und Erweiterbarkeit aus. Das Wiederverwendungspotential der Modell-Software wird anhand eines realen Wiederverwendungsszenarios in Form eines Modellvergleichs aufgezeigt. Fallstudie II beschreibt den Prozess der Zusammenstellung von Datensätzen mit Messdaten aus Meereisbohrkernen zu einer umfassenden und analysebereiten Datenbank. Diese Fallstudie zeigt die Herausforderungen, die bei der manuellen Harmonisierung und Kombination von verteilten und heterogenen Datensätzen auftreten. Beide Fallstudien legen besonderen Wert auf die Transparenz der Methoden und die Wiederverwendbarkeit der generierten Forschungsprodukte.

Die Fallstudien zeigen deutlich, dass die Wiederverwendung von Kryosphären-Daten und -Software nicht trivial und noch nicht ausschließlich maschinell durchführbar ist. Die Bereitstellung von Produkten der Kryosphärenforschung folgt individuellen Präferenzen und es fehlt an Standardisierung von Daten und Metadaten sowie an Qualitätsinformationen, die beide die Verständlichkeit und Interoperabilität der Produkte beeinflussen. Die Fallstudien zeigen: (1) die Notwendigkeit, den Wiederverwendungsbedarf in Form von Wiederverwendungsszenarien klar zu kommunizieren; (2) die große Diskrepanz zwischen den vielen Herausforderungen, die mit der manuellen Wiederverwendung einer Ressource verbunden sind, und der FAIRness der Ressource, die ihre maschinelle Wiederverwendbarkeit widerspiegelt; und (3) die Tatsache, dass die Herausforderungen der manuellen Wiederverwendung auch Maschinen betreffen werden.

In Zukunft wird es wichtig sein, Wiederverwendungsszenarien zu kommunizieren, um die Perspektive der Wiederverwendung systematisch aufzuzeigen. Darüber hinaus sollte das Potenzial kleiner, heterogener und verteilter Datensätze und deren Kombination untersucht werden. Die Verbesserung der Wiederverwendbarkeit erfordert eine transparentere und umfassendere Dokumentation von Daten und Software mit Metadaten, die Entwicklung von gemeinsam vereinbarten Standards für Terminologie und Formate und eine allgemeine Priorisierung sowie Dokumentation der Daten- und Softwarequalität. Solche Entwicklungen kommen manuellen Wiederverwendungen zugute und unterstützen die Entwicklung von Technologien, die eine autonomere Wiederverwendung ermöglichen. Konkret sollte die zukünftige Forschung die Anwendung von Large Language Models zur Automatisierung und Erleichterung von Wiederverwendungsszenarien untersuchen, einschließlich der Untersuchung der Anwendungsgrenzen.

Acknowledgments

I am grateful for the many people that supported me and positively influenced my PhD project, and I would like to express my gratitude to some of them in particular.

Prof. Julia Kowalski from the RWTH Aachen planted the seed of doing a PhD in my mind, while I was writing my Master's thesis with her. I am very thankful that she offered me a PhD position, which gave me the freedom to pursue my interests. On a personal note, I particularly appreciate her open-mindedness, flexibility, and optimism. After each of our meetings, I was energized by her curiosity and left full of motivation. On a professional level, our discussions helped me overcome the many challenges I faced and ensured that I stayed on track. I am very grateful for her scientific support and guidance that were invaluable for the success of my PhD project.

I thank Dr. Anil Yildiz from the RWTH Aachen, who is not only a co-author of my second article, but he has also collaborated with me on several other contributions. His scientific feedback and guidance have been invaluable for the submission of the article. Our discussions have inspired my ideas about the transparent management of research products. He has always been available to answer all kinds of questions and to give me scientific advice.

I want to thank Dr. Henning Löwe from the SLF in Switzerland, who is a co-author of my first article. I really enjoyed our collaboration in the beginning of my PhD. His knowledge on snow physics and his scientific feedback have advanced my PhD project significantly. I will never forget his live-support for the submission of my first article.

I would like to thank Dr. Achim Basermann from the German Aerospace Center (DLR) for his consistent enthusiasm about my research project since day one. His suggestions and support have been very important to me.

I thank Dr. Marc Boxberg from the RWTH Aachen for his guidance during the development of the sea ice core database. I am thankful for the collaborations on many conference contributions that he led.

I want to thank Dr. Johanna Kerch from the University of Göttingen for our discussions on cryospheric data management. The exchange with her has showed me the relevance of better data management and its potential.

I thank the members of the doctoral commission. I would like to thank Prof. Florian Wellmann from the RWTH Aachen, who was the second examiner and provided valuable feedback on the thesis; and Prof. Stefan Back from the RWTH Aachen, who chaired the defense.

I would also like to express my gratitude to the Glaciology Section of the Alfred Wegener Institute, where I spent a few weeks as part of the Helmholtz Information & Data Science (HIDA) Trainee Network. I especially want to thank Dr. Maria Hörhold, Dr. Johannes Freitag, Dr. Daniela Jansen, Dr. Rémi Dallmayr, Prof. Olaf Eisen, and Prof. Frank Wilhelms. During my stay, they shared with me a variety of perspectives on meteoric ice cores, ranging from data management and measurement methods to ice laboratories and ice core logistics. These insights was very valuable to me.

I would like to thank Dr. Sebastian Mieruch from the Alfred Wegener Institute and Taco de Bruin from the Royal Netherlands Institute for Sea Research for reassuring me of the relevance of my research project. Their feedback has strengthened my resolve to continue. Furthermore, I would like to thank Dr. Sebastian Mieruch for integrating the RESICE database into the MOSAiC WebODV.

I thank my graduate school Helmholtz School for Data Science in Life, Earth, and Energy (HDS-LEE), and especially Dr. Ramona Kloß for coordinating the graduate school over the past 5 years. HDS-LEE provided a part of my funding and offered many courses, seminars, and retreats that allowed me to broaden my knowledge in Data Science and to exchange with other PhD students.

I also want to thank the NFDI4Earth Academy that I have been part of since spring 2024. As part of the academy, I had the opportunity to attend several workshops and events and to exchange with other earlier career researchers.

Dr. Ann-Kathrin Edrich has been my office mate during the past two years. I thank her for always being there for me and to give me feedback and support when in doubt. Her motivational speeches have always encouraged me.

Last but not least, I want to express my gratitude to the MBD team. I enjoy spending my day-to-day working life with you, and I have benefited from the nice working atmosphere, the mutual support, and many discussions on all our different research projects.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Sharing research products in an open and reusable way allows other scientists to build on existing work, and it drives scientific progress and collaborations (UNESCO, 2022). Reusable research products promote transparency, accelerate scientific progress (Figueiredo, 2017; Pronk, 2019), and preserve scientific results for future investigations. Reusability of research products is particularly important because researchers have different areas of expertise. Some researchers are involved in field campaigns, while others receive special training in the development of complex models. Therefore, resources such as data sets and software should be interdisciplinary reusable by all, so that everyone can benefit and advance scientific progress.

Making research products available on the Web is a first step toward reusability, and the movement toward a more open and sustainable science has led to the sharing of vast amounts of research products (Pampel and Dallmeier-Tiessen, 2014; Tenopir et al., 2018). These resources are published alongside scientific articles but also independently. In order to be valuable and reusable for new purposes and by other researchers, resources must also be searchable, findable and have sufficient metadata. Metadata serves as a description of the resource and provides contextual details about it so that people other than the creators can find, understand, and reuse it (Jacobsen et al., 2020b). The reusability of research products is generally supported by the use and adoption of repositories and standards (Kim, 2021) that make resources and their metadata available and ensure consistency of format, structure, and terminology.

The demand for reusable research products has been embedded in data and code sharing policies of journals (Christensen et al., 2019; Vasilevsky et al., 2017), and has furthermore been adopted by founders and institutions so that the term *reuse* is mentioned in the European Union Data Act (EU, 2023). Over the past decade, the FAIR Principles (Wilkinson et al., 2016) for Findable, Accessible, Interoperable and Reusable research products have been used as a guide to increase the reusability of digital resources in many domains such as medicine (Holub et al., 2018) and cryospheric sciences (Frémand et al., 2023).

1.1 Has FAIR solved reusability?

Reusability of research products is the ultimate goal of implementing the FAIR (Findable, Accessible, Interoperable and Reusable) Principles. Since the initial publication of the *FAIR Guiding Principles for Scientific Data Management and Stewardship* in 2016, many FAIR guidelines, products, and methods have been developed. These include the *FAIR Principles for Research Software* (FAIR4RS) (Barker et al., 2022), *FAIR computational workflows* (Goble et al., 2020; Wolf et al., 2021), *FAIR digital objects* (De Smedt et al., 2020), *FAIR data points* (Da Silva Santos et al., 2023), and *FAIR metrics* (Wilkinson et al., 2019). The FAIR community is growing organically and many domains are slowly paving their FAIR way.

While reusability in the sense of FAIR also supports reuse efforts that are carried out by humans, the main focus of FAIR is to enable reuse of digital resources by machines (Mons et al., 2017). In the vision of FAIR, machine-reuse of digital resources should allow computational agents to autonomously carry out analysis and integration tasks at the request of humans (Wilkinson et al., 2016). Specifically, following Wilkinson et al. (2016), machines should be able to autonomously "determine if it [a resource] is useful within the context of the agent's current task by interrogating metadata and/or data elements." Jacobsen et al. (2020b) furthermore specify machine-actionability as the capabilities of machines to decide whether "a digital resource should be reused (i.e., is it relevant to the task at-hand?)" and "under what conditions (i.e., do I fulfill the conditions of reuse?)." This future state, in which humans can assign tasks to machines that then execute them using linked existing digital resources, has been referred to as the *Internet of FAIR data and services* (Mons et al., 2017; Jacobsen et al., 2020b).

Technologies such as repositories and standards, required for the realization of the described FAIR vision, have been developed for the past decades (Sansone et al., 2019). The findability and accessibility of resources can usually be solved with existing generic solutions such as by uploading the digital resource on a repository and assigning it a Digital Object Identifier (DOI). In contrast, the interoperability and reusability of resources require domain specific standards such as terminologies, metadata schemes, and formats that are still lacking in many communities (Boté and Térmens, 2019; Kinkade and Shepherd, 2021; Simmonds et al., 2022; Lush et al., 2024).

Despite the evolution toward FAIR, reusability remains a challenge for researchers. Some of the challenges encountered for data reuse have been gathered in a survey by Bishop et al. (2019). Participants reported on challenges including the non-persistence of the data location, missing or incomprehensible metadata, and the lack of metadata standards such as standardized terminology. In some cases, the data sets authors had to be contacted due to inaccessible data. Sometimes a decision on the adoption of a data set or modeling software can only be made after contacting the authors or the hosting repository (Yoon, 2017). Differences between individual models and data sets may not be obvious due to unspecific metadata. In case of modeling software, reusers often need a long time to create the compatible software environment including dependencies and packages for execution of the code, or they have to deal with insufficient documentation. Research products often miss quality information (Peng et al., 2021b) and sufficient context so it can be

fully understood as it would be necessary (Faniel et al., 2013; Faniel et al., 2019). Furthermore, relevant data may be distributed across data sets and repositories (Hughes et al., 2023) and lack harmonization and standards (Miller et al., 2015; Bavay and Fiddes, 2020; Duerr et al., 2024).

Instead of a quick integration of research products, researchers often have to deal with a variety of models and data sets without knowing whether the resources are actually suitable for their specific task. Although there are many positive examples of reuse, the attempt to reuse existing resources often is a frustrating and time-intensive experience, which may lead to failed reuse (Yoon, 2016b; Bishop et al., 2019). As a result, reuse is often hampered, making it time consuming and ineffective. The challenges also highlight that reuse remains a task to be performed by humans, i.e., humans find, analyze, and integrate data according to their needs. In this thesis, I refer to this human perspective of reuse as *manual reuse*, in contrast to the *machine reuse* envisioned by FAIR.

Stewardship of scientific products is currently in a FAIR transition phase (Bloemers and Montesanti, 2020), many technologies from repositories such as Zenodo and GitHub to standards such as controlled vocabularies and metadata schemes are being developed and increasingly used (Thompson et al., 2020). Thanks to increased FAIRification, reusability of and especially awareness for reusable research products is growing. However, the FAIRification process and the uptake of repositories and standards is only advancing slowly and shows a large discrepancy between generally positive data sharing attitudes of researchers and low adoption of good practices for managing research products (Thompson et al., 2020; David et al., 2020; Tenopir et al., 2020). Already existing historic data sets will most likely never reach such future FAIR state (Easterday et al., 2018). The majority of research products still lacks reusability. It remains to be investigated how FAIR-aligned solutions like standards and repositories are implemented and adopted by the scientific community, and how they contribute to the reusability of research products.

1.2 Reuse scenarios as context and purpose for reuse

The reuse of research products has a context and purpose in form of a task. Such tasks derive from research questions and rely on already existing resources. Yet, the usefulness and relevance of a resource for a task depends on its suitability with the specifications of the task (Wilkinson et al., 2016; Jacobsen et al., 2020b). In this thesis, I refer to these tasks as *reuse scenarios*. The following are generic, exemplary reuse scenarios with a focus on existing physics-based process models and data sets:

- **Conduct comparative studies** of the physical properties of a medium using existing modeled and/or measured data.
- **Validate model performance** of a physics-based process model through model validation studies using existing measurement data.
- **Train data-based models** to classify or predict physical properties of a medium based on existing data.

- **Interpret measurement data** acquired from a physical process by leveraging existing physics-based process models, for instance, to infer dominant processes.
- **Couple models** to account for interactive processes by combining different physics-based process models describing the same phenomenon.

Each of these scenarios involves the use of one or more existing research products. These products must suit the requirements of the respective scenario in order to be useful for it. The following description of the reuse scenarios of a comparative study and a model validation illustrate the different dimensions of the scenarios' requirements for existing research products.

In a comparative study, for example, the reuser would define the measurable and/or observable quantities of interest, the measurement methods, the temporal and spatial dimensions, and the boundary and initial conditions. Existing data sets must meet adequate measurement and modeling conditions. The boundary conditions must be within the range of interest, and the measurables and observables must be the same or transformable for a valid comparison. It is important to note that such requirements have to be reflected in the metadata of research products to efficiently determine their suitability for a reuse scenario.

When reusing software of physics-based models to interpret measurement data, the software must be flexible enough to adapt the model's geometry and initial and boundary conditions to the measurement conditions to provide a meaningful interpretation. To resolve dominant processes, the software should also allow the investigation of different physical processes in isolation.

Reuse scenarios are often carried out agnostically with respect to the existence and properties of research products since the reusers are primarily motivated by the research question and not by existing resources. This means that reuse scenarios are often performed by people who were not involved in creating the research product, i.e., collecting the data or developing the model, and who come from domains other than the research product's origin. Thus, I refer to such reusers as *agnostic reusers*.

1.3 Relevance of reusability for cryospheric research products

High reusability is particularly relevant for research products from areas that are undergoing drastic and rapid change due to global warming, such as cryospheric sciences (Johnson et al., 2015). The cryosphere encompasses all occurrences of frozen water on Earth including glaciers, shelf ice, sea ice, snow cover, and permafrost. Changes of the cryosphere critically impact our planet's climate system so that 6 out of 9 identified global core tipping elements belong to the cryosphere (Armstrong McKay et al., 2022). Therefore, measurements of sea ice and glaciers need careful preservation as they cannot be repeated. Predictions of the impact of climate change need to be performed and require easily reusable models for their coordinated reuse.

As the cryosphere is in constant interplay with the four other components of the Earth climate system biosphere, hydrosphere, atmosphere, and lithosphere (WMO, n.d.[a]), cryospheric research products also need to be interdisciplinary reusable. Atmosphere radiation, for instance, depends on the reflectivity of the Earth referred to as Albedo. Fresh dry snow reflects 85% to 95% and sea ice 30% to 40% of the incoming sunlight while tundra soil without snow has only 18% reflectivity (Budyko, 1974). Less reflection means more absorption of the Earth's surface, leading to a temperature increase of land and ocean as well as the atmosphere. The hydrosphere is affected by seasonal melt water discharge that elevates water levels in lakes and rivers (Chen et al., 2019). Furthermore, ocean water circulation is affected by sea ice, for instance, through ventilation from brine rejected during sea ice formation (Shcherbina et al., 2003), and the deglaciation of ice sheets and glaciers leads to sea level rise.

Consequently, ongoing research that involves the cryosphere is interdisciplinary and poses overarching questions such as:

- How does the melting of ice sheets affect sea level rise (Dutton et al., 2015)?
- How does a reduced sea ice extent affect weather and climate (Vihma, 2014)?
- How does a decreasing snow precipitation affect water availability (Barnett et al., 2005)?

Many reuse scenarios can be derived from these research questions and their investigation benefits from easily reusable cryospheric models and data across disciplines and types of methodological expertise, ranging from experimentalists to modelers. Answering these research questions requires interdisciplinary approaches that jointly investigate phenomena of the cryosphere with those of other spheres. These questions also call for reuse scenarios that combine of models and data within and across geoscience disciplines. Yet, in the past, the demand for reusable research products that can be easily integrated into reuse scenarios has not been adequately addressed in geoscience and its subfield cryospheric sciences.

To improve the reuse of geoscience data sets for interdisciplinary research, the geoscience community recently introduced the *Principles for Integrated, Coordinated, Open, and Networked* (ICON) Science (Goldman et al., 2022) as a complement to FAIR. Following a call from Goldman et al. (2021) contributions of the majority of the 25 sections of the American Geophysical Union (AGU) have been published that highlight current examples and directions of ICON research of the individual disciplines. The AGU Cryosphere section also contributed a commentary (Brugger et al., 2022) discussing a lack of uniformity and conciseness in data reporting formats, such as field and laboratory protocols, and in the resulting data sets. This is in accordance with Peng et al. (2021a), who highlight challenges of the management of geophysical research products; they are distributed across repositories, use manifold standards and formats, and have incomplete metadata. All of theses challenges impede coordinated reuse scenarios such as comparative studies or model validations.

An insufficient implementation and availability of standards for cryospheric sciences has also been highlighted by Bavay et al. (2020), who conducted a survey of 50 cryospheric data reusers. They

reported on incorrect or missing standards in cryospheric data sets, which may be due to the complexity of such standards, while at the same time often failing to meet the specific needs of cryospheric data. The lack of standardization of cryospheric data poses problems to the reusers, as data are not readily available for other applications due to a lack of interoperability. Communication with the data collector was often necessary, which highlights the general time intensity of cryospheric data reuse. Furthermore, Miller et al. (2015) emphasized the lack of harmonized standards for measurement records of bio-geochemical sea ice measurements that hinder a coordinated and meaningful combined database. They call for the establishment and use of basic standards for sample classification.

1.4 Challenges and objectives

As in many other scientific disciplines, cryospheric research products are currently not seamlessly reusable. In order to reuse existing cryospheric research products, reusers must overcome many barriers. This thesis addresses this lack of reusability. Specifically, I focus on three observations and the related gaps.

- On the one hand, the reuse of published cryospheric resources such as data sets and models remains challenging due to the lack of appropriate standardization of reporting formats and methods (Brugger et al., 2022) and the distribution of resources across repositories and databases. On the other hand, many repositories as well as standards have been developed, among which it is difficult to select those that are relevant for a specific research product (Sansone et al., 2019; Bavay et al., 2020). **GAP 1:** It remains to be investigated how the needs of cryospheric research products are represented in generic and domain specific standards and repositories, how they have already been adopted in cryospheric research products, and how they can be improved.

- To implement the FAIR vision of automated reuse, specific community standards need to be developed and defined (Wilkinson et al., 2019). The development process is not trivial and will take time (Kinkade and Shepherd, 2021; Brugger et al., 2022). A study performed by Bishop et al. (2019) showed that natural scientist lack understanding of the FAIR principles and standards required for their implementation. Consequently, implementation (if at all) of FAIR is heterogeneous, especially in communities generating rather small in contrast to big data (David et al., 2020) like many studies from cryospheric sciences. As a consequence, the reuse of research products remains a manual task. **GAP 2:** The current implementation of FAIR Principles in cryospheric research products and their effect on manual and automated reuse remain to be analyzed.

- While data sharing practices have been studied extensively (Tenopir et al., 2018; Gärtner-Roer et al., 2022), studies of reuse practices are sparse (Faniel and Jacobsen, 2010; Curty et al., 2017) except from questionnaires or interviews on general reuse behaviors of researchers (Gregory, 2020; LaFlamme et al., 2022; Bishop and Collier, 2022). Concrete examples putting

reuse scenarios in the focus are lacking although the FAIR principles generally consider the perspective of a reuse purpose (Wilkinson et al., 2016; Jacobsen et al., 2020b). This observation also applies to cryospheric sciences. **GAP 3:** There is a need to document and communicate reuse scenarios and practices from the cryospheric sciences to raise awareness for the benefits of highly reusable research products within the community.

The aim of this thesis is to contribute to the advancement of the reusability and FAIRness of cryospheric research products by addressing the described gaps. The main objectives of this thesis are as follows:

- Investigate the representation of the needs of cryospheric sciences in standards and repositories and their adoption in cryospheric research products.
- Present concrete examples of reuse scenarios for data sets and modeling software from cryospheric sciences including obstacles encountered.
- Compare the FAIRness of cryospheric research products, an estimate of FAIR-compliant machine reusability, with experiences obtained through the manual reuse of the products.
- Provide recommendations and practical solutions to improve the reusability and transparency of cryospheric research products.

1.5 Thesis outline

Chapter 2 summarizes terminology and concepts for reusability and reuse of data and software from the literature and introduces the FAIR principles. Furthermore, FAIR metrics and assessment tools to check a resource's compliance with FAIR are presented. The improvement of reusability and realization of FAIR requires technological solutions. Chapter 3 presents such digital solutions, namely repositories and standards. The focus is on the representation of cryospheric sciences.

Chapters 4 and 5 are two case studies from cryospheric sciences that serve to demonstrate current practices of reusing cryospheric research products in the form of modeling software and data sets. Cryospheric Case Study I in Chapter 4 describes a physics-based process model for coupled water vapor transport and settling in the snowpack. The model is characterized by modularity and extendability, which is also reflected in the modeling software. The software's reuse potential is highlighted with potential reuse scenarios and at the example of a conducted model comparison. Cryospheric Case Study II in Chapter 5 describes the process of compiling data sets containing measurements from sea ice cores into a comprehensive and analysis-ready database. This case study is motivated by two reuse scenarios, and it demonstrates the challenges encountered when harmonizing and combining distributed and heterogeneous data sets. Both case studies put a special focus on the transparency of the methods and the openness of the generated research products.

Chapter 6 provides a synthesis of the concepts of reusability and FAIR, drawing on the experiences of case studies. I discuss the potential of communicating and systematically documenting reuse scenarios. Next, I asses the FAIRness of the resources used in the case studies by applying automatic FAIRness assessment tools and compare the results with the experience of manual reuse from the case studies. Furthermore, I discuss the challenges of manual reuse and its implications for machine reuse. Chapter 7 contains concluding remarks on the discussion and provides an outlook on the application of large language models for reuse efforts. Finally, I present future perspectives on feasible efforts to improve the reusability of cryospheric research products.

Chapter 2

Reusable and FAIR data and software for sustainable research

Scientists base their work on other scientists' research. In the past, this was often limited to results published in written form, such as in articles, theses, and reports. Without access to the data and software used, building on the methods and results described in these publications is often tedious. In recent years, initiatives for Open Data (Murray-Rust, 2008; Mauthner and Parry, 2013) and Open Source (Wu and Lin, 2001) as well as journal sharing policies encouraged authors to also share their research data and software (Vasilevsky et al., 2017) and make them reusable for other purposes. Reusable research products foster scientific progress by making their reuse in other scientist's work time and cost efficient (EC DGRI, 2018; Pronk, 2019).

Although the terms *reuse* and *reusability* are widely used, they do not have standardized definitions. In this chapter, I first summarize definitions from the literature, including the delineation between reusability and reproducibility. Next, I describe the FAIR Principles for Findable, Accessible, Interoperable and Reusable data and research software, as well as the FAIR metrics and assessments developed to evaluate the *FAIRness* of research products.

2.1 Basic concepts and terminologies

In the context of data and software, *reuse* has been increasingly promoted since the publication of the FAIR Principles (Groth et al., 2020; Dempsey et al., 2022). However, concepts for the reuse of data and software are not new and have been topics of scientific discussion for decades (Krueger, 1992; Nakayama et al., 2006; Zimmerman, 2008). Alongside *reuse* and *reusability*, also the terms *reproducibility* is often mentioned when dealing with software product, which also does not have a uniformly applicable and standardized definition (Barba, 2018; Sandt et al., 2019).

2.1.1 Reuse and reusability

Before investigating the specific concepts of the terms in relation to software and data, it is instructive to consider their generic definitions in dictionaries. The Oxford English Dictionary defines the nouns *use* and *reuse* as "the act of putting something to work, or employing or applying a thing, for any purpose" (OED, 2024) and "second or further use" (OED, 2023b) respectively. The noun *reusability* is not defined, but the adjective *reusable* refers to being "able to be used again" and "suitable for second or further usage" (OED, 2023a). According to these definitions, *use* and *reuse* are acts with different purposes, while reusability is a property of a product that considers not only the general ability to be reused, but also the suitability of the product for reuse. Next, I summarize definitions and concepts of the terms in relation to data and software from the literature, and I conclude with the definitions applicable to the remainder of this thesis.

Definitions and concepts for the reuse and reusability of data

In the context of increased sharing possibilities for research data, concepts and definitions for *data reuse* have been formulated. Concerning data, Pasquetto et al. (2017) define reuse by discriminating it from *use*. Accordingly, data use is carried out by the person or group of persons, who collected the data, and it is related with the investigation of a defined original research question. If the original collectors fall back to their data set at a later time, this act is still considered a use since it is used in their context. Reuse, on the contrary, is described as data usage for a project other than the original project. This new project is carried out by persons other than the creators, and it involves access to the data through repositories. This definition assigns reuse two characteristics: 1) a person conducting reuse, and 2) a project that defines the purpose of the reuse. The purpose has to be different than that of the original project, and the reuser(s) has/have to be other than the creator(s). Pasquetto et al. (2017) also distinguish between the *independent reuse* of a single data set and *data integration* that combines data sets for a reuse purpose. While independent reuse of well-documented data sets may be challenging, data integration is considered an even more complex and demanding task.

Table 2.1: Concepts of the term data reuse with respect to the person conducting data reuse and the purpose. Definitions are based on publications gathered by Sandt et al. (2019).

Publication	**Data reuse person**		**Data reuse purpose**	
	original researcher	other researcher	new purpose	original purpose new methods
Szabo and Strang (1997)	✓	✓	✓	✓
Zimmerman (2008)	-	-	✓	-
Faniel and Jacobsen (2010)	-	✓	✓	-
Curty and Qin (2014)	-	-	✓	✓
Faniel et al. (2015)	-	-	✓	-
Yoon (2016a)	-	-	✓	-
Sun and Khoo (2017)	-	-	✓	-
Pasquetto et al. (2019)	-	✓	✓	-

Sandt et al. (2019) extensively reviewed the term reuse in relation to data by investigating its definitions and current uses in existing publications. In addition to the characteristics person (1) and purpose (2), they assign to the term *reuse* the characteristics *character of the data* (3), referring to a possible combination of data sets for reuse, and *time* (4), considering the assumption that reuse occurs after use. Sandt et al. (2019) studied 20 publications that use and define the term reuse. Most definitions strictly tie data reuse to a new purpose (Zimmerman, 2008; Sun and Khoo, 2017; Faniel et al., 2019). Two definitions also include original purposes if combined with new methods (Szabo and Strang, 1997; Curty and Qin, 2014). Similar to Pasquetto et al. (2019), a few publications define the person conducting the reuse to be different than the original data collectors (Szabo and Strang, 1997; Faniel and Jacobsen, 2010), while the majority does not specify the reuser in the definition. An overview on the definitions is provided in Table 2.1.

Based on their analysis, Sandt et al. (2019) claim that a generalized definition of reuse requires a distinction to the term use with respect to the four defined characteristics. They find that the characteristics are difficult to objectively measure. This is illustrated with a research data life cycle in Figure 2.1, which considers the time and purpose characteristics of reuse. Team A collects data to investigate Research Question A. Team A publishes the data set in a repository shortly after acquisition and before finishing the accompanying Publication A. In the meantime, Team B finds the data on the repository and uses it solve Research Question B, and they publish Publication B before Team A publishes Publication A. In this case, the first measurable *use* of the data in time would be Publication B. Yet, Publication B serves a different purpose than that of data acquisition, and it could thus also be considered a reuse. Sandt et al. (2019) conclude that a strict distinction between reuse and use cannot be made and propose "accepting a more dynamic, chaotic and less categorizable research world" and to not differentiate between *reuse* and *use*. Instead they propose the statement "(re)use of research resources equals their usage."

Data *reusability*, following Faniel and Jacobsen (2010), mainly concerns the relevance of the data for a specific context as well as the understandability and trustworthiness of the data. This is in alignment with Nusser et al. (2021), who describe data reusability as the "ability of a new user to find, understand, evaluate, and appropriately incorporate the shared data in their analyses." If these conditions are met, data reuse is facilitated. Data reusability is defined by Thanos (2017) as "the ease of using data" by reusers from different communities than the creator's.

Definitions and concepts for reuse and reusability of software

The concept of *software reuse* was first presented during the NATO Software Engineering Conference in 1968 by McIlroy (1968), who proposed the standardization of software components in form of high quality routines so that the same routine can be used by different machines and users. Software reuse has been inherent to software engineering ever since and has been followed by many definitions and concepts. The *IEEE Standard for Information Technology–System and Software Life Cycle Processes–Reuse Processes* defines software reuse as "capitalization on existing software and systems to create new products" (IEEE, 2010). For software reusability, the standard provides two definitions.

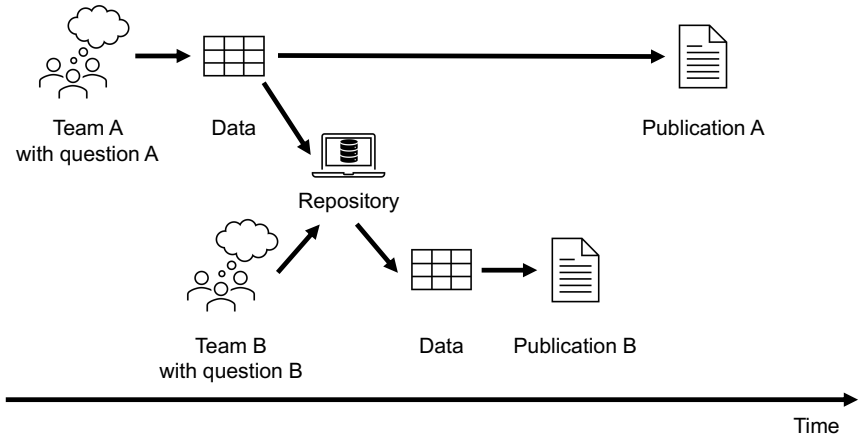


FIGURE 2.1: Example of a data life cycle with focus on the time and purpose of reuse. Team A collected data to solve question A and published the data on a repository. Team B finds the data in a repository and uses it to solve question B. Publication B is finished before Publication A. The figure is adjusted from Sandt et al. (2019).

1. "The degree to which an asset can be used in more than one software system or in building other assets."

2. "In a reuse library, the characteristics of an asset that make it easy to use in different contexts, software systems, or in building different assets."

Further definitions of software reuse and software reusability gathered from the literature are provided in Table 2.2. All listed publications define software reuse as using existing software (components) to create new software, while IEEE (2010) does not specify the product as software but refers to it as "asset." Some of the publications (Frakes and Terry, 1996; Frakes and Kang, 2005) also include the reuse of software knowledge.

Regarding software reusability, Frakes and Terry (1996) and Gill and Sikka (2011) agree with the first definition from IEEE (2010) by referring to a *degree* of reuse. Such degree could be interpreted as the flexibility and adaptability of a software or the portion of an existing software that suitable for combination with other software. Frakes and Kang (2005) define reusability of software as a property related to the probability of its reuse, and the second definition from IEEE (2010) defines reusability as characteristics of software libraries that contribute to the library's ease of reuse.

Several approaches for software reuse have been developed over the last 50 years and are common practice in today's software development. Sommerville (2016) describes four levels of software reuse that give some general ideas of the strategies in detail. Please refer to Sommerville (2016) for more details.

1. **Abstraction level:** Reuse of the concepts and architectures of a software without direct reuse of software.

2. **Object level:** Reuse of software objects from libraries instead of coding the respective objects, methods, and functions.

3. **Component level:** Reuse of components of an existing application, which often involves adjustment and extension of the component for the new application.

4. **System level:** Reuse of a whole application system for which the original system may be integrated into another system without changes or adjusted to serve a different application.

In the case of modeling software, reuse on the abstraction the level could refer to following the same software compartmentalization of another modeling software, for example, separating a solver from the assignment of initial conditions in different files. Reuse at the object level occurs when programming libraries such as numpy (Harris et al., 2020) or pandas (Pandas team, 2020) are used. Component level reuse would then be the reuse of a software component that represents the computation of a particular process in another modeling software. Lastly, system-level reuse would be the reuse of an entire model, for example, when coupling it with another one. To enable these levels of software reuse, technological solutions such as programming libraries need to be available and coding practices, such as modular development must be followed.

Differences between the concepts for data and software

Data reuse definitions focus on a delineation to data use with regards to the person who carries out the reuse and the purpose of reuse. Both may have to be different from the circumstances of data collection. Neither a specific person nor a new purpose are specifically considered in definitions

Table 2.2: Definitions for software reuse and software reusability from the literature. The entry by Lanergan and Grasso (1984) marked with an asterisk has been interpreted and is not a direct citation.

Publication	**Software reuse**	**Software reusability**
Lanergan and Grasso (1984)*	use of standardized previously redundant code modules	-
Krueger (1992)	process of creating software systems from existing software rather than building software systems from scratch	-
Frakes and Terry (1996)	use of existing software artifacts or knowledge to create new software	degree to which a thing can be reused
Frakes and Kang (2005)	use of existing software or software knowledge to construct new software	property of a software asset that indicates its probability of reuse
Gill and Sikka (2011)	use of existing artifacts to create new software	degree to which the artifacts can be reused
Imoize et al. (2019)	act involving the recycling of components and entities in the process of software development	-

for software reuse. Furthermore, the delineation of use and reuse with respect to software is not discussed in the literature. Data reusability concepts include the understandability of data and the ability to evaluate their trustworthiness and suitability with respect to a specific context, which increase in case reusability is fulfilled. Such characteristics seem not be an integral part of software reusability.

Software reuse usually refers to compatibility and interoperability with other software, so it includes integration and combination aspects with the goal of creating new software. While combining and integrating multiple data sets may be part of the data reuse process, the goal of data reuse is usually to solve a research question, not to create new data. However, a combined data set may be a by-product. Furthermore, software reusability considers the degree to which software and its components can be reused. If the degree is interpreted as the amount of software that can be reused, then data reusability lacks an analogue property as it does not consider how many of the data points in a data set are included for reuse.

Conventions for this thesis

In the remainder of this thesis, reuse and reusability of research products will cover both data and software. Therefore, definition are needed that apply to both.

The term reuse is used to highlight the use of data and software by people other than the creators. These reusers are often research product agnostic as they may come from a different domain and have a other methodological skills than the creators of a particular research product. It is therefore important to specifically address the challenges faced by agnostic reusers, as these are different from those faced by the original creators. In this definition, reuse is independent of time and purpose; these may be the same as or different from those of the original product creation. Furthermore, I define reuse scenarios, which were introduced in Sect. 1.2, as the context of reuse. These scenarios are purpose-directed and each has specific requirements.

Reusability takes into account the ability to find, understand, evaluate, and integrate data and software. As each of these criteria facilitates reuse, reusability is referred to as the ease with which a research product can be reused. To some extent, reusability can be assessed on the basis of the resource alone, such as its findability and understandability. In the context of reuse scenarios, reusability also depends on a product's ability to integrate with other data or software following the respective scenarios' requirements.

2.1.2 Reproducibility versus reusability

Reusability and reproducibility are individual properties of research products, but they are often mentioned together (Vasilevsky et al., 2017; Wagner et al., 2024). In this section, I summarize the concept of reproducibility from the literature and place it in the context of reusability.

Reproducibility has received increased attention at least since the discussion of a reproducibility crisis. Baker (2016) revealed that more than 50% of scientists have experienced failed attempts to reproduce their own experiments, and more than 70% have failed to reproduce other researchers' experiments. However, the use of the term reproducibility is inconsistent across domains and methodological expertise, ranging from computational science to experimental science (Barba, 2018; Plesser, 2018). Reproducibility can be used interchangeably with *replicability*, or the two terms are used with opposite meanings.

Barba (2018) classifies three different cases (A, B1, and B2) of definitions for reproducibility and replicability from the existing literature. A refers to an interchangeable use of the terms; B1 and B2 distinguish between the terms. B1 defines reproducibility as obtaining the same results for a research question by using the same data and methods from a previous study, while it defines replicability as obtaining the same findings or consistent results for the same research question but using different data and possibly a different method. For B2, the definitions of reusability and replicability are reversed with respect to B1. This thesis follows the definitions of B1, which is consistent with the use in computational science and defines *replicability* of scientific findings as the *gold standard* (Peng, 2011).

For Thanos (2017), the potential for reproducibility derives from the origin of the data intended to be reproduced. Following his definition, observational data can neither be reproduced nor recollected due to continuously changing conditions of the observational environment. Experimental data are reproducible but may not be *accurately* reproduced as it may be impossible to reestablish the same experiment conditions. Computational data should be naturally reproducible as they derive from computer programs that can be run again.

Reproducibility for computational research

The original idea of reproducible research for computational science dates back to the 1990s when Claerbout and Karrenbach (1992) began to demand the publication of code associated with computational results in the field of seismic exploration. Specifically, they demonstrated how data and code accompanying an article in form of an electronic document on a CD-ROM could be used to reproduce the results, such as a figure from an article.

Reproducibility in relation to computational results is also defined in the report on *Reproducibility and Replicability in Science* by the US-American National Academies of Sciences, Engineering and Medicine (NASEM, 2019). Following the definition of the report, "Reproducibility is obtaining consistent results using the same input data; computational steps, methods, and code; and conditions of analysis." Accordingly, reproducibility requires transparent documentation of the method and provision of all necessary code, data, and potentially other materials necessary to reproduce the results.

Freire and Chirigati (2018) distinguish between *accurate* and *approximate* reproducibility of computational data. Thereby, exact reproducibility is achieved by following the same steps, using the

same data and operating in the same environment, i.e., operating system and hardware; it generates the exact same results. Approximate reproducibility is achieved when a similar result with respect to the original is obtained by following and using the same or different data, computational steps, and environment. For both types of reproducibility, consistency of the results, so that the same conclusion can be drawn, is strictly required. The definition of approximate reproducibility is necessary because the specific hardware used for the original results may not be available, or an adjustment of the sequence of steps may be necessary.

Provision of original software and data is necessary together with documentation and specification of the respective hardware is required to make the related result reproducible (Gruenpeter et al., 2021). However, reproducibility of a result does neither say something about the credibility of the result nor about the quality of the products themselves, i.e., the reusability of the software or data (Peng, 2011; NCS, 2021).

The quality aspect of reusability

Wagner et al. (2024) define reusability of software in the context of models for deep-learning as software that is "well-documented and easy to integrate into existing workflows and models should be robust toward noise and generalizable toward data from different sources." Accordingly, obtaining reusability is more effort than providing reproducibility. This definition is similar to that defined by the editors of the journal *Nature Computational Science* (NCS, 2021). They highlight the need of code to be reusable beside providing reproducibility of results. Reusable code can accelerate scientific progress when other researchers can use it for the same purpose, evaluate the same software methodology with new data, or use it and build on it to solve new questions.

In practice, reusable code is well documented, including installation, and software and hardware requirements are clearly stated. The code can be executed directly with defined instructions and does not require the adjustment of path variables. In addition, optional settings of the code, such as different parametrizations to choose from, are distinguishable, for example, by including them as input to the main function that executes the code (NCS, 2021).

The advantages of reusable software have been demonstrated by the journal *Nature Machine Intelligence* (NMI), which provides the article type Reusability report (Machine Intelligence, 2022). Reusability reports are based on code of a previously accepted or published article in NMI. This code is given to a group of researchers uninvolved in the code development. The group gets the task to build on the code, for instance, by running it with new data, improve, or extend it.

Conventions for this thesis

Reproducibility for computational research begins with the simple provision of the code and data that lead to a result, which is consistent with the one of the original publication. Reproducibility creates transparency. Reproducibility is furthermore improved by detailed documentation of the code, and it also benefits from easy installation and quick execution of the code. However, both

these properties are also characteristics of reusability. Reproducibility thus supports reusability, but preparing code in a reusable manner requires much more effort and generates a higher level of transparency than reproducibility. Reusability is thus directly related to the quality of the software.

But how do these concepts translate to non-computational data that are not based on a computer program that can be executed infinite times and still provides the exact same result? Reproducibility of the experimental data can only be obtained if the measurements are holistically described including environmental conditions, measurement devices, and preparation of the sample so that the experiment can in principle be set up again. This kind of transparency required for reproducibility at the same time supports the reusability of the data product by providing the contextual information needed to understand the data. Reusability of data can thus also be understood as a quality characteristics of the provided metadata. If the quality of the metadata is low, the data is hardly reusable.

In summary, reproducibility increases the transparency of the workflow, which was followed to create a specific product, but it is not a quality characteristic of the generated data or software. Reusability, on the contrary, is directly related to the quality of the research product. Reproducibility can exist without reusability, but a highly reusable data product is usually sufficiently well described to be reproducible.

2.2 The FAIR Principles for optimizing the reuse of research products

Nowadays, the FAIR Principles are widely recognized as guidelines for good stewardship of research products by enabling their findability, accessibility, interoperability, and reusability. The number of data sets, software, and articles using the FAIR acronym to describe research products, methods, standards, and tools is growing. Google Scholar registers 719 resources with the term "FAIR Principles" in their title or abstract for the period 2016 to 2017 and 8,150 resources for the period 2022 to 2023. In geoscience, FAIR has been used to describe data such as *Open and FAIR mineralogy data* (Ma et al., 2023) and repositories such as the *FAIR National Environmental Data Repository for Earth Observation Open Science* (Giuliani et al., 2021). In cryospheric sciences, there exist publications on workflows such as for *FAIR snow mapping* (Alnaim and Sun, 2022) and on data such as *FAIR data of Antarctic bed maps* (Frémand et al., 2023). The implications of the FAIR Principles for the management of polar data have recently been discussed (Vey et al., 2024).

The acronym FAIR is also used to describe the general FAIR development of resources. The term *FAIRification* refers to actions that make research products FAIR (Jacobsen et al., 2020a) or more FAIR compliant (Luna et al., 2022). FAIRification should furthermore increase the *FAIRness* of a product, which represents the FAIR status of a research product. FAIRness improves as the resource complies with more and more FAIR Principles (Jacobsen et al., 2020a). Several tools have been developed to assess FAIRness of research products (Wilkinson et al., 2019; Gaignard et al.,

2023; Devaraju and Huber, 2024). These tools use metrics to measure compliance of a research product with specific principles.

Adherence to the principles should maximize the findability and ultimately the reuse potential of published research products by humans and computers with a special focus on enabling their machine-actionability (Mons et al., 2017; Wilkinson et al., 2018). In fact, many products claim to be FAIR, even though they lack full machine readability and interoperability. Both properties are indispensable for FAIR (Wilkinson et al., 2018; Wyborn, 2023). Within this dynamic FAIR environment, misinterpretations cause a risk to deviate from the original vision (Mons et al., 2017) and risk incompatibility of technological solutions used for FAIR-compliance (Jacobsen et al., 2020b). In the following, I will first introduce the FAIR Principles for Data and Research Software, and then continue with FAIR metrics and the FAIR assessment of data and software.

2.2.1 The FAIR Principles for research data and software

The *FAIR Guiding Principles for Scientific Data Management and Stewardship*, in the following referred to as FAIR Data Principles, were published in 2016. The formal publication followed their first proposal during a Lorentz workshop in Leiden, the Netherlands, in 2014 (Wilkinson et al., 2016). The FAIR Data Principles are the combined result of several initiatives aimed at promoting data management and sharing practices in the scientific community with the goal to enable machine actionability (Jacobsen et al., 2020b; Mons et al., 2020).

An increase in sharing of data has led to a diversification of sharing options and practices. As a result, deposited data is often not harmonized and does not integrate with existing data. The FAIR Principles are intended to address this diversification and improve the reuse potential of resources for both humans and machines (Wilkinson et al., 2016). Once published, FAIR quickly gained momentum. Institutions such as the G20 (G20, 2016) and the European Commission (EC DGRI, 2016) started to incorporate the principles in their policies, and the scientific community began with the implementation (Mons et al., 2017). Although FAIR principles are often mentioned together with Open Science, FAIR is not equal to open. Accordingly, data access may require authentication or authorization and still be FAIR-compliant.

Although only data and their metadata are mentioned in the FAIR Data Principles, the concept of *Findable, Accessible, Interoperable and Reusable* was not limited to data from the beginning. It should also apply to other digital assets such as tools and software (Wilkinson et al., 2016; Wilkinson et al., 2018). A better understanding of the needs for FAIR research software by the scientific community has led to the publication of the *FAIR Principles for research software* (FAIR4RS Principles). *Research software* includes source code files, algorithms, scripts, computational workflows, and executables that were created during the research process or for a research purpose. Software components, such as operating systems, libraries, dependencies, packages, and scripts that are used for research but were not created during or with a clear research intent should be considered software in research, not research software (Gruenpeter et al., 2021). This differentiation may vary between disciplines.

Table 2.3 (a) lists the FAIR principles for data (Wilkinson et al., 2016) and (b) for research software (Barker et al., 2022). There are 15 FAIR principles for data and 17 for research software. I will refer to the superordinate *Findable, Accessible, Interoperable and Reusable* as the main principles and the subordinate principles such as F1 or R1.2 as sub-principles. The principles are not only addressed to the resource, i.e., data and software, but also to its metadata. The reader should note that this separate consideration is necessary since metadata may be accessible independent of the availability of the resource it describes. In the following, I will use the terms resources or products, when referring to both data and software. As the FAIR principles themselves do not provide context on their respective interpretation the following description of the principles is combined with more detailed explanation for the FAIR Data Principles by GOFAIR (2023).

Findability

Findability of the resource and its metadata should generally be possible for both humans and machines. Accordingly, metadata of a resource has to be machine-readable (GOFAIR, 2023). F1 requires the provision of globally unique and persistent identifiers. These identifiers are links and fulfill important functions. First, they remove ambiguity for both understanding and referencing the resource. Understanding the resource is facilitated by providing identifiers for metadata such as specific measurements, concepts, and terminology that guide the machine or human to the corresponding definitions. Referencing of the resource has to be unique, such as Globally Unique Identifiers (GUID), so that nobody else can use the same identifier for a different resource. Second, identifiers should ensure persistence, such as Persistent Identifiers (PID), so that the resource remains reachable via the link assigned to the identifier for long time frames. F2 calls for a rich and comprehensive description of the resource in the metadata. This is important as the resource's findability depends, for instance, on keywords provided in the metadata. As metadata and resource often come in separate files, metadata should provide the identifier of the resource, i.e., the address via which the resource is accessed (F3). Furthermore, the resource itself has to be registered so that it can be found (F4). This is often ensured by repositories that provide a DOI registry service. The same applies to the content of the metadata such as used terminology, which has to be registered and indexed, and in the best case be FAIR itself. Software furthermore requires that components and versions of the software each to be assigned with individual persistent identifiers (FAIR4RS F1.1 and F1.2).

Accessibility

The resource and its metadata should be accessible via their identifiers (A1) and through standardized communication protocols that are free, open and universal such as HTTP (A1.1). In case the resource is not open, authentication and authorization for access has to be enabled via the protocol (A1.2). Furthermore, metadata has to remain persistently available and accessible also if the described resource is removed (A2).

TABLE 2.3: Principles for FAIR data (a) and FAIR research software (b). (Meta)data refers to data and metadata. The introducing sentences in bold in (a) are based on GOFAIR (2023).

FAIR (a) (Wilkinson et al., 2016)	**FAIR4RS (b)** (Barker et al., 2022)
Findable	
Metadata and data should be easy to find for both humans and computer.	**Software, and its associated metadata, is easy for both humans and machines to find.**
F1 (Meta)data are assigned a globally unique and persistent identifier.	**F1** Software is assigned a globally unique and persistent identifier. **F1.1** Components of the software representing levels of granularity are assigned distinct identifier. **F1.2** Different versions of the software are assigned distinct identifiers.
F2 Data are described with rich metadata.	**F2** Software is described with rich metadata.
F3 Metadata clearly and explicitly include the identifier of the data it describes.	**F3** Metadata clearly and explicitly include the identifier of the software they describe
F4 (Meta)data are registered or indexed in a searchable resource.	**F4** Metadata are FAIR, searchable and indexable.
Accessible	
The user needs to know how to access the data, possibly including authentication and authorization.	**Software, and its metadata, is retrievable via standardized protocols.**
A1 (Meta)data are retrievable by their identifier using a standardized communications protocol. **A1.1** The protocol is open, free, and universally implementable. **A1.2** The protocol allows for an authentication and authorization procedure, where necessary.	**A1** Software is retrievable by its identifier using a standardized communications protocol. **A1.1** The protocol is open, free, and universally implementable. **A1.2** The protocol allows for an authentication and authorization procedure, where necessary.
A2 Metadata are accessible, even when the data are no longer available.	**A2** Metadata are accessible, even when the software is no longer available.
Interoperable	
The data needs to be integrated with other data and interoperated with applications or workflows for analysis, storage, and processing.	**Software interoperates with other software by exchanging data and/or metadata, and/or through interaction via application programming interfaces (APIs), described through standards.**
I1 (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	**I1** Software reads, writes and exchanges data in a way that meets domain-relevant community standards.
I2 (Meta)data use vocabularies that follow FAIR principles.	**I2** Software includes qualified references to other resources.
I3 (Meta)data include qualified references to other (meta)data.	
Reusable	
Metadata and data should be well-described so that they can be replicated and/or combined in different settings.	**Software is both usable (can be executed) and reusable (can be understood, modified, built upon, or incorporated into other software).**
R1 Meta(data) are richly described with a plurality of accurate and relevant attributes. **R1.1** (Meta)data are released with a clear and accessible data usage license. **R1.2** (Meta)data are associated with detailed provenance. **R1.3** (Meta)data meet domain-relevant community standards.	**R1** Software is described with a plurality of accurate and relevant attributes. **R1.1** Software is given a clear and accessible license. **R1.2** Software is associated with detailed provenance.
	R2 Software includes qualified references to other software.
	R3 Software meets domain-relevant community standards.

Interoperability

Data interoperability ensures seamless integration with other data and its interaction with and use in applications or existing analysis workflows. Therefore, data and metadata have to represent knowledge in a language that refers to defined terminologies that are themselves digitally accessible. Furthermore, this knowledge has to be represented in a format understandable to machines (FAIR I1). These terminologies must themselves be FAIR (FAIR I2), i.e., equipped with identifiers, documented for both humans and machines, and, at best, defined and controlled by the respective community (GOFAIR, 2023). The use of uniquely identifiable terminologies is the basis for later integration with other data. For software, interoperability requires standardized interfaces that allow software to interoperate with other software through Application Programming Interfaces (APIs). More specifically software has to apply community standards for reading, writing, and exchanging data (FAIR4RS I1). Furthermore, data and its metadata as well as software should reference other resources (FAIR I3, FAIR4RS I2) providing contextual information (GOFAIR, 2023).

Reusability

Reusability is generally supported by a detailed, correct, and relevant description of the resource and its metadata (R1). Based on this description, data should be replicable and combinable with other resources in a meaningful way, and software should be executable and reusable. Software reuse can involve modification, further development or incorporation into other software. Therefore, this description should contain contextual information including the purpose, limitations, and processing stage provided by the creator, and it should neglect assumptions on potential reuse (GOFAIR, 2023). Software is often based on other software, which should be referenced (FAIR4RS R2). Additionally, the legal use constraints of the resource should be clarified with a license (R1.1), and provenance of the resource should be transparent (R1.2). Resources should be provided in a way that facilitates reuse and therefore comply with community standards (FAIR R1.3, FAIR4RS R2). Community standards should ensure that resources of the same type are similar. The standards can include specific data types, formats, and templates as well as the use of specific vocabularies and the general organization of resources (GOFAIR, 2023). There may be a favored specific tabular structure for data or a folder structure for software.

Machine-actionability

The FAIR Principles describe characteristics of digital resources that make them reusable, not only by humans, but especially by computers. The FAIR Principles should allow computers to interact with digital resources in order to autonomously decide on their suitability for a given task, and to ultimately perform the task (Mons et al., 2017; Jacobsen et al., 2020b). In the future, computational agents could perform tasks like the autonomous discovery and analysis of data (Wilkinson et al., 2016). Machine-actionability is thought to potentially lead to a speed up of scientific discovery as researchers will spend less time on time-intensive preparation and combination of resources from different sources. Furthermore, machines may overtake tasks exceeding the scales of human

capabilities (Wilkinson et al., 2016). However, this FAIR state has not been reached yet and may never be reached. Mons et al. (2020) describe FAIR as "a community journey" toward "machine-assisted research and innovation."

The vision proposed by FAIR describes a continuum of machine-enabled research products, linked with each other and to tools and services. This continuum was later referred to as the *Internet of FAIR data and services* (IFDS) (Mons et al., 2017; Jacobsen et al., 2020b). The IFDS's future state is described in the report of the European Commission on *Realizing the European open science cloud* (DGRI, 2016). It should take the shape of a "federated, globally accessible environment where researchers, innovators, companies and citizens can publish, find and re-use each other's data and tools for research, innovation and educational purposes." The European Open Science Cloud is a project of the European Commission to facilitate data-driven research, and it is designed as European contribution to the IFDS.

Aspirational character of the principles

From design, the interpretation and implementation of the FAIR principles is left to discussion by the science community as the principles are aspirational. The principles do not suggest specific technological solutions or concrete examples for their realization, which has stimulated a wide range of interpretations for the implementation of FAIR principles (Jacobsen et al., 2020b), which also includes misinterpretations of the general intention of FAIR. Misinterpretations have been clarified by Mons et al. (2017). They emphasized what FAIR is not: it is not a standard or a technological solution, it is not only directed to serve humans but mainly machines, and it is not equal to open.

While the FAIR principles are domain-independent, their implementation takes place within domains and should be tailored to respective needs and constraints (Wilkinson et al., 2016; Wilkinson et al., 2018). The resulting diversity of implementations is intended, as the FAIR principles are permissive and do not suggest specific tools or standards. Mons et al. (2017) regard FAIR as *spectrum*, which emphasizes the suitability of FAIR-compliant tools and standards depends on the application and domain.

The FAIR vision has evolved over the last decade, but much development remains to be done, especially in terms of domain-specific implementation of the principles. Such developments require community efforts, which is a time-intensive process and needs clear responsibilities (David et al., 2020; Kinkade and Shepherd, 2021). Furthermore, researchers have to be trained a new skill set for the generation of FAIR-complying research resources (Mons et al., 2020). Another obstacle for the implementation of FAIR principles, is the lack of a measure to check compliance with the principles, which stems from their aspirational character.

2.2.2 FAIR metrics and FAIRness evaluation

The FAIR Principles deliberately do not prescribe tools or properties to reach ultimate FAIRness for digital resources. Consequently, research communities have developed individual interpretations

for the evaluation of FAIRness (Cox and Yu, 2017; Dunning et al., 2017). There is a need for an objective and transparent assessment of the FAIRness of digital resources. FAIRness assessment is of interest to data collectors who want to know the FAIR-compliance of published resources and how its FAIRness can be improved; it is also relevant to journals, repositories, and founders, who have incorporated FAIR into their policies and need a means to monitor and verify the actual compliance of research products with the principles (Wilkinson et al., 2018).

Wilkinson et al. (2018) and Wilkinson et al. (2019) took the request of the research community to measure FAIRness into account and proposed the development of FAIR metrics for FAIRness assessment. In general, a FAIR metric describes a property of a digital resource that is derived from a specific FAIR principle. FAIR metrics are measurable. Fulfillment of this property implies compliance with the corresponding FAIR principle. The process of checking a resource's consistency with a metric is referred to as FAIRness assessment, and can be realized through compliance tests incorporated in FAIRness assessment tool. FAIR metrics propose a remedy to the non-prescriptive character of the FAIR principles as they concretize the interpretation of the principles and implementation choices. Wilkinson et al. (2018) and Wilkinson et al. (2019) emphasize that FAIRness should not be understood as a competition.

Design of FAIR metrics and FAIRness assessment tools

Wilkinson et al. (2018) and Wilkinson et al. (2019) described the general design principles for FAIR metrics and FAIRness assessment tools with a focus on enabling machines to perform FAIRness evaluation autonomously. The design principles are summarized in the following.

FAIR metrics should be clear to anyone, reflect all FAIR principles, and apply to all types of digital resources. Accordingly, anyone should be able to check the FAIRness of any digital object with FAIRness assessment tool and without any extra knowledge. Each FAIR metric should be discriminating by representing at best not more than one principle. FAIR metrics may vary from domain to domain since different communities require different standards and tools. Thus, universal metrics need to be complemented by community-specific metrics, and FAIRness evaluation tools need to take these community-specific metrics, including their specific recommendations, into account when evaluating a digital resource from the respective domain. The FAIR metrics, assessment tools and their results have to comply with FAIR themselves, which means they have to be understandable by machines. As FAIR metrics and assessment tools may change with developing technological solution or community needs, they have to be provided in a way so they can be reviewed and versioned. Measures used to check the compliance of a resource with a FAIR metric must be computable to support autonomous and objective FAIRness assessment of resources. Machine-actionability of FAIR metrics and FAIRness assessment is an integral part of the FAIR vision.

In addition, FAIR metrics should be designed so that they allow not only the assessment of overall compliance, but also the evaluation of the degree of compliance for each metric. Therefore, the output of the FAIRness assessment should be a FAIRness score accompanied by a report specifying the score per metric. This report should document the inputs and outputs for each FAIR metric

TABLE 2.4: (a) FAIR metrics from FAIRsFAIR used in F-UJI and (b) FAIR metrics defined by FAIR-Checker tool

	(a) **FAIRsFAIR Metrics v0.5** (Devaraju et al., 2020)	**(b)** **FAIR-Checker metrics** (Gaignard et al., 2023)
	Findable	
F1	**FsF-F1-01D:** Data is assigned a GUID **FsF-F1-02D:** Data is assigned a persistent identifier	**F1A:** Unique IDs **F1B:** Identifier persistence
F2	**FsF-F2-01M:** Metadata includes descriptive core elements (creator, title, data identifier, publisher, publication date, summary and keywords) to support data findability	**F2A:** Structured metadata **F2B:** Shared vocabularies for metadata
F3	**FsF-F3-01M:** Metadata includes the identifier of the data it describes	
F4	**FsF-F4-01M:** Metadata is offered in such a way that it can be retrieved by machines	
	Accessible	
A1	**FsF-A1-01M:** Metadata contains access level and access conditions of the data **FsF-A1-02D:** Metadata is accessible through a standardized communication protocol **FsF-A1-03M:** Data is accessible through a standardized communication protocol	
A1.1		**A1.1:** Open resolution protocol
A1.2		**A1.2:** Authorization procedure or access rights
A2	**FsF-A2-01M:** Metadata remains available, even if the data is no longer available	
	Interoperable	
I1	**FsF-I1-01M:** Metadata is represented using a formal knowledge representation language	**I1:** Machine readable format
I2	**FsF-I2-01M:** Metadata uses semantic resources	**I2:** Use shared ontologies
I3	**FsF-I3-01M:** Metadata includes links between the data and its related entities	**I3:** External links
	Reusable	
R1	**FsF-R1-01MD:** Metadata specifies the content of the data	
R1.1	**FsF-R1.1-01M:** Metadata includes license information under which data can be reused	**R1.1:** Metadata includes license
R1.2	**FsF-R1.2-01M:** Metadata includes provenance information about data creation or generation	**R1.2:** Metadata includes provenance
R1.3	**FsF-R1.3-01M:** Metadata follows a standard recommended by the target research community **FsF-R1.3-02D:** Data is available in a file format recommended by the target research community	**R1.3:** Community standards

test and provide clear recommendations for potential improvements of FAIRness. In addition, the progress of a resource's FAIR compliance over time should be demonstrable with the FAIRness assessment tool. FAIR metrics should be realistic so resources can adopt respective measures based on recommendations from FAIRness assessment. FAIRness assessment must measure the FAIRness outcome of a resource and not the intended FAIRness of the resource as it could be suggested in a data management plan.

FAIR metrics for research data and software

First, Wilkinson et al. (2018) suggested 14 exemplar metrics, and then Wilkinson et al. (2019) proposed 15 FAIR Maturity Indicators. Since then more FAIR metrics based on the FAIR Data Principles have been proposed from individual groups. There are, for instance, the FAIR data maturity model indicators from the Research Data Alliance (RDA) (Bahim et al., 2020b), the FAIRsFAIR Data Object Assessment Metrics, referred to as FAIRsFAIR metrics in the following (Devaraju et al., 2020), and the FAIR Metrics for the European Open Science Cloud (EC DGRI EOSC, 2021). All of the named metrics focus on digital resources in form of data sets alone.

The FAIRsFAIR metrics are listed per FAIR Data Principle in Table 2.4 (a). Furthermore, Table 2.4 (b) lists the FAIR-Checker metrics used by the FAIR-Checker FAIRness assessment tool (Gaignard et al., 2023). The FAIR-Checker metrics are only defined within the tool and are intended to apply to any digital object. In total, FAIRsFAIR provides 17 metrics, while FAIR-Checker provides 12. FAIRsFAIR metrics indicate whether they apply to data or its metadata by adding a *D* or an *M* to the name of the metric. Each metric is furthermore described together with specific requirements for the compliance test in an accompanying document (Devaraju et al., 2020) as requested by (Wilkinson et al., 2018). Such documentation is not provided for the metrics used by the FAIR-Checker tool.

Specifically for software, the FAIR-IMPACT project developed the Metrics for automated FAIR software assessment, and they uniquely apply to software (FAIR-IMPACT, n.d.). They consist of 17 metrics, directly refer to the FAIR4RS principles, and are formulated as questions. For reference, the metrics are provided in Table A.1 in the Appendix.

2.2.3 FAIRness assessment tools for research data and software

Based on the many FAIR metrics, several FAIRness assessment tools have been developed. Candela et al. (2024) reviewed 20 FAIRness assessment tools. The types of resources to be evaluated with the tools vary between data sets (9), repositories (3), semantic artifacts (2), software (1), and several types of digital resources (5). The level of automation of the assessment varies between manual (7), automatic (11), and a combination of both (2). Manual assessment often involves some type of checklist or survey such as is the case for the FAIR data maturity model indicators from the RDA (Bahim et al., 2020b) that are available as a survey on a webpage (Bahim et al., 2020a).

The majority of the tools reviewed by Candela et al. (2024) use universally applicable metrics and neither check discipline nor community-specific metrics.

Lang et al. (2023) evaluated FAIRness assessment tools and discriminate them into the four categories summarized below.

- **Regular list tools** are static checklists of FAIRness criteria without any interactivity. They can be used to manually check the FAIRness of a given resource. These lists may not provide sufficient explanation or specific examples for the criteria, so additional knowledge is required to understand the criteria. Static lists may be versioned, and they can be integrated into automatic tools
- **Improved survey tools** are lists of questions for which predefined answers can be selected, for example through a drop down menu. The evaluation process is manual and time-intensive. The result may include specific suggestions for improving FAIRness.
- **Fully configurable tools** allow the evaluation of a resources based on self defined metrics or predefined registered metrics. Self-configuration takes time and requires technical knowledge.
- **Automatic tools** check FAIRness based on the the machine-readable metadata assigned to a resource. Evaluation is quick by pasting the resource's DOI on a website. The result includes a FAIRness score and an overview on failed and passed technical tests, which require technical knowledge to be fully understood.

The future direction should be toward providing automatic tools "as-a-service," i.e., in the form of a ready to use web app without the need for local installation (Candela et al., 2024). Therefore, some automatic tools for FAIRness assessment will be described in the following.

Automated FAIRness assessment tools

According to Wilkinson et al. (2019), automatic FAIRness assessment has three main components:

1. A set of FAIR metrics that specify a property of a resource that is measurable and therefore can be evaluated automatically.
2. Compliance tests representing the technical realization of the metrics. They test the compliance of the resource with the metric.
3. A web service that can be accessed to automatically check the compliance of a resource with a set of registered FAIR metrics. The service also generates a result for the FAIRness of the resource together with a report of the evaluation.

Two automatic, as-a-service tools for FAIRness assessment are the F-UJI (Devaraju and Huber, 2024) and the FAIR-Checker (Gaignard et al., 2023). F-UJI uses the FAIRsFAIR metrics listed in

Table 2.5: General properties of the automated FAIRness assessment tools F-UJI and FAIR-Checker gathered from the service websites and GitHub documentations.

(a) **F-UJI**	(b) **FAIR-Checker**
Service website	
https://www.f-uji.net/	https://fair-checker.france-bioinformatique.fr/check
Type of resource and identifier used for assessment	
Data set PID/URL of data set landing page	Any digital resource URL/DOI of the digital resource
Metadata included in assessment	
Aggregated metadata; this includes metadata embedded in the data (landing) page, metadata retrieved from a PID provider (e.g., DataCite) and other services (e.g., RE3).	Metadata extracted from landing page
Result presentation	
1) Total score in percent 2) Total FAIR level (incomplete, initial, moderate, advanced) 3) Score per main FAIR principle 4) FAIR level per main FAIR principle 5) Score per metric 6) FAIR level per metric 7) Score per test	1) Total score in percent 2) Score in percent per main FAIR principle 3) Score per metric, which can be 0/2, 1/2 or 2/2
Report	
- Overview on passed and failed tests - Application log with debug messages - List of tests for each metric - Output of passed tests - Available as json-file	- Description of the metric - Application log - Recommendations for improvement - Available as csv-file
Number of FAIR metrics included in the assessment: total and per main FAIR principle	
16 F:5, A:3, I:3, R:5	12 F:4, A:2, I:3, R:3

Table 2.4 (a). The FAIR-Checker tool does not refer to a specific set of FAIR metrics. Instead it refers directly to the FAIR Data Principles based on which the 12 metrics listed in Table 2.4 (b) are derived. F-UJI is valid for the assessment of data sets only, and the FAIR-Checker (Gaignard et al., 2023) can be used for any digital object. Several properties of both tools are summarized in Table 2.5. Automated FAIRness assessment tools use the available metadata that computers can extract from a resource. Both tools aggregate metadata from the landing page. F-UJI furthermore uses metadata assigned to the PID of the resource and from services such the Registry of Research Data Repositories (RE3). Concerning the results, F-UJI provides 7 individual types of results varying from the level that considers the entire resource to the performance of individual tests. FAIR-Checker has 3 different types of results not going deeper than the metric level. The distribution of the metrics per main FAIR data principle shows that the amount of metrics per principle

follows different distributions for each tool, which is a general observation also for other FAIRness assessment tools (Candela et al., 2024). This is also highlighted in Wilkinson et al. (2022) where the tools FAIR Evaluator and F-UJI are compared with each other revealing that FAIR Evaluator includes 7 compliance tests for interoperability and 2 for reusability while it is 4 and 10 for F-UJI.

The general process of an automatic, as-a-service FAIRness assessment based on the tool F-UJI including its relation to the FAIRsFAIR metrics and the FAIR principles is illustrated in Fig. 2.2. FAIR principles such as F1, which is defined as "(Meta)data are assigned globally unique and persistent identifier," are interpreted in terms of FAIR metrics. Several FAIR metrics can derive from one principle. FsF-F1-02D is one metric derived from F1, which defines that data has to be assigned a PID. In the next step metrics have to be assigned compliance tests, which clarify of how the metric can be measured. In the example in Fig. 2.2, two compliance tests are defined for the metric FsF-F1-02D. Next, these compliance tests have to be technically implemented in software so they can be executed in practice. These tests then become part of the FAIRness evaluation

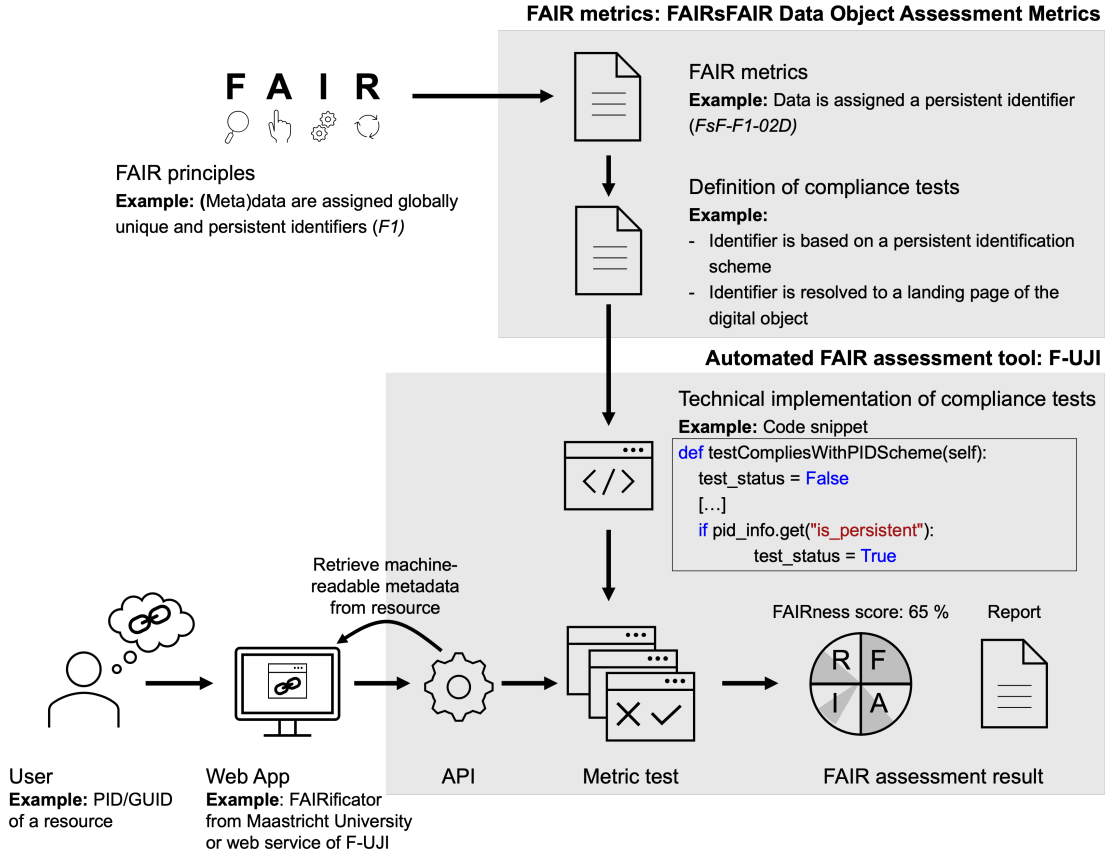


FIGURE 2.2: FAIRness evaluation workflow with the automated FAIRness assessment tool F-UJI (Devaraju and Huber, 2024). FAIR metrics are derived from FAIR principles, here exemplified for sub-principle F1. *FsF-F1-02D* is one of the FAIR metrics derived for F1 in the FAIRsFAIR Data Object Assessment Metrics (Devaraju et al., 2020). The *FsF-F1-02D* related compliance tests are technically implemented using a programming language within F-UJI. A potential user can now insert the PID or GUID of a resource into a web service such as the FAIRificator provided by Maastricht University (IDS, 2020) or the web service of F-UJI. The metric tests are executed, and the FAIRness assessment result is displayed as described in Table 2.5. This figure is adapted and adjusted from Sun et al. (2022).

tool. Practical representation in form of code for each of the compliance tests have to be developed for each of the metics. Now, a potential user with a PID or GUID of a resources can use a Web App such as the FAIRificator (IDS, 2020) or the web service provided by F-UJI itself to assess a resource's FAIRness. Specifically, web services communicate with F-UJI via API to run the assessment. The result is then provided in the respective Web service. Figure 2.2 emphasizes that automated FAIRness evaluation tools and FAIR metrics can be regarded as independent from another. The automated FAIRness assessment tool FAIR shake (Clarke et al., 2019), for instance, allows the use of several or self-configured sets of metrics (Lang et al., 2023).

Although FAIR metrics for research software have been formulated (Chue Hong et al., 2023), but they have not yet been implemented in automated FAIRness assessment tools. Candela et al. (2024) only include the tool *howfairis*, which explicitly applies to the assessment of software. Howfairis is not provided as-a-service tool, and does not explicitly refer to the individual FAIR Principles. Instead it is a command line tool based on five recommendations for FAIR software provided by fair-software.nl (NLeSC and DANS, 2020). The tests are derived from these recommendations, which are listed in the following.

- **(1/5) Code repository:** The code should be available from public accessible repositories with version control such as GitHub. Code repositories should be used from the beginning of the project to support early collaborations, allow reproducibility, and contribute to transparency.
- **(2/5) License:** A software license should be attached to the repository, because even if the code is available and accessible, it cannot be used without permission. The license defines the copyright owner and describes by whom and in which context the software may be used. Reuse is facilitated by permissive licenses such as MIT, Apache License 2.0, and GNUv3.
- **(3/5) Community registry:** The repository should be findable through community registries for software. These registries contain metadata of the code in order to make it findable for a specific community. Besides community registries also generic, national and language-specific repositories exist (NLeSC, 2023).
- **(4/5) Citation:** Use of software requires citation of its creators. The correct citation of the software is not always clear from a GitHub repository alone. Citation is facilitated if the creators provide a citation file or a DOI with all the relevant information for a specific release via, for example, Zenodo.
- **(5/5) Checklist:** Good quality code facilitates its reuse. Software quality checklists guide the creation of good quality code, such as its documentation, design standards, and best practices, and also help to report on the work status of the software project. One example is the Open Source Security Foundation's *Best Practices Badge Program*, which provides automated best practice compliance testing (OpenSSF, 2024).

Howfairis is based on compliance tests that can be automatically checked for public software repositories that are hosted by GitHub and GitLab. Each test is checked separately. After running

howfairis a badge illustrating the FAIRness result can be added to the readme-file of the repository. The command line result shows a list the five recommendations together with accepted solutions for each of them. Citation, for instance, has the options *has citation file*, *has citation .cff file*, or *has zenodo badge*. Howfairis assessment results provided in the command line do not provide recommendations directly, but the related website fair-software.nl (NLeSC and DANS, 2020) provides many suggestions that describe how compliance can be assured. The *Metrics for automated FAIR software assessment* (Chue Hong et al., 2023) have not been integrated in automated FAIRness assessment tools, but they are planned to be implemented into F-UJI and howfairis in the future.

Limitations of automated FAIRness assessment

Many different automatic FAIRness assessment tools based on various FAIR metrics have been developed (Moser et al., 2023). Consequently, the results of automated FAIRness assessment tools for the same resource are not the same. For the tools F-UJI and the FAIR Evaluator, the differences in the scores can be as drastic as a score of 20/22 obtained with FAIR Evaluator and a score of 2/24 obtained with F-UJI (Wilkinson et al., 2022).

Two main bottlenecks in the workflow of automatic FAIRness assessment tools (Fig. 2.2) lead to these different results. They derive from the interpretative steps. First, the static FAIR principles have to be reflected in FAIR metrics, which gives rise to diverging interpretations manifested in the metrics. Second, the implementation of the compliance tests into software leaves freedom for different code design, which may result in diverging FAIRness scores with different tools also if the same metrics are used.

To date, an objective FAIRness assessment remains difficult given the unspecific nature of the FAIR Principles and the resulting variation in interpretations. Diverging implementations have to be adjusted so that consistency between assessment results is given. Different FAIRness results for the same resource challenge objectivity and interoperability of the metrics (Jacobsen et al., 2020b; Moser et al., 2023). The differences between interpretations is also highlighted by Candela et al. (2024), who finds that the intent of metrics may relate to a FAIR principle, but the practical tests assigned may actually not be consistent with the principle.

Chapter 3

Digital infrastructures for reusable research products

The realization of the FAIR vision and the reusability of research products require technological solutions. While the FAIR Metrics are interpretations of the FAIR Principles that suggest how FAIRness can be achieved, technological solutions such as repositories, identifiers, and terminologies provide concrete implementation choices for data products that allow them to conform to a metric.

In general, technological solutions should focus on the sharing of research products and their standardization. I would like to illustrate the importance of both with an example. If a data set or modeling software used for a published article is stored only on an institute's server, it is not available to anyone without access to the server. By moving the resource to the publicly accessible part of the institute's website, it becomes accessible, but only to people who know the relevant link to the resource, which may be provided in the related article. The resource would still be invisible to anyone else who has not read the article. Even if someone finds the resource without knowing about the article, he or she would only be able to make sense of the resource if it provides sufficient contextual information in the form of metadata. By sharing the resource publicly through a repository, it becomes available, and by adding metadata, it becomes findable through keywords and understandable through context. If the resource, i.e., data or software, and its metadata also comply with community-agreed upon standards, the understanding and reuse would be even easier. Technological solutions and infrastructures, namely repositories and standardization efforts, support reusability and FAIRness of research products.

Over the past decades, the FAIR Principles and a strong emphasis on the management of research products have contributed to the development of a variety of repositories and standards. Decisions about the appropriate repository should be based on the publisher, institutional, and community recommendations and guidelines. A community or domain-specific repository should be preferred to an institutional repository, and the latter should be preferred to a general purpose repository (OpenAIRE, 2023; ARDC, 2024). Prior to publication, the structure, format, and terminology of the digital resource and its metadata should be conformed to standards (EU, 2021). This should

allow the digital resource to be integrated with other resources, reduce tacit knowledge, and make it easier to understand. Navigating and selecting among repositories and standards is difficult and best practices of specific domains are often not obvious. Look-up services such as the Registry of Research Data Repositories (RE3) (RE3, 2013b; Pampel et al., 2023) and FAIRsharing (Sansone et al., 2019) provide an overview and allow searching for suitable repositories and standards applicable to specific domains. RE3 lists ∼3,300 repositories not limited to data, and the FAIRsharing has ∼1,500 ready-to-use standards registered.

3.1 Repositories for sharing research products

Repositories form the interface between the creators of resources and the potential reusers of those resources, and they should make resources discoverable through search queries and accessible through graphical user interfaces or APIs. Repositories often have a DOI registration service that assigns a DOI to published resources. Many different repositories have emerged over the past decades to share all types of digital resources, and many of these are covered by RE3. In terms of content type, ∼290 repositories listed by RE3 allow the upload of *source code*, ∼510 repositories allow the upload of *software applications* and ∼2,070 repositories allow the upload of *scientific and*

Table 3.1: General properties of the four repositories GitHub (RE3, 2013a), Zenodo (RE3, 2023b), PANGAEA (RE3, 2023a) and AADC (RE3, 2021). The content is provided from the Registry of Research Data Repositories (RE3, 2013b), and the table is adapted from Simson et al. (2025c).

GitHub	**Zenodo**	**PANGAEA**	**AADC**
		Focus	
Code sharing, collaborative coding and version control	Research results of all sorts and from all domains	Georeferenced data from the Earth, environmental, and biodiversity sciences	Science data from Australia's Antarctic research
		Repository type	
Neither institutional nor disciplinary	Neither institutional nor disciplinary	Disciplinary	Disciplinary
		Content types	
	Standard office documents	Standard office documents	Standard office documents
	Images	Images	Images
	Plain text	Plain text	Plain text
	Audiovisual data	Audiovisual data	
	Archived data	Archived data	
Source code	Source code	Source code	
	Scientific and statistical data formats		Scientific and statistical data formats
	Raw data		
	Other		
Databases			
Software applications			
Configuration data			
		Quality control	
Optional automatic	None	Manual by data steward	Manual by data steward

statistical data formats. In terms of subject areas, ∼900 repositories are related to *geosciences (including geography)*. Cryospheric sciences are not listed as a subject. A simple search for *cryosphere* in the registry returns ∼30 entries.

In the following, I first summarize the sharing of geoscientific and cryospheric data in repositories. Next, I look at repositories that provide research software. I put a main focus on the repositories Zenodo, PANGAEA, Australian Antarctic Data Center (AADC), and GitHub as they are discussed in the case studies, but I will also provide other examples. An overview of the four repositories, including their main focus, repository types, content types allowed, and quality control is provided in Table 3.1.

3.1.1 Data repositories for data sets from geosciences and cryospheric sciences

Geoscience data can be found in many repositories. There are, for instance, repositories that cover a broad spectrum of different geoscience disciplines. The Data Publisher for Earth and Environmental Science PANGAEA hosts a total of more than 400,000 resources (Felden et al., 2023). Most of these are related to chemistry (∼73,900), lithosphere (∼51,000) and atmosphere (∼33,000). The Data Catalogue Service from the British National Environmental Research Council (NERC) (NERC, n.d.) manages ∼16,000 resources and the US-American National Centers for Environmental Information (NCEI) (NCEI, n.d.) provide several Data Discovery Tools of which the NOAA OneStop of the National Oceanic and Atmospheric Administration hosts over 100,000 resources. Additionally, also generic repositories host geoscience data. In Zenodo, ∼1,100 data set can be found when searching for *geoscience*, and it is ∼16,500 for Figshare.

Properties of repositories

Each repository has its own focus and curation process, provides metadata in different formats and schemes, uses different terminology, allows for specific data file formats, and has different filtering and querying options. Repositories may contain research products only from a specific region such as the Northern California Earthquake Data Center (NCEDC, 2014) or only from specific national institutions such as the AADC, which hosts data from Antarctica gathered by Australian research institutions. This also applies to the data repositories from the NCEI and NERC that focus on research products primarily generated by institutions in the United States and United Kingdom respectively.

Besides general purpose repositories and repositories with contributions from the broad geosciences there are also discipline-specific repositories relevant for geoscience. Highly specific repositories are the Crystallography Open Database (COD) (COD, 2023), which only accepts crystal structures of organic, inorganic, metal-organic compounds, and minerals in Crystallographic Information File (CIF) format, and the EarthChem Library (Lehnert et al., 2000), which focuses on geochemical data and provides submission file formats for specific measurements. The World Glacier Monitoring Service (WGMS) (WGMS, 2024) brings together data on glaciers and focuses on measurements

TABLE 3.2: Submission requirements and search possibilities of the data repositories Zenodo (Zenodo, n.d.[b]), PANGAEA (PANGAEA, 2024), and AADC. Fields annotated with an asterisk were clearly obligatory. It is not clear which further metadata is obligatory for AADC in the metadata record. Table is adapted from Simson et al. (2025c).

Zenodo	**PANGAEA**	**AADC**
Fields in data submission mask		
Title*		
Description*		
Resource type* (Publication, image, code, etc.)	Title*	Title*
Publication date*	Description	Description*
Creators*	Authors*	Submission type* (New data or replacement)
License	Keywords	
Contributors	License*	Release status* (Public, embargoed, review, etc.)
Keywords	References	
Languages	Projects	Project
Dates	Grants	Metadata record*
Version		
Publisher		
etc.		
Further metadata to be provided with data set		
		Temporal coverage
		Spatial coverage
		Purpose
	Date or time*	Quality
	Coordinates*	Access
	Corresponding publication	Science keywords
	Project	Additional keywords
	Parameter units	Locations
	Instruments	Platforms
	Methods	Instruments
	etc.	Researchers
		Use constraints
		etc.
Use of references, vocabularies or ontologies		
	Parameter names linked to, e.g., Environment Ontology, dbPedia, Wikipedia, and PANGAEA wiki	Global Change Master Directory (GCMD) keywords
Data set search filter options		
	Date and coordinate coverage	
	Author	
	Basis (Vessel, land)	Date coverage
Access status	Publication year	Coordinate coverage
Resource types	Topic	Researchers
Subjects	Projects	Source (Vessel, laboratory, field)
File type	Method or device	Keywords
	Campaign	
	Location (in words)	
Searchable metadata		
Advanced search based on field name metadata	Advanced search based on field names and further metadata	Simple search through metadata content

of glacier area, elevation, and mass balance. WGMS favors submission via a provided Google sheet or Excel templates. PANGAEA focuses on uploads of geo-referenced tabular data as tab- or xlsx-files but also allows other upload formats PANGAEA (2023a). AADC allows upload of data sets in common, non-proprietary formats AADC (n.d.). In contrast, generic repositories Zenodo and Figshare accept digital resources from all domains without prescribing or suggesting a specific digital representation format. They allow, but are not limited to, data sets, papers, pre-prints, software, presentations, and figures (Zenodo, n.d.[a]; Figshare, 2024).

Curation procedures may be manually performed by data stewards, as practiced by the NCEI, AADC, EarthChem Library, PANGAEA, and WGMS. COD performs automatic checks during the submission process. The general purpose repositories Zenodo and Figshare do not curate or quality check uploads, so resources are published directly after submission.

Data repositories request different metadata to be provided by the data collector in the submission mask and alongside or within the data file as summarized for AADC, PANGAEA, and Zenodo in Table 3.2. The metadata is then provided in different metadata formats and schemes, and using specific terminology. For example, PANGAEA provides metadata in *JSON-LD (Linked Data) format* following the *schema.org* schema, or in XML format for both *DataCite* and *Dublin Core* schemes (PANGAEA, 2020). AADC uses *Directory Interchange Format* (DIF) and Zenodo uses schemes such as *Dublin Core* and others. For domain-specific terminology, AADC uses the keywords of the *Global Change Master Directory* (GCMD) (GCMD, 2024), and PANGAEA uses, among others, the *Environment Ontology* (ENVO) (Jackson et al., 2021), the vocabularies of the *Chemical Entities of Biological Interest* (Hastings et al., 2016), and the *QUDT ontology* (FAIRsharing Team, 2015), which defines terms for units and quantities. Both metadata formats and schemes as well as terminology will be further discussed in Sect. 3.2.2 and Sect. 3.2.3

Filtering and searching options usually depend on the scope and focus of the repository. While general purpose repositories usually offer filter options for file format and subject, repositories that focus on georeferenced data often offer filtering for data from specific areas as practiced by PANGAEA and AADC. The availability of searchable metadata together with query and filter options strongly affects the search for resources on repositories. The efficiency of the search also depends on whether the searchable metadata is directly linked to the data file, i.e., whether the metadata consistently describes the actual file content. Searchable metadata is generated through Zenodo's submission mask and are not directly linked to actual data file content. Each data files consistency with the metadata has to be checked. This is similar to the AADC, where the metadata record in DIF format and using GCMD terminology is created separately for each record. The content of the metadata record is searchable but may differ from the actual content of the data file. In PANGAEA, predefined keywords such as parameter, method, and author can be used in the search. These keywords are consistent with the fields in the data file due to *relationalization* (Felden et al., 2023), so that search results for specific parameter names return only records with a corresponding column label.

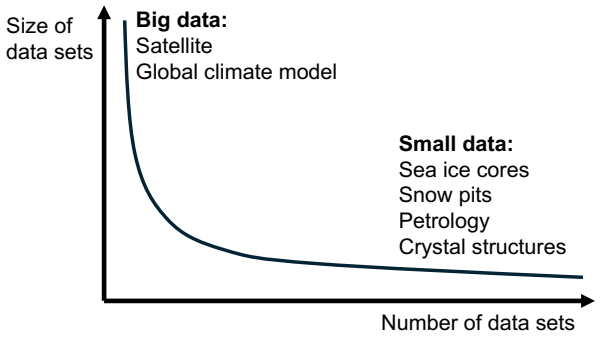


FIGURE 3.1: The long tail of geoscience data shows the distribution of the number of data set generated in geoscience with respect to their size. Some examples are provided.

Representation of cryospheric data in repositories

Of particular interest for cryospheric data are the polar-related data repositories. The Norwegian Polar Data Center (NPI, n.d.) is a repository of the Norwegian Polar Institute. It currently hosts ∼440 resources. They range from time series webcam images in Svalbard (Pedersen, 2013) to snow pit records (Itkin et al., 2024), but they also provide hunting restrictions for different areas in Svalbard (Sysselmannen, 2016). The Netherlands Polar Data Center (NPDC, n.d.) contains ∼50 individual data sets ranging from dissolved metal concentrations in the polar oceans (Middag, 2009) to glacier ice velocities (Tijm-Reijmer, 2010). The US National Snow and Ice Data Center (NSIDC) lists ∼1,400 data sets ranging from freeboard sea ice derived from ICE sat2 satellite data (Kwok et al., 2023) to ice core measurements (Bender, 2002). NSIDC does not host all data sets but provides the metadata of resources and then guides to the external location of the data such the U.S. Antarctic Program Data Center. PANGAEA hosts a total of ∼1,900 records related to cryospheric sciences, and AADC hosts ∼3,200 records originating from Antarctica. In addition, the generic repositories Zenodo and Figshare contain cryospheric data sets. Zenodo contains ∼280 records and Figshare contains ∼132 records. The amount of data sets per repository emphasizes the distribution of cryospheric data sets.

Challenges of managing long tail geoscience and cryospheric data in repositories

Geoscience data and thus also cryospheric data vary from many large data sets to an even larger number of small data sets that exhibit a long tail (Malik and Foster, 2012; Wyborn, 2023) as shown in Fig. 3.1. Large data sets are derived, for example, from Earth observation with satellites such as CryoSat-2 (Wingham et al., 2006), which is available through the ESA Earth Online portal (ESA, n.d.) and from global coupled climate models whose output is available through the Earth System Grid Federation (Cinquini et al., 2014). Small data sets contain only a few data points per data set, such as data from sea ice cores, snow pit observations, crystal structure, and geochemical analysis of rocks. Some of these small data sets can be deposited in discipline-specific and well-curated repositories such as the COD (COD, 2023) and the EarthChem Library (Lehnert et al., 2000). As shown in Fig. 3.2, these repositories, which specialize in specific types of small data, support the interoperability data sets and thus its potential reusability as a larger, combined data set. However,

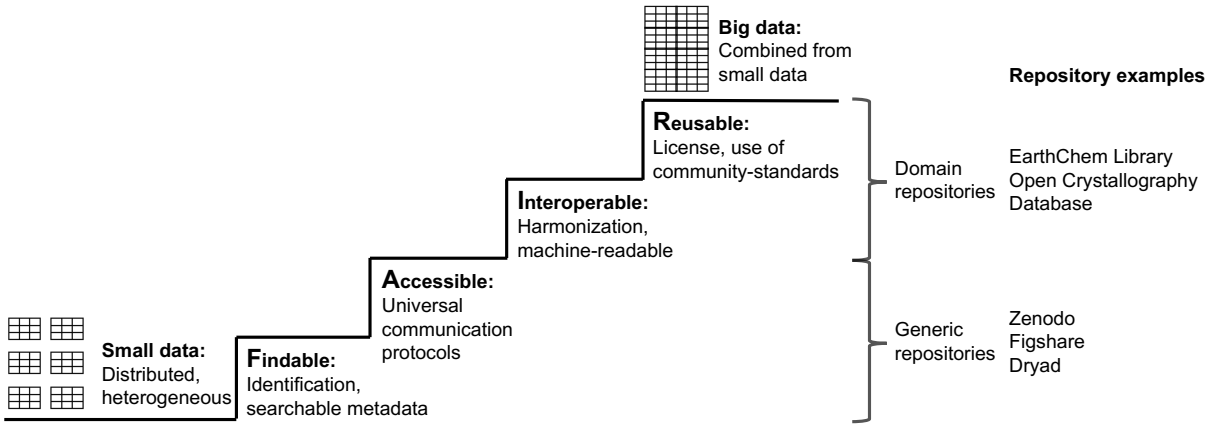


FIGURE 3.2: Combining small data to larger data by making them interoperable with repositories. Figure is adapted from Lehnert (2015).

such highly specialized repositories are often not available for small data sets such as measurements from snow pits (e.g., Schöber et al. (2019), Bouchard et al. (2022) and sea ice cores (e.g., Katlein et al. (2020a)). These data sets are then deposited in broader domain repositories, institutional repositories, or generic repositories. While these repositories ensure findability and accessibility of the data, they do often not facilitate its interoperability and thus do not support its efficient reuse. On the one hand, these generic repositories often offer flexibility, allowing data authors to design metadata and data files according to their needs. On the other hand, this flexibility contributes to inconsistencies between individual data sets complicating their combined reuse. Thus, the potential of small data often remains small and unexplored.

3.1.2 Code repositories for research software from geosciences and cryospheric sciences

Published code is distributed among many types of repositories. There are many domain specific repositories registered in RE3 that also host source code and software applications such as the World Ocean Database (Boyer et al., 2018) and also PANGAEA. In addition, the generic repositories Zenodo and Figshare also allow submission of code. Zenodo hosts ∼1,200 geoscience-related software assets and ∼130 cryosphere-related software assets. For Figshare, it is ∼110 and ∼10 respectively. In addition to these general-purpose repositories, there are repositories specialized for all kinds of software, such as GitHub and GitLab. GitHub hosts more than 420 million repositories and has more than 100 million registered members (GitHub, n.d.). A search for *cryosphere* on GitHub shows the term is used over 17,000 times in code, and about 150 repositories use it in their name or description. There may be duplicates between the repositories provided via Zenodo and Github, as Zenodo is often used to register the software by depositing a zip-file of the software and assigning it a DOI via the Zenodo Github integration (GitHub, 2024). Additionally, Zenodo also makes the software findable if it is described with metadata.

Besides repositories researchers can make their software products findable through software registries as they are listed by the the Netherlands eScience Center (NLeSC, 2023). For geoscientific research software only a few registries exist such as the portal of the Community Surface Dynamics Modeling System ((Geosoft, n.d.)), which also lists three software projects from cryospheric sciences.

3.2 Standardization efforts for the consistency of research products

The provision of resources and their metadata in standardized forms is a prerequisite for making their content understandable, comparable, and combinable. Standards should minimize the tacit knowledge of a resource, allow its unambiguous reuse in different contexts, and support its findability.

Standards are relevant from the creation of a resource to its publication. Standardized procedures, such as measurement guidelines and protocols for measurement data, ensure that the same steps have been followed when measuring a particular variable and that the same metadata is documented (Halbritter et al., 2019). For software development, there may be standards for specific folder structures such as a generic structure for Python packages (PPA, 2024). These standards should be used from the creation of a data set or software project to make them easily understandable and reusable for others. When resources are published on repositories, further technical standards are used to digitally represent the metadata for efficient discovery, citation, and reuse of the resources. While the quality of a resource greatly benefits from the use of standards from the very start, it is often the case that standardization is only applied when the resources are made available in repositories.

Crystal-Ornelas et al. (2022) distinguish between strict and rigid *formal standards*, such as those certified by the International Organization for Standardization (ISO), and more flexible, adaptable, and evolving *reporting formats*. Reporting formats are often derived from best practices and may eventually be incorporated into formal standards. In the following, I will refer to standards as the ensemble of formal standards and reporting formats, including, for example, ontologies, controlled vocabularies, metadata formats and schemes, data file formats, and guidelines for the creation of research products. Standards can be discipline-specific or general-purpose, and "standards are fragmented, with gaps and duplication" (Sansone et al., 2019), which also means that the distinction between different standards is often not clear and transitions are smooth.

In recent decades, the emphasis on FAIR and improved stewardship of digital resources has led to the development of new standards, with a particular focus on their machine-readability. Machine-readable standards are the foundation of FAIR by enabling the automatic processing of metadata and potentially resources. Standards are constantly being developed, updated, or deprecated, and they are difficult to navigate. The FAIRsharing service provides an overview of existing standards to facilitate community navigation and adoption.

FAIRsharing contains four different types of standards (Sansone et al., 2019; Sandström et al., 2023):

- *Reporting guidelines* state the required, contextual information of a digital object or the behaviors to be followed to create a digital object in narrative form such as checklists and guiding principles.
- *Models and formats* define the technical representation of digital resource and especially their metadata. They enable their machine usability, retrieval, and exchange between systems based on metadata schemes and transmission formats.
- *Terminology artifacts* form semantic layers of digital objects that allow for interpretation by machines and humans. They can be used to query for digital objects through distinctive identification of concepts and terms such as controlled vocabularies, taxonomies, and ontologies.
- *Identifier schemes* identify information in a unique, machine-readable way through PIDs that provide links between digital research objects and their definitions, authors, organizations and infrastructures.

The reader should note that these four types are not separate from each other. Assigning unique identifiers to terminology artifacts helps to uniquely distinguish terms. Metadata models may be developed to satisfy a specific reporting guideline, and terminology artifacts can be technically represented according to a structure suggested by metadata models and encoded in accordance with a metadata format.

In the following, I will provide some examples of standards from FAIRsharing, following the distinction into the four types of standards as proposed by Sansone et al. (2019). While FAIRsharing mainly registers standards that are already digitally represented and partially machine-readable, many community standards such as measurement guidelines, documentation templates, and preferred terminology are still only understandable and usable by humans, as they lack sufficient digital representation. Specific practices are often hidden on institute or stakeholder websites, for example as PDF documents or Excel files, making them difficult to find for researchers from other disciplines or new to the field. Therefore, I will supplement the standards from FAIRsharing with further examples.

3.2.1 Reporting guidelines

Reporting guidelines narratively describe the information to be provided or the instructions to be followed when generating a digital research product. Adherence to the guidelines should provide complete and concise contextual information and allow to understand the described research product (Sansone et al., 2019; Sandström et al., 2023). Reporting guidelines comprise, for instance, minimal requirements for metadata or a sequence of steps to be followed when publishing a digital research product; they may be formulated for the general community or specific domains and for

specific repositories and infrastructures. The reader should note that reporting guidelines also exist for metadata standards themselves such as for terminology artifacts (Matentzoglu et al., 2018). Guidelines may also not be explicitly called as this but rather like the standard they describe.

Generic reporting guidelines as listed by FAIRsharing.org are, for instance, the FAIR Principles for data and for research software (Wilkinson et al., 2016; Barker et al., 2022) as well as the CARE principles for Indigenous Data Governance (Carroll et al., 2020). Specific guidelines for infrastructures include those for data managers providing instructions for archiving data to ensure compatibility with OpenAIRE (FAIRsharing Team, 2022b), and for federal agencies to select an appropriate repository for acquired data (OSTP, 2022). An example from geoscience are the Marine Environmental Data and Information Network guidelines that list all relevant information to be provided with marine data (MEDIN, 2024).

Geoscientific reporting guidelines

In the field of geoscience, Simmonds et al. (2022) provide a guideline for the publication of inputs, metadata, model code and output from terrestrial models. Crystal-Ornelas et al. (2021) provide a *Guide to Using GitHub for Developing and Versioning Data Standards and Reporting Formats* with a special focus on geoscience. Furthermore, Peng et al. (2021b) call for the development of community guidelines for sharing quality information of Earth Science data. A response to this call is the first draft of the *International Community Guidelines for Sharing and Reusing Quality Information of Individual Earth Science Data sets* that provides recommendations for data set preparation and curation.

Measurement guidelines for cryospheric sciences

For cryospheric sciences, there are guidelines for standardized measurements of the cryosphere by the World Meteorological Organization (WMO) such as the *Guide to Hydrological Practices* (WMO, 2008) or the *Guide to Instruments and Methods of Observation Volume II – Measurement of Cryospheric Variables* (WMO, 2023) which both describe measurements of snow depth. Mahoney and Gearheard (2008) describes best practices for field techniques for sea ice research.

Sea ice measurement reporting guidelines

For sea ice, several measurement guidelines are available. The ASPECT (Antarctic Sea ice processes and climate) expert group provides some guidelines for sea ice measurements on their website, such as for in-situ measurements and ship-based observations of sea ice. They also provide a standardized template for documenting sea ice measurements such as salinity, thickness, temperature, chlorophyll, and structure (ASPeCT, n.d.). In addition, Miller et al. (2015) present a guideline for metadata that should be provided with biogeochemical studies of sea ice cores. They suggest adding the coordinates, date and time, weather conditions, water depth, and sea ice development stage for each core. For each measurement along that core, they suggest to include the depth of the measurement

```
{
  "@context": "http://schema.org",
  "@id": "https://doi.org/10.5281/zenodo.3779867",
  "@type": "https://schema.org/Dataset",
  "identifier": "https://doi.org/10.5281/zenodo.3779867",
  "name": "Physical properties of summer sea ice in the Pacific sector of the Arctic
    [...]",
  "creator": [
    {
      "name": "Qingkai Wang",
      "familyName": "Qingkai Wang",
      "affiliation": [
        {
          "@type": "Organization",
          "name": "Dalian University"
        }
      ],
      "@type": "Person"
    }
  ],
  "publisher": {
    "@type": "Organization",
    "name": "Zenodo"
  },
  "contentSize": "5.99 MB",
  "license": "https://creativecommons.org/licenses/by/4.0/legalcode",
  "description": "<p>This dataset includes the data of physical [...] ice cores <\\p>",
  "url": "https://zenodo.org/records/3779867",
  "distribution": [
    {
      "@type": "DataDownload",
      "contentUrl": "https://zenodo.org/api/records/3779867/files/readme.txt/content",
      "encodingFormat": "text/plain"
    }
  ]
}
```


FIGURE 3.3: JSON-LD format (identified by @context: "http://schema.org") is used to describe the data set Wang et al. (2020b) with Schema.org vocabularies like @type (Dataset, Person, Organization), name, creator, publisher, and license, detailing key information such as DOI, creator (Qingkai Wang), and publisher (Zenodo). JSON-LD-file was downloaded from Zenodo.

in the core and the thickness of snow or meltwater layers above the ice surface. Miller et al. (2015) also emphasize that the distance between several cores in close proximity should be specified.

3.2.2 Models and formats

FAIRsharing defines models and formats as means to digitally represent metadata, ensuring that both the metadata and the resource it describes are machine-readable and usable (Sansone et al., 2019; Sandström et al., 2023). These models and formats are technological solutions that specify the structure and encoding of the digital representation of metadata. They form the information layer that repositories use to query digital resources and facilitate metadata harvesting. Automatic FAIR evaluation tools use metadata files that are based on a combination of metadata models and formats for the evaluation of a resource. Information stored in these metadata files provides the basis for automatic reuse of digital resources.

Generic models and formats

A model defines the structure and relationships between different pieces of information, serving as a conceptual framework for metadata organization and description. This ensures that metadata is structured consistently across systems. Common metadata models include:

- Schema.org (Schema.org, 2024) provides a metadata model that allows for the structured description of various types of web resources, including data sets, publications, people, organizations, and more. It offers a set of standardized terms that ensure metadata is machine-readable and consistent across platforms.
- DataCite (DataCite, 2024b) provides a metadata model specifically for citing research products, using terms such as title, creator, and DOI.
- Dublin Core is a widely used metadata model (DCMI, 2020) for describing a wide range of resources. It uses a set of 15 core elements such as title, creator, subject, date, and format. Dublin Core is often used for general purpose metadata across disciplines and systems.
- CodeMeta (Jones et al., 2023) is a scheme for science software and code and provides a vocabulary that can be used to standardize the exchange of software metadata across repositories and organizations.
- Resource Description Framework (RDF) uses triples (subject-predicate-object) to define relationships, where each part can be identified by a Uniform Resource Identifier (URI) (W3C, 2024). RDF facilitates linked data in the Web Ontology Language (OWL). Simple Knowledge Organization System (SKOS) builds on RDF. RDF is usually encoded in XML or JSON format.

Formats define how metadata is encoded and transferred between systems, enabling interoperability across platforms. Common formats include:

- JSON-LD encodes linked data in JSON, which is commonly used with Schema.org vocabularies. Fig. 3.3 provides an example of a metadata file in JSON-LD format.
- Extensible Markup Language (XML) is a widely used format for the encoding of various metadata models. Fig. 3.6 shows an example of a metadata file in XML format.

By integrating models with formats, metadata standards provide interoperability, often through mappings between combinations of standards, for instance, for DataCite and Schema.org (DataCite, 2024a). This enables systems to exchange metadata efficiently. Zenodo provides metadata in a variety of formats including JSON, JSON-LD, DataCite JSON, DataCite XML, Codemeta and Dublin Core XML. An example of JSON-LD format using the Schema.org model is provided in Fig. 3.3. This variety of formats allows Zenodo to be interoperable with multiple systems and to support a wide range of metadata exchange needs. PANGAEA, provides metadata, for instance, in JSON-LD format following the schema.org scheme or in XML format for both DataCite and Dublin Core Schemes (PANGAEA, 2020).

Geoscience standards

The models and formats discussed previously are generic. In domain-specific contexts, appropriate formats are required. There exist geoscience-related models and formats. The Open Geospatial Consortium (OGC) has developed many models and formats for geospatial data, such as the Water Markup Language (OGC, n.d.), which is a model for describing water observations, such as time series of water levels in a river. There are also formats such as GeoJSON, which represents geographic features in JSON format (FAIRsharing Team, 2022a). ISO 19115 (ISO, 2014) specifies the metadata content for geospatial data, and ISO 19139 (ISO, 2019) provides an XML schema for the encoding of the metadata after ISO 19115. Furthermore, relevant for geoscience is the Observation Data Model 2 (ODM2) (Horsburgh et al., 2016). ODM2 has been developed to facilitate exchange of data sets across disciplines, and it is designed in a way to fit to the requirements of a broad range of geoscience data from, e.g., hydrology, geochemistry, biodiversity, and originating from different sampling sites, e.g., weather stations, bore holes. Furthermore, NASA has developed standards for Earth Science data sets. Tow examples are the Directory Interchange Format (DIF) (NASA, 2015) and the Unified Metadata Model (UMM) (NASA, 2021).

Among other standards the repository PANGAEA, provides metadata using ISO 19115 formatted based on ISO 19139 and additionally DIF-files formatted in XML (PANGAEA, 2020). AADC provides metadata as DIF-files in XML, and Zenodo provides metadata in GeoJSON format.

3.2.3 Terminology artifacts

Terminology artifacts comprise semantic representation of knowledge, also referred to as knowledge organization systems (KOS). They include controlled vocabularies, taxonomies and ontologies (Mazzocchi, 2018). The major functions of the three terminology artifacts following Zeng (2008) are listed in Table 3.3. Figure 3.4 is a simplified illustration of a controlled vocabulary, taxonomy and ontology based on the classes of the Environment Ontology (Jackson et al., 2021). Controlled vocabularies are lists of preferred terms within a community; they enable consistency and reduce ambiguity. Taxonomies classify the defined terms in form of 1D hierarchical relationships as parent or child. Ontologies conceptualize the terms as classes with multi-dimensional and semantic relationships. These classes are characterized by their relationships to other classes and their specific attributes such as a class-specific properties. While there are indefinite ways to arrange these terms into a conceptualization, an ontology should aim at its reuse by clearness, coherence and consistency, and it should provide the opportunity for extension (Jashapara, 2011).

Terminology artifacts are digitally represented, for instance, using formats like OWL or RDF-based Simple Knowledge Organization System (SKOS). These formats constitute a perquisite for the description of digital objects with machine understandable metadata, enabling autonomous findability and interpretability. However, many controlled vocabularies are rather available in form of txt-, csv-, pdf- or html-files since they were not intended for use by machines at the time of creation. As domain knowledge grows, semantic representations have to be updated and maintained (Sansone et al., 2019).

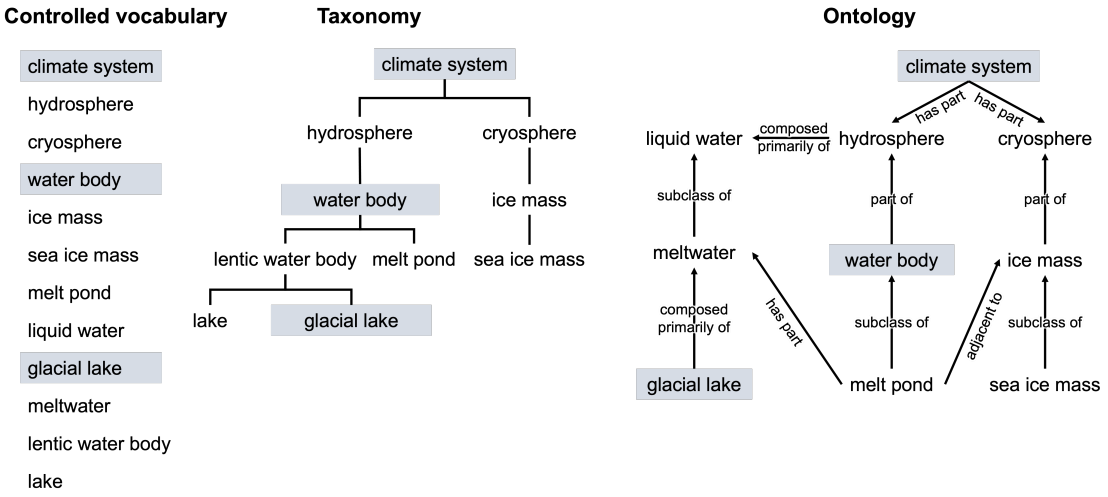


FIGURE 3.4: Comparison of a controlled vocabulary, taxonomy, and ontology for terms related to cryospheric sciences from the Environment Ontology (ENVO).

FAIRsharing currently contains ∼40 ready and subject-agnostic terminology artifacts. They are for example the Unified Code for Units of Measure (FAIRsharing Team, 2018c) and the Data Use Ontology (FAIRsharing Team, 2018b). Several geoscience related naming standards are listed in Table 3.4. The representation of cryospheric terms is indicated. The Environment Ontology (ENVO) contains the most terms related to snow and sea ice. Besides ENVO also the Semantic Web for Earth and Environmental Terminology (SWEET) (SWEET, 2022) is an important semantic source for the geoscientific and especially the cryospheric community (Duerr et al., 2024). Figure 3.5 shows the terms for snow and sea ice available in SWEET and ENVO. The arrows show the references between them. For sea ice SWEET points to ENVO and for snow ENVO references similar terms in SWEET.

The Ontology Lookup Service (OLS) (EMBL-EBI, 2023) and the BioPortal (Whetzel et al., 2011) help to navigate between ontologies and at finding suitable terms. Both have a major focus on biomedical research. OLS provides ∼270 ontologies, and the BioPortal (Whetzel et al., 2011) hosts ∼1350 ontologies including ENVO and SWEET. Both platforms make terms and ontologies searchable and enable visualization of the terms including their semantic relationships to other terms in form of graphs. The Linked Open Vocabulary (LOV) (OEG, 2024) makes ∼860 vocabularies and also ontologies searchable in a combined way. Duerr et al. (2024) describe the process of harmonizing cryospheric semantics, which is required due to the inconsistencies between terminologies,

TABLE 3.3: Major functions of knowledge organization systems adapted from Zeng (2008).

Major functions	**Controlled vocabularies**	**Taxonomy**	**Ontology**
Eliminating ambiguity	✓✓✓	✓✓	✓✓
Controlling synonyms		✓✓	✓✓
Establishing hierarchical relationships		✓✓✓✓	✓✓✓
Establishing associative relationships			✓✓✓✓✓
Presenting properties			✓✓✓✓✓

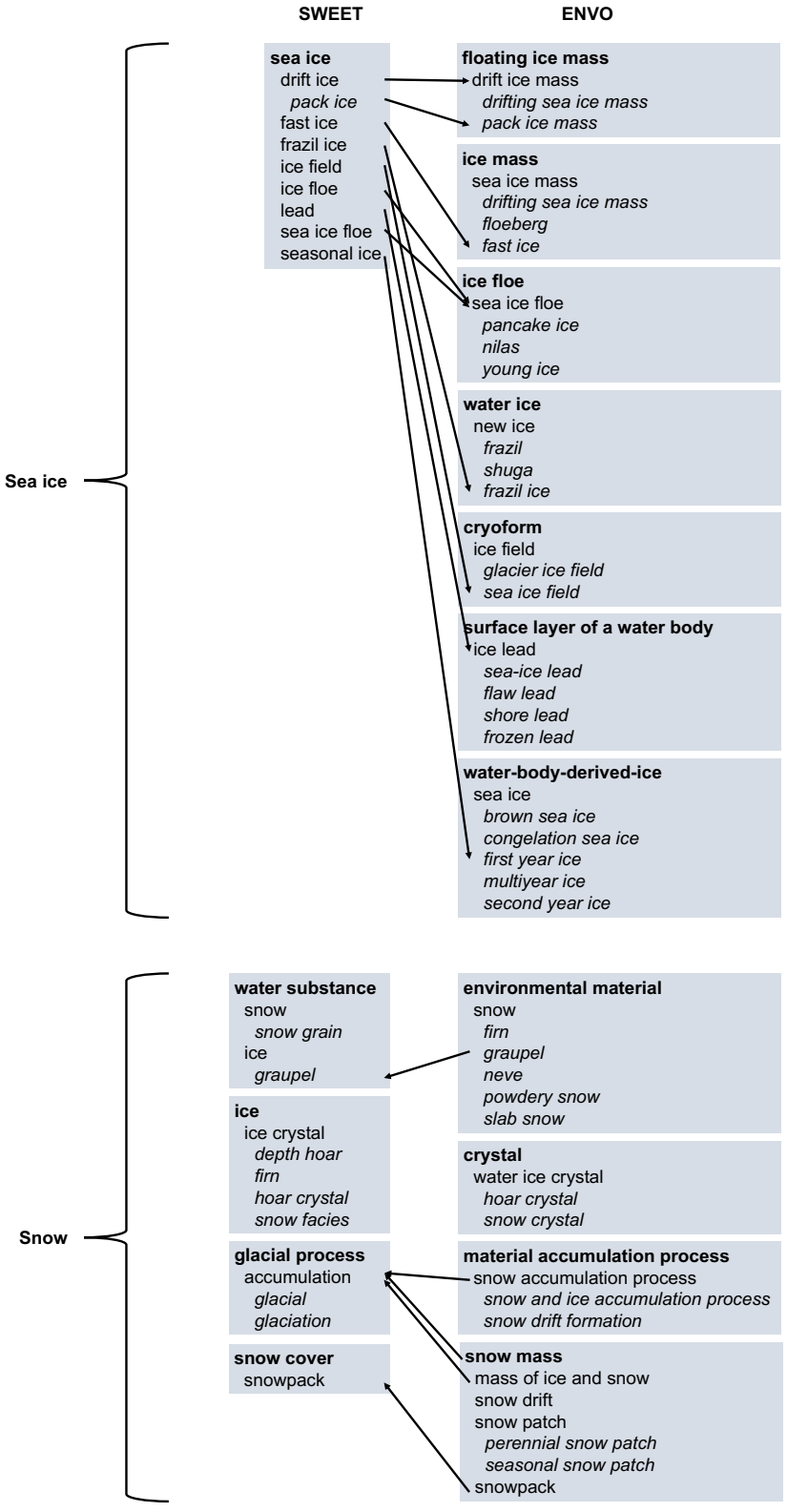


FIGURE 3.5: List of snow and sea ice related terms that are available in the ontologies SWEET and ENVO. Arrows show cross-ontology references to similar or identical terms from SWEET to ENVO for sea ice and vice versa for snow.

TABLE 3.4: Selection of terminology artifacts relevant for geosciences and indicating their availability of cryospheric terms.

Terminology artifact	**Focus**	**Cryospheric terms**
Observations Data Model 2 Controlled Vocabulary (ODM2 CV)	Spatially discrete Earth observations	✓
Paleoenvironmental Standard Terms (PaST) Thesaurus	Variables measured by the World Data Service for Paleoclimatology	✓
Community Surface Dynamics Modeling System (CSDMS) Standard Names (CSN)	Identification and categorization of scientific variables for modeling	✓✓✓
Environment Ontology (ENVO)	Environmental entities of all kinds, from microscopic to intergalactic scales	✓✓✓
Marine Regions	Relational list of mainly marine geographic names, coupled with information and maps	✓
Semantic Web for Earth and Environmental Terminology Ontology (SWEET)	Domain-specific ontologies for Earth System Science	✓✓
Global Change Master Directory Keywords (GCMD Keywords)	Controlled vocabulary for Earth Science	✓✓

definitions, and metadata standards specifically between the *Glossary of the Global Cryosphere Watch* (GGCW) (WMO, n.d.[b]) provided by the World Meteorological Organization (WMO) and the ontologies SWEET and ENVO. The GGCW lists terms from several sources such as the WMO Sea Ice Nomenclature (WMO, 2014) and the Illustrated Glossary of Snow and Ice (Armstrong et al., 1966). Duerr et al. (2024) raise the need for the representation of numeric values from definitions provided in the GGCW into ontologies such as the size of an ice floe or the age of sea ice. These properties are currently not provided in a machine readable way, so that human readable definitions are more accurate than the machine readable one.

3.2.4 Identifier schemes

Identifier schemes are used to uniquely identify digital objects on the Internet and to link digital resources (Sansone et al., 2019). They are used in metadata to identify and define labels, people, organizations, methods, terminologies and places. These identifiers should be persistent in the form of Persistent Identifiers (PIDs) to ensure permanent accessibility. Identifier schemes are digital in nature. FAIRsharing lists ∼70 identifier schemes. Some generic ones are:

- Uniform Resource Identifier (URI) is a resource identifier consisting of a specific combination of digits and characters. These resources are typically accessible over the Internet. The rules of URI syntax are inherited by all URI subtypes. URIs can be defined for, e.g., websites, folders, email addresses. (Berners-Lee et al., 2005)

```
1   <identifier identifierType="DOI">10.5281/zenodo.6997449</identifier>
2   <alternateIdentifiers>
3     <alternateIdentifier alternateIdentifierType="URL">https://zenodo.org/records
      /6997449</alternateIdentifier>
4   </alternateIdentifiers>
5   <creators>
6     <creator>
7       <creatorName nameType="Personal">Omatuku Ngongo, Emmanuel</creatorName>
8       <givenName>Emmanuel</givenName>
9       <familyName>Omatuku Ngongo</familyName>
10      <nameIdentifier nameIdentifierScheme="ORCID">0000-0002-8802-0262</nameIdentifier>
11    </creator>
12  <creators>
13 [...]
14  <relatedIdentifiers>
15    <relatedIdentifier relatedIdentifierType="DOI" relationType="IsVersionOf">10.5281/
      zenodo.6997448</relatedIdentifier>
16    <relatedIdentifier relatedIdentifierType="URL" relationType="IsPartOf">https://zenodo
      .org/communities/scale_south_africa</relatedIdentifier>
17  </relatedIdentifiers>
18 [...]
19  <rightsList>
20    <rights rightsURI="https://creativecommons.org/licenses/by/4.0/legalcode"
      rightsIdentifierScheme="spdx" rightsIdentifier="cc-by-4.0">Creative Commons
      Attribution 4.0 International</rights>
21  </rightsList>
```

FIGURE 3.6: XML format with DataCite schema describes the data set Omatuku Ngongo et al. (2022) using tags like ⟨identifier⟩ (DOI: 10.5281/zenodo.6997449), ⟨creatorName⟩ (Emmanuel Omatuku Ngongo), and ⟨rights⟩ (CC BY 4.0). DataCite vocabularies such as nameType, relationType, and rightsIdentifier define metadata details, including the creator's ORCID and related identifiers. XML-file was downloaded from Zenodo.

- Uniform Resource Locator (URL) is a subtype of URI and defines the location of the resource. URLs are typically used to address web pages using the HTTP protocol (Berners-Lee et al., 1994). Fig. 3.6 provides examples of an URL linked in the metadata file in lines 3 and 16.

- Universally Unique Identifier (UUID) is also called Globally Unique Identifier (GUID). UUIDs have a size of 128 bits, and they are designed to uniquely identify resources, such as their name in the form of a Uniform Resources Name (URN). UUIDs do not require an official registration process because they are generated by an algorithm. (Davis et al., 2024)

- Open Researcher and Contributor ID Registry (ORCID) is used to transparently acknowledge persons that were involved in the creation of research products (ORCID). Once registered in the ORCID repository, researchers can link their research products with their unique ORCID ID. Fig. 3.6 provides an example of an ORCID linked in the metadata file in line 10.

- Digital Object Identifier (DOI) (FAIRsharing Team, 2018a) was first introduced in 1997, and it has been represented by the International Standard ISO 26324 since 2012. DOIs are typically assigned to digital resources once they are published to make them uniquely identifiable and permanently accessible. Using the DOI metadata of the corresponding resource can be accessed and retrieved. So far over 90 billion DOIs have been registered. Fig. 3.6 provides an example of a DOI linked in a metadata file in line 1.

In geoscience, the International Generic Sample Number (IGSN) is used (FAIRsharing Team, 2023). IGSNs uniquely reference physical samples such as outcrops, rock or plant samples. Initially developed for geoscience the concept is now also used in other domains. Each IGSN is assigned with metadata providing details on the sample (Klump et al., 2021).

Chapter 4

Cryospheric Case Study I: Physics-based process model for snow

Software development is becoming easier, and many people with and without programming and software development training are developing code to model physical processes. These software assets need to be documented transparently in order to make the best possible contribution to scientific progress. Accordingly, all specifications of the methods used, such as the numerical approximation and the boundary conditions, must be described. Furthermore, the developed code must be published and legally reusable. Reusable modeling software has a high potential for comparative studies, model coupling and interpretation of measurement data. In Cryospheric Case Study I, I present an example of a modular and extendable snow model and its modeling software with a special focus on its detailed and transparent documentation to enhance the reuse potential.

This case study is based on the article *Elements of future snowpack modeling – Part 2: A modular and extendable Eulerian–Lagrangian numerical scheme for coupled transport, phase changes and settling processes* published in *The Cryosphere* (Simson et al., 2021). The article is published under the Creative Commons Attribution 4.0 License. Parts of this article are integrated in this thesis with permission from the copyright holders, namely the authors. Sect. 4.1 is partly adopted with adaptations for better alignment with this thesis. Most of Sect. 4.2 has been adopted as is from the article, except for a few structural changes and omissions for better accordance with this thesis. Sect. 4.3 is new, and has been specifically written for this thesis. The reader should note that the article's Sect. 5 *Summary and Conclusions* and Sect. 5 *Future work and challenges* are only partly reproduced in this thesis. The related software asset *Eulerian–Lagrangian snow solver* is available in a GitHub repository (Simson and Kowalski, 2021a) with a MIT license, and it is registered via Zenodo (Simson and Kowalski, 2021b).

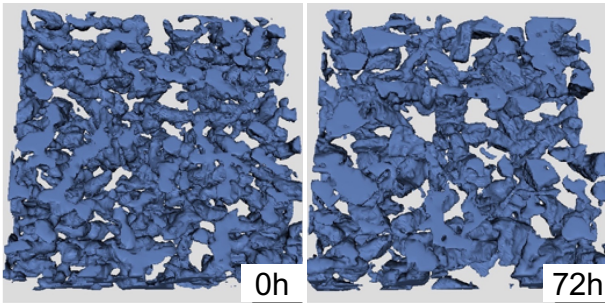


FIGURE 4.1: Metamorphic changes of the snow microstructure after 72 h imaged with a micro-CT scanner. The width of the snow sample is 3.6 mm, and the ambient temperature gradient is 50 Km^{-1}, with the warmer temperature at the bottom. The figure is adapted from Pinzer et al. (2012) with license CC BY 3.0.

4.1 Introduction to the case study

Snow contributes to the runoff of melt water, controls the albedo of the Earth, insulates the soil, and is subject to avalanches. Research questions are interdisciplinary (see Sect. 1.3) and concern predictions on the amount of melt water runoff (Lang, 1986), the effect of snow cover on Earth's energy balance (Meng, 2016) and on permafrost (Park et al., 2014), as well as the forecast of snow avalanches (Schweizer et al., 2003).

Snow models simulate physical processes and are available to study these questions. Snow density is probably the most important prognostic variable of snow models (Krinner et al., 2018, and references therein). Changes in snow density are related to mechanical compaction and phase change processes in form of sublimation, melting, deposition, and freezing, and they are described with the ice mass balance. In the last decade, observations of snow in micro-CT scanners have shown that snow can change its entire microstructure within hours (Pinzer et al., 2012). This process is called *snow metamorphism*, and Fig. 4.1 shows examples of micro-CT images. Snow metamorphism is driven by diffusive water vapor transport and strongly influences the thermal and mechanical properties of the snow. Yet, the exact nature of the non-linear physical interaction of water vapor transport with heat transport and stress-induced settling remains to be investigated in depth.

SNOWPACK (Bartelt and Lehning, 2002; Lehning et al., 2002) and Crocus (Brun et al., 1989; Brun et al., 1992; Vionnet et al., 2012) are so call *detailed* snowpack models. They have been developed for avalanche forecasting and are now also widely used for applications in climate science (Bavay et al., 2013; Spandre et al., 2019) and hydrology (Eidhammer et al., 2021). While heat transport, mechanical settling and processes due to the presence of liquid water have been incorporated into SNOWPACK and Crocus for a long time, water vapor transport and its effect on snow density has only been considered recently (Touzeau et al., 2018; Jafari et al., 2020). Two characteristics of detailed snowpack models challenge the coupling with water vapor diffusion.

First, SNOWPACK and Crocus only implicitly incorporate the ice mass conservation equation, which is the most important conservation law for snow and describes changes in snow density. The ice mass conservation equation is not even explicitly mentioned in the technical documentations of

both models (Brun et al., 1989; Brun et al., 1992; Bartelt and Lehning, 2002; Lehning et al., 2002), and only a more detailed inspection reveals how ice mass conservation is accounted for. Changes of ice mass from settling processes are integrated rather indirectly by stating a settling law for individual layers that are resorted with a "Lagrangian coordinate system that moves with the ice matrix" (Bartelt and Lehning, 2002). This scheme translates the snow deformation into a thickness change of the layers (Brun et al., 1989; Vionnet et al., 2012). While this procedure has been well established for a long time, it is without numerical ambiguities only in the absence of water vapor transport. Another disadvantage of this implicit ice mass conservation is that it hinders the isolation of the numerical scheme encompassing all coupled non-linear partial differential equations to further develop the numerical core. Accordingly, both models lack flexibility for the implementation of new parametrizations or coupling schemes such as water vapor transport.

Second, resolving diffusive water vapor transport requires short temporal (1 min) and small spatial scales (0.1 cm), which do not match with the typical scales of Crocus and SNOWPACK. For SNOWPACK, time steps are on the order of 15 min or longer (Bartelt and Lehning, 2002), and a typical layer thickness is 2 cm (Wever et al., 2016). For Crocus, time steps are on the order of 15 min (Viallon-Galinier et al., 2020) to 1 h (Vionnet et al., 2012), and the minimum layer thickness is 0.5 cm (Brun et al., 1989).

Modeling studies that exclusively focus on the non-linear coupling of diffusive water vapor transport and heat conduction have been put forward by Calonne et al. (2014) and Hansen and Foslien (2015a). They are based on upscaled and homogenized continuum mechanical process models and provide different flavors of how to set up the underlying mathematical model. However, both approaches investigate water vapor diffusion in the absence of settling and neglect its feedback on snow density. Furthermore, the digital representation of both approaches in from of modeling software is not sufficient. Calonne et al. (2014) does not provide modeling software as they use COMSOL Multiphysics for their computation. Hansen and Foslien (2015a) provide the MATLAB code of the model as supplementary material to their article (Hansen and Foslien, 2015b). However, the code does not provide sufficient modularity and documentation.

A major difference between SNOWPACK and Crocus with the models of Calonne et al. (2014) and Hansen and Foslien (2015a) is the spatial discretization. Detailed snowpack models divide the snowpack into layers of different thicknesses and assign physical properties to each layer. Accordingly, a spatial increment is literally a layer of snow originating from the same precipitation event. This is why Crocus and SNOWPACK are referred to as layer-based schemes in the following. In contrast, Calonne et al. (2014) and Hansen and Foslien (2015a) use a continuous discretization of the snowpack with the same spatial increments, i.e., an equidistant computational grid. Thus, a snow layer is represented by several spatial increments, and physical properties change within this layer.

The investigation of the coupling of water vapor transport with other snow processes would benefit from modular and flexible physics-based process models based on simple, mathematically rigorous numerical approximations. More specifically it would facilitate the investigation of competing non-linear effects from distinct processes, the quality assessment of the numerical approximations

and the testing of different parametrization. However, current models lack the necessary flexibility to easily test new coupling schemes or adjust parametrization, and they also lack fully transparent documentation of numerical approximations and related software assets. Both limiting their reusability and understandability. While significant progress has been made in understanding the coupling of water vapor and heat transport within snow, particularly through the companion article *Elements of Future Snowpack Modeling – Part 1: A physical instability arising from the non-linear coupling of transport and phase changes* by (Schürholt et al., 2022), there remains a critical gap in the development of snow models that systematically integrate these processes with mechanical settling.

In response to the limitations of existing snow models to perform a systematic study, the snow model proposed in this case study has the following objectives:

- Generic coupling of the multi-physics process equations comprising water vapor diffusion, heat conduction and mechanical settling around the ice mass balance.
- Stability through a simple, mathematically rigorous numerical approximation of the process equations.
- Modularity and extendability through a consistently formulated flexible numerical scheme.
- Transparency and understandability through detailed documentation of the model including numerical approximations.
- Reusability of the modeling software though consistency with the computational approach described in the article.

Sect. 4.2 describes the physical and numerical foundations of the modular model and discusses its results, and Sect. 4.3 describes the published modeling software. Sect. 4.4 highlights the reuse potential of the modular and extendable modeling software at the example of fictive reuse scenarios and a real-word reuse example of the modeling software for a model comparison conducted by Brondex et al. (2023b).

4.2 A snow model to couple heat and water vapor transport with mechanical settling

The aim of this section is twofold. First, I describe the multi-physics process equations (Sect. 4.2.1) followed by the applied numerical strategy for a phase-changing snowpack (Sect. 4.2.2). The numerical scheme is hybrid, in the sense that it clearly discriminates between a solution of the mechanical settling operator by means of a Lagrangian approach and a solution to the heat and water vapor transport operator by means of an Eulerian approach. To some degree, the numerical model description must be understood as a rigorous re-formulation of the numerical schemes from existing

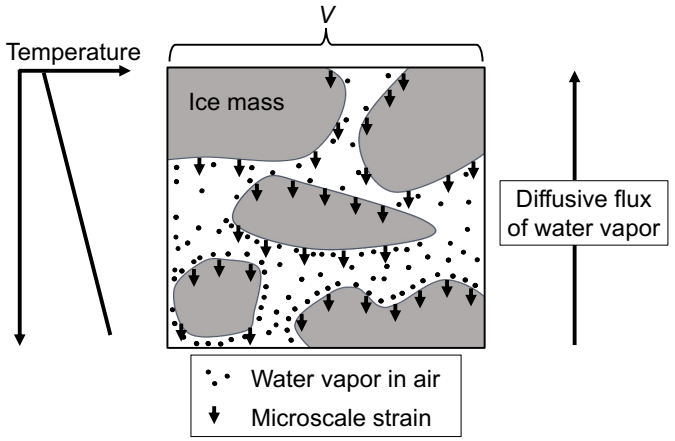


FIGURE 4.2: Representative elementary volume (V) of dry snow consisting of ice mass and water vapor and air. The ice mass changes due to diffusive flux of water vapor related to sublimation and deposition.

computational snowpack models SNOWPACK and Crocus. Yet, in addition to existing schemes the proposed model (a) explicitly separates the Eulerian and Lagrangian part of the solver to facilitate a later modular adaption and (b) provides a full finite-difference formulation including correction terms due to the deforming (non-uniform) mesh that are typically omitted. Second, I present a systematic model cascade of different process modules, which in their most comprehensive version, correspond to the fully coupled system of heat conduction, water vapor diffusion and mechanical settling (Sect. 4.2.3). The computational potential is demonstrated by applying the model cascade in form of eight benchmarks and analyzing the simulation results for an idealized two-layer, dry-snow situation (Sect. 4.2.4).

4.2.1 Physical model of dry snow

In the most general case, snow is a mixture of ice, air, water vapor, and liquid water, and the snow density is given as a mixture of the respective pure densities (Morland et al., 1990; Bader and Weilenmann, 1992). The amount of ice in one reference volume of snow is $\phi_i \rho_i V$ in which ϕ_i denotes the ice's volume fraction, ρ_i its pure density, and V the volume of the reference volume. The structure and volume fraction of the ice can change over time either due to strain-induced settling processes or due to transient phase changes, such as sublimation, deposition, melting, and freezing. The proposed model allows for only water vapor as secondary phase besides ice, which constitutes a dry snow situation so that snow density can be approximated as $\phi_i \rho_i$.

As a common starting point, snow models take a macroscale perspective that volume averages (Bader and Weilenmann, 1992; Bartelt and Lehning, 2002; Hansen and Foslien, 2015a) or homogenizes (Calonne et al., 2014) the snowpack's microstructural state into macroscale variables. If not stated otherwise, all state variables are assumed to be macroscale variables. State variables, model parameters, and constants used in the context of the model are summarized in Table 4.1.

The system of generic process equations of the proposed model is built around the ice mass balance, which is an indicator of snow density. The ice mass balance is introduced first, followed by mechanical settling and the transport of heat and water vapor.

Ice mass balance

The ice volume fraction $\phi_i = \phi_i(z,t)$ within a spatio-temporally evolving snowpack of varying snow height $H(t)$ is governed by the ice mass balance and reads

$$\partial_t \phi_i + \nabla \cdot (\mathbf{v}\, \phi_i) = \frac{c}{\rho_i}, \tag{4.1}$$

with velocity field $\mathbf{v}$, source term c, and ice density ρ_i (Bader and Weilenmann, 1992; Hansen and Foslien, 2015a).

In a 1D situation, the vertical position z is the only relevant spatial coordinate ($z \in [0, H(t)]$), and the velocity field $\mathbf{v}$ reduces to vertical velocity $v = v(z,t)$ for each time and position within the column. Velocity is negative for snow height decrease and positive for snow height increase. Vertical motion results either from mechanical settling, hence a consolidation or compaction of the snowpack, or alternatively it is a continuity response to changes in ice volume from sublimation and deposition via the source term $c = c(z,t)$ (Bader and Weilenmann, 1992). The continuity response varies with time and position and leads to a minor vertical decrease/increase of snow height. In the following, c will be referred to as deposition rate. c is positive (production) if new ice is built, namely water vapor deposits, and it is negative if ice is lost, namely sublimates. Although effects due to consolidation of snow may be significantly more pronounced than those due to phase change processes in the pore space, the latter needs to be accounted for to acknowledge mass conservation of the complete system. At this point in time, I do not consider any additional increases of snow height due to precipitation, and I guide the interested reader to the discussion of such an implementation in Sect. 6 *Future work and challenges* of the comprehensive article by Simson et al. (2021). Finally, ρ_i denotes the constant pure density of ice and serves as a scaling factor.

The ice mass balance (Eq. (4.1)) couples mechanical settling ($\mathbf{v}$) and phase change processes (c). Considering the equation in its full form is essential for the goal to model and eventually analyze the interplay between these processes. The reader should note that the structure of the ice mass balance resembles an advection-reaction equation that can conveniently be solved by means of Lagrangian type computational methods, such as the method of characteristics (see Sect. 4.2.2). Yet in order to do so, a closure for both vertical velocity v and deposition rate c is needed.

A closure for the velocity field

Velocity v represents mechanical deformation in the snowpack. Its idealized relation to the strain rate is given by

$$\nabla v = \dot{\epsilon}. \tag{4.2}$$

It should be noted that this is simplified with respect to more general, tensorial formulations of 1D consolidation theories, see for instance Audet and Fowler (1992). Yet even the idealized formulation Eq. (4.2) will be sufficient for the purpose, as it resembles the approach implicitly chosen in detailed snowpack models (Bartelt and Lehning, 2002; Vionnet et al., 2012).

In general, one would expect that porous snow inherits the non-linear constitutive behavior of ice (Kirchner et al., 2001), which leads to

$$\dot{\epsilon} = \frac{1}{\eta}\sigma^m \tag{4.3}$$

and is a variant of Glen's law. Here, η denotes the compactive viscosity of snow, and σ denotes the stress. The choice of the Glen exponent m in earlier works depends on both the physical regime and the computational feasibility. The linear form of Glen's law ($m = 1$) is chosen in Vionnet et al. (2012) and Bartelt and Lehning (2002). For the sake of comparability, I thus mainly use a linear version of Glen's law, hence $m = 1$. The framework, however, also works with the non-linear relation, such as $m = 3$, which is also used in one of the benchmarks.

The compactive viscosity η depends on the snow's microstructure, and it is challenging to determine from experiments (Wiese and Schneebeli, 2017). In snow models, it is typically provided as a parametrized closure for a specific physical situation, and it strongly correlates with the choice for the Glen exponent m. This fact clearly constrains its universal applicability and makes any transfer of a model validated viscosity parametrization to other physical situations challenging. The benchmarks in Sect. 4.2.4 consider both constant and dynamic viscosities. The dynamic viscosity is based on an empirical viscosity closure $\eta = \eta(\phi_i, T)$ provided by Vionnet et al. (2012). The respective equations and the full procedure to derive viscosity can be found in Appendix B.1.3.

In the absence of strong horizontal deformation and deviatoric stress components, it is reasonable to assume a stress-free condition at the snow's surface and a hydrostatic stress condition in its interior:

$$\nabla\sigma = g\,\rho_{snow}. \tag{4.4}$$

g is the gravitational acceleration and ρ_{snow} refers to the snow's density, which is clearly dominated by the ice fraction via $\rho_{snow} \approx \phi_i(z)\,\rho_i$. It varies with the position z in the snow column due to a vertically varying ice volume fraction $\phi_i(z)$. Integration of Eq. (4.4) and combination with Eqs. (4.2) and (4.3) yields an expression for the velocity gradient:

$$\partial_z v = \frac{1}{\eta}\left(g\int_z^{H(t)} \phi_i\,(\zeta)\,\rho_i\,d\zeta\right)^m. \tag{4.5}$$

ζ is the integration variable. A second integration along the vertical axis finally yields an expression for the velocity at position z in the snow column:

$$v(z) = -\int_0^z \frac{1}{\eta}\left(g\int_{\tilde{z}}^{H(t)} \phi_i(\zeta)\rho_i d\zeta\right)^m d\tilde{z}, \tag{4.6}$$

in terms of total height $H(t)$, ice volume fraction $\phi_i(z,t)$, and with velocity $v(z=0,t) \equiv 0$. This definition of the vertical velocity yields a process that complies with the obvious physical constraints: a) The velocity vanishes at the bottom of the snow column, hence prevents artificial penetration into the ground, This is similar to displacement requirements in SNOWPACK (Bartelt and Lehning, 2002). b) The vertical velocity accumulates with height, which prevents any artificial disaggregation of the snowpack, and c) the vertical velocity relaxes towards zero as the ice volume fraction tends towards its maximum volume fraction $\phi_i < \phi_{i,max} < 1$. In the remainder of this chapter, I use Eq. (4.6) to account for the mechanical settling of the snowpack.

Heat and water vapor transport

The ice deposition rate c typically depends on coupled heat and mass transport for the involved phases ice, water and water vapor. The proposed model considers the process model from Hansen and Foslien (2015a) that reflects a dry snow condition in which void space is filled by water vapor only. It should be noted however, that this coupled process model could readily be substituted or extended by another one, e.g., Calonne et al. (2014), Jafari et al. (2020), Schürholt et al. (2022). Next, I state the essential aspects and process equations of the model proposed in Hansen and Foslien (2015a), and I describe how it can be used to recover the ice deposition rate.

Assuming a dry snow condition, the ice production is solely determined by mass transport between water vapor and ice. The water vapor mass balance reads

$$\partial_t \left(\rho_v \left(1 - \phi_i \right) \right) - \nabla \cdot \left(D_{eff} \, \nabla \rho_v \right) = -c, \tag{4.7}$$

in which ρ_v denotes the water vapor density and D_{eff} the effective water vapor diffusion coefficient. Water vapor production corresponds to negative ice deposition rate $-c$ that represents sublimation. Following Hansen and Foslien (2015a), water vapor density in the pore space can be assumed to be at saturation density ρ_v^{eq}, so that $\rho_v \equiv \rho_v^{eq}$. The latter is well investigated, and empirical relations exist that specify its temperature dependency $\rho_v^{eq}(T)$. In this work, I employ an empirical relation from Libbrecht (1999). The full expression can be found in Appendix B.1.1. Due to the closure for water vapor density ρ_v^{eq}, the water vapor mass balance (Eq. (4.7)) can be rewritten using the temperature dependence of the equilibrium water vapor density

$$\left(1 - \phi_i \right) \, \frac{d\rho_v^{eq}}{dT} \, \partial_t T - \nabla \cdot \left(D_{eff} \, \frac{d\rho_v^{eq}}{dT} \nabla T \right) = -c. \tag{4.8}$$

Assuming the snow is in thermal equilibrium at the microscale, the energy balance can likewise be written in terms of the temperature, which reads

$$\left(\rho C \right)_{eff} \, \partial_t T - \nabla \cdot \left(\, k_{eff} \, \nabla T \, \right) = c \, L. \tag{4.9}$$

The parameters $(\rho C)_{eff}$ and k_{eff} stand for the effective heat capacity of snow and effective thermal conductivity respectively. Both parameters depend on the ice volume fraction, and their definition

is stated in Appendix B.1.2. The right hand side of the heat equation (Eq. (4.9)) accounts for latent heat release, which is coupled to phase change processes.

The system of the two equations, Eqs. (4.8) and (4.9), and the two unknowns, temperature T and deposition rate c, is solved by replacing c in Eq. (4.9) with Eq. (4.8), which yields a non-linear equation for temperature

$$\left((\rho C)_{eff} + (1 - \phi_i) \, \frac{d\rho_v^{eq}(T)}{dT} \, L \right) \partial_t T = \nabla \cdot \left(\left(L \, D_{eff} \, \frac{d\rho_v^{eq}(T)}{dT} + k_{eff} \right) \nabla T \right). \tag{4.10}$$

The spatio-temporal temperature evolution is then used to recover the ice deposition rate c from either Eq. (4.8) or Eq. (4.9).

TABLE 4.1: Terminology of state variables, model parameters and constants

Symbol	Name	Equation/Value	Unit
	State variables		
ϕ_i	Ice volume fraction	Eq. (4.1)	-
ρ_v	Water vapor density	Eq. (B.1)	$\mathrm{kg\,m^{-3}}$
v	Vertical velocity	Eq. (4.6)	$\mathrm{m\,s^{-1}}$
c	Ice deposition rate	Eq. (4.8)	$\mathrm{kg\,m^{-3}\ s^{-1}}$
T	Temperature	Eq. (4.10)	K
	Model parameters of snow		
$\dot{\epsilon}$	Strain rate	Eq. (4.3)	$\mathrm{s^{-1}}$
η	Viscosity	Eq. (B.5)	$\mathrm{Pa\,s}$
σ	Stress	Eq. (4.4)	$\mathrm{Pa\,m^{-2}}$
ρ_{snow}	Density	Eq. (4.4)	$\mathrm{kg\,m^{-3}}$
D_{eff}	Water vapor diffusion coefficient	Eq. (B.1.1)	$\mathrm{m^2\,s^{-1}}$
$(\rho C)_{eff}$	Heat capacity	Eq. (B.4)	$\mathrm{J\ m^{-3}\ K^{-1}}$
k_{eff}	Thermal conductivity	Eq. (B.3)	$\mathrm{W\ m^{-1}\ K^{-1}}$
	Constants (Calonne et al., 2014)		
ρ_i	Ice density $\phi_i = 1$	917	$\mathrm{kg\ m^{-3}}$
L	Ice latent heat of sublimation	2835333	$\mathrm{J\ kg^{-1}}$
C_i	Ice heat capacity $\phi_i = 1$	2000	$\mathrm{J\ kg^{-1}\ K^{-1}}$
ρ_a	Air density $\phi_i = 0$	1.335	$\mathrm{kg\ m^{-3}}$
C_a	Air heat capacity $\phi_i = 0$	1005	$\mathrm{J\ kg^{-1}\ K^{-1}}$

4.2.2 Modular and extendable Eulerian–Lagrangian computational approach

It is well known that any numerical strategy that aims at simulating simultaneous settling-induced deformation of the snowpack and (arbitrary) diffusive transport requires a special computational treatment to couple both. Diffusive transport is best modeled by taking an Eulerian perspective, hence on a static mesh. In this section, I present the advantage of a clear and explicit separation into a Lagrangian deformation module that accounts for mechanical settling and a Eulerian heat and water vapor transport module.

The complete process model is given by the ice mass balance Eq. (4.1), its mechanically induced vertical velocity Eq. (4.6), and the coupled system for temperature Eq. (4.10) and ice deposition

rate determined by either Eq. (4.8) or Eq. (4.9). Each of the equations will be solved in a separate module. The ice mass balance in conjunction with the vertical velocity has the form of a non-linear advection equation, whereas the remaining equations are of parabolic nature. Based on the distinction into diffusion and advection dominated processes, the model follows a two-step solution scheme:

Step 1 accounts for the mesh deformation and solves the advection dominated mechanical settling, i.e., the ice mass balance Eq. (4.1), by means of a Lagrangian approach that tracks the movement of the coordinates including changes from metamorphism.

Step 2 determines the spatio-temporal evolution of temperature and deposition rate fields as introduced in Sect. 4.2.1 based on an Eulerian approach that solves the diffusion dominated transport of heat and water vapor via a finite difference implementation on a deformed (unstructured) mesh.

A finite difference method is employed because it provides a feasible algorithm that is applicable to the benchmarks considered in Sect. 4.2.3 using a 1D snow column. It also naturally integrates with the Lagrangian part of the solution (Step 1), as it can use the same mesh. In principle, it is also possible to couple the two-step approach with a finite element solution for temperature and deposition rate, for instance, when aiming for a 2D or 3D model in a complex geometry that incorporates realistic mountain slope topographies. When using a finite element solver, it has to be kept in mind that deposition rate and temperature fields need to be reconstructed from the solution at each time step. Especially when wanting to use higher order elements this might limit computational feasibility.

The solution scheme alternates both steps via straightforward first order operator splitting. This works well for the benchmarks, yet could be readily exchanged with a higher order splitting scheme such as a second order Strang splitting (LeVeque, 2002), if required.

The computational model is implemented in Python and it is modular and extendable, in the sense that each module can separately be activated and deactivated. This not only simplifies the verification of individual process building blocks but also allows an in-depth investigation of the various coupling effects and the model's non-linear feedback. Alternative formulations of, for instance, the parametrized velocity field are implemented and can easily be exchanged. Finally, the modular structure facilitates the implementation of additional closure relations or the integration of entire new process modules.

Computational grid

The proposed model considers a 1D snow column, which is discretized into $nz+1$ spatial mesh nodes denoted by z_k with $k \in \{0, 1, .., nz\}$. The computational domain consists of 101 computational nodes ($nz = 100$) except for some simulations that required a higher resolution of 251 nodes ($nz = 250$). The mesh is non-uniform in general, meaning that the distance between neighboring nodes $z_{k+1} - z_k$ varies throughout the snow column and with time. The reader should note that the z-axis is oriented opposing gravitational acceleration, such that z_0 denotes the position of the ground and z_{nz} the

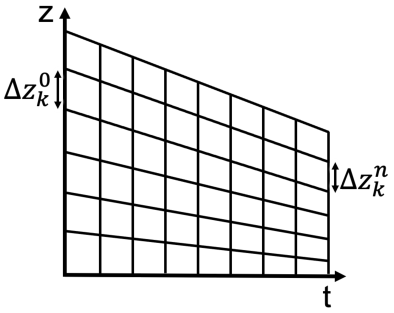


FIGURE 4.3: Computational mesh. The snowpack height varies with time, i.e., it shrinks due to settling of the snow. This has to be incorporated into the computational mesh, which undergoes deformation due to the downward movement of the free surface. The initially equidistant mesh does not uniformly change, which results in a mesh of varying node distances, so that in general $\Delta z_k^0 \neq \Delta z_k^n$ and $\Delta z_k^n \neq \Delta z_{k+1}^n$.

position of the snowpack's free surface. Time increments are denoted by t_n with $n \in \{0, 1, .., nt\}$ and nt being the maximum number of time steps in a complete simulation run. For each of the field variables subscript k denotes the vertical coordinate and superscript n denotes the time step hence $T(z_k, t_n) = T_k^n$.

Lagrangian solution of the ice mass balance

When the snowpack is subject to vertical motion in form of settling, its physical height decreases, hence its vertical extent shrinks. One option to reflect this in a computational method is to adjust the spatial node coordinates accordingly. The challenging fact in this situation is that the vertical motion within the snow column (non-linear advection) is coupled to phase changes, i.e., a change in ice volume fraction via the source term in the ice mass balance (Eq. (4.1)). The method of characteristics is a suitable method to solve such a non-linear advection equation with source term. It can be interpreted as a simultaneous motion tracking of snow material elements or *particles*, referred to as the integration along so-called characteristics, while also accounting for its phase change-induced metamorphism along the trajectory. By construction, the method correctly tracks the snowpack's moving free surface. Due to the fact that the snow column's evolution is determined with respect to a material particle that moves vertically at speed v in the snowpack, the method of characteristics is called a Lagrangian approach.

In order to derive the specific update rule for the ice mass balance (Eq. (4.1)) the product rule is applied to its initial Eulerian version

$$\partial_t \phi_i + v\, \partial_z \phi_i = \frac{1}{\rho_i} c - \phi_i\, \partial_z v, \tag{4.11}$$

and then the equation is re-formulated to fit with a Lagrangian reference frame, hence with respect to nodes moving at the vertical velocity v. Changing to the moving reference frame effectively

compensates the advection term in Eq. (4.11) and yields

$$\partial_t \phi_i = \frac{1}{\rho_i} c - \phi_i \, \partial_z v \tag{4.12}$$

$$\partial_t z = v. \tag{4.13}$$

Equation (4.13) accounts for the settling of material particles within the snowpack. The equation is used to update the coordinates of the mesh nodes directly, which results in a continuous mesh deformation as illustrated in Fig. 4.3. Equation (4.12) captures the evolution of the ice volume fraction along the trajectory of a moving ice material particle within the snowpack. It accounts for volume changes due to a) mass production/loss in response to phase changes, and b) vertical variation of the vertical velocity. Further details and generalizations of the method of characteristics can be found in Farlow (1993).

Equations (4.12) and (4.13) can be solved analytically for a constant vertical velocity and deposition rate. In this case, the velocity closure is provided by Eq. (4.6) and the deposition rate results from solving yet another process model (Eqs. (4.9) and (4.10)), which requires a numerical solution. The response of the ice volume fraction can be expected to be slow with respect to other processes in the system, so that a first order explicit Euler time integration scheme should be sufficient:

$$\phi_{i,k}^{n+1} = \phi_{i,k}^{n} + \Delta t^n \left(\frac{1}{\rho_i} c_k^n - \phi_{i,k}^n \, \partial_z v_k^n \right) \tag{4.14}$$

$$z_k^{n+1} = z_k^n + \Delta t^n \, v_k^n, \tag{4.15}$$

The update of the mesh coordinates according to Eq. (4.15) for the vertical velocity closure derived before requires a numerical approximation of Eq. (4.6) at each node z_k. This approximation results in

$$v(z_k) = -\sum_{j=0}^{k} \left(\frac{1}{\eta} \sigma_j^m \right) \Delta z_j^n \tag{4.16}$$

with $\Delta z_j^n := z_{j+1}^n - z_j^n$ with $j \in [0, nz[$, m Glen exponent, η viscosity and σ_j denoting the stress exerted by the overburdened snow mass

$$\sigma_j = \sum_{l=j}^{nz} g \, \phi_{i,l} \, \rho_i \, \Delta z_l^n. \tag{4.17}$$

with g gravitational acceleration. The reader should not that the stress at the uppermost node $k = nz$ is zero, so that velocity $v(z_{nz})$ is only controlled by the movement below, and it is thus equivalent to the velocity at the next lower node ($v(z_{nz-1})$). The forward Euler scheme of Eqs. (4.14) and (4.15) via the method of characteristics combined with the velocity update (Eq. (4.16)) essentially resembles the treatment of mass conservation as it is presently done in SNOWPACK. However, the explicit formulation and numerical treatment of Eqs. (4.14) and (4.15) allows to employ also other (e.g., higher order, implicit) solution schemes for both equations, if this was required to capture detailed aspects of the spatio-temporal coupling of phase changes and settling via $\partial_z v$. To solve

Eq. (4.14), the velocity's spatial derivative $\partial_z v$ is discretized, which corresponds to the strain rate $\dot{\epsilon}_k^n = \frac{1}{\eta}\,\sigma_k^m$ given via Eq. (4.3). This is beneficial as it avoids numerically approximating the velocity gradient. The complete numerical update of ice volume fraction ϕ_i and mesh coordinates z can now concisely be written as

$$\phi_{i,k}^{n+1} = \phi_{i,k}^n + \Delta t^n \left(\frac{1}{\rho_i} c_k^n + \frac{1}{\eta} \left(\sum_{l=k}^{nz} g\, \phi_{i,l}^n\, \rho_i c, \Delta z_l^n \right)^m \phi_{i,k}^n \right) \tag{4.18}$$

$$z_k^{n+1} = z_k^n + \Delta t^n \left(\sum_{j=0}^{k} \frac{1}{\eta} \left(\sum_{l=j}^{nz} g\, \phi_{i,l}^n\, \rho_i\, \Delta z_l^n \right)^m \Delta z_j^n \right). \tag{4.19}$$

Similar to existing layer-based schemes (see for instance Sect. 3.4. in Bartelt and Lehning (2002) or its recent extension Jafari et al. (2020)) the method of characteristics provides information on the settling of layers within the snowpack. Yet in addition, it serves as a basis for a fully modular and flexible computational strategy that a) accounts by construction for the two-way feedback between the ice volume fraction and mass production or decay rates resulting from phase changes as a response to transport processes within the snowpack, b) allows for a flexible adaptation/extension of the process model, used to determine c, and the velocity closure. The latter could for instance serve as a pathway to integrate a data-driven velocity closure (or assimilation) from measurements. Such flexibility in numerical tools will be important in the future to conduct model comparisons, such as presented in Schürholt et al. (2022) within holistic snowpack models. A remaining difficulty now is to provide a Eulerian numerical scheme for diffusive processes that can operate on a spatially varying unstructured mesh.

Eulerian solution of transport and phase changes on a moving mesh

The process model accounting for water vapor transport and heat transport (Eqs. (4.8) and (4.10)) has to be solved with respect to a moving computational mesh according to Eq. (4.15). Both equations have the same generic structure, namely

$$\alpha\, \partial_t T - \partial_z \left(\beta\, \partial_z\, T \right) = \gamma \tag{4.20}$$

with $\alpha = \alpha_T = (\rho C)_{eff} + (1-\phi_i)\,\frac{d\rho_v^{eq}(T)}{dT} L\,, \quad \beta = \beta_T = k_{eff} + L\, D_{eff}\,\frac{d\rho_v^{eq}(T)}{dT}$ and $\gamma = \gamma_T = 0$ for heat equation Eq. (4.10) and $\alpha = \alpha_c = (1-\phi_i)\,\frac{d\rho_v^{eq}(T)}{dT}\,, \quad \beta = \beta_c = D_{eff}\,\frac{d\rho_v^{eq}(T)}{dT}$ and $\gamma = \gamma_c = -c$ for water vapor transport equation Eq. (4.8). An implicit first order finite difference approximation of Eqs. (4.8) and (4.10) for a spatially varying mesh of increments Δz_k^n results in

$$\begin{aligned} \alpha_{T,k}^n \frac{T_k^{n+1} - T_k^n}{\Delta t^n} = & \frac{2\,\beta_{T,k}^n \left(\left(T_{k+1}^{n+1} - T_k^{n+1} \right) - \left(T_k^{n+1} - T_{k-1}^{n+1} \right) \right)}{\left(\Delta z_k^n \right)^2 + \left(\Delta z_{k-1}^n \right)^2} + \\ & \frac{\beta_{T,k+1}^n - \beta_{T,k-1}^n}{\Delta z_k^n + \Delta z_{k-1}^n} \frac{T_{k+1}^{n+1} - T_{k-1}^{n+1}}{\Delta z_k^n + \Delta z_{k-1}^n} + E_T(T_{k+1}^{n+1}, T_{k-1}^{n+1}) \end{aligned} \tag{4.21}$$

$$\alpha_{c,k}^{n}\frac{T_k^{n+1}-T_k^n}{\Delta t^n} = \frac{2\,\beta_{c,k}^{n}\,\left(\left(T_{k+1}^{n+1}-T_k^{n+1}\right)-\left(T_k^{n+1}-T_{k-1}^{n+1}\right)\right)}{\left(\Delta z_k^n\right)^2+\left(\Delta z_{k-1}^n\right)^2} - c_k^{n+1}+\frac{\beta_{c,k+1}^{n}-\beta_{c,k-1}^{n}}{\Delta z_k^n+\Delta z_{k-1}^n}\frac{T_{k+1}^{n+1}-T_{k-1}^{n+1}}{\Delta z_k^n+\Delta z_{k-1}^n}+E_c(T_{k+1}^{n+1},T_{k-1}^{n+1}). \tag{4.22}$$

The parameters α_f and β_f for $f \in \{T, c\}$ vary in space and time, and they are explicitly derived based on the snowpack's state at time n. The terms E_c and E_T are higher-order mesh errors for the water vapor and temperature equations. These higher-order mesh errors account for the necessary correction due to non-uniformity of the mesh and are controlled by the temperature gradient; they vanish for equidistant meshes or constant temperatures. The complete form of the higher-order mesh errors is given in Appendix B.2, and their effect on the accuracy of the simulation is discussed in Sect. 4.2.4.

The complete numerical update can be concisely written in matrix form, which matches with the way it is implemented in the software

$$\vec{T}^{n+1} = \left(\mathbf{A_T}+\mathbf{E_T}\right)^{-1}\left(\mathbf{B_T}\vec{T}^{n}\right) \tag{4.23}$$

$$\vec{c}^{n+1} = \left(\mathbf{A_c}+\mathbf{E_c}\right)\vec{T}^{n+1}+\mathbf{B_c}\vec{T}^{n}. \tag{4.24}$$

First, Eq. (4.23) is solved for temperature T^{n+1}. Next, the updated temperature is used to solve Eq. (4.24) for the deposition rate c^{n+1}. The complete matrix definitions are given in Appendix B.3. Formally, it would be possible to add up matrices $\mathbf{A_T}$ and $\mathbf{E_T}$ as well as $\mathbf{A_c}$ and $\mathbf{E_c}$. The separation by this particular form stresses the similarity of this formulation with a standard finite difference approximation on an equidistant mesh in which $\mathbf{E_T}$ and $\mathbf{E_c}$ vanish.

Iterative coupling of Eulerian and Lagrangian solutions

The derived numerical update routines for temperature, deposition rate, vertical velocity and ice volume fraction comprise the four main modules that are sequentially called to update the respective state variables for one time step. A schematic illustration is given in Fig. 4.4. The equations for heat and water vapor transport have already been implemented by Calonne et al. (2014) and Hansen and Foslien (2015a). A feedback on the ice volume fraction in the absence of a vertical velocity has been investigated in the companion article (Schürholt et al., 2022). The modules for vertical velocity and the coupled update of ice volume fraction and mesh coordinates, through the method of characteristics is novel in this approach. The implementation is modular in the sense that it allows for a coupling with other process models that comply with a non-uniform mesh.

The time step size for the next time step $n+1$ is dynamically updated in the computational scheme. Since diffusive processes are dominant, mesh Fourier number is based on heat diffusivity $\frac{\beta_T}{\alpha_T}$ of the current time step n

$$\Delta t^{n+1} = \min_k\left(0.5\,\frac{\alpha_{T,k}^{n}}{\beta_{T,k}^{n}}\left(\Delta z_k^n\right)^2\right). \tag{4.25}$$

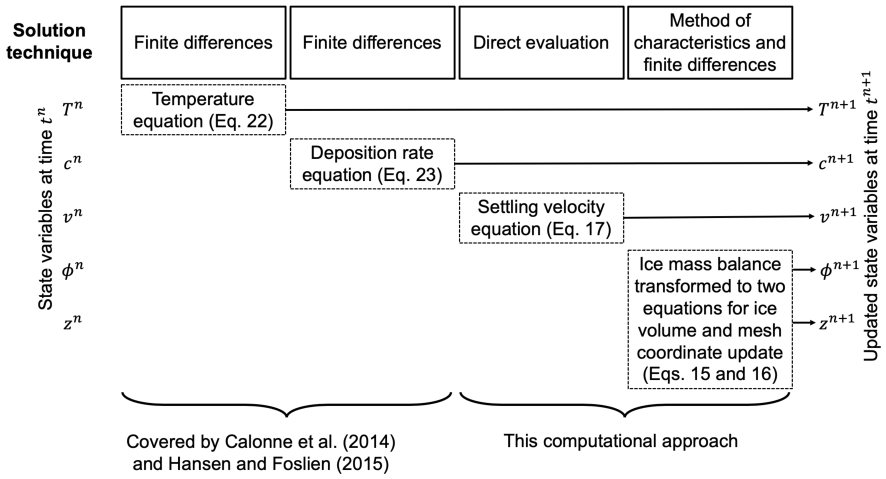


FIGURE 4.4: Illustrates the modular computational workflow of one iteration. The state variables at time t^n, depicted on the left hand side, are updated through the modules annotated as dashed boxes. After each update the state variables at time t^{n+1} are retrieved. The equations of the modules are implemented into the computational model through the respective solution technique stated in the solid boxes on the top row. The computational steps are carried out from top to bottom. The iterative approach can be summarized as: 1) determine time step size Δt according to Eq. (4.25), 2) update the temperature field based on Eq. (4.23), 3) compute the deposition rate with the temperature field based on Eq. (4.24), 4) determine the vertical velocity with Eqs. (4.16) and (4.17), and 5) update the ice volume fraction and the mesh coordinates simultaneously based on Eqs. (4.18) and (4.19). While 2) and 3) is a re-implementation of an existing approach previously published by Hansen and Foslien (2015a) and Calonne et al. (2014), their coupling to 4) and 5) constitutes the novelties of this work.
Note that 4) is computed as part of 5) in the code.

This choice for the time step computation did not yield instabilities. Thus water vapor's diffusivity is excluded for the time step computation. In response to settling processes, the mesh sizes vary and decrease (see Fig. 4.3) with time, and so does the time step.

In the following, the modularity of the model is leveraged to asses the individual effect of the different process building blocks by a strategical activation and deactivation of the modules illustrated in Fig. 4.4.

4.2.3 Application of the model

The developed numerical scheme is used to perform several simulations with varying combinations of *activated* and *deactivated* advection- and diffusion-type process building blocks. Theses combinations consider heat conduction in interaction with water vapor diffusion, such as also considered in Schürholt et al. (2022), and additionally the observation of both processes in the presence of settling. Furthermore, this scheme allows the numerical verification of separate building blocks. While the benchmarks are still idealized, they demonstrate the robustness of the Eulerian–Lagrangian scheme against a selection of varying sub-sets of model components. Table 4.2 provides an overview of the various combinations considered as benchmarks in Simson et al. (2021). While all of the benchmarks in Table 4.2 are introduced in this thesis, only results of the benchmarks printed in bold (Cases

TABLE 4.2: List of the various benchmarks, referred to as *cases*. Each benchmark represents a different combination of activated process modules considering either a constant or dynamic viscosity closure. Heat transport induces water vapor transport. Therefore, water vapor transport is always active in combination with heat transport. Cases 5, 7 and 8 are also referred to as *fully coupled processes*.

Case	**Heat transport (Eq.** (4.23)**)**	**Water vapor transport (Eq.** (4.24)**)**	**Mechanical settling (Eqs.** (4.18) **and** (4.19)**)**			
			Viscosity Eq. (B.5)		**Glen's law (Eq.** (4.16)**)**	
			$\eta = const$	$\eta(\phi_i, T)$	$m = 1$	$m = 3$
Case 1			✓		✓	
Case 2	✓					
Case 3	✓		✓		✓	
Case 4	✓	✓				
Case 5	✓	✓	✓		✓	
Case 6				✓	✓	
Case 7	✓	✓		✓	✓	
Case 8	✓	✓	✓			✓

1-6) are presented and discussed as they sufficiently demonstrate the advantages of the modularity of the approach. The interested reader can refer to Sect. 4 *Results and Discussion* in Simson et al. (2021) for the simulation results of the excluded benchmarks.

Benchmark overview

Firstly, Case 1 focuses on the effects due to pure mechanical settling on the snowpack. Case 2 considers isolated heat transport, and Case 3 considers additionally the interplay of heat transport with settling processes. Case 4 considers coupled heat and water vapor transport in the absence of settling, and then Case 5 couples both processes with settling. At the example of Case 5, the effect of included or excluded higher-order mesh errors $\mathbf{E}_T$ and $\mathbf{E}_c$ (see Sect. 4.2.2) on the temperature profiles is evaluated. Case 1, Case 3, and Case 5 consider the constant viscosity for linear Glen's law ($m = 1$) $\eta_{const,m=1}$, as introduced in Sect. 4.2.1. Furthermore, the impact of an empirical, temperature and ice volume fraction controlled, viscosity closure (Eq. (B.5)) is investigated. Case 6 considers mechanical settling active only, and Case 7 refers to the fully coupled processes. The general approach can be combined with the non-linear Glen's law (Eq. (4.3)) by using $m = 3$ (Case 8) and an accordingly adjusted constant viscosity $\eta_{const,m=3}$. For a detailed explanation of the general derivation of the viscosity values see Appendix B.1.3 and B.1.4.

Computational setup, initial and boundary conditions

Initial condition: The initial ice volume fraction ϕ_i reflects a layered situation as depicted in Fig. 4.5, with two snow layers of equal thickness. The bottom layer has an initial snow density of 150 kgm^{-3}, and the upper layer's density is 75 kgm^{-3}. The transition from the upper layer to the lower layer is linearly smoothed out over 2 cm, which for a grid defined according to Sect. 4.2.2 corresponds to 5 computational nodes for the coarser and 11 computational nodes for the finer discretization. The snow densities are in the range of *damped new snow* and *new snow* respectively

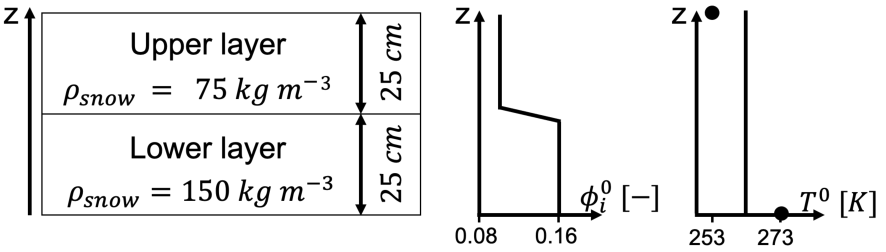


FIGURE 4.5: Shows the initial condition of the snowpack regarding snow density at the left hand side and profile plots of the initial ice volume fraction (ϕ_i^0) and temperature (T^0) at the right hand side. There are two snow layers with equal thickness of 25 cm, yielding a snowpack of height 50 cm height. The bottom layer has a higher density of 150 kgm^{-3} and the upper layer's density is 75 kg m^{-3}. The z-axis of the 1D model increases in upward direction, so that $z = 0$ denotes the ground. Downward directed movements are thus described by negative velocities. The vicinity of the interface between the two layers is referred to as transition area. The initial ice volume fraction is derived from initial snow density. Its profile ($\phi_{i,0}$) shows the linear decrease over 2 cm of ice volume fraction in the transition area from the lower to the upper layer. The initial temperature profile (T_0) is constant at 263 K. The black dots mark the constant temperature boundary conditions: 273 K at the bottom; and 253 K at the top.

(Paterson, 1994). Snow densities in this range are expected for new snow in the European Alps (Helfricht et al., 2018) or in the Rocky mountains (Judson and Doesken, 2000). This layered snowpack ensures an extreme and very active snow regime with a strong dynamical coupling of the processes. Temperature is initially constant at 263 K throughout the whole snowpack. Deposition rate is directly deduced from temperature (see Eq. (4.24)) and therefore requires neither initial nor boundary conditions. Constant viscosity values are derived from the initial condition. They are $\eta_{const,m=1} \approx 9.1 \times 10^7$ Pas and $\eta_{const,m=3} \approx 16 \times 10^{12}$ Pas (see also Appendix B.1.4).

Boundary condition: The temperature is constantly 273 K at the bottom boundary and constantly 253 K at the free surface. **Simulation time:** Simulation times are 2 days (48 h), 3 days (62 h), and 4 days (96 h).

4.2.4 Results and discussion

The following results are all based on the two-layered snowpack described in Sect. 4.2.3 using different combinations of activated process modules as listed in Table 4.2. In the last example, the proposed modeling approach is compared with the detailed snowpack models by mimicking a layer-based scheme.

Settling (Case 1)

The effects of mechanical settling on the snowpack are investigated based on Case 1 (Table 4.2) with a particular focus on the evolution of the vertical velocity (Fig. 4.6 (a)) and the ice volume fraction (Fig. 4.6 (b)). The vertical velocity decreases from top to bottom and relaxes during the first 48 h. Vertical velocity varies more in the lower layer compared to the upper layer within one

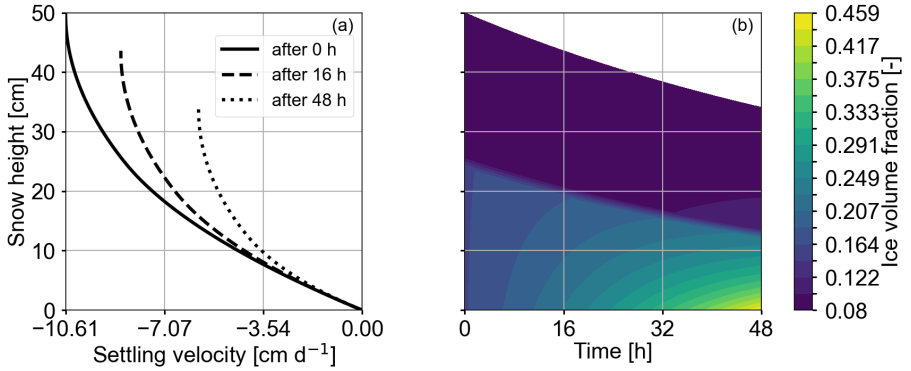


FIGURE 4.6: The plots show the results of Case 1 (Table 4.2), which corresponds to isolated settling effects. (a) shows vertical velocity profiles at the state of the snowpack at initiation, after 16 h and after 48 h. It is clearly visible that vertical velocity and absolute snow height decrease with time. (b) shows ice volume fraction over time. The lower snow layer, where changes in ice volume fraction are clearly discernible can be interpreted as the lower layer and the darker, upper part of the plot as the upper layer. Changes in ice volume fraction are visible in the lower layer; it increases the most at the bottom of the snowpack.

time step. This pattern remains prominent as time proceeds, while the overall velocity variation decreases. This effect is due to the increase of the overburdened snow mass from top to bottom. Settling proceeds the fastest just after the start of the simulation, when the snowpack is at its maximum height, and correspondingly its snow density is at its lowest. Over time, the ice volume fraction increases faster in the lower layer than in the upper layer, and it is the highest at the bottom of the snowpack (Fig. 4.6 (b)). This observation is also visualized in Fig. 4.7, which shows snow density profiles for several times. Furthermore, the extent of the upper layer decreases only slightly, approximately 3.5 cm, over the simulation time, while the lower layer reduces to half of its initial height to approximately 12.5 cm.

The described observations meet expectations, and they reflect the correlation between the amount of compaction and the total overburdened mass. The total settlement of 14 cm after 2 days is a 30% snow height reduction. Bartelt and Christen (1999) simulated 11.6 to 54.8 cm snow height reduction after 5 days for an initially 90 cm high snowpack of 115 kgm^3 density, which is a snow height reduction of 12% to 60%. Taking into account that snow settles slower with increasing density, the results fit to the highest settling rate derived by Bartelt and Christen (1999).

Heat transport in the absence and presence of settling (Cases 2 and 3)

For this simulation, heat transport is considered alone. This simulation is based on Case 2 in Table 4.2. Temperature (Fig. 4.8 (a)), and temperature gradients (Fig. 4.8 (c)) reach a stationary state after approximately 60 h. Heat flux differences between the two layers are clearly visible in the temperature gradient plot. Next, heat transport is superposed by mechanical settling (Fig. 4.8 (b) and (d)), representing Case 3. As a result, the snow height decreases while the internal temperature profiles evolve. Active mechanical processes result in a steeper temperature gradient and hence a higher value of heat flux (Fig. 4.8 (d)). This effect can be attributed to:

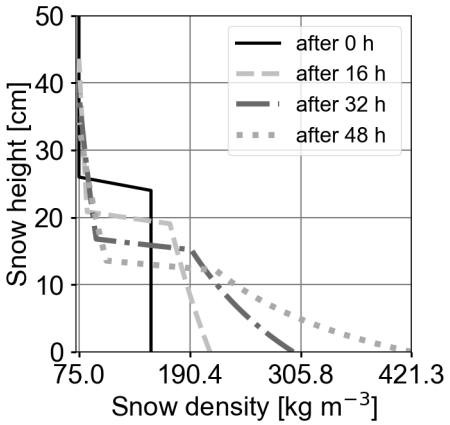


FIGURE 4.7: Snow density profiles at the beginning of the simulation and after 16 h, 32 h and 48 h. The snowpack height decreases and the snow density increases with time at all locations. The increase is especially strong in the lower part of the snowpack.

- the decrease of snow height while keeping the temperatures at the boundaries fixed, and
- the permanent change of thermal conductivity and thermal diffusivity due to their dependency on variations in the ice volume (Eq. (B.1.2)).

The temperature profile reaches a stationary state once the ice volume fraction has reached its maximum value.

Heat and water vapor transport in the absence and presence of settling (Cases 4 and 5)

When considering the water vapor equation from Hansen and Foslien (2015a), diffusion of water vapor requires an apparent temperature gradient, such that the evolution of water vapor transport can only be considered in conjunction with heat transport. Fig. 4.9 (a) shows deposition rate (negative for sublimation) due to heat and water vapor transport only (Case 4 in Table 4.2), and it also shows deposition rate with additionally activated settling processes, representing the fully coupled processes (Case 5 in Table 4.2). Both profiles are characterized by moderate deposition rates throughout the snow column with a pronounced negative (sublimation) peak at the center of the snow column, which is located in the transition area of the layers. The profile for the fully coupled processes shows a sublimation peak approximately 4 times higher than for Case 4. Figure 4.9 (b) shows the temporal evolution of the fully coupled processes (Case 5). In the first hours, sublimation is low in the transition area. After approximately 6 hours, the pronounced sublimation rate peak, as already described for (a), develops and increases until the end of the simulation (48 h). The increased sublimation in the layer transition area may be driven by strong water vapor density gradients (Fig. 4.9 (c)) above the transition area that can be inferred from a strong, local temperature gradient (Fig. 4.9 (d)). This temperature gradient is further enhanced (Fig. 4.8 (d)) by compaction due to settling for Case 5. Case 5 yields even stronger variations of the material properties in the transition area than without compaction and explains the stronger sublimation rates for the fully coupled processes.

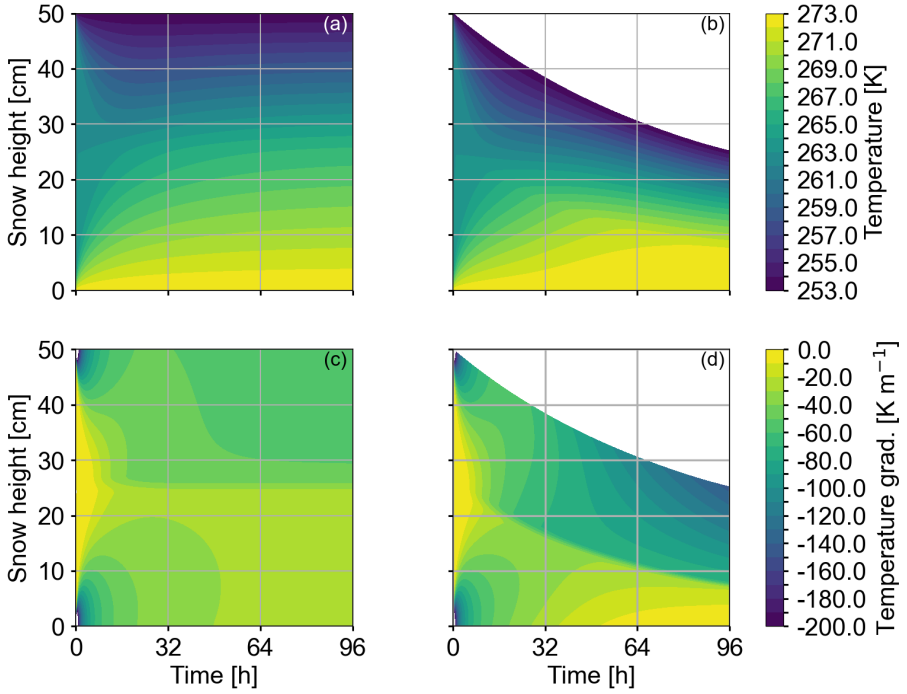


FIGURE 4.8: (a) and (c) show the results for Case 2 of Table 4.2 corresponding to heat transport solely active. (b) and (d) show the results for Case 3 of Table 4.2 corresponding to active heat transport and mechanical settling. For each plot, y-axis represents snow height and x-axis time. (a) and (b) show the temperature evolution, and (c) and (d) show the respective temperature gradients. In (a) and (c), settling is inactive so that the boundary of upper layer and lower layer is at the snowpack center. In (b) and (d), the upper, darker part, which is characterized by higher gradients, is interpreted as the upper layer and the lower, lighter area with lower gradients as the lower layer. The initial condition for both cases is equivalent (see Fig. 4.5). In Case 2, the temperature profile (a) has reached the stationary, piecewise linear profile after approximately 60 h. For Case 3, the temperature profile (b) is not yet stationary at the end of the simulation (96 h) as mechanical processes are still yielding a change in ice volume fraction. The temperature gradient (c) will become constant, only when the maximum ice volume fraction has been reached.

Furthermore, both profiles in Figure 4.9 (a) show a small peak of deposition rate (positive x-direction) just above the aforementioned sublimation rate peak. This peak is very weak for Case 4 and more prominent for Case 5. This deposition rate peak is highly interesting as it is interpreted as the onset of spatio-temporal oscillations as observed and investigated in greater detail in the companion article (Schürholt et al., 2022). Schürholt et al. (2022) describe these wiggles as “smooth oscillations" that are “intrinsic features" of the equations. The results in Fig. 4.9 (a) nicely demonstrate that a) the Eulerian–Lagrangian scheme can capture this behavior, and b) that the instability prevails and even increases in the presence of settling processes. The results suggest that mechanics likely increase local phase change activity in the vicinity of layer boundaries, which potentially has a large effect on weak layer formation.

The deposition rates obtained with the model are between -2 and 2 kgm^{-3}d^{-1}, which fits to the range of -1.728 to 1.728 kgm^{-3}d^{-1} presented in Jafari et al. (2020). Sublimation rate peaks on the order of 0.1 to 1.2 kgm^{-3}d^{-1} have also been computed with the numerical test cases by Hansen and Foslien (2015a). For comparison with experiments, deposition rates can be derived via $SSA \cdot v_n \cdot \rho_i$, with v_n the ice crystal's interface growth velocities and SSA the ice's specific surface area (see

Eq. (21) in Calonne et al. (2014)). For a simple characteristic scale analysis, a SSA in the range of 0.6×10^4 to 10^4 m^{-1} (Schleef et al., 2014) is combined with an interface growth velocity on the order of $1\times10^{-9}ms^{-1}$ (Calonne et al., 2014; Krol and Löwe, 2016). Combining these literature values yields deposition rates on the order of 0.5×10^4 $kgm^{-3}d^{-1}$), which is significantly larger than the simulation results. Interface growth velocities on the order of 1×10^{-13} or 1×10^{-14} ms^{-1} would match with the simulated magnitudes for deposition rate.

Lastly, the impact of included higher-order mesh errors E_T and E_c (see Sect. 4.2.2) on the temperature distribution is evaluated. The error is determined by computing the temperature deviation between the solution that considers higher order mesh errors, and the solution that does not. The deviation is then quantified in a L1 norm. The error increases with simulation time and is 0.13 K after 24 h, 0.23 K after 36 h, and 0.28 K after 48 h. After 48 h the deviation is highest for the computational nodes just above the layer transition, where high temperature gradients are present (see Fig. 4.8). The reader should note that the error for deposition rate could be derived similarly. From the temperature error, the deposition rate error can be derived as deposition rate is directly derived from temperature via the water vapor transport equation. Therefore, one error measure is considered sufficient to emphasize the impact of mesh errors.

Settling-induced evolution of the ice volume fraction in the absence and presence of water vapor and heat transport (Cases 1 and 5)

Figure 4.10 compares isolated settling (Case 1 in Table 4.2) and the fully coupled processes (Case 5 in Table 4.2) with respect to their impact on the evolving ice volume fraction. Figure 4.10 shows the corresponding ice volume fraction profiles after 2 days. In Fig. 4.10 (a), both profiles are very similar, which suggests that the density evolution is dominated by settling processes and coupled heat and water vapor transport play a minor role. When focusing on the upper boundary of the transition area (Fig. 4.10 (b)), there is a locally decreased ice volume fraction for the fully coupled processes (Case 5). This suggests a local ice volume decay for active water vapor transport from phase changes. This observation is consistent with the enhanced sublimation rate observed in Fig. 4.9 and indicates the formation of a density heterogeneity.

Comparison against layer-based schemes (based on Case 6)

Lastly, I compare results of the proposed Eulerian–Lagrangian scheme with conventional layer-based models, e.g., SNOWPACK, Crocus. I would like to emphasize that a two-layer snowpack model certainly constitutes an extremely simplified case, as layer-based schemes are usually operated with a significantly higher number of snow layers. It is yet informative to conduct this analysis to point out differences, as these can certainly accumulate during long simulation times.
In layer-based snowpack models, state variables are assigned layerwise, and the two-layer snowpack (Fig. 4.5) would have three computational nodes at the following locations: at the bottom of the lower layer, at the top of the lower layer, and at the top of the upper layer. The two nodes located at the top of the layers would then represent the physical state of the lower and the upper layer,

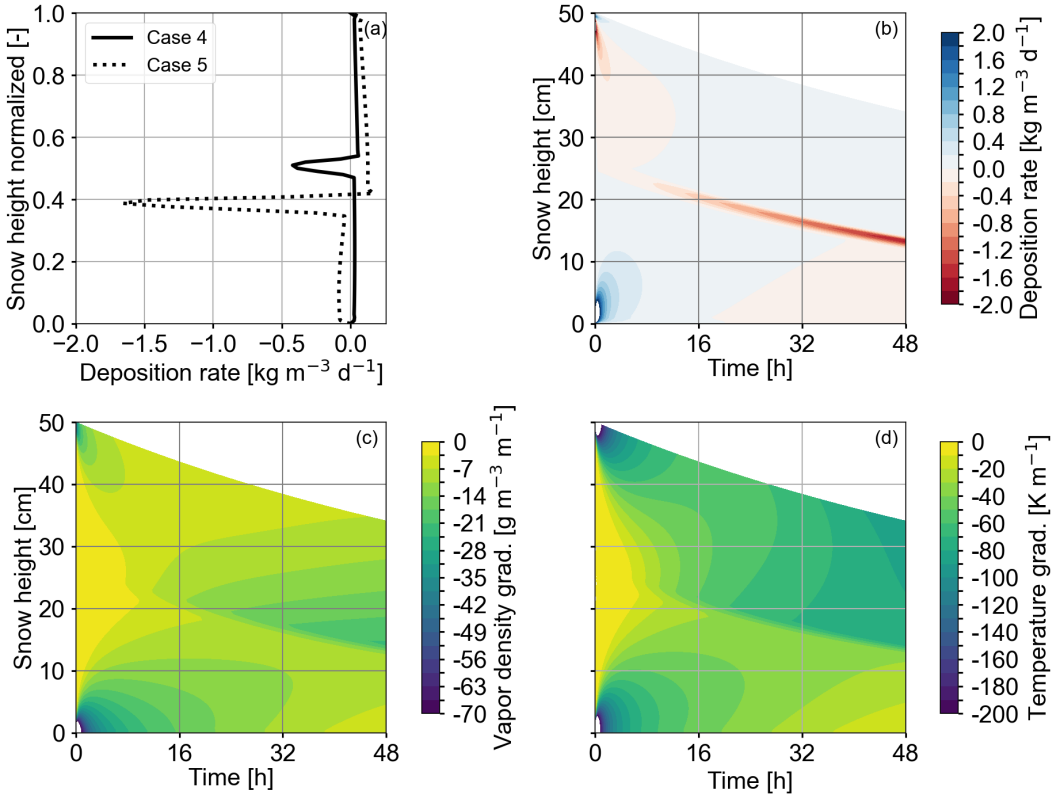


FIGURE 4.9: (a) shows two deposition rate profiles over normalized snow height after 2 days. The solid line represents the results of heat and water vapor transport in the absence of settling for Case 4 of Table 4.2. The dashed line refers to Case 5 of Table 4.2 that additionally accounts for settling. Sublimation rates (negative deposition rates) for Case 5 (fully coupled processes) are increased by approximately a factor of 4 with respect to Case 4 without settling. At the top of the sublimation peak for both cases, a slight peak in deposition rate is visible. (b) shows the deposition rate profile evolution for Case 5. A pronounced sublimation rate peak in the transition area is first visible after approximately 6 h and increases with time. I interpret this area of increased sublimation (red line in the center) as the boundary of the upper and lower layer. (c) shows the evolution of the water vapor density gradient. The gradient at the bottom of the upper layer (at approximately 20 cm and 15 cm height after 16 h and 48 h) increases with time. (d) shows the evolution of the temperature gradient with time. Overall the temperature gradient is higher in the upper compared to the lower layer. The lobes at the top and bottom at the start of the simulation in (b), (c) and (d) are due to the strong phase change activity and heat flux triggered by the initial and boundary conditions.

respectively. Velocity is again derived from stress exerted by the overburden snow mass. Since the upper layer is represented by the computational node at the top, it is unloaded and requires a special treatment for stress. I adopt the approach by Vionnet et al. (2012) and apply a 'non-physical stress' equivalent to half of the layer's own weight, yet apply it to the uppermost computational node (Sect. 3.4 in Vionnet et al. (2012)). Next, vertical velocity is computed likewise with Eq. (4.6) and viscosity with Eq. (B.5). I compare both schemes, namely layer-based and continuous, based on Case 6 of Table 4.2, hence in the presence of mechanical settling only and for a dynamic viscosity closure. Since heat and water vapor transport are neglected, the viscosity changes over time are solely due to the evolution of ice volume fraction alone.

In Fig. 4.11, the layer-based scheme sustains a layer-wise vertical velocity (a) and ice volume fraction

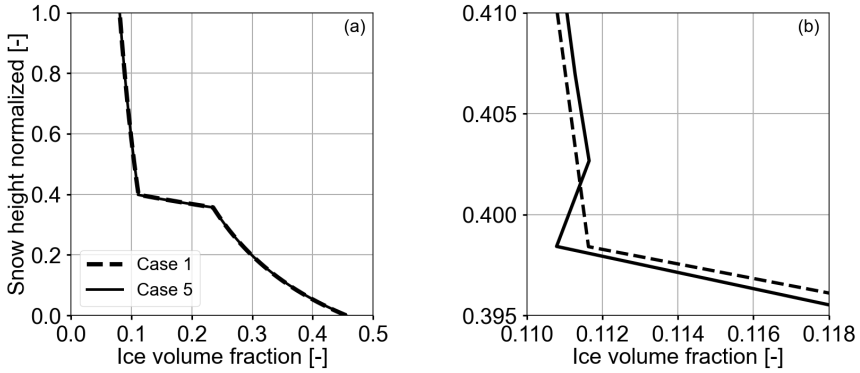


FIGURE 4.10: The plots show ice volume fraction profiles over normalized snow height after 2 days. (a) Case 1 (Table 4.2) depicts ice volume fraction corresponding to solely active settling, and Case 5 refers to the fully coupled processes. (b) zooms into the density transition area of (a). The kink in the profile of Case 5 shows the effect of the increased sublimation in Fig. 4.9 that yields a local decrease in ice volume fraction. In order to better resolve the kink of Case 5, I increased the number of grid nodes to 251.

evolution (c): One value for the velocity and one value for the ice volume fraction describe an entire layer. In contrast, using the generalized Lagrangian approach described in Sect. 4.2.2, yields a sublayer resolution of the vertical velocities (b) and ice volume fractions (d). For both schemes the vertical velocity is higher in the top part of the snowpack and zero at the bottom. For early times, the layer-based scheme determines a vertical velocity that is one order of magnitude higher than values computed with the Eulerian–Lagrangian scheme. This may be related to the comparably high, non-physical stress at the top of the upper layer. At the end of the simulation, the snowpack has settled almost twice as much with the layer-based scheme, which highlights the impact of this conceptual difference. This effect may result from an overestimation of velocity with layer-based schemes. Following the proposed continuous method, ice volume fraction is higher in the lower part of the snowpack and reaches higher values (Fig. 4.11 (d)). Furthermore, ice volume fraction at the top of the snowpack does not change during the simulation since there is no stress from overburden mass. In contrast, for the layer-based scheme ice volume fraction grows at this location (Fig. 4.11 (c)). This is again due to the stress condition at the top. Of course this discrepancy reduces as for an increase of the number of layers. However, this slight offset in the stress condition will always be present and lead to uncertainties. In the proposed computational approach based on a continuous layer scheme, the spatial resolution of processes can be easily changed to assess its impact on snowpack evolution. In a future study, it might be interesting to quantitatively compare results against Jafari et al. (2020), who also rely on a rather fine spatial resolution.

4.3 Publication of the modeling software Eulerian–Lagrangian snow solver

Process models are usually described in articles, but they are only complete with their associated software assets. In the case of the presented snow model, it is the *Eulerian–Lagrangian snow solver* written in Python that puts the computational approach into practice. The Eulerian–Lagrangian

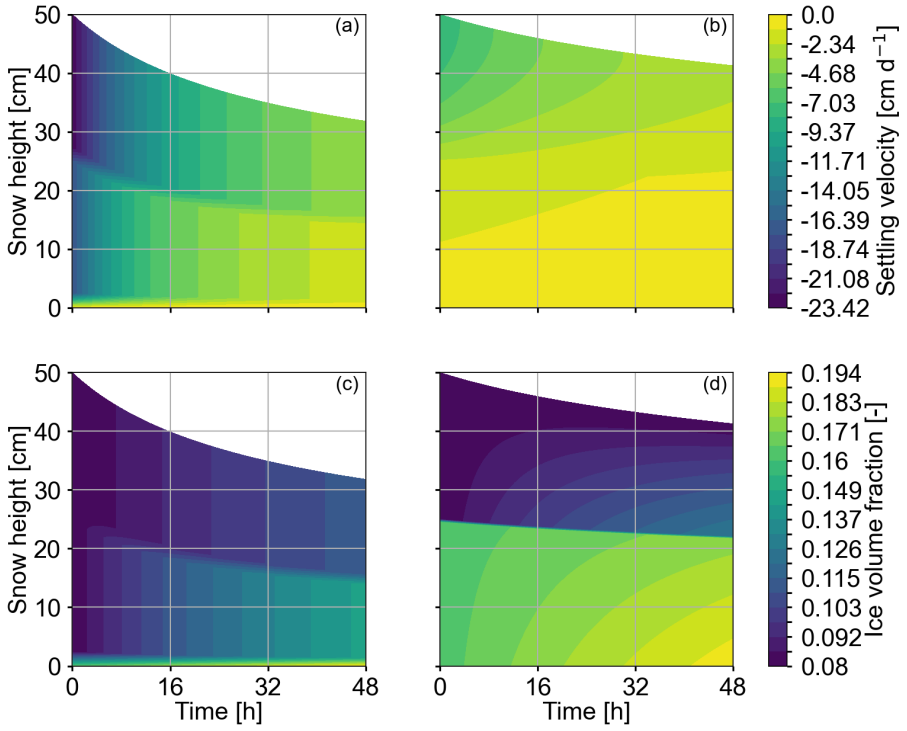


FIGURE 4.11: The plots show the temporal evolution of vertical velocity (top row) and ice volume fraction (bottom row) for Case 6 of Table 4.2 corresponding to solely active mechanical settling. Y-axis depicts snow height. For (b) and (d), I applied the highly discretized settling scheme, and for (a) and (c), I mimicked the layer-wise discretization of layer-based schemes. Snow viscosity is controlled by ice volume fraction alone, since heat transport is inactive. In (a) and (c) the lower and upper layer are resolved as the darker upper an brighter lower parts of the snowpack. Their respective values refer to the computational nodes at the top and between the two layers. The values retrieved for the lowest node, do not represent an entire layer and are depicted at height zero. For the layer-based scheme, one velocity or ice volume fraction value represents the movement or density of the entire layer. In contrast, with this approach vertical velocity varies throughout each layer in (b) so that ice volume fraction increases within layers and develops a gradual pattern (d). For (b) and (d), I interpret the locations of the upper and lower layers based as the darker upper and brighter lower areas respectively in (d).

snow solver has been made available alongside the article (Simson et al., 2021) as a public GitHub repository (Simson and Kowalski, 2021a). The version of the software at the time of the article's publication is tagged as a release with the name *final paper submission TC*. Furthermore, a zip-file of the release has been published on Zenodo (Simson and Kowalski, 2021b) to register a DOI. The reader should note that there have been no major updates to the software since this release. While it may be updated in the future, this specific release is the subject of this section.

The folder structure of the GitHub repository is provided in Fig. 4.12. The *Model* folder contains individual Python files containing functions such as the solvers for the state variables temperature and deposition rate, the application of boundary and initial conditions, selection of the model geometry as the update of the ice volume fraction, and the physical model parameters. The descriptive naming of the files aims at quick finding of specific functions. Generally, the software design follows the modular approach as described Fig. 4.4. Temperature and deposition rate are solved with `T_solver.py`

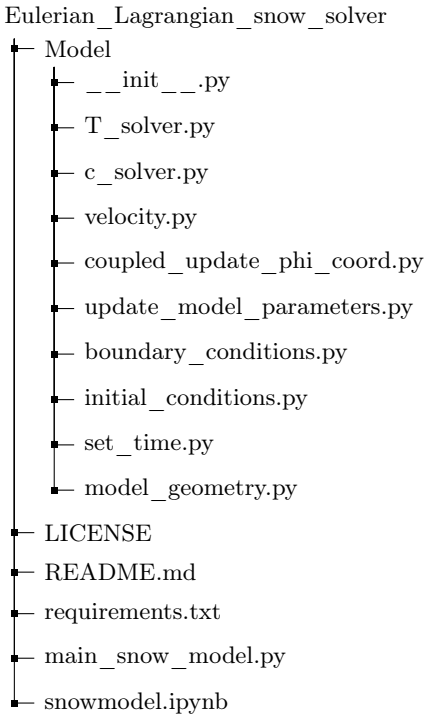


FIGURE 4.12: Folder structure of the Eulerian–Lagrangian snow solver repository on GitHub (Simson and Kowalski, 2021a).

and `c_solver.py`. Next, velocity is solved with `velocity.py` followed by the update of ice volume fraction and vertical coordinates of the deforming mesh with `coupled_update_phi_coord.py`. There are two possibilities to execute the model. First, with the script in `main_snow_model.py`, which calls the individual functions with a default setting. Second, with the marked up jupyter notebook `snowmodel.ipynb`.

With the goal to facilitate consistency between article and code, similar names for parameters, vectors, and matrices were chosen. This is true, for instance, for the matrices and vectors of the deposition rate and temperature equations (Eqs. 4.24 and 4.23). The matrices $\mathbf{E_T}$ and $\mathbf{E_c}$ from the article that account for higher-order mesh errors due to the non-equidistance mesh are equivalent to `ET` and `Ec` in the code.

Different computational options such as viscosity parametrizations, predefined sets of initial and boundary conditions, and the maximum number of iterations can be selected through the tags in the input of the `main` function in `main_snow_model.py` or by following the step-by-step explanation in the Jupyter notebook. The software includes more options than those mentioned in the article, for instance, equations used for the update of the model parameters of snow as listed in Table 4.1 can also be selected from Hansen and Foslien (2015a) additionally to equations by Calonne et al. (2014) used for the simulations in the article. The default option is in accordance with the article.

Given the perspective that the code has been written in 2021, many things could be changed and improved from today's perspective. The highlighted deactivation of individual modules of

the software is realized through out-commenting of the respective functions, which may be error-prone. However, I want to highlight that this sort of software is a real-world example for code developed by individual researchers. According to Barnes (2010), the publication of any software code, regardless of quality, benefits other researchers in understanding and working with scientific results and improves transparency. The Eulerian–Lagrangian snow solver was published with the goal to allow others to understand and reproduce the results presented and in that way to contribute to the anticipated development of new snow models that provide modularity and extendability.

The software is assigned a MIT license to legally allow others researchers to reuse it and build new research on it. The `README.md` file provides a detailed documentation of the code to facilitate reuse, it provides a badge with the registered DOI to link the repository with the Zenodo entry (Simson and Kowalski, 2021a) and to make it citable, and it contains a `requirements.txt` file with the necessary libraries to run the code. In the course of writing this thesis, I also added a badge to the readme-file that indicates the FAIRness of the software using the automated FAIRness assessment tool howfairis introduced in Sect. 2.2.2. The commit history contains ∼150 commits, but the full development of the software from its start is not publicly available.

4.4 Reuse potential of the Eulerian–Lagrangian snow solver

My goal in releasing the modeling software Eulerian–Lagrangian snow solver is to make the best possible contribution to the development of snowpack models. This goal implies that the model and its software should be used by others. In this section, I would like to advance this general idea of reusing the Eulerian–Lagrangian snow solver. First, I present potential, not yet implemented, reuse scenarios. The first is formulated based on an extension proposed in(Simson et al., 2021), and the second and third are independent of the article. Next, I highlight a published reuse example of the software from Brondex et al. (2023b), which demonstrates the real-world reuse of the Eulerian–Lagrangian snow solver for developing new snow models.

4.4.1 Reuse scenarios exploiting the model's modularity and extendability

The following three reuse scenarios reflect potential reuse purposes and contexts of the Eulerian–Lagrangian snow solver. The scenarios emphasize the potential of the model's modularity and extendability, and the variety of people who would benefit from its high reusability.

Reuse Scenario A1

Master student Sam wants to write his master's thesis in Geophysics on snowpack modeling. His supervisor recently read the article by Simson et al. (2021) and proposes him to extend the snow model to investigate the interaction of water vapor transport with further processes. His supervisor suggests him to follow the ideas proposed in the *Future work and challenges* section of the article Simson et al. (2021), which outlines the potential integration of liquid water in the model equations.

While including potential phase changes from melting and freezing could be straight-forwardly implemented via the source term c, it is the advective transport of liquid water that is more demanding. Liquid water transport is commonly modeled via the Richards equation (Wever et al., 2014), which would benefit from existing hybrid Eulerian–Lagrangian solution strategies such as shown for saturated media without mechanical settling (Huang et al., 1994). Furthermore, a model for wet snow requires a second energy balance to account for the temperature of the liquid water.

Sam decides to change the physical model of dry snow (Sect. 4.2.1) to allow for wet snow and extend the existing Eulerian–Lagrangian snow solver accordingly. First, Sam explores the corresponding GitHub repository by Simson and Kowalski (2021b) and gets an overview of the different functions and important interfaces for the planned extension. Once set up, he integrates the changes into the computational workflow (Fig. 4.4) and couples the new software module with the Eulerian–Lagrangian snow solver.

Reuse Scenario A2

Passionate skier Younes has read that water vapor transport can reduce the mechanical stability of the snowpack, which increases the thread of avalanches. He is very curious and wants to improve his physical understanding of snow. Through a google search, he discovers the Eulerian–Lagrangian snow solver. He remembers some Python basics from his studies, reads the documentation and then runs the `main_snow_model.py` script for different boundary conditions and model geometries. Younes tests different combinations of activated processes by leveraging the flexible modularity of the model.

Reuse Scenario A3

PhD student Zoe acquired a time series of snow density and temperature data from thin Arctic snowpacks known for significant water vapor transport (Domine et al., 2016; Domine et al., 2019). As a next step, she plans to use the modeling software Eulerian–Lagrangian snow solver to track the changes observed in the snow profiles. Based on the simulations, she aims to quantify the amplitudes of sublimation and deposition and to localize changes in snow density related to mechanical settling and diffusive water vapor transport. Therefore, she changes the model geometry to match the observed snowpacks and adjusts the boundary and initial conditions to match the observed air and soil temperatures. She then runs simulations, activating and deactivating water vapor diffusion and mechanical settling, and she compares the results with the measured data.

4.4.2 Conducted reuse example in form of a model comparison

The model description by Simson et al. (2021) and the accompanying modeling software Eulerian–Lagrangian snow solver were published in 2021. Afterwards, Brondex et al. (2023b) continued to advance the development of a generic and modular dry snow model that couples heat conduction, water vapor diffusion and mechanical settlement. They developed the Python-based finite-element

framework *IvoriFEM* (Brondex et al., 2023c). Brondex et al. (2023b) build on the Eulerian–Lagrangian snow solver and use the corresponding article Simson et al. (2021) as technical documentation. During the reuse process, several issues arose that the authors documented in the article. In the following, I first summarize the software asset IvoriFEM and the reuse issues faced by (Brondex et al., 2023b) when reusing the Eulerian–Lagrangian snow solver. Next, I present the concrete reuse example of the Eulerian–Lagrangian snow solver by (Brondex et al., 2023b).

IvoriFEM

IvoriFEM combines different variations of the process model equations proposed for water vapor diffusion and heat conduction by Hansen and Foslien (2015a) and Calonne et al. (2014). Furthermore, it couples the equations with the ice mass balance as proposed by Schürholt et al. (2022) and Simson et al. (2021), and it adopts the settling approach as proposed by Simson et al. (2021) (see Sect. 4.2.2). The focus of IvoriFEM is the investigation of the numerical challenges arising from the coupling of the Partial Differential Equations of the process model. Similar as the Eulerian–Lagrangian snow solver it provides flexibility through modularity of the code. The code of IvoriFEM is linked in the corresponding article (Brondex et al., 2023b) to a Zenodo entry (Brondex et al., 2023a) that contains a zip-file of the software and provides a DOI for identification. The software is furthermore available via GitHub (Brondex et al., 2023c) with an extensive documentation of the code. However, the GitHub repository itself does not provide the DOI or a link to the corresponding Zenodo entry. The Zenodo entry of the software assigns it a CC-BY license. Yet, the GitHub repository does not provide a license. Furthermore, the commit history of the repository and thus the software development process is not public.

Reuse issues

Brondex et al. (2023b) mention several inconsistencies between the article by Simson et al. (2021) and the code (Simson and Kowalski, 2021b) that they found while investigating the mechanical settlement scheme of the Eulerian–Lagrangian snow solver.

In their Section 3.3.2, Brondex et al. (2023b) mention that the definition of the spatial discretization step in the article is not consistent with the formulation in the code. Namely the space step in the integrated velocity equation (4.16) is defined as $\Delta z_j^n = \Delta z_{j+1}^n - \Delta z_j^n$ although it is implemented as $\Delta z_j^n = \Delta z_j^n - \Delta z_{j-1}^n$ in the code. In Brondex et al. (2023b) Appendix D1, they mention another inconsistency, which concerns the computation of velocity from strain rate. While in the article by Simson et al. (2021) the velocity of a node is calculated from the strain exerted by the overburden mass, the same velocity in the code is found to be assigned to the next lower node. Furthermore, the deposition rate at the boundaries is forced to zero in Simson and Kowalski (2021b), which is not described in the respective section in the article by Simson et al. (2021). This discrepancy is not explicitly mentioned by Brondex et al. (2023b) but becomes clear after reading their article.

Furthermore, Brondex et al. (2023b) find that other than stated in the article (Simson et al., 2021), the model implemented in the code is not strictly mass conserving. Accordingly, ice mass is not fully

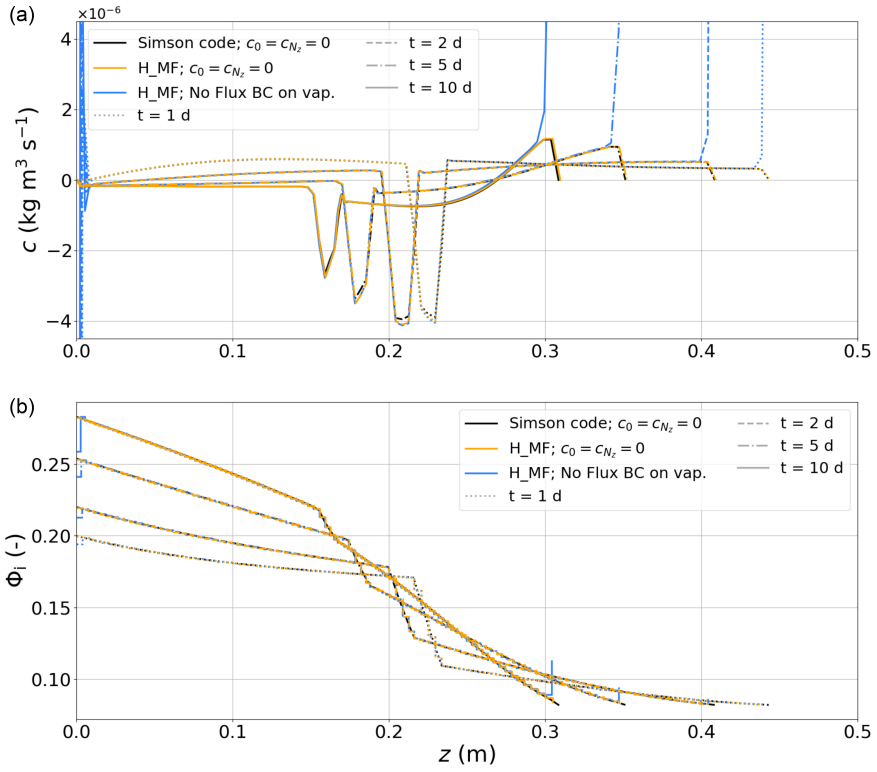


FIGURE 4.13: Reuse example of the Eulerian–Lagrangian snow solver in a model comparison with IvoriFEM by Brondex et al. (2023b) based on conditions of Case 7 in Table 4.2. The figure is taken from Fig. 8 in Brondex et al. (2023b) as is (CC BY 4.0). H_MF refers to the model by Hansen and Foslien (2015a). Depth axis is switched in comparison to figures by Simson et al. (2021).

conserved with the Eulerian–Lagrangian snow solver. Instead a total loss of 20 g of snow for a snow pack of 56,25 kg mass is observed within 20 days. According to (Brondex et al., 2023b), this issue arises from the explicit temporal discretization use for the update of for ice volume fraction (see eq. 4.12). Ice mass conservation is fulfilled by changing the explicit scheme to an implicit scheme (see Eq. (19) in Brondex et al. (2023b)). The reader should note that described issues, except from the deposition rate boundaries, have been corrected in a new version of the code (v1.1.0) in 2025 (Simson and Kowalski, 2021a).

When other researchers try to build their own research on top of existing software, they quickly get into details that are not covered in the technical documentation, which makes reuse difficult. Other challenges arise from inconsistencies between the described model and its technical implementation in software. Often these inconsistencies remain undetected also because the software provided as supporting documentation is typically not reviewed. The inconsistencies found by Brondex et al. (2023b) are minor issues that can be resolved by changing the indexes or adding the appropriate information in the article or other documentation, and they do not significantly affect the results. However, they cause confusion when trying to understand another researcher's work or reproduce published results. These types of challenges affect many researchers, but they are not often the subject of discussion. In the best case, the article and software documentation describe the model sufficiently and consistently so that the software is understandable to other researchers. However,

this is often not the case, even when researchers take great care to make their documentation understandable and consistent, as exemplified by the Eulerian-Lagrangian snow solver.

Model comparison reuse examples

As Brondex et al. (2023b) built their settlement approach based on that proposed in Simson et al. (2021), they conducted two model comparisons. The reader should note that Brondex et al. (2023b) modified the Eulerian–Lagrangian snow solver to overcome the inconsistencies mentioned above before carrying out two comparative studies.

The first comparative study (Fig. 7 (b) in Brondex et al. (2023b)) is based on Case 6 in Table 4.2. Case 6 only considers mechanical settling, which is based on a dynamic ice volume fraction dependent viscosity (Eq. (B.5)). Several profiles of settling velocity obtained with both models are plotted. Eulerian–Lagrangian snow solver shows slightly higher settling velocities. This inconsistency is related to different ways of computing the deformation rate as explained by Brondex et al. (2023b).

The second comparative study (Fig. 8 in Brondex et al. (2023b)) is based on Case 7 in Table 4.2. Case 7 considers the fully coupled system of heat conduction, water vapor diffusion and mechanical settling, and it applies a dynamic viscosity that depends on ice volume fraction and temperature. The comparison of the models is based on deposition rate and ice volume fraction profiles for several times between 1 day and 20 days. The profiles of both models are consistent when the deposition rate in IvoriFEM is forced to zero at the boundaries and the model equations of Hansen and Foslien (2015a) are applied, both of which are also implemented in the Eulerian-Lagrangian Snow Solver.

Chapter 5

Cryospheric Case Study II: Compilation of sea ice core data sets

Acquisition of data in fieldwork and experiments is laborious and expensive. Once the data has been utilized to explore the hypotheses that motivated its collection, there is an increasing expectation for scientists to publish their datasets in data repositories and to adhere to appropriate metadata standards. However, as described in Chapter 3, selecting a repository and metadata standards for a specific type of data set can be challenging. This difficulty is due to unavailable or inadequate discipline-specific repositories and standards for the respective data.

Consequently, data sets originating from different scientists yet representing the same type of measurement are often distributed across data repositories and lack or have heterogeneous standardization. This issue is especially pronounced for smaller data sets (see Sect. 3.1). To unlock the full potential, it is required to combine these individual data sets into larger compilations. Such compilations facilitate the reuse of the combined data in reuse scenarios, such as comparative studies, validation studies, or the training of data-based models.

The compilation task itself is often not trivial as individual data sets have different contents, formats, and standards. In Cryospheric Case Study II, I present RESICE, a database compiled from small, distributed, and heterogeneously standardized sea ice core data sets. A main focus is on the transparent documentation of the compilation approach including the manual reuse challenges faced.

This case study is based on the article *Reusability-targeted enrichment of sea ice core data* published in Scientific Data (Simson et al., 2025c). The article is accompanied by the database *RESICE* (Simson and Kowalski, 2025a; Simson and Kowalski, 2025b; Simson and Kowalski, 2025c) and the software asset in from of the pyresice Python package on GitLab (Simson et al., 2025a). All accompanying resources are available on Zenodo.

5.1 Introduction to the case study

Sea ice forms from freezing sea water, and it floats on the ocean, where it acts as an insulator from the atmosphere. Sea ice is a multi-phase and multi-component porous medium (Worster and Rees Jones, 2015; Feltham et al., 2006). Pure ice is the solid phase of sea ice and forms its matrix, and the pores are filled with brine and air. Large parts of the Arctic and Antarctic oceans are covered with sea ice, which typically persists for one to five years (Stroeve and Notz, 2018). In recent decades, the extent of sea ice cover and its thickness have decreased in response to climate change (Sumata et al., 2023; Stroeve and Notz, 2018; Purich and Doddridge, 2023). The reduction in sea ice cover is leading to a shrinking habitat for polar bears (Pagano et al., 2021), poses logistical (besides many other) challenges for indigenous communities as sea ice trails disappear (Huntington et al., 2017), and it impacts ocean circulation and the oceanic food web (Rossi et al., 2019).

Measurements of sea ice are essential to provide an overview of the status quo of sea ice and to observe changes in its properties. Sea ice extent, concentration, thickness, and age can be inferred from satellite data (Sandven et al., 2023). Satellite observational data serves as initial conditions or is used for assimilation of climate models. Furthermore, model validation efforts use satellite data. Notz and Stroeve (2016) used satellite observations to predict the total area of Arctic sea ice cover in order to validate predictions of climate models. Remote sensing satellite observations are commonly used in cryospheric sciences, and they are complemented by in-situ measurements of sea ice, for instance, on the basis of sea ice cores.

Scientists from all over the world acquire data from sea ice cores drilled in Earth's polar regions (Wang et al., 2020a; Johnson et al., 2023; Duprat et al., 2019; Katlein et al., 2020b). An increasing number of the data sets acquired are becoming publicly available in repositories allowing other researchers to use it to address new research questions in form of reuse scenarios. In the realm of sea ice core data, exemplary reuse scenarios are, for instance, the validation of physics-based process models for sea ice evolution or the data-driven classification of qualitative sea ice characteristics. Such scenarios require a database, which contains data from many, different sea ice cores. Yet, no comprehensive sea ice core database that combines data sets collected from different scientists and measurement campaigns exists. Instead, reusers, who want to conduct such reuse scenario, have to find, integrate, and combine individual sea ice core data sets themselves to unlock their full potential.

The FAIR principles (Wilkinson et al., 2016) should facilitate these agnostic reuses as they have been formulated to ensure findability, accessibility, interoperability, and ultimately reusability of data. If data is published in FAIR-compliant repositories, agnostic reusers (see Sect. 2.1.1) could assume that published sea ice core data are easily integrable into their scenarios. In reality, the combined cross-resource and -repository reuse of multiple sea ice core data sets is not possible without significant engagement of the reuser and time investment.

Besides the issues deriving from the cross-repository distribution, sea ice core data sets do not align with reuse needs, as they are structured to best reflect the data acquisition and not to best

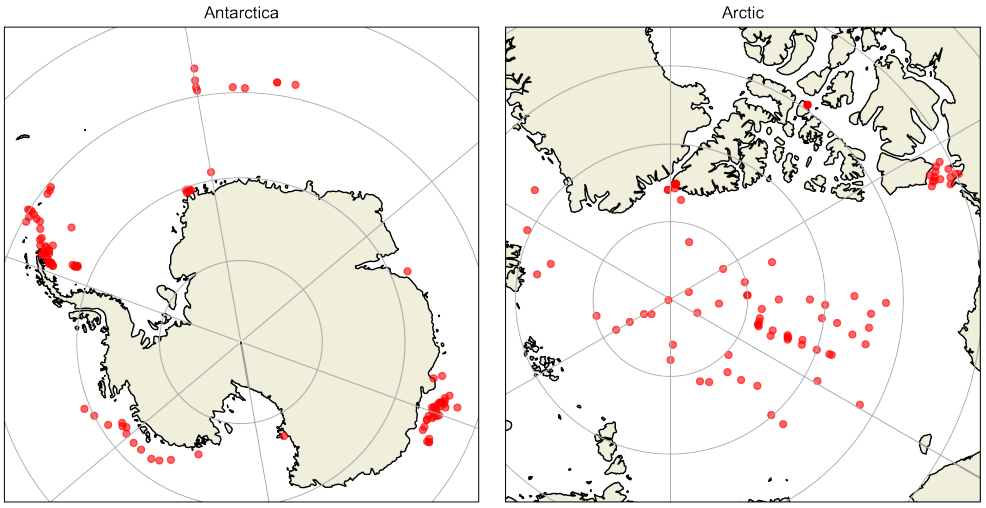


FIGURE 5.1: Locations of all sea ice cores incorporated in RESICE. Color intensity indicates density of available sea ice cores. The map is generated with Cartopy (Met Office, 2010).

serve reuse purposes. The combination of data sets into a homogeneous form is impeded by the heterogeneity within and between data sets. The data sets vary in quality, descriptiveness, content, label names, units, and formats. Relevant data and metadata are not available from data sets but from text-based, context providing resources, such as articles or expedition reports. Furthermore, the data and metadata of the same sea ice core can be distributed across several data sets, and it can be redundant on different repositories.

In Cryospheric Case Study II, I introduce the Reusability-targeted Enriched Sea Ice Core Database (RESICE), which tackles these challenges and combines data and metadata of 287 sea ice cores. RESICE enrichment increases reusability of all combined data sets and is achieved in two ways. First, RESICE combines data and metadata from 138 sources, including 107 data sets from Zenodo, AADC, and PANGAEA; 23 articles and reports; and 8 instrument manuals. Second, RESICE contains additional automatically generated metadata tailored to the requirements of the specific reuse scenarios and following standard naming. RESICE allows traceability of all data points to their original resources as it is demonstrated in the data2source traceabilty plot in Fig. 5.2 (b). The database is accessible via Zenodo and the MOSAiC webODV, and it is extendable by other researchers through the pyresice Python package (Simson et al., 2025a).

In this case study, I will first describe the reusability-targeted approach using two reuse scenarios as motivating example. The approach involves the curation and enrichment of a combined cross-resource and cross-repository database. Second, I describe the RESICE database and its reuse pathways, which facilitate reuse scenarios in the future.

5.2 Reusability-targeted approach to compile data sets

The reusability-targeted approach followed to compile RESICE can be summarized as follows:

Step 1 Formalization of the reuse perspective with

- **(a)** reuse scenarios
- **(b)** reuse scope, i.e., the required data and metadata, combined referred to as elements

Step 2 Assembly of reuse scope relevant data and metadata by

- **(a)** searching for resources
- **(b)** matching elements of the reuse scope that are available from resources

Step 3 Plausibility checks of the resources

Step 4 Technical combination of the resources

Step 5 Automatic metadata enrichment of unavailable reuse scope elements

Intuitively, one assumes all information contained in the reuse scope is available from resources. Accordingly, agnostic data reusers would dedicate most time for Step 2 (a) assuming Step 3 and Step 4 would be easy. In reality, data are distributed across resources and are difficult to combine. Therefore, significant time is necessary for Step 2 (b), namely matching reuse scope elements with resource availabilities, and for Step 3 and Step 4. After combining all data and metadata available from existing sources in Step 4, missing reuse scope elements are automatically enriched on the basis of Python routines developed by the data reuser in Step 5. The following section describes the combination of the database RESICE and is aligned with the five steps of the reusability-targeted approach. It is followed by a description of the published data record. Lastly, the RESICE reuse potential is highlighted with its three reuse pathways of different complexity.

5.2.1 Step 1: Formalization of the reuse perspective with reuse scenarios and scopes

Step 1 (a): Reuse scenarios

Reuse scenarios vary in purposes and require individual inputs in form of data and metadata. In this case study, I define Reuse Scenario B1, which is the validation of a sea ice evolution model, and Reuse Scenario B2, which is the training of a classification algorithm to detect qualitative features of sea ice. The validation of a heat flux simulation, for instance, requires temperature measurement data as input. Depending on the scenario, these inputs can be further specified by constraints, e.g., minimum mean distance between measurements. Further examples for constraints are specific units, formats, value limits, maximum allowed measurement errors, compliance with

TABLE 5.1: List of inputs and constraints for Reuse Scenarios B1 and B2. This table constitutes the formalization of the reuse perspective for both reuse scenarios. ppt is parts per thousand. SIN is sea ice nomenclature. Vertical profile means measurements are assigned a depth.

	Reuse Scenario B1 (a)		**Reuse Scenario B2 (b)**	
Elements	**Input**	**Constraints**	**Input**	**Constraints**
Date			✓	Format: YYYY-MM-DD
Coordinates			✓	Format: decimal degrees
Water body			✓	with SeaVoX class names
Salinity sea ice (vertical profile)	✓	Ratio (salinity): ppt Unit (depth): m Mean distance: below 0.2 m Measurement error: below 5%		
Solid fraction sea ice (vertical profile)	✓	Unit (solid fraction): - Unit (profile location): m Mean distance: below 0.2 m Measurement error: below 5 %		
Temperature sea ice (vertical profile)	✓	Unit (temperature): K Unit (depth): m Mean distance: below 0.2 m Measurement error: below 5 %		
Temperature air	✓	Unit: K		
Salinity sea water	✓	Ratio: ppt		
Temperature sea water	✓	Unit: K		
Thickness snow	✓	Unit: m		
Thickness sea ice	✓	Unit: m Value: below 1 m	✓	Unit: m
Measurement errors	✓	Ratio: %		
Standard deviations of repeated measurements	✓	Ratio: %		
Mean salinity sea ice			✓	Ratio: ppt
Mean temperature sea ice			✓	Unit: K
Form sea ice			✓	with SIN class names
Development stage sea ice			✓	with SIN class names

specific naming standards and others. Inputs have to comply with the constraints to be suitable for the scenario. This is best exemplified based on the two defined Reuse Scenarios B1 and B2. The inputs and constraints of both scenarios are listed in Table 5.1, which constitutes the formalization of the reuse perspective.

Reuse Scenario B1 PhD student Ada implemented a 1D physics-based process model of sea ice that simulates heat and salt transport and phase change processes in vertical direction of sea ice such as proposed by Buffo et al. (2018). In a next step, she wants to validate the process model by comparing the simulated vertical profiles for this task of sea ice consisting of the state variables temperature, salinity, and solid fraction with measurement data. Sea ice core data is suitable as it often provides several measurements along the core (profile). In order to be useful for the task, measurement data should have the same unit or ratio as the state variables in the model. Ada would like to include uncertainties in form of measurement errors and standard deviations. She wants to

ensure fidelity by only including data with a measurement error below 5%. The extent of the model domain is 1 m, and the spatial discretization is 0.01 m. The mean distance of the measurement data does typically not coincide with the spatial discretization of the domain. Measurement data may has to be interpolated to match with the location of the computational grid. For a sufficiently good interpolation, Ada restricts the task to measurement data with an averaged mean distance of less than 0.2 m. She needs to know the thickness of the sea ice to ensure sea ice is thinner than domain extent. Ada would like to use suitable boundary conditions for the model, namely temperature and salinity of the underlying sea water and temperature of the overlaying air. The upper boundary may be affected from insulating snow covers. Thus, she also needs the thickness of potentially overlaying snow to approximate these effects. The inputs and constraints of Reuse Scenario B1 are listed in Table 5.1.

Reuse Scenario B2 Liza is a master's student. As part of her thesis project she wants to train a classification algorithm based on sea ice core data. She defines the target variables *form* and *development stage* based on the Sea Ice Nomenclature (SIN) provided by the World Meteorological Organization (WMO) (WMO, 2014). She refers to *form* as the definitions provided in SIN Section 1.1. Sea ice is *fast ice* (1.1.1) if it is attached to the coast. Non-attached occurrence of ice are either *drift ice* or *pack ice* (both 1.1.2). *Drift ice* and *pack ice* are distinguished based on the sea ice concentration. Sea ice concentration above 70% indicates pack ice and below drift ice. She refers to *development stage* as the definitions provided in SIN Section 2 (Development). Section 2 has up to three sub-categories of which Liza considers the first two. All sea ice development stages of the first two sub-categories are listed in Table 5.2. On the sub-category level, development stage has classes *New ice* (2.1), *Nilas* (2.2), *Pancake ice* (2.3), *Young ice* (2.4), *First-year ice* (2.5), and *Old ice* (2.6). On the sub-subcategory level, classes, for instance, for *First-year ice* are *Thin first-year ice* (2.5.1), *Medium first-year ice* (2.5.2), and *Thick first-year ice* (2.5.3). Sea ice development stage is a property typically assigned by the person(s) who drilled the sea ice sore. In the SIN, each of the classes except from *New ice* are assigned with characteristic thicknesses such as '30 cm - 2 m' for *First year-ice* and 'up to 2.5 m and sometimes more' for *Second-year ice*. The thicknesses assigned to each level 1 and level 2 sea ice development stages are also provided in Table 5.2. Liza requires the target variables to only consist of classes defined in the SIN. The predictor variables are the mean values of the measurements for sea ice temperature and salinity that have been acquired along the central axis of a core. Furthermore, she wants to include the thickness of sea ice at each coring location, the date of core retrieval as well as the coordinates and the name of the water body, i.e., the sea or ocean, of the coring location. The names of the water bodies should be consistent with the terminology provided by the controlled vocabulary of *The SeaVoX Salt and Fresh Water Body Gazetteer* from the BODC (2023). For a quick integration into the training script, she wants the date to be in *YYYY-MM-DD* format and the coordinates to be in decimal degrees. The inputs and constraints of Reuse Scenario B2 are listed in Table 5.1.

2	Development	Sea ice thickness from SIN (WMO, 2014)	Intervals
2.1	New ice	-	-
2.1.1	Frazil ice	-	-
2.1.2	Grease ice	-	-
2.1.3	Slush	-	-
2.1.4	Shuga	-	-
2.2	Nilas	'up to 10 cm'	[0.0, 0.1]
2.2.1	Dark Nilas	'under 5 cm'	[0.0, 0.05]
2.2.2	Light Nilas	'more than 5 cm'	[0.05, 0.1]
2.2.3	Ice Rind	'to about 5 cm'	-
2.3	Pancake ice	'up to about 10 cm'	[0.0, 0.1]
2.4	Young ice	'10-30 cm'	[0.1, 0.3]
2.4.1	Grey ice	'10-15 cm'	[0.1, 0.15]
2.4.2	Grey-white ice	'15-30 cm'	[0.15, 0.3]
2.5	First-year ice	'30 cm -2 m'	[0.3, 2.0]
2.5.1	Thin first-year ice	'50-70 cm'	[0.3, 0.7]
2.5.2	Medium first-year ice	'70-120 cm'	[0.7, 1.2]
2.5.3	Thick first-year ice	'over 120 cm'	[1.2, 2.0]
2.6	Old ice	'up to 3 m or more'	[0.3, 4.0]
2.6.1	Residual ice	'30 to 180 cm'	[0.3, 1.8]
2.6.2	Second-year ice	'up to 2.5 m and sometimes more'	[2.0, 2.5]
2.6.3	Multi-year ice	'up to 3 m or more'	[2.5, 4.0]

TABLE 5.2: All subcategories (referred to as 2.x and level 1) and sub-subcategories (referred to as 2.x.y and level 2) of the Sea Ice Nomenclature (SIN) Section 2 Development, which I refer to as sea ice development stages, together with the characteristic sea ice thicknesses as assigned to the definitions in the SIN (WMO, 2014). The sea ice thickness intervals indicate how the development stages are represented in the automatic enrichment routine, which is explained in Step 5.

Step 1 (b): Reuse scope

Reuse scope is defined by an agnostic data reuser, i.e., a person uninvolved in data collection, based on the inputs and constraints of the formalized reuse scenario(s). The reuse scope formalizes the desired content of the database. For RESICE, it is the data and metadata required per sea ice core to conduct the scenarios. It should be noted that parts of the reuse scope can be used as a constraint and as an input in the same scenario. For instance, measurement error is an input when combined directly with sea ice temperature for a high-fidelity model validation. At the same time, measurement error is a constraint when only temperature data below a certain error threshold is considered for the validation. Therefore, I refer to both constraints and inputs uniformly as reuse scope elements and to reuse scope also as scope. In this study, the scope is combined from Reuse Scenarios B1 and B2 as formalized in Table 5.1. The scope is extended by the elements ID, campaign, and polar region to allow filtering of the database. Reuse scope consists of the following elements:

- unique ID of the sea ice core,
- name of associated campaign,
- date, coordinates, name of the polar region, and water body of the coring location,

- salinity, solid fraction, and temperature of sea ice assigned with a depth indicating the measurement position along the core,
- temperature of the air,
- salinity and temperature of the sea water,
- thickness of snow cover on the sea ice surface,
- thickness of sea ice,
- measurement errors and standard deviations of repeated measurements,
- mean distances and mean values of the measurements,
- form and development stage of the sea ice,
- unit of all measurement data, and
- naming standards, such as controlled vocabularies, used to classify water body and form and development stage of sea ice.

The scope is the search target for the content of RESICE. A database with this particular scope would furthermore enable other researcher's secondary usages. The granularity of reuse scopes typically implies a structure that does not comply with the content of the resources in a one to one relationship. Therefore, the final content and structure of the database may be different from the initially defined reuse scope. Scope elements are assembled from resources in a search process described in the following.

5.2.2 Step 2: Assembly of reuse scope relevant data and metadata

Step 2 (a): Searching for resources

Search for resources is initiated on data repositories. At best, the reuse scope elements per sea ice core would be provided by one data set alone, and data repositories would have filtering options to allow selection of only those elements that meet defined constraints. In reality, data repositories do not provide filtering options for all of the defined constraints, and data sets are not a one by one representation of the reuse scope. Thus, the reuse scope cannot be comprehensively populated from a single data set. The search for missing elements is continued in other data sets and also includes further resources such as articles and reports providing context for specific sea ice core measurements, and instrument application notes providing general information. The different resource types are grouped, and I refer to data sets as primary resources, sea ice core or campaign specific articles and reports as secondary resources and general information as tertiary resources. Table C.2 in the Appendix lists all original sources an overview on the original data sets selected per repository found during the search for sources.

The collection of reuse scope elements per sea ice core is not completed after source search. Instead it is followed by linking relevant parts of the source content with the corresponding scope elements, since not all data and metadata provided in a source are required in the scope. I refer to this combinatorial process as element availability matching. While availability matching is element specific and explained in the next subsection, source search can be generalized and is explained in this subsection.

Primary resources There is a variety of repositories for sharing all types of research results, and I constrain the search for data sets to the three repositories Zenodo, PANGAEA, and Australian Antarctica Data Centre (AADC) that majorly publish under licenses that allow reuse and republication, such as CC0, CC BY, or CC BY-SA Creative Commons, 2023. General properties of these data repositories have been summarized in Sect. 3.1.1 and specifically in Table 3.1 and Table 3.2. Here, I focus on the search routines and selected data sets.

In Zenodo, a search for *sea ice salinity* with filter options *access* set to *open* and *resource type* to *data set* has around 9,500 results. The first data set from Oggier (2019) provides data from lab experiments; it is excluded. The subsequent three results contain elements of the scope and are therefore selected. More data sets from Zenodo are not included as the subsequent search results are either laboratory data, 2D satellite data, or modeled data. Other combinations of elements in the search query did not improve the first results. Only three data sets were considered fit for purpose.

In PANGAEA, advanced search for *sea ice, parameter:salinity AND parameter:temperature AND parameter:depth ice/snow* and more granular element-wise versions yields around 230 results. The provided filter options for the search results were not useful to further restrict the results. 21 data sets were selected. It should be noted that PANGAEA data sets Kramer et al. (2010e), Kramer et al. (2010d), Kramer et al. (2010b), and Kramer et al. (2010c) each represent several sources. They are all indicated in the references. The same holds forLange et al. (2015b).

In AADC, search for *sea ice salinity temperature* yields around 1,000 results. Only five data sets were considered fit for purpose.

Secondary resources Data sets are often published as supplements to articles. Such articles provide context to the measurement data and may contain missing elements of the reuse scope. In PANGAEA, articles are often directly linked in the data set, such as the article Torstensson et al. (2018a) in the corresponding data set by Torstensson et al. (2018b). The Zenodo data set from Omatuku Ngongo et al. (2022) provides a reference to an article by Skatulla et al. (2022) in the description of the accompanying PDF file. If no such resource is referenced in the data set, a google search of the campaign name and the data set authors may reveal related articles. For example, the data set from Arndt et al. (2021b) does not reference a specific article, but a google search for the campaign name *PS 118* and the first author's name *Arndt* reveals a journal article by Arndt et al. (2021a) that describes *PS118*. In addition to articles, there are expedition reports, which provide an overview on entire measurement campaigns and may contain reuse scope elements missing in

the data sets. The electronic Publication Information Center (ePIC) is an official repository of the Alfred Wegener Institute (AWI) and publishes its expedition reports.

Tertiary resources Elements that are still missing after searching in primary and secondary resources may be matched from resources unrelated to the specific sea ice core measurements. Tertiary resources are mostly instrument manuals that provide specifications of the instruments. Manuals can be found by a google search of the instrument names or by searching in the manufacturer online shops. Other tertiary resources used in this article are naming standards. They are the Sea Ice Nomenclature, which provides definitions for sea ice and is findable in the World Meteorological Organization's e-Library, and the *The SeaVoX Salt and Fresh Water Body Gazetteer* (SeaVoX) from the British Oceanographic Data Centre (BODC, 2023), which is a controlled vocabulary for water body names. The latter provides the shapefile *Polygon data set of the extent of water bodies* delineating Earth's water bodies into distinct polygons, each tagged with attributes defining its respective ocean or sea. As tertiary resources are independent of specific sea ice cores, matching elements from tertiary resources requires elements already matched from primary or secondary resources. For example, to find an instrument's missing measurement error, the instrument's name must be available in primary or secondary resources to find the corresponding manual, i.e., the tertiary resource.

Step 2 (b): Matching available elements

Element availability matching is necessary due to significant mismatches between structure and content as anticipated in the reuse scope and as actually available from resources. It cannot be assumed that the elements of interest are comprehensively available in data sets. Instead reuse scope elements have to be combined from different resources per sea ice core as illustrated in Figure 5.3. Through element availability matching, data and metadata from resources are first identified as relevant for the reuse scope and then linked to the respective reuse scope elements for each sea ice core. Any element that can be matched from a resource is a hit, and I distinguish between direct and indirect hits. Direct hits occur when the relation of a resource to a reuse scope element is unambiguously understandable to the reuser, such as a column in a data file, in the data set's metadata, or in a table from a secondary resource. The respective data or metadata have to be assigned with explicit, easy to understand, labels. Hits are indirect when additional common sense or context has to be applied for the matching, such as missing or not explicit labels, or in case a graphic provides the respective data or metadata. Availability depends on the resource group from which the element is matched. Primary resources are direct resources and secondary and tertiary resources are indirect resources. Direct hits with direct availability are desired.

Availability matching is a sequential and iterative process. It is sequential because the matching process begins by finding a data set that provides elements for a core or selection of cores. This data set together with further resources is used to find as many elements as possible for this core or selection of cores. Only then does the matching process begin for a new core or selection of cores from another data set. The process is iterative because more scope elements are matched by moving

from primary to secondary to tertiary resources. It should be noted that the matching process is subjective as it depends on the reuser's search method and resources considered. An element may appear to be missing from the resources, but it may eventually be available from a resource that remained undiscovered by the reuser.

Scope elements are available as defined, have to be adjusted to fit the purpose or are unavailable. Further elements may be added during availability matching; they have not been anticipated in the reuse scope but are available from the resources and add value to the database. The checkerboard in Fig. 5.2 (a) shows the availability per element and per sea ice core. In the following, I discuss availability matching separately per reuse scope element as listed in Step 1 (b). The matching process is carried out manually. An automation of the process is challenged by the heterogeneity of the resources, which requires interpretation within the context and core specific resources of interest. The manual matching process is complemented by an automatic enrichment process for elements that are not available as required from the sources. This process is described in detail in Step 5.

ID The ID is a sea ice core-specific name or number used to differentiate sea ice cores. Usually data sets provide a name or a number for each core and location, also called station, in the metadata or as a specified column entry. I compose IDs based on the provided information.

Campaign The name of the campaign, expedition, cruise, or project is available from primary resources for ∼95% of the cores. Direct hits occur when PANGAEA's metadata label *Campaign* or other repository labels such as *Cruise* as in Audh et al. (2022) and Omatuku Ngongo et al. (2022) are provided. In all other situations with direct availability, the campaign name is also a direct hit but from the data set's metadata general information (e.g., Wang et al. (2020b)). For the remaining ∼5% of the cores, which are all sea ice cores from Torstensson et al. (2018b) data set, campaign is indirectly available from contextual information provided in the secondary resource by Torstensson et al. (2018a).

Date Date is available from primary resources for ∼95% of the cores. It is often a direct hit as a column in a data file (e.g., Omatuku Ngongo et al. (2022)), metadata label (e.g., Kramer et al. (2010b)) or both (e.g., Peeken et al. (2018a)). PANGAEA data sets always provide date as metadata except from the data set by Torstensson et al. (2018b), where date is indirectly available as direct hit from a table with column label *Date* provided in the accompanying secondary resource by Torstensson et al. (2018a). The AADC data set from Duprat (2019), provides the column label *Julian Day* in the data file, which would result in an indirect hit because it has to be transformed into a date first. The data set's metadata, however, provides the date so that it is matched as direct hit from there. Secondary resources usually provide the date, when describing the sea ice cores.

Coordinates Coordinates are always provided with direct availability. Most of the hits are direct as data file column labels *Longitude* and *Latitude* in Omatuku Ngongo et al. (2022) or *Location* in the data set by Wang et al. (2020b). For all PANGAEA data sets, coordinates are

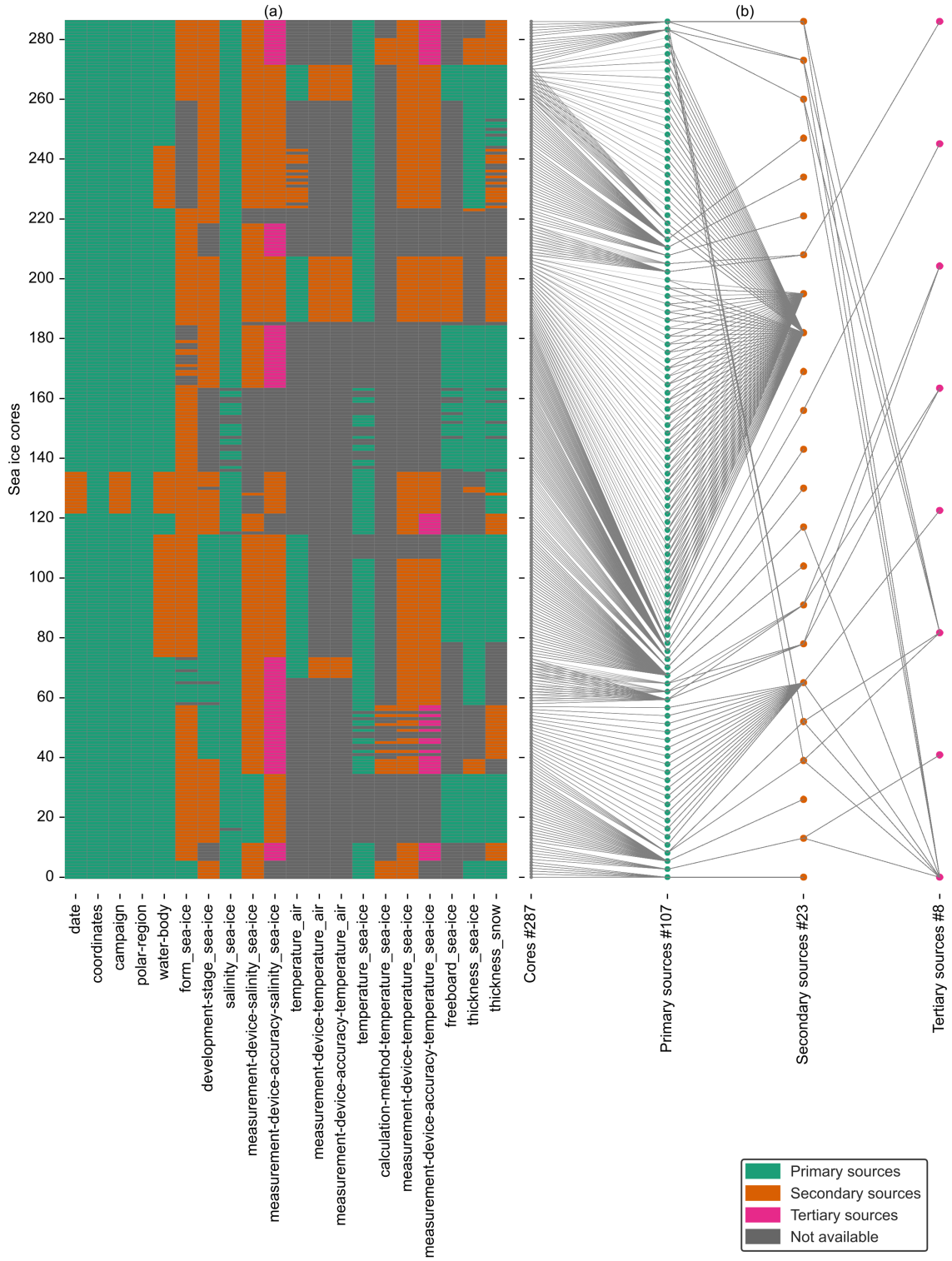


FIGURE 5.2: (a) Overview on the availability of the fields in each YAML-file per sea ice core (y-axes) of the RESICE extendable database. Green, orange and pink refer to availability from a primary, secondary or tertiary resource respectively. Gray indicates the field is not available. (b) Data2source traceability plot for RESICE. Each gray node represents one YAML-file or sea ice core. An interactive version of this figure is available in svg-format (https://www.doi.org/10.5281/zenodo.10866408)(Simson et al., 2025b).

provided in the metadata with labels *Longitude* and *Latitude* (e.g., Lange et al. (2015b)). Coordinates can also be redundant, when available in the metadata and in the data file, which is the case for the data set by Lannuzel (2016a). Meiners (2019) data set consists of several data files in xlsx-format each containing data from different measurements. Instead of directly providing coordinates for the measurements, the names and numbers of the station are assigned per measurement. One data file *PS117_IceStations_Positions_PS117* provides labels for station names *Station* and station numbers *St#* assigned with coordinates *Lat S* and *Long W*. The data file *PS117_IceStations_Chlorophyll_a* providing salinity measurements contains columns named *Station* and *Site*, and the data file *PS117_Ice_Temp_Profiles* providing temperature measurements assigns a *Core_Site* for each core, and it also provides unlabeled comments for each core, e.g., *Ice1*, *Ice2*. Salinity measurements can be linked to coordinates with the *Station* column in form of a direct hit. The linkage of temperature measurements with the corresponding coordinates is ambiguous since the entries for *Core_Site* do not coincide with the entries for *Station* or *St#*; coordinates cannot be easily linked. After closer assessment, I assume that the mentioned unlabeled comments per core such as *Ice1* refer to the station numbers provided with column *Station*. I match the coordinates accordingly as indirect hits. Temperature and salinity data of Meiners (2019) are included as separate cores in the database, since I cannot verify whether the cores were taken in close proximity or not.

Polar region Polar region is always directly available. Most cores have Arctic or Antarctic in the data set title (e.g., Torstensson et al. (2018a); Wang et al. (2020b)) or the data set description (e.g., Omatuku Ngongo et al. (2022); Duprat (2019)), so it is a direct hit. In all other cases, the polar region can be derived from the coordinates, so it is an indirect hit with direct availability.

Water body The name of the sea or ocean, where the core was retrieved, is provided with direct availability for ∼73% of the sea ice cores. Direct hits occur with metadata label *Location* for data sets from PANGAEA and AADC, which is assigned a name for the water body such as *Southern Ocean* (Meiners, 2019), *Scotia Sea* (Kramer et al., 2010e) or *Lincoln Sea* (Lange et al., 2015b). For the other cases, the name of the water bodies are indirectly available as indirect hits from accompanying articles such as for the sea ice cores from Torstensson et al. (2018b) and Wang et al. (2020b), which are matched from maps in Torstensson et al. (2018a) and Wang et al. (2020a) respectively. The water body of the sea ice cores from Duprat (2019) is missing. Information on the use of standardized naming schemes for the water bodies, for instance, in form of controlled vocabularies is not provided for the agnostic reuser.

Salinity sea ice Salinity of the sea ice, along with the depths of bottom and top or center of the respective section (salinity is measured from melted section of the core) are directly available for ∼95% of the cores. For the remaining ∼5%, salinity measurements were not available. The hits are mostly direct because of the distinctive label names for salinity sea ice such as *Salinity/PSU* (Wang et al., 2020b), *Ice bulk-salinity* (Meiners, 2019), *Salinity (psu)* (Duprat, 2019), *Salinity* (Lannuzel, 2016a), and *Sea ice salinity* (Arndt et al., 2021b). The depth label for temperature measurements is often *depth ice/snow*, it can also be *Av. Depth (cm)* (Duprat, 2019), *Ice depth [m]* (Omatuku

Ngongo et al., 2022). The depth label corresponding to salinity sea ice is ambiguous in Meiners (2019) and Torstensson et al. (2018b), where it is labeled *HPLC Core* and *ice thickness* respectively. These are indirect hits since additional context was used for the matching.

Solid fraction sea ice The solid fraction of the sea ice is not a typical measurable. It is therefore not available and missing for all sea ice cores. The reader should note that solid fraction is available in some data sets, where it is post-processed as a function of salinity sea ice and temperature sea ice. Since it is not a measured value, I neglect it.

Temperature sea ice Temperature of the sea ice, along with its measurement depth, is available from primary sources for ∼73% of the cores. Most of the hits are direct because of the distinctive label names for temperature such as *Temperature/°C*(Wang et al., 2020b), *t [°C]*(Omatuku Ngongo et al., 2022) and *Temperature, ice/snow* (Lannuzel, 2016a). Depth labels are similar to those listed for salinity sea ice. A special case is the data set from Torstensson et al. (2018b), where the label name for temperature is *Temperature, water*. However, I understand them as temperature sea ice since the values range between -0.25 and -2.65°C with the lower temperature being too cold for seawater to stay liquid. Furthermore, the corresponding article (Torstensson et al., 2018a) describes the measurement of sea ice temperature directly after core retrieval, but it does not mention temperature measurements of water. Furthermore, the article describes that brine salinity is computed using temperature sea ice and salinity sea ice, which also implies it is the temperature of the ice. Thus,the temperature sea ice is matched as an indirect hit from the column *Temperature, water*. For the remaining ∼27% of sea ice cores, temperature is not available. This is the case for Arndt et al. (2021b), Mundy et al. (2010) and some of the cores from Meiners (2019).

Calculation method temperature sea ice In some cases the temperature data provided for the sea ice cores are linearly interpolated in the data sets. This detail was revealed during the search process and is therefore added to the reuse scope. I found two reasons for linear interpolations. First, temperature measurement data in the sea ice may be linearly interpolated to match the locations of the salinity measurements. This is the case for data sets Lannuzel (2016d) and Lannuzel (2016b) and Duprat (2019), and it is described in the respective articles (van der Merwe et al., 2011b; Duprat et al., 2019). Second, temperatures may be linearly interpolated between one measurement at the surface and an estimate of the temperature at the sea ice bottom. This is the case for the data set from Lange et al. (2015b), described in the accompanying article by Lange et al. (2015a). It is important to note that information on linear interpolation was never directly available from a data set. Instead it is only indirectly available through indirect hits in secondary resources.

Temperature air Temperature of the air is not available for ∼66% of the cores. For ∼29%, it is directly available. Wang et al. (2020b) provides air temperature for each sea ice core as a direct hit with direct availability with label *Air temperature (°C)*. Isleifson et al. (2010a) and Kramer et al. (2010d) provide air temperature with label *temperature, air*. For ∼5% of the sea ice cores, which are all from data set by Audh et al. (2022), air temperature is provided with indirect availability

from secondary resource Johnson et al. (2023) as direct hit, where it is provided in a table. Air temperature is included whenever available.

Temperature and salinity of the seawater Both elements are unavailable for the sea ice cores considered.

Thickness snow Thickness of snow cover on top of the sea ice sea is directly available for ∼51% of the cores. It is as direct hit, for instance, with labels *snow thickness* (Mundy et al., 2010; Arndt et al., 2021b; Kramer et al., 2010a) and *Snow [m]* (Meiners, 2019). In Duprat (2019) and Torstensson et al. (2018b), snow thickness is an indirect hit because it has label *Av. Depth [m]* and *Section* and *Depth ice/snow* so that context had to be used for matching. Thickness of snow is indirectly available for ∼28% of the cores. It is a direct hit for sea ice cores from Lannuzel (2016c), where it is matched from a column with label *Snow* provided in the article Lannuzel et al. (2008) and for sea ice cores from Audh et al. (2022), where it is matched from a table in the article from Johnson et al. (2023) *Snow depth (cm)*. For the sea ice cores from Kramer et al. (2010e) and Kramer et al. (2010d), snow thickness is an indirect hit with indirect availability as it is measured from a graphical representation of the sea ice cores from Haas et al. (2009), which is a chapter in the corresponding expedition report. It is not clear to agnostic reusers whether the absence of snow thickness data is equivalent to missing snow cover.

Thickness sea ice Thickness of sea ice is directly available for ∼64% of the core. It is a direct hit with direct availability for Mundy et al. (2010), Kramer et al. (2010a) and Wang et al. (2020b). Sea ice thickness (Arndt et al., 2021b) is the sum of snow thickness and sea ice thickness, so that first snow thickness has to be subtracted; it is an indirect hit with direct availability. Sea ice thickness is labeled *total core length* in data set by Meiners (2019) according to the provided readme-file, which is an indirect hit. Sea ice thickness is indirectly available for ∼5% of the cores. Direct hits from secondary resources are for sea ice cores from Lannuzel (2016c) since they are available from a table in an article by Lannuzel et al. (2008). The article by Torstensson et al. (2018a) provides sea ice thickness as a direct hit with indirect availability for the data set by Torstensson et al. (2018b). Sea ice thickness is missing in data set from Lannuzel (2016e), but it can be measured from a graphical representation of sea ice from Lannuzel et al. (2016b) not referenced in the data set. For the sea ice cores from Kramer et al. (2010e) and Kramer et al. (2010d) data sets, sea ice thickness is an indirect hit with indirect availability. It is measured from a graphical representation of the sea ice cores provided in a chapter (Haas et al., 2009) in the corresponding expedition report. Sea ice thickness is missing for several of the sea ice cores such as from Nicolaus et al. (2012a) and Nicolaus et al. (2012b) and Audh et al. (2022). Missing values of sea ice thickness were (manually) filled with the deepest measurement location (∼31%) from the salinity or temperature profile, which is an indirect hit with direct availability. For the sea ice cores from Duprat (2019), sea ice thickness is not explicitly reported. Figure 2 in the corresponding article (Duprat et al., 2019) shows the measurement locations along the core. The lowest measurement is 2.5 cm above the ice water interface. Therefore, 2.5 cm was added to the lowest measurement of each core.

Freeboard sea ice Sea ice thickness is often accompanied with sea ice freeboard, which is the extent of sea ice above the water level. Sea ice freeboard is added to the reuse scope since it could be useful for other reuse scenarios. It is available as direct hit with direct availability (∼40% of the cores) in Mundy et al. (2010), Kramer et al. (2010a), Meiners (2019), Arndt et al. (2021b) and Wang et al. (2020b). For the sea ice cores from Kramer et al. (2010e) and Kramer et al. (2010d), sea ice freeboard is a direct hit with indirect availability as it is available in a table of the secondary resource by Haas et al. (2009).

Measurement error Measurement errors for salinity and temperature measurements are not available. Primary resources from AADC discuss reasons for uncertainties in the measurements such as Lannuzel et al. (2017) or outliers such as Duprat (2019). However, the context and related impact on the measurement error is neither quantifiable nor interpretable for data reusers.

Instrument accuracy While a general measurement error is not available, instrument accuracy is available, and it is added to the reuse scope. Instrument accuracy is never directly available. For salinity sea ice, it is provided as direct hit from secondary resources for ∼49% of the cores and for temperature sea ice it is ∼27%. In these cases, instrument accuracy is often combined with the name of the instrument. For ∼28% of the cores, instrument accuracy of salinity, and for ∼16% of the cores instrument accuracy of temperature sea ice is matched from the respective instruments manual.

Instrument name Measurement error is not available as such and instead replaced by instrument accuracy. In some cases, a secondary resource provides the name of the instrument but not the accuracy. Therefore, the name of the instrument is added to the scope, so that it is documented for a subsequent search for accuracy specifications in tertiary sources. The names of the instruments used to measure salinity and temperatures are never available from primary resources except from the sea ice cores provided in Mundy et al. (2010) data set, which contains the name of the salinity measurement device. For the majority of sea ice cores, the instrument for salinity sea ice (∼78%) and temperature sea ice (∼86%) measurements is indirectly available as direct hit, and for the rest it is not available.

Standard deviation Standard deviation for repeated measurements of salinity and temperature measurements is never available. Wang et al. (2020b) data set is the only primary resource that provides standard deviation, in this case for sea ice thickness; it is included in the database. Meiners (2019) provides repeated measurements of snow thickness without inferring standard deviation. The standard deviation is calculated for Meiners (2019) and then integrated in RESICE.

Mean distances of temperature and salinity sea ice The mean distances of measurement locations along the core of the temperature and salinity of sea ice is never available as metadata.

Mean values of temperature and salinity sea ice The mean values of temperature and salinity of sea ice are never provided except the data set from Wang et al. (2020b).

Form and development stage sea ice The elements sea ice form and development stage are often available in a combined form or with similar label names. Therefore, they are searched for together during availability matching. Next, they are disentangled where necessary into two independent elements. Development stage is available from primary resources for ∼25% and from secondary resources for ∼58% of the cores. For sea ice form, it is ∼6% and ∼75% respectively. Indirect hits with direct availability are matched from the column name *ice type* for Pućko et al. (2010b), where it is a combination of both elements *landfast first-year ice*, for Wang et al. (2020b), where it is equivalent to sea ice development stage, and for Peeken et al. (2018a), where it is equivalent to sea ice form. Furthermore, Lannuzel (2016d) provides both elements in a combined form in the comment section, namely *first year pack ice (granular columnar)*. The data set from Duprat (2019) reports that the data represents *land fast sea ice*. Indirect hits with indirect availability are, for example, sea ice development stage for the cores from Audh et al. (2022), matched from information provided in the article from Johnson et al. (2023), where it is called *first-year Antarctic sea ice*. Sea ice form and development stage for the data set from Kramer et al. (2010b) and Kramer et al. (2010c) are also matched from the accompanying article from Kramer et al. (2011), which states sea ice was *first-year ice* and that all cores were *drifting pack ice* except from *IO-5* which was *offshore fast ice*. Furthermore, the use of terms for sea ice development stage often mixes categorical levels of the SIN. Wang et al. (2020b) use *first-year ice*, which is level sub-category *2.5* and *multi-year ice*, which is level sub-subcategory *2.6.3* to define their *ice type* column.

When sea ice form is missing, the SIN can be used to manually fill the element (∼17%) by interpreting sea ice form based on sea ice concentration if available. In this manner, the SIN was used for the data set Wang et al. (2020b) sea ice cores. The accompanying article (Wang et al., 2020a) states that the cores were taken from *vast ice floes* that had *diameters of several kilometers*. Therefore, sea ice concentration is assumed to be above ∼70%, which is consistent with sea ice form *pack ice*. The same holds for the data set by Audh et al. (2022) for which locations the accompanying article (Johnson et al., 2023) shows a map of sea ice concentration. All coring locations appear to be above ∼70%. Lastly, none of the data sets refers explicitly to the Sea Ice Nomenclature (SIN) or other naming standards. Only in the article Skatulla et al. (2022), the standard is referenced. They state "With reference to WMO (code 3739) ice age ID 5 applied for the southernmost ice station and ice age ID 3 for the most northerly station." The mentioned *WMO (code 3739)* defines the development stage of sea ice (WMO, n.d.[c]). Accordingly, ID 3 refers to "predominantly new and/or young ice with some first-year ice" and ID 5 to "all thin first-year ice (30 - 70 cm thick)".

Units and ratios Data sets usually provide units together with the measurement data. This can be in form of the *Unit* column in PANGAEA data sets or combined with the column label in data sets from Zenodo and AADC such as *Depth (m)*, *Ice Temp ©* and *Snow (cm)* in Meiners (2019), *Av. Depth (cm)*, *Temperature. (oC)* and *Salinity (psu)* in Duprat (2019) or *Ice depth [m]* and *sal [PSU]* in Omatuku Ngongo et al. (2022). The ratio for salinity is often not mentioned in the data sets. In this case, it is assumed to have ratio parts per thousand (ppt).

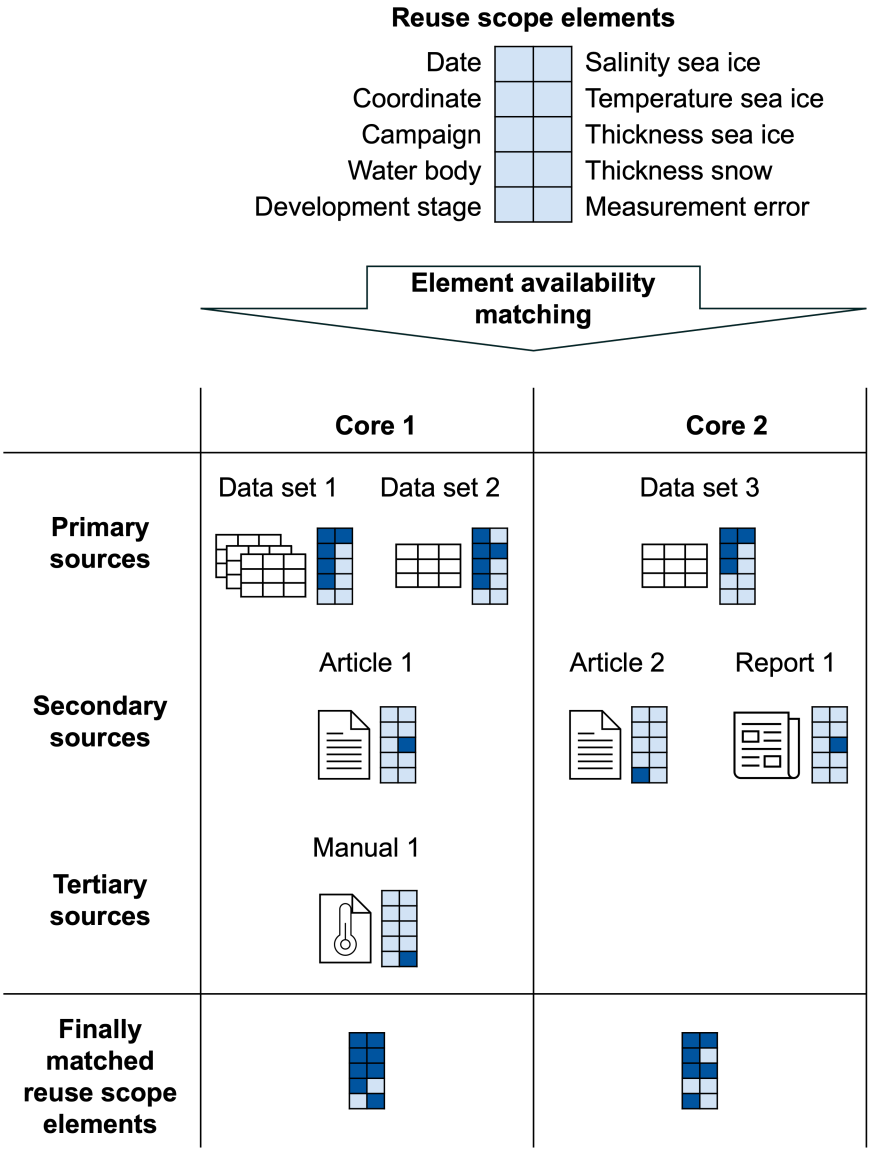


FIGURE 5.3: Element availability matching for two exemplary sea ice cores based on a reduced reuse scope w.r.t. Table 5.1. The ten light blue boxes at the top represent the reuse scope elements. During element availability matching, reuse scope elements are matched. While moving from primary to secondary to tertiary resources, more elements of each core can be matched. Dark blue boxes represent matched elements.

Naming standards Metadata of AADC data sets (Duprat, 2019; Meiners, 2019; Trull et al., 2011; Lannuzel et al., 2017) use Global Change Master Directory keywords to specify the location and the water body. Pućko et al. (2010b) provide a classification scheme used to classify *ice type* column of their data set, which is a combination of *development stage sea ice* and *form sea ice*. The article from Arndt et al. (2021a) provides a classification scheme for the column *ice age classification*, which is similar to the development stages class names from SIN. Furthermore, the article by Skatulla et al. (2022) refers explicitly to the SIN to classify the sea ice development stage as described above. I did not find references to other naming standards.

5.2.3 Step 3: Plausibility and sanity checks of the resources

Several observations in the original resources required special attention for consistent integration into the database. As a result, some original resources were omitted or had to be adapted. Several challenges identified are described in the following sections.

Redundant data sets across repositories

The same data sets may be available from different repositories. I found that PANGAEA data sets by Lannuzel (2016d) and Lannuzel (2016e) are duplicates of the AADC data sets by Trull et al. (2011) and Lannuzel et al. (2017) as they provide equivalent measurements for the same coordinates and the same campaign. The data sets do not reference each other. The measurement values are equivalent, but there are differences between the repository entries. The sea ice core with id *SIPEX-01*, for example, has one more temperature sea ice measurement in the PANGAEA data set, while the AADC data set provides more context on data quality. Furthermore, the PANGAEA data sets by Kramer et al. (2010b) and by Kramer et al. (2010c) appear to be redundant with the PANGAEA data set by Lannuzel (2016d) as they all provide data for the same campaign name, namely *SIPEX*, and dates. However, a closer inspection shows that the coordinates and measurement values are not equivalent between the data sets. Consequently, it is assumed that different measurements took place during the same campaign. All three data sets by Kramer et al. (2010b), Kramer et al. (2010c), and Lannuzel (2016d), are included in the database.

Duplicates within the same data set

Equivalent measurements may appear duplicated in the same data set. One example is the data set from Torstensson et al. (2018b), which has repeated measurements for Fucoxanthin concentration at the same location. Salinity and temperature measurements are not repeated. Instead, they are duplicated in the same data set and per Fucoxanthin measurement. In this case, temperature and salinity data is only included once per location in RESICE. Another example is Lannuzel (2016c), where cores *05*, *06*, and *07* are assigned equivalent temperature measurements, and cores *07* and *08* are assigned equivalent salinity measurements. Here, all data was kept, as it is not clear from which location the measurements originate from. Omatuku Ngongo et al. (2022) assign equivalent snow thickness measurements to repeated salinity and temperature measurements from the same location. In this case, all snow thicknesses are kept and added to RESICE. Katlein et al. (2020a) provide potential duplicates in the data set. Each *Depth ice/snow* value is duplicated except from the first and the last ones. The measurement values seem doubled but shifted by one depth value, and each measurement value is included only once in RESICE.

Incorrect metadata

The metadata provided in the resources may be incorrect. This is the case for the name of the salinity measurement device in Torstensson et al. (2018b), which is provided as *Cond 310i* from

manufacturer *WTW*. After searching for the related instrument accuracy via google, the manufacturer's homepage and consulting the costumer service of WTW, the conclusion was that there exist instruments *Cond 315i* and *Cond 3110* but no *Cond 310i*. The article Skatulla et al. (2022) also provides a non-existing instrument for temperature measurements for data set Omatuku Ngongo et al. (2022), which is called *GMH 3750-GE logger* from manufacturer *Testo*. However, this instrument cannot be found via Google or the manufacturer's homepage. Instead an instrument with this name is available from the manufacturer *Greisinger*. I changed the name of the manufacturer in RESICE.

Inconsistencies within the same data sets

In the data set by Meiners (2019) data and metadata is stored in separate files, and the connection between them is ambiguous due to inconsistent naming. It is not clear if salinity and temperature measurements originate from the same core, individual cores in close proximity, or individual cores at different locations. Temperature and salinity measurements are therefore stored in separate YAML-files to differentiate them. In the data set by Mundy et al. (2010), there is an inconsistency in the depths assigned to the salinity measurements. Salinity measurements are made for a melted section of the core, so that one salinity measurement is associated with two depths, one at the top and one at the bottom of each section. Generally, both depths values are used to compute the center of the section, which is the depth assigned to the salinity measurement in RESICE. In Mundy et al. (2010), the depth for the top of the section often has a higher value than the bottom of the section, for instance, it is 0.900 m for the top and 0.110 m for the bottom for one of the cores. I assume that the bottom depth has a typo and should be 1.100 m. The center of the section is calculated with the corrected value. Another inconsistency is in the data sets by Lannuzel (2016a), Lannuzel (2016b), and Lannuzel (2016e), where sea ice and snow thickness as well as sea ice form and development stage are provided as metadata in the comment section. After checking with the measurements, I found that values for sea ice thickness are implausible as the measurement depths go beyond the provided sea ice thickness. Furthermore, the assigned development stages are inconsistent. Comment sections of these data sets are neglected.

Inconsistencies between resources

For many sea ice cores more than one source provides relevant data and metadata, which may arise inconsistencies between these sources. This could be, for instance, a naming inconsistency making the connection between the sources difficult. This is the case for the names of the sea ice cores in data set by Lannuzel (2016a) and accompanying article by Lannuzel et al. (2007). I detected this inconsistency through sea ice thickness values from Table 1 in the article by Lannuzel et al. (2007), which did not fit to the measurement depths of the respective cores from the data set by Lannuzel (2016a). Apparently, there was a mix-up in the naming, so that sea ice cores *V*, *IX*, and *VII* from the article by Lannuzel et al. (2007) are equivalent to *IX*, *VII*, and *V* in the data set by Lannuzel (2016a). I used the naming of the article. Another naming inconsistency was found between data set by Lannuzel (2016b) and article by van der Merwe et al. (2011a), where sea ice core *XX* from the data set, is matched with data and metadata such as sea ice thicknesses from

sea ice core *10* in the article. Other inconsistencies include different dates in different resources for the same core. For example, data set by Lannuzel (2016d) assigns 17/09 to sea ice core *05*, while the article by Lannuzel et al. (2016b) assigns 18/09. In RESICE, I use the date from the data set. Another inconsistency concerning sea ice form was found for the cores from Meiners (2019) from campaign PS 117. The corresponding expedition report states that 5 of the 8 ice stations were located in the *Eastern Weddell sea*, which was a *pack-ice zone*, and 3 stations were located in the *Western/North-Western Weddell sea*, which was a *marginal ice zone* (Castellani et al., 2019). However, after checking all locations on the map, 8 of the sea ice core locations lay in the eastern Weddell Sea. Another example is an inconsistency for the sea ice development stages for the sea ice cores from Kramer et al. (2010e) and Kramer et al. (2010d). The accompanying article Kramer et al. (2011) defines sea ice development stage for the four sea ice cores *WS-4*, *WS-7*, *WS-11*, and *WS-21* as *multi-year ice* and the rest as *first-year ice*. The corresponding expedition report by (Haas et al., 2009) provides sea ice development stage for each date in a table, which assigns *WS-21* *first-year ice* and *WS-4*, *WS-7*, *WS-11* are all *second-year ice*.

Misleading semblance of accuracy

The given decimal numbers of a measurement may exceed the actual possible accuracy imposed by the instrument. The data set by Lange et al. (2015b) provides sea ice salinity and temperature with three decimal numbers. The instruments used have accuracies of 0.1 for salinity and 0.2° for temperature. The same holds for Torstensson et al. (2018b), who provide sea ice temperature with 3 decimal numbers although the reported accuracy of the measurement device is 0.1°. Also the data sets from Lannuzel (2016a), Lannuzel (2016b), Lannuzel (2016c), Lannuzel (2016d), and Lannuzel (2016e) give two decimal places for sea ice temperature and three decimal numbers for sea ice salinity, although the accuracy of the salinity instruments are 0.4 ppt and 0.06 ppt and the accuracy of the temperature instrument is 0.2°. In this case, all measurement values are rounded to 1 decimal place before integrating them into RESICE.

Reference depth unclear

The reference depth of the measurements along the central axis of the sea ice cores is not always clearly defined. In RESICE, reference depth of 0 m should be at the interface of snow and sea ice. Measurements in the sea ice should be defined a positive value. Omatuku Ngongo et al. (2022), Audh et al. (2022) and Arndt et al. (2021b) define reference depth as requested and assign measurements in the sea ice a positive depth. In some data sets reference depth is not explicitly defined, but it can be derived from the context. For Duprat (2019), measurements in the snow are consistently assigned a positive depth and measurements in the sea ice a negative depth. The signs of the depth values were switched assuming reference depth is again at the snow sea ice interface. Wang et al. (2020b) do not explicitly define reference depth. Since the reported sea ice thickness is equivalent to the respective deepest sea ice core measurement, I conclude that the reference depth is at the requested position. Meiners (2019) does not define reference depth but always provides snow thickness measurements. Thus, I conclude reference depth is again at the interface of sea ice and snow. Lannuzel (2016c) does not define reference depth, but the accompanying article (Lannuzel

et al., 2008) provides a table, where snow thickness is assigned a negative depth. Therefore, I assume reference depth is as requested. The majority of data sets do not define the reference depth. In these data sets (Mundy et al., 2010; Kramer et al., 2010b; Kramer et al., 2010c; Kramer et al., 2010d; Kramer et al., 2010e; Pućko et al., 2011b; Nicolaus et al., 2012b; Lange et al., 2015b; Lannuzel, 2016a; Lannuzel, 2016b; Lannuzel, 2016d; Lannuzel, 2016e; Torstensson et al., 2018b; Katlein et al., 2020a), I assume reference depth of 0 m to be at the intersection of snow and sea ice. The reader should note that PANGAEA data sets such as by Peeken et al. (2018a) and Katlein et al. (2020a) link the measurement depth along the core, i.e., *depth ice/snow*, with the geocode wiki of PANGAEA (PANGAEA, 2023b). The figure provided in the wiki does not clearly enough define reference depth.

5.2.4 Step 4: Technical combination of the resources

In the next step, the selected resources are technically combined. The availability matching follows a logic that is individual per core and cannot easily be automated (if at all). Therefore, each core should be represented in a separate file instance. This file should allow transparent documentation of the matching process and flexibly store elements of different nature, e.g., strings, tabulated data, scalars. I choose the YAML format due to its flexibility. The elements matched from sources have to be made available in Python and then written into YAML-files. The ensemble of YAML-files is referred to as RESICE extendable database. It is extendable since it is hosted in a public GitLab repository, and, therefore, allows for pull requests and issue reporting from the community in the future. The instantiated YAML-files are all set up in the same scheme, so they can be subsequently combined into the RESICE tabular database by merging the YAML-files into a tabular data frame in Python. Figure 5.4 highlights the major differences between the extendable and tabular databases. RESICE extendable database is the result of Steps 1 to 4 and only contains data and metadata from existing sources, while the RESICE tabular database also contains automatically enriched metadata as will be explained in Step 5. Thus, the databases represent two different enrichment stages of RESICE, namely enrichment stages 1 and 2, as annotated in Fig. 5.4. The RESICE tabular database constitutes the final product of the compilation process. I refer to elements in context of reuse scope, fields in context of YAML-files and columns in context of the tabular database. In the following, the creation of the YAML-files is explained.

Supplying resources to Python

The process of loading primary resources into Python for the initiation of the YAML-file writing process depends on the data set formats and structures as they are provided on the data repositories. Data sets from PANGAEA can be loaded directly into Python as a Pandas DataFrame using the panageapy package (Huber et al., 2020). The data files from Zenodo and AADC data sets are downloadable via API, for instance, with the datahugger Python package. However, the downloaded data files are not readily available in Python. The data files first have to be read in a next step. Another option is to download the data files directly from the data repositories.

Some data files from Zenodo and AADC are not efficiently machine readable and have to be manually adjusted. Examples are the data sets from Meiners (2019) and Wang et al. (2020b), where the data files are structured in an unsystematic form with repeated column names for each core. Once the data is supplied to Python, it is written to YAML-files. Different scripts for YAML-file writing are required depending on the content and naming scheme of the individual data sets. Elements matched from secondary and tertiary sources are manually provided to the script.

Harmonization

Matched elements have to be harmonized, when transferred to the YAML-files. Harmonization includes label names, classes, units, coordinates, dates, and measurement depths.

- Label names used for the same elements vary between resources, but they have to be consistent across all YAML-files. Therefore, a naming convention for the YAML-files is necessary. I choose an underscore as the join between the property name (e.g., date, temperature, thickness) and the material (e.g., sea ice, air) such as *temperature_ air*. Properties or material names that consist of several words such as *measurement-device-accuracy-temperature_ sea-ice* are separated with a hyphen.

- For categorical elements such as *form_ sea-ice* and *development-stage_ sea-ice*, there exist different naming for the same classes, which requires name harmonization. Examples are *multi-year ice* and *multiyear ice* as well as *landfast ice* (Peeken et al., 2018a) and *fast ice*. I adapt the respective class names to a SIN classes if it can be assigned.

- Units and ratios need to be consistent across all YAML-files. Temperature in Celsius (T_C) has to be converted to temperature in Kelvin (T_K) following $T_K = T_C + 273.15$. If the unconverted T_C temperature has only one decimal place, the converted temperature in K will be rounded accordingly. Salinity is in principle unit less, but it is assigned a ratio. The data is given in either practical salinity unit (psu) or parts per thousand (ppt), which is equivalent to $g\,kg^{-1}$. Salinity is converted from psu (S_{psu}) to ppt (S_{ppt}), using the formula $S_{ppt} = S_{psu}\frac{35.16504}{35}\,g\,kg^{-1}$ from Millero et al. (2008), which may effect changes of the second decimal place.

- Coordinates must be converted if they are not in decimal degrees. This is the case for Audh et al. (2022) and Meiners (2019), who use degrees, minutes, and seconds.

- The date format has to be converted if it does not follow YYYY-MM-DD, such as in Meiners (2019).

- Temperature is measured point wise at locations along the core, and salinity is measured volume wise per melted core sections. For salinity measurements, the measurement depth has to be adjusted if two depths, i.e., the bottom and the top of a section, are assigned to one measurement. The center of the section is calculated and used as depth.

YAML-file format and scheme

The elements matched from the resources are combined in one YAML-file per sea ice core. Each YAML-file follows a structure of fields and sub-fields. The assignment of reuse scope elements to YAML-file fields follows the reuse scope, unless a hierarchical combination of elements within one field and its sub-fields is logical. For instance, instrument accuracy of a thermometer would be hierarchically a sub-field of the *temperature_ sea-ice* field, and standard deviation of sea ice thickness would be a sub-field of the *thickness_ sea-ice* field. However, instrument accuracy usually comes from a different resource than the temperature data. Thus, it is represented as a separate field, while the relationship of the two fields is obvious from the naming convention. On the contrary, standard deviation for sea ice thickness is provided in the same resource as the measurement data in the data set by Wang et al. (2020b). Thus, standard deviation is added as sub-field to the *thickness_ sea-ice* field.

The general structure of the fields and sub-fields together and with examples for *temperature_ sea-ice* and *thickness_ sea-ice* are provided in Table 5.3. Each field's type is defined by sub-field *type*, which can be coordinate, string, scalar, or tabulated. The type defines the form of the sub-field *value*. If type is tabulated, as for *temperature_ sea-ice*, sub-field *value* is of form dictionary with each key representing the measurement depth and each value representing a measurement. The sub-fields *unit_ str* and *unit* define the unit of *value*, where the first is human-readable version and the latter defines the unit with systematically documented base SI-Units so that `[0 0 0 1 0 0 0]` represents the unit Kelvin according to `[ kg m s K A mol cd ]`. The same holds for the units of *variable*, which defines the unit of the keys for *type* tabulated. In case of *temperature_ sea-ice*, it would be *depth ice/snow*. Each field has a single resource. The resource of each field is provided with its name as sub-field *source* and its DOI as sub-field *doi*. If no DOI is available, it is filled with the URL.

Traceability

RESICE requires reproducibility of the availability matching and potential changes with respect to their original resources due to plausibility checks or harmonization. This is implemented in RESICE through several traceability options. First, each YAML-file field has sub-fields *source* and *doi*, where the name and DOI or URL of the original resources are stored. These fields allow to reach the original resources quickly. For elements matched with direct hits and direct availability, these fields are sufficient to ensure the process is comprehensible and reproducible.

As soon as elements are matched through indirect hits or indirect availability or their units need to be harmonized, more options for a traceable documentation are required. Examples are elements that were matched from a table of a secondary source, which has to be documented, or subjective decisions in the matching process, such as when missing sea ice thickness is matched with the depth of the lowest measurement. For these cases, I use the sub-field *comment* in the YAML-file, where the process is commented (e.g., table number, excerpt quote). The *comment* field is also used to comment changes due to harmonization and plausibility checks.

TABLE 5.3: General description of YAML-file sub-fields (a) with examples (b) for fields *temperature_sea-ice* and *thickness_sea-ice*. The sub-fields marked with an asterisk are compulsory for every field. Other sub-fields are required depending on the element type and the peculiarities of the availability matching process (mainly sub-field *comment*). The attributes involving *variable* are only required for elements of type *tabular*.

General (a)		**Examples (b)**	
Sub-field	**Description**	**temperature_sea-ice**	**thickness_sea-ice**
type*	Defines the type of the element as string/scalar/tabulated	tabulated	scalar
value*	Text/float/dictionary	{0.05: 271.50, 0.15: 271.25, 0.25: 270.70, [...]}	0.825
unit_str	Standard unit symbol for the value	K	m
unit	Unit of the value in machine readable format [kg m s K A mol cd]	[0 0 0 1 0 0 0]	[0 1 0 0 0 0 0]
comment	Matching and plausibility - relevant excerpt, table or figure - inconsistencies or adjustments	from C to K	Sea ice thickness is not explicitly available. Substituted by depth of the lowest measurement.
variable	Labels for the dictionary keys if type is tabulated	depth ice/snow	-
variable_unit_str	Standard symbols for the units of the keys if type is tabulated	m	-
variable_unit	Unit of the dictionary key if type is tabulated in machine readable format [kg m s K A mol cd]	[0 1 0 0 0 0 0]	-
source*	Abbreviation of the resource as listed in Table C.2	Torstensson_et_al_2018a	Audh_et_al_2022
doi*	DOI or URL of the resource	10.1594/PANGAEA.924295	10.5281/zenodo.6997630
adjusted	Indicates adjusted temperature and salinity measurement data with respect to the original resource adjusted: 1.0, else: 0.0	1.0	

A second traceability option is required to transparently document the inconsistent provision of the same element from different sources, for instance, two different water bodies or sea ice development stages. In this case, an extra field is added to the YAML-file, which is named after the actual field name combined with the suffix *option*. The field without the suffix would be the official field used for the final RESICE database. Table 5.4 shows an example of the traceability options to ensure transparency of the matching process for the field *development-stage_sea-ice*. The element can be matched from two resources that provide conflicting classes for sea ice development stages. Therefore, both development stages are stored in the YAML-file with separate fields, one with suffix *option*. Additionally, the sub-field *comment* provides details on the excerpt and the table from where the fields were matched. The sub-field *adjusted* is added to highlight changes of the measurement data as saved in the YAML-file with respect to the original resources. It is only provided for temperature and salinity measurements, and it is 1.0 if changes such as unit conversion

TABLE 5.4: Example for the transparent consideration of inconsistencies in the matching process in YAML-files for the element *development-stage_ sea-ice* of sea ice core with *id PS69_ 584-1_ WS-21*. *comment* sub-field contains an excerpt for *development-stage_ sea-ice* and for *development-stage_ sea-ice_ option* in provide the table from where the information development stage was matched. The *option* suffix is necessary, since the same element was available from two different resources.

Sub-field	**development-stage_ sea-ice**	**development-stage_ sea-ice_ option**
type	string	string
value	multi-year ice	first-year ice
comment	The samples from stations WS-4, WS-7, WS-11, and WS-21 were multi-year ice covered with second-year snow, whereas the samples from all other stations were first-year ice (Haas et al., 2009, Willmes et al., in press).	from Table 1
source	Kramer et al. (2011)	Haas et al. (2009)
DOI	10.1016/j.dsr2.2010.10.029	10.2312/BzPM_0586_2009

and plausibility checks have been conducted. In all other cases it is 0.0.

Creation of the tabular database

The RESICE tabular database is created by reading in the YAML-file fields including the sub-fields *value*, *source*, *doi*, and if available *comment* and *adjusted*, into a Pandas dataframe in Python. Each sub-field is transferred to a column with a name that combines YAML-file field and sub-field names (e.g., *temperature_ sea_ ice_ source*). If the field has a sub-field specifying the unit, the unit is combined with the column name (e.g., *temperature_ sea_ ice [K]*). All hyphens of the YAML-file field names are changed to underscores. The fields with suffix *option* are neglected. Instead of combining the measurement depth in tabulated form directly with the measurement values, the tabular database has a *depth* column with all measurement depths that are unique per core. YAML-file field *coordinate* is split into *Latitude* and *Longitude* columns in the tabular database. The interested reader can find the functions used to merge the YAML-files into the tabular database in the `read` module of pyresice Simson et al. (2025a). After the automatic metadata enrichment in Step 5, the tabular database is exported as csv-files as illustrated in the green in Figure 5.4.

5.2.5 Step 5: Automatic metadata enrichment of reuse scope elements

The YAML-files in the RESICE extendable database combine all data and metadata found in primary, secondary, and tertiary sources. Yet, not all elements as requested in the reuse scopes could be matched from sources in Step 2 (b) and are therefore not included in the final RESICE tabular database. In some cases the data reuser can infer these unavailable elements from already matched elements conserved in the YAML-files. For RESICE, Python routines are proposed to systematically derive missing elements in an automatic way. It is important to note that the enriched elements are not added to the YAML-files. They are directly provided as columns to the RESICE tabular database, which is created by merging the YAML-file contents in a tabular dataframe. The workflow is illustrated in Figure 5.4, and the automatic enrichment routines are provided in the

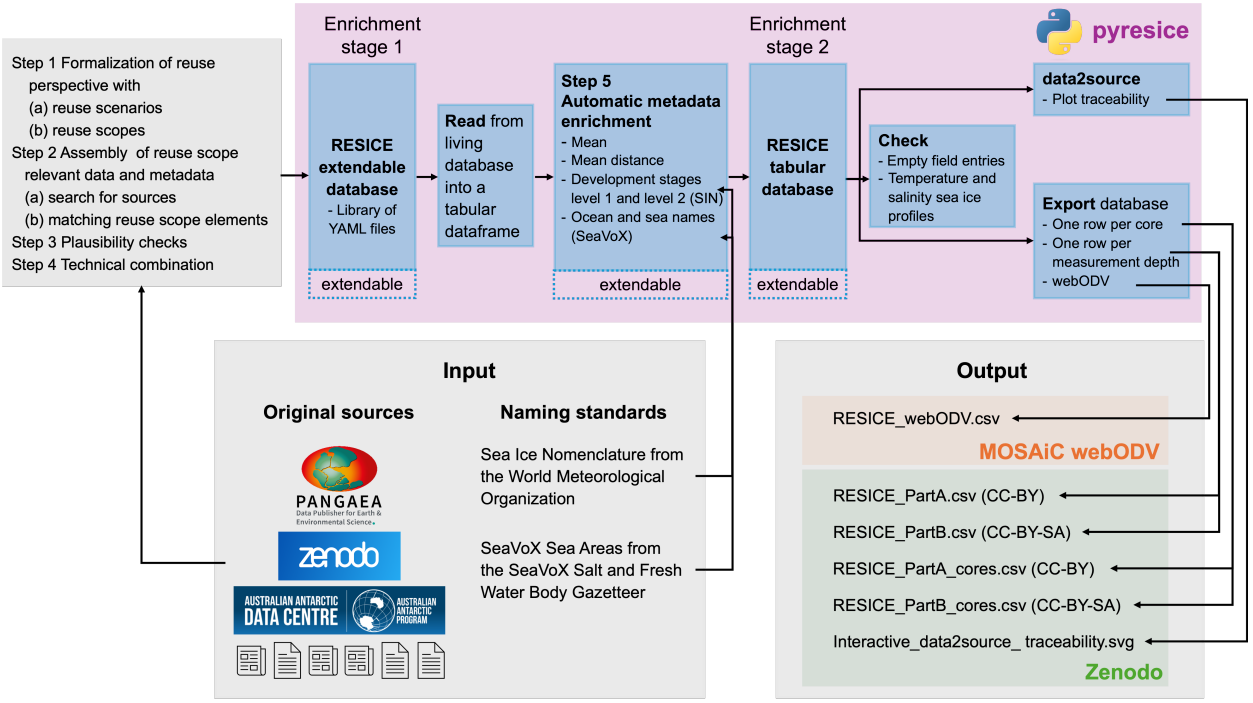


FIGURE 5.4: Shows the RESICE tabular database preparation steps. Starting from the original resources, the RESICE extendable database is created following Steps 1 to 4. The YAML-files of the extendable database are then merged into a tabular dataframe, which is automatically enriched in Step 5. The RESICE tabular database is then checked for empty field entries, and consistency of the measurement profiles. Next, the output files are exported. They are available in Zenodo and the MOSAiC webODV. The blues boxes in the purple box indicate the modules of the pyresice Python Package. RESICE extendable and tabular databases can be extended either by adding new YAML-files to the RESICE extendable database or new fields to existing YAML-files, or by adding new routines to the `automatic_enrichment` module. Lastly, the `data2source` module contains the function to create the plot `Interactive_data2source_traceability.svg`.

`automatic_enrichment` module of the pyresice Python package. The columns *mean_ salinity_ sea-ice*, *mean_ temperature_ sea_ ice*, *mean_ distance_ measurements_ salinity_ sea_ ice* and *mean_ distance_ measurements_ temperature_ sea-ice*, *sea_ SeaVoX*, *ocean_ SeaVoX*, *development_ stage_ SIN _ level_ 1_ sea_ ice* and *development_ stage_ SIN_ level_ 2_ sea_ ice* are automatically enriched as described in the following.

Mean distances of temperature and salinity sea ice To enable Reuse Scenario B1, mean distance between the profile measurements is calculated and then averaged. The mean distance of the measurements is automatically enriched for all sea ice cores that provide the respective temperature and salinity measurements.

Mean values of temperature and salinity sea ice To enable Reuse Scenario B2, the mean value is calculated from the profile measurements of sea ice temperature and salinity. The mean value of the measurement data is automatically enriched for all sea ice cores that provide the respective temperature and salinity measurements.

Development stage sea ice from SIN level 1 and level 2 To enable Reuse Scenario B, development stage of sea ice has to follow the classes of the Sea Ice Nomenclature (SIN)WMO, 2014 in a consistent way. However, the development stages for sea ice provided in the sources do not always match a SIN class, or they are used in a cross-categorical manner as explained in Step 2 (b). Therefore, two columns are automatically enriched in the RESICE tabular database. They are *development-stage-SIN-level-1_ sea-ice*, which refers to all classes on the *2.x* level, and *development-stage-SIN-level-2_ sea-ice*, which refers to all classes on the *2.x.y* level of the SIN. More specifically, *development-stage-SIN-level-1_ sea-ice* and *development-stage-SIN-level-2_ sea-ice* are derived based on the YAML-file fields *development-stage_ sea-ice* and/or *thickness_ sea-ice*, which were matched from sources as explained in Step 2 (b). A dictionary of all SIN class names of level 1 and 2 is used to derive the level of the class of *development-stage_ sea-ice*. If it corresponds to level 1, subordinate level 2 is derived using *thickness_ sea-ice* by matching it with the characteristic thickness ranges of the level 2 classes as listed in Table 5.2. If it corresponds to level 2, superordinate level 1 is easily derived as it is the parent of level 2. It may occur that the YAML-file field *development-stage_ sea-ice* is consistent with a level 1 or level 2 class from the SIN, while the YAML-file field *thickness_ sea-ice* does not match the corresponding characteristic thicknesses of the respective SIN class. In this case, the YAML-file field *development-stage_ sea-ice* is neglected, and *development-stage-SIN-level-1_ sea-ice* and *development-stage-SIN-level-2_ sea-ice* are derived based on the *thickness_ sea-ice* alone, i.e., the corresponding level 2 class is matched based on the thickness following the thickness ranges in Table 5.2 and then superordinate level 1 is derived. The same applies to the case when *development-stage_ sea-ice* is not available from a source or it does not match with a SIN class. In these cases, a warning is issued and stored in the RESICE tabular database column *INFO_ SIN*. The reader should note that the SIN classes corresponding to new ice, pancake ice, and ice rind are excluded from the automatic enrichment routine because they are not associated with sea ice thicknesses or intersect with other characteristic thicknesses. Furthermore, the characteristic thickness of the level 2 class *residual ice* intersects that of *first-year ice*. To avoid conflicts of class assignment with the automatic enrichment routine, *development-stage-SIN-level-2_ sea-ice* can only be *residual ice* if the YAML-file field *development-stage_ sea-ice* is equal to *old ice* or *residual ice*. If the automatic enrichment is based only on the *thickness_ sea-ice* field due to missing details on the development stage from sources, it will only assign classes of *first-year ice* even if it could also be residual ice. In this case the result should be treated with caution. A warning is issued in the column *INFO_ residual_ ice*.

Sea and ocean from SeaVoX The names for water bodies as they are provided in the sources and stored in the YAML-file field *water-body* do not comply with a controlled vocabulary. Liza's Reuse Scenario B requires names defined with the *SeaVoX Salt and Fresh Water Body Gazetteer* (BODC, 2023). Therefore, the two elements *ocean-SeaVoX* and *sea-SeaVoX* are automatically enriched, and they are assigned the SeaVoX attributes *OCEAN* and *SUB_ REGION*. More specifically, the *Polygon data set of the extent of water bodies* shapefile was loaded into Python using GeoPandas and then evaluated each core's coordinates as stored in the YAML-files for the polygon attributes *REGION*, which is equivalent to *ocean-SeaVoX*, and *SUB_ REGION*, which is equivalent to *sea-SeaVoX*.

5.3 Publication of the database RESICE

The static version of the RESICE tabular database is available in its newest version in csv-format from Zenodo (Simson and Kowalski, 2025b; Simson and Kowalski, 2025c). The full RESICE tabular database has the length of the amount of unique measurement depths of the sea ice salinity and temperature measurements of all cores. It includes 2745 sea ice temperature measurements and 2862 sea ice salinity measurements. In total, it has 4327 rows since many salinity and temperature measurements of the core are assigned the same depth.

As different licenses apply to the original sources of the data, the RESICE tabular database is split into two Zenodo entries. The data sets are provided in RESICE - Reusability Enhanced Sea Ice Core Database - Part A (Simson and Kowalski, 2025b) with CC-BY licenses and Part B (Simson and Kowalski, 2025c) with CC-BY-SA license. Part A and Part B of the database are linked via RESICE - Reusability Enhanced Sea Ice Core Database - General Information (Simson and Kowalski, 2025a) providing metadata and information on the column labels. The Zenodo entries of Part A and Part B contain the following csv-files.

- `RESICE_PartA.csv`/`RESICE_PartB.csv`: These csv-files contain the database with one row per measurement depth of the sea ice salinity and temperature measurements. The column labels of the csv-files are listed in Appendix C.1.
- `RESICE_PartA_cores.csv`/`RESICE_PartB_cores.csv`: These csv-files constitute a reduced version of the RESICE tabular database. In this case, the columns representing the profile measurement data, i.e., *depth*, *salinity_sea_ice* and *temperature_sea_ice* are neglected. Part A and B combined have 287 rows so that one row represents one sea ice core.
- `sources.csv`: This csv-file lists the respective sources, licenses and dois/urls used to create the csv-files.

Additionally to the tabular database, the RESICE extendable database, i.e., the ensemble of YAML-files, is provided in the pyresice Python package available via GitLab (Simson et al., 2025a). Specifically it is provided in module `RESICE_extendable_database` in the folder `src/pyresice`. The YAML-file field names are listed and explained in Appendix C.1 together with the column labels of the tabular database.

5.4 Reuse pathways of RESICE

The main goal of RESICE is to enable reuse scenarios that require sea ice core data by lowering the data access threshold. Therefore, the use of RESICE is facilitated by providing it on three different platforms. 1) RESICE tabular database is interactively viewable and analyzable on the web in the MOSAiC webODV. 2) RESICE tabular database is published in csv-format in Zenodo. 3)

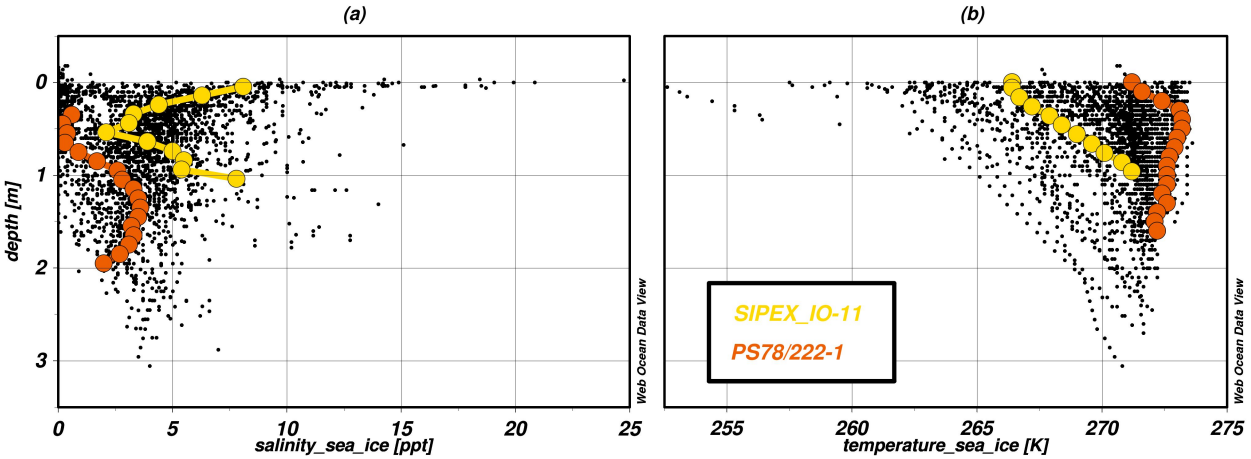


FIGURE 5.5: All sea ice temperature (a) and salinity (b) measurements of the RESICE tabular database combined in a scatter plot. The profiles for the two sea ice cores with IDs SIPEX_IO-11 and PS78/222-1 are highlighted. This figure was prepared with webODV and is available in an interactive version (https://mvre.webodv.cloud.awi.de/DataExploration/id/DVevtE7c/sal_temp_scatter).

RESICE extendable database is provided in the pyresice Python package in GitLab, which includes all routines required to generate the RESICE tabular database.

5.4.1 Interactive webtool: MOSAiC webODV

RESICE has been added to the interactive online visualization and analysis tool MOSAiC webODV, where it is available as supporting data for analysis of data from the MOSAiC campaign (Mieruch et al., 2023). Several figures are accessible via the collection (https://mvre.webodv.cloud.awi.de/DataExploration/id/DVevtE7c). The reader should note that before accessing RESICE, a webODV login is necessary. Anonymous login is possible. After reaching the RESICE collection, figures can be selected via *view* on the top right and then via *load views*. Additionally to the scatter plot in Figure 5.5, correlation plots for Antarctica (https://mvre.webodv.cloud.awi.de/DataExploration/id/DVevtE7c/antarctic_correlation) and the Arctic (https://mvre.webodv.cloud.awi.de/DataExploration/id/DVevtE7c/arctic_correlation) as well as a plot for the development stages (https://mvre.webodv.cloud.awi.de/DataExploration/id/DVevtE7c/dev_stage) are available. The user can furthermore generate custom plots. The csv-file used to integrate RESICE in MOSAiC webODV can be generated with the respective function of the `export` module in pyresice.

5.4.2 Tabular database: RESICE on Zenodo

If users plan to use RESICE in their work, the download of the tabular database from Zenodo is suggested. As explained in the Data Records, two Zenodo entries for the RESICE database were required due to the different licenses of the original data resources. Therefore, it is suggest to load the csv-files of Part A and Part B into Python by creating a Pandas DataFrame, merge the data

frames, and reduce the database to the columns required for the specific scenario. If the RESICE extendable database is changed, a new version of the tabular database will be provided via Zenodo.

It should be noted that all column labels with suffixes *doi*, *source*, and *comment* provide metadata on the origin of the respective data point and are equivalent to the content of the respective YAML-file sub-fields. When working with the files `RESICE_PartA.csv`/`RESICE_PartB.csv` users should note that negative depths in the *depth* columns indicate that *temperature_sea-ice* and *salinity_sea-ice* were measured in the snow.

5.4.3 Extendable database and automatic enrichment: pyresice Python package

The pyresice Python package should be used for extending RESICE or reproducing the output data files and the interactive data2source traceability plot (see Fig. 5.2). The Python package is available via GitLab. It currently is a local Python package created based on the template provided by Cookiecutter. The package can then be imported using Poetry (https://github.com/python-poetry/poetry). The package is licensed with GPLv3 since the RESICE extendable database also contains data licensed by CC-BY-SA. The general package structure is illustrated in the purple box on the top right in Figure 5.4. Each blue box represents a module of the package, i.e., a sub-folder in the `src/pyresice` folder of the package.

The modules 1) `RESICE_extendable_database` or 2) `automatic_enrichment` can be used to extend the database RESICE.

1) The RESICE extendable database can be extended by adding new parameters to a sea ice core, i.e., adding new fields to existing YAML-files. For instance, snow thickness for sea ice core with *id PS78/230-1* is missing. A user, who knows a resource that provides the snow thickness for this core, could add a *thickness_snow* field in the respective YAML-file. When adding a new field to the YAML structure as explained in Table 5.3 and the naming scheme explained in the readme file (https://git.rwth-aachen.de/mbd/pyresice/-/blob/main/src/pyresice/RESICE_extendable_database/yaml_db/readme.md) provided in the module `RESICE_extendable_database` should be followed. Furthermore, the RESICE extendable database can be extended by instantiating a new YAML-file and adding a new sea ice core. For instance, if users have been part of a field campaign and measured salinity of sea ice cores that are not yet part of RESICE, they could add a new YAML-file for each new core to the database. For the extension of RESICE with additional sea ice cores the procedure is the same as described in the Reusability-targeted approach. The original resources have to be checked for plausibility, supplied to Python, harmonized, and then combined in a YAML-file. The reader should note that a YAML-file can also be generated manually and without a Python script.

2) By adding functions to the `automatic_enrichment` module, the RESICE tabular database can be extended. There may exist another scheme for the classification of the development

stages of sea ice that is of interest for a data reuser. This scheme could be transferred into a Python routing and then added to the module.

Generally, extensions can then be added through a pull request via the GitLab repository. The pull request would be granted after a quality check of the new data that verifies the standard format is met.

If users add new YAML-files, they should note that the automatic enrichment routine `enrich_seaVoX` for the elements *sea-SeaVoX* and *ocean-SeaVoX* is very time intensive. Therefore, a mapping for all coordinates that are currently part of the database is provided. If users add a new sea ice core with new coordinates to the extendable database, this mapping has to be extended by running first the `check_for_new_coordinates`, second the `map_new_coordinates`, and third the `extend_mappings` functions. The functions are all part of the `automatic_enrichment` module.

Pyresice furthermore contains a `check` module, which contains the function `overview_10` to create a matrix plot for all rows and columns of the database. The entries of the matrix are either zero or one, depending on whether the entry is empty or not. If one of the columns is zero for all cores, there may be a spelling mistake in the code. The `check` module also includes the function *plot_ temp_ salinity_ combined*, which creates an interactive plot that allows users to scroll through the salinity and temperature profiles of all sea ice cores.

The `export` module creates two different versions of the RESICE tabular database. One is the full database (`RESICE_PartA.csv` and `RESICE_PartB.csv`) and the other one is the reduced version without profile measurements (`RESICE_PartA_cores.csv` and `RESICE_PartB_cores.csv`). Additionally, the `export` module creates a version of RESICE for the integration in MOSAiC webODV. Lastly, the `data2source` module provides the functions used to generate the interactive data2source traceability plot in Figure 5.2.

Chapter 6

Reusability and FAIRness of cryospheric research products in light of the case studies

The reuse scenarios presented in the Cryospheric Case Studies in Chapters 4 and 5 showcase various considerations that need to be taken into account when preparing existing digital research products for reuse. The scenarios emphasize that combinatorial and interpretative skills are necessary for reuse, posing challenges to the automated reuse suggested by the FAIR principles. Furthermore, the scenarios demonstrate that the content, structure, and description of digital resources often differ from the needs of reusers.

In the following, I present three areas for discussion by combining insights on concepts and infrastructures for FAIR and reusable research products in Chapters 2 and 3 with the lessons learned in the case studies in Chapters 4 and 5.

1. **Reuse scenarios for an improved reuse experience:** The case studies showed the potential discrepancy between requirements of reuse scenarios and the state of existing research products and repositories. The communication and documentation of well-defined reuse scenarios could guide resource creators and managers to anticipate the reuse perspective in products and in search and filter options of repositories.

2. **Manual reuse perspective in automated FAIRness assessment:** FAIRness assessment tools provide a means to evaluate a resource's FAIRness by operating on its machine-readable metadata (see Sect. 2.2.2). Yet, the results from FAIRness assessment tools often inaccurately reflect a resource's readiness for reuse as they fail to detect manual reuse issues.

3. **Implications of manual reuse issues for automated reuse:** The goal of the FAIR principles is to enable autonomous, machine-based reuse of research products. In the future, the challenges encountered by manual reuse will have to be resolved by machines alone.

6.1 Formalized reuse scenarios for an improved reuse experience

In this thesis, I introduce the concept of reuse scenarios to reflect the reuse perspective. Reuse scenarios describe the context and purpose, which motivate the reuse of existing research products. Reuse scenarios are based on self-contained research questions and come with specified requirements for the research products. That means if a specific research product can be reused depends on its accordance with the demands of the reuse scenario, which is exemplified with Reuse Scenario B1 from Cryospheric Case Study II (see Sect. 5.2.1).

In Reuse Scenario B1 of Case Study II (see Sect. 5.2.1), Ada wants to perform a high-fidelity model validation of a sea ice evolution model. Therefore, she wants to include uncertainties of the measurement data in form measurement errors, instrument accuracies, and standard deviations. However, the measurement data of sea ice temperature and salinity majorly not includes such uncertainty information. Consequently, Ada can perform the model validation only by disregarding uncertainties. The data is therefore not reusable for Ada's specific reuse scenario. As a consequence of insufficient documentation, reuse scenarios may need to be adjusted to match with existing research products, or scenarios may be deprecated.

Also for research software, a fast check for suitability with a reuse scenario is often hindered by insufficient documentation. Software documentation is often lacking criteria for convergence, accuracy, and resolution. Research software then has to be manually tested and investigated in depth such as highlighted by Brondex et al. (2023b) when carrying out a comparative analysis of the models Eulerian–Lagrangian snow solver (Simson and Kowalski, 2021a) and IvoriFEM (Brondex et al., 2023c) as described in Cryospheric Case Study I (see Sect. 4.4.2).

To circumvent troublesome or even failed reuse, the content, structure, and description of research products should reflect reuse requirements in the best possible way. All relevant contextual information should be documented. The observed insufficient representation of the reuser perspective in products may also be related to creators' lack of awareness of their products' reuse options since reuse scenarios are not communicated efficiently. Failed and successful reuse scenarios are not systematically documented as such in the literature. A potential documentation format could follow the structure of the reuse scopes suggested in Cryospheric Case Study II (see Sect. 5.2.1).

The reuse scope summarizes the reuse requirements in a formalized, tabular structure and defines, in case of Reuse Scenario B1, the scenario's *inputs*, such as temperature, salinity, and snow thickness, as well as the input's *constraints*, such as allowed uncertainties, units, and value ranges. While Cryospheric Case Study II explicitly refers to data sets, similar reuse scopes could also be defined for modeling software. In a first step, reuse scopes could be used to increase suitability of research products to reuse scenarios. Additionally they bear potential to improve the reuse experience in other aspects.

In a next step, the search and selection process of suitable research products could be improved based on a collection of reuse scopes. This process currently is not trivial and time-intensive task (see Sect. 5.2.2 and 5.2.2), and it could be accelerated by systematically addressing reuse

scopes in metadata standards of research products. This metadata should then be searchable in repositories on the *input* level through search bars and, if applicable, selectable on the *constraint* level in the search filters. To date, often only parts of the provided metadata is searchable and filters do not appropriately reflect reuse requirements. PANGAEA, for instance, allows to filter products from specific authors but not to constrain temperature data to be in a specific value range. Query options reflecting the reuser perspective would constrain search results as much as possible and ensure that the results are appropriate to the requirements of a reuse scenario. Specifically, search and filter options for measurements data would include the variables of interest, value ranges, temporal and spatial resolutions, uncertainties, instruments, measurement methods, and the formats of the resource. For the modeling software of a physics-based process model, it would include the process equations included in the computational model, the numerical method, initial and boundary conditions, temporal and spatial discretizations, and the software used.

In the future, reuse scopes that are digitally represented in a machine-readable way could also serve as a communication tool for reuse requirements and for assigning search tasks to machines.

6.2 Representation of the manual reuse perspective in automated FAIRness assessment

FAIRness assessment should guide the implementation of FAIR-complying digital resources and thus ensure their reusability as explained in Sect. 2.2.2. FAIRness assessment tools should be understandable and executable by anyone, and the results should be accompanied by recommendations for an improvement of the resource's FAIRness (Wilkinson et al., 2018; Wilkinson et al., 2019). In the following, two questions regarding the automated FAIRness assessment of cryospheric research products with respect to the manual reuse experience are investigated in depth: 1) How do FAIRness results compare with manual reuse experience? 2) Do recommendations from FAIRness assessment tools support the improvement of FAIRness and reusability of digital resources? I present results from the FAIRness assessment tools F-UJI, FAIR-Checker, and howfairis to discuss these questions. F-UJI is specifically designed for data sets, howfairis is for software, and FAIR-Checker is for any kind of digital object. The details of the tools are explained and discussed in Sect. 2.2.2.

6.2.1 Comparing FAIRness scores with manual reuse experience

In Cryospheric Case Study I and II, I report on several reuse issues encountered when manually reusing cryospheric modeling software and data. In this section, I analyze whether the FAIRness scores obtained with automated FAIRness assessment tools coincide with manual reuse experiences. By comparing both FAIRness scores and manual reuse issues, I aim to identify whether FAIRness scores can be used as indicators for low or high manual reusability.

TABLE 6.1: Results from FAIRness assessment for modeling software using the howfairis and FAIR-Checker tools. Calonne et al. (2014) does not provide code, and Hansen and Foslien (2015b) provide the code as supplementary material to the article (Hansen and Foslien, 2015a). For the Eulerian–Lagrangian snow solver (Simson and Kowalski, 2021b) and IvoriFEM (Brondex et al., 2023a) the link to the GitHub repositories was used for both assessments.

Software	**Repository**	**howfairis**					**FAIR-Checker**
		(1/5) repository	(2/5) license	(3/5) registry	(4/5) citation	(5/5) checklist	
Calonne et al. (2014)	None	-	-	-	-	-	-
Hansen and Foslien (2015b)	Supplement	-	-	-	-	-	-
Simson and Kowalski (2021a)	GitHub	✓	✓	-	✓	-	37.5%
Brondex et al. (2023c)	GitHub	✓	-	-	-	-	37.5%

Modeling software from Cryospheric Case Study I

Table 6.1 lists the modeling software discussed in Cryospheric Case Study I and their form of publication. Furthermore, it includes the results of the FAIRness assessment obtained with the tools howfairis and FAIR-Checker. The software used by Calonne et al. (2014) has not been published as they used the proprietary software COMSOL Multiphysics. However, Calonne et al. (2014) describe the numerical details as well as the boundary conditions and system of Partial Differential Equations with great detail so the results may be reproducible. The software used for the model by Hansen and Foslien (2015a) is provided as supplementary material of the article (Hansen and Foslien, 2015a) and is hosted as zip-file by the journal The Cryosphere from where it can be downloaded. The FAIRness assessment tool howfairis is a command line tool that can only be used with links to public GitHub or GitLab repositories. Therefore, FAIRness assessment can only be carried out for the modeling software Eulerian–Lagrangian snow solver (Simson and Kowalski, 2021b) and IvoriFEM (Brondex et al., 2023c), which are both available on GitHub. Table 6.1 illustrates nicely the increased use of GitHub for the development for cryospheric modeling software.

The five compliance tests of howfairis are introduced in Sect. 2.2.3. The results of howfairis in Table 6.1 show that while both modeling software pass the test for a publicly accessible repository with version control, only the software of of Simson and Kowalski (2021b) passes the tests for license and citation, and none of the software repositories pass the checks for registry and checklist. The GitHub repository of (Brondex et al., 2023c) was manually cross-checked for citation and license, but no citation and license was found. However, the article Brondex et al. (2023b) cites a Zenodo entry containing a version of the code provided as zip-file. This entry assigns the code a CC-BY license. The DOI provided by Zenodo, as well as the license are then however neither specified nor linked in the GitHub repository. The readme of the IvoriFEM GitHub repository mentions that a license will be added, but the last change to the repository was in 2023. Modeling software is not legally reusable without a permissive license allowing its reuse. Regarding the modeling software provided by Hansen and Foslien (2015b), it is not clear whether the license is the same as the one assigned to the article.

With a score of 37.5%, the result of the FAIR-Checker is the same for both GitHub code repositories although Simson et al. (2021), in contrast to Brondex et al. (2023c), provide a license and a DOI.

TABLE 6.2: FAIRness scores obtained with the automated FAIRness assessment tools F-UJI and FAIR-Checker for several data sets from Cryospheric Case Study II, and a selection of issues encountered when manually reusing these data sets (see Sect. 5.2.2 and 5.2.3). The following reuse issues are listed: label names unclear (L), redundancy across repositories (R), duplication within data set (D), metadata and/or data is incorrect and/or inconsistent (I), misleading semblance of accuracy (A), reference depth unclear (RD), tabular format inconsistent (T).

Repository	**Data set**	**FAIRness score**		**Reuse issues (Sect. 5.2)**						
		F-UJI	FAIR-Checker	L	R	D	I	A	RD	T
Zenodo	Audh et al. (2022)	79%	79.17%							
	Omatuku Ngongo et al. (2022)	79%	79.17%			×				
	Wang et al. (2020b)	79%	91.67%						×	×
AADC	Duprat (2019)	37%	83.33%	×					×	
	Meiners (2019)	37%	83.33%	×			×		×	×
PANGAEA	Arndt et al. (2021b)	91%	83.33%							
	Kramer et al. (2010b)	91%	79.19%		×				×	
	Katlein et al. (2020a)	91%	79.19%			×			×	
	Lannuzel (2016e)	91%	91.67%		×		×	×	×	
	Torstensson et al. (2018b)	91%	91.67%	×		×		×	×	

Apparently, FAIR-Checker does neither find a license nor a DOI in the GitHub repository of the Eulerian–Lagrangian snow solver (Simson and Kowalski, 2021b). FAIR-Checker cannot resolve differences between GitHub Repositories and does not seem to be applicable to GitHub yet.

Both FAIRness assessment tools do not explicitly check resource characteristics relevant to manual reusability such as the detailed technical documentation of the code in the readme-file, high quality modular code and the availability of demonstrating examples. Brondex et al. (2023c) and Simson and Kowalski (2021a) both provide detailed documentation of the modeling software in the readme-files, modular codes and demonstrating examples. These features improve the manual reusability of the code. However, none of the FAIRness evaluation tools has an explicit check for these properties.

FAIRness assessment tools provide different results for the same assets and do not explicitly check characteristics important for manual software reusability. Accordingly reuse issues such as inconsistencies between the technical documentation and the software implementation or insufficient documentation cannot be resolved. In general, howfairis searches for specific files that contain license and citation or specific badges added to the readme-file, while FAIR-Checker uses the metadata provided by the landing page of the GitHub repository. At the current stage, both tools cannot be used to obtain an estimate of the manual reusability of modeling software publicly available via GitHub.

Data sets from Cryospheric Case Study II

Ten data sets used in Cryospheric Case Study II were randomly selected and evaluated with the FAIRness assessment tools F-UJI and FAIR-Checker. Table 6.2 presents the results. F-UJI returns the same FAIRness score for all data sets from a given repository: 37% for AADC, 79% for Zenodo, and 91% for PANGAEA. In contrast, FAIR-Checker's results vary between data sets from the same repository. However, out of the data sets tested, F-UJI and FAIR-Checker return only three

different scores each. FAIRness scores of both tools are partly similar for the data sets from Audh et al. (2022), Omatuku Ngongo et al. (2022), Lannuzel (2016e), and Torstensson et al. (2018b), but the scores also show strong variations for the same data sets such as for Duprat (2019) and Meiners (2019) for which the result is either 37% with F-UJI or 83.33% with FAIR-Checker. In general, PANGAEA data sets obtain the highest F-UJI scores.

Furthermore, Table 6.2 indicates the issues encountered during manual reuse of the data sets in Cryospheric Case Study II. Reuse issues are extracted from the *Element availability matching* in Sect. 5.2.2 and the *Plausibility and sanity checks* in Sect. 5.2.3. The reuse issues are: label names unclear, redundancy across repositories, duplication within data sets, metadata and/or data is incorrect and/or inconsistent, misleading semblance of accuracy, reference depth undefined, and tabular format inconsistent. Each of theses issues reduced the ease of manually reusing the individual data set. Table 6.2 shows that all except two of the data sets listed are affected by one or more reuse issues. Meiners (2019), Lannuzel (2016e) and Torstensson et al. (2018b) each have four reuse issues. Omatuku Ngongo et al. (2022) has one and Audh et al. (2022) and Arndt et al. (2021b) have none. The most frequent reuse issue is an unclear definition of reference depth of the sea ice core measurements, which is the case for 7 out of 10 data sets. The other issues each have a frequency of 2 or 3.

Experienced manual reusability does not coincide with the FAIRness scores so that data sets with no reuse issues may have a lower score than data sets with reuse issues and vice versa. For example, Lannuzel (2016e) has a higher score than Arndt et al. (2021b), although manual reusability is perceived to be significantly higher for Arndt et al. (2021b). I experienced the least reusability issues with the data sets form Arndt et al. (2021b) and Audh et al. (2022) and the most issues with data set from Lannuzel (2016e) and Meiners (2019). Against intuition, Lannuzel (2016e) is assigned one of the highest FAIRness scores achieved for both F-UJI and FAIR-Checker. Consequently, automated FAIRness assessment tools are not yet ready to assess the manual reusability of the cryospheric data sets that are of interest for Cryospheric Case Study II.

6.2.2 Recommendations from FAIRness assessment tools for reusability improvement

As stated in Sect. 2.2.2, automated FAIRness assessment should be understandable for everyone, and provide the user with recommendations for improved FAIRness of the tested resource (Wilkinson et al., 2018). Therefore, one might assume that by checking a resource's FAIRness with an automated assessment tool and following such recommendations, it would be easy to increase the resource's FAIR level and thus its reusability. In the following, the FAIRness assessment results of F-UJI and FAIR-Checker for two exemplary data sets from Cryospheric Case Study II are discussed on this aspect. General properties of the tools F-UJI and FAIR-Checker are provided in Table 2.5.

F-UJI

The sea ice core data set by Wang et al. (2020b) hosted on Zenodo is checked with the F-UJI tool. F-UJI uses the FAIRsFAIR metrics defined in Table 2.4 (a). The results obtained for the metric FsF-R1-01MD in form of the metric score and its FAIRness level together with the results of the individual practical tests and a list of debug messages are presented in Table 6.3.

The F-UJI metric FsF-R1-01MD is interpreted from the FAIR principle R1 and considers the consistency between the metadata specifications and the actual content of the data it describes. Compliance with this metric is checked by means of 4 practical tests (FsF-R1-01MD-1 - FsF-R1-01MD-4) listed and described in Table 6.3. A check mark indicates that a test is passed. FsF-R1-01MD-1 tests if the minimal information of the data is provided in the metadata, such as the specification of the resource type. In addition, FsF-R1.3-01M-2 checks whether specifications of properties of the data file are available in the metadata, such as the file format, the size of the file, and the names of the measured or observed variables stored in the data file. The practical tests FsF-R1.3-01M-3 and FsF-R1.3-01M-4 then verify whether the information provided in the metadata matches the actual data file and its contents.

The data set by Wang et al. (2020b) passes only one of the four practical for the metric FsF-R1-01MD, and it is assigned an *initial* FAIR level. While the minimal information of the data is retrievable from metadata (FsF-R1-01MD-1), the metadata lack information on the size and types of the data files and the definitions of the measured variables (FsF-R1-01MD-2) so that further tests for checking the consistency of the metadata with the actual data file cannot be performed.

The result of the FAIRness evaluation should transparently outline the reasons for failed and passed tests, i.e., the specific property of a record that caused it to fail or pass a test. The test descriptions provided in Table 6.3 are comprehensive enough to understand the general characteristics of the metadata of the data set that gave rise to passing a test. The descriptions do not contain specialized vocabulary, so they should be understandable to people who have not been specifically trained in FAIR data management.

The FAIR level and the FAIRness score for this metric are low for the tested data set from Wang et al. (2020b). Naturally, the testing person would be interested in modifying the metadata and data to pass more tests and thus reach a higher FAIR level. Yet, the description of the tests and the indication of a pass or fail does not provide the necessary instructions.

As a next step, an inspection of the debug messages can be considered for more helpful and guiding instructions. The debug messages are of types *INFO*, *WARNING* and *SUCCESS*, and they provide a more technical description of the test application and result. After reading the debug messages, one may understand that the tested metadata does not specify the file size and measured variables. However, a large part of the messages can also be confusing due to the use of technical terminology such as error specifications like *HTTPError code 404*, status codes like *200*, and metadata standards like schema.org, DCMI and DataCite (see Sect. 3.2.2). In addition, there is no explicit instruction describing how to adjust the data set and its metadata to pass the tests.

TABLE 6.3: FAIRness assessment results including debug messages obtained for the data sets by Wang et al. (2020b) hosted on Zenodo. The table shows the results for the FAIRsFAIR metric FsF_R1-01MD.

Test Result	**Description**
FsF-R1-01MD	**Metadata specifies the content of the data**
Score: 1/4	
FAIR level: 1/3	
FsF-R1-01MD-1	Minimal information about available data content is given in metadata
✓	
a	Resource type (e.g. data set) is given in metadata
✓	
b	Information about data content (e.g. links) is given in metadata
✓	
FsF-R1-01MD-2	Verifiable data descriptors (file info, measured variables or observation types) are specified in metadata
-	
a	File size and type information are specified in metadata
-	
b	Measured variables or observation types are specified in metadata
-	
FsF-R1-01MD-3	Data content matches file type and size specified in metadata
-	
FsF-R1-01MD-4	Data content matches measured variables or observation types specified in metadata
-	
Debug messages:	
Level	**Message**
INFO	Trying to complete local file name with full path info using landing page URI
WARNING	Content identifier inaccessible -: https://zenodo.org/records/file://, HTTPError code 404
INFO	Successfully parsed data object file using TIKA
INFO	File request status code -: 200
INFO	Successfully parsed data file(s) -: https://zenodo.org/api/records/3779867/files/readme.txt/content
INFO	Object landing page accessible status -: True
SUCCESS	Valid resource type (e.g., subtype of schema.org/CreativeWork, DCMI Type, or DataCite resourceType) specified -: data set
SUCCESS	Valid resource type (e.g., subtype of schema.org/CreativeWork, DCMI Type, or DataCite resourceType) specified -: webpage
WARNING	NO info about file size available in given metadata -:
WARNING	NO measured variables found in metadata, skip 'measured_variable' test.

As the data set by Wang et al. (2020b) is published on Zenodo, the FAIRness assessment is carried out based on the machine-readable metadata generated from the contextual information provided in the predefined submission mask. Consequently, data set authors are constrained to the metadata fields predefined by Zenodo and listed in Table 3.1. The recommendations provided with the FAIRness assessment do not explain how to modify and extend these metadata fields. Therefore, it is not trivial to improve the test results for research products hosted by Zenodo. The same applies to other repositories that generate metadata from predefined submission masks.

Lastly, I manually cross-checked the metadata file of the Wang et al. (2020b) data set for the

TABLE 6.4: FAIRness assessment results with FAIR-Checker of metric R1.3 for the Zenodo data set Omatuku Ngongo et al. (2022) including application log-type comments and a guideline for FAIRness improvement

Description	
R1.3	**Community standards**
Weak	FAIR-Checker verifies that at least one used ontology class or property are known in major ontology registries (OLS, BioPortal, LOV)
Strong	FAIR-Checker verifies that all used ontology class or properties are known in major ontology registries (OLS, BioPortal, LOV)
Score:	**Score: 1/2**
Comments:	
INFO - Evaluating metrics Community standards	
INFO - Strong evaluation:	
INFO - Checking if all classes used in RDF are known in OLS, LOV, or BioPortal	
INFO - All classes found in those ontology registries	
WARNING - http://schema.org/scheme property not known in OLS, LOV, or BioPortal	
INFO - Weak evaluation:	
INFO - Checking if at least one class used in RDF is known in OLS, LOV, or BioPortal	
INFO - http://schema.org/DataDownload known in Linked Open Vocabularies (LOV)	
Guideline:	
You should express all your metadata with properties coming from interoperable ontologies and vocabularies: use Ontology Lookup Service, BioPortal or Linked Open Vocabularies to find the most suitable classes you want to use.	

unavailability of file size and type resulting from metric FsF-R1-01MD-2. In contrast to the result of F-UJI, the metadata file in JSON-LD format specifies the sizes of all files stored in the data set in bytes. It seems that F-UJI is not able to process this information, although the tool states to use metadata in JSON-LD format (Devaraju and Huber, 2024).

FAIR-Checker

The sea ice core data set from Omatuku Ngongo et al. (2022) hosted by Zenodo is checked with the FAIR-Checker. FAIR-Checker uses its own metrics, which are only specified in the tool and refer directly to the FAIR principles. The metrics are summarized in Table 2.4 (b). The results obtained for metric R1.3 in the form of the metric score, a description of the practical tests, a list of application log type comments, and a recommendation are presented in Table 6.4.

Metric R1.3 is derived from the FAIR principle R1.3 and takes into account the use of community standards in resources. FAIR-Checker uses two practical tests to check the compliance of the resource with the metric. Both practical tests check whether the metadata of the resource can be found in registries of ontologies, namely OLS, BioPortal and LOV (see Sect. 3.2.3). The first test checks whether at least one property or class can be found in such ontologies and is referred to as the *weak* formulation. The second, which is referred to as *strong* formulation, checks whether all properties and classes in the metadata are available in one of the ontology registries OLS, BioPortal or LOV. Omatuku Ngongo et al. (2022) only passes the weak test, so the score for this metric is 1/2.

In order to understand the assessment process, the comments accompanying the results can be studied. These comments show that FAIR-Checker first checks if the classes used in RDF (see Sect. 3.2.2) can be found in one of the named ontology registries, which is passed. Then it issues a warning stating that the property "http://schema.org/scheme", is not found in the ontology registries, so the strong test fails and only the weak test passes. Additionally, the result of this FAIRness assessment includes the recommendation that "all your metadata with properties from interoperable ontologies and vocabularies: use Ontology Lookup Service, BioPortal or Linked Open Vocabularies to find the most appropriate classes you want to use."

The comments and recommendations provided by FAIR-Checker are not instructive enough to improve the FAIRness. Based on the given recommendation, the data set's creator could look for terms in the ontology registries. Yet, it still remains unclear how to integrate such terms and to modify data and metadata so the data sets passes the strong formulation of the test. To understand the comments of the test, a person needs to know RDF and the difference between classes and properties of RDF. Furthermore, they need to know how to find and assign properties and classes in and from ontologies, and how to implement them correctly in metadata. Again, from the perspective of a data collector, who publishes data sets via Zenodo and creates metadata via the submission mask, implementing these recommendations is nearly impossible and may cause confusion.

6.3 Implications of manual reuse issues for automated, machine-based reuse

The case studies demonstrated the need to interpret, adjust and combine research products to make them ready for reuse. So far, tasks like downloading, installing, and adapting modeling software for application in a reuse scenario is manually performed. The same applies to the combination and harmonization and majorly also to the consistency checks of data sets. Several manual reuse issues had to be resolved in the case studies, as noted for the examples of cryospheric modeling software in Sect. 4.3 and cryospheric data sets in Sect. 5.2.2 and in Sect. 5.2.3. Future machine-based reuse needs to autonomously perform these interpretations and plausibility checks. In the following, I summarize the challenges faced in manual reuse and their implications for machine reuse for six key aspects: discoverability, machine readability, combinability, consistency and correctness, terminology standardization, and quality.

6.3.1 Discoverability

There are many repositories for sharing research products. It is challenging to navigate them to find a particular product suitable for a reuse scenario. Reusers have to find repositories that host the research products they are interested and familiarize themselves with the varying search and filter options. In order to efficiently find a research product, it has to be equipped with searchable metadata that specifically describes it.

Deciding whether a modeling software is suitable for a reuse scenario only based on minimal metadata like its title and author, can be challenging. However, even CodeMeta metadata files, which were developed to describe software specifically, do not provide much more than this for IvoriFEM and the Eulerian-Lagrangian snow solver on Zenodo (Simson and Kowalski, 2021b; Brondex et al., 2023a). Without knowing the articles describing the respective software, it is difficult to locate relevant resources, and without descriptive metadata, it is challenging to determine the suitability of a modeling software for a particular reuse scenario. To improve findability for humans and machines, software should have comprehensive, searchable metadata to enable easier discovery for specific purposes. Additionally, more extensive search and filter options can facilitate the identification of relevant resources in repositories.

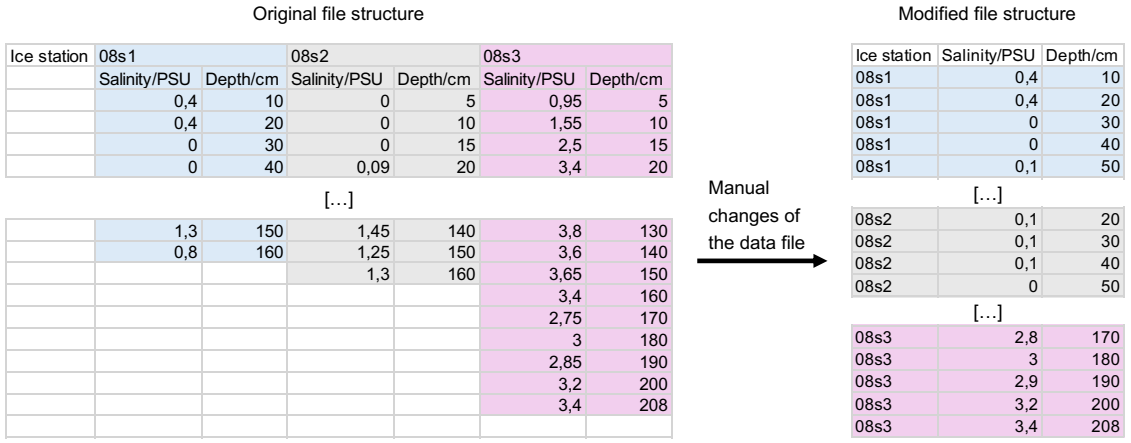
The search for suitable data sets for Cryospheric Case Study II was an iterative process, repeatedly adapting to different combinations of desired parameters. As described in Sect. 5.2.2, some data sets contained only two parameters of interest for one sea ice core. To find more parameters of the same core, I had to search for varying combination of parameters. Furthermore, the search had to be adapted to the search options of the respective repository, and it also extended to other resources such as articles and manuals describing the product. In addition to repositories, I searched for resources via the Google search engine. Such search queries usually considered a specific author in combination with a campaign name or the name of a measurement instrument. Different combination were tested. Machines would have to follow similar iterative approaches when searching for suitable resources. Similar as for humans, it may be especially challenging to find relevant resources related to a software or data set if they are not explicitly linked to it, as observed in Cryospheric Case Study II (see Sect. 5.2.2).

6.3.2 Machine readability

For research products to be used by machines, they must be readable by machines. The steps to run the modeling software presented in Cryospheric Case Study I, namely IvoriFEM (Brondex et al., 2023c) and the Eulerian-Lagrangian snow solver (Simson and Kowalski, 2021b), are described in the readme-files in the respective GitHub repositories. The readme-files describe the steps to install and run the software in a human understandable way. These files do not provide sufficient machine-readable instructions for the modeling software to tun automatically.

In Cryospheric Case Study II, two tabular data files from Zenodo (Wang et al., 2020b) and AADC (Meiners, 2019) were not sufficiently prepared to be machine-readable. They had to be manually modified. An example of the necessary changes is illustrated in Fig. 6.1. Humans can understand such unsystematic structures of data files and combine them according to their needs using their experience and context. For Wang et al. (2020b), the modifications were rather easy, but for Meiners (2019), who provided an even less ordered tabular spreadsheet, it was not without ambiguity.

If machines are to interact with modeling software and data files such as those used in the case studies, they need standardized, machine-understandable instructions on how to use the software.



Ice station	08s1		08s2		08s3	
	Salinity/PSU	Depth/cm	Salinity/PSU	Depth/cm	Salinity/PSU	Depth/cm
	0,4	10	0	5	0,95	5
	0,4	20	0	10	1,55	10
	0	30	0	15	2,5	15
	0	40	0,09	20	3,4	20
			[...]			
	1,3	150	1,45	140	3,8	130
	0,8	160	1,25	150	3,6	140
			1,3	160	3,65	150
					3,4	160
					2,75	170
					3	180
					2,85	190
					3,2	200
					3,4	208

Ice station	Salinity/PSU	Depth/cm
08s1	0,4	10
08s1	0,4	20
08s1	0	30
08s1	0	40
08s1	0,1	50
	[...]	
08s2	0,1	20
08s2	0,1	30
08s2	0,1	40
08s2	0	50
	[...]	
08s3	2,8	170
08s3	3	180
08s3	2,9	190
08s3	3,2	200
08s3	3,4	208

FIGURE 6.1: Manually performed modifications of the tabular structure of the data set from Wang et al. (2020b) to allow machine-readability

They also need the means to correctly process unsystematically structured data files so that they can read and use them.

6.3.3 Combinability

The integration of existing research products into reuse scenarios often requires the combination of products with each other and their adjustment to the specific purpose of the scenario. I refer to combinability as the ability of products like data sets or modeling software to be A) merged or coupled into, for instance, a unified database or model, and B) modified or pre-processed to the constraints of a reuse scenario.

Setting up a new database for a data-driven modeling task may involve combining different data sets. For Reuse Scenario B2 from Cryospheric Case Study II (see Sect. 5.2.1), Master student Liza requires data from sea ice cores for a classification task including temperature, salinity and thickness of the sea ice as well as its development stage. As exemplified in Cryospheric Case Study II, the necessary data is usually not available in a comprehensive way from one singular database. Instead, temperature and salinity data is distributed across data sets, and the development stage and thickness of sea ice may be solely provided in text-based form in articles and not with the data sets. The combination process for data compilations is illustrated for RESICE in Cryospheric Case Study II, where it is referred to as *Element availability matching* (see Sect. 5.2.2). First, several data sets and other resources must be combined for each sea ice core, for which a YAML file is used. Then, these YAML files must be combined to form the tabular compilation.

Such data combinations processes are challenging. They are time-intensive because they typically involve intensive engagement of the reuser. Furthermore, data sets are heterogeneous, such that several harmonization steps such as for units, labels, and reference depths have to be performed. The combinatorial task suffers from the lack of interoperability of data sets and their metadata hindering a seamless combination of data sets with the same properties. Another challenge is the potential redundancy of data sets from different repositories. Machines would have to merge and

harmonize relevant metadata and data in a reasonable form, and they have to detect redundancy even if the respective data sets were not linked to each other. In case of redundancy, they would have to decide on the data set to select as well as possibly merge metadata that is only available from one of the data sets.

Combinatorial tasks are also required for software. Examples include combining different software modules for model coupling, and modifying a model's initial and boundary conditions to suit the measurement data of interest in a given scenario. In Reuse Scenario A1 (see Sect. 4.4.1), Sam wants to couple a process module for water percolation to the Eulerian–Lagrangian snow solver. Therefore, he needs to define an interface to combine the two models and adjust them accordingly. In Reuse Scenario A3 (see Sect. 4.4.1), Zoe combines measurement data with simulations from the Eulerian–Lagrangian snow solver. In this case, the combinatorial task involves the adjustment of the model geometry as well as boundary and initial conditions of the model to mimic the environment conditions of the measurements. Similarly as for data sets, these combinatorial tasks require engagement of the reuser with the modeling software. In order to operate independently, a machine would require clear definitions of where to adjust the initial boundary conditions and model geometry, and where to couple an extension to the model.

6.3.4 Consistency and correctness

The case studies described inconsistencies and errors detected in research products that should be addressed before integrating them in reuse scenarios. Cryospheric Case Study I describes inconsistencies of the modeling software Eulerian–Lagrangian snow solver (Simson and Kowalski, 2021b) and its technical documentation in the article (Simson et al., 2021). These inconsistencies were detected based on plausibility checks carried out by Brondex et al. (2023b) (see Sect. 4.4.2), and they required changes of the code before performing the reuse scenario.

In Cryospheric Case Study II, several consistency checks had to be performed on the data sets, their metadata, and other related resources. These consistency checks detected incorrect metadata, such as the data sets from Lannuzel (2016a), Lannuzel (2016b), and Lannuzel (2016e), which report incorrect sea ice thicknesses in the metadata that do not match with the actual data. Other examples are the data set from Mundy et al. (2010), which reports measurement depths with the decimal point in the wrong position, the data set from Torstensson et al. (2018b) that has an incorrect label name and reports ice temperatures as water temperatures. Several other data sets such as Lange et al. (2015b) and Torstensson et al. (2018b) provide measurements with a higher accuracy than the instrument used can guarantee.

All of the consistency checks described were performed manually. When researchers want to build their research on existing resources, they naturally check their correctness and consistency. In the case of future high-quality machine reuse, such consistency checks would need to be performed in an automated manner. At the very least, machines should be able to report potential inconsistencies in software, data, and their metadata, even if they cannot resolve them. So far, FAIRness assessment tools do not include such checks because they focus mainly on metadata and not on the data files

themselves. Cross-checking the consistency of data and metadata will be particularly difficult when, in addition to data sets, text-based resources such as articles or unlabeled metadata in the form of comments are involved, as illustrated in Cryospheric Case Study II. These consistency issues also highlight the need to focus on high quality code and data. Rather than discovering these problems after the software and data have been published, automated software and data checks could help to identify inconsistencies before they occur.

6.3.5 Terminology standardization

Without explicitly and unambiguously defining the meaning of a variable or parameter in data sets or modeling software, reuse issues arise for both humans and machines. Such issues can be minimized by using and referencing defined terminologies such as the semantic artifacts described in Sect. 3.2.3. The modeling software discussed in Cryospheric Case Study I does not use standardized terminology to describe and define their cryospheric variables. In Cryospheric Case Study II, reuse issues arise from varying label names for the same properties among data sets, and the same applies to class names such as those used to describe the development stage of sea ice. Different data sets use different terms without specifying the respective terminology standards and classification schemes, or they define their own classification such as Arndt et al. (2021a).

Reference to discipline-specific terminology standards is often not given in research products. This relates to the literature study, which showed that cryospheric terms are not sufficiently and detailed enough represented in geoscience related terminology standards. Besides the unavailability of standards, the insufficient use of standardized terms in the descriptions of data sets and software may also be related to the metadata options of the repositories. Zenodo submission mask does not by default provide reference to disciplines-specific naming standards. AADC uses geoscience related keywords in their metadata files that are listed in the GCMD (GCMD, 2024). These names are then however not always consistently used in the data files. PANGAEA links parameter names with terms registered in ontologies to improve the standardization of the terms. This could in principle support both humans and machines to understand the data, but it seems as it is not yet meaningfully implemented. An example is the parameter *Conductivity* in the data set from Arndt et al. (2021b), which is the electrical conductivity measured in the meltwater of sea ice, and it is assigned the definition "a physical property inherent in a carrier by virtue of the carrier's disposition to transmit an entity through a medium" (PTO, n.d.) in the Phenotype and Trait Ontology. Lannuzel (2016e) provides another example for the parameter *Temperature ice/snow*. The term *temperature* is linked to "temperature (thermodynamic temperature)" on dbPedia (dbPedia, n.d.), and the term *ice* is linked to the Environment Ontology, which defines ice as "Ice formed from water" (ENVO, n.d.). These kinds of links did not improve my understanding of the resource. From my perspective, a definition of the entire term may be better understandable.

The described observations of the insufficient use of standards match with those of Duerr et al. (2024), who claim that cryospheric terminology is not yet sufficiently digitally represented, and it is also in line with Miller et al. (2015), who address the lack of standardization of biogeochemical

investigations of sea ice including sea ice cores. Section 3.2.3 also demonstrated that the terminology standards ENVO and SWEET do not yet detailed enough represent all cryospheric properties. They do for instance not reflect the property *snow water vapor density* or *sea ice temperature*, which would be important to be added in the metadata of the Eulerian–Lagrangian snow solver and sea ice core data sets respectively.

Because there is a lack of terminology standardization in data sets and software, understanding research products is often not possible from metadata alone. In the case studies, interpretation using context, domain knowledge, and common sense was required for meaningful reuse, and yet the context of some resources was still not fully understood. For machines, the context of a resource needs to be explicitly described in the metadata. Otherwise, it will be difficult for machines to decide whether a particular asset is suitable to their task and how to correctly use it.

6.3.6 Quality

The quality of resources is an important indicator for the suitability with a reuse scenario. Quality in the context of modeling software could refer to verification measures such as the discrepancy of a simulation result with respect to an analytical solution or experimental data, or it could refer to the order of accuracy of the numerical method. Similarly, the quality of a data set could be represented as the accuracy of the used measurement device, the standard deviation of a measurement, or its spatiotemporal resolution.

The research products considered in the case studies provide little information on quality. Only the data sets from AADC provide a section in the metadata with comments on the quality. Meiners (2019) states in this section "Sampling and analyses followed standard procedures (e.g. Miller et al. 2015). No problems were encountered.", and Duprat (2019) comments on the quality as "High outlier values of sFe and DFe for S5 was found at station 4 and 5 when compared to the neighboring sections and to all top and intermediate layer (S1-S4) values. We suspect that possible contamination could have occurred during the melting process of the S5 sections on these days." While the provided information is generally of interest for the reuser, it does not provide quality in a quantitative form, which is for instance of interest for Liza in Reuse Scenario B1 (see Sect. 5.2.1). She needs measurement device accuracies and standard deviations of the measurements, and these information are typically not provided in the considered data sets. Instead, accuracy of the measurement device had to be obtained from other sources such as articles and instrument manuals.

If quality information is lacking, humans evaluate a resource on experience and decide whether to trust it or not. In case of uncertainty, machines would have to make similar decisions, which may be challenged when standardized quality information is not available in a structured, machine-friendly way. Certainly, a machine should avoid using low-quality data and models. Lack of quality information challenges the reuse of modeling software and data. The need for standardized quality information for geoscience research products has also been emphasized by Peng et al. (2021b).

Chapter 7

Concluding remarks and outlook

This thesis explores the reusability of research products from various perspectives. Chapter 2 summarizes definitions of data and software reusability from the literature, and it describes reusability as it is defined by the FAIR principles. As these principles aim to enable the autonomous, machine-based reuse of research products, I refer to the FAIR envisioned reusability as *machine reusability* (Sect. 2.2). Related FAIRness assessment tools (see Sect. 2.2.2) evaluate the reusability of a resource based on their inherent, machine-readable properties and metadata, i.e., *independent* of the scenario within it is intended to be reused. In contrast, the case studies in Chapters 4 and 5 demonstrate that reusability also *depends* on the specific reuse scenario in which it is intended to be reused, i.e., a data set may be reusable in one scenario but not in another. Lastly, the case studies provide examples for manually performed reuse of research products. In this case, I refer to the ease of reusing a product as *manual reusability* in contrast to automated *machine reusability*. The discussion in Chapter 6 highlights the role of reuse scenarios, FAIRness assessment results, and machine reusability with respect to manual reusability. In this final chapter of the thesis, I will first present my concluding remarks and then provide an outlook on the future potential of artificial intelligence for reuse, along with three feasible suggestions for improving reusability.

7.1 Concluding remarks

The reuse scenarios of the case studies in Chapters 4 and 5 illustrate the manual reuse of cryospheric research products, and they provide examples for several reuse issues listed in Sect. 6.3. These issues concern the discoverability, machine-readability, combinability, consistency, correctness, terminology standardization, and quality of research products. Existing digital infrastructures such as repositories and standardization efforts are proposed as solutions to improve FAIRness and reusability of research products in Chapter 3. Resource creators should select a suitable repository and implement domain specific standards for data and metadata. While generally improving the products' reusability, the case studies demonstrated that existing digital infrastructures are not yet sufficient to enable seamless, automated reuse. The effectiveness of standards and repositories

depends critically on their general availability and suitability to the respective resources, as well as their consistent and correct implementation in research products by providers and curators.

Case Study II (Chapter 5) exemplified that sea ice core data sets are shared on various generic and domain-specific repositories (see Sect. 3.1.1) and use heterogeneous or no standards for structure, format, and terminology. This observation shows that repositories do not yet support interoperability, such as the seamless integration of individual data sets with another into a larger database by combining data of the same kind. Such databases would unlock the combined potential of the data sets, which is especially beneficial for small data as illustrated in Fig. 3.2. Instead, data sets such as sea ice core data must be compiled manually to increase their potential for reuse. This lack of interoperability can be related to the insufficient availability and adoption of domain-specific standards for format, structure, and terminology of the data and metadata. Interoperability could be improved, for instance, by linking different label names of the same property with a PID that uniquely references its definition in a domain-specific ontology or controlled vocabulary. So that data of the same measurable can be combined even if names are heterogeneously used in data sets. It would be furthermore improved if the structure and format of the data and metadata would be consistent and followed the same standards, for instance, by using the same order of column labels in data files. References and consistency would facilitate finding and combining data of the same type. This would be especially beneficial for smaller data sets, facilitating faster integration with other smaller data sets and increasing their potential for reuse applications.

While the repositories Zenodo, AADC, and PANGAEA use consistent standards for the structure and format of metadata, PANGAEA also has a consistent tabular data structure for the data. Additionally, PANGAEA has begun linking parameter names with PIDs of vocabulary from registered ontologies as described in Sect. 3.2.3. Currently, linking parameters with PIDs is a beta feature of PANGAEA. Based on experience from Case Study II, this feature does not yet make parameter names fully or uniquely understandable or interoperable when reused. Features aimed at the increease of interoperability are hampered by the insufficient standardization and digital representation of community terminology making it difficult to select suitable terminology.

Understandable and detailed documentation in the form of metadata, descriptions, or readme-files should allow the easy understanding and application of resources, and they should contain quality information on the research product and its context. However, documentation of research products do often miss relevant information that are critical for their reuse. In both case studies, for instance, quality information like the measurement error or model accuracy were not documented in the products, i.e., tacit knowledge remains with the resource creator and is not transferred to potential reusers. For data sets, it means that this knowledge is unreachable for any future reuser. In case of software, reusers may derive quality information themselves, for instance, by running a verification or validation. For powerful and efficient filtering, the content of the documentation including quality information of the products should be directly addressable via the search and selection process.

Metadata files, readme-files, and other documentation must be carefully curated, as incomplete or inaccurate details can hinder effective reuse. Correctness and consistency have to be ensured so

that metadata and documentation is accurate and aligns with the respective data set or modeling software it describes. For measurement data, this means careful collection of metadata, as missing information cannot be added later. For modeling software, this could mean making the software development process transparent in the form of public git commit history. Software transparency would benefit from the practice of openness and digital interaction between researchers from the start, and version-controlled repositories like GitHub already bear a lot of potential often not yet fully leveraged for software development. Brondex et al. (2023c), for example, find bugs in the code by Simson and Kowalski (2021b), but they do not use tools of the public GitHub repository to report them.

This may also be due to missing domain-specific solutions for standardized documentation defining and constraining the necessary information to be provided. Research communities should consolidate standards for metadata, description, and structure, to prevent loss of relevant information of research products. Furthermore, standardized terminology like controlled vocabularies and ontologies need to be further developed so they provide terms at the level of detail required for the description of research products and their parameters.

Like reusability, FAIR developments suffer from insufficient domain representations. The domain-agnostic foundation of the FAIR principles has yet to be tailored into discipline-specific interpretations and metrics. While the recent focus on FAIR has led to an improvement of reusability by suggesting solutions for the findability and accessibility of resources, interoperability and reusability of research products remains partly unsolved. Furthermore, FAIR initiatives and tools are unclear and inaccessible to many researchers. For a significant enhancement of reusability, FAIR has to be adoptable by all researchers, regardless of their experience with technical standards or metadata. The need for an improved understandability of FAIR has also been highlighted by David et al. (2020). FAIR principles and assessment tools should support, rather than discourage, researchers in their efforts to make their research products reusable.

The literature review of FAIR in Sect. 2.2 indicates the challenges of manual reuse are currently not specifically addressed by FAIR developments. Consequently, also automated FAIRness assessment tools cannot capture manual reusability of research products within particular fields as it has been repeatedly shown for cryospheric research products in Sect. 6.2.1. FAIRness scores do usually not align with manual reusability experiences from the case studies. In fact, data sets with high FAIRness scores may exhibit low manual reusability, and vice versa. Furthermore, current automated assessment tools evaluate resources only based on the metadata and thus miss critical reuse issues pronounced on the file level. Although reusability is the ultimate goal of the FAIR principles, automated assessment tools can assign high FAIRness scores to resources that are far from being easily reusable. Strikingly only three out of the 20 FAIRness assessment tools examined by Candela et al. (2024) offer discipline-specific evaluations, and none are dedicated to cryospheric science. These tools collectively include over 1,000 metrics, yet fail to address key aspects of manual reuse challenges. Given these observations it seems as FAIRness assessment is several steps ahead of current community developments, which again highlights the need for the consolidation and convergence of community specific standards. To make FAIRness assessments truly meaningful,

these tools must evolve to reflect specific community standards, enabling evaluations and instructive recommendations tailored to the characteristics of the respective resource.

The discussion of automated FAIRness assessment in Sect. 6.2.2 demonstrates another shortcoming of current FAIR developments regarding the general understandability of the principles and related developments. Other than expected and defined by Wilkinson et al. (2018), automated FAIRness assessment tools F-UJI and FAIR-Checker fall short on the understandability of the evaluation process and at providing feasible suggestions for improving resource FAIRness. While these tools display debug messages and, in some cases, generic recommendations, understanding them often requires technical expertise, and the feedback is not practical for researchers without training in FAIR data management. This observation is a general indication that the consistent enforcement of FAIR remains a challenge and an objective of current research.

7.2 Outlook

As described in the concluding remarks, the goal of seamless and automated reuse of research products envision in the FAIR principles is not solved yet. Manual reuse challenges also affect machine reuse as discussed in Chapter 6, specifically Sect. 6.3. Trustworthy machine-based reuse in a FAIR-way has to overcome shortcomings experienced in manual reuse. Like humans, machines require interpretation and combination skills to perform the reuse of research products. So far, *machines* have not been further specified in this thesis. Recent Artificial Intelligence developments suggest that large language models (LLMs) (Naveed et al., 2025) could automate the reuse process. Due to their interpretive abilities, LLMS have the potential to facilitate reuse and to overcome current reuse issues.

The application of LLMs has the potential to increase the reusability of research products, not only during reuse, but also may also support the steps like the publication of the product. Although only a few publications specifically address the use of LLMs for enhancing reusability (Berenguer et al., 2024), many capabilities of LLMs could be leveraged to improve reuse in the future.

- **Code generation:** LLMs have been trained to automatically write code (Jiang et al., 2025). In the future, such models could be used to program interfaces, for instance, for the coupling of existing modeling software.
- **Terminology harmonization:** Santos et al. (2025) introduce an interactive method to harmonize terminology in data sets based on LLMs. The use of LLMs to harmonize terminology bears would be beneficial for reuse, especially when many data sets need to be combined.
- **Metadata extraction:** LLMs have been used to extract metadata from articles (Yu et al., 2025; Jiang et al., 2025). Similar models could be used to also extract metadata from other text-based sources like read-me files or descriptions. It is also tangible that LLMs could extract model information like temporal and spatial discretization as well as numerical methods from modeling software.

- **Standard development:** LLMs can support the development of standards such as ontologies, controlled vocabularies, metadata models and formats, as well as data structures (Lippolis et al., 2025). The use of such domain-specific standards in research products would then improve their reusability.

While LLMs bear a lot of potential to improve and automate reuse, they have limitations, and some reusability issues will persist. Missing or incorrect documentation cannot be recreated with certainty. Missing quality information, such as measurement accuracy, could be predicted based on training data, but it will still be uncertain. In general, the potential of LLMs depends on the quality of the data used for their training.

A major component of future development with or without the use of LLMs should be the advancement and consolidation of community standards, the convergence of interpretations of FAIR, and the implementation of standards in research products. Community discussions should furthermore solve questions on to be shared resources, for instance, which parts of a simulation are relevant (Simmonds et al., 2022) for sharing, and what quality and context is considered sufficient. However, the development and harmonization of such standards will take time as responsibilities are not clear (Kinkade and Shepherd, 2021) and adoption usually is inert (Tenopir et al., 2020). In addition, standardization will likely never reach a final state because the needs of ever evolving research questions and methodologies are difficult to capture in standardized formats.

LLM-based applications and harmonized standards will support and automate the reuse process. Currently, neither is readily available for operation, and most reuse steps are still carried out manually. Therefore, at the end of this thesis, I recommend three feasible key efforts that prioritize the manual reuse perspective but will also support automated reuse once available. The first two key efforts are general and refer to modeling software and data. The last one is specifically directed to the management of small, heterogeneous, and distributed data.

1. **Improving the quality and transparency of resources:** Full standardization, especially at the domain level, will take time and may never be feasible. At this stage, it is not easy to find suitable standards for a particular research product (Sansone et al., 2019). On the one hand, one has to work through a multitude of standards, which is a tedious task, and on the other hand the implementation of standards requires technical understanding. Therefore, the focus should be on high quality, comprehensive documentation and transparency of digital resources, without mandatory use of specific standards. This requires more transparent and comprehensive documentation of data and software, as well as a general prioritization of data and software quality, and it includes the communication of quality within metadata or directly in the data. The reported practice of uploading modeling software and data sets with insufficient context and poor quality, potentially just to comply with the guidelines of funding agencies, should be avoided. Creators of digital resources should focus on the quality and understandability of the resources, also following their common sense. As the case studies show, these practices can greatly support reuse. Although the Eulerian-Lagrangian snow model does not follow any defined terminology or quality standards for modeling software, it has proven

to be reusable due to its detailed technical documentation and modular software. In the case of data sets, it has been shown that with good documentation and clear definition of terminology, even if self-defined, they are easier to reuse than others. Transparent communication of quality should focus on quantifiability of the quality as highlighted in Sect. 6.3.6. General comments about quality issues are usually not helpful, as reusers lack the knowledge to assess the impact and include such uncertainties. Good documentation today will also facilitate the integration of the resource into future community-tailored standards by machines.

2. **Communicate reuse perspective with reuse scenarios:** The reuse perspective is not yet sufficiently reflected in the creation and curation of digital resources, which may be due to missing awareness for the reuse potential. Therefore, reuse scenarios and their requirements should be represented and communicated more explicitly. A systematic documentation of reuse scenarios in digital form could raise awareness for the variety of reuse possibilities for resources. For instance, one could follow the form of tabular reuse scopes in Sect. 5.2.1. The structured formalization of reuse requirements may enhance the inclusion of specific and relevant details in the metadata of research products improving the experience of reusers. As discussed in Sect. 6.1, formalizing reuse scenarios could also support standardized communication and documentation of reuse requirements. Such standardized documentation could be used as a means to assign reuse tasks to machines in the future so they do exactly what we want them to do. But even before autonomous reuse is realized, formalized reuse requirements could support the development of systematic and effective queries of data repositories that reflect the reuse perspective.

3. **Address potential of smaller data:** Much potential remains unlocked in the combination of rather small data like it is demonstrated for sea ice core data sets in Case Study II (Chapter 5). Repositories like Zenodo, AADC, and PANGAEA do not succeed at making such small data sets easily discoverable and combinable. Therefore, I developed the reusability-targeted approach for data compilations, which is described in Sect. 5.2. While it has been originally developed for sea ice cores, the approach could be similarly followed for the integration of other small, heterogeneous, and distributed data sets, such as snow pit measurements. Small data communities could also exploit the approach on a larger scope to guide efforts toward combined databases. By following the approach, data sets with varying standards could be transparently harmonized and enriched. Using GitHub as a basis, it could guide flexible and transparent collaborative efforts for combined databases as it has been for instance proposed for biogeochemical sea ice data by Miller et al. (2015). Thereby, the approach would enable the traceability of the origin of all used resources and of manual changes to the original data sets in the compilation process. Figures like the data2source traceability plot in Fig. 5.2 could be used to provide an overview on the properties available in the database and provide a means to credit contributing resources in an interactive way. The transparency of such compiled databases is important, as other researchers will only adopt existing resources including their compiled versions if they trust them.

Bibliography

Alnaim, A. and Z. Sun (2022). Using Geoweaver to Make Snow Mapping Workflow FAIR. In: *2022 IEEE 18th International Conference on e-Science (e-Science)*. IEEE. DOI: 10.1109/eScience55777.2022.00062.

Antarctic Sea Ice Processes and Climate **(ASPeCT)** (n.d.). ASPeCT resources. Accessed: 2024-10-23.

Armstrong, T., B. Roberts, and C. Swithinbank (1966). *Illustrated Glossary of Snow and Ice*. 4. Scott Polar Research Institute.

Armstrong McKay, D. I., A. Staal, J. F. Abrams, R. Winkelmann, B. Sakschewski, S. Loriani, I. Fetzer, S. E. Cornell, J. Rockström, and T. M. Lenton (2022). Exceeding 1.5°C global warming could trigger multiple climate tipping points. In: *Science* 377.6611. DOI: 10.1126/science.abn7950.

Arndt, S., C. Haas, H. Meyer, I. Peeken, and T. Krumpen (2021a). Recent observations of superimposed ice and snow ice on sea ice in the northwestern Weddell Sea. In: *The Cryosphere* 15.9, pp. 4165–4178. DOI: 10.5194/tc-15-4165-2021.

Arndt, S., C. Haas, I. Peeken, E. Allhusen, and H. Meyer (2021b). Physical ice core properties from ice stations during Polarstern cruise PS118. PANGAEA [data set]. DOI: 10.1594/PANGAEA.928948.

Audet, D. and A. Fowler (1992). A mathematical model for compaction in sedimentary basins. In: *Geophysical Journal International* 110.3, pp. 577–590. DOI: 10.1111/j.1365-246x.1992.tb02093.x.

Audh, R. R., S. Johnson, M. Hambrock, R. Marquart, J. Pead, T. Rampai, S. Skatulla, and M. Vichi (2022). Sea ice core temperature and salinity data collected during the 2019 SCALE Spring Cruise. Zenodo [data set]. DOI: 10.5281/ZENODO.6997630.

Australian Antarctic Data Centre **(AADC)** (n.d.). Data and Metadata Upload. AADC. URL: https://data.aad.gov.au/about/help-and-resources/metadata.

Australian Research Data Commons **(ARDC)** (2024). Guide to Choosing a Data Repository. Accessed: 2024-10-20. URL: https://ardc.edu.au/resource/guide-to-choosing-a-data-repository/.

Bader, H.-P. and P. Weilenmann (1992). Modeling temperature distribution, energy and mass flow in a (phase-changing) snowpack. I. Model and case studies. In: *Cold Regions Science and Technology* 20.2, pp. 157–181. DOI: 10.1016/0165-232x(92)90015-m.

Bahim, C., C. Casorrán-Amilburu, M. Dekkers, E. Herczog, N. Loozen, K. Repanas, K. Russell, and S. Stall (2020a). fairdat. Accessed: 2024-10-20. URL: https://www.surveymonkey.com/r/fairdat.

– (2020b). The FAIR Data Maturity Model: An Approach to Harmonise FAIR Assessments. In: *Data Science Journal* 19. DOI: 10.5334/dsj-2020-041.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. In: *Nature* 533.7604, pp. 452–454. DOI: 10.1038/533452a.

Barba, L. A. (2018). Terminologies for Reproducible Research. arXiv [preprint]. DOI: 10.48550/ARXIV.1802.03311.

Barker, M. et al. (2022). Introducing the FAIR Principles for research software. In: *Scientific Data* 9.1. DOI: 10.1038/s41597-022-01710-x.

Barnes, N. (2010). Publish your computer code: it is good enough. In: *Nature* 467.7317, pp. 753–753. DOI: 10.1038/467753a.

Barnett, T. P., J. C. Adam, and D. P. Lettenmaier (2005). Potential impacts of a warming climate on water availability in snow-dominated regions. In: *Nature* 438.7066, pp. 303–309. DOI: 10.1038/nature04141.

Bartelt, P. and M. Christen (1999). A computational procedure for instationary temperature-dependent snow creep. In: *Advances in Cold-Region Thermal Engineering and Sciences*. Ed. by K. Hutter, Y. Wang, and H. Beer. Springer Berlin Heidelberg, pp. 367–386.

Bartelt, P. and M. Lehning (2002). A physical SNOWPACK model for the Swiss avalanche warning: Part I: numerical model. In: *Cold Regions Science and Technology* 35.3, pp. 123–145. DOI: 10.1016/s0165-232x(02)00074-5.

Bavay, M., T. Grünewald, and M. Lehning (2013). Response of snow cover and runoff to climate change in high Alpine catchments of Eastern Switzerland. In: *Advances in Water Resources* 55, pp. 4–16. DOI: https://doi.org/10.1016/j.advwatres.2012.12.009.

Bavay, M. and J. Fiddes (2020). Global Cryosphere Watch data survey. EnviDat. DOI: 10.16904/ENVIDAT.133.

Bavay, M., J. Fiddes, and Ø. Godøy (2020). Automatic Data Standardization for the Global Cryosphere Watch Data Portal. In: *Data Science Journal* 19. DOI: 10.5334/dsj-2020-006.

Bender, M. (2002). Concentration and Isotopic Composition of O2 and N2 in Trapped Gases of the Vostok Ice Core. DOI: 10.7265/N5862DCW.

Berenguer, A., A. Morejón, D. Tomás, and J.-N. Mazón (2024). Using Large Language Models to Enhance the Reusability of Sensor Data. In: *Sensors* 24.2. DOI: 10.3390/s24020347.

Berners-Lee, T., R. T. Fielding, and L. M. Masinter (2005). Uniform Resource Identifier (URI): Generic Syntax. RFC 3986. DOI: 10.17487/RFC3986.

Berners-Lee, T., L. M. Masinter, and M. P. McCahill (1994). Uniform Resource Locators (URL). RFC 1738. DOI: 10.17487/RFC1738.

Bishop, B. W. and H. Collier (2022). Fitness for use of data: scientists' heuristics of discovery and reuse behaviour framed by the FAIR Data Principles. In: *Information Research: an international electronic journal* 27.3. DOI: 10.47989/irpaper942.

Bishop, B. W., C. Hank, J. Webster, and R. Howard (2019). Scientists' data discovery and reuse behavior: (Meta)data fitness for use and the FAIR data principles. In: *Proceedings of the Association for Information Science and Technology* 56.1, pp. 21–31. DOI: 10.1002/pra2.4.

Bloemers, M. and A. Montesanti (2020). The FAIR Funding Model: Providing a Framework for Research Funders to Drive the Transition toward FAIR Data Management and Stewardship Practices. In: *Data Intelligence* 2.1-2, pp. 171–180. DOI: 10.1162/dint_a_00039.

Boté, J.-J. and M. Térmens (2019). Reusing Data: Technical and Ethical Challenges. In: *DESIDOC Journal of Library & Information Technology* 39.6, pp. 329–337. DOI: 10.14429/djlit.39.06.14807.

Bouchard, B., D. F. Nadeau, and F. Domine (2022). Snow pit dataset from "Comparison of snowpack structure in gaps and under the canopy in a humid boreal forest". Zenodo [data set]. DOI: 10.5281/ZENODO.6469035.

Boyer et al. (2018). World Ocean Database 2018. Ed. by A. Mishonov. URL: https://www.ncei.noaa.gov/sites/default/files/2020-04/wod_intro_0.pdf.

British Oceanographic Data Centre **(BODC)** (2023). Polygon dataset of the extent of water bodies from the SeaVoX Salt and Fresh Water Body Gazetteer (v19). Marineregions.org [data set]. United Kingdom. DOI: 10.14284/590.

Brondex, J., K. Fourteau, M. Dumont, P. Hagenmuller, N. Calonne, F. Tuzet, and H. Löwe (2023a). Supplementary to "A finite-element framework to explore the numerical solution of the coupled problem of heat conduction, water vapor diffusion and settlement in dry snow (IvoriFEM v0.1.0)". Zenodo [software]. DOI: 10.5281/ZENODO.7941767.

Brondex, J., K. Fourteau, M. Dumont, P. Hagenmuller, N. Calonne, F. Tuzet, and H. Löwe (2023b). A finite-element framework to explore the numerical solution of the coupled problem of heat conduction, water vapor diffusion, and settlement in dry snow (IvoriFEM v0.1.0). In: *Geoscientific Model Development* 16, pp. 7075–7106. DOI: 10.5194/gmd-16-7075-2023.

– (2023c). ivori_model_homemadefem. GitHub [software]. URL: https://github.com/jbrondex/ivori_model_homemadefem.

Brown, K. A., L. A. Miller, C. J. Mundy, T. Papakyriakou, R. Francois, M. Gosselin, G. Carnat, K. Swystun, and P. D. Tortell (2015). Inorganic carbon system dynamics in landfast Arctic sea ice during the early-melt period. In: *Journal of Geophysical Research: Oceans* 120.5, pp. 3542–3566. DOI: 10.1002/2014JC010620.

Brugger, S. O., A. A. Jimenez, L. Ponsoni, and C. Todd (2022). Cryosphere Sciences Perspectives on Integrated, Coordinated, Open, Networked (ICON) Science. In: *Earth and Space Science* 9.4. DOI: 10.1029/2021ea002111.

Brun, E., P. David, M. Sudul, and G. Brunot (1992). A numerical model to simulate snow-cover stratigraphy for operational avalanche forecasting. In: *Journal of Glaciology* 38.128, pp. 13–22. DOI: 10.1017/s0022143000009552.

Brun, E., E. Martin, V. Simon, C. Gendre, and C. Coleou (1989). An Energy and Mass Model of Snow Cover Suitable for Operational Avalanche Forecasting. In: *Journal of Glaciology* 35.121, pp. 333–342. DOI: 10.3189/s0022143000009254.

Budyko, M. (1974). "II Methods for Determining the Components of the Heat Balance". In: *Climate and Life*. Ed. by D. H. . MILLER. Vol. 18. International Geophysics. Academic Press, pp. 41–139. DOI: 10.1016/S0074-6142(09)60007-3.

Buffo, J. J., B. E. Schmidt, and C. Huber (2018). Multiphase Reactive Transport and Platelet Ice Accretion in the Sea Ice of McMurdo Sound, Antarctica. In: *Journal of Geophysical Research: Oceans* 123.1, pp. 324–345. DOI: 10.1002/2017JC013345.

Calonne, N., F. Flin, S. Morin, B. Lesaffre, S. R. du Roscoat, and C. Geindreau (2011). Numerical and experimental investigations of the effective thermal conductivity of snow. In: *Geophysical Research Letters* 38.23, p. L23501. DOI: 10.1029/2011GL049234.

Calonne, N., C. Geindreau, and F. Flin (2014). Macroscopic Modeling for Heat and Water Vapor Transfer in Dry Snow by Homogenization. In: *The Journal of Physical Chemistry B* 118.47, pp. 13393–13403. DOI: 10.1021/jp5052535.

Candela, L., D. Mangione, and G. Pavone (2024). The FAIR Assessment Conundrum: Reflections on Tools and Metrics. In: *Data Science Journal* 23, p. 33. DOI: 10.5334/dsj-2024-033.

Carroll, S. R. et al. (2020). The CARE Principles for Indigenous Data Governance. In: *Data Science Journal* 19. DOI: 10.5334/dsj-2020-043.

Castellani, G. et al. (2019). "Sea ice production and ecology study (SIPES2)". In: *The Expedition PS117 of the Research Vessel POLARSTERN to the Weddell Sea in 2018/2019*. Ed. by O. Boebel. Vol. 732. Reports on Polar and Marine Research. Bremerhaven, Germany: Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research. DOI: 10.2312/BzPM_0732_2019.

Chen, H., Y. Chen, W. Li, and Z. Li (2019). Quantifying the contributions of snow/glacier meltwater to river runoff in the Tianshan Mountains, Central Asia. In: *Global and Planetary Change* 174, pp. 47–57. DOI: 10.1016/j.gloplacha.2019.01.002.

Christensen, G., A. Dafoe, E. Miguel, D. A. Moore, and A. K. Rose (2019). A study of the impact of data sharing on article citations using journal policies as a natural experiment. In: *PLOS ONE* 14.12, pp. 1–13. DOI: 10.1371/journal.pone.0225883. URL: 10.1371/journal.pone.0225883.

Chue Hong, N. et al. (2023). D5.2 - Metrics for automated FAIR software assessment in a disciplinary context. DOI: 10.5281/ZENODO.10047401.

Cinquini, L. et al. (2014). The Earth System Grid Federation: An open infrastructure for access to distributed geospatial data. In: *Future Generation Computer Systems* 36, pp. 400–417. DOI: 10.1016/j.future.2013.07.002.

Claerbout, J. F. and M. Karrenbach (1992). Electronic Documents Give Reproducible Research a New Meaning. In: *1992 SEG Annual Meeting*, SEG-1992–0601. DOI: 10.1190/1.1822162.

Clarke, D. J. et al. (2019). FAIRshake: Toolkit to Evaluate the FAIRness of Research Digital Resources. In: *Cell Systems* 9.5, pp. 417–421. DOI: 10.1016/j.cels.2019.09.011.

Control Company (2016). Control Company 4000 Instruction Manual. [Manual]. Webster, Texas, USA. URL: https://www.novatech-usa.com/pdf/Control%20Company%204000%20Instruction%20Manual.pdf.

Cox, S. and J. Yu (2017). OzNome 5-star Tool: A Rating System for making data FAIR and Trustable. In: *eResearch Australasia 2017, Brisbane, Australia.*

Creative Commons (2023). About CC Licenses. URL: https://creativecommons.org/about/cclicenses/.

Crystal-Ornelas, R. et al. (2022). Enabling FAIR data in Earth and environmental science with community-centric (meta)data reporting formats. In: *Scientific Data* 9.1. DOI: 10.1038/s41597-022-01606-w.

Crystal-Ornelas, R. et al. (2021). A Guide to Using GitHub for Developing and Versioning Data Standards and Reporting Formats. In: *Earth and Space Science* 8.8. DOI: 10.1029/2021ea001797.

Crystallography Open Database **(COD)** (2023). Crystallography Open Database. URL: https://www.crystallography.net/cod/.

Curty, R. G., K. Crowston, A. Specht, B. W. Grant, and E. D. Dalton (2017). Attitudes and norms affecting scientists' data reuse. In: *PLOS ONE* 12.12. Ed. by C. R. Sugimoto, e0189288. DOI: 10.1371/journal.pone.0189288.

Curty, R. G. and J. Qin (2014). Towards a model for research data reuse behavior. In: *Proceedings of the American Society for Information Science and Technology* 51.1, pp. 1–4. DOI: 10.1002/meet.2014.14505101072.

Da Silva Santos, L. O. B., K. Burger, R. Kaliyaperumal, and M. D. Wilkinson (2023). FAIR Data Point: A FAIR-Oriented Approach for Metadata Publication. In: *Data Intelligence* 5.1, pp. 163–183. DOI: 10.1162/dint_a_00160.

DataCite (2024a). How can I map different metadata formats to the DataCite schema? Accessed: 2024-10-23. URL: https://support.datacite.org/docs/how-can-i-map-different-metadata-formats-to-the-datacite-xml.

DataCite Metadata Working Group **(DataCite)** (2024b). DataCite Metadata Schema for the Publication and Citation of Research Data and Other Research Outputs v4.5. DOI: 10.14454/ZNVD-6Q68.

David, R. et al. (2020). FAIRness Literacy: The Achilles' Heel of Applying FAIR Principles. In: *Data Science Journal* 19. DOI: 10.5334/dsj-2020-032.

Davis, K. R., B. Peabody, and P. Leach (2024). Universally Unique IDentifiers (UUIDs). RFC 9562. DOI: 10.17487/RFC9562.

dbPedia (n.d.). Temperature. URL: https://dbpedia.org/page/Temperature.

De Smedt, K., D. Koureas, and P. Wittenburg (2020). FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units. In: *Publications* 8.2. DOI: 10.3390/publications8020021.

Dempsey, W. P., I. Foster, S. Fraser, and C. Kesselman (2022). Sharing Begins at Home: How Continuous and Ubiquitous FAIRness Can Enhance Research Productivity and Data Reuse. In: *Harvard Data Science Review*. DOI: 10.1162/99608f92.44d21b86.

Devaraju, A. and R. Huber (2024). F-UJI - An Automated FAIR Data Assessment Tool. Zenodo [software]. DOI: 10.5281/ZENODO.6361400.

Devaraju, A., R. Huber, M. Mokrane, P. Herterich, L. Cepinskas, J. de Vries, H. L'Hours, J. Davidson, and A. White (2020). FAIRsFAIR Data Object Assessment Metrics. Version 0.4. DOI: 10.5281/zenodo.4081213.

Domine, F., M. Barrere, and D. Sarrazin (2016). Seasonal evolution of the effective thermal conductivity of the snow and the soil in high Arctic herb tundra at Bylot Island, Canada. In: *The Cryosphere* 10.6, pp. 2573–2588. DOI: 10.5194/tc-10-2573-2016.

Domine, F., G. Picard, S. Morin, M. Barrere, J.-B. Madore, and A. Langlois (2019). Major Issues in Simulating Some Arctic Snowpack Properties Using Current Detailed Snow Physics Models: Consequences for the Thermal Regime and Water Budget of Permafrost. In: *Journal of Advances in Modeling Earth Systems* 11.1, pp. 34–44. DOI: 10.1029/2018ms001445.

Dublin Core Metadata Initiative Usage Board **(DCMI)** (2020). Dublin Core Metadata Initiative Metadata Terms. URL: http://dublincore.org/specifications/dublin-core/dcmi-terms/2020-01-20/.

Duerr, R., P. L. Buttigieg, G. B. Cross, K. L. Blumberg, B. Whitehead, N. Wiegand, and K. Rose (2024). Harmonizing GCW Cryosphere Vocabularies with ENVO and SWEET. Towards a General Model for Semantic Harmonization. In: *Data Science Journal* 23.practices.

Dunning, A. (, M. de Smaele, and J. K. Böhmer (2017). Evaluation of data repositories based on the FAIR Principles for IDCC 2017 practice paper. DOI: 10.4121/UUID:5146DD06-98E4-426C-9AE5-DC8FA65C549F.

Duprat, L., N. Kanna, J. Janssens, A. Roukaerts, F. Deman, A. T. Townsend, K. M. Meiners, P. van der Merwe, and D. Lannuzel (2019). Enhanced Iron Flux to Antarctic Sea Ice via Dust Deposition From Ice-Free Coastal Areas. In: *Journal of Geophysical Research: Oceans* 124.12, pp. 8538–8557. DOI: 10.1029/2019jc015221.

Duprat, L. P. (2019). Davis sea ice Nov. 2016. AADC [data set]. DOI: 10.26179/5CAD74D6A3179.

Dutton, A., A. E. Carlson, A. J. Long, G. A. Milne, P. U. Clark, R. DeConto, B. P. Horton, S. Rahmstorf, and M. E. Raymo (2015). Sea-level rise due to polar ice-sheet mass loss during past warm periods. In: *Science* 349.6244, aaa4019. DOI: 10.1126/science.aaa4019.

Easterday, K., T. Paulson, P. DasMohapatra, P. Alagona, S. Feirer, and M. Kelly (2018). From the Field to the Cloud: A Review of Three Approaches to Sharing Historical Data From Field Stations Using Principles From Data Science. In: *Frontiers in Environmental Science* 6. DOI: 10.3389/fenvs.2018.00088.

Eidhammer, T., A. Booth, S. Decker, L. Li, M. Barlage, D. Gochis, R. Rasmussen, K. Melvold, A. Nesje, and S. Sobolowski (2021). Mass balance and hydrological modeling of the Hardangerjøkulen ice cap in south-central Norway. In: *Hydrology and Earth System Sciences* 25.8, pp. 4275–4297. DOI: 10.5194/hess-25-4275-2021.

Environment Ontology **(ENVO)** (n.d.). water ice. URL: http://purl.obolibrary.org/obo/ENVO_01000277.

European Commission: Directorate-General for Research and Innovation **(EC DGRI)** (2016). Guidelines on FAIR Data Management in Horizon 2020. Accessed: 2024-10-20. URL: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.

– (2016). *Realising the European Open Science Cloud – First report and recommendations of the Commission high level expert group on the European Open Science Cloud*. Publications Office. DOI: 10.2777/940154.

– (2018). *Cost-benefit analysis for FAIR research data: cost of not having FAIR research data*. Publications Office. DOI: 10.2777/02999.

European Commission: Directorate-General for Research and Innovation and EOSC Executive Board **(EC DGRI EOSC)** (2021). Recommendations on FAIR metrics for EOSC. DOI: 10.2777/70791.

European Molecular Biology Laboratory - European Bioinformatics Institute **(EMBL-EBI)** (2023). Ontology Lookup Service. Accessed: 2024-10-18. URL: http://www.ebi.ac.uk/ols4/.

European Parliament, Council of the European Union **(EU)** (2023). Regulation (EU) 2023/2854 of the European Parliament and of the council (Data Act). Official Journal of the European Union. URL: http://data.europa.eu/eli/reg/2023/2854/oj.

European Space Agency **(ESA)** (n.d.). ESA Earth Online. Accessed: 2024-10-20. URL: https://earth.esa.int/eogateway/.

FAIR-IMPACT (n.d.). Domain-agnostic metrics for software assessment. [internet]. Version 1.0. URL: https://fair-impact.eu/metrics-software.

FAIRsharing Team (2015). FAIRsharing record for: Quantities, Units, Dimensions and Types. DOI: 10.25504/FAIRSHARING.D3PQW7.

– (2018a). FAIRsharing record for: Digital Object Identifier. FAIRsharing. DOI: 10.25504/FAIRSHARING.HFLKCN.

– (2018b). FAIRsharing record for: The Data Use Ontology. FAIRsharing. DOI: 10.25504/FAIRSHARING.5DNJS2.

– (2018c). FAIRsharing record for: The Unified Code for Units of Measure. FAIRsharing. DOI: 10.25504/FAIRSHARING.27W8K0.

– (2022a). FAIRsharing record for: GeoJSON. FAIRsharing. DOI: 10.25504/FAIRSHARING.4D969B.

– (2022b). FAIRsharing record for: OpenAIRE Guidelines for Data Archives. FAIRsharing. DOI: 10.25504/FAIRSHARING.123197.

– (2023). FAIRsharing record for: International Generic Sample Number. FAIRsharing. DOI: 10.25504/FAIRSHARING.C7F365.

Faniel, I., E. Kansa, S. Whitcher Kansa, J. Barrera-Gomez, and E. Yakel (2013). The challenges of digging data: a study of context in archaeological data reuse. In: *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*. JCDL '13. Indianapolis, Indiana, USA: Association for Computing Machinery, pp. 295–304. DOI: 10.1145/2467696.2467712.

Faniel, I. M., R. D. Frank, and E. Yakel (2019). Context from the data reuser's point of view. In: *Journal of Documentation* 75.6, pp. 1274–1297. DOI: 10.1108/jd-08-2018-0133.

Faniel, I. M. and T. E. Jacobsen (2010). Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data. In: *Computer Supported Cooperative Work* 19.3–4, pp. 355–375. DOI: 10.1007/s10606-010-9117-8.

Faniel, I. M., A. Kriesberg, and E. Yakel (2015). Social scientists' satisfaction with data reuse. In: *Journal of the Association for Information Science and Technology* 67.6, pp. 1404–1416. DOI: 10.1002/asi.23480.

Farlow, S. J. (1993). *Partial Differential Equations for Scientists and Engineers*. Reprint of the John Wiley & Sons, New York, USA, 1982 edition. Dover Publications, Mineola, New York, USA. 414 pp. ISBN: 048667620X.

Felden, J., L. Möller, U. Schindler, R. Huber, S. Schumacher, R. Koppe, M. Diepenbroek, and F. O. Glöckner (2023). PANGAEA - Data Publisher for Earth & Environmental Science. In: *Scientific Data* 10.1. DOI: 10.1038/s41597-023-02269-x.

Feltham, D. L., N. Untersteiner, J. S. Wettlaufer, and M. G. Worster (2006). Sea ice is a mushy layer. In: *Geophysical Research Letters* 33.14. DOI: 10.1029/2006GL026290.

Figshare (2024). Figshare about. Accessed: 2024-10-28. URL: https://knowledge.figshare.com/about.

Figueiredo, A. S. (2017). Data Sharing: Convert Challenges into Opportunities. In: *Frontiers in Public Health* 5. DOI: 10.3389/fpubh.2017.00327.

Frakes, W. and K. Kang (2005). Software reuse research: status and future. In: *IEEE Transactions on Software Engineering* 31.7, pp. 529–536. DOI: 10.1109/tse.2005.85.

Frakes, W. and C. Terry (1996). Software reuse: metrics and models. In: *ACM Computing Surveys* 28.2, pp. 415–435. DOI: 10.1145/234528.234531.

Freire, J. and F. Chirigati (2018). Provenance and the Different Flavors of Computational Reproducibility. In: *IEEE Data Engineering Bulletin* 41, pp. 15–26. URL: http://sites.computer.org/debull/A18mar/A18MAR-CD.pdf.

Frémand, A. C. et al. (2023). Antarctic Bedmap data: Findable, Accessible, Interoperable, and Reusable (FAIR) sharing of 60 years of ice bed, surface, and thickness data. In: *Earth System Science Data* 15.7, pp. 2695–2710. DOI: 10.5194/essd-15-2695-2023.

G20 (2016). G20 Leaders' Communique Hangzhou Summit. [internet]. Accessed: 2024-10-20. URL: https://ec.europa.eu/commission/presscorner/detail/en/STATEMENT_16_2967.

Gaignard, A., T. Rosnet, F. De Lamotte, V. Lefort, and M.-D. Devignes (2023). FAIR-Checker: supporting digital resource findability and reuse with Knowledge Graphs and Semantic Web standards. In: *Journal of Biomedical Semantics* 14.1. DOI: 10.1186/s13326-023-00289-5.

Gärtner-Roer, I., S. U. Nussbaumer, B. Raup, F. Paul, E. Welty, A. K. Windnagel, F. Fetterer, and M. Zemp (2022). Democratizing Glacier Data – Maturity of Worldwide Datasets and Future Ambitions. In: *Frontiers in Climate* 4. DOI: 10.3389/fclim.2022.841103.

Geosoft (n.d.). Community Surface Dynamics Modeling System Portal. Accessed: 2024-09-26. URL: https://csdms.ontosoft.org/.

Gill, N. and S. Sikka (2011). Inheritance Hierarchy Based Reuse & Reusability Metrics in OOSD. In: *International Journal on Computer Science and Engineering* 3.6.

GitHub (2024). Issuing a persistent identifier for your repository with Zenodo. Accessed: 2024-10-22. URL: https://docs.github.com/en/repositories/archiving-a-github-repository/referencing-and-citing-content#issuing-a-persistent-identifier-for-your-repository-with-zenodo.

– (n.d.). About GitHub. URL: https://github.com/about.

Giuliani, G., H. Cazeaux, P.-Y. Burgi, C. Poussin, J.-P. Richard, and B. Chatenoux (2021). SwissEnvEO: A FAIR National Environmental Data Repository for Earth Observation Open Science. In: *Data Science Journal* 20. DOI: 10.5334/dsj-2021-022.

Global Change Master Directory **(GCMD)** (2024). GCMD Keywords. Version 19.8 Greenbelt, MD: Earth Science Data and Information System, Earth Science Projects Division, Goddard Space Flight Center, NASA. Accessed: 2024-11-02. URL: https://forum.earthdata.nasa.gov/app.php/tag/GCMD+Keywords.

Goble, C., S. Cohen-Boulakia, S. Soiland-Reyes, D. Garijo, Y. Gil, M. R. Crusoe, K. Peters, and D. Schober (2020). FAIR Computational Workflows. In: *Data Intelligence* 2.1–2, pp. 108–121. DOI: 10.1162/dint_a_00033.

GOFAIR (2023). FAIR principles. Accessed: 2024-09-13. URL: https://www.go-fair.org/fair-principles/.

Goldman, A. E., S. R. Emani, L. C. Pérez-Angel, J. A. Rodríguez-Ramos, and J. C. Stegen (2022). Integrated, Coordinated, Open, and Networked (ICON) Science to Advance the Geosciences: Introduction and Synthesis of a Special Collection of Commentary Articles. In: *Earth and Space Science* 9.4. DOI: 10.1029/2021ea002099.

Goldman, A., S. Emani, L. Pérez-Angel, J. Rodríguez-Ramos, J. Stegen, and P. Fox (2021). Special Collection on Open Collaboration Across Geosciences. In: *Eos* 102. DOI: 10.1029/2021eo153180.

Gregory, K. (2020). A dataset describing data discovery and reuse practices in research. In: *Scientific Data* 7.1. DOI: 10.1038/s41597-020-0569-5.

Groth, P., H. Cousijn, T. Clark, and C. Goble (2020). FAIR Data Reuse – the Path through Data Citation. In: *Data Intelligence* 2.1–2, pp. 78–86. DOI: 10.1162/dint_a_00030.

Gruenpeter, M. et al. (2021). Defining Research Software: a controversial discussion. Zenodo. DOI: 10.5281/ZENODO.5504016.

Haas, C., A. Friedrich, Z. Li, M. Nicolaus, A. Pfaffling, and T. Toyota (2009). Regional variability of sea ice properties and thickness in the Northwestern Weddell Sea obtained by in-situ and satellite measurements. In: *The Expedition of the Research Vessel Polarsternto the Antarctic in 2006 (ANT-XXIII/7)*. Ed. by P. Lemke. Vol. 586. Reports on Polar and Marine Research. DOI: 10.2312/BzPM_0586_2009.

Hach (2000). sensION5 Conductivity Meter Manual. [Manual]. Loveland, Colorado, USA. URL: https://www.fondriest.com/pdf/hach_sension5_manual.pdf.

Halbritter, A. H. et al. (2019). The handbook for standardized field and laboratory measurements in terrestrial climate change experiments and observational studies (ClimEx). In: *Methods in Ecology and Evolution* 11.1. Ed. by R. Freckleton, pp. 22–37. DOI: 10.1111/2041-210x.13331.

Hansen, A. and W. Foslien (2015a). A macroscale mixture theory analysis of deposition and sublimation rates during heat and mass transfer in dry snow. In: *The Cryosphere* 9.5, pp. 1857–1878. DOI: 10.5194/tc-9-1857-2015.

– (2015b). Supplementary material of "A macroscale mixture theory analysis of deposition and sublimation rates during heat and mass transfer in dry snow". DOI: 10.5194/tc-9-1857-2015-supplement.

Harris, C. R. et al. (2020). Array programming with NumPy. In: *Nature* 585.7825, pp. 357–362. DOI: 10.1038/s41586-020-2649-2.

Hastings, J., G. Owen, A. Dekker, M. Ennis, N. Kale, V. Muthukrishnan, S. Turner, N. Swainston, P. Mendes, and C. Steinbeck (2016). ChEBI in 2016: Improved services and an expanding collection of metabolites. In: *Nucleic acids research* 44.D1, pp. D1214–9. DOI: 10.1093/nar/gkv1031.

Helfricht, K., L. Hartl, R. Koch, C. Marty, and M. Olefs (2018). Obtaining sub-daily new snow density from automated measurements in high mountain regions. In: *Hydrology and Earth System Sciences* 22.5, pp. 2655–2668. DOI: 10.5194/hess-22-2655-2018.

Hendricks, S., M. Nicolaus, R. Ricker, M. Hoppman, P. Hunkeler, and C. Katlein (2012). Sea ice physics. In: *The Expedition of the Research Vessel Polarstern to the Arctic in 2011 (ARK-XXVI/3 - TransArc)*. Ed. by U. Schauer. Vol. 649. Bremerhaven, Germany: Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research. DOI: 10.2312/BzPM_0649_2012.

Holub, P. et al. (2018). Enhancing Reuse of Data and Biological Material in Medical Research: From FAIR to FAIR-Health. In: *Biopreservation and Biobanking* 16.2, pp. 97–105. DOI: 10.1089/bio.2017.0110.

Horsburgh, J. S. et al. (2016). Observations Data Model 2: A community information model for spatially discrete Earth observations. In: *Environmental Modelling & Software* 79, pp. 55–74. DOI: 10.1016/j.envsoft.2016.01.010.

Huang, K., R. Zhang, and M. T. van Genuchten (1994). An Eulerian-Lagrangian approach with an adaptively corrected method of characteristics to simulate variably saturated water flow. In: *Water Resources Research* 30.2, pp. 499–507. DOI: 10.1029/93WR02881.

Huber, R., E. Gordeev, M. Stocker, A. Balamurugan, and U. Schindler (2020). PANGAEApy - a Python module to access and analyse PANGAEA data. Zenodo [software]. Exeter, Devon. DOI: 10.5281/zenodo.4013940.

Hughes, L. D. et al. (2023). Addressing barriers in FAIR data practices for biomedical data. In: *Scientific Data* 10.1. DOI: 10.1038/s41597-023-01969-8.

Huntington, H. P., S. Gearheard, L. K. Holm, G. Noongwook, M. Opie, and J. Sanguya (2017). Sea ice is our beautiful garden: indigenous perspectives on sea ice in the Arctic. In: *Sea Ice*. John Wiley & Sons, Ltd. Chap. 25, pp. 583–599. DOI: 10.1002/9781118778371.ch25.

IEEE (2010). IEEE Standard for Information Technology–System and Software Life Cycle Processes–Reuse Processes. In: *IEEE Std 1517-2010 (Revision of IEEE Std 1517-1999)*, pp. 1–51. DOI: 10.1109/IEEESTD.2010.5551093.

Imoize, A., D. Idowu, and T. Bolaji (2019). A Brief Overview of Software Reuse and Metrics in Software Engineering. In: *World Scientific News*.

Institute of Data Science at Maastricht University **(IDS)** (2020). FAIRificator. URL: https://maastrichtu-ids.github.io/fairificator/.

International Organization for Standardization **(ISO)** (2014). *ISO 19115. Geographic information — Metadata*. Tech. rep. Geneva, CH: International Organization for Standardization.

– (2019). *ISO 19139. Geographic information — XML schema implementation*. Tech. rep. Geneva, CH: International Organization for Standardization.

Isleifson, D., B. Hwang, D. G. Barber, R. K. Scharien, and L. Shafai (2010a). (Table I) Physical properties of sea ice, frost flowers and brine at ArcticNet, CFL and CASES ice stations. PANGAEA [data set]. DOI: 10.1594/PANGAEA.811539.

– (2010b). C-Band Polarimetric Backscattering Signatures of Newly Formed Sea Ice During Fall Freeze-Up. In: *IEEE Transactions on Geoscience and Remote Sensing* 48.8, pp. 3256–3267. DOI: 10.1109/TGRS.2010.2043954.

Itkin, P., A. Steer, A. Cristea, B. Raffel, D. Divine, and S. Gerland (2024). Snowpit records from the Nansen Legacy Project. npolar.no [data set]. DOI: 10.21334/NPOLAR.2024.6A0BF7B6.

Jackson, R. et al. (2021). OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies. In: *Database* 2021. DOI: 10.1093/database/baab069.

Jacobsen, A., R. Kaliyaperumal, L. O. B. da Silva Santos, B. Mons, E. Schultes, M. Roos, and M. Thompson (2020a). A Generic Workflow for the Data FAIRification Process. In: *Data Intelligence* 2.1–2, pp. 56–65. DOI: 10.1162/dint_a_00028.

Jacobsen, A. et al. (2020b). FAIR Principles: Interpretations and Implementation Considerations. In: *Data Intelligence* 2.1–2, pp. 10–29. DOI: 10.1162/dint_r_00024.

Jafari, M., I. Gouttevin, M. Couttet, N. Wever, A. Michel, V. Sharma, L. Rossmann, N. Maass, M. Nicolaus, and M. Lehning (2020). The Impact of Diffusive Water Vapor Transport on Snow Profiles in Deep and Shallow Snow Covers and on Sea Ice. In: *Frontiers in Earth Science* 8, p. 249. DOI: 10.3389/feart.2020.00249.

Jashapara, A. (2011). *Knowledge management : an integrated approach / Ashok Jashapara*. 2. edition. Harlow: Pearson. ISBN: 9780273728191.

Jiang, J., F. Wang, J. Shen, S. Kim, and S. Kim (2025). A Survey on Large Language Models for Code Generation. In: *ACM Trans. Softw. Eng. Methodol.* Just Accepted. DOI: 10.1145/3747588.

Johnson, B. R., A. Leon, and S. J. S. Khalsa (2015). Data management in the ERA of a rapidly changing cryosphere. In: *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE. DOI: 10.1109/igarss.2015.7326028.

Johnson, J. B. (2011). "Snow Deformation". In: *Encyclopedia of Snow, Ice and Glaciers*. Ed. by V. P. Singh, P. Singh, and U. K. Haritashya. Springer, Dordrecht, Netherlands, pp. 1041–1045. DOI: 10.1007/978-90-481-2642-2_501.

Johnson, S., R. R. Audh, W. de Jager, B. Matlakala, M. Vichi, A. Womack, and T. Rampai (2023). Physical and morphological properties of first-year Antarctic sea ice in the spring marginal ice zone of the Atlantic-Indian sector. In: *Journal of Glaciology*, pp. 1–14. DOI: 10.1017/jog.2023.21.

Jones, M. B. et al. (2023). CodeMeta: an exchange schema for software metadata. Version 3.0. URL: https://w3id.org/codemeta/v3.0.

Judson, A. and N. Doesken (2000). Density of Freshly Fallen Snow in the Central Rocky Mountains. In: *Bulletin of the American Meteorological Society* 81.7, pp. 1577–1587. DOI: 10.1175/1520-0477(2000)081<1577:DOFFSI>2.3.CO;2.

Katlein, C., P. Itkin, and D. V. Divine (2020a). Salinity measured on sea ice core PS122/2_24-114 during MOSAiC Leg 2. PANGAEA [data set]. DOI: 10.1594/PANGAEA.919474.

Katlein, C. et al. (2020b). Platelet ice Under Arctic pack ice in winter. In: *Geophysical Research Letters* 47.16. DOI: 10.1029/2020gl088898.

Kim, Y. (2021). A study of the roles of metadata standard and data repository in science, technology, engineering and mathematics researchers' data reuse. In: *Online Information Review* 45.7, pp. 1306–1321. DOI: 10.1108/oir-09-2020-0431.

Kinkade, D. and A. Shepherd (2021). Geoscience data publication: Practices and perspectives on enabling the FAIR guiding principles. In: *Geoscience Data Journal* 9.1, pp. 177–186. DOI: 10.1002/gdj3.120.

Kirchner, H. O. K., G. Michot, H. Narita, and T. Suzuki (2001). Snow as a foam of ice: Plasticity, fracture and the brittle-to-ductile transition. In: *Philosophical Magazine A* 81.9, pp. 2161–2181. DOI: 10.1080/01418610108217141.

Klump, J., K. Lehnert, D. Ulbricht, A. Devaraju, K. Elger, D. Fleischer, S. Ramdeen, and L. Wyborn (2021). Towards Globally Unique Identification of Physical Samples: Governance and Technical Implementation of the IGSN Global Sample Number. In: *Data Science Journal* 20. DOI: 10.5334/dsj-2021-033.

Kramer, M., K. M. Swadling, K. M. Meiners, R. Kiko, A. Scheltz, M. Nicolaus, and I. Werner (2010a). (Table 2.2.1) Average salinity and temperature, and integrated pigment concentrations of sea-ice cores of Aurora Australis cruise SIPEX. PANGAEA [data set]. DOI: 110.1594/PANGAEA.734526.

– (2010b). Vertical profile of bulk salinity and relative brine volume of sea-ice cores IO [1-3, 5,7-11,13,14,15b]. PANGAEA [data set]. DOI: 10.1594/PANGAEA.734461[61-63,65,66,68-75].

Kramer, M., K. M. Swadling, K. M. Meiners, R. Kiko, A. Scheltz, M. Nicolaus, and I. Werner (2010c). Vertical temperature and brine salinity profiles of sea-ice cores IO [1-3, 5,7-11,13,14,15b]. PANGAEA [data set]. DOI: 10.1594/PANGAEA.734417 [17-29].

– (2010d). Vertical temperature and brine salinity profiles of sea-ice cores PS69/[542-2, 543-1, 546-1, 549-1, 551-2, 554-2, 556-1, 558-1, 562-1, 564-1, 565-1, 567-4, 568-1, 572-1, 574-1, 576-1, 577-1, 578-1, 579-1, 581-1, 584-1, 585-1]. PANGAEA [data set]. DOI: 10.1594/PANGAEA.734389 [389,391-397,401-413].

– (2010e). Vertical profile of bulk salinity and relative brine volume of sea-ice cores PS69/[542-2, 543-1, 546-1, 549-1,551-2, 554-2, 556-1, 558-1, 562-1, 564-1, 565-1, 567-4, 568-1, 572-1, 574-1, 576-1, 577-1, 578-1,
579-1, 581-1, 584-1, 585-1]. PANGAEA [data set]. DOI: 10.1594/PANGAEA.734439 [39-53, 55-60, 77].

– (2011). Antarctic sympagic meiofauna in winter: Comparing diversity, abundance and biomass between perennially and seasonally ice-covered regions. In: *Deep Sea Research Part II: Topical Studies in Oceanography* 58.9-10, pp. 1062–1074. DOI: 10.1016/j.dsr2.2010.10.029.

Krinner, G. et al. (2018). ESM-SnowMIP: assessing snow models and quantifying snow-related climate feedbacks. In: *Geoscientific Model Development* 11.12, pp. 5027–5049. DOI: 10.5194/gmd-11-5027-2018.

Krol, Q. and H. Löwe (2016). Analysis of local ice crystal growth in snow. In: *Journal of Glaciology* 62.232, pp. 378–390. DOI: 10.1017/jog.2016.32.

Krueger, C. W. (1992). Software reuse. In: *ACM Computing Surveys* 24.2, pp. 131–183. DOI: 10.1145/130844.130856.

Kwok, R., A. A. Petty, G. Cunningham, T. Markus, D. Hancock, A. Ivanoff, J. Wimert, M. Bagnardi, N. Kurtz, and T. I.-2. S. Team (2023). ATLAS/ICESat-2 L3A Sea Ice Freeboard, Version 6. DOI: 10.5067/ATLAS/ATL10.006.

LaFlamme, M., M. Poetz, and D. Spichtinger (2022). Seeing oneself as a data reuser: How subjectification activates the drivers of data reuse in science. In: *PLOS ONE* 17.8. Ed. by S. Fàbregues, e0272153. DOI: 10.1371/journal.pone.0272153.

Lanergan, R. G. and C. A. Grasso (1984). Software Engineering with Reusable Designs and Code. In: *IEEE Transactions on Software Engineering* SE-10.5, pp. 498–501. DOI: 10.1109/tse.1984.5010273.

Lang, H. (1986). Forecasting Meltwater Runoff from Snow-Covered Areas and from Glacier Basins. In: *River Flow Modelling and Forecasting*. Ed. by D. A. Kraijenhoff and J. R. Moll. Dordrecht: Springer Netherlands, pp. 99–127. DOI: 10.1007/978-94-009-4536-4_5.

Lang, K., C. Assmann, N. Neute, R. Gerlach, and J. Rex (2023). FAIR Assessment Tools Overview. In: *3. Sächsische FDM-Tagung, Leipzig, 22. September 2022*. Zenodo. DOI: 10.5281/ZENODO.7701941.

Lange, B. A., C. Michel, J. F. Beckers, J. A. Casey, H. Flores, I. Hatam, G. Meisterhans, A. Niemi, and C. Haas (2015a). Comparing Springtime Ice-Algal Chlorophyll a and Physical Properties of Multi-Year and First-Year Sea Ice from the Lincoln Sea. In: *PLOS ONE* 10.4. Ed. by C. Lovejoy, e0122418. DOI: 10.1371/journal.pone.0122418.

Lange, B. A., C. Michel, J. Beckers, J. A. Casey, H. Flores, I. Hatam, G. Meisterhans, A. Niemi, and C. Haas (2015b). Ice-algal chlorophyll a and physical properties of multi-year and first-year sea ice of cores CASIMBO [1-10,1-12,3-10,4-10,5-10,5-12,6-12,7-12]. PANGAEA [data set]. DOI: 10.1594/PANGAEA.842359[59,61,65,68,70,71,73,75,76].

Lannuzel, D. (2016a). Iron in sea ice during campaign ARISE 2003. PANGAEA [data set]. DOI: 10.1594/PANGAEA.865023.

– (2016b). Iron in sea ice during campaign CASEY. PANGAEA [data set]. DOI: 10.1594/PANGAEA.865026.

– (2016c). Iron in sea ice during campaign ISPOL. PANGAEA [data set]. DOI: 10.1594/PANGAEA.865027.

– (2016d). Iron in sea ice during campaign SIPEX. PANGAEA [data set]. DOI: 10.1594/PANGAEA.865031.

– (2016e). Iron in sea ice during campaign SIPEX2. PANGAEA [data set]. DOI: 10.1594/PANGAEA.865035.

Lannuzel, D., A. R. Bowie, F. Chever, J. Jannssens, A.-J. Cavagna, and A. Roukaerts (2017). Sea ice Trace Metals sampling during the SIPEX II voyage of the Aurora Australis, 2012. AADC [data set]. DOI: 10.4225/15/59b0ddc2e4bd5.

Lannuzel, D., F. Chever, P. C. van der Merwe, J. Janssens, A. Roukaerts, A.-J. Cavagna, A. T. Townsend, A. R. Bowie, and K. M. Meiners (2016a). Iron biogeochemistry in Antarctic pack ice during SIPEX-2. In: *Deep Sea Research Part II: Topical Studies in Oceanography* 131, pp. 111–122. DOI: 10.1016/j.dsr2.2014.12.003.

Lannuzel, D., V. Schoemann, J. de Jong, L. Chou, B. Delille, S. Becquevort, and J.-L. Tison (2008). Iron study during a time series in the western Weddell pack ice. In: *Marine Chemistry* 108.1–2, pp. 85–95. DOI: 10.1016/j.marchem.2007.10.006.

Lannuzel, D., V. Schoemann, J. de Jong, J.-L. Tison, and L. Chou (2007). Distribution and biogeochemical behaviour of iron in the East Antarctic sea ice. In: *Marine Chemistry* 106.1-2, pp. 18–32. DOI: 10.1016/j.marchem.2006.06.010.

Lannuzel, D., M. Vancoppenolle, P. van der Merwe, J. de Jong, K. Meiners, M. Grotti, J. Nishioka, and V. Schoemann (2016b). Iron in sea ice: Review and new insights. In: *Elementa: Science of the Anthropocene* 4. Ed. by J. W. Deming and L. A. Miller. DOI: 10.12952/journal.elementa.000130.

Lehnert, K., Y. Su, C. H. Langmuir, B. Sarbas, and U. Nohl (2000). A global geochemical database structure for rocks. In: *Geochemistry, Geophysics, Geosystems* 1.5. DOI: 10.1029/1999gc000026.

Lehnert, K. (2015). Making Small Data Big. Slideshare [presentation]. URL: https://www.slideshare.net/slideshow/lehnert-making-small-data-big-iacs-april2015/46850055.

Lehning, M., P. Bartelt, B. Brown, C. Fierz, and P. Satyawali (2002). A physical SNOWPACK model for the Swiss avalanche warning: Part II. Snow microstructure. In: *Cold Regions Science and Technology* 35.3, pp. 147–167. DOI: 10.1016/s0165-232x(02)00073-3.

LeVeque, R. J. (2002). *Finite Volume Methods for Hyperbolic Problems*. Cambridge Texts in Applied Mathematics. Cambridge University Press. DOI: 10.1017/CBO9780511791253.

Libbrecht, K. G. (1999). Physical properties of ice. Accessed: 29 July 2021. URL: http://www.cco.caltech.edu/~atomic/snowcrystals/ice/ice.htm (visited on 2020).

Lippolis, A. S., M. J. Saeedizade, R. Keskisärkkä, S. Zuppiroli, M. Ceriani, A. Gangemi, E. Blomqvist, and A. G. Nuzzolese (2025). Ontology Generation Using Large Language Models. In: *The Semantic Web*. Springer Nature Switzerland, pp. 321–341.

Luna, R. A., J. Zubcoff, I. Garrigós, and H. Gonz (2022). FAIRification of Citizen Science Data. In: *Web Engineering*. Ed. by T. Di Noia, I.-Y. Ko, M. Schedl, and C. Ardito. Cham: Springer International Publishing, pp. 450–454. DOI: 10.1007/978-3-031-09917-5_34.

Lush, V., L. Bastin, K. Otsu, and J. Masó (2024). Assessing FAIRness of citizen science data in the context of the Green Deal Data Space. In: *International Journal of Digital Earth* 17.1, p. 2344587. DOI: 10.1080/17538947.2024.2344587.

Ma, X., J. Ralph, J. Zhang, X. Que, A. Prabhu, S. M. Morrison, R. M. Hazen, L. Wyborn, and K. Lehnert (2023). OpenMindat: Open and FAIR mineralogy data from the Mindat database. In: *Geoscience Data Journal* 11.1, pp. 94–104. DOI: 10.1002/gdj3.204.

Machine Intelligence (2022). Revisiting code reusability. DOI: 10.1038/s42256-022-00554-9.

Mahoney, A. and S. Gearheard (2008). Handbook for community-based sea ice monitoring. en. DOI: 10.25607/OBP-855.

Malik, T. and I. Foster (2012). Addressing data access needs of the long-tail distribution of geoscientists. In: *2012 IEEE International Geoscience and Remote Sensing Symposium*, pp. 5348–5351. DOI: 10.1109/IGARSS.2012.6352399.

Marine Environmental Data and Information Network **(MEDIN)** (2024). MEDIN guidelines. URL: https://medin.org.uk/data-standards/medin-data-guidelines.

Matentzoglu, N., J. Malone, C. Mungall, and R. Stevens (2018). MIRO: guidelines for minimum information for the reporting of an ontology. In: *Journal of Biomedical Semantics* 9.1. DOI: 10.1186/s13326-017-0172-7.

Mauthner, N. S. and O. Parry (2013). Open Access Digital Data Sharing: Principles, Policies and Practices. In: *Social Epistemology* 27.1, pp. 47–67. DOI: 10.1080/02691728.2012.760663.

Mazzocchi, F. (2018). Knowledge Organization System (KOS): An Introductory Critical Account. In: *Knowledge Organization* 45.1, pp. 54–78. DOI: 10.5771/0943-7444-2018-1-54.

McIlroy, M. D. (1968). Mass-Produced Software Components. In: *Software Engineering Concepts and Techniques (1968 NATO Conference of Software Engineering)*. Ed. by J. M. Buxton, P. Naur, and B. Randell. Accessed: 2024-11-02. NATO Science Committee, pp. 88–98. URL: https://www.cs.dartmouth.edu/~doug/components.txt.

Meiners, K. M. (2019). Sea-ice core and under-ice optical measurements from ice stations conducted during RV Polarstern PS117 voyage. AADC [data set]. DOI: 10.26179/5D9AC6A8CECC6.

Meng, C. (2016). Quantifying the impacts of snow on surface energy balance through assimilating snow cover fraction and snow depth. In: *Meteorology and Atmospheric Physics* 129.5, pp. 529–538. DOI: 10.1007/s00703-016-0486-5.

Met Office (2010). Cartopy: a cartographic Python library with a Matplotlib interface. Exeter, Devon. URL: https://scitools.org.uk/cartopy.

Middag, R. (2009). Dissolved Mn in the Southern Ocean. NIOZ Royal Netherlands Institute for Sea Research [data set]. URL: https://npdc.nl/dataset/d56083b3-6f4e-501c-bb65-51ee65e547b5.

Mieruch, S., I. Linck Rosenhaim, and R. Schlitzer (2023). The MOSAiC webODV: Interactive online data exploration, visualization and analysis. In: *EGU23, the 25th EGU General Assembly, held*

23-28 April, 2023 in Vienna, Austria and Online. Copernicus GmbH. DOI: 10.5194/egusphere-egu23-13807.

Miller, L. A. et al. (2015). Methods for biogeochemical studies of sea ice: The state of the art, caveats, and recommendations. In: *Elementa: Science of the Anthropocene* 3. DOI: 10.12952/journal.elementa.000038.

Millero, F. J., R. Feistel, D. G. Wright, and T. J. McDougall (2008). The composition of Standard Seawater and the definition of the Reference-Composition Salinity Scale. In: *Deep Sea Research Part I: Oceanographic Research Papers* 55.1, pp. 50–72. DOI: 10.1016/j.dsr.2007.10.001.

Mons, B., C. Neylon, J. Velterop, M. Dumontier, L. O. B. da Silva Santos, and M. D. Wilkinson (2017). Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. In: *Information Services & Use* 37.1, pp. 49–56. DOI: 10.3233/isu-170824.

Mons, B., E. Schultes, F. Liu, and A. Jacobsen (2020). The FAIR Principles: First Generation Implementation Choices and Challenges. In: *Data Intelligence* 2.1–2, pp. 1–9. DOI: 10.1162/dint_e_00023.

Morland, L., R. Kelly, and E. Morris (1990). A mixture theory for a phase-changing snowpack. In: *Cold Regions Science and Technology* 17.3, pp. 271–285. DOI: 10.1016/s0165-232x(05)80006-0.

Moser, M. N., J. M. Werheid, T. Hamann, A. Abdelrazeq, and R. H. Schmitt (2023). Which FAIR are you? In: *1st Conference on Research Data Infrastructure (CoRDI) - Connecting Communities : 12 – 14 September 2023, Karlsruhe (Germany).* Ed. by Y. Sure-Vetter and C. Goble. Vol. 1. DOI: 10.52825/cordi.v1i.401.

Mundy, C. J., M. Gosselin, Y. Gratton, K. A. Brown, V. Galindo, K. Campbell, M. Levasseur, D. Barber, T. N. Papakyriakou, and S. Bélanger (2010). Sea ice chemistry of Arctic-ICE2010. PANGAEA [data set]. DOI: 10.1594/PANGAEA.845798.

Murray-Rust, P. (2008). Open Data in Science. In: *Nature Precedings.* DOI: 10.1038/npre.2008.1526.1.

Nakayama, K., M. Shibata, S. Yazawa, Y. Kobayashi, M. Maekawa, and T. Okamoto (2006). Sharing and Searching History: Collaborative Component and Data Reuse. In: *2006 International Symposium on Communications and Information Technologies*, pp. 580–583. DOI: 10.1109/ISCIT.2006.340015.

National Academies of Sciences, Engineering, and Medicine **(NASEM)** (2019). *Reproducibility and Replicability in Science.* National Academies Press. DOI: 10.17226/25303.

National Aeronautics and Space Administration **(NASA)** (2015). Directory Interchange Format Schemas. NASA Earth Data. Bitbucket. Version 10.0. Accessed: 2024-10-24. URL: https://git.earthdata.nasa.gov/projects/EMFD/repos/dif-schemas/browse/10.x.

– (2021). Unified Metadata Model. NASA Earth Data. Bitbucket. Accessed: 2024-10-24. URL: https://git.earthdata.nasa.gov/projects/EMFD/repos/unified-metadata-model/browse.

National Centers for Environmental Information **(NCEI)** (n.d.). Discovery Tools. Accessed: 2024-10-22. URL: https://www.ncei.noaa.gov/access.

National Environmental Research Council **(NERC)** (n.d.). Data Catalogue Service. Accessed: 2024-10-22. URL: https://data-search.nerc.ac.uk/geonetwork/srv/eng/catalog.search.

Nature Computational Science **(NCS)** (2021). But is the code (re)usable? In: *Nature Computational Science* 1.7, pp. 449–449. DOI: 10.1038/s43588-021-00109-9.

Naveed, H., A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian (2025). A Comprehensive Overview of Large Language Models. In: *ACM Trans. Intell. Syst. Technol.* DOI: 10.1145/3744746.

Netherlands eScience Center **(NLeSC)** (2023). GitHub. Accessed: 2024-09-26. URL: https://github.com/NLeSC/awesome-research-software-registries/tree/main.

Netherlands eScience Center and Dutch national centre of expertise and repository for research data **(NLeSC and DANS)** (2020). Five Recommendations for FAIR Software. Accessed: 2024-10-20. URL: https://fair-software.nl/.

Netherlands Polar Data Center **(NPDC)** (n.d.). Netherlands Polar Data Center. Accessed: 2024-10-22. URL: https://npdc.nl/.

Nicolaus, M., S. Hendricks, M. Hoppmann, R. Ricker, C. Katlein, and P. A. Hunkeler (2012a). Salinity measured on sea ice cores during POLARSTERN cruise ARK-XXVI/3 (TransArc). PANGAEA [data set]. DOI: 10.1594/PANGAEA.773276.

– (2012b). Temperature measured on sea ice cores during POLARSTERN cruise ARK-XXVI/3 (TransArc). PANGAEA [data set]. DOI: 10.1594/PANGAEA.773277.

Northern California Earthquake Data Center **(NCEDC)** (2014). Northern California Earthquake Data Center. Accessed: 2024-10-20. DOI: 10.7932/NCEDC.

Norwegian Polar Institute **(NPI)** (n.d.). Norwegian Polar Data Centre. Accessed: 2024-10-22. URL: https://data.npolar.no/about.

Notz, D. and J. Stroeve (2016). Observed Arctic sea-ice loss directly follows anthropogenic CO 2 emission. In: *Science* 354.6313, pp. 747–750. DOI: 10.1126/science.aag2345.

Nusser, S. M., J. E. Cutcher-Gershenfeld, A. M. Mikytuck, and G. Korkmaz (2021). Fostering Data Reusability: Increasing Impact and Ease in Sharing and Reusing Research Data - Workshop Report and Action Steps. en. DOI: 10.5281/ZENODO.5390123.

Oggier, M. (2019). RSOI: Sea ice properties collected during the detection of oil on-in-and-under ice experiment. Zenodo [data set]. DOI: 10.5281/ZENODO.3237873.

Omatuku Ngongo, E., R. R. Audh, B. Hall, S. Skatulla, K. MacHutchon, T. Rampai, and M. Vichi (2022). Sea ice core temperature and salinity data collected during the 2019 SCALE Winter Cruise. Zenodo [data set]. DOI: 10.5281/zenodo.6997448.

Ontology Engineering Group **(OEG)** (2024). Linked Open Vocabularies. URL: https://lov.linkeddata.es/.

Open Geospatial Consortium **OGC** (n.d.). WaterML. Accessed: 2024-10-23. URL: https://www.ogc.org/publications/standard/waterml/.

Open Source Security Foundation **(OpenSSF)** (2024). OpenSSF Best Practices Badge Program. Accessed: 2024-10-20. URL: https://www.bestpractices.dev/.

OpenAIRE (2023). How to find a trustworthy repository for your data. Accessed: 2024-10-20. URL: https://www.openaire.eu/find-trustworthy-data-repository.

ORCID (n.d.). URL: https://orcid.org/.

Oxford English Dictionary **(OED)** (2023a). reusable (adj.) Accessed 2024-10-02. DOI: 10.1093/OED/8560598864.

– (2023b). reuse (n.) Accessed 2024-10-02. DOI: 10.1093/OED/5298793648.

– (2024). use (n.), sense I.1.a. Accessed 2024-10-02. DOI: 10.1093/OED/4502352485.

Pagano, A. M., G. M. Durner, T. C. Atwood, and D. C. Douglas (2021). Effects of sea ice decline and summer land use on polar bear home range size in the Beaufort Sea. In: *Ecosphere* 12.10. DOI: 10.1002/ecs2.3768.

Pampel, H. and S. Dallmeier-Tiessen (2014). Open Research Data: From Vision to Practice. In: *Opening Science: The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing*. Ed. by S. Bartling and S. Friesike. Cham: Springer International Publishing, pp. 213–224. DOI: 10.1007/978-3-319-00026-8_14.

Pampel, H. et al. (2023). re3data – Indexing the Global Research Data Repository Landscape Since 2012. In: *Scientific Data* 10.1. DOI: 10.1038/s41597-023-02462-y.

PANGAEA (2020). Metadata. PANGAEA wiki. URL: https://wiki.PANGAEA.de/wiki/Metadata.

– (2023a). Format. PANGAEA wiki. URL: https://wiki.PANGAEA.de/wiki/Format.

– (2023b). Geocode. PANGAEA wiki. URL: https://wiki.PANGAEA.de/wiki/Geocode.

– (2024). Data submission. PANGAEA wiki. URL: https://wiki.PANGAEA.de/wiki/Data%5C_submission.

Park, H., A. N. Fedorov, M. N. Zheleznyak, P. Y. Konstantinov, and J. E. Walsh (2014). Effect of snow cover on pan-Arctic permafrost thermal regimes. In: *Climate Dynamics* 44.9–10, pp. 2873–2895. DOI: 10.1007/s00382-014-2356-5.

Pasquetto, I. V., C. L. Borgman, and M. F. Wofford (2019). Uses and Reuses of Scientific Data: The Data Creators' Advantage. In: *Harvard Data Science Review* 1.2. DOI: 10.1162/99608f92.fc14bf2d.

Pasquetto, I. V., B. M. Randles, and C. L. Borgman (2017). On the Reuse of Scientific Data. In: *Data Science Journal* 16, p. 8. DOI: 10.5334/dsj-2017-008.

Paterson, W. S. B. (1994). *The Physics of Glaciers*. Pergamon international library of science, technology, engineering and social studies. Elsevier. ISBN: 9780080379449.

Pedersen, C. (2013). Zeppelin Webcamera Time Series. npolar.no [data set]. DOI: 10.21334/NPOLAR.2013.9FD6DAE0.

Peeken, I., S. Primpke, B. Beyer, J. Guetermann, C. Katlein, T. Krumpen, M. Bergmann, L. Hehemann, and G. Gerdts (2018a). Microplastic and environmental data from Arctic sea ice. PANGAEA [data set]. DOI: 10.1594/PANGAEA.886593.

Peeken, I., S. Primpke, B. Beyer, J. Gütermann, C. Katlein, T. Krumpen, M. Bergmann, L. Hehemann, and G. Gerdts (2018b). Arctic sea ice is an important temporal sink and means of transport for microplastic. In: *Nature Communications* 9.1. DOI: {10.1038/s41467-018-03825-5}.

Peng, G., D. Smith, S. Wingo, and R. Ramachandran (2021a). Stewardship Best Practices for Improved Discovery and Reuse of Heterogeneous and Cross-Disciplinary Earth System Data. In: *AGU Fall Meeting 2021, 13–17 Dec 2021*. DOI: 10.1002/essoar.10509377.1.

Peng, G. et al. (2021b). Call to Action for Global Access to and Harmonization of Quality Information of Individual Earth Science Datasets. In: *Data Science Journal* 20. DOI: 10.5334/dsj-2021-019.

Peng, R. D. (2011). Reproducible Research in Computational Science. In: *Science* 334.6060, pp. 1226–1227. DOI: 10.1126/science.1213847.

Phenotype And Trait Ontology **(PTO)** (n.d.). conductivity. URL: http://purl.obolibrary.org/obo/PATO_0001585.

Pinzer, B. R., M. Schneebeli, and T. U. Kaempfer (2012). Vapor flux and recrystallization during dry snow metamorphism under a steady temperature gradient as observed by time-lapse micro-tomography. In: *The Cryosphere* 6.5, pp. 1141–1155. DOI: 10.5194/tc-6-1141-2012.

Plesser, H. E. (2018). Reproducibility vs. Replicability: A Brief History of a Confused Terminology. In: *Frontiers in Neuroinformatics* 11. DOI: 10.3389/fninf.2017.00076.

Pronk, T. E. (2019). The Time Efficiency Gain in Sharing and Reuse of Research Data. In: *Data Science Journal* 18. DOI: 10.5334/dsj-2019-010.

Publications Office of the European Union **(EU)** (2021). Data.europa.eu data quality guidelines. DOI: 10.2830/333095.

Pućko, M., G. A. Stern, D. G. Barber, R. W. Macdonald, and B. Rosenberg (2010a). The international polar year (IPY) circumpolar flaw lead (CFL) system study: The importance of brine processes for α- and γ-hexachlorocyclohexane (HCH) accumulation or rejection in sea ice. In: *Atmosphere-Ocean* 48.4, pp. 244–262. DOI: 10.3137/oc318.2010.

Pućko, M., G. A. Stern, R. W. Macdonald, B. Rosenberg, and D. G. Barber (2011a). The influence of the atmosphere-snow-ice-ocean interactions on the levels of hexachlorocyclohexanes in the Arctic cryosphere. In: *Journal of Geophysical Research* 116.C2. DOI: 10.1029/2010JC006614.

Pućko, M., G. A. Stern, D. G. Barber, R. W. Macdonald, and B. Rosenberg (2010b). (Table 1) Physical properties, and alpha- and gamma-Hexachlorocyclohexane concentrations of sea-ice and water samples, eastern Beaufort Sea. PANGAEA [data set]. DOI: 10.1594/PANGAEA.818523.

Pućko, M., G. A. Stern, R. W. Macdonald, B. Rosenberg, and D. G. Barber (2011b). (Table 4) Physical properties in different layers of sea-ice, Beaufort Sea. PANGAEA [data set]. DOI: 10.1594/PANGAEA.818647.

Purich, A. and E. W. Doddridge (2023). Record low Antarctic sea ice coverage indicates a new sea ice state. In: *Communications Earth & Environment* 4.1. DOI: 10.1038/s43247-023-00961-9.

Python Packaging Authority **(PPA)** (2024). Packaging Python Projects. Accessed: 2024-10-23. URL: https://packaging.python.org/en/latest/tutorials/packaging-projects/.

Registry of Research Data Repositories **(RE3)** (2013a). GitHub. en. DOI: 10.17616/R3559G.

– (2013b). Home Re3data.org. Accessed: 2024-09-16. DOI: 10.17616/R3D.

– (2021). Re3data.org: Australian Antarctic Data Centre. Accessed: 2024-09-12. DOI: 10.17616/R3NP43.

– (2023a). Re3data.org: PANGAEA. Accessed: 2024-09-12. DOI: 10.17616/R3XS37.

– (2023b). Re3data.org: Zenodo. Accessed: 2024-09-12. DOI: 10.17616/R3QP53.

Rossi, L., S. Sporta Caputi, E. Calizza, G. Careddu, M. Oliverio, S. Schiaparelli, and M. L. Costantini (2019). Antarctic food web architecture under varying dynamics of sea ice cover. In: *Scientific Reports* 9.1. DOI: 10.1038/s41598-019-48245-7.

Sandström, M., A. Lister, and S.-A. Sansone (2023). FAIRsharing content: standards overview. Zenodo. DOI: 10.5281/ZENODO.8186982.

Sandt, S. van de, S. Dallmeier-Tiessen, A. Lavasa, and V. Petras (2019). The Definition of Reuse. In: *Data Science Journal* 18. DOI: 10.5334/dsj-2019-022.

Sandven, S., G. Spreen, G. Heygster, F. Girard-Ardhuin, S. L. Farrell, W. Dierking, and R. A. Allard (2023). Sea Ice Remote Sensing—Recent Developments in Methods and Climate Data Sets. In: *Surveys in Geophysics* 44.5, pp. 1653–1689. DOI: 10.1007/s10712-023-09781-0.

Sansone, S.-A., P. McQuilton, P. Rocca-Serra, A. Gonzalez-Beltran, M. Izzo, A. L. Lister, and M. Thurston (2019). FAIRsharing as a community approach to standards, repositories and policies. In: *Nature Biotechnology* 37.4, pp. 358–367. DOI: 10.1038/s41587-019-0080-8.

Santos, A., E. H. M. Pena, R. Lopez, and J. Freire (2025). Interactive Data Harmonization with LLM Agents. arXiv [preprint]. DOI: 10.48550/arXiv.2502.07132.

Schauer, U. (2012). Summary and Itinerary. In: *The Expedition of the Research Vessel Polarstern to the Arctic in 2011 (ARK-XXVI/3 - TransArc)*. Ed. by U. Schauer. Vol. 649. Bremerhaven, Germany: Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research. DOI: 10.2312/BzPM_0649_2012.

Schema.org (2024). Schema.org. Accessed: 2024-10-22. URL: https://schema.org/.

Schleef, S., H. Löwe, and M. Schneebeli (2014). Influence of stress, temperature and crystal morphology on isothermal densification and specific surface area decrease of new snow. In: *The Cryosphere* 8.5, pp. 1825–1838. DOI: 10.5194/tc-8-1825-2014.

Schöber, J., A. Klebelsberg, P. Schattan, K. Helfricht, and C. Fey (2019). Snow density in snow pits around Weisssee snow research site (2014-2018). PANGAEA [data set]. DOI: 10.1594/PANGAEA.898226.

Schürholt, K., J. Kowalski, and H. Löwe (2022). Elements of future snowpack modeling – Part 1: A physical instability arising from the nonlinear coupling of transport and phase changes. In: *The Cryosphere* 16.3, pp. 903–923. DOI: 10.5194/tc-16-903-2022.

Schweizer, J., J. B. Jamieson, and M. Schneebeli (2003). Snow avalanche formation. In: *Reviews of Geophysics* 41.4, p. 1016. DOI: 10.1029/2002rg000123.

Semantic Web for Earth and Environmental Terminology **(SWEET)** (2022). SWEET ontologies. GitHub [software]. Version v3.5.0. Accessed: 2024-10-24. URL: https://github.com/ESIPFed/sweet.

Shcherbina, A. Y., L. D. Talley, and D. L. Rudnick (2003). Direct Observations of North Pacific Ventilation: Brine Rejection in the Okhotsk Sea. In: *Science* 302.5652, pp. 1952–1955. DOI: 10.1126/science.1088692.

Simmonds, M. B. et al. (2022). Guidelines for Publicly Archiving Terrestrial Model Data to Enhance Usability, Intercomparison, and Synthesis. In: *Data Science Journal* 21.1, p. 3. DOI: 10.5334/dsj-2022-003.

Simson, A., H. Löwe, and J. Kowalski (2021). Elements of future snowpack modeling – Part 2: A modular and extendable Eulerian–Lagrangian numerical scheme for coupled transport, phase changes and settling processes. In: *The Cryosphere* 15.12, pp. 5423–5445. DOI: 10.5194/tc-15-5423-2021.

Simson, A., Q. Chen, M. S. Boxberg, and J. Kowalski (2025a). Pyresice Python package (v0.1.1). Zenodo [software]. DOI: 10.5281/zenodo.11198658.

Simson, A. and J. Kowalski (2021a). Eulerian_Lagrangian_snow_solver. GitHub [software]. URL: https://github.com/geo-fluid-dynamics/Eulerian_Lagrangian_snow_solver.

– (2021b). geo-fluid-dynamics/Eulerian_Lagrangian_snow_solver: final paper submission TC. Zenodo [software]. DOI: 10.5281/ZENODO.5588307.

– (2025a). RESICE - Reusability-targeted Enriched Sea Ice Core Database - General Information (v3). Zenodo [data set]. DOI: 10.5281/zenodo.10866346.

Simson, A. and J. Kowalski (2025b). RESICE - Reusability-targeted Enriched Sea Ice Core Database - Part A (v3). Zenodo [data set]. DOI: 10.5281/zenodo.10866363.

– (2025c). RESICE - Reusability-targeted Enriched Sea Ice Core Database - Part B (v3). Zenodo [data set]. DOI: 10.5281/zenodo.10866371.

Simson, A., A. Yildiz, and J. Kowalski (2025b). RESICE - Reusability-targeted Enriched Sea Ice Core Database - Interactive data2source Traceability (v3). Zenodo [figure]. DOI: 10.5281/zenodo.10866408.

– (2025c). Reusability-targeted enrichment of sea ice core data. DOI: 10.1038/s41597-025-04665-x.

Skatulla, S. et al. (2022). Physical and mechanical properties of winter first-year ice in the Antarctic marginal ice zone along the Good Hope Line. In: *The Cryosphere* 16.7, pp. 2899–2925. DOI: 10.5194/tc-16-2899-2022.

Sommerville, I. (2016). Software engineering. In: Tenth edition. Always learning. Boston: Pearson. Chap. Software reuse, pp. 437–453. ISBN: 9781292096148.

Spandre, P., H. François, D. Verfaillie, M. Lafaysse, M. Déqué, N. Eckert, E. George, and S. Morin (2019). Climate controls on snow reliability in French Alps ski resorts. In: *Scientific Reports* 9.1. DOI: 10.1038/s41598-019-44068-8.

Stroeve, J. and D. Notz (2018). Changing state of Arctic sea ice across all seasons. In: *Environmental Research Letters* 13.10, p. 103001. DOI: 10.1088/1748-9326/aade56.

Sumata, H., L. de Steur, D. V. Divine, M. A. Granskog, and S. Gerland (2023). Regime shift in Arctic Ocean sea ice thickness. In: *Nature* 615.7952, pp. 443–449. DOI: 10.1038/s41586-022-05686-x.

Sun, C., V. Emonet, and M. Dumontier (2022). A Comprehensive Comparison of Automated FAIRness Evaluation Tools. In: *Semantic Web Applications and Tools for Health Care and Life Sciences. 13th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences - Online, Leiden, Netherlands*. URL: https://cris.maastrichtuniversity.nl/files/151906569/Dumontier-2022-A_comprehensive_comparison_of_automated.pdf.

Sun, G. and C. S. Khoo (2017). Social science research data curation: issues of reuse. In: *Libellarium: journal for the research of writing, books, and cultural heritage institutions* 9.2, pp. 59–80. DOI: 10.15291/libellarium.v9i2.291.

Sysselmannen (2016). Jaktforbudssoner Svalbard. npolar.no [data set]. DOI: 10.21334/NPOLAR.2016.7C5EE27D.

Szabo, V. and V. R. Strang (1997). Secondary Analysis of Qualitative Data. In: *Advances in Nursing Science* 20.2, pp. 66–74. DOI: 10.1097/00012272-199712000-00008.

Tenopir, C., L. Christian, S. Allard, and J. Borycz (2018). Research Data Sharing: Practices and Attitudes of Geophysicists. In: *Earth and Space Science* 5.12, pp. 891–902. DOI: 10.1029/2018EA000461.

Tenopir, C., N. M. Rice, S. Allard, L. Baird, J. Borycz, L. Christian, B. Grant, R. Olendorf, and R. J. Sandusky (2020). Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. In: *PLOS ONE* 15.3. Ed. by S. Lozano, e0229003. DOI: 10.1371/journal.pone.0229003.

Testo (2024). Testo 720 - Temperature meter. [Manual]. Titisee-Neustadt, Germany. URL: https://static.testo.com/image/upload/Instruction-manual-and-Software/Instruction-manuals/testo-720-instruction-manual-7808.pdf.

Thanos, C. (2017). Research Data Reusability: Conceptual Foundations, Barriers and Enabling Technologies. In: *Publications* 5.1, p. 2. DOI: 10.3390/publications5010002.

The pandas development team **(Pandas team)** (2020). pandas-dev/pandas: Pandas. Version latest. DOI: 10.5281/zenodo.3509134. URL: https://doi.org/10.5281/zenodo.3509134.

Thompson, M., K. Burger, R. Kaliyaperumal, M. Roos, and L. O. B. da Silva Santos (2020). Making FAIR Easy with FAIR Tools: From Creolization to Convergence. In: *Data Intelligence* 2.1-2, pp. 87–95. DOI: 10.1162/dint_a_00031.

Tijm-Reijmer, C. H. (2010). Seasonal and Inter-annual Fluctuations of Ice Velocities on Five Svalbard Glaciers. npolar.no [data set]. data set. URL: https://npdc.nl/dataset/c0158fbb-f642-5978-bd03-0fe233ddf4f9.

Torstensson, A., A. Fransson, K. Currie, A. Wulff, and M. Chierici (2018a). Microalgal photophysiology and macronutrient distribution in summer sea ice in the Amundsen and Ross Seas, Antarctica. In: *PLOS ONE* 13.4. Ed. by D. Lannuzel, e0195587. DOI: 10.1371/journal.pone.0195587.

Torstensson, A., A. Fransson, K. I. Currie, A. Wulff, and M. Chierici (2018b). Seawater carbonate chemistry and Microalgal photophysiology and macronutrient distribution in summer sea ice in the Amundsen and Ross Seas, Antarctica. PANGAEA [data set]. DOI: 10.1594/PANGAEA.924295.

Touzeau, A., A. Landais, S. Morin, L. Arnaud, and G. Picard (2018). Numerical experiments on vapor diffusion in polar snow and firn and its impact on isotopes using the multi-layer energy balance model Crocus in SURFEX v8.0. In: *Geoscientific Model Development* 11.6, pp. 2393–2418. DOI: 10.5194/gmd-11-2393-2018.

TPS (2024a). Conductivity and temperature kit Aqua-C. [Manual]. Brendale, Australia. URL: https://cdn.shopify.com/s/files/1/0552/9924/4191/files/Aqua_C_Manual.pdf.

– (2024b). Conductivity and temperature kit WP-84. [Manual]. Brendale, Australia. URL: https://cdn.shopify.com/s/files/1/0552/9924/4191/files/WP-84.pdf.

Trull, T. W., A. R. Bowie, P. van der Merwe, and D. Laannuzel (2011). SIPEX data iron site - Aurora Australis voyage Sept-Oct 2007. AADC [data set]. DOI: 10.4225/15/514A9368882AA.

UNESCO (2022). Records of the General Conference, 41st session, Paris, 9-24 November 2021, volume 1: Resolutions. English. Document code: 41 C/RESOLUTIONS VOL.1 + CORR., Catalog Number: 0000380399. Paris, France. URL: https://unesdoc.unesco.org/ark:/48223/pf0000380399.

van der Merwe, P., D. Lannuzel, A. R. Bowie, and K. M. Meiners (2011a). High temporal resolution observations of spring fast ice melt and seawater iron enrichment in East Antarctica. In: *Journal of Geophysical Research* 116.G3. DOI: 10.1029/2010jg001628.

van der Merwe, P., D. Lannuzel, A. Bowie, C. Mancuso Nichols, and K. Meiners (2011b). Iron fractionation in pack and fast ice in East Antarctica: Temporal decoupling between the release of dissolved and particulate iron during spring melt. In: *Deep Sea Research Part II: Topical Studies in Oceanography* 58.9–10, pp. 1222–1236. DOI: j.dsr2.2010.10.036.

van der Merwe, P., D. Lannuzel, C. M. Nichols, K. Meiners, P. Heil, L. Norman, D. Thomas, and A. Bowie (2009). Biogeochemical observations during the winter–spring transition in East Antarctic sea ice: Evidence of iron and exopolysaccharide controls. In: *Marine Chemistry* 115.3-4, pp. 163–175. DOI: 10.1016/j.marchem.2009.08.001.

Vasilevsky, N. A., J. Minnier, M. A. Haendel, and R. E. Champieux (2017). Reproducible and reusable research: are journal data sharing policies meeting the mark? In: *PeerJ* 5, e3208. DOI: 10.7717/peerj.3208.

Vey, G., W. Van Wychen, C. Verhey, P. Pulsifer, and E. LeDrew (2024). Polar Research Data Management: Understanding Technical Implementation and Policy Decisions in the Era of FAIR Data. In: *Library and Information Sciences in Arctic and Northern Studies*. Ed. by S. Acadia. Cham: Springer International Publishing, pp. 175–190. DOI: 10.1007/978-3-031-54715-7_8.

Viallon-Galinier, L., P. Hagenmuller, and M. Lafaysse (2020). Forcing and evaluating detailed snow cover models with stratigraphy observations. In: *Cold Regions Science and Technology* 180, p. 103163. DOI: 10.1016/j.coldregions.2020.103163.

Vihma, T. (2014). Effects of Arctic Sea Ice Decline on Weather and Climate: A Review. In: *Surveys in Geophysics* 35.5, pp. 1175–1214. DOI: 10.1007/s10712-014-9284-0.

Vionnet, V., E. Brun, S. Morin, A. Boone, S. Faroux, P. L. Moigne, E. Martin, and J.-M. Willemet (2012). The detailed snowpack scheme Crocus and its implementation in SURFEX v7.2. In: *Geoscientific Model Development* 5.3, pp. 773–791. DOI: 10.5194/gmd-5-773-2012.

W3C Semantic Web Standards **(W3C)** (2024). Resource Description Framework (RDF). Accessed: 2024-10-22. URL: https://www.w3.org/RDF/.

Wagner, S. J., C. Matek, S. Shetab Boushehri, M. Boxberg, L. Lamm, A. Sadafi, D. J. Winter, C. Marr, and T. Peng (2024). Built to Last? Reproducibility and Reusability of Deep Learning Algorithms in Computational Pathology. In: *Modern Pathology* 37.1, p. 100350. DOI: 10.1016/j.modpat.2023.100350.

Wang, Q., P. Lu, M. Leppäranta, B. Cheng, G. Zhang, and Z. Li (2020a). Physical Properties of Summer Sea Ice in the Pacific Sector of the Arctic During 2008–2018. In: *Journal of Geophysical Research: Oceans* 125.9. DOI: 10.1029/2020JC016371.

Wang, Q., P. Lu, M. Leppäranta, B. Cheng, and Z. Li (2020b). Physical properties of summer sea ice in the Pacific sector of the Arctic in 2008-2018. Zenodo [data set]. DOI: 10.5281/ZENODO.3779867.

Wever, N., C. Fierz, C. Mitterer, H. Hirashima, and M. Lehning (2014). Solving Richards Equation for snow improves snowpack meltwater runoff estimations in detailed multi-layer snowpack model. In: *The Cryosphere* 8.1, pp. 257–274. DOI: 10.5194/tc-8-257-2014.

Wever, N., C. V. Valero, and C. Fierz (2016). Assessing wet snow avalanche activity using detailed physics based snowpack simulations. In: *Geophysical Research Letters* 43.11, pp. 5732–5740. DOI: 10.1002/2016GL068428.

Whetzel, P. L., N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, and M. A. Musen (2011). BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. In: *Nucleic Acids Research* 39.suppl, W541–W545. DOI: 10.1093/nar/gkr469.

White House Office of Science and Technology Policy **(OSTP)** (2022). *Desirable Characteristics of Data Repositories for Federally Funded Research*. Tech. rep. Executive Office of the President of the United States. DOI: 10.5479/10088/113528.

Wiese, M. and M. Schneebeli (2017). Early-stage interaction between settlement and temperature-gradient metamorphism. In: *Journal of Glaciology* 63.240, pp. 652–662. DOI: 10.1017/jog.2017.31.

Wilkinson, M. D., S.-A. Sansone, G. Marjan, J. Nordling, R. Dennis, and D. Hecker (2022). FAIR Assessment Tools: Towards an "Apples to Apples" Comparisons. Zenodo [report]. DOI: 10.5281/ZENODO.7463421.

Wilkinson, M. D., S.-A. Sansone, E. Schultes, P. Doorn, L. O. Bonino da Silva Santos, and M. Dumontier (2018). A design framework and exemplar metrics for FAIRness. In: *Scientific Data* 5.1. DOI: 10.1038/sdata.2018.118.

Wilkinson, M. D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. In: *Scientific Data* 3.1. DOI: 10.1038/sdata.2016.18.

Wilkinson, M. D. et al. (2019). Evaluating FAIR maturity through a scalable, automated, community-governed framework. In: *Scientific Data* 6.1. DOI: 10.1038/s41597-019-0184-5.

Wingham, D. et al. (2006). CryoSat: A mission to determine the fluctuations in Earth's land and marine ice fields. In: *Advances in Space Research* 37.4. Natural Hazards and Oceanographic Processes from Satellite Data, pp. 841–871. DOI: 10.1016/j.asr.2005.07.027.

Wolf, M., J. Logan, K. Mehta, D. Jacobson, M. Cashman, A. M. Walker, G. Eisenhauer, P. Widener, and A. Cliff (2021). Reusability First: Toward FAIR Workflows. In: *2021 IEEE International Conference on Cluster Computing (CLUSTER)*. IEEE. DOI: 10.1109/cluster48925.2021.00053.

World Glacier Monitoring Service **(WGMS)** (2024). Fluctuations of Glaciers Database. DOI: 10.5904/WGMS-FOG-2024-01.

World Meteorological Organization **(WMO)** (2008). Guide to Hydrological Practices, Volume I. Hydrology – From Measurement to Hydrological Information. WMO e-Library. URL: https://library.wmo.int/idurl/4/35804.

– (2014). Sea ice nomenclature. WMO e-Library. URL: https://library.wmo.int/idurl/4/41953.

– (2023). Guide to Instruments and Methods of Observation Volume II. Measurement of Cryospheric Variables. WMO e-Library. URL: https://library.wmo.int/idurl/4/68660.

– (n.d.[a]). Climate. Accessed: 2024-11-04. URL: https://wmo.int/topics/climate.

– (n.d.[b]). Glossary of the Global Cryosphere Watch. Accessed: 2024-10-24. URL: https://globalcryospherewatch.org/reference/glossary.php.

– (n.d.[c]). WMO Meteorological codes. Accessed: 2025-05-27. URL: https://artefacts.ceda.ac.uk/badc_datadocs/surface/code.html.

Worster, M. G. and D. W. Rees Jones (2015). Sea-ice thermodynamics and brine drainage. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 373.2045, p. 20140166. DOI: 10.1098/rsta.2014.0166.

WTW (2004). Bedienungsanleitung – WTW Cond 315i. [Manual]. Weilheim, Germany. URL: https://www.labworld.at/wp-content/uploads/2014/10/Cond_315i.pdf.

– (2008a). Bedienungsanleitung – WTW Cond 3110. [Manual]. Weilheim, Germany. URL: https://www.labworld.at/wp-content/uploads/2014/10/Cond_3110.pdf.

– (2008b). Bedienungsanleitung – WTW Cond 3300i-3400i. [Manual]. Weilheim, Germany. URL: https://www.labworld.at/wp-content/uploads/2017/09/Bedienungsanleitung-WTW-Cond-3300i-3400i.pdf.

Wu, M.-W. and Y.-D. Lin (2001). Open source software development: an overview. In: *Computer* 34.6, pp. 33–38. DOI: 10.1109/2.928619.

Wyborn, L. (2023). One Geoscience: Providing FAIR global Access to all Geoscience Data - are we there yet? FID GEO [presentation]. DOI: 10.23689/FIDGEO-5815.

Yoon, A. (2016a). Data reusers' trust development. In: *Journal of the Association for Information Science and Technology* 68.4, pp. 946–956. DOI: 10.1002/asi.23730.

– (2016b). Red flags in data: Learning from failed data reuse experiences. In: *Proceedings of the Association for Information Science and Technology* 53.1, pp. 1–6. DOI: 10.1002/pra2.2016.14505301126.

– (2017). Role of communication in data reuse. In: *Diversity of Engagement: Connecting People and Information in the Physical and Virtual Worlds*. Ed. by S. Erdelez and N. Agarwal. Vol. 54. Associtaion for Information Science & Technology. Wiley, pp. 463–471. DOI: 10.1002/pra2.2017.14505401050.

Yu, W., K. Ito, and S. Matsubara (2025). Capabilities and Challenges of LLMs in Metadata Extraction from Scholarly Papers. In: *Sustainability and Empowerment in the Context of Digital Libraries*. Springer Nature Singapore, pp. 280–287.

Zeng, M. L. (2008). Knowledge Organization Systems (KOS). In: *Knowledge Organization* 35.2/3, pp. 160–182. DOI: 10.5771/0943-7444-2008-2-3-160.

Zenodo (n.d.[a]). General Policies. Version 1.0. Accessed: 2024-11-02. URL: https://about.zenodo.org/policies/.

– (n.d.[b]). Search guide. Accessed: 2024-11-02. URL: https://help.zenodo.org/guides/search/.

Zimmerman, A. S. (2008). New Knowledge from Old Data: The Role of Standards in the Sharing and Reuse of Ecological Data. In: *Science, Technology, & Human Values* 33.5, pp. 631–652. DOI: 10.1177/0162243907306704.

Appendix A

Supplements to FAIR metrics

A.1 Metrics for automated FAIR software assessment

TABLE A.1: Metrics for automated FAIR software assessment developed by FAIR-IMPACT (n.d.) and the respective FAIR4RS principle (Chue Hong et al., 2023) addressed with the metric

Metrics for automated FAIR software assessment	**FAIR4RS Principle**
FRMS-01: Does the software have a globally unique and persistent identifier?	F1 R3
FRMS-02: Do the different components of the software have their own identifiers?	F1 F1.1
FRMS-03: Does each version of the software have a unique identifier?	F1 F1.2 R3
FRSM-04: Does the software include descriptive metadata which helps define its purpose?	F2 R1 R3
FRSM-05: Does the software include development metadata which helps define its status?	F2 R1 R3
FRSM-06: Does the software include metadata about the contributors and their roles?	F2 R3
FRSM-07: Does the software metadata include the identifier for the software?	F3 R3
FRSM-08: Does the software have a publicly available, openly accessible and persistent metadata record?	F4 A2 R3
FRSM-09: Is the software developed in a code repository/forge that uses standard communication protocols?	A1.1 A1.2 R3
FRSM-10: Are the formats used by the data consumed or produced by the software open and a reference provided to the format?	I1 I2
FRSM-11: Does the software use open APIs that support machine-readable interface definition?	I1
FRSM-12: Does the software provide references to other resources that support the use?	I2
FRSM-13: Does the software describe what is required to use it?	R1 R2
FRSM-14: Does the software come with test cases to demonstrate it is working?	R1
FRSM-15: Does the software source code include licensing information for the software and any bundled external software?	R1.1
FRSM-16: Does the software metadata record include licensing information?	R1.1
FRSM-17: Does the software include provenance information that describe the development of the software?	R1.2

Appendix B

Supplements to Cryospheric Case Study I

B.1 Formulas required for the process model

B.1.1 Water vapor saturation density

An empirical expression for the water vapor saturation density $\rho_v^{eq}(T)$ in terms of temperature T is formulated based on the empirical formulation for vapor saturation pressure from Libbrecht, 1999 and reads

$$\rho_v^{eq}(T) = \frac{exp(-\frac{T_{ref}}{T})}{f\,T}(a_0 + a_1(T - T_m) + a_2(T - T_m)^2) \tag{B.1}$$

with coefficients $a_0 = 3.6636 \times 10^{12}$ kgm^{-1}s^{-2}, $a_1 = -1.3086 \times 10^8$ kgm^{-1}s^{-2}K^{-1}, $a_2 = -3.3793 \times 10^6$ kgm^{-1}s^{-2}K^{-2}, $f = 461.31$ J kg^{-1}K^{-1}, $T_m = 273.15$ K $T_{ref} = 6150$ K. f is the specific gas constant for water vapor. Note that division by fT account for the conversion from pressure Pa (as used in Part 1) to density kgm^{-3}.

B.1.2 Model parameters in the heat and water vapor transport equations

The effective vapor mass diffusion coefficient $D_{eff}(\phi_i)$ in terms of ice volume fraction ϕ_i is taken from Calonne et al., 2011, but is extended by the Heaviside function Θ to hinder vapor diffusion for ice volumes above $\frac{2}{3}$

$$D_{eff}(\phi_i) = D_0(1 - \frac{3}{2}\,\phi_i)\,\Theta\left(\frac{2}{3} - \phi_i\right), \tag{B.2}$$

with $D_0 = 2.036 \times 10^{-5}$ m^{2}s^{-1} the vapor diffusion constant in air.
The effective thermal conductivity $k_{eff}(\phi_i)$ in terms of ice volume fraction ϕ_i is taken from Calonne et al., 2011 and reads

$$k_{eff}(\phi_i) = a_0 + a_1(\phi_i\,\rho_i) + a_2(\rho_i\,\phi_i)^2 \tag{B.3}$$

with coefficients $a_0 = 0.024$, $a_1 = -1.23 \times 10^{-4}$ and $a_2 = 2.5 \times 10^{-6}$ and ice density ρ_i.
The effective heat capacity $(\rho C)_{eff}(\phi_i)$ in terms of ice volume fraction ϕ_i is taken from Calonne et al., 2014 and Hansen and Foslien, 2015a and reads

$$(\rho C)_{eff}(\phi_i) = \phi_i \, \rho_i \, C_i + (1 - \phi_i) \, \rho_a \, C_a, \tag{B.4}$$

with C_i ice heat capacity, C_a air heat capacity, ρ_i ice density and ρ_a air density.

B.1.3 Dynamic viscosity

The empirical closure for dynamic viscosity from Vionnet et al., 2012 is

$$\eta(\phi_i, T) = f \, \eta_0 \, \frac{\rho_i \, \phi_i}{c_\eta} exp \left(a_\eta \, (T_{ph} - T) + b_\eta \, \rho_i \, \phi_i \right), \tag{B.5}$$

with state variables temperature T and ice volume fraction ϕ_i and constants ice density ρ_i, phase change temperature $T_{ph} = 273$ K and further constants $\eta_0 = 7.62237 \times 10^6$ kg s^{-1}, $a_\eta = 0.1$ K^{-1}, $b_\eta = 0.023$ m^3 kg^{-1}, $c_\eta = 250$ kg m^{-2}. Finally, f reflects properties of the snow microstructure, i. e., the angularity and the size of the grains, and it and is assumed to be 1 in our case. The constant viscosity value applied to linear Glen's law $\eta_{const,m=1}$ is derived with intermediate values for ice volume fraction and temperature of the respective initial conditions. These values are plugged into the empirical closure Eq. (B.5) to solve for viscosity. The same procedure cannot be applied to derive a constant viscosity value for the non-linear version of Glen's law $\eta_{const,m=3}$ since the viscosity closure (Eq. (B.5)) has initially been calibrated to the linear form of Glen's law ($m = 1$). Instead we choose a snow deformation rate from the literature ($\dot{\epsilon} = 10^{-6}$ s^{-1} (Johnson, 2011)) and determine the maximum stress value from the initial snow density. These strain rate and stress values are then inserted into the constitutive relation (Eq. (4.3)) which is finally solved for viscosity. To avoid infinite ice volume growth above physical values ($\phi_i > 1$) the viscosity must tend to infinity for $\phi_i \to 1$. Therefore, the constant viscosity values are restricted to ice volumes below 0.95 by multiplication with an ice volume fraction dependent power law (Appendix (B.7)). This power law yields ≈ 1 for $\phi_i \leq 0.95$ and exponentially increases for higher ice volumes. Multiplied with the constant viscosity values viscosity remains constant below $\phi_i < 0.95$ and exponentially increases above, which stops further densification and settling. This procedure does not intend to reproduce the correct physics for low porosity ice, which mathematically leads though to a similar crossover behavior.

B.1.4 Constant viscosity for the two-layer case

B.1.4.1 Linear Glen's law, $\eta_{const,m=1}$

We derived intermediate ice volume fraction $\phi_{i,const} = 0.1125$ and temperature $T_{const} = 263$ K values from the initial condition of the two-layer case and insert them as constants into Eq. (B.5).

B.1.4.2 Non-linear Glen's law $\eta_{const,m=3}$

Equation (B.5) does not hold for Glen exponent $m = 3$, therefore we derive an adjusted constant viscosity $\eta_{const,m=3}$ via the constitutive equation (Eq. (4.3))

$$\dot{\epsilon}_{lit} = \frac{1}{\eta_{const,m=3}} \sigma_{max}^{m}, \tag{B.6}$$

with $\dot{\epsilon}_{lit} \equiv 10^{-6}$ s^{-1} a strain rate value from the literature Johnson, 2011 and $\sigma_{max} \equiv 547.71$ Pa the maximum stress value obtained from the initial snow density profile of the two-layer case. Eq. (B.6) is then solved for the constant viscosity $\eta_{const,m=3}$.

B.1.4.3 Restrict infinite ice volume growth

To hinder infinite ice volume growth, the constant viscosity $\eta_{const,m}$ is combined with a power law that yields exponential growth of viscosity for cells with $\phi_i > 0.95$

$$PL(\phi_i) = exp\left(pl1\,\phi_i - pl2\right) + 1, \tag{B.7}$$

with $pl1 = 690$ and $pl2 = 650$. The constant viscosity is then multiplied with the power law ($\eta_{const,m}\, PL(\phi_i)$), so that computational nodes with $\phi_i > 0.95$ are assigned and viscosity grows exponentially. Note that for better readability the multiplication with the power law is omitted in the equations of this paper.

B.2 Higher-order mesh errors to correct for non-uniform mesh

For the temperature equation Eq. (4.21) the higher-order mesh error is

$$E_T(T_{k+1}^{n+1}, T_{k-1}^{n+1}) = \frac{\Delta t^n\, 2\, \beta_{T,k}}{(\Delta z_k^n)^2 + (\Delta z_{k-1}^n)^2} \frac{\Delta z_k^n - \Delta z_{k-1}^n}{\Delta z_k^n + \Delta z_{k-1}^n} \left(T_{k+1}^{n+1} + T_{k-1}^{n+1}\right), \tag{B.8}$$

and for the vapor transport equation Eq. (4.22) it is

$$E_c(T_{k+1}^{n+1}, T_{k-1}^{n+1}) = \frac{2\, \beta_{c,k}}{(\Delta z_k^n)^2 + (\Delta z_{k-1}^n)^2} \frac{\Delta z_k^n - \Delta z_{k-1}^n}{\Delta z_k^n + \Delta z_{k-1}^n} \left(T_{k+1}^{n+1} + T_{k-1}^{n+1}\right). \tag{B.9}$$

For the matrix equations Eqs. (4.24) and (4.23) the higher-order mesh errors are defined as $\mathbf{E}_T$ and $\mathbf{E}_C$.

B.3 Matrices from temperature and vapor transport equations

Matrix **A** is defined as follows:

$$\mathbf{A} = \begin{pmatrix} A^n_{m,0} & 0 & 0 & \cdots & 0 & 0 & 0 \\ A^n_{l,1} & A^n_{m,1} & A^n_{u,1} & \cdots & 0 & 0 & 0 \\ 0 & A^n_{l,2} & A^n_{m,2} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & A^n_{m,nz-2} & A^n_{u,nz-2} & 0 \\ 0 & 0 & 0 & \cdots & A^n_{l,nz-1} & A^n_{m,nz-1} & A^n_{u,nz-1} \\ 0 & 0 & 0 & \cdots & 0 & 0 & A^n_{m,nz} \end{pmatrix} \tag{B.10}$$

For the heat equation (Eq. (4.23)) ($\mathbf{A}_T$) the entries are

$$A^n_{l,k} = \Delta t^n \left(\frac{\beta^n_{T,k+1} - \beta^n_{T,k-1}}{(\Delta z^n_k + \Delta z^n_{k-1})^2} - D^n_{T,k}\right) \tag{B.11}$$

$$A^n_{u,k} = -\Delta t^n \left(\frac{\beta^n_{T,k+1} - \beta^n_{T,k-1}}{(\Delta z^n_k + \Delta z^n_{k-1})^2} + D^n_{T,k}\right) \tag{B.12}$$

$$A^n_{m,k} = \alpha^n_{T,k} + 2\,\Delta t^n D^n_{T,k} \text{ with } A^n_{m,0} = 1 \text{ and } A^n_{m,nz} = 1 \tag{B.13}$$

and for the vapor transport (Eq. (4.24)) ($\mathbf{A}_c$) the entries are

$$A^n_{l,k} = \frac{\beta^n_{k+1} - \beta^n_{k-1}}{(\Delta z^n_k + \Delta z^n_{k-1})^2} - D^n_{c,k} \tag{B.14}$$

$$A^n_{u,k} = -\frac{\beta^n_{k+1} - \beta^n_{k-1}}{(\Delta z^n_k + \Delta z^n_{k-1})^2} - D^n_{c,k} \tag{B.15}$$

$$A^n_{m,k} = \frac{\alpha^n_{c,k}}{\Delta t^n} + 2\,D^n_{c,k} \text{ with } A^n_{m,0} = -\frac{\alpha^n_{c,0}}{\Delta t^n} \text{ and } A^n_{m,nz} = -\frac{\alpha^n_{c,nz}}{\Delta t^n} \tag{B.16}$$

with

$$D^n_{f,k} = \frac{2\ \beta^n_{f,k}}{(\Delta z^n_k)^2 + (\Delta z^n_{k-1})^2} \,\text{for} f \in \{T, c\} \tag{B.17}$$

Matrix **B** is defined as follows:

$$\mathbf{B} = \begin{pmatrix} \alpha^n_{m,0} & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \alpha^n_{m,1} & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & \alpha^n_{m,2} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \alpha^n_{m,nz-2} & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & \alpha^n_{m,nz-1} & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & \alpha^n_{m,nz} \end{pmatrix} \tag{B.18}$$

For Eq. (4.23) ($\mathbf{B}_T$) the entries are

$$\alpha^n_{m,k} = \alpha^n_{T,k} \tag{B.19}$$

and for Eq. (4.24) ($\mathbf{B}_c$)

$$\alpha^n_{m,k} = \frac{\alpha^n_{c,k}}{\Delta t^n}. \tag{B.20}$$

Matrix $\mathbf{E}$ is defined as follows

$$\mathbf{E} = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ E^n_{l,0} & 0 & E^n_{u,1} & \cdots & 0 & 0 & 0 \\ 0 & E^n_{l,1} & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & E^n_{u,nz-2} & 0 \\ 0 & 0 & 0 & \cdots & E^n_{l,nz-1} & 0 & E^n_{u,nz-1} \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 \end{pmatrix}, \tag{B.21}$$

consisting of the following terms for the heat equation ($\mathbf{E}_T$) (Eq. (4.23))

$$E^n_{l,k} = -\Delta t^n D^n_{T,k} \frac{\Delta z^n_k - \Delta z^n_{k-1}}{\Delta z^n_k + \Delta z^n_{k-1}} \tag{B.22}$$

$$E^n_{u,k} = \Delta t^n D^n_{T,k} \frac{\Delta z^n_k - \Delta z^n_{k-1}}{\Delta z^n_k + \Delta z^n_{k-1}} \tag{B.23}$$

and the vapor equation ($\mathbf{E}_c$) (Eq. (4.24))

$$E^n_{l,k} = -D^n_{c,k} \frac{\Delta z^n_k - \Delta z^n_{k-1}}{\Delta z^n_k + \Delta z^n_{k-1}} \tag{B.24}$$

$$E^n_{u,k} = D^n_{c,k} \frac{\Delta z^n_k - \Delta z^n_{k-1}}{\Delta z^n_k + \Delta z^n_{k-1}} \tag{B.25}$$

Note that $\beta = \beta_c$ for Eq. (4.24) and $\beta = \beta_T$ for Eq. (4.23), as explained in Sect. 4.2.2

Appendix C

Supplements to Cryospheric Case Study II

C.1 List of scope elements, YAML file fields, and tabular database columns

TABLE C.1: Scope elements with description and distinction into YAML file fields and tabular database columns. Elements in gray colored rows are automatically enriched. Naming convention follows that of the YAML files. Hyphens are replaced by underscores in the tabular database.

Field/Column	Unit	Description	YAML	Tabular
average-spatial-resolution... -salinity_sea-ice and -temperature_sea-ice	m	average distance between measurements of one core salinity sea ice and temperature sea ice		✓
depth	m	one depth value (instead of one depth value per measurement of sea ice temperature and salinity)		✓
development-stage-SIN... -level-1_sea-ice and -level-2_sea-ice		term from Development Section of SIN (WMO, 2014) sub-category 2.x and sub-sub-category 2.x.y		✓
latitude		in decimal degrees		✓
longitude		in decimal degrees		✓
mean... -salinity_sea-ice and -temperature_sea-ice	 ppt K	mean values of all measurements of salinity of sea ice and temperature of sea ice		✓
ocean-SeaVox		name of ocean from SeaVoX (BODC, 2023)		✓
sea-SeaVox		name of sea from SeaVoX (BODC, 2023)		✓
[*column name*]_adjusted		indicates adjustment of values from original sources		✓
[*column name*]_comment		excerpt/figure/table/inconsistencies/adjustments		✓
[*column name*]_doi		doi (sometimes url) of the source		✓
[*column name*]_source		name of the source as listed in Table C.2		✓
calculation-method- temperature_sea-ice		potential computational manipulations of the measurement data (e.g. interpolations)	✓	✓
campaign		name of campaign/expedition/project	✓	✓
date		date of core retrieval in YYYY-MM-DD format	✓	✓
development-stage_sea-ice		sea ice age classification	✓	✓
form_sea-ice		pack ice or drift ice or fast ice as called in the source	✓	✓
freeboard_sea-ice	m	sea ice thickness above water level	✓	✓
id		combination of campaign and core number	✓	✓
measurement-device-accuracy... -salinity_sea-ice, -temperature_air and -temperature_sea-ice	 ppt K K	accuracy of measurement device salinity sea ice, temperature air and temperature sea ice	✓	✓
measurement-device... -salinity_sea-ice, -temperature_air and -temperature_sea-ice		name of measurement device for salinity sea ice, temperature air and temperature sea ice	✓	✓
polar-region		Arctic or Antarctica	✓	✓
salinity_sea-ice	ppt	bulk salinity of sea ice in YAML combined with depth	✓	✓
temperature_air	K	temperature of air in YAML combined with depth	✓	✓
temperature_sea-ice	K	temperature of sea ice in YAML combined with depth	✓	✓
thickness_sea-ice	m	thickness of sea ice	✓	✓
thickness_snow	m	thickness of snow cover on top of sea ice	✓	✓
water-body		name of water body at the coring location	✓	✓
name		name of the core	✓	
coordinates		"N": 74.70933, "E": -95.24408 in decimal degrees	✓	
[*field name*]_option		in case of two sources providing the same field	✓	

C.2 List of original sources used in RESICE

TABLE C.2: Lists all original sources used in RESICE. Sources are ordered horizontally by source group (primary, secondary, tertiary) and vertically by repository from which the primary source is available. One row represents the combination of sources that provide fields to one or several YAML files, indicating they are from the same campaign. TThe column label *#YAML file* indicates the number of files originating from the respective combination of resources. Redundant data sets are neglected in RESICE. Secondary sources assigned an asterisk are referenced in the respective primary source.

Primary sources	**Secondary sources**	**Tertiary sources**	**#YAML files**
	PANGAEA		
Arndt et al. (2021b)	Arndt et al. (2021a)	WTW (2008a)	21
Katlein et al. (2020a)	Katlein et al. (2020b)*		1
Kramer et al. (2010e) Kramer et al. (2010d)	Kramer et al. (2011)* Haas et al. (2009)		22
Kramer et al. (2010b) Kramer et al. (2010c) Kramer et al. (2010a)	Kramer et al. (2011)*		12
Lange et al. (2015b)	Lange et al. (2015a)*	WTW (2008b)	18
Lannuzel (2016a)	Lannuzel et al. (2007) Lannuzel et al. (2016b)*	Testo (2024) TPS (2024b)	6
Lannuzel (2016b)	Lannuzel et al. (2016b) van der Merwe et al. (2011a)	Testo (2024) TPS (2024a)	5
Lannuzel (2016c)	Lannuzel et al. (2016b)* Lannuzel et al. (2008)		7
Lannuzel (2016d)	Lannuzel et al. (2016b) van der Merwe et al. (2009), van der Merwe et al. (2011b)	Testo (2024) TPS (2024a)	9
Lannuzel (2016e)	Lannuzel et al. (2016b)* Lannuzel et al. (2016a)	TPS (2024a) Testo (2024)Testo (2024)	6
Mundy et al. (2010)	Brown et al. (2015)*		23
Nicolaus et al. (2012a) Nicolaus et al. (2012b)	Hendricks et al. (2012) Schauer (2012)	WTW (2004)	11
Peeken et al. (2018a)	Peeken et al. (2018b) *		5
Pućko et al. (2010b) Pućko et al. (2011b) Isleifson et al. (2010a)	Pućko et al. (2010a)* Pućko et al. (2011a) Isleifson et al. (2010b)	Hach (2000) Control Company (2016)	16
Torstensson et al. (2018b)	Torstensson et al. (2018a)*		14
	Zenodo		
Audh et al. (2022)	Johnson et al. (2023)		21
Omatuku Ngongo et al. (2022)	Skatulla et al. (2022)*		15
Wang et al. (2020b)	Wang et al. (2020a)		41
	AADC		
Duprat (2019)	Duprat et al. (2019)		6
Lannuzel et al. (2017)			redundant
Meiners (2019)	Castellani et al. (2019)		28
Trull et al. (2011)			redundant