Informatik Spektrum (2022) 45:271–277
https://doi.org/10.1007/s00287-022-01489-6

HAUPTBEITRAG

Cross domain fusion for spatiotemporal applications: taking interdisciplinary, holistic research to the next level

Matthias Renz[1] · Peer Kröger[1] · Agnes Koschmider[1,2] · Olaf Landsiedel[1] · Nelson Tavares de Sousa[1]

Accepted: 10 July 2022 / Published online: 29 September 2022

Abstract
Exploiting the power of collective use of complementing data sources for the discovery of new correlations and findings offers enormous additional value compared to the summed values of isolated analysis of the individual information sources. In this article, we will introduce the concept of "cross domain fusion" (CDF) as a machine learning and pattern mining driven and multi-disciplinary research approach for fusing data and knowledge from a variety of sources enabling the discovery of answers of the question to be examined from a more complete picture. The article will give a basic introduction in this emerging field and will highlight examples of basic CDF tasks in the field of marine science.

Introduction

The increased availability of data and knowledge as well as the trend towards more interdisciplinary and transdisciplinary research approaches call for solutions to the paradigm of cross domain fusion (CDF). CDF is a data-driven and multi-disciplinary research approach that leads towards a holistic "big picture" supporting and integrating multiple scientific views. In our definition, *views* represent the various perspectives on a scene of an excerpt of the real world. Thereby, views vary among different disciplines, different data sets, data types, and data sources including models, and different abstraction of data ranging from measurement records over patterns up to knowledge representations.

For example, physicists and biologists have different perspectives on some scenes in the ocean, such as the behavior of saline concentration and behaviour of living organisms. Also within the same scientific discipline, there may be many varying views on entities of a general research topic—for example, oxygen deficits in coastal regions measured by remote sensing raster data vs. series of in situ sensor measurements. Since most scientific views are traditionally explored in isolation, they are also restricted to specific findings and often do not allow a broad understanding of relationships and functional dependencies between multiple concepts such as structural relationships among different scientific views.

Defining cross domain fusion

Cross domain fusion differs from traditional concepts for data fusion studied in the database community in terms of focusing on the fusion of data at higher data abstraction levels (including patterns and knowledge) rather than schema mapping and data merging conducted at the (lowest) data level.

Cross domain fusion (CDF) *is a systematic approach for data analysis techniques organically fusing data and knowledge from various data sources referring to different abstraction levels, including scientific models and user (stakeholder/expert) knowledge, through a (semi)automatic big data analytics pipeline and interactive exploration.*

✉ Matthias Renz
mr@informatik.uni-kiel.de

Peer Kröger
pkr@informatik.uni-kiel.de

Agnes Koschmider
agnes.koschmider@uni-bayreuth.de

Olaf Landsiedel
ol@informatik.uni-kiel.de

Nelson Tavares de Sousa
ntd@informatik.uni-kiel.de

[1] Department of Computer Science, Christian-Albrechts-Universität zu Kiel (Kiel University), Kiel, Germany

[2] University of Bayreuth, Bayreuth, Germany

Let us note that the focus here is on applications with spatial and spatiotemporal data (e.g., climate research, smart cities, marine science, archaeology, etc.); however, CDF is a general paradigm not limited to these disciplines.

Examples of cross domain fusion

CDF, for example, enables the consideration of the interplay between a scientific model, e.g., a physical ocean current model, and observational data sensed from the real world, e.g., ocean current sensor data.

Considering both views induced by the data sources, respectively, in an isolated way often results in a mismatch between the model and the real world. The reason is that models cannot cover all aspects of the real world while the sensed data are usually highly incomplete in space and time. The fusion of both views covering the interplay between both is the key to overcoming the two above-mentioned shortcomings.

Another example of CDF is the integration of multiple heterogeneous views of data, e.g., spatiotemporal data augmented with other, not necessarily spatial, context data including text, images, sensor measurements, etc. For example, ocean current driven trajectories of floating particles in combination of water mass attribute data, like saline concentration, temperature, etc. gives ocean scientists valuable new insights into water mass transportation in the ocean and helps to understand the dynamics and stability of the gulf stream. Again, the integration of the combined view on multiple different data sources into the data analysis procedure is the key to discovering more insights into a real world scene than from the single data sources alone.

Motivation and outline

Recent advances in exploiting the power of machine learning to fuse data/views that increase the potential of data analysis are summarized in [1]. While most approaches have been proposed in the field of urban analytics [2–4], it is increasingly attracting further scientific disciplines [5, 6]. However, most existing solutions are isolated, often too narrow, ad-hoc approaches designed for specific applications. We need more general, broader, and systematic fusion concepts, such as CDF aims for. CDF offers significant potential for fundamental research in Computer Science, e.g., it provides next generation methods for knowledge acquisition and data science. In this article, we sketch out why CDF as a new scientific paradigm would be beneficial for gaining scientific wisdom and discuss the key aspects of CDF from the Computer Science point of view. In particular, we describe typical sources of data and knowledge that needs to be fused. Furthermore, we categorize the basic algorithmic approaches and methods to implement CDF and sketch preliminary ideas and opportunities for CDF. Finally, we discuss technical challenges surrounding the execution of CDF methods. We will point out possible research challenges for Computer Science, particularly data science as the core research area.

Cross domain fusion: why (now)?

Over decades, scientific knowledge and wisdom have been obtained by transforming small scale observations of real-world phenomena into—mostly very specific—scientific models. Scientific models typically represent a small and isolated abstraction of the real world. The process of gaining wisdom from observations/data typically follows the four stages of the “DIKW” (data, information, knowledge, wisdom) pyramid [7] (see left side of Fig. 1). Scientific models may represent the upper stages, i.e., information (such as patterns in the observations), knowledge (e.g., in the form of scientific models), or wisdom. We use the term “domain” for the more general concept representing any stage of the DIKW pyramid (including data, patterns, models, etc.). Domains usually represent specific scientific views (disciplines), but there may be multiple views on the same abstraction within one discipline. Often, the domain of scientific models are used for further exploitation to ultimately understand the real world. Since most scientific views and their corresponding models are traditionally explored in isolation, they are also restricted to limited findings and do not allow the derivation of a holistic understanding of functional networks between multiple concepts such as relationships between entities among different scientific views and disciplines.

The reasons for isolated research may be manifold, e.g., missing data and knowledge, limited expressiveness of scientific models, or lacking awareness in domain experts about data science methods (which in turn may be hard to apply for inexperienced users). Many views are even supposed to be incompatible to each other so far, because of representational gaps in models, technical boundaries (e.g., missing tools, systems, application programming interfaces (APIs)), semantic discrepancies of data sources (e.g., quantitative versus qualitative data), etc. In order to derive exhaustive insights, it is usually no longer sufficient to apply data analysis tools in an ad-hoc manner to data (or more general: domains) from isolated views or to a single source of truth integrating all relevant domains, because these domains are heterogeneous, dynamically changing, and mostly unstructured.

However, the generation, accessing, and processing of diverse data sources together with the proliferation of data science methods for the processing and analysis of these data sources requires a multi-disciplinary, holistic research

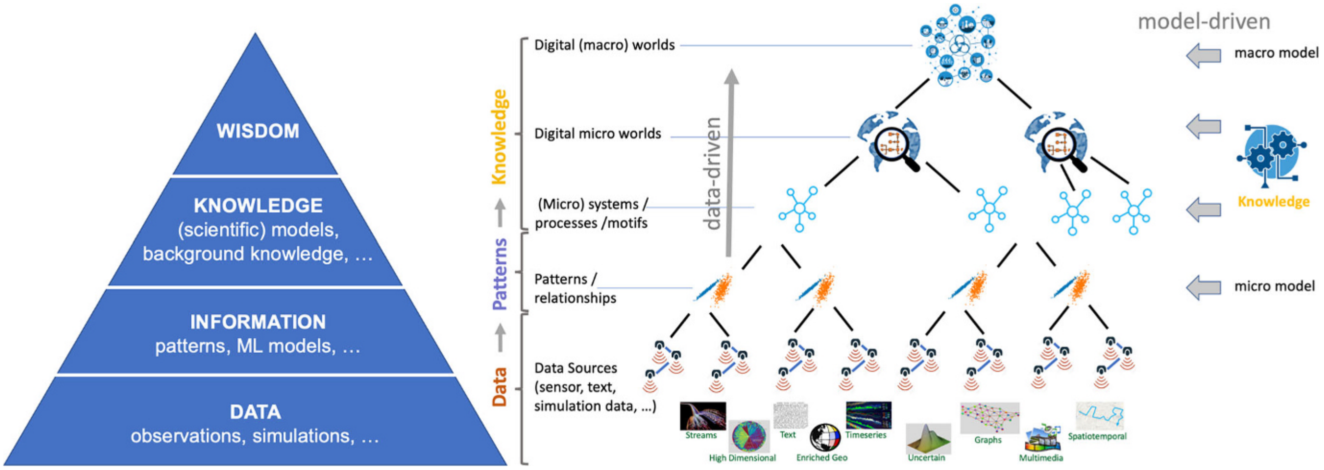

Fig. 1 Cross domain fusion (*right*) integrates different stages of the data information knowledge wisdom (DIKW) pyramid (*left*) from different scientific views and domains using the next generation of data science methods and tools. *ML* machine learning

approach. The paradigm of CDF calls for the next generation of data science methods to combine multiple domains from various scientific views. CDF systematically combines multiple ingredients (data, patterns, models, and knowledge) to derive comprehensive wisdom (cf. Fig. 1). In particular, knowledge in the form of patterns (e.g., derived from multiple data sources using data mining and machine learning) or existing models reflecting expert knowledge (e.g., knowledge graphs or simulation models) are combined with domains (data, information, knowledge) from different views to derive higher-level models and, finally, a higher level of wisdom. For example, to get insights for water mass transportation, clusters of water masses (low level patterns) are used to embed floating trajectories (further low level pattern), which in turn are used to identify patterns of water mass transitions (high-level patterns).

Sources for cross domain fusion

One key data science aspect of CDF is what sources of data and knowledge need to be fused. In general, the set of sources for CDF is unbounded; however, we will focus on prevalent sources in spatiotemporal applications and discuss potential challenges.

Data

In virtually all scientific disciplines researchers observe quantitative and/or qualitative data specific to the topic of interest.

Raw data contains quantitative measurements derived directly from some source of origin, including any types of sensors (optical, acoustic, chemical, etc.), but also humans (recording some standard measurements, e.g., the size of a bone, etc.). Often the raw data is already in some representation that is suitable for data science methods, but depending on the source of origin this usually requires some implicit (non-transparent) processing or transformation. Raw data is also often inherently fuzzy (uncertain) due to technical constraints of a sensor or due to personal bias. This uncertainty needs to be taken into account when used within a CDF framework.

Processed data contains quantitative data that has been derived from raw data through explicit pre-processing. Typical pre-processing includes aggregations (e.g., along some concept hierarchy), normalization, transformations (e.g., the derivation of new features), etc.

Interpreted data contains qualitative aspects derived from raw or processed data through an additional interpretation step that usually introduces some bias. For example, the aggregation of a scientific questionnaire involves interpretations by humans. Another example is the (temporal, cultural, etc.) classification of an archaeological finding by an expert.

While the majority of data science methods focus on the analysis of either quantitative data or qualitative data only, one particular challenge for CDF is the fusion of qualitative and qualitative data.

Information and knowledge

Scientific models Scientific models describe more or less complex systems that cover a particular part or feature of the real world, aiming at making the world (or particular aspects of the world) easier to understand, define, quantify, visualize, or simulate. These models are usually built on existing and commonly accepted domain knowledge and are represented using some mathematical formalism, e.g., some differential equations. As mentioned above, these scientific models are usually limited in the following aspects: First, the models do not cover all aspects of the real scene, so they do not exactly match the data observed from the real world. Second, the models are inferred by (individual) interpretations of observations or built on (individual) hand selected features. This often leads to multiple concurrent and sometimes even contradicting models. Third, models for particular scenes are still missing altogether. In such cases, general approximate models derived from some hypothesis that may be significantly removed from the real world are frequently used.

As a consequence, we typically see a mismatch between the data a scientific model generates and the real world, i.e., the observations about the concept the model represents. A possible solution to overcome this mismatch is to integrate the two different views on the real world (scientific model and observational data) using the paradigm of CDF.

Categories of cross domain fusion methods

CDF ranges over multiple data abstraction levels and falls into the following fusion categories.

Cross domain data fusion

To pave the path towards CDF, a novel, fundamental, and systematic framework of methods is needed that enable the fusion of data, patterns, knowledge, etc., i.e., the fusion of multiple potentially heterogeneous views and stages of data from diverse domains. Existing approaches to cross domain data fusion can be classified into three categories: stage-based, feature-based, and semantic-based data fusion methods [2, 8]. Most of these approaches have been applied in the field of urban analytics fusing traffic data with other urban attributes [2–4]. In the following, we give a brief overview of these fusion categories.

Stage-based fusion Stage-based methods apply tailored data mining algorithms to different views and stages, e.g., combine raw trajectory data with patterns derived from spatial data representing points of interest. In general, different views can be combined at any stage without any constraints regarding the consistency of their modalities, syntax, and semantics.

Feature-based fusion Feature-based methods extract features from the involved views. The simplest way to do this feature extraction is to use “manually” designed features, to concatenate these features into one common feature space, and to apply standard data mining or machine learning techniques to this common feature space. Since each (extracted) feature is treated equally, this approach may be too simple if the representation, distribution, and scale of the extracted features is very different. Instead of simply concatenating manually designed features, recent representation learning approaches that automatically learn a common feature representation, typically using deep learning, have been investigated. A downside of these approaches is that the learned feature spaces are usually hard to interpret.

Semantic-based fusion Semantic-based methods try to account for the semantics of each view and the relations between features across different views and, thus, they are typically very specific and tailored to a given problem. They rely on understanding that the fusion has a semantic meaning (and insights) derived from the ways that people think of a problem with the help of multiple views. General methods used in this category include multi-view-based, similarity-based, probabilistic dependency-based, and transfer learning-based methods.

Knowledge fusion through inter-active systems

As the fusion of information from diverse data sources has tremendous potential for the discovery of new and relevant knowledge, it becomes clearly obvious that the incorporation of background knowledge from the user or stakeholder would significantly increase this potential.

Knowledge fusion based on human computer interaction addresses three key questions: 1) what objects (data or information) the user is allowed to interact with, 2) how the data is presented to the user, and 3) how the user can interact with the objects. While most work in this field has been done concerning the two latter questions mainly addressing visualization and modes of human computer interaction, the identification of objects or information presented to the user and enabled for interaction is a far underrepresented field of study. Usually, the decision on what the user can interact with are predefined and often hard-coded. With interaction-driven knowledge fusion, however, we have a different situation as different users want to interact with different things and in a different way. Traditional information systems require exact specification of the information to be extracted by the system, requiring precise knowledge about the data and information organized in the system, which is infeasible for data-rich systems. Instead, techniques capable of filtering the data that is most relevant for the users or stakeholders are required. Machine learning-driven recommender engines, mostly used for product recommendation like amazon, or recommendation in social networks [9] like linked-in or Facebook, could be a promising blueprint for

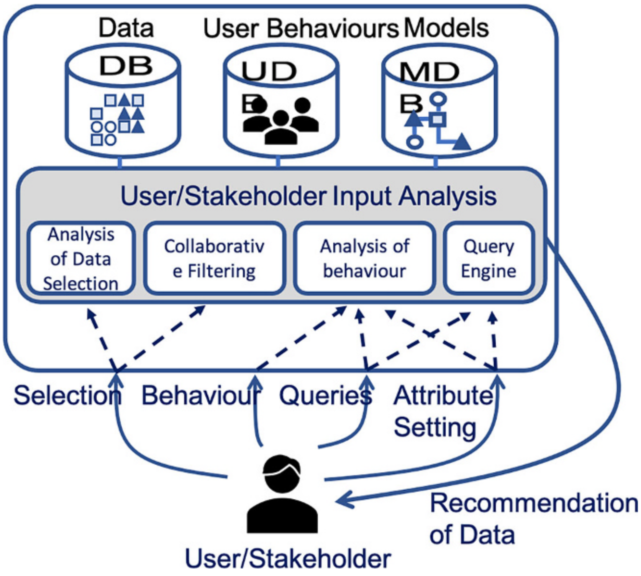

Fig. 2 Data recommendation framework enabling interactive knowledge fusion. *DB* database, *UDB* user behavior database, *MDB* model data base

a solution to the aforementioned problem. The idea of such a system, adapted from ordinary recommendation architectures [10], is depicted in Fig. 2. The backbone of the system is a user input analysis layer with a component for data selection analysis, collaborative filtering, a component for user behavior analysis, and a query engine. The components of this layer are connected to data and model repositories managing the data and models the user wants to interact with, as well as a database managing the data about the interaction between the user/stakeholder and the system. All interaction activities including selection and manipulation of certain data or models, data/model exploration tasks, queries, and attribute-setting of models will be monitored and maintained in the user behaviour database. This database will be used to predict further data and/or models that are potentially relevant for the user based on certain user–system interaction events supported by machine learning techniques. The identified data is recommended to the user, potentially in a fashion like "Based on your interaction with dataset A, you might be also interested in dataset B" and together with explanations of the relationship between these two datasets.

Fusion of data and models

The fusion of pure data-driven models using machine learning approaches and scientific models enables the combination between the real world and our parametric description of the world. By means of bidirectional interfaces such hybrid approach aims to achieve a synergetic effect between formalized and data-driven models. Two key questions arise in this context: 1) how can we advance models by data measured from the real world? and 2) how can we augment measured data with data from models?

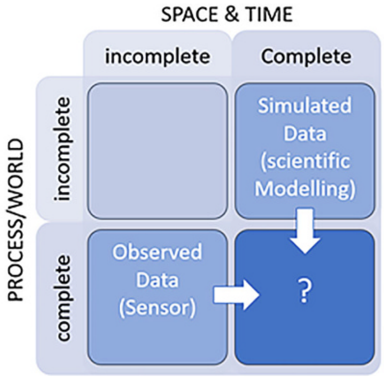

Fig. 3 Completeness of data vs model along spatial and temporal domain and real world coverage

The interplay of the two different views represented by scientific models and by observation (data) is key to CDF. Scientific models for spatiotemporal abstractions such as differential equations are (almost) complete in space and time, i.e., for any input from the continuous space and time dimensions, the models will be able to compute an output. However, due to the mismatch between model and observation mentioned above, the view of scientific models is incomplete with respect to the real world, i.e., the output of the model often does not match the real world. In contrast, data observed from the real world is inherently incomplete (sparse) in space and time since we only have a limited set of observations from the infinite space-time continuum. However, this view is (obviously by design) complete with respect to the real world. One key challenge for CDF will be to fill the gaps of both views by fusing these views. The fusion of the different views will aim at deriving a more complete view, i.e., a model that is shifted towards completeness in both dimensions, space and time, as well as the real world (see Fig. 3).

Examples for CDF tasks

Interplay of scientific models and observational data In marine science, scientific models describe an abstraction of the real world using some mathematical formalism, e.g., a set of differential equations. While these models are by design (almost) complete in space and time, i.e., provide an output for any input from the continuous space and time dimensions, the models suffer from an observational gap, i.e., a mismatch between the data a scientific model generates and the real world. The reasons for this gap may be manifold, e.g., the models cannot cover all aspects of the real world, the models are inferred by (individual) interpretations of observations or built on (individual) hand selected features (which often leads to multiple concurrent and sometimes even contradicting models), as well as models for particular aspects are still missing altogether (general approximate models derived from some hypothesis which may be significantly removed from the real world are frequently used in such cases). In contrast, observational data obviously does not suffer from an observational gap, but is inherently incomplete in space and time (we will never have observations at *any* possible spatiotemporal coordinate), i.e., has a spatiotemporal gap. To overcome these different gaps, the two different views (scientific model and observational data) need to be integrated using the paradigm of CDF.

Integrating multiple views and/or stages into data-driven analysis In spatiotemporal data analysis, integrating heterogeneous views and stages is a common challenge. Pure spatiotemporal data is usually enriched by other data modalities such as text, images, general sensor readings, etc. The integration of these different, typically very heterogeneous modalities into the data analysis process is often the key to success. This integration often needs to be done along multiple stages. A simple example of such a stage-based CDF approach is described in the following (cf. Fig. 4). The shown stage-based CDF approach aims at identifying insights of water mass transportation. The approach is based on two input data sets, Eulerian water mass data (data set 1) and Lagrangian water flow trajectories (data set 2). In the first fusion stage (stage 1), Eulerian water mass data (saline concentration and temperature) is transformed into categories by means of clustering and mapping techniques. In stage 2, Lagrangian water flow trajectory data is fused with the water mass patterns derived from stage 1 and embedded into sequences of water mass states that are subsequently used to derive the final water mass transition model in stage 3. This example illustrates how different data sources can be fused on different data abstraction levels with the help of machine learning techniques using the stage-based fusion concept.

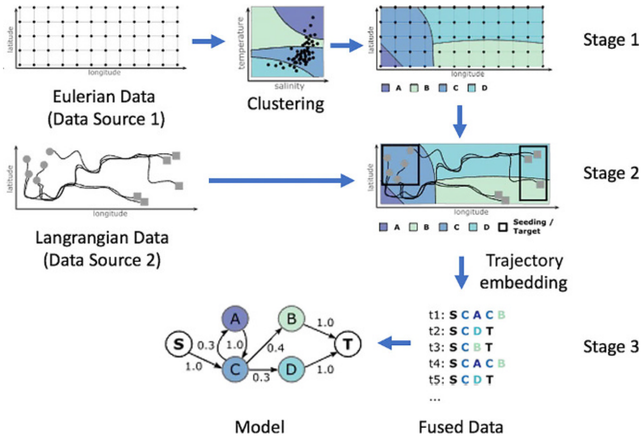

Fig. 4 Discovery of water mass transportation model based on stage-based cross domain fusion paradigm as in [11]

Conclusion

From the data science perspective, CDF offers a plethora of new research challenges. The general ideas of cross domain fusion are not entirely new, particularly in spatiotemporal applications. Triggered by the big data era, early incarnations of these ideas in the spatiotemporal domain include the paradigm of "enriched geo-spatial data." However, existing work in this direction consists of several, mostly isolated and application-specific approaches for fusing different views. Rather, to leverage the paradigm of CDF, a fundamental and systematic framework of new data science methods that enable the fusion of multiple views and stages is needed. This includes (but is not limited to) innovative concepts dealing with the combined analysis of heterogeneous and unstructured data sources, data types, and data formats, the fusion of any stage from data to knowledge including the fusion of different stages with research methods and tools, the management and analysis of fusion provenance, fusion quality, and fusion reliability, as well as interactive fusion (e.g., "human-in-the-loop" concepts) and the visualization of the fusion.

Acknowledgements We acknowledge the support from Kiel Marine Science (KMS) – Zentrum für interdisziplinäre Meereswissenschaften at the Christian-Albrechts-Universität zu Kiel – and MarDATA – Helmholtz School for Marine Data Science for our work on Cross-Domain Fusion.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

1. Meng T, Jing X, Yan Z, Pedrycz W (2020) A survey on machine learning for data fusion. Inf Fusion 57:115–129
2. Zheng Y (2015) Methodologies for cross-domain data fusion: an overview. IEEE Trans Big Data 1(1):16–34
3. Liu J, Li T, Xie P, Du S, Teng F, Yang X (2020) Urban big data fusion based on deep learning: an overview. Inf Fusion 53:123–133

4. Khan S, Nazir S, García-Magariñob I, Hussain A (2021) Deep learning-based urban big data fusion in smart cities: towards traffic monitoring and flow-preserving fusion. Comput Electr Eng 89:106906
5. Soldi G, Gaglione D, Forti N, Millefiori LM, Braca P, Carniel S, Di Simone A, Iodice A, Riccio D, Daffin'a FC et al (2021) Space-based global maritime surveillance. part ii: artificial intelligence and data fusion techniques. IEEE Aerosp Electron Syst Mag 36(9):30–42
6. Karagiannopoulou A, Tsertou A, Tsimiklis G, Amditis A (2022) Data fusion in earth observation and the role of citizen as a sensor: a scoping review of applications, methods and future trends. Remote Sens 14(5):1263
7. Rowley J (2007) The wisdom hierarchy: representations of the dikw hierarchy. J Inf Commun Sci 33(2):163–180
8. Zhang L, Xie Y, Xidao L, Zhang X (2018) Multi-source heterogeneous data fusion. In: 2018 International conference on artificial intelligence and big data (ICAIBD). IEEE, pp 47–51
9. Raghuwanshi SK, Pateriya R (2019) Recommendation systems: techniques, challenges, application, and evaluation. In: Soft computing for problem solving. Springer, Berlin, pp 151–164
10. Eckhardt A (2009) Various aspects of user preference learning and recommender systems. Dateso 2009, pp. 56–67. ISBN 978-80-01-04323-3
11. Trahms C, Wölker Y, Handmann P, Visbeck M, Renz M (2022) Data fusion for connectivity analysis between ocean regions. In: Submitted to: 2022 IEEE 18th International Conference on eScience (eScience)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.