Canonical Workflows in Simulation-based Climate Sciences Postprint

Authors: Ivonne, Anders, Karsten, Peters-von Gehlen, Hannes, Thiemann, Ivonne, Anders, Karsten, Peters-von Gehlen

Date: 2022-11-28T00:00:00+00:00

Abstract

In this paper we present the derivation of Canonical Workflow Modules from current workflows in

simulation-based climate science in support of the elaboration of a corresponding framework for simulation#2;

based research. We first identified the different users and user groups in simulation-based climate science

based on their reasons for using the resources provided at the German Climate Computing Center (DKRZ).

What is special about this is that the DKRZ provides the climate science community with resources like high

performance computing (HPC), data storage and specialised services, and hosts the World Data Center for

Climate (WDCC). Therefore, users can perform their entire research workflows up to the publication of the

data on the same infrastructure. Our analysis shows, that the resources are used by two primary user types:

those who require the HPC-system to perform resource intensive simulations to subsequently analyse them

and those who reuse, build-on and analyse existing data. We then further subdivided these top-level user

categories based on their specific goals and analysed their typical, idealised workflows applied to achieve

the respective project goals. We find that due to the subdivision and further granulation of the user groups,

the workflows show apparent differences. Nevertheless, similar "Canonical Workflow Modules" can be

clearly made out. These modules are "Data and Software (Re)use", "Compute", "Data and Software Storing",

"Data and Software Publication", "Generating Knowledge" and in their entirety form the basis for a Canonical

Workflow Framework for Research (CWFR). It is desirable that parts of the workflows in a CWFR act as FDOs,

but we view this aspect critically. Also, we reflect on the question whether the derivation of Canonical Workflow

modules from the analysis of current user behaviour still holds for future systems and work processes.

Full Text

Preamble

Canonical Workflows in Simulation-based Climate Sciences

Ivonne Anders†, Karsten Peters-von Gehlen† & Hannes Thiemann
German Climate Computing Center (DKRZ), Bundesstraße 45a, D-20146 Hamburg, Germany

Keywords: Canonical Workflow; FAIR Digital Objects; HPC infrastructures; Data services; Climate science

Citation: Anders, I., et al.: Canonical workflows in simulation-based climate sciences. Data Intelligence 4(2), 212-225 (2022). doi: 10.1162/dint_a_{00127}

Received: August 2, 2021; **Revised:** December 1, 2021; **Accepted:** February 4, 2022

Abstract

In this paper we present the derivation of Canonical Workflow Modules from current workflows in simulation-based climate science to support the development of a corresponding framework for simulation-based research. We first identified the different users and user groups in simulation-based climate science based on their reasons for using the resources provided at the German Climate Computing Center (DKRZ).

What is special about this is that DKRZ provides the climate science community with resources like high-performance computing (HPC), data storage, and specialized services, and hosts the World Data Center for Climate (WDCC). Therefore, users can perform their entire research workflows up to the publication of data on the same infrastructure. Our analysis shows that the resources are used by two primary user types: those who require the HPC system to perform resource-intensive simulations to subsequently analyze them, and those who reuse, build upon, and analyze existing data. We then further subdivided these top-level user categories based on their specific goals and analyzed their typical, idealized workflows applied to achieve respective project goals. We find that due to the subdivision and further granulation of user groups, the workflows show apparent differences. Nevertheless, similar "Canonical Workflow Modules" can be clearly identified. These modules are "Data and Software (Re)use," "Compute," "Data and Software Storing," "Data and Software Publication," "Generating Knowledge," and in their entirety form the basis for a Canonical Workflow Framework for Research (CWFR). While it would be desirable for parts of the workflows in a CWFR to act as FAIR Digital Objects (FDOs), we take a critical view of this aspect. We also reflect on whether the derivation of Canonical Workflow modules from the analysis of current user behavior will still hold for future systems and work processes.

Corresponding authors: Ivonne Anders (Email: anders@dkrz.de; ORCID: 0000-0001-7337-3009); Karsten Peters-von Gehlen (Email: peters@dkrz.de; ORCID: 0000-0003-0158-2957)

---

1. Introduction

In addition to long-term meteorological measurements, results obtained from simulations performed with climate models form the main basis for research and statements on past and possible future global, regional, and local climate. These models are based on mathematical equations that express fundamental physical relationships and laws, such as the laws of conservation of mass, momentum, and energy [e.g., 1, 2, 3].

Running these models, for example at high spatial resolution, is computationally expensive and requires high-performance computing (HPC) infrastructure. Depending on the experiment, the data volume for simulations over long periods of time and/or at high spatial resolution can reach sizes of TeraBytes (TB) to PetaBytes (PB) [4, 5]. Therefore, the geoscience community is working to develop and establish effective and reproducible scientific workflows [6]. The aim is to enable researchers to spend more time on actual scientific work [7]. Furthermore, several people or teams often work in projects of coordinated climate simulations where tasks are shared, which once again highlights the need for standardized workflows.

With the help of climate models, it is possible to investigate and determine interactions of individual components of the climate system, but they also serve as input data for impact models—that is, calculations on how climate change can affect ecosystems, urban development, or various other systems [e.g., 8, 9, 10].

Climate models and climate data are major components of different disciplines of climate science and form the basis for assessing the risks and opportunities of future climate change and for developing adaptation measures [e.g., 11, 12, 13]. Various statistical methods are used for scientific evaluation as well as for processing and using climate model data [e.g., 14, 15, 16]. Further, the ever-increasing data volume and complexity of analysis workflows has spawned the development of novel data access and analysis infrastructures [e.g., 17, 18, 19]. Upcoming major funding initiatives facilitating a step-change in climate science already cast their shadows [e.g., 20] and will allow for major investments in infrastructure needed to transform the way simulation-based climate science is performed [21].

High Performance Computing (HPC) Centers like the German Climate Computing Center (DKRZ) are major partners for climate research, as they provide essential infrastructure such as HPC resources, data storage, and tailored services to support simulation-based climate science. It should be pointed out that climate scientists can and do perform the entire suite of their data-intensive workflows using DKRZ infrastructure and services—ranging from planning and setting up model simulations, analyzing model output, reusing existing large-volume datasets, to data publication and long-term archival. This allows us to analyze DKRZ user behavior for the purpose of devising generalized analysis workflows in simulation-based climate science. These are then amenable to further abstraction to support the development, adaptation, and dissemination of the Canonical Workflows For Research (CWFR) [22] concept.

In the first step, we present the results of our user analysis. Based on this, we present the CWFR modules we derived from it. We then take a critical look at the use of FAIR Digital Objects (FDOs) [23, 24] in the CWFR process with its possibilities and limitations. Finally, we summarize and offer first solutions.

2.1 User Types Analysis

As a basis for deriving CWFR modules in simulation-based climate science, we first analyzed the current workflows of DKRZ users. We distinguish two main types of DKRZ users: the modellers and the data (re)users. However, it is also possible that the modeller switches roles to become a data (re)user. Modellers are further divided into modellers with little to no data use in the modelling process itself and modellers with unconditional data use. These are users who run models requiring existing input data. Both subgroups of modellers are further subdivided into modellers pursuing a scientific goal, model developers, and modellers performing production runs. The latter two groups usually do not pursue a scientific objective. Of course, there are also users here who nevertheless pursue a research goal afterwards—that is, change their role.

The group of data (re)users is divided into three subgroups: researchers pursuing a scientific goal, (climate) service providers, and impact modellers.

[Figure 1: see original paper] *User groups and further subdivisions. It should be noted that the above presents an idealized view of the DKRZ user groups. Each group has been analyzed in detail. In each group there are sub-groups that may not carry out parts of the workflows or may carry them out repeatedly. This can lead to very complex, interlocking and interdependent workflows.*

2.2 Detailed Workflow Examples

In the following, we illustrate the idealized workflows of just two specific user groups for the sake of brevity: climate modellers with unconditional use of existing data in the modelling process itself and data re-users. In each described case, both users pursue a scientific goal.

Example 1: Climate Modeller pursuing a scientific question

The climate modeller in this example (Figure 2 [Figure 2: see original paper]) (re)uses data to drive the climate model. This data is either already present in the local infrastructure or is copied from external resources (Step 1 in Figure 2).

[Figure 2: see original paper] *Idealized linear workflow of a climate modeller pursuing a scientific question. Steps from left to right: reusing existing data, carrying out model simulations at HPC-system, storing model output data, analysing model results and comparing existing data (reuse), storing output data from analysis, publishing model output data and analyzing data in the repository, and generating knowledge; light grey boxes indicate the converting processes.*

The data is then converted (e.g., by an input model) to comply with the input format required by the climate model. Before executing the model simulation, the model has to be adapted and compiled according to the HPC system specifications. In the example shown, the scientist runs several model simulations with the same input data but different model configurations. Output data is obtained for each model run. The model needs a converter to write the output data in certain formats. In subsequent analyses and calculations on a computer server, the scientist compares the data from the different simulations, but also to other already existing datasets, e.g., observation data. This comparative data usually also requires reformatting (converting) prior to analysis. Eventually, new data is created, stored, and—like the output data of the simulations—published. A converter adapting the data to the standards required by the repository is usually required as well. The workflow is finalized with answering the scientific question and generation of new knowledge.

The scheme shown in Figure 2 would look identical if not one model but different models were used, as for example in a Model Intercomparison Project like CMIP6 [4].

Example 2: Data (re)user for reasons of analysis

The workflow of a scientist reusing existing data and pursuing a scientific question is shown in Figure 3 [Figure 3: see original paper].

The data either exists in the local infrastructure or is copied from external sources (light orange object and dashed line in Figure 3). The data, usually available following different (meta)data formats/standards, are read in for analysis by specific routines that work as converters. Reuse of such routines or converters is uncommon—that is, almost every scientist conceives an individual solution. Next, the prepared data are analyzed by the scientist. For this, the scientist employs the local software environment and computing infrastructure, e.g., analysis servers or even HPC if massive parallel operations are needed. As a result of the analyses, new data is generated, stored, and published. Another converter is used to adapt the data to the standards required by the repository. The scientist answers the scientific question and thus generates new knowledge by, for example, summarizing the research results in a scientific publication.

[Figure 3: see original paper] *Idealized linear workflow of data reuse following a research question. Parts from left to right: reusing existing data, analysing and comparing existing data (reuse), storing output data from analysis, publishing analysis data in the repository, and generating knowledge; light grey boxes indicate the converting processes.*

---

3. Deriving Canonical Workflows from User Behaviour

The two illustrated workflows show apparent differences but also similarities. Specifically, we see that steps 4 to 7 in example 1 (Figure 2) are identical to steps 2 to 5 in example 2 (Figure 3). Therefore, "Canonical Workflow Modules" (as we call them) can be clearly identified and are amenable to abstraction.

The modules we define are "Data and Software (Re)use," "Compute," "Data and Software Storing," "Data and Software Publication," "Generating Knowledge" (compare Figure 4 [Figure 4: see original paper]). In a specific workflow, these modules can be repeated individually or in complex interaction. However, individual modules can be omitted or not run through. Between the Canonical Workflow modules, "Converters" must be used, which can be Canonical Workflows themselves or FAIR Digital Objects [23, 24].

[Figure 4: see original paper] *Canonical Workflow modules and identified submodules.*

3.1 Data and Software (Re)use

Data and software (re)use is an important Canonical Workflow module. Data is accessed either directly at the local computing infrastructure or obtained from a repository/archive. In the latter case, additional steps are needed to either copy the data to the computing infrastructure or access and process it remotely. In the modelling process, input data must conform to a certain form or standard (see also Section 3.6).

Software sharing is still not very established and can be challenging [25]. Specifically, sharing of model code is not common practice across institutions in climate science [26] due to intricate licensing frameworks and non-trivial code adaptations necessary to apply it on different computing infrastructures.

3.2 Compute

The CWFR module "Compute" incorporates all kinds of computing infrastructure use. We distinguish between two types of infrastructure: HPC and Analysis Servers. HPC environments are used for performing model simulations, whereas Analysis Servers are used for pre- and postprocessing of data as well as for analysis. Of course, very computationally intensive analyses, like deep learning approaches, must also be performed on an HPC system. The "Compute" module also includes a distinction between the use of local or external infrastructure, which has an impact on the used converters (compare Section 3.6).

3.3 Data and Software Storing

As described in the examples of scientific user workflows, output data but also new software (new parts in the model code or analysis software) are created in the process. Writing and storing data and software is thus another potential CWFR module. Again, we distinguish between the extent to which data is stored close to the user' s computer infrastructure or transferred to external resources.

3.4 Data and Software Publication

The process of data or software publication is complex, as many aspects have to be taken into account. Institutional definitions of standards and licences have to be reconciled with those of funders and repositories. Software is subject to a different legal framework than data. There are currently discussions within the RDA [27], but also at the national level, on how to deal with research software publication to increase awareness for its relevance [28] and provide guidance. In the context of publication, converters play a decisive role because they are used for mapping metadata.

From a less technical perspective, it should be noted that publishing data is most often only carried out when requested by journals, funders, or institutions. This is because data publication (and sharing) is currently still not considered from the beginning of the research project, and the data preparation and publication process is considered too time-consuming given the lack of evident scholarly benefits [29, 30, 31, 32]. The integration of the publication process into the CWFR framework is therefore crucial to incentivize increased data publication.

3.5 Generating Knowledge

Achieving new knowledge is not a workflow step in the sense that it leads to technical implementation. Nevertheless, we include it in the workflow chain because it is an important link between research phases in the scientific cycle and plays a role in the further development of theories and solutions. In addition to pure knowledge gain, dissemination—that is, the sharing of knowledge with other scientists—can also be considered here, for example through a scientific publication or presentation of results at a scientific conference.

3.6 Converter

Converters are very important elements in the entire workflow and therefore in a CWFR. They occupy a special position because they are not bound to the individual research workflow but can be used and reused in a variety of generalized ways as described in Section 2. In Figures 2 and 3 they are shown in light grey. Currently, they are individually generated and applied by scientists themselves, which takes valuable time that should be more usefully spent on the scientific question. Therefore, converters offer the greatest potential for technical standardization to make them more transparent and act as FDOs (see also Section 4).

Converters themselves can be individual programs but also more complex processes consisting of several technical elements. For example, the process of publishing consists of transforming metadata standards, adapting data formats, assigning PIDs, creating entries in repository databases, etc.

---

4. FDOs in CWFRs of Data Intensive Climate Science

CWFRs are the overarching constructs containing the identified workflow modules (Section 3). In these, FDOs are the data products resulting from CWFR modules, as well as the converters and routines used in the CWFR modules (tools). The workflow itself is also an FDO, thereby facilitating reuse and reproducibility.

In simulation-based climate science, the use of FDOs in every step of the entire process chain would be groundbreaking, and the advantages are obvious:

Provenance tracking and full reproducibility: For each FDO, complete provenance information is stored (in the PID profile and in the metadata). Source data, converters, processing software or model and model specifications, which are also specified as FDOs, are indexed. Ultimately, the last FDO in the workflow contains the information of the entire workflow. If software is viewed as an FDO, this object contains not only the usual metadata such as authors and licences but also all information on the infrastructure, compiler, and compiler specifications.

Creation of globally federated databases: FDOs have a unique globally accessible PID, allowing for global access and data (re)use. Globally federated databases of FDOs building on existing infrastructure concepts [e.g., 17] would thus be achieved.

Supporting Big Data approaches: One part of Big Data approaches is the combined use of heterogeneous data. To achieve this, data must be standardized and described in terms of its heterogeneity to facilitate machine-actionability. FDOs have a decisive role here.

Automated selection of analysis tools and libraries: With regard to non-expert users, interfaces that automatically select analysis tools and libraries in the background are necessary for a desired workflow to run through. These interfaces can also be designed as FDOs.

However, the full implementation of the CWFR concept in climate science would be challenging (Section 5.2).

5.1 Status Quo Workflow vs. Future Workflows

We have identified Canonical Workflow modules based on current user behavior, but do they hold up for the future? The modules are designed so that they continue to exist as such. However, it is expected that users will not actively engage with the implementation or processing of certain modules in the near to distant future. This is especially true for the Data and Software Storing and Publication parts, including the standardization of metadata. However, we expect that with the further establishment of globally federated databases, the distinction between "local" and "external" will technically no longer make a difference. The aspect will remain, but the question is how the information of the overall workflow will be preserved. A system is needed to log individual location-independent workflow steps and record them in FDOs.

5.2 Critical View on the Use of FDOs in Simulation Based Climate Science

Numerical models are the essential component of simulation-based climate science. The model code is often not freely available and is subject to strict licensing and usage agreements. Only a few models are freely accessible in their source code [26] and can thus truly be declared as FDOs.

Further, simulation results depend on the infrastructure used for the simulation [33]. HPC systems are usually replaced after 4–6 years. With the discontinuation of the infrastructure used, the data are also no longer reproducible. This would mean that indexing of the model and model setup as an FDO would be possible, but no longer repeatedly executable. To overcome this issue, efforts to establish the use of workflow tools and container approaches in HPC environments with concrete applications in simulation-based climate science are being considered [e.g., 34, 35, 15, and references therein]. Storing the workflows themselves instead of data would also be a very good approach in terms of saving resources (storage space).

In order to truly introduce FDOs in the area of software (model and analysis tools), decisive prerequisites must be created, e.g., git commits. Furthermore, there are currently no standards for PID profiles. Without standardization in this area, the introduction and use of FDOs is very difficult, as PID profiles form an essential basis for the FAIRness of data objects and contain important information about the data or software object.

---

6. Summary and Outlook

We have used DKRZ user behavior as an example to devise building blocks of CWFRs. DKRZ provides users with computing power and storage capacity but is also a repository for climate data (model and observation data), which enables data publication but also direct data reuse. Thus, scientists can carry out all steps of their research work in one overall system. This makes it possible to develop and implement the CWFR framework in an interlocked manner.

We identified different users and user groups based on their reasons for using the supplied resources and then examined the workflows they apply in their work. Our analysis shows that the resources are used by two primary user types: those who require the HPC system to perform resource-intensive simulations to subsequently analyze them, and those who reuse, build upon, and analyze existing data stored either on local file systems or reused from external resources. We found that there are superordinate CWFR modules that several user groups use in the same way or that are repeated again and again. These offer potential for automation or the use of FDOs. However, we have also taken a critical look at the use of FDOs in this subject area as well as possible future changes to workflows. Because climate science is rapidly proceeding toward exascale computing [20], current scientific workflows will soon not be fit-for-purpose anymore [21]. Furthermore, the reproducibility of simulation results is very limited due to constant changes in HPC infrastructures [33, 25].

We are currently facing these challenges: The Free Evaluation System Framework for Earth System Modelling [Freva, 18] is being further developed at DKRZ. Freva can be used to orchestrate all steps of scientific workflows necessary to perform research in data-intensive climate science. It builds on a standardized database, provides a programming interface, tracks provenance information, and can be run from the command line or a web interface. Freva is currently undergoing massive updates and improvements, such as direct data access to repositories or handling numerical model simulations, to support future research projects requiring a step-change in the way climate science is performed.

Overall, the complexity of the individual CWFR modules we identified here poses a great challenge, as it is not sufficient to provide individual solutions but to always keep an eye on the interaction of all components.

Looking at the defined Canonical Workflow Modules and corresponding sub-modules, one finds that they are transferable to all disciplines that follow simulation-based approaches (data-, simulation-, and/or analysis-intensive disciplines). These are not only fields from physics or chemistry but also political science, economic research, etc.

Acknowledgements

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’ s Excellence Strategy—EXC 2037 CLICCS—Climate, Climatic Change, and Society—Project No. 390683824.

Author Contributions

I. Anders (anders@dkrz.de) and K. Peters-von Gehlen (peters@dkrz.de) collected all information on users and how these used the existing systems. They analyzed the individual work processes and grouped the users into different classes and derived the workflow modules. They also defined the structure of the paper and wrote the main part. H. Thiemann (thiemann@dkrz.de) contributed significantly to sharpening the content and thus the improvement. All authors have made meaningful and valuable contributions to revising and proofreading the manuscript.

References

[1] Washington, W.: Three dimensional numerical simulation of climate: The fundamentals. In: von Storch, H., Flöser, G. (eds.) Anthropogenic Climate Change, pp. 37-59. Springer, Berlin (1999)

[2] Roeckner, E., et al.: The atmospheric general circulation model ECHAM 5. PART I: Model description (2003). Available at: https://www.researchgate.net/publication/247784697_{The}
Accessed 1 December 2021

[3] Zängl, G., et al.: The ICON (ICOsahedral Non-hydrostatic) modelling framework of DWD and MPI-M: Description of the non-hydrostatic dynamical core. Quarterly Journal of the Royal Meteorological Society 141(687), 563-579 (2015)

[4] Balaji, V., et al.: Requirements for a global data infrastructure in support of CMIP6. Geoscientific Model Development Discussions 11(9), 3659-3680 (2018)

[5] Stevens, B., et al.: DYAMOND: The DYnamics of the atmospheric general circulation modeled on non-hydrostatic domains. Progress in Earth and Planetary Science 6(1), 1-17 (2019)

[6] ECMWF: Workshop: Building reproducible workflows for earth sciences. Available at: https://www.ecmwf.int/en/learning/workshops/building-reproducible-workflows. Accessed 26 July 2021

[7] Weigel, T., et al.: Making data and workflows findable for machines. Data Intelligence 2(1-2), 40-46 (2020)

[8] Bopp, L., et al.: Multiple stressors of ocean ecosystems in the 21st century: Projections with CMIP5 models. Biogeosciences 10(10), 6225-6245 (2013)

[9] Lauwaet, D., et al.: Detailed urban heat island projections for cities worldwide: Dynamical downscaling CMIP5 global climate models. Climate 3(2), 391-415 (2015)

[10] Carvalho, D., et al.: Potential impacts of climate change on European wind energy resource under the CMIP5 future climate projections. Renewable Energy 101, 29-40 (2017)

[11] Howden, S.M., et al.: Adapting agriculture to climate change. PNAS 104(50), 19691-19696 (2007)

[12] Lim, W.H., et al.: Long-term changes in global socioeconomic benefits of flood defenses and residual risk based on cmip5 climate models. Earth' s Future 6, 938-954 (2018)

[13] Giordano, R., et al.: Urban adaptation to climate change: Climate services for supporting collaborative planning. Climate Services 17, 100100 (2020)

[14] von Storch, H., Zwiers, F.W.: Statistical analysis in climate research. Cambridge University Press, Cambridge (2002)

[15] Kadow, C., Hall, D.M., Ulbrich, U.: Artificial intelligence reconstructs missing climate information. Nature Geoscience 13(6), 408–413 (2020)

[16] Reichstein, M., et al.: Deep learning and process understanding for data-driven earth system science. Nature 566(7743), 195–204 (2019)

[17] Cinquini, L., et al.: The earth system grid federation: An open infrastructure for access to distributed geospatial data. Future Generation Computer Systems 36, 400–417 (2014)

[18] Kadow, C., et al.: Introduction to Freva–A free evaluation system framework for earth system modeling. Journal of Open Research Software 9, Article No. 13 (2021)

[19] Abernathey, R.P., et al.: Cloud-native repositories for big scientific data. Computing in Science and Engineering 23(2), 26–35 (2021)

[20] Bauer, P., Stevens, B., Hazeleger, W.: A digital twin of earth for the green transition. Nature Climate Change 11(2), 80–83 (2021)

[21] Lawrence, B.N., et al.: Crossing the chasm: How to develop weather and climate models for next generation computers? Geoscientific Model Development 11(5), 1799–1821 (2018)

[22] Hardisty, A., Wittenburg, P. (eds.): Canonical Workflow Framework for Research (CWFR)—position paper—version 2, December 2020. Working paper. Available at: https://osf.io/9e3vc/. Accessed 28 July 2021

[23] Schultes, E., Wittenburg, P.: FAIR principles and digital objects: Accelerating convergence on a data infrastructure. In: Manolopoulos, Y., Stupnikov, S. (eds.) Data Analytics and Management in Data Intensive Domains, pp. 3–16. Springer International Publishing, Cham (2019)

[24] Schwardmann, U.: Digital objects-FAIR digital objects: Which services are required? Data Science Journal 19(1), Article No. 15 (2020)

[25] Peer, L., et al.: Challenges of curating for reproducible and FAIR research output (2021). Available at: https://www.rd-alliance.org/group/cure-fair-wg/outcomes/challenges-curating-reproducible-and-fair-research-output. Accessed 26 July 2021

[26] Añel, J.A., García-Rodríguez, M., Rodeiro, J.: Current status on the need for improved accessibility to climate models code. Geoscientific Model Development Discussions 14(2), 923–934 (2021)

[27] RDA: FAIR for research software (FAIR4RS) WG. Available at: https://www.rd-alliance.org/groups/fair-research-software-fair4rs-wg. Accessed 28 July 2021

[28] Davenport, J.H., Grant, J., Jones, C.M.: Data without software are just numbers. Data Science Journal 19(1), Article No. 3 (2020)

[29] Borgman, C.L.: The conundrum of sharing research data. Journal of the American Society for Information Science and Technology 63(6), 1059–1078 (2012)

[30] Mongeon, P., et al.: Incorporating data sharing to the reward system of science: Linking DataCite records to authors in the Web of Science. Aslib Journal of Information Management 69(5), 545–556 (2017)

[31] Cousijn, H., et al.: Bringing citations and usage metrics together to make data count. Data Science Journal 18(1), Article No. 9 (2019)

[32] Pronk, T.E.: The time efficiency gain in sharing and reuse of research data. Data Science Journal 18(1), Article No. 10 (2019)

[33] Geyer, B., Ludwig, T., von Storch, H.: Limits of reproducibility and hydrodynamic noise in atmospheric regional modelling. Communications Earth & Environment 2, Article No. 17 (2021)

[34] Manubens-Gil, D., et al.: Seamless management of ensemble climate prediction experiments on HPC platforms. In: 2016 International Conference on High Performance Computing Simulation (HPCS), pp. 895–900 (2016)

[35] Canon, R.S., Younge, A.: A case for portability and reproducibility of HPC containers. In: 2019 IEEE/ACM International Workshop on Containers and New Orchestration Paradigms for Isolated Environments in HPC (CANOPIE-HPC), pp. 49–54 (2019)

Author Biography

Ivonne Anders has been working at the Data Management Department at the German Climate Computing Center (DKRZ) since 2019. She initially worked in the CLICCS cluster of excellence in collaboration with the University of Hamburg, where she developed concepts for data management in large-scale projects and dealt with process optimization. She is a member of various national and international working groups on the topics of discipline-specific data management plans, data management plans in general, but also in relation to research software. She has been co-chair of the Research Data Alliance (RDA) Working Group "Discipline-specific Guidance for DMPs" since 2019. Since February 2022 she has also been co-chair of the Technical Specification and Integration Group in relation to the definition and implementation of FAIR Digital Objects within the FDO Forum. This year, Ivonne Anders joined the National Research Data Infrastructure project, where she is designing a service platform for RDM for Earth system data. She studied cartography, geodesy, and software development at the Technical University of Dresden. Her doctoral thesis was on regional numerical climate modelling and model evaluation. She then worked for 11 years at the Austrian Weather Service in the field of climate research and climate services.

ORCID: 0000-0001-7337-3009

Karsten Peters-von Gehlen works as Research Data Management (RDM) Service Communicator in the Data Management Department of the German Climate Computing Center (DKRZ). In this role, he is the main DKRZ contact for climate scientists with any questions or concerns regarding their RDM needs. Further, he is an active contributor to numerous national and international activities and interest groups, e.g., in the framework of the Research Data Alliance (RDA) or the FAIR Digital Object (FDO) Forum, fostering RDM education and the general push toward real acceptance and operationalization of the FAIR principles in the everyday work of (climate) scientists. He also holds RDM lectures to master and undergraduate students as well as senior scientists to bring about a change in RDM culture. Before joining DKRZ and becoming involved in RDM in 2018, he attained M.Sc. (2008) and Ph.D. (2011) degrees in Meteorology at the University of Hamburg in Germany and spent six years working as a PostDoc at Monash University in Australia and Max-Planck-Institute for Meteorology in Germany in that field.

ORCID: 0000-0003-0158-2957

Hannes Thiemann has been Head of the Data Management Department at the German Climate Computing Center (DKRZ) since 2019. After completing his studies in geophysics at the University of Hamburg in Germany, he soon specialized in the field of data management at DKRZ. For over 20 years, he has been involved in numerous e-infrastructures and (inter)national projects, including e.g., IS-ENES, EUDAT, NFDI4Earth, and the European Open Science Cloud (EOSC). He is the director of the thematic research data archive World Data Center for Climate (WDCC) and is supporting Open Science and FAIR-Data activities in numerous functions.

ORCID: 0000-0002-2329-8511