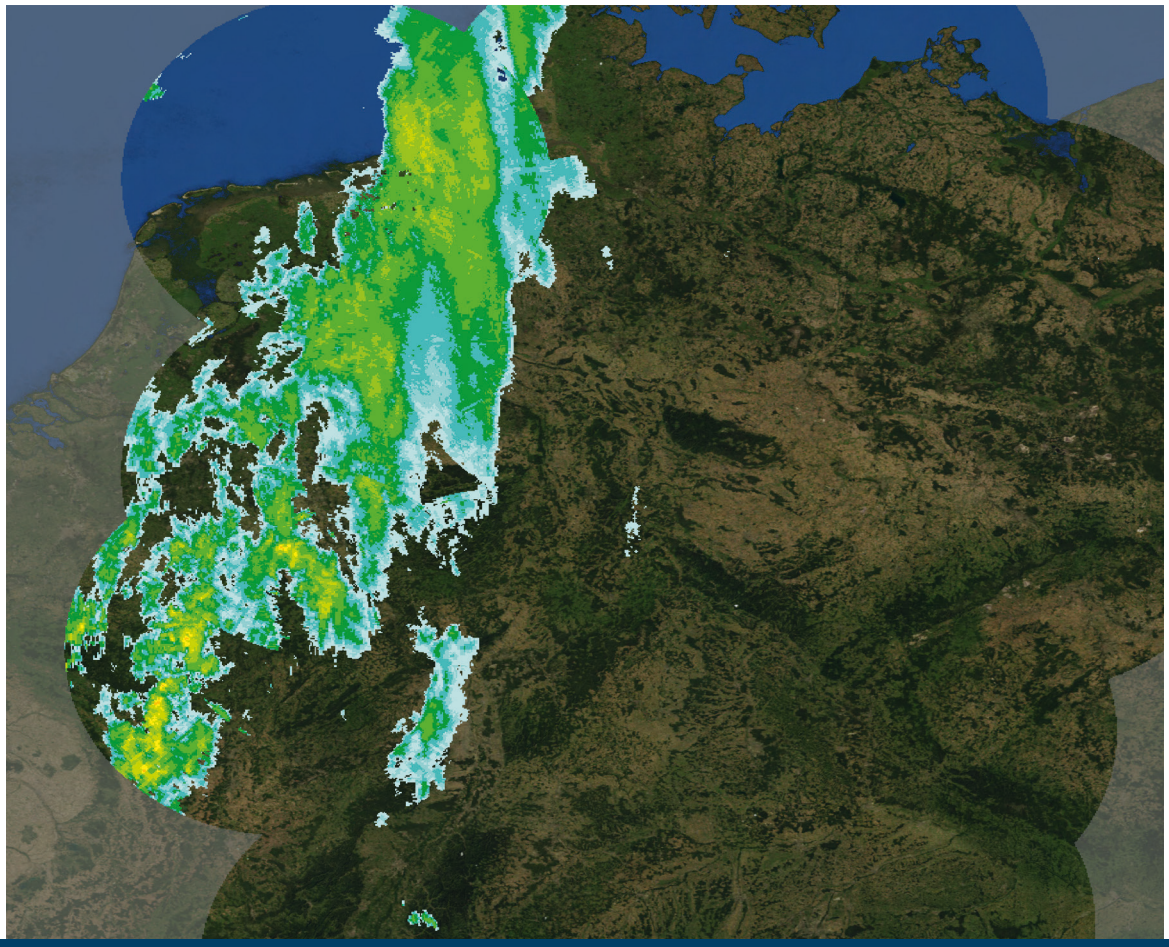


Final report of the DeepRain project
Abschlußbericht des DeepRain Projektes

IAS Series
Band / Volume 51
ISBN 978-3-95806-675-5

Mitglied der Helmholtz-Gemeinschaft

JÜLICH
Forschungszentrum

Schriften des Forschungszentrums Jülich
IAS Series Band / Volume 51

Forschungszentrum Jülich GmbH
Institute for Advanced Simulation (IAS)
Jülich Supercomputing Centre (JSC)

Final report of the DeepRain project
Abschlußbericht des DeepRain Projektes

Schriften des Forschungszentrums Jülich
IAS Series

Band / Volume 51

ISSN 1868-8489

ISBN 978-3-95806-675-5

Bibliografische Information der Deutschen Nationalbibliothek.
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der
Deutschen Nationalbibliografie; detaillierte Bibliografische Daten
sind im Internet über http://dnb.d-nb.de abrufbar.

Herausgeber und Vertrieb:	Forschungszentrum Jülich GmbH Zentralbibliothek, Verlag 52425 Jülich Tel.: +49 2461 61-5368 Fax: +49 2461 61-6103 zb-publikation@fz-juelich.de www.fz-juelich.de/zb
Umschlaggestaltung:	Grafische Medien, Forschungszentrum Jülich GmbH Jülich Supercomputing Centre, Forschungszentrum Jülich GmbH
Titelbild:	DWD Radar image generated with https://maps.dwd.de/geoserver, Nov 27th, 2022.
Druck:	Grafische Medien, Forschungszentrum Jülich GmbH
Copyright:	Forschungszentrum Jülich 2022

Schriften des Forschungszentrums Jülich
IAS Series, Band / Volume 51

ISSN 1868-8489
ISBN 978-3-95806-675-5

Vollständig frei verfügbar über das Publikationsportal des Forschungszentrums Jülich (JuSER)
unter www.fz-juelich.de/zb/openaccess.

Final report of the DeepRain project

Abschlußbericht des DeepRain Projektes

October, 1st, 2018 – March, 31st 2022 / 1. Oktober 2018 bis 31. März 2022

Funded by BMBF through grant 01IS18047 A-E

Gefördert durch das BMBF unter dem Kennzeichen 01IS18047 A-E

A collaborative project of / Ein Verbundprojekt der Partner

Forschungszentrum Jülich GmbH

Deutscher Wetterdienst

Jacobs Universität Bremen

Universität Osnabrück

Universität Bonn

Content / Inhalt:

Federal Ministry
of Education
and Research

DeepRain

Final Report

Grant no. 01IS18047
Oct. 2018 – Mar. 2022

UNIVERSITÄT OSNABRÜCK

JACOBS UNIVERSITY

UNIVERSITÄT BONN

JÜLICH Forschungszentrum

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

DWD

The research described in this report was funded by the Federal Ministry for Education and Research of Germany under grant no 01IS18047. The authors are fully responsible for the content of this publication.

Authors:

Martin G. Schultz, Forschungszentrum Jülich (PI, editor)
Amirpasha Mozaffari, Forschungszentrum Jülich (co-editor)
Michael Langguth, Forschungszentrum Jülich (co-editor)

Peter Baumann, Jacobs University Bremen
Otoniel Campos, Jacobs University Bremen
Rita Glowienka-Hense, University of Bonn
Bing Gong, Forschungszentrum Jülich
Andreas Hense, University of Bonn
Yan Ji, Forschungszentrum Jülich
Jan Keller, Hans-Ertel-Zentrum/DWD
Gordon Pipa, University of Osnabrück
Rodolfo Adrián Rojas-Campos, University of Osnabrück
Martin Wittenbrink, Hans-Ertel-Zentrum/DWD

Jülich, September 2022

Executive Summary

The DeepRain project was launched to develop new approaches to combine modern machine learning methods with high performance IT systems for data processing and dissemination in order to produce high-resolution spatial maps of precipitation over Germany. The foundation of this project was the multi-year archive of ensemble model forecasts from the numerical weather model COSMO of the German Weather Service (DWD). Six trans-disciplinary research institutions worked together in DeepRain to develop an end-to-end processing chain which could potentially be used in the future operational weather forecasting context. The project proposal had identified several challenges which had to be overcome in this regard. Next to the technical challenges in establishing a novel data fusion of rather diverse data sets (numerical model data, radar data, ground-based station observations), building scalable machine learning solutions and optimising the performance of data processing and machine learning, there were various scientific challenges related to the small local-scale structures of precipitation events, difficulties with finding robust evaluation methods for precipitation forecasts and non-gaussian precipitation statistics combined with highly imbalanced data sets.

When DeepRain started, the application of machine learning to weather and climate data was still very new and there were hardly any publications or software codes available to build upon. DeepRain thus pioneered the use of modern deep learning models in the domain of weather forecasting. Concurrently, one could observe an exponential increase in the number of publications in this new field over the past three years; very often these were studies conducted in North America or China. Global players like Google, Amazon, NVidia, or Microsoft have in the meantime established groups of scientists and engineers to advance research on “Weather AI” and develop (marketable) weather and climate applications with deep learning. Therefore, the DeepRain project was very timely as it established a baseline for machine learning in weather and climate in Germany and it allowed the consortium to explore the potential of deep learning in context with the gigantic data processing that is needed and to keep pace with the international developments in this rapidly growing field of research.

While DeepRain could not complete the final deliverable, i.e. the construction of a prototype end-to-end workflow for high-resolution precipitation forecasts based on deep learning, all of the related research questions have been answered and all the necessary building blocks for such a workflow have been developed. In particular, modern datacube technology has been used successfully for establishing four to six-dimensional atmospheric simulation datacubes based on DWD data available for extraction and analytics.

In addition to the anticipated challenges described above there were severe issues materialising during the project: 1) a large scale data loss due to hardware failures in spring 2021, 2) the Covid-19 pandemic from March 2020 until now, and 3) difficulties to find highly-skilled personnel - especially in times when most work had to be done in a home office setting.

The main accomplishments of DeepRain are:

- Petabyte-scale data transfer of archived COSMO-DE EPS forecasts from tape drives of DWD and of RADKLIM dataset from OpenData-server to the file system JUST at JSC/FZ Jülich, organisation and cleaning of these data and granting data access to all project partners,
- Parallelized processing of COSMO-EPS and RADKLIM data (ensemble statistics, remapping for data fusion and for ingestion to rasdaman),
- Implementation of Rasdaman datacube array database servers at FZ Jülich and ingestion of several TBytes of weather data,

- Establishing links from the Jülich Rasdaman servers to the EarthServer datacube federation,
- Further developments of Rasdaman to accelerate data ingestions and retrieval, define new user-defined functions for analysis of topographic data, define a new coordinate reference system for rotated pole coordinates, and prepare for interfacing machine learning workflows,
- Development of statistical downscaling techniques and machine learning models to:
    - Generate dichotomous and quantitative precipitation forecasts at station sites,
    - Generate area forecasts at the RADKLIM radar data resolution,
- Exploration of new verification statistics based on partial correlations and regression boosting.

In this report, we provide a detailed overview on the work and the achievements within the DeepRain project. This report is organised in five sections: In Section 1, we present the deliverable plan from the project proposal and provide information on the delivery state of each task separately to allow for a compact comparison between the project plan and its output. In Section 2, we then outline in detail the work carried out in the project for each work package individually. The expected outcome from the project achievements as well as possible future benefits are discussed in Section 3. In Section 4, we give a general overview on the progress made in the research fields related to DeepRain; specifically, these are: machine learning for precipitation forecasting, precipitation forecast evaluation methods, big data handling and FAIR data practices. Finally, Section 5 lists all the journal publications, data sets and software packages and planned submissions resulting from the DeepRain project.

1) The essential accounting evidence

Table 1: Overview of project deliverables and accomplishments

Task number	Deliverable in proposal	Tangible deliverables	Delivery state
1.1	Project coordination: ● monitoring project progress ● reporting progress ● meeting organization including kick-off, yearly and final meetings	● Organization of yearly report ● Organization final report ● Organization of yearly meetings, virtual meetings and the final meeting	Complete
1.2	Communication: ● Creation of project website ● news publication ● public relation work	● Website of the project is available under https://www.deeprain-project.de ● News publication has been performed regularly ● Public communication	Complete
1.3	Project management: ● final cooperation agreement ● monitor expenditure ● communication & risk management	● Monitoring the expenses ● Cost-free extension of the project	Complete
2.1	Provision of model, radar, lightning and station data from the DWD includes format description and specifications ● Part of the DWD weather data is available in Jülich. ● Data format description and specification.	● Registration of 15.24TB of COSMO remapped data for 6 relevant meteorological variables in rasdaman instance at Jülich (EnterpriseCube link) ● Documentation of the data format and description of the registered COSMO variables at the service landing page for the coverages in EarthServer federation Julich node.	Complete
2.2	Topography data accessible via rasdaman service endpoint via OGC WCS / WCPS / WMS requests. ● Download and process SRTM Topography data ● Consider interpolation and calculation of relevant features for the ML	● Ingestion scripts for SRTM topography data ● Extension of the *project()* function in rasdaman to support customized Interpolation. ● Implementation of slope/aspect/hillshade functions as user-defined functions (UDFs) in rasdaman (link).	Complete
2.3	Setting up a rasdaman array database instance in JÜLICH, the definition of the data imports from	● Established rasdaman instance at Jülich with incorporation into the EarthServer federation. ● Established ingestion scripts for	Complete

	tasks 2.1 and 2.2 and performance optimization: ● COSMO data available via rasdaman service endpoint via OGC WCS / WCPS / WMS requests. ● Topographic functions available in rasdaman.	COSMO data (link) ● Support for rotated grid CRS (link) ● Enhanced open-source geo-libraries PROJ/GDAL with rotated CRS definition ● Development of a standardized rotated CRS definition (link) ● Improved rasdaman performance: ○ Faster GRIB data import ○ Optimized case statement ○ Established benchmarks for data ingestion on using different parameters in the rasdaman ingestion code.	
2.4	Setting up a relational database for the provision of station data for the evaluation	● Registration of 2.95 TB of precipitation data into the rasdaman instance for the years 2015-2018.	Complete
2.5	Development of new database query operators for radar and lightning data and optimization of the interface to the neural network	● Since the consortium decided to refrain from a direct coupling of the rasdaman database with the neural networks, Python-based query methods have been set-up and shared among the project partners	Partial
2.6	Implementation of the data flow for the output data of the neural network, including embedding in the JOIN web interface	● Due to the decision of refraining from an explicit coupling of the rasdaman database with the developed neural networks (see 2.5), this task has been dropped	Not done
3.1	Downscaling to station locations with deep learning	● First experimental deep learning algorithms for downscaling of precipitation registered by one rain station	Complete
3.2	Implementation of first version of neural network in Jülich	● Trained deep learning models stored in Jülich	Complete
3.3	Downscaling of rain prediction considering spatial component	● Paper “Deconvolutional and generative models for generation of precipitation maps based on Numerical Weather Prediction” submitted to Journal Geoscientific Model Development ● Public repository with the code for the deep learning models on github	Complete
3.4	Downscaling of rain prediction considering spatial and temporal components	● Paper “Post-processing of NWP precipitation forecasts using deep learning” submitted to Journal Weather and Forecasting ● Public repository with the code for the models on github	Complete

3.5	Downscaling of storm prediction considering spatial and temporal components	• Impossible to attain due to the highly imbalanced number of samples on extreme events in the training data.	Not done
4.1	Database for version changes of the COSMO numerical weather model	• Documented on the COSMO model web site	Complete
4.2	Implementation of classical downscaling methods	• Logistic regression, generalized linear model and analog ensemble methods have been implemented in Python based on the DeepRain data set as classical downscaling methods.	Complete
4.3	Documentation of the method for consistency verification	• see 5.1 / 5.3	shifted to 5.1 , 5.3
4.4	Database with consistent prediction variables and with variables that undergo significant changes due to version changes.	• see 5.1 / 5.3	shifted to 5.1 , 5.3
4.5	Post-processed predictions for evaluation	• The classical downscaling methods have been applied to the data • The produced data sets have been provided as reference for comparison to the DL approaches • Paper “Post-processing of NWP precipitation forecasts using deep learning” submitted to Journal Weather and Forecasting	Complete
5.1	Documentation of bootstrap procedures and selection of probabilistic evaluation scores	• PhD thesis: Wahl (2015) with open access • Methods are part of the R-package “verification” by Gilleland (2014): R-package ”verification”: Weather forecast verification utilities. NCAR - Research Applications Laboratory, version 1.41. • More details are provided in Dorninger et al. (2018) • Presentation “Evaluation of model simulations”, March 10th, 2021 that is published via B2SHARE Presentation at project meeting in April 2022 on “free energy” based measures and Laplace’s approximation of posterior probability densities that are published via B2SHARE	Complete

5.2	First version of the evaluation toolbox available	• Glowienka-Hense et al. (2020) • Presentations published via B2SHARE	Complete
5.3	Prototype for graphical output of evaluation results	• presentations during web meetings between March 2020 and April 2022, entropy-based measures, MSE/MAE based measures, • Presentations that were given in March and September 2022 are published via B2SHARE	Complete
5.4	Documentation of procedures for optimal input variable selection, information criteria, extreme values, with results.	• Not possible due to only partially available data sets, extreme events could not be evaluated due to absence of large enough sample sizes	Not done
6.1	First workflow and data flow analysis utilising the Jülich HPC system	• Developed PyStager as a scalable workflow solution for parallelized large-scale data processing on HPC systems • Preprocessing of COSMO-EPS & RADKLIM data using PyStager, Code published on github • Preprocessing of COSMO-EPS for ingestion to rasdaman (cf. task 2.1), Code published on github • Developed Jupyter notebooks to demonstrate user-friendly HPC data handling and visualisation	Complete
6.2	Final version of the project's internal data flow and workflow analysis for HPC system	• Viewgraphs of workflow architecture created and presented at project meeting in November 2020 and published via B2SHARE • Contribution to machine learning codes to facilitate handling of massive amounts of weather data • Direct access to data for ML applications from rasdaman could not be implemented during the project due to technical issues and access restrictions	Partial
6.3	Publication of the DeepRain system architecture and workflow	• Conceptual design of the DeepRain workflow for FAIR and reproducibility described in journal Data Intelligence (see Section 5) • A complete implementation of the concept has not been possible within the project	Partial
6.4	Workflow analysis and system design for a possible	• Components of the possible operating system have been created:	Partial

	operationalization of the DeepRain process.	○ Pre-processing workflow using HPC ○ Data stream query from federated rasdaman datacube via web-browser and API query ○ Jupyter Notebook to demonstrate interaction with rasdaman and visualisation ● Viewgraph of the operational workflow have been presented in internal meetings and published via B2SHARE	

2) The performed work during the project

The central aim of the DeepRain project was the development of advanced machine learning (ML) methods, i.e. in particular deep learning (DL), for improved, local precipitation forecasts based on the output of numerical weather prediction (NWP) models. Specifically, DeepRain always had a potential operational application in mind and therefore focused on the weather model of the German Weather Service (DWD). The current NWP model chain at DWD is built upon the ICON model. The global model configuration of ICON became operational in January 2015, whereas it took until February 2021 to replace the convection-permitting model configuration COSMO-D2 with its successor, ICON-D2. Since a large number of highly resolved model forecasts at kilometre-scale are a necessary prerequisite for machine learning on precipitation forecasts, the consortium decided to work with output from the predecessor model COSMO as it was originally planned in the proposal. Specifically, the high-resolution COSMO ensemble prediction system (COSMO-EPS) was chosen as the primary source of input data to the DeepRain machine learning models.

Due to the large amount of available data from COSMO-EPS, efficient data provisioning and processing systems constituted a necessary prerequisite to building any machine learning application that could potentially be operationalised. Two main lines of work have been followed during the project: 1) parallelisation of data pre-processing on HPC-systems, and 2) set-up and population of the big data analytics server rasdaman. The first task resulted in the development of the Python software package PyStager, which is now used in several ML applications at the Jülich Supercomputing Centre. For rasdaman, data pre-processing scripts had to be developed and rasdaman had to be extended to support the Coordinate Reference System (CRS) of the COSMO-EPS data with rotated pole coordinates. The latter involved extensive discussions with core developers of the widely used geoinformation software packages GDAL and PROJ and resulted in a subcontract with Spatialsys and Geomatys to implement the new rotated grid CRS. Related to the data processing tasks was the development of big data workflows to build the end-to-end processing chain that is needed in an operational application. Various workflow concepts have been designed and demonstrator elements were written in the form of Jupyter notebooks, which provide flexible, user-friendly access to the required functionality and can easily be refactored into code that can be run automatically by batch systems.

The development of modern machine learning methods for the task of precipitation downscaling had to overcome specific challenges due to the highly non-Gaussian distribution of rainfall amount. Besides the necessity to carefully select suitable machine learning methods for this purpose, the non-gaussian data distributions had two important consequences: 1) the

overwhelming part of the data shows no precipitation (amount = 0), and this leads to a highly imbalanced training data set (Wahl, 2015); 2) extreme events, which are particularly relevant because of their high potential for damage, occur very seldomly so that there are very few samples available for the training of a data-hungry neural network. Rainfall statistics provided by DWD showed that most locations in Germany experienced at most one or two extreme rainfall events during a 20-year observation period. Mathematically, precipitation is best approximated with a right-skewed Gamma distribution (see e.g. Zolina et al., 2004 and Martinez et al., 2019). These statistical properties of precipitation data made it necessary to explore different loss functions in the deep neural networks, because the standard loss functions (typically mean square error) implicitly assume that the data are more or less normally distributed (see discussion in Schultz et al., 2021).

A third major aspect of DeepRain was the quantification of uncertainties and the development of suitable evaluation metrics to assess the quality of ML-based precipitation forecasts. To establish meaningful competitor models, classical downscaling methods (logistic regression, generalised linear model and analogue ensemble) were implemented on the Jülich HPC system and further enhanced. Furthermore, new statistical procedures based on partial correlation and entropy analysis were designed to investigate the robustness and quality of the ML forecasts and novel verification techniques (e.g. Dorninger et al., 2018) were adopted in the evaluation workflows.

In the following, we provide a summary of the main achievements within the six work packages of the DeepRain project.

Work Package 1: Project coordination - monitoring of progress, reports, organisation of project meetings (FZ Jülich)

Continuous monitoring and reporting of the progress was ensured with the help of regular meetings with all DeepRain partners throughout the project period (Table 2). In addition to the regular project meetings, which occurred roughly every 6 months, three special meetings were organised (see Table 2). The spring project meetings were scheduled about one month prior to the reporting deadlines to facilitate the collection of content. As classical in-person meetings became impossible after March 2020 due to the CoVid-19 pandemic, the consortium began to organise online meetings instead. This led to an increase in meeting frequency. While the 6-monthly project meetings were re-organized into three online sessions of 3 hours each, monthly meetings were held in addition to discussing open issues and ensuring progress. Finally, in September 2021, an in-person meeting became possible again, which was used to discuss the strategy and plans for the final project phase.

All partners delivered their regular financial and progress reports. As some partners faced difficulties finding suitable personnel at the beginning, some rearrangements of the budget and project deliverables had to be made. In February 2021, a major security incident at supercomputing centres across Europe caused serious delays in several DeepRain activities. Together with the initial delays in finding personnel and reduced efficiency in collaboration due to the home office regulations during the CoVid-19 pandemic, this prompted the consortium to request a cost-neutral project extension from 30th September 2021 to 31st March 2022. This request was approved in May 2021.

For external communication and dissemination, a project website (https://www.deeprain-project.de) was set-up via a service contract to a web development company located in Aachen. The website was regularly updated, including several news posts reporting on the progress and achievements in DeepRain and a continuously growing publication list. Postdoctoral and doctoral researchers on the project had various conference presentations at national and European meetings; for details, please see the DeepRain website. As described

in the project proposal, a meeting with DWD was arranged in February 2021 to explain the general DeepRain strategy and results and discuss a possible operationalisation of DeepRain results (SM03, see Table 3). Even though it became clear around that time that the project will not achieve the construction of a pre-operational prototype system, the discussions with DWD were regarded as highly informative and influenced the launch of a DWD seminar series on machine learning, which is now organised every 6 months.

Table 2: List of DeepRain project meetings

Meeting number	Date(s)	Location	Meeting number	Date(s)	Location
PM01	Nov 2018	Jülich	WM06	Sep 2020	online
SM01(*)	Jan 2019	Bremen	PM05/WM07	Nov 2020	online
PM02	Mar 2019	Osnabrück	WM08	Dec 2020	online
SM02(**)	Jun 2019	Dortmund	WM09	Feb 2021	online
PM03	Nov 2019	Bonn	SM03(***)	Feb 2021	online
WM01	Jun 2019	online	PM06/WM10	Apr 2021	online
WM02	Aug 2019	online	WM11	Jun 2021	online
PM04/WM03	Mar 2020	online	WM12	Jul 2021	online
WM04	Jul 2020	online	PM07	Sep 2021	Cologne
WM05	Aug 2020	online	PM08	Apr 2022	Osnabrück

PM: project meeting, WM: web meeting, SM: special meeting
(*) Rasdaman training course, (**) Statusseminar (***) Information exchange with DWD

Work Package 2: Data processing and provision (Jacobs U Bremen and FZ Jülich)

Earth system science and deep learning (DL) both constitute research fields which are inherently related to big data. While observational as well as modelling data of the Earth system typically comprise datasets in the order of several Tera- or Petabytes, the performance of neural networks on specific applications heavily depends on the amount of available (training) data. Thus, data management, preparation and staging constitute key challenges in the DeepRain project, where extensive model and observational datasets are applied to improve local precipitation forecasts with the help of DL methods.

Particularly, the forecasts from the COSMO ensemble prediction system (COSMO-EPS) as well as the rain-gauge adjusted radar observations of the RADKLIM dataset, are the main data sources within the DeepRan project. COSMO-DE EPS constitutes the former ensemble prediction system, which ran operationally at DWD from May 2011 to May 2018[1]. The COSMO-DE model constitutes a convection-permitting, regional numerical weather prediction (NWP) model, which deploys a rotated pole grid with a horizontal spacing of 2.8 km and with 50 vertical layers. The model domain covers all Germany and also includes parts of the

[1]https://www.dwd.de/DE/leistungen/nwv_cosmo_de_eps_aenderungen/nwv_cosmo_de_eps_aenderungen.html

surrounding countries. COSMO-DE EPS then consists of 20 model members based on varying boundary conditions and perturbed physics, which create probabilistic 27-hour forecasts eight times per day (every three hours). This rapid update cycle allows quick assimilation of recent observations such as radar data which is highly valuable for short-range forecasts of convective precipitation.

In May 2018, COSMO-DE was replaced by its successor, COSMO-D2, which deploys a finer grid spacing of 2.2 km, more vertical layers and an increased model domain. COSMO-D2 EPS was then run operationally at DWD before it was replaced by ICON-D2 in February 2021.

The COSMO model data is complemented by the RADKLIM dataset which constitutes a radar-based precipitation climatology based on the RADOLAN method. The “RW” product of the RADKLIM dataset provides gridded accumulations of precipitation on an hourly basis. The data result from a blending of radar reflectivity with rain gauge measurements. Several corrections are applied to remove clutter pixels, to compensate for shading effects due to mountains and to remove other artefacts from the raw data. The RADKLIM dataset also comprises the “YW” product which provides five-minute precipitation rates. The YW product is disaggregated from the RW data and is therefore referred to as a quasi rain-gauge adjusted precipitation product. The data is available on a 1 km-grid using a polar stereographic projection and covers the time span of 2001 until 2021.

The proposal envisioned the development of a holistic workflow in which the data are integrated into rasdaman datacubes, which then allows subsequent data streaming to train deep neural networks. However, due to delays in the development of the machine learning models, we decided to split the workflow into several segments that could be pursued at the University of Bremen and at JSC separately. The first segment organised the data retrieval of COSMO model data and RADKLIM radar data from DWD. The second segment involved the set-up, maintenance and improvements of the rasdaman server. Finally, in the third segment, the University of Bremen and JSC collaborated to develop some components for a prospective future coupling of rasdman to the ML workflows. The following sections provide a detailed description of the respective work accomplishments.

a) Retrieval of DWD data, data management and data processing

To allow direct access of all project partners to the aforementioned datasets on the HPC facility of Jülich, data transfer of 650 TBytes from DWD to the Jülich Supercomputing Centre (JSC) constituted a necessary prerequisite, which was undertaken jointly with WP6. As a hardware incident at JSC in early 2021 led to corruption of multiple data racks, the data had to be transferred twice. Thus, in total about 1.3 Petabytes of data were transferred from DWD to the Juelich filesystem JUST through the X-Win net of the DFN. The proof that data transfer over the internet on this scale is possible constitutes an important milestone of the DeepRain project with implications for the system design of future weather and climate services, e.g. in DestinE.

While a lot of the monitoring of the data transfer had to occur manually during the first transmission, the procedures could be largely automated for the second transfer. This was in part due to a hardware upgrade at DWD, allowing for larger chunks of data to be stored temporarily on spinning disks, but also because of a better understanding of data structures and development of functional processing scripts. Together, these changes made it possible to re-transfer all data within 2 weeks instead of 3 months, as during the first transfer. DWD granted Forschungszentrum Jülich the permission to store a copy of the data and make it publicly available. Therefore, all COSMO-EPS data can now be made available upon request via the MeteoCloud hosted at JSC in the original format and some datasets are publicly available through the rasdaman Earth Server federation.

Storage of the RADKLIM data was a comparatively minor problem. The data were retrieved from DWD's OpenData server and converted from the original binary data format to the more convenient netCDF format. This resulted in a total data volume of 4 TBytes.

b) Data fusion and development of high-performance big data workflows

The DeepRain data constitute multidimensional data, which are naturally modelled as data cubes. Data cubes with up to 6 dimensions have been established and processing scripts were developed to insert them into the rasdaman database system. Besides latitude, longitude, vertical level and time a second time axis was defined for the forecast lead time, and an ensemble dimension to capture the 20 model simulations for the same forecasting time. Two instances of the rasdaman data cubes have been configured and deployed at JSC. Other geodata, such as digital elevation models (DEMs), are also stored on these instances, and this allows for an ad-hoc fusion of such data during the rasdaman query. The instance of the data cube that houses COSMO-DE EPS precipitation forecasts has been integrated into the EarthServer datacube federation.[2] This integration allows distributed data fusion of climate, remote sensing, and Copernicus data from several data centres (Fig. 1).

The incorporation of COSMO data into rasdaman required the additional development of a new Coordinate Reference System (CRS) to support the "rotated pole" grid of the COSMO model. It was therefore necessary update the rasdaman software and to subcontract some work to Spatialys and Geomatys for the implementation of the new grid definition into common open-source tools (GDAL 3.4.0, PROJ 8.2.0). This remapping onto an unrotated geographical grid was necessary to avoid interpolation artefacts when using the data in machine learning applications. A rotated Grid CRS definition will also be published via OGC in the near future (see github).

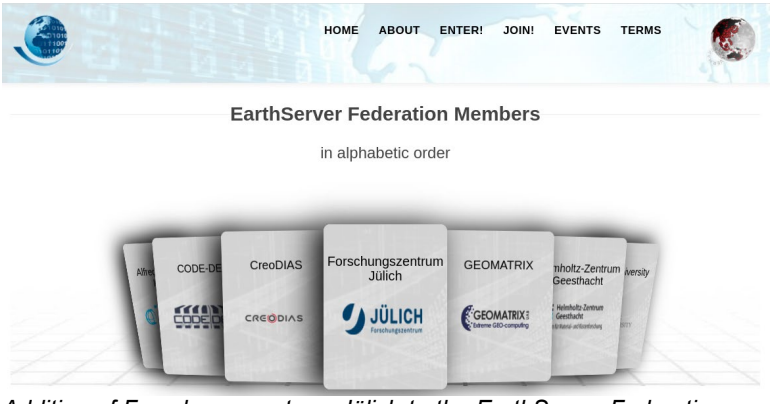


Figure 1: Addition of Forschungszentrum Jülich to the EarthServer Federation members list. (Source: https://earthserver.eu/)

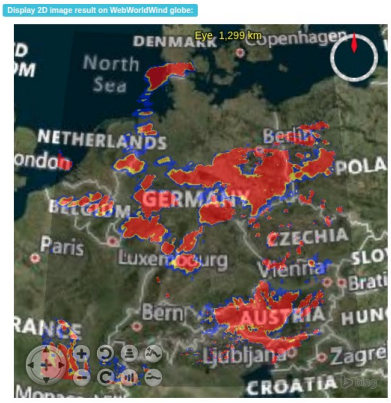
To speed up data processing, the Rasdaman data cube engine has been improved with respect to data import and retrieval. This led to improvements of a factor of 15 for data ingestion. On average, ingesting a one-year dataset with five variables now requires less than 24h on the JSC cloud server. The Rasdaman performance was also assessed by the Free University of Bozen-Bolzano (see on gitlab).

[2] http://fz-juelich.earthserver.xyz/rasdaman/ows#/services

We also developed Jupyter-Notebooks and a dashboard to assist data retrievals from the rasdaman datacube and to help users in the visualisation of the COSMO-EPS data (Fig. 2). In the course of the DeepRain project, the experience and best practices with data cubes were documented in the corresponding git-project to enable efficient re-use of the project's developments.

Another Rasdaman development in the context of the DeepRain project was the definition of user-defined functions (UDFs) to derive higher-level aggregated variables from high-resolution digital elevation models, such as hill slope and terrain roughness. These functions are available on the DeepRain project repository on github.

The achievements of work package 2 were documented in several journal articles and conference papers (see Section 5).




```
wwd(44.9875,2.956405722,56.2625,17.735782202)>>for $c in ( COSMO_DE_EPS__deeprain__precipitation_data )
let $dom := $c[ansi("2016-06-01"),ensemble(2), forecast_hour(10),
                Lat(44.9875:56.2625),Long(2.956405722:17.735782202)]
return
 encode(
  switch
   case $dom=0
    return {red: 0; green: 0; blue: 0; alpha: 60}
   case 0.01 > $dom
    return {red: 0; green: 0; blue: 255; alpha : 127}
   case 0.03 > $dom
    return {red: 255; green: 255; blue: 0; alpha : 127}
   case 0.05 > $dom
    return {red: 255; green: 140; blue: 0; alpha : 127}
   default
    return {red: 255; green: 0; blue: 0; alpha : 127},
"image/png")
```

Figure 2: Precipitation map for Germany and its surroundings on 2016-06-01 using WCPS and NASA World Wind from the Rasdaman endpoint at Jülich

c) Development of a deep learning workflow

As DeepRain was supposed to build a prototype for an end-to-end future weather forecasting system based on deep learning technologies, substantial effort was devoted to the development of concepts for exploiting large-scale data cubes as user-friendly repositories for terabyte-scale datasets, including on-demand processing capabilities. Based on the rasql interface to rasdaman, we have developed an HPC-oriented workflow for extracting climate and weather data as input data in machine learning models. This has been implemented as Jupyter notebooks, which can be executed on the HPC-aware Jupyter-JSC system. The workflow allows for the merging of RADKLIM radar data with the COMOS-EPS data and includes an automated remapping during the extraction procedure.

Work Package 3: Method development for the machine learning (U Osnabrück, FZ Jülich, DWD)

Developing DL techniques with a deep knowledge of the climate and weather domain was one of the main goals of the DeepRain project. The machine learning solutions obtained from work package 3 concentrated on two highly relevant meteorological problems related to precipitation forecasting. The first problem was the post-processing of the COSMO-EPS to offer refined and more accurate precipitation predictions at specific locations, i.e. stations with rain gauge measurements. Two types of predictions were developed: dichotomous rain / no rain predictions and rain amount predictions. After substantial investigation of suitable neural network architectures and extensive hyperparameter tuning, the artificial neural networks (ANNs) showed vital skills to improve the precipitation quality in both types of predictions.

The artificial neural network (ANN) predictions yielded superior results compared to the classical statistical evaluation methods after substantial investment in hyperparameter tuning and architecture optimisation. Figure 3 shows the quantitative comparison between ANNs and statistical post-processing (logistic regression and generalised linear model fitting) for point-based forecasts at four selected stations in Northern Germany. The ANN results generally show a substantial improvement compared to the statistical postprocessing metods (positive skill scores). Our results suggest furthermore that training the ANN with data from all stations improves the generalisation capabilities of the ANN.

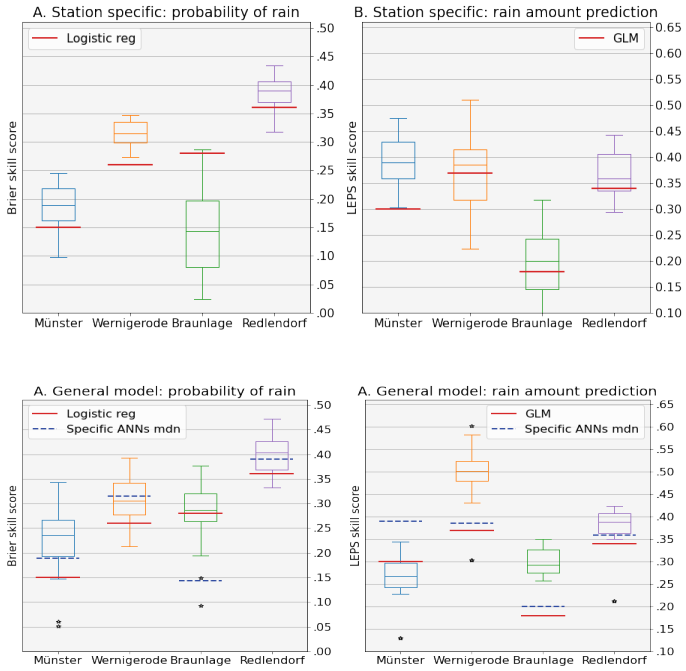
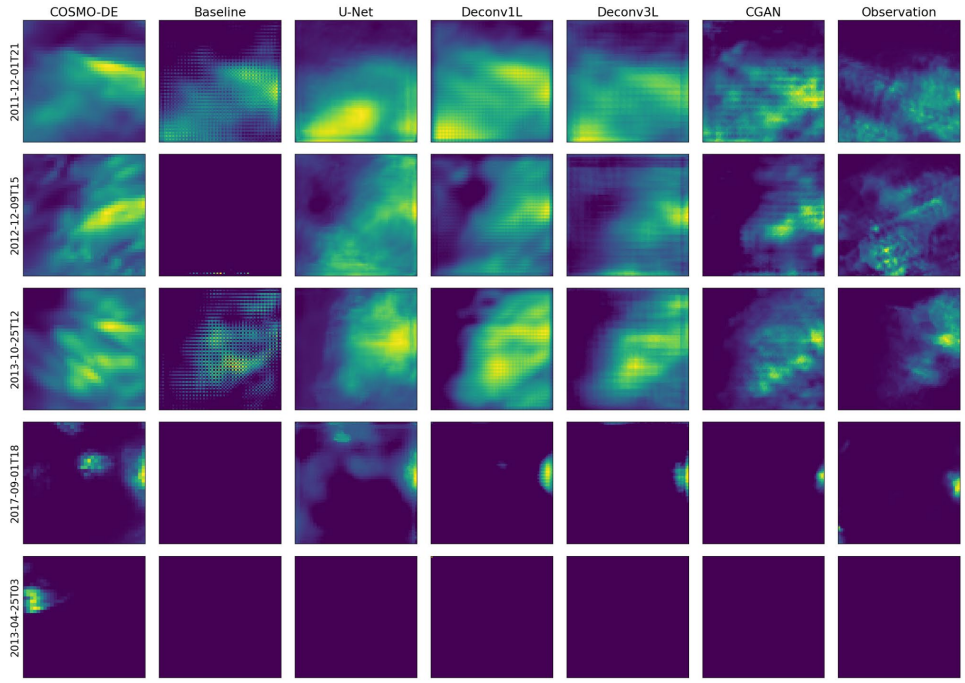


Figure 3: Performance evaluation in terms of the Equitable Threat Score (ETS) for precipitation events (first row) and the Linear Error in Probability Space (LEPS) of the precipitation amount (second row). The performance of a logistic regression and a generalized linear model (GLM) is compared against the best-performing artificial neural network. The models in the first row are trained independently for each station site, whereas models in the second row are generalized over all stations.

The second scientific problem addressed in WP 3 was the spatial downscaling (“super resolution”) of the precipitation forecasts offered by COSMO-EPS. We developed an innovative approach that combines the super-resolution task with bias correction, offering high resolution and corrected precipitation maps for the area of interest. To achieve this, we trained and tested different deep learning algorithms using TensorFlow (Abadi et al., 2015) and developed a processing chain on the FZ Jülich HPC systems in collaboration with WP6.



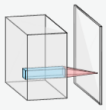
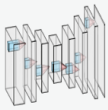
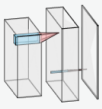
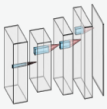
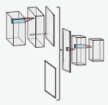
					
Name	Baseline	U-Net	Deconv1L	Deconv3L	CGAN
Layers	2dDeconv(1)	See Appendix A	2dDeconv(32) BatchNorm Conv(1)	Max pooling(2) 2dDeconv(32) BatchNorm 2dDeconv(16) Conv(1)	Deconv3L + discriminator
Kernel size	7x7		7x7	5x5	5x5
Stride	2		2	2	2
Parameters	7,008	153,849	224,673	127,841	127,841

Figure 4: Top: comparison of the COSMO-DE output (left column) with various deep learning models (middle columns) and the RADKLIM observations (right column). Each row represents an independent sample, i.e. a precipitation event. Bottom: Summary of the deep learning models developed and tested for radar downscaling

Various convolutional neural network (CNN) based architectures were explored (Figure 4). The best results were obtained with the Conditional Generative Adversarial Network (CGAN) architecture (Rojas-Campos et al., 2022). The algorithmic approach involves a non-linear combination of COSMO-EPS variables and exploitation of the ensemble statistics. With this, improved precipitation maps with 3 hours lead time could be generated at the spatial resolution of 1.4 km pixel size, i.e. a factor of two refinement compared to the original COSMO grid resolution of 2.8 km. Note that the baseline model (second column in the top plot of Figure 4) has the fewest trainable parameters and clearly performs worst. This demonstrates that certain network complexity is needed to describe precipitation patterns correctly.

Discussions with DWD during an information meeting in February 2021 showed a general interest in adopting the DeepRain super-resolution approach for operational post-processing after further evaluation, as it demonstrated superior performance compared to the classical statistical methods. The trans-disciplinary collaboration in DeepRain, therefore, allowed reaching the major objective of developing practicable machine learning solutions for high-resolution precipitation forecasts. However, such an operational implementation would still require substantial preparatory work to adapt the method for use with the ICON model, which constitutes the current operational weather model at DWD.

Work Package 4: Data consistency and classical downscaling algorithms (DWD, U Bonn)

The objectives of this WP were twofold. First, DL methods rely heavily on the consistency of the underlying statistical patterns of the input data to extract meaningful patterns that can be used in the post-correction of the precipitation forecasts. Thus, data consistency influences the quality of the predictions. Therefore, we carried out comprehensive quality control on the COSMO-DE EPS to identify data gaps and inconsistencies. Together with work package 5 (see below), an overview of missing data, mixed-up time series or incomplete GRIB time stamps was established and the database was cleaned accordingly.

The second objective of WP4 was to provide benchmarks or reference results for the DL approach through implementation and further development of more conventional statistical methods. For point-based observations, we implemented and tuned a classical logistic regression (LR) approach for dichotomous forecasts and a generalised linear model (GLM) for precipitation amount forecasts. We first employed the LASSO technique in order to identify relevant input parameters for correcting precipitation predictions with both methods. As input parameters, we used a 5x5 grid point neighbourhood from the COSMO-EPS forecasts around the point in question (i.e. the station location) to allow for the GLM to account for information from the local environment of the station. The point-based approaches proved to be efficient in comparison to the COSMO-DE-EPS forecasts, but they performed slightly worse than the ANN network as discussed above (Figure 3). More detailed information on the algorithm and results is provided in the aforementioned publication currently under review (Rojas-Campos et al., 2022).

For the spatial precipitation forecasts, we first adopted the analog ensemble (AnEn) technique to forecast 2-dimensional fields. The AnEn uses a predefined metric to rank all data points in the training data set with respect to their similarity to the current situation. Then for the best matching cases (i.e. the cases with the most similar weather patterns), the corresponding observations (here, radar precipitation estimates from the RADKLIM data set) are taken as the members of the AnEn. To allow for more variability in the spatial output fields we used a tiling approach. While the AnEn is able to produce more fine-grained estimates than other classical methods, the available training data set is too small for reliable precipitation

forecasts. Despite substantial efforts to tune the AnEn (various settings for tilings size, metric, etc.) and even with the use of wavelets (see below), the AnEn produced on average similar or only marginally better results than the original COSMO-DE-EPS model and performed worse than the CGAN model described above.

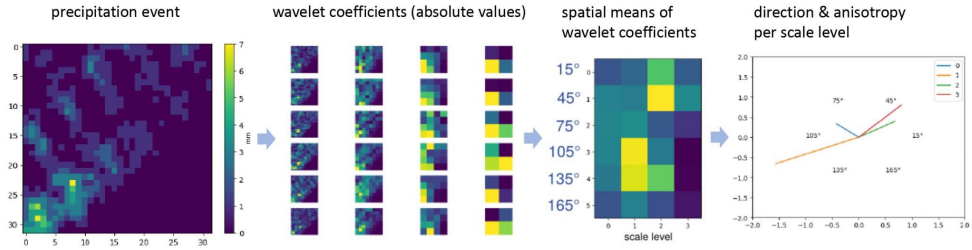


Figure 5: Exemplary illustration of the derivation of structure information for the analog ensemble method using wavelet transforms

In order to provide a reference data set for a spatial DL downscaling, we then used a pixel-wise GLM approach similar to the one described above for time series. Then, we enhanced this approach by deriving wavelet coefficients from the COSMO-DE-EPS data as predictors in the GLM to account for spatial correlations in the data. The wavelet coefficients are determined from the precipitation field (or a tile of the field) using a dual-tree complex wavelet transform. Then the spatial means for the six different orientation angles and four different spatial scale levels are calculated. To further reduce the complexity, we determined the values of direction and anisotropy per scale level (cf. Figure 5).

An exemplary comparison of the ground truth (RADKLIM data), numerical model (COSMO-DE-EPS), GLM wavelet and the DL methods is shown in Fig. 6. While the results indicate some performance gain especially with respect to small scale structures, the tiles (and their edges) still remain quite visible, while the DL model described in WP3 offers a continuous field which resembles the target RADKLIM data reasonably well.

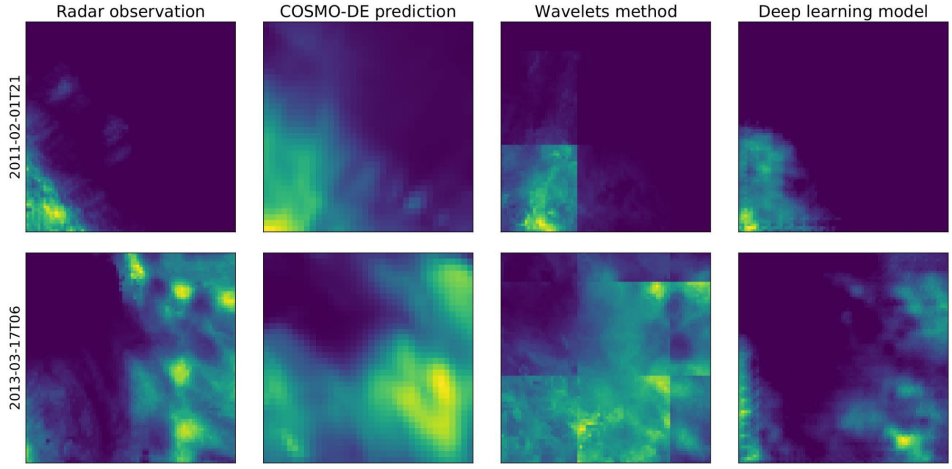


Figure 6: Comparison of the analog ensemble method and the CGAN neural network for two precipitation episodes. Left column: RADKLIM radar observations, second coumn: original COSMO-DE model output, third column: GLM wavelets prediction, and right column: prediction from the deep learning model (CGAN).

Work Package 5: Evaluation and comparison of results (U Bonn, DWD, FZ Jülich)

The collection of the entire COSMO-EPS dataset in Jülch allowed for a rigorous assessment of the data quality and consistency (see also WP4). As the analysis tools of WP5 were tailored to work with the original GRIB files from DWD, WP5 established a catalogue of missing data, mixed-up time series, and incomplete GRIB timestamps in the early phase of the project. Together with WP4, a clean dataset was then created.

This cleaned COSMO-DE EPS dataset was then used to develop a first post-processing system consisting of a logistic regression model (GLM whose predictand follows a Bernoulli distribution) for threshold exceedance of 3 h precipitation (e.g. 0.5 mm/3h) at selected observation sites. Furthermore, a set of suitable predictor variables from the COSMO-EPS was deduced, which are presented in the Table 3.

Table 3: Predictor variables for a generalised linear model regression

Grib code	Variable	Space-time dimensions	MKS unit
113	Convective Available Potential Energy	2-dim, instant	J kg^{-1}
129	Convective Inhibition Energy	2-dim,instant	J kg^{-1}
137	Cloud Cover (0 - 400 hPa)	3dim, instant	%
161	Total Cloud Cover	2-dim, instant	%
170	Geopotential	3-dim, instant	m^2 s^{-2}
273	Vertical velocity	3-dim instant	Pa s^{-1}
361	Surface pressure	2-dim instant	Pa
385	Relative humidity	3-dim instant	%
458	2 metre temperature	2-dim instant	K
540	atm. Temperature	3-dim instant	K
561	Total Precipitation	2-dim accum	kg m^{-2}
665	10 metre U wind component	2-dim instant	m s^{-1}
721	10 metre V wind component	2-dim instant	m s^{-1}

The post-processing models developed in DeepRain were evaluated systematically with the help of partial correlations based on entropy values. This evaluation technique allows us to decompose the target information (i.e. the prediction from the post-processing model) into its correct redundant, its false redundant and its added value parts. The decomposition has been expanded by the false redundant part in the study of Glowienka-Hense et al. (2020) as inspired by Williams and Beer (2010) (Fig. 7).

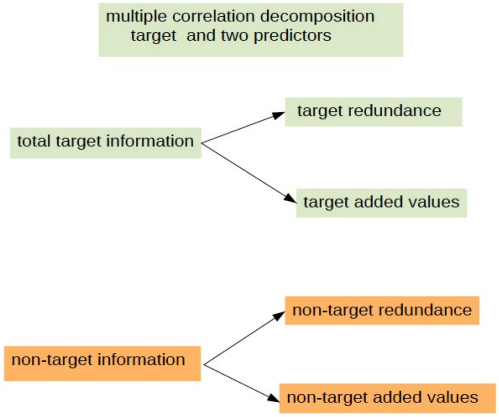


Figure 7: Graph of the novel multiple decomposition technique for comparison of two forecast systems vs observations (=target) as an extension to Glowienka-Hense et al. (2020). In general, information is measured by negative entropy, in the case of Gaussian random variables, it can be replaced by variance. Target redundancy is observed information/variance jointly predicted by both forecast systems, target added values are the increase of information/variance of one forecast system relative to the other, and non-target redundancy is the information/variance in both forecast systems not represented in the observations ("joint forecast error"), non-target added value is the information/variance in one forecast system over the other not represented in the observations.

Further work of WP5 concerned refinements of the underlying statistical regression approaches classical logistic regression and DL. Single station analysis showed in the cases of high predicted probabilities (>50%) of the logistic model that the ratio of correct to false positive events is much higher than for a COSMO grid point in the neighbourhood of that station. A misfit was calculated analogously to the Brier Skill Score BSS, but with the probabilities from the logistic regression set to zero and one if they were less than or larger than 0.5, respectively. This number is thus more stringent than the BSS and is a measure of the relative number of incorrect yes/no decisions based on the logistic model compared to the reference case. A negative value means that correct predictions are made more frequently with the assumption of no precipitation. The total number of precipitation events correctly predicted by the model to the total number of all precipitation events is typically around 0.4, so the logistic regression model captures only a portion of precipitation events. To increase this proportion, further logistic models were sought, following the method of principal component analysis (Anderson, 1984, p.454). For this purpose, the time points classified as precipitation events by the first model were eliminated from the data. This method is called boosting in the field of machine learning (Friedman et al. 2000). As a result, it is recommended to use the output of the logistic model only in the case of the forecast of a precipitation event.

For those cases of predicted precipitation events, the probability of occurrence is generally increased by boosting. A third logistic regression boosting step was also tested after eliminating the events for which the second regression predicted a precipitation probability greater than 50%. Again, this boosting step yielded a similarly informative model. This gives rise to the so-called conditional evaluation: „Given that threshold, exceedance is predicted", the probability of observed precipitation is very high. However, the regression models based on the deep learning neural network ML approach (NNM) cannot be treated in the same way as the logistic regression models with the boosting procedure. But an ensemble interpretation of the stochastic single models NNM obtained by initialising the optimisation procedure with different random conditions can be treated such that this leads to similar results as the boosted

logistic regression model. Comparing the logistic regression with boosting and the ensemble interpretation of the NNM, a missing predictability gap is identified: it appears that multiple solutions (Nandwani et al., 2020; Holzinger et al., 2021) to the regression problem might exist in the sense that the same threshold of precipitation is exceeded by different meteorological flow configurations, a feature which is well known to any operational weather forecaster.

Further work in WP5 was devoted to the definition of the loss function, especially in the case of binary observations like threshold exceedances of precipitation. The approach is based on a strict application of Bayesian statistics to evaluate the posterior (conditional) probability of a forecast system at a fixed lead time given the set of binary observations. It combines a Gaussian model prior to describing the forecast ensemble with a Bernoulli-type probability mass function as the likelihood for the binary events. The posterior is obtained by a high dimensional integral over the phase space of the forecasts. To avoid a costly Markov Chain Monte Carlo evaluation of such integrals, the saddle-point/Laplace approximation (Reich and Cotter, 2015) is proposed instead, which can be calculated semi-analytically. It defines as a loss function the free energy of the posterior density. The approach can be generalised to any likelihood, which can be written as an exponential family probability density function (Wainwright and Jordan, 2008).

Work Package 6: System design and workflow analysis (FZ Jülich and Jacobs U Bremen)

The initial work of WP6 concentrated on transferring the DWD COSMO-EPS and RADKLIM datasets to the JSC storage system (see description in WP2 above).

To facilitate the handling of massive data as incurred during DeepRain, we developed an efficient and easy-to-deploy scalable software in Python for parallel computation on HPC systems (PyStager, see github). PyStager is now further developed in the EuroHPC-JU project MAELSTROM. After the data incident in the spring of 2021, a monitoring workflow based on PyStager was developed to detect the potential corruption of data blocks. This helped to avoid data errors during the second data transfer in May 2021 (see WP2 above). PyStager was also used to pre-process and remap the RADKLIM dataset to facilitate merging with COSMO-EPS data.

The primary goal of work package 6 was to design and assemble an end-to-end workflow for the handling of terabyte-scale data in ML applications in weather and climate. Several concepts were developed, which guided the work in WP2 and WP3, and software tools were developed to assist the project partners with efficient data handling (Fig. 8). Together, the workflow components developed in DeepRain cover most aspects of the end-to-end precipitation forecast application that was envisioned in the proposal. However, integration of these components into one coherent software framework could not be achieved due to delays in other work packages and eventual lack of time.

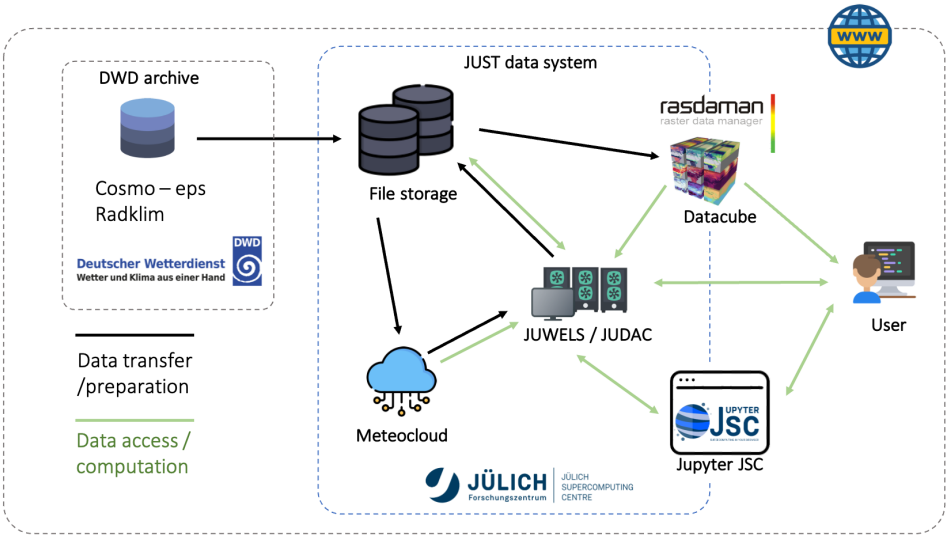


Figure 8: Data transfer, preparation, user access and computation patterns for DeepRain data

To go beyond classical workflow development, we investigated the potential of modern software concepts to enhance FAIRness of meteorological data processing workflows. FAIRness aims at making data and metadata findable, accessible, interoperable, and reusable (Wilkinson et al., 2016). We proposed a novel concept based on FAIR digital objects (FDO) (De Smed et al., 2020). This concept makes use of the cutting-edge Jupyter-JSC infrastructure in conjunction with the gitlab server (Fig. 9; Mozaffari et al., 2022).

As described in Mozaffari et al. (2022), Jupyter notebooks as such are not well suited for reproducible workflow design as their state and content changes with every user interaction. Therefore, version control, and thus documentation of changes, is provided by the gitlab server. A dashboard for machine learning experiments, similar to modern tools like Weights and biases or MLFlow acts as the FDO. Distributed launchers (for example, using the Python library papermill) initiate a suite of machine learning experiments with pre-defined sets of parameters coming from the dashboard. As a consequence, easy-to-use and reproducible workflows can be built and shared among fellow researchers.

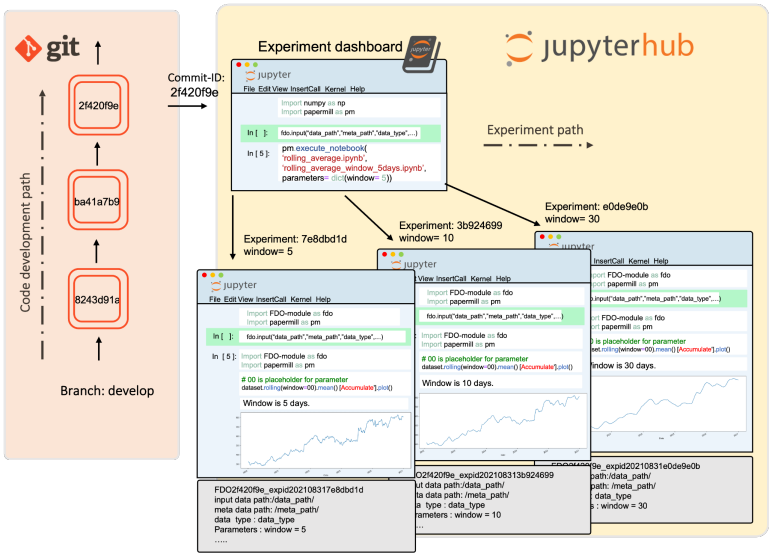


Figure 9: Integration of git and experiment dashboard utilising papermill to execute each experiment in an individual instance of the notebook with passed parameters and the creation of corresponding FDO (Mozaffari et al., 2022)

3) Outcomes and future perspective

The DeepRain project demonstrated the potential of modern deep learning methods to improve local-scale precipitation forecasts based on ensemble prediction forecasts from a numerical weather model, radar and station observations. It also showed the importance of exploiting state-of-the-art IT technologies for Big Data management and parallel processing. In contrast to the original planning, it was impossible to develop all components into one coherent service architecture that could be put into operation at DWD. Nevertheless, the project did lay the groundwork for this, so that with sufficient funding available, modern machine learning concepts could become part of the operational weather forecasting in Germany in the near future.

Necessary steps to accomplish this objective include:

- Adaptation of the DeepRain data processing chain to the operational ICON model at DWD,
- Adaptation of the DeepRain data processing chain to the hardware architecture and storage systems at DWD,
- Retraining of the deep learning models of DeepRain with ICON data,
- Further improvements and evaluation of the deep learning models employed in DeepRain,
- Development of robust, automated output processing routines,
- System integration and testing.

Besides the potential to operationalise DeepRain components, many of the DeepRain results, software packages and data processing strategies will be used in forthcoming academic research and teaching. There is already an uptake of DeepRain results in the EuroHPC-JU project MAELSTROM and the ERC Advanced grant IntelliAQ. The interfacing of rasdaman with machine learning workflows that was started in DeepRain is already continuing in the follow-up projects EU H2020 CENTURION, BMWI AI-Cube, and EU H2020 FAIRiCUBE. The rasdaman servers and EarthServer federation node installed in DeepRain are now also used in the Rhenish bioeconomy region.

In detail, the DeepRain project partners plan to use DeepRain tools and results as follows.

UBonn and DWD:

Advance the conventional forecasting method through a combination of approaches following suggestions by Buschow and Friederichs (2021) and Brune et al. (2021) on the use of wavelets to describe the spatial structure of multiscale fields like precipitation in an effective way.

All methods and results of the DeepRain project will be integrated into future courses of the Master's study program “Physics of Earth and Atmosphere” at University Bonn. This will not only cover the necessary theoretical background but can also involve computer lab exercises using COSMO data sets and DWD observations by synop stations and RADOLAN retrievals. Jupyter notebooks with R or Python are already a standard teaching tool in the study programs at University Bonn, easing the transition from the research character of the DeepRain results into the teaching use.

U Osnabrück and FZ Jülich:

Further pursue the development and application of deep neural networks for precipitation forecasting and downscaling in the context of MAELSTROM and other projects. Different variants of GAN models will be explored as well as the new deep learning method of Swin Transformers (see, e.g., Liang et al., 2021 and Zhang et al., 2022). The methods will also be applied to forecast other meteorological fields like temperature or wind. The application of GANs will be extended to a probabilistic framework to account for the chaotic nature of precipitation processes (Gilleland et al., 2009). FZ Jülich also plans to investigate if the inclusion of physical constraints to the optimisation procedure can help to ensure that the generated precipitation forecast products remain realistic.

The HPC-ready machine learning workflows and data processing pipelines developed in DeepRain will be used in related research projects and in training courses.

Jacobs University:

will continue to lead and drive the open-source rasdaman community project to enhance the use and foster uptake of the data cube technology. Additionally, the EarthServer federation curated by Jacobs University will be further improved and extended. In particular, the integration of data cubes within AI workflows will be further explored.

Jacobs University and FZ Jülich:

are pursuing further work on large-scale climate datacube services in the Digitales Geosystem Rheinisches Revier (DG-RR). The DG-RR project's first datasets are already integrated into the federated rRasdaman node and are available on the Earth Server federation. These can be easily retrieved using the Web Coverage Service (WCS) and Web Map Service (WMS) standards and processed using the Web Coverage Processing Service (WCPS) standard,

thus greatly decreasing the data extraction and processing times and thereby reducing the skills barrier for users.

FZ Jülich:

In the context of the workflow design, HPC system capability, development and implementation of the DeepRain project in the JSC infrastructure showed the possibility of an end-to-end ML data pipeline. The recently funded BMBF project WarmWorld will explore a case of an operational data storage system inspired by the MARS system already in use at ECMWF. In addition, integration of the FREVA software will bring the concept of the federated data centre to a new level, connecting the data centres of ECMWF, DKRZ and Jülich.

Novel concepts for reproducible large-scale data analysis workflows conceived during DeepRain shall be further developed in follow-up research projects. Specifically, it is planned to develop functional software to test the FAIR Digital Object concept in real-life workflows concerning Earth system data (Mozaffari et al., 2022b).

4) External developments in the field during the project's lifetime

During the lifetime of the DeepRain project, many advances in ML applications for weather and climate, parallel processing of the big data, and Earth system data services occurred.

Open access, large-scale Earth system data services have been set up at various meteorological agencies and research institutions worldwide. Examples are: the Climate Data Store of ECMWF, ESA, NASA and NOAA. Also, the German Weather Service opened a substantial portion of their archive and made it web accessible through the DWD geo portal. While these sites generally provide good functionality to search, visualize and download large datasets, they offer minimal on-demand processing capabilities, primarily when deriving new information from several input variables. This is the particular strength of data cubes, as exemplified in the EarthServer datacube federation service. Specifically, on array databases, a comprehensive analysis of a large number of approaches has been published by Jacobs University (Baumann et al, 2021).

Significant technical enhancements have occurred in the context of the parallel data processing for weather and climate data. ECMWF produces approximately 120TBytes of raw daily weather data from high-resolution and ensemble forecasts. A tailor-made data handling solution by ECMWF, called Meteorological Archival and Retrieval System (MARS), enables users to explore and retrieve meteorological data in GRIB or NetCDF from the massive archive. The raw data are also stored in MARS (the world's largest meteorological archive, currently holding over 300 PBytes of primary data). In Germany, the research project RegIKlim, funded by the Federal Ministry of Education and Research (BMBF), aims to develop decision-relevant knowledge on climate change in municipalities and regions and create a sound basis for regionally specific information and evaluation services. RegIKlim created the Free Evaluation System Framework (FREVA), a standardised data and evaluation system - developed at the DKRZ and FUB in Germany. Freva offers efficient and comprehensive access to the model database and evaluation data sets. The application system is conceived as an easy-to-use low-end application minimising technical requirements for users and tool developers.

The field of deep machine learning continued its exponential growth, and especially ML applications in the weather and climate domain received increased attention by leading deep learning method developers, including giant corporations such as Google, Amazon, Microsoft, or NVidia. While at the start of the DeepRain project, convolutional neural networks (CNNs) were considered state-of-the-art, in the meantime, generative adversarial networks (GANs) have taken their role (see e.g. Price and Rasp, 2022; Harris et al., 2022, and Jeong and Yi, 2022). Combining ensemble predictions from NWP with deep neural networks leads to significant improvements in forecast quality (Grönquist et al., 2020). Moreover, graph neural networks have also shown promises of a significant breakthrough in forecasting global weather (Keisler, 2022). A recent development is tapping into the enormous capabilities of extremely large transformer models (Vaswani et al., 2017). With sufficient training data, these models can learn a general representation of the atmosphere, which can then be exploited in a variety of so-called downstream applications. Precipitation downscaling is one of them. Jülich is involved in the atmorep initiative, which is developing a prototype transformer model based on ERA5 data (Hoffmann and Lessig, 2022).

During the DeepRain project, various improvements of conventional weather and climate forecasting occurred. ECMWF adopted a precipitation post-processing method (EcPoint) that uses techniques similar to an Analog Ensemble (Hewson et al., 2021). Further improvements in verifying the precipitation forecast and error comparison (Stein, J., & Stoop, F., 2019; Buschow and Friederichs, 2021), and enhancement in statistical post-processing for weather forecasts were published (Vannitsem et al., 2021). As the current trend of the analogue ensemble technique and DL application for precipitation forecasts (Sha et al., 2022) suggests, a hybrid analogue ensemble for precipitation forecast will emerge in the coming years.

In the field of FAIR practices in research, Fair Digital Objects (FDO; De Smed et al., 2020) have been designed to enhance reproducibility of scientific studies. FDOs provide data, code and documentation as a single independent and self-explanatory piece of information. Through the use of persistent identifiers (PID), data, metadata and machine-actionable capabilities can be bundled together and archived analysis workflows can be reproduced. FDOs are widely expected to become a new standard for Open Data repositories. It is expected that the government of the Netherlands will recognize FDO as the new standard of data in October 2022. Other community-driven efforts toward granular FAIR segments, such as RO-CRATE, have been expanding recently (Soiland-Reyes et al., 2021).

5) Publications resulting from the DeepRain project

Work package 2:

- Baumann P., (2021), "Towards a Model-Driven Datacube Analytics Language," IEEE International Conference on Big Data (Big Data), pp. 3740-3746, https://doi.org/10.1109/BigData52589.2021.9672038. [**veröffentlicht**]
- Baumann, P. (2021). A General Conceptual Framework for Multi-Dimensional Spatio-Temporal Data Sets. Environmental Modelling & Software. 143. 105096. https//doi.org/10.1016/j.envsoft.2021.105096. [**veröffentlicht**]
- Baumann, P., Misev, D., Merticariu, V. et al. (2021), Array databases: concepts, standards, implementations. J Big Data 8, 28. https://doi.org/10.1186/s40537-020-00399-2 [**veröffentlicht**]
- Campos Escobar O. J. , Misev D. and Baumann P., (2020), "Making an Array Database Language Server-Side Extensible," IEEE International Conference on Big

Data (Big Data), pp. 2743-2750, https://10.1109/BigData50022.2020.9378108. [**veröffentlicht**]
- Villarroya S. and Baumann P. (2020), "On the Integration of Machine Learning and Array Databases," IEEE 36th International Conference on Data Engineering (ICDE), pp. 1786-1789, https://10.1109/ICDE48307.2020.00170. [**veröffentlicht**]
- Villarroya, S., Baumann, P. A survey on machine learning in array databases. Appl Intell (2022). https://doi.org/10.1007/s10489-022-03979-2. [**veröffentlicht**]

Work package 3:
- Kesselheim, S., Herten, A., Krajsek, K., Ebert, J., Jitsev, J., Cherti, M., ... & Lippert, T. (2021, June). JUWELS Booster–A Supercomputer for Large-Scale AI Research. In *International Conference on High Performance Computing* (pp. 453-468). Springer, Cham. https://doi.org/10.1007/978-3-030-90539-2_31 [**veröffentlicht**]
- Schultz, M. G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L. H., ... & Stadtler, S. (2021). Can deep learning beat numerical weather prediction?. *Philosophical Transactions of the Royal Society A*, *379*(2194), 20200097. https://doi.org/10.1098/rsta.2020.0097. [**veröffentlicht**]
- Gong, B., Langguth, M., Ji, Y., Mozaffari, A., Stadtler, S., Mache, K., & Schultz, M. G. (2022). Temperature forecasting by deep learning methods. *Geoscientific Model Development Discussions*, 1-35. https://doi.org/10.5194/gmd-2021-430. [**zur Veröffentlichung angenommen**]
- Ji, Y., Gong, B., Langguth, M., GAN-based video prediction models for precipitation nowcasting **[eingereicht]**
- Rojas-Campos, A., Langguth, M., Wittenbrink, M. & Pipa, G. (2022). Deep learning model for generation of precipitation maps based on Numerical Weather Prediction. EGUsphere. https://doi.org/10.5194/egusphere-2022-648 **[in Begutachtung, Preprint verfügbar]**

Work packages 3 und 4:

- Rojas-Campos, A., Wittenbrink, M., Nieters, P., Schaffernicht, E., Keller, J. D. & Pipa, G. (2021). Post-processing of NWP precipitation forecasts using deep learning. *Weather and Forecast*. [**in Begutachtung**]

Work package 4:
- Wittenbrink, M., Keller, J. D. (2022). A two-dimensional analog ensemble approach for precipitation forecasting based on wavelet transforms [**geplant**]

Work package 5:
- Glowienka-Hense, R., Hense, A., Brune, S., and Baehr, J (2020): Comparing forecast systems with multiple correlation decomposition based on partial correlation, Adv. Stat. Clim. Meteorol. Oceanogr., 6, 103–113, https://doi.org/10.5194/ascmo-6-103-2020 [**veröffentlicht**]
- Glowienka-Hense, R., Hense A. (2022): Evaluating models sensitivities with partial multiple correlation decomposition, [**geplant**]

Work package 6:
- Mozaffari, A., Langguth, M., Gong, B., Ahring, J., Rojas-Campos, A., Nieters, P., Campos Escobar, O.J., Wittenbrink, M., Baumann, P., Schultz, M. (2022) ; HPC-oriented Canonical Workflows for Machine Learning Applications in Climate and Weather Prediction. *Data Intelligence*; https://doi.org/10.1162/dint_a_00131 [**veröffentlicht**]

6) References

Abadi, M., et al. (2015) *TensorFlow: Large-scale machine learning on heterogeneous systems*. https://www.tensorflow.org/

Anderson, T.W. 1984: An introduction to multivariate statistical analysis (second edition). Wiley series in probability and mathematical statistics.John Wiley and Sons, New York, 675pp

Bach, Liselotte; Schraff, Christoph; Keller, Jan D. and Hense, Andreas (2016): Towards a probabilistic regional reanalysis system for Europe: evaluation of precipitation from experiments - Tellus A, Vol. 68, No. 1, https://doi.org/10.3402/tellusa.v68.32209

Buschow, S, Friederichs, P. (2021) SAD: Verifying the scale, anisotropy and direction of precipitation forecasts. *QJR Meteorol Soc*.; https://doi.org/10.1002/qj.3964

Brune, S., S. Buschow and P. Friederichs (2021): The Local Wavelet-based Organization Index - Quantification, Localization and Classification of Convective Organization from Radar and Satellite Data. *Q. J. R. Meteorol. Soc*. **147**, 1853-1872, https://doi.org/10.1002/qj.3998

De Smedt K, Koureas D, Wittenburg P (2020) FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units. Publications 8: 21. https://doi.org/10.3390/publications8020021

Dmitrienko, V. D., Zakovorotnyi, A. Y., Leonov, S. Y., & Khavina, I. P. (2014). Neural Networks Art: Solving problems with multiple solutions and new teaching algorithm, The Open Neurology Journal, 8, 15

Dorninger, M., P. Friederichs, S. Wahl, M. P. Mittermaier, C. Marsigli, and B. G. Brown (2018): Editorial: Forecast verification methods across time and space scales – Part I. - Meteorologische Zeitschrift 27, 433 - 434, https://doi.org/10.1127/metz/2018/0955

Feng, Y., Zhou, M. & Tong, X. (2020). Imbalanced classification: a paradigm-based review. *arXiv*. https://doi.org/10.48550/arxiv.2002.04592

Friedman, J., Hastie, T. and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*, *28*, 337-407.

Gilleland, E., Ahijevych, D., Brown, B. G., Casati, B., & Ebert, E. E. (2009). "*Intercomparison of spatial forecast verification methods*". Weather and forecasting, 24(5), 1416-1430.

Grönquist, P. Yao, C., Ben-Nun, T., Dryden, N., Dueben, P., Li, S., Hoefler, T, (2021), "Deep learning for post-processing ensemble weather forecasts",Philosophical Transactions of the Royal Society, https://doi.org/10.1098/rsta.2020.0092

Harris, L., McRae, A. T., Chantry, M., Dueben, P. D., & Palmer, T. N. (2022). "*A Generative Deep Learning Approach to Stochastic Downscaling of Precipitation Forecasts*". *arXiv preprint arXiv:2204.02028*.

Hewson, T. D., Pillosu, F. M. (2021) *"A low-cost postprocessing technique improves weather forecasts around the world"* Communications Earth & Environment https://doi.org/10.1038/s43247-021-00185-9.

Hoffmann, S., and Lessig, C. (2021) Towards representation learning for atmospheric

dynamics, arXiv:2109.09076

Holzinger, A., Malle, B., Saranti, A., & Pfeifer, B. (2021), Towards multi-modal causability with Graph Neural Networks enabling information fusion for explainable AI. Information Fusion, 71, 28-37.

Jeong, C.-H. and Yi, M. Y. (2022) *Correcting rainfall forecasts of a numerical weather prediction model using generative adversarial networks*, The Journal of Supercomputing, https://doi.org/10.1007/s11227-022-04686-y .

Keisler, R. (2022). Forecasting Global Weather with Graph Neural Networks. *arXiv preprint arXiv:2202.07575*.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). "*Swin transformer: Hierarchical vision transformer using shifted windows*". Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 10012-10022).

Martinez-Villalobos, Cristian, and J. David Neelin. "Why do precipitation intensities tend to follow gamma distributions?." *Journal of the Atmospheric Sciences* 76.11 (2019): 3611-3631.

Mozaffari, A., Selke, N., Schultz M., (2022b) Advancing caching and automation with FDO (in press)

Price, I., and Rasp, S. (2022) *"Increasing the accuracy and resolution of precipitation forecasts using deep generative models*." International Conference on Artificial Intelligence and Statistics. PMLR, 2022. Available at https://doi.org/10.48550/arXiv.2203.12297.

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., ... & Mohamed, S. (2021). "*Skilful precipitation nowcasting using deep generative models of radar"*. Nature, 597(7878), 672-677.

Raissi, M., Yazdani, A. and Karniadakis, G. E. (2020). Hidden fluid mechanics: Learning velocity and pressure fields from flow visualizations. *Science*, *367* (6481), 1026-1030

Reich, S. and Cotter, C. (2015). Probabilistic forecasting and Bayesian data assimilation. Cambridge University Press. 296pp

Stein, J., and Stoop, F. (2019). Neighborhood-Based Contingency Tables Including Errors Compensation, *Monthly Weather Review;* https://doi.org/10.1175/MWR-D-17-0288.1

Soiland-Reyes, S., et al.: (2021) Packaging research artefacts with RO-Crate. arXiv preprint arXiv: 2108.06503

Nandwani, Y., Jindal, D., & Singla, P. (2020), Neural Learning of One-of-Many Solutions for Combinatorial Problems in Structured Output Spaces. arXiv preprint arXiv:2008.11990.

Vannitsem et al. (2021). Statistical Postprocessing for Weather Forecasts: Review, Challenges, and Avenues in a Big Data World, *Bulletin of the American Meteorological Society;* https://doi.org/10.1175/BAMS-D-19-0308.1

Sabrina Wahl (2015): Uncertainty in mesoscale numerical weather prediction: probabilistic forecasting of precipitation, 2015, Bonner Meteorologische Abhandlung, 108 S, https://hdl.handle.net/20.500.11811/6560

Shorten, C., & Khoshgoftaar, T. M. (2019). "*A survey on image data augmentation for deep learning.*" Journal of big data, 6(1), 1-48.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., Polosukhin, I. (2017) *Attention is all you need*, Proceedings of the 31st International Conference on Neural Information Processing Systems, Red Hook, NY, USA, ISBN 978-1-5108-6096-4.

Wainwright, M. J., and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, *1*(1–2), 1-305.

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.1

Williams, P. L. and Beer, R. D. (2010): Nonnegative Decomposition of Multivariate Information, arXiv [preprint], arXiv:1004.2515,14 April 2010

Zhang, B., Gu, S., Zhang, B., Bao, J., Chen, D., Wen, F., ... & Guo, B. (2022). "*Styleswin: Transformer-based gan for high-resolution image generation*". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11304-11314).

Zolina, O., Kapala, A., Simmer, C. and Gulev, S. K. (2004). Analysis of extreme precipitation over Europe from different reanalyses: a comparative assessment. Global and Planetary Change, 44(1-4), 129-161.

Bundesministerium
für Bildung
und Forschung

DeepRain

Abschlussbericht

Förderkennzeichen 01IS18047
Oktober 2018 – März 2022

UNIVERSITÄT OSNABRÜCK

JACOBS UNIVERSITY

UNIVERSITÄT BONN

JÜLICH Forschungszentrum

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

DWD

Das diesem Bericht zugrunde liegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IS18047A-E gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei der Autorin/beim Autor. Dieser Text wurde mit Hilfe von DeepL aus dem Englischen ins Deutsche übersetzt und anschließend manuell überarbeitet.

Autoren:

Martin G. Schultz, Forschungszentrum Jülich (PI, Editor)
Amirpasha Mozaffari, Forschungszentrum Jülich (Co-editor)
Michael Langguth, Forschungszentrum Jülich (Co-editor)

Peter Baumann, Jacobs University Bremen
Otoniel Campos, Jacobs University Bremen
Rita Glowienka-Hense, University of Bonn
Bing Gong, Forschungszentrum Jülich
Andreas Hense, University of Bonn
Yan Ji, Forschungszentrum Jülich
Jan Keller, Hans-Ertel-Zentrum/DWD
Gordon Pipa, University of Osnabrück
Rodolfo Adrián Rojas-Campos, University of Osnabrück
Martin Wittenbrink, Hans-Ertel-Zentrum/DWD

Jülich, September 2022

Zusammenfassung

Das DeepRain-Projekt zielte auf die Entwicklung neuer Ansätze für die Kombination aus modernen Methoden des maschinellen Lernens mit leistungsstarken IT-Systemen für die Datenverarbeitung und -verbreitung kombinieren, um verbesserte hochauflösende räumliche Karten des Niederschlags über Deutschland zu erstellen. Grundlage für dieses Projekt war das mehrjährige Archiv von Ensemble-Modellvorhersagen des numerischen Wettermodells COSMO des Deutschen Wetterdienstes (DWD). Sechs transdisziplinäre Forschungseinrichtungen arbeiteten in DeepRain zusammen, um eine durchgängige Verarbeitungskette zu entwickeln, die potenziell in der zukünftigen operationellen Wettervorhersage eingesetzt werden kann. Der Projektantrag hatte mehrere Herausforderungen identifiziert, die es in diesem Zusammenhang zu bewältigen galt. Neben den technischen Herausforderungen bei der Schaffung einer neuartigen Datenfusion von recht unterschiedlichen Datensätzen (numerische Modelldaten, Radardaten, Beobachtungen von Bodenstationen), dem Aufbau skalierbarer maschineller Lernlösungen und der Optimierung der Leistung der Datenverarbeitung und des maschinellen Lernens gab es verschiedene wissenschaftliche Herausforderungen im Zusammenhang 1. mit den kleinräumigen Strukturen von Niederschlagsereignissen, 2. Schwierigkeiten bei der Suche nach robusten Bewertungsmethoden für Niederschlagsvorhersagen und 3. den nicht-normalverteilten Niederschlagsstatistiken in Verbindung mit stark unausgewogenen Datensätzen.

Zum Projektstart von DeepRain war die Anwendung von maschinellem Lernen auf Wetter- und Klimadaten noch sehr neu und es gab kaum Veröffentlichungen oder Softwarecodes, auf denen man aufbauen konnte. DeepRain leistete somit Pionierarbeit bei der Anwendung moderner Deep-Learning-Modelle im Bereich der Wettervorhersage. Gleichzeitig konnte man in den letzten drei Jahren einen exponentiellen Anstieg der Zahl der Veröffentlichungen in diesem neuen Bereich beobachten. Sehr oft handelte es sich dabei um Studien, die in Nordamerika oder China durchgeführt wurden. Globale Unternehmen wie Google, Amazon, NVidia oder Microsoft haben inzwischen Gruppen von Wissenschaftlern und Ingenieuren gegründet, um die Forschung zu "Wetter-KI" voranzutreiben und marktfähige Wetter- und Klimaanwendungen mit Deep Learning zu entwickeln. Daher kam das DeepRain-Projekt zur rechten Zeit, da es eine Basis für maschinelles Lernen im Bereich Wetter und Klima in Deutschland geschaffen hat. DeepRain ermöglichte es dem Konsortium, das Potenzial von Deep Learning im Zusammenhang mit der erforderlichen gigantischen Datenverarbeitung zu erforschen und mit den internationalen Entwicklungen in diesem schnell wachsenden Forschungsbereich Schritt zu halten.

DeepRain konnte das geplante Ergebnis, d. h. den Bau eines Prototyps für einen durchgängigen Arbeitsablauf für hochauflösende Niederschlagsvorhersagen auf der Grundlage von Deep Learning, zwar nicht vollständig erzielen, aber es wurden alle damit verbundenen Forschungsfragen beantwortet und alle erforderlichen Bausteine für einen solchen Arbeitsablauf wurden entwickelt. Beispielsweise wurde die moderne Datenwürfel-Technologie erfolgreich eingesetzt, um vier- bis sechsdimensionale atmosphärische Simulationsdatenwürfel auf der Basis von DWD-Daten für die Extraktion und Analyse bereitzustellen.

Zusätzlich zu den oben beschriebenen erwarteten Herausforderungen traten während des Projekts die folgenden schwerwiegenden Probleme auf: 1. ein weitreichender Datenverlust aufgrund von Hardwareausfällen im Frühjahr 2021, 2. die Covid-19-Pandemie von März 2020 bis heute und 3. Schwierigkeiten, hochqualifiziertes Personal zu finden - insbesondere in Zeiten, in denen die meiste Arbeit im Home-Office erledigt werden musste.

Die wichtigsten Ergebnisse von DeepRain sind:

- Datentransfer im Petabyte-Bereich von archivierten COSMO-DE-EPS-Vorhersagen von Bandlaufwerken des DWD und des RADKLIM-Datensatzes vom OpenData-Server zum Dateisystem JUST am JSC/FZ Jülich, Organisation und Bereinigung dieser Daten und Gewährleistung des Datenzugangs für alle Projektpartner,
- Parallelisierte Verarbeitung von COSMO-EPS- und RADKLIM-Daten (Ensemblestatistik, Remapping für Datenfusion und für das Einfügen in Rasdaman),
- Implementierung von Rasdaman Datenwürfel Array Datenbankservern am FZ Jülich und Ingestion von mehreren TBytes an Wetterdaten,
- Aufnahme des Jülicher Rasdaman-Servers in den EarthServer-Datenwürfel-Verbund,
- Weiterentwicklung von Rasdaman zur Beschleunigung des Dateneinfügens und -abrufs, Definition neuer benutzerdefinierter Funktionen für die Analyse topographischer Daten, Definition eines neuen Koordinatenreferenzsystems für gedrehte Polkoordinaten und Vorbereitung der Anbindung von Prozessierungsketten für maschinelles Lernen,
- Entwicklung von statistischen Downscaling-Techniken und maschinellen Lernmodellen, um:
    - dichotomen und quantitativen Niederschlagsvorhersagen an Stationsstandorten zu generieren und
    - Gebietsvorhersagen in der Auflösung der RADKLIM-Radardaten zu erzeugen,
- Erforschung neuer Verifikationsstatistiken auf der Grundlage partieller Korrelationen und des Regression Boostings.

In diesem Bericht geben wir einen detaillierten Überblick über die Arbeit und das Erreichte im Rahmen des DeepRain-Projekts. Dieser Bericht ist in fünf Abschnitte gegliedert: In Abschnitt 1 stellen wir den Arbeitsplan aus dem Projektantrag vor und geben Informationen über den Stand der Erbringung jeder einzelnen Aufgabe, um einen kompakten Vergleich zwischen dem Projektplan und seinen Ergebnissen zu ermöglichen. In Abschnitt 2 werden dann die im Rahmen des Projekts durchgeführten Arbeiten für jedes einzelne Arbeitspaket detailliert beschrieben. In Abschnitt 3 werden die Projektergebnisse und deren mögliche künftige Nutzung erörtert. In Abschnitt 4 geben wir einen allgemeinen Überblick über die außerhalb des Projektes erfolgten Fortschritte in den Forschungsbereichen, die mit DeepRain in Verbindung stehen. Im Einzelnen sind dies: maschinelles Lernen für die Niederschlagsvorhersage, Methoden zur Bewertung von Niederschlagsvorhersagen, Umgang mit Big Data und FAIR-Datenpraktiken. Schließlich werden in Abschnitt 5 alle Zeitschriftenveröffentlichungen, Datensätze und Softwarepakete sowie geplante Einreichungen aufgeführt, die aus dem DeepRain-Projekt hervorgegangen sind. Abschnitt 6 beinhaltet das Literaturverzeichnis.

1) Vergleich der Projektplanung mit den geleisteten Arbeiten und erzielten Ergebnissen

Tabelle 1: Überblick über die geplanten Projektergebnisse und die erbrachten Leistungen

Auf-gabe	**Geplante Arbeiten/Ergebnisse**	**Geleistete Arbeiten/Erzielte Ergebnisse**	**Status**
1.1	Projektkoordination: • Überwachung des Projektfortschritts • Berichterstattung über den Fortschritt • Organisation von Sitzungen, einschließlich Auftakt-, Jahres- und Abschlusssitzungen	• Organisation des Jahresberichts • Organisation des Abschlussberichts • Organisation der jährlichen Treffen, der virtuellen Treffen und des Abschlusstreffens	Vollständig
1.2	Kommunikation: • Erstellung einer Projekt-Website • Veröffentlichung von Neuigkeiten • Öffentlichkeitsarbeit	• Die Website des Projekts ist unter https://www.deeprain-project.de verfügbar. • Es werden regelmäßig Neuigkeiten veröffentlicht • Öffentliche Kommunikation	Vollständig
1.3	Projektleitung: • endgültige Kooperationsvereinbarung • Überwachung der Ausgaben • Kommunikation und Risikomanagement	• Überwachung der Kosten • Kostenneutrale Verlängerung des Projekts	Vollständig
2.1	Bereitstellung von Modell-, Radar-, Blitz- und Stationsdaten des DWD mit Format-beschreibung und Spezifikationen • Ein Teil der DWD-Wetter-daten ist in Jülich verfügbar. • Beschreibung und Spezifikation des Datenformats.	• Registrierung von 15.24TB an interpolierten COSMO Daten für sechs relevante meteorologische Variablen in der rasdaman Instanz in Jülich (EnterpriseCube) • Dokumentation des Datenformats und Beschreibung der registrierten COSMO-Variablen auf der Service-Landingpage für die Coverages im EarthServer-Verbundknoten Jülich.	Vollständig
2.2	• Topographiedaten sind über den rasdaman Service Endpunkt über OGC WCS / WCPS / WMS Anfragen zugänglich. • Herunterladen und Verarbeiten von SRTM-Topographiedaten • Berücksichtigung der	• Datenimport-Skripte für SRTM-Topographiedaten • Erweiterung der Funktion project() in rasdaman zur Unterstützung benutzerdefinierter Interpolation. • Implementierung von Slope/Aspect/Hillshade-Funktionen als benutzerdefinierte Funktionen (UDFs) in rasdaman (Link).	Vollständig

	Interpolation und Berechnung relevanter Merkmale für die ML		
2.3	Einrichtung einer rasdaman-Array-Datenbankinstanz in JÜLICH, die Definition der Datenimporte aus den Aufgaben 2.1 und 2.2 und Performance-Optimierung: • COSMO-Daten verfügbar über rasdaman Service Endpoint via OGC WCS / WCPS / WMS Anfragen. • Topographische Funktionen in rasdaman verfügbar.	• Einrichtung der rasdaman-Instanz in Jülich mit Einbindung in die EarthServer-Föderation. • Ingestion-Skripte für COSMO-Daten eingerichtet (Link) • Unterstützung für rotierte Gitter CRS (Link) • Erweiterung der Open-Source-Geobibliotheken PROJ/GDAL um die Definition von gedrehten CRS • Entwicklung einer standardisierten rotierten CRS-Definition (Link) • Verbesserte rasdaman Leistung: ○ Schnellerer GRIB-Datenimport ○ Optimierte Case-Anweisung ○ Erstellung von Benchmarks für die Datenaufnahme unter Verwendung verschiedener Parameter im rasdaman-Aufnahmecode.	Vollständig
2.4	Aufbau einer relationalen Datenbank zur Bereitstellung von Stationsdaten für die Auswertung	• Registrierung von 2,95 TB an Niederschlagsdaten in der rasdaman-Instanz für die Jahre 2015-2018.	Vollständig
2.5	Entwicklung neuer Datenbankabfrageoperatoren für Radar- und Blitzdaten und Optimierung der Schnittstelle zum neuronalen Netz	• Statt einer direkten Kopplung der rasdaman-Datenbank mit den neuronalen Netzen, wurden Python-basierte Abfragemethoden entwickelt und von den Projektpartnern gemeinsam genutzt; dadurch konnte die Entwicklung der ML Verfahren beschleunigt werden.	Teilweise
2.6	Implementierung des Datenflusses für die Ausgabedaten des neuronalen Netzes, einschließlich der Einbettung in das JOIN-Webinterface	• Aufgrund des Verzichts auf eine explizite Kopplung der rasdaman-Datenbank mit den entwickelten neuronalen Netzen (siehe 2.5) wurde diese Aufgabe nicht durchgeführt	Nicht erledigt
3.1	Downscaling auf Stationsstandorte mit Deep Learning	• Erste experimentelle Deep-Learning-Algorithmen für das Downscaling des von einer Regenstation registrierten Niederschlags	Vollständig
3.2	Implementierung der ersten Version des neuronalen Netzes in Jülich	• Trainierte Deep-Learning-Modelle in Jülich gespeichert	Vollständig
3.3	Downscaling der Regenvorhersage unter Berücksichtigung der räumlichen Komponente	• Paper "Deconvolutional and generative models for generation of precipitation maps based on Numerical Weather Prediction"	Vollständig

		eingereicht bei Journal Geoscientific Model Development • Öffentliches Repository mit dem Code für die Deep-Learning-Modelle auf github	
3.4	Downscaling der Regenvorhersage unter Berücksichtigung räumlicher und zeitlicher Komponenten	• Paper "Post-processing of NWP precipitation forecasts using deep learning" eingereicht bei Journal Weather and Forecasting • Öffentliches Repository mit dem Code für die Modelle auf github	Vollständig
3.5	Downscaling von Sturmvorhersagen unter Berücksichtigung räumlicher und zeitlicher Komponenten	• Unmöglich zu erreichen aufgrund der sehr geringen Anzahl von Stichproben zu extremen Ereignissen in den Trainingsdaten.	Nicht erledigt
4.1	Datenbank für Versionsänderungen des numerischen Wettermodells COSMO	• Dokumentiert auf der COSMO-Modell-Website	Vollständig
4.2	Implementierung klassischer Downscaling-Methoden	• Logistische Regression, verallgemeinertes lineares Modell und analoge Ensemble-Methoden wurden auf der Grundlage des DeepRain-Datensatzes als klassische Downscaling-Methoden in Python implementiert.	Vollständig
4.3	Dokumentation der Methode zur Konsistenzprüfung	• siehe 5.1 / 5.3	verlagert auf 5.1 , 5.3
4.4	Datenbank mit konsistenten Vorhersagevariablen und mit Variablen, die aufgrund von Versionsänderungen erhebliche Änderungen erfahren.	• siehe 5.1 / 5.3	verlagert auf 5.1 , 5.3
4.5	Nachbearbeitete Vorhersagen für die Bewertung	• Die klassischen Downscaling-Methoden wurden auf die Daten angewandt. • Die erzeugten Datensätze wurden als Referenz für den Vergleich mit den DL-Ansätzen bereitgestellt • Paper "Post-processing of NWP precipitation forecasts using deep learning" eingereicht bei Journal Weather and Forecasting	Vollständig
5.1	Dokumentation der Bootstrap-Verfahren und Auswahl der probabilistischen Bewertungsergebnisse	• Dissertation: Wahl (2015) mit open access • Die Methoden sind Teil des R-Pakets "verification" von Gilleland (2014): R-Package "verification": Weather forecast verification utilities. NCAR .-Research Applications Laboratory, Version 1.41.	Vollständig

		• Weitere Details sind in Dorninger et al. (2018) zu finden. • Präsentation "Evaluation of model simulations", 10. März 2021, die über B2SHARE veröffentlicht wird. • Präsentation auf der Projektsitzung im April 2022 über Maßnahmen auf der Grundlage der "freien Energie" und der Laplace-Approximation der posterioren Wahrscheinlichkeitsdichten, die über B2SHARE veröffentlicht wird	
5.2	Erste Version der Bewertungs-Toolbox verfügbar	• Glowienka-Hense et al. (2020) • Über B2SHARE veröffentlichte Präsentationen	Vollständig
5.3	Prototyp für die grafische Ausgabe von Bewertungsergebnissen	• Präsentationen während Web-Meetings zwischen März 2020 und April 2022, Entropie-basierte Maßnahmen, MSE/MAE-basierte Maßnahmen, • Präsentationen, die im März und September 2022 gehalten wurden, werden über B2SHARE veröffentlicht	Vollständig
5.4	Documentation of procedures for optimal input variable selection, information criteria, extreme values, with results.	• Nicht möglich, da nur teilweise Datensätze verfügbar, Extremereignisse konnten mangels ausreichender Stichproben nicht ausgewertet werden	Nicht erledigt
6.1	Erste Workflow- und Datenflussanalyse mit dem Jülicher HPC-System	• Entwicklung von PyStager als skalierbare Workflow-Lösung für die parallelisierte Verarbeitung großer Datenmengen auf HPC-Systemen • Vorverarbeitung von COSMO-EPS- und RADKLIM-Daten mit PyStager, • Code auf github veröffentlicht • Vorverarbeitung von COSMO-EPS für die Einspeisung in rasdaman (vgl. Aufgabe 2.1), Code auf github veröffentlicht • Entwicklung von Jupyter-Notebooks zur Demonstration der benutzerfreundlichen HPC-Datenverarbeitung und -Visualisierung	Vollständig
6.2	Endgültige Version der projektinternen Datenfluss- und Workflow-Analyse für das HPC-System	• Erstellung von Schaubildern der Workflow-Architektur, die auf dem Projekttreffen im November 2020 vorgestellt und über B2SHARE veröffentlicht werden • Beitrag zu Codes für maschinelles Lernen, um den Umgang mit großen Mengen an Wetterdaten zu erleichtern	Teilweise

		• Der direkte Zugang zu Daten für ML-Anwendungen von rasdaman konnte während des Projekts aufgrund technischer Probleme und Zugangsbeschränkungen nicht implementiert werden.	
6.3	Veröffentlichung der DeepRain-Systemarchitektur und des Arbeitsablaufs	• Konzeption des DeepRain-Workflows für FAIR und Reproduzierbarkeit, beschrieben in der Zeitschrift Data Intelligence (siehe Abschnitt 5) • Eine vollständige Umsetzung des Konzepts war im Rahmen des Projekts nicht möglich	Teilweise
6.4	Workflow-Analyse und Systemdesign für eine mögliche Operationalisierung des DeepRain-Prozesses.	• Komponenten des möglichen Betriebssystems sind erstellt worden: • Vorverarbeitungs-Prozessierungsketten mit HPC • Abfrage des Datenstroms aus dem föderierten rasdaman-Datenwürfel über Webbrowser und API-Abfrage • Jupyter Notebook zur Demonstration der Interaktion mit rasdaman und der Visualisierung • Viewgraph des operativen Workflows wurde in internen Meetings vorgestellt und über B2SHARE veröffentlicht	Teilweise

2) Detaillierte Beschreibung der im Projekt geleisteten Arbeiten und erzielten Ergebnisse

Das zentrale Ziel des DeepRain-Projekts war die Entwicklung fortschrittlicher Methoden des Maschinellen Lernens (ML) und speziell des Deep-Learnings (DL) für verbesserte, lokale Niederschlagsvorhersagen auf der Grundlage von Simulationen numerischer Wettervorhersagemodelle. Vor dem Hintergrund einer zukünftigen operationellen Anwendung dieser Methoden wurden Simulationsdaten aus der operationellen Wettervorhersage-modellkette des Deutschen Wetterdiensts (DWD) herangezogen. Die aktuelle Modellkette des DWD baut auf dem ICON-Modell auf (eine kompakte Beschreibung ist hier verfügbar). Während die globale Modellkonfiguration von ICON bereits im Januar 2015 in Betrieb ging, wurde die regionale Modellkonfiguration COSMO-D2 für konvektions-erlaubende Vorhersagen auf der Kilometer-Skala erst im Februar 2021 durch den Nachfolger ICON-D2 ersetzt. Da große Datensätze eine notwendige Voraussetzung für ML-Anwendungen auf Niederschlagsvorhersagen darstellen, beschloss das Konsortium, mit den (historischen) Vorhersagedaten des Vorgängermodells COSMO zu arbeiten, wie es auch ursprünglich im Antrag vorgesehen war. Konkret wurde das hochauflösende COSMO-Ensemble Vorhersagesystem (COSMO-EPS) als primäre Quelle von Eingangsdaten für die ML-Modelle in DeepRain gewählt.

Aufgrund der großen Menge an verfügbaren COSMO-EPS Daten wurden effiziente Datenbereitstellungs- und –verarbeitungssysteme benötigt, die die effiziente Entwicklung der ML-Anwendungen sowie deren potenzielle, zukünftige Operationalisierung ermöglichen. Während des Projekts wurden zwei Hauptarbeitslinien verfolgt: 1) Parallelisierung der Datenvorverarbeitung auf HPC-Systemen und 2) Einrichtung und Betrieb des Big-Data-Analyse Servers Rasdaman. Die erste Aufgabe führte zur Entwicklung des Python-Softwarepakets PyStager, das nun in mehreren ML-Anwendungen am Jülich Supercomputing Centre eingesetzt wird.

Neben der Entwicklung von Skripten zur Vorverarbeitung und Einspeisung der Daten in Rasdaman, musste die Funktionalität des Big-Data-Analyse Servers erweitert werden, um das Koordinatenreferenzsystem (engl.: Coordinate Reference System, CRS) der COSMO-EPS-Daten mit gedrehten Polkoordinaten zu unterstützen. Letzteres erforderte umfangreiche Diskussionen mit den Hauptentwicklern der weit verbreiteten Geoinformationssoftwarepakete GDAL und PROJ und führte zu einem Untervertrag mit Spatialsys und Geomatys zur Implementierung der CRS für rotierte Polkoordinaten. Mit den Aufgaben der Datenverarbeitung verknüpft war die Entwicklung von Big-Data-Workflows zum Aufbau einer durchgängigen Verarbeitungskette, wie sie für eine operationelle Anwendung benötigt wird. Es wurden verschiedene Workflow-Konzepte entworfen und Demonstrationselemente in Form von Jupyter-Notebooks bereitgestellt, die einen flexiblen, benutzerfreundlichen Zugang der erforderlichen Funktionen bieten und darüber hinaus leicht angepasst werden können, um benutzerdefinierte Anwendungen auf Basis von automatisierten Batch-Systemen zu ermöglichen.

Zur Entwicklung von statistischen Downscaling-Verfahren für Niederschlagsvorhersagen mithilfe von ML-Methoden mussten anwendungsspezifische Herausforderungen bewältigt werden, die sich aus der ausgeprägten nicht-gaußschen Verteilung von Niederschlagsereignissen ergeben. Aus statistischer Sicht lässt sich Niederschlag mit einer positiv asymmetrischen Gamma-Verteilung beschreiben (siehe z. B. Zolina et al., 2004 und Martinez et al., 2019), aus der sich unmittelbar folgende Konsequenzen für die statistische Datenverarbeitung ergeben: 1) Die Wahrscheinlichkeitsdichte-Verteilung wird durch Ereignisse ohne Niederschlag dominiert, so dass sich ein stark unausgewogener Trainingsdatensatz ergibt. 2) Starke Niederschlagsereignisse treten nur sehr selten auf, obwohl sie aufgrund ihres großen Schadenpotentials besonders relevant sind. Die vom DWD zur Verfügung gestellte Niederschlagsstatistik zeigte, dass die meisten Orte in Deutschland in einem 20-jährigen Beobachtungszeitraum höchstens ein oder zwei extreme Niederschlagsereignisse erlebten. Die besonderen statistischen Eigenschaften von Niederschlag, insbesondere die starke Unterrepräsentierung von starken Ereignissen, erforderten die Untersuchung von Techniken zur Datentransformation und von angepassten Kostenfunktionen, die zur Optimierung von neuronalen Netzwerken benötigt werden. Während die Datentransformation darauf abzielt, die Dominanz von Ereignissen ohne Niederschlag zu reduzieren, ergibt sich der letzte aufgeführte Punkt aus der Tatsache, dass standardmäßig angewendete Verlustfunktionen (z.B. in Termen des mittleren quadratischen Fehlers) implizit davon ausgehen, dass die Daten mehr oder weniger normal verteilt sind (siehe Diskussion in Schultz et al., 2021).

Ein dritter wichtiger Aspekt von DeepRain war die Quantifizierung von Unsicherheiten und die Entwicklung geeigneter Bewertungsmetriken zur Beurteilung der Qualität ML-basierter Niederschlagsprognosen. Um aussagekräftige Vergleichsmodelle zu den ML-basierten Methoden zu etablieren, wurden klassische Downscaling-Methoden (logistische Regression, generalisierte lineare Modelle und Analog-Ensembles) auf dem Jülicher HPC-System implementiert und weiterentwickelt. Darüber hinaus wurden neue statistische Verfahren basierend auf partieller Korrelations- und Entropieanalyse entwickelt, um die Robustheit und Qualität der ML-Prognosen zu untersuchen und um neuartige Verifikationstechniken (z.B. Dorninger et al., 2018) in die Prozessketten für die Evaluierung aufzunehmen.

Im Folgenden geben wir einen Überblick über die wichtigsten Ergebnisse der sechs Arbeitspakete des DeepRain-Projekts.

Arbeitspaket 1: Koordination und Projekt-Management (FZ Jülich)

Die kontinuierliche Beaufsichtigung und Berichterstattung über die Fortschritte wurde mit Hilfe regelmäßiger Treffen mit allen DeepRain-Partnern während des gesamten Projektzeitraums sichergestellt (Tabelle 2). Zusätzlich zu den regelmäßigen Projekttreffen, die etwa alle sechs Monate stattfanden, wurden drei Sondertreffen organisiert (siehe Tabelle 2). Die Projekttreffen im Frühjahr wurden etwa einen Monat vor Ablauf der Berichtsfristen angesetzt, um die Erfassung der Berichtsinhalte zu erleichtern.
Da aufgrund der CoVid-19-Pandemie persönliche (Projekt-) Treffen nach März 2020 nicht mehr möglich waren, begann das Konsortium Online-Sitzungen zu organisieren. Um die eingeschränkten, direkten Kommunikationsmöglichkeiten zu kompensieren, wurden die Sitzungsintervalle verkürzt. Während die halbjährlichen Projekttreffen in drei Online-Sitzungen zu je 3 Stunden umorganisiert wurden, fanden zusätzlich monatliche Treffen statt, um offene Fragen zu besprechen und den gemeinsamen Entwicklungsfortschritt sicherzustellen. Im September 2021 war schließlich wieder die Durchführung eines persönlichen Projekttreffens möglich, bei dem die Strategie und die Pläne für die letzte Projektphase besprochen wurden.

Alle Partner stellten regelmäßigen Finanz- und Fortschrittsberichte bereit. Da einige Partner zu Beginn Schwierigkeiten hatten, geeignetes Personal zu finden, mussten Umverteilungen beim Budget und bei den Projektleistungen vorgenommen werden. Im Februar 2021 führte ein größerer Sicherheitsvorfall an Supercomputing-Zentren in ganz Europa zu erheblichen Verzögerungen bei mehreren DeepRain-Aktivitäten. In Anbetracht der anfänglichen Verzögerungen bei der Personalsuche und der Einschränkungen bei der transdisziplinären Zusammenarbeit aufgrund der Vorschriften des Innenministeriums während der CoVid-19-Pandemie veranlasste das Konsortium, eine kostenneutrale Projektverlängerung vom 30. September 2021 bis zum 31. März 2022 zu beantragen. Dieser Antrag wurde im Mai 2021 genehmigt.

Für die externe Kommunikation und Verbreitung wurde eine Projektwebsite (https://www.deeprain-project.de) über einen Dienstleistungsvertrag mit einer in Aachen ansässigen Web-Entwicklungsfirma eingerichtet. Die Website wurde regelmäßig aktualisiert und enthält mehrere Nachrichten über die Fortschritte und Errungenschaften von DeepRain sowie eine ständig wachsende Publikationsliste. Postdoktoranden und Doktoranden des Projekts hielten verschiedene Konferenzvorträge auf nationalen und europäischen Tagungen. Nähere Einzelheiten sind auf der DeepRain-Website dokumentiert. Wie im Projektantrag beschrieben, wurde im Februar 2021 ein Treffen mit dem DWD organisiert, um die allgemeine DeepRain-Strategie und die Ergebnisse zu erläutern und eine mögliche Operationalisierung der Methoden im DeepRain-Projekt zu diskutieren (SM03, siehe Tabelle 3). Auch wenn zu diesem Zeitpunkt bereits feststand, dass die Erschaffung eines Prototyps für das präoperationelle System von ML-basierten Niederschlagsvorhersagen im Projekt nicht möglich sein würde, wurden die Gespräche mit dem DWD als sehr informativ angesehen. Die Ergebnisse des DeepRain-Projekts beeinflussten zudem wesentlich die Initiative einer DWD-Seminarreihe zum Thema maschinelles Lernen, die nun alle sechs Monate organisiert wird.

Tabelle 2: Liste der DeepRain Projekttreffen.

Meeting nummer	Datum	Ort	Meeting nummer	Datum	Ort
PM01	Nov 2018	Jülich	WM06	Sep 2020	online
SM01(*)	Jan 2019	Bremen	PM05/WM07	Nov 2020	online
PM02	Mar 2019	Osnabrück	WM08	Dec 2020	online
SM02(**)	Jun 2019	Dortmund	WM09	Feb 2021	online
PM03	Nov 2019	Bonn	SM03(***)	Feb 2021	online
WM01	Jun 2019	online	PM06/WM10	Apr 2021	online
WM02	Aug 2019	online	WM11	Jun 2021	online
PM04/WM03	Mar 2020	online	WM12	Jul 2021	online
WM04	Jul 2020	online	PM07	Sep 2021	Köln
WM05	Aug 2020	online	PM08	Apr 2022	Osnabrück

PM: project meeting, WM: web meeting, SM: special meeting
(*) Rasdaman Trainingskurs, (**) Statusseminar (***) Informationsaustausch mit dem DWD

Arbeitspaket 2: Datenprozessierung und –bereitstellung (Jacobs Universität Bremen und FZ Jülich)

Sowohl die Erdsystemforschung als auch Deep Learning stellen Forschungsbereiche dar, die von Natur aus mit großen Datenmengen verbunden sind. Während Beobachtungs- und Modellierungsdaten des Erdsystems typischerweise Datensätze in der Größenordnung von mehreren Tera- oder Petabytes umfassen, hängt die Leistung neuronaler Netze für bestimmte Anwendungen stark von der Menge der verfügbaren (Trainings-) Daten ab. Daher stellten Datenmanagement, -aufbereitung und -bereitstellung zentrale Herausforderungen im DeepRain-Projekt dar, bei dem umfangreiche Modell- und Beobachtungsdatensätze verwendet wurden, um lokale Niederschlagsvorhersagen mit Hilfe von DL-Methoden zu verbessern.

Wie bereits erwähnt, stellen die Vorhersagen des COSMO-Ensemble Vorhersagesystems (COSMO-EPS) sowie die adjustierten Radarbeobachtungen des RADKLIM-Datensatzes die Hauptdatenquellen im DeepRan-Projekt dar. COSMO-DE EPS bezeichnet das frühere Ensemble-Vorhersagesystem, das von Mai 2011 bis Mai 2018 beim DWD im operationellen Betrieb war. Das Modell COSMO-DE ist ein konvektionserlaubendes, regionales numerisches Wettervorhersagemodell (NWP), das ein rotiertes Polgitter mit einem horizontalen Abstand von 2,8 km und 50 vertikalen Schichten verwendet. Das Modellgebiet deckt ganz Deutschland sowie Teile der Nachbarländer ab. COSMO-DE EPS erstellt dann insgesamt 20 modelbasierte Szenarien, mit denen achtmal am Tag probabilistische 27-Stunden-Vorhersagen erstellt werden. Die unterschiedlichen Modellrealisierungen (die Ensemble-Mitglieder) basieren dabei auf variierenden Rand- und Initialbedingungen sowie auf physikalisch plausiblen Permutationen in den parametrisierten atmosphärischen Prozessen. Der schnelle Aktualisierungszyklus (alle drei Stunden) ermöglicht eine rasche Assimilation aktueller Beobachtungen wie Radardaten, die für kurzfristige Vorhersagen von konvektivem Niederschlag sehr wertvoll sind.

Im Mai 2018 wurde COSMO-DE durch seinen Nachfolger COSMO-D2 ersetzt, der ein feineres Gitter von 2,2 km, mehr vertikale Schichten und einen größeren Modellbereich verwendet. COSMO-D2 EPS war dann Teil der operationellen Modell-Vorhersagekette beim DWD, bevor es im Februar 2021 durch ICON-D2 ersetzt wurde.

Die COSMO-Modelldaten werden durch den RADKLIM-Datensatz ergänzt, der eine radarbasierte Niederschlagsklimatologie auf der Grundlage des RADOLAN-Verfahrens bereitstellt. Das "RW"-Produkt des RADKLIM-Datensatzes stellt gerasterte Niederschlagsbeobachtungen auf stündlicher Basis bereit, wobei die Daten aus einer optimierten Eichung der Radarmessungen mithilfe von bodenbasierten Stationsbeobachtungen resultieren. Hierzu werden mehrere Korrekturen vorgenommen, um Störpixel zu entfernen, Abschattungseffekte durch Berge zu kompensieren und andere Artefakte aus den Rohdaten zu entfernen. Der RADKLIM-Datensatz umfasst auch das "YW"-Produkt, das Niederschlagsraten mit einer zeitlichen Auflösung von 5 Minuten liefert. Das YW-Produkt ergibt sich mittels Disaggregation der oben erwähnten RW-Daten und wird daher als quasi-adjustiertes Niederschlags-Beobachtungsprodukt angesehen. Die Daten sind auf einem 1 km-Raster unter Verwendung einer polaren stereographischen Projektion verfügbar und decken die Zeitspanne von 2001 bis 2021 ab.

Der Antrag sah die Entwicklung einer einheitlichen und umfassenden Prozesskette vor, bei der alle Daten in Rasdaman-Datenwürfel integriert werden, die dann ein anschließendes Daten-Streaming zum Training tiefer neuronaler Netze ermöglichen sollten. Aufgrund von Verzögerungen bei der Entwicklung der ML-Modelle wurde jedoch beschlossen, den Arbeitsablauf in mehrere Segmente aufzuteilen, die an der Universität Bremen und am JSC getrennt verfolgt werden konnten. Im ersten Teil wurde die Beschaffung der COSMO und RADKLIM Daten vom DWD organisiert. Der zweite Teil befasste sich mit dem Aufsetzen, dem Betrieb und der Verbesserung der Rasdaman-Server, und im dritten Teil wurden einige gemeinsame Arbeiten im Hinblick auf eine mögliche zukünftige Ankopplung von rasdaman an die ML Prozesskette durchgeführt. Eine detaillierte Beschreibung der drei Teile erfolgt in den folgenden Abschnitten.

a) Beschaffung der DWD Daten, Datenmanagement und -prozessierung

Um allen Projektpartnern den direkten Zugriff auf die oben genannten Datensätze auf dem Jülicher HPC-System zu ermöglichen, war ein Datentransfer von 650 Terabyte vom DWD zum Jülich Supercomputing Centre (JSC) nötig. Der Datentransfer wurde gemeinsam mit AP6 durchgeführt. Da ein Hardware-Störfall am JSC Anfang 2021 zu einer Beschädigung mehrerer Datensegmente führte, mussten die Daten zweimal übertragen werden. So wurden insgesamt etwa 1,3 Petabyte Daten vom DWD über das X-Win-Netz des DFN auf das Jülicher Dateisystem JUST übertragen. Der Nachweis, dass ein Datentransfer über das Internet in dieser Größenordnung möglich ist, stellt einen wichtigen Meilenstein des DeepRain-Projekts dar und hat Auswirkungen auf das Systemdesign zukünftiger Wetter- und Klimadienste, wie z.B. in DestinE.

Während bei der ersten Übertragung ein Großteil der Überwachung der Datenübertragung manuell erfolgen musste, konnten die Abläufe bei der zweiten Übertragung weitgehend automatisiert werden. Dies lag zum Teil an einer Hardwareaufrüstung beim DWD, die es ermöglichte, größere Datenmengen auf rotierenden Festplatten zwischenzuspeichern, aber auch an einem besseren Verständnis der Datenstrukturen und der Entwicklung funktionaler Verarbeitungsskripte. Durch diese Änderungen war es möglich, alle Daten innerhalb von 2 Wochen statt, wie bei der ersten Übertragung, innerhalb von 3 Monaten erneut zu übertragen. Der DWD hat dem Forschungszentrum Jülich die Erlaubnis erteilt, eine Kopie der Daten zu speichern und diese öffentlich zugänglich zu machen. Daher können nun alle COSMO-EPS-

Daten auf Anfrage über die am JSC gehostete MeteoCloud im Originalformat zur Verfügung gestellt werden, während eine Auswahl der Daten über die Rasdaman Earth Server-Föderation öffentlich zugänglich sind.

Die Speicherung der RADKLIM-Daten war ein vergleichsweise geringes Problem. Die Daten wurden vom OpenData-Server des DWD abgerufen und vom ursprünglichen binären Datenformat in das praktischere netCDF-Format konvertiert. Dies führte zu einer Gesamtdatenmenge von 4 Terabytes.

b) Datenfusion und Entwicklung hochperformanter Big Data Arbeitsabläufe

Bei den DeepRain-Daten handelt es sich um mehrdimensionale Strukturen, welche typischerweise als Datenwürfel modelliert werden. Es wurden Datenwürfel mit bis zu 6 Dimensionen erstellt und Verarbeitungsskripte entwickelt, um sie in das Datenbanksystem Rasdaman einzufügen. Die 6 Dimensionen umfassen neben den drei räumlichen Achsen (Längen- und Breitengrad sowie Vertikale) und der Modellinitialzeit noch eine weitere Zeitachse für die relative Vorhersagezeit und für die 20 Ensemble-Mitglieder des COSMO Ensemble-Vorhersagesystems.

Zwei Instanzen der Rasdaman-Datenwürfel wurden am JSC konfiguriert und eingesetzt. Neben den DeepRain Daten enthalten die Rasdaman-Instanzen auch andere Geodaten, wie z. B. digitale Höhenmodelle (DEMs), die dann in Abfragen verknüpft werden können. Die Instanz des Datenwürfels, der die COSMO-DE EPS-Niederschlagsvorhersagen umfasst, wurde in den EarthServer-Verbund integriert, womit eine über mehrere Datenzentren distributierte Fusion von Klima-, Fernerkundungs- und Copernicus-Daten ermöglicht wird (Abb. 1).

Die Einbindung der COSMO-Daten erfolgte initial nach einer Reprojektion der Daten vom rotierten Polgitter auf ein gewöhnliches, sphärisches Gitter. Zur Vermeidung von inhärenten Verlusten in der Datenqualität aufgrund von Interpolation und zur Unterstützung des nativen, rotierten Polgitters des COSMO-Modells in Rasdaman war jedoch die zusätzliche Entwicklung eines neuen Koordinatenreferenzsystems (CRS) notwendig. Hierfür musste die Rasdaman-Software aktualisiert werden, wobei einige Arbeitsschritte von Seiten der externen Software-Dienstleister Spatialys und Geomatys realisiert werden mussten, um die neue Gitterdefinition in die zugrundeliegenden open-source Software zu implementieren (GDAL 3.4.0, PROJ 8.2.0). Eine CRS-Definition des rotierten Polgitters wird in naher Zukunft ebenfalls über OGC veröffentlicht (siehe github).

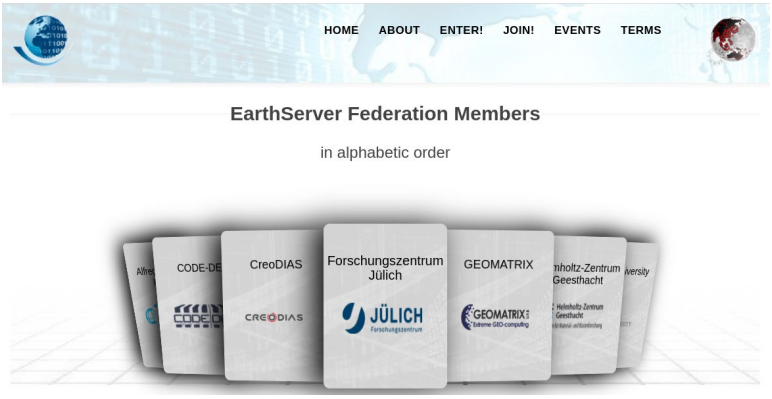


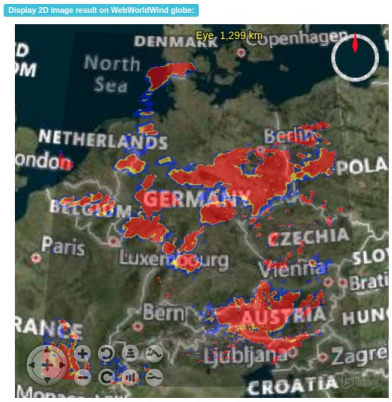
Abbildung 1: Beleg für die Mitgliedschaft des Forschungszentrums Jülich als neues Mitglied der EarthServer Federation. (Source: https://earthserver.eu/)

Um die Datenverarbeitung zu beschleunigen, wurde Rasdaman im Hinblick auf den Datenimport und -abruf optimier. Dies führte zu Beschleunigungen um den Faktor 15 bei der Datenaufnahme. Im Durchschnitt benötigt das Einlesen eines einjährigen Datensatzes mit fünf Variablen nun weniger als 24 Stunden auf dem Cloud-Server des JSCs. Die Leistung von Rasdaman wurde zusätzlich von der Freien Universität Bozen (gitlab) extern evaluiert.

Um nutzerspezifische Datenabfragen aus Rasdaman zu vereinfachen und den Benutzern bei der Visualisierung der COSMO-EPS-Daten zu helfen (Abb. 2), wurden ein Dashboard sowie ein Tutorial auf Basis eines Jupyter Notebooks erstellt. Im Zuge des DeepRain-Projekts wurden die Erfahrungen und Best Practices mit Datenwürfeln in dem entsprechenden git-Projekt dokumentiert, um eine effiziente Wiederverwendung der Entwicklungen im Projekts zu ermöglichen.

Eine weitere Rasdaman-Entwicklung im Rahmen des DeepRain-Projekts war die Definition von benutzerdefinierten Funktionen (UDFs) zur Ableitung übergeordneter aggregierter Variablen aus hochauflösenden digitalen Höhenmodellen, wie z. B. der Hangneigung und Gelände-Rauigkeit. Diese Funktionen sind im DeepRain-Projekt auf github verfügbar.

Die Ergebnisse von Arbeitspaket 2 wurden in mehreren Zeitschriftenartikeln und Konferenzbeiträgen dokumentiert und publiziert (siehe Abschnitt 5).




```
wwd(44.9875,2.956405722,56.2625,17.735782202)>>for $c in ( COSMO_DE_EPS__deeprain__precipitation_data )
let $dom := $c[ansi("2016-06-01"),ensemble(2), forecast hour(10),
                Lat(44.9875:56.2625),Long(2.956405722:17.735782202)]
return
 encode(
  switch
   case $dom=0
    return {red: 0; green: 0; blue: 0; alpha: 60}
   case 0.01 > $dom
    return {red: 0; green: 0; blue: 255; alpha : 127}
   case 0.03 > $dom
    return {red: 255; green: 255; blue: 0; alpha : 127}
   case 0.05 > $dom
    return {red: 255; green: 140; blue: 0; alpha : 127}
   default
    return {red: 255; green: 0; blue: 0; alpha : 127},
"image/png")
```

Abbildung 2: Beispielhafte Visualisierung einer COSMO-DE EPS Niederschlagsvorhersage für den 01. Januar 10 UTC (Ensemble-Mitglied 2, Modell-Initialzeit 01.Januar 2016 00 UTC) mithilfe von WCPS und NASA World Wind der Rasdaman-Instanz in Jülich.

c) Konzeptionelle Entwicklung eines Deep Learning Arbeitsablaufs

Da mit DeepRain ein Prototyp für eine zukünftige, vollständige Prozesskette für Wettervorhersagen auf der Grundlage von Deep Learning-Technologien entwickelt werden sollte, wurde ein Konzept für die Nutzung großer Datenwürfel als benutzerfreundliche Speicher-Systeme für Terabyte-große Datensätze, inklusive On-Demand-Verarbeitungsfunktionen, entwickelt. Basierend auf der rasql-Schnittstelle zu Rasdaman wurde ein HPC-orientierter Arbeitsablauf für die Extraktion von Klima- und Wetterdaten als Eingabedaten für ML-Modelle entwickelt. Dieser wurde als Jupyter-Notebooks implementiert, die auf dem HPC-fähigen Jupyter-JSC System ausführbar sind. Der Prozessablauf ermöglicht die Zusammenführung von RADKLIM-Radardaten mit den COSMO-EPS-Daten und beinhaltet ein automatisiertes Re-Mapping während der Extraktion.

Arbeitspaket 3: Methodenentwicklung für das Maschinelle Lernen (U Osnabrück, FZ Jülich, DWD)

Eines der Hauptziele des DeepRain-Projekts war die Entwicklung von ML-Anwendungen für typische meteorologische Problemstellungen. Die im Rahmen des Arbeitspakets 3 erarbeiteten Lösungen des maschinellen Lernens konzentrierten sich auf zwei hochrelevante meteorologische Probleme im Zusammenhang mit der Niederschlagsvorhersage.

Das erste Problem war die Nachbearbeitung von COSMO-DE EPS-Vorhersagen, um verbesserte Niederschlagsprognosen an bestimmten Orten, sogenannte Punktvorhersagen für Stationen mit Regenmessern, zu erhalten. Es wurden neue Ansätze für zwei Vorhersageprodukte in diesem Zusammenhang entwickelt: 1) dichotome Regen/Trocken-Vorhersagen und 2) quantitative Regenmengen-Vorhersagen.

Nach intensiver Optimierung der Hyperparameter ermöglichten die künstlichen neuronalen Netze (artificial neural network = ANNs) eine merkliche Verbesserung der Prognosequalität für beide Vorhersageprodukte im Vergleich zu klassischen statistischen Nachbereitungsmethoden. Abbildung 3 zeigt einen quantitativen Vergleich zwischen ANNs und klassischer statistischer Nachbearbeitung (logistische und verallgemeinerte lineare Regression) für vier ausgewählte Stationen im norddeutschen Raum. Die neuronalen Netzwerke zeigen in den meisten Fällen eine deutliche Verbesserung gegenüber den COSMO-DE EPS Rohdaten (positive Skill Score-Werte). Die Ergebnisse zeigen außerdem, dass das Training der ANN mit Daten von allen Stationen die Generalisierungsfähigkeiten der neuronalen Netze erheblich verbessern kann (zweite Spalte in Abbildung 3).

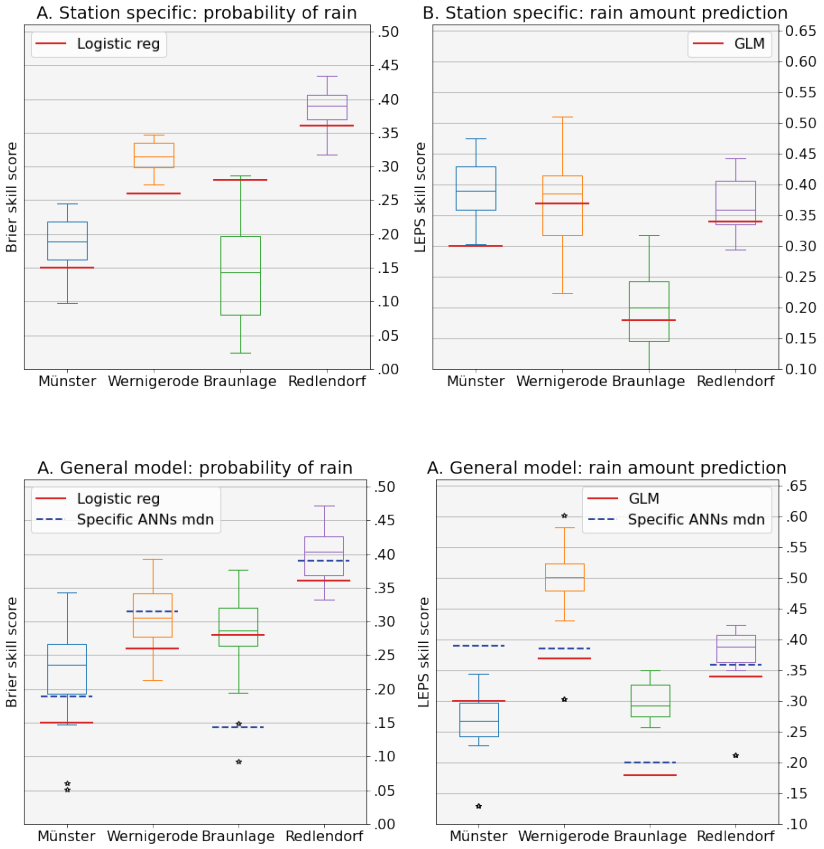
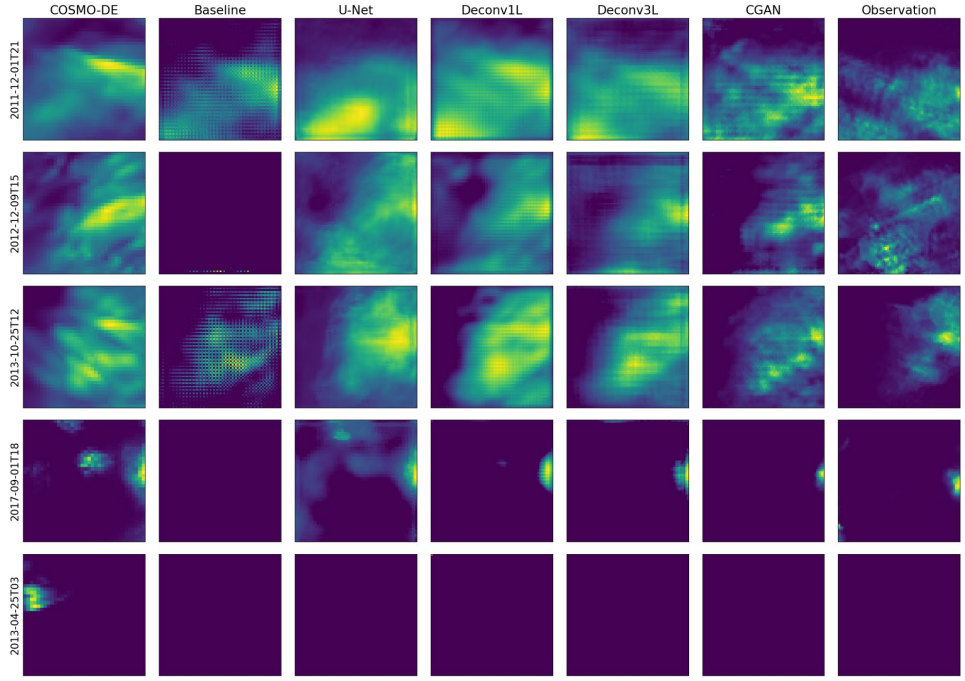


Abbildung 3: Evaluierung der Verbesserung der Vorhersage für Niederschlagsereignisse als Brier Skill Score (erste Zeile) und des linearen Fehlers im Wahrscheinlichkeitsraum (Linear Error in Probability Space, LEPS) der Niederschlagsmengen-Vorhersagen (zweite Zeile) für vier ausgewählte Stationen im norddeutschen Raum. Die Vorhersagequalität der logistischen Regression sowie eines generalisierten linearen Modells (generalized linear model, GLM) sind jeweils gegen ein trainiertes neuronales Netz aufgetragen. Die Modelle in der ersten Zeile wurden einzeln für jede Station trainiert, wohingegen die Modelle in der zweiten Zeile für Anwendungen auf alle untersuchten Stationen generalisiert wurden.

Als zweites Vorhersageproblem wurde im Rahmen von AP3 das so genannte räumliche Downscaling von Niederschlagsfeldern aus COSMO-DE EPS Vorhersagen bearbeitet. Hierzu wurden neuronale Netzwerke entwickelt, die mithilfe von RADKLIM-Beobachtungen nicht nur die räumliche Auflösung der prognostizierten Niederschlagsfelder erhöhen, sondern auch eine Fehlerkorrektur auf Basis historischer Vorhersagen realisieren. Hierzu wurden vier verschiedene neuronale Netzwerkarchitekturen mit TensorFlow (Abadi et al., 2015) trainiert und getestet. Die hierfür notwendigen Operatoren und Vorprozessierungsschritte zum Trainieren der neuronalen Netzwerke auf den HPC-Systemen am FZ Jülich wurden unter Zusammenarbeit mit AP 6 entwickelt.

Es wurden verschiedene neuronale Netzwerke auf Basis von Faltungsoperationen, so genannte Convolutional Neural Networks (CNNs) getestet (Abbildung 4). Die besten Ergebnisse wurden mithilfe einer Kopplung mit einem konditionalen generativen Netzwerk, einem so genannten Conditional Generative Adversarial Network (CGAN), erzielt (Rojas-Campos et al., 2022). Die jeweiligen Ansätze ermöglichen eine nichtlineare Kombination von COSMO-EPS-Variablen unter Nutzung der Ensemblestatistik. Auf diese Weise konnten verbesserte Niederschlagskarten mit einer Vorhersagezeit von 3 Stunden auf einem Gitter mit einer Maschenweite von rund 1,4 km erstellt werden, also einer linearen Verfeinerung um den Faktor 2 gegenüber dem originalen COSMO Modellgitter von etwa 2,8 km Auflösung. Einige ausgewählte Vorhersagen der getesteten neuronalen Netzwerke sind zusammen mit einer vereinfachten Visualisierung der verwendeten Netzwerkarchitekturen in Abbildung 4 dargestellt. Quantitative Auswertungen zeigen, dass das Basismodell (zweite Spalte im oberen Diagramm von Abbildung 4) mit der geringsten Anzahl an trainierbaren Parameter am schlechtesten abschneidet. Dies zeigt, dass eine gewisse Netzwerkkomplexität erforderlich ist, um Niederschlagsmuster korrekt zu beschreiben.

Der Informationsaustausch mit dem DWD im Februar 2021 zeigte ein allgemeines Interesse daran, die in DeepRain entwickelten Downscaling-Ansätze nach weiterer Evaluierung für die operationelle Nachbearbeitung zu übernehmen. Die transdisziplinäre Zusammenarbeit in DeepRain ermöglichte es daher, das Hauptziel zu erreichen, praktikable Lösungen des maschinellen Lernens für hochauflösende Niederschlagsvorhersagen zu entwickeln. Eine operationelle Umsetzung würde jedoch noch erhebliche Vorarbeiten erfordern, um die Methode u.a. für die Verwendung mit dem ICON-Modell anzupassen, welches das derzeitige operationelle Wettermodell des DWD darstellt.



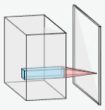
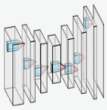
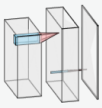
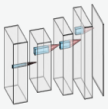
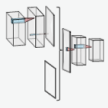
					
Name	Baseline	U-Net	Deconv1L	Deconv3L	CGAN
Layers	2dDeconv(1)	See Appendix A	2dDeconv(32) BatchNorm Conv(1)	Max pooling(2) 2dDeconv(32) BatchNorm 2dDeconv(16) Conv(1)	Deconv3L + discriminator
Kernel size	7x7		7x7	5x5	5x5
Stride	2		2	2	2
Parameters	7,008	153,849	224,673	127,841	127,841

Figure 4: Top: comparison of the COSMO-DE output (left column) with various deep learning models (middle columns) and the RADKLIM observations (right column). Each row represents an independent sample, i.e. a precipitation event. Bottom: Summary of the deep learning models developed and tested for radar downscaling

Arbeitspaket 4: Datenkonsistenz und klassische Downscaling-Algorithmen (DWD, U Bonn)

Im Rahmen dieses Arbeitspakets wurden zwei Hauptziele verfolgt. Das erste Ziel beruht auf der Tatsache, dass ML-Methoden stark von der Konsistenz der zugrundeliegenden Eingabe-Daten abhängen, um aussagekräftige Muster zu extrahieren, die bei der Nachkorrektur der Niederschlagsvorhersagen genutzt werden können. Die Konsistenz der Daten beeinflusst damit direkt die Möglichkeiten der Nachbereitung zur Prognosequalitäts-Verbesserung. In Anbetracht dessen wurde zu Beginn des Projekts eine umfassende Qualitätskontrolle der COSMO-DE EPS Daten durchgeführt, um Datenlücken und Inkonsistenzen zu identifizieren. Zusammen mit Arbeitspaket 5 (siehe unten) wurde eine Übersicht über fehlende Daten, vertauschte Zeitreihen oder unvollständige GRIB-Zeitstempel erstellt und die Datenbank entsprechend bereinigt.

Das zweite Ziel des Arbeitspakets 4 war die Bereitstellung von Benchmarks bzw. von Referenzergebnissen für die ML-Ansätze durch die Implementierung und Weiterentwicklung konventionellerer statistischer Methoden. Für punktbasierte Beobachtungen haben wir einen klassischen logistischen Regressionsansatz (LR) für dichotome Vorhersagen und ein verallgemeinertes lineares Modell (GLM) für Niederschlagsmengenvorhersagen implementiert und optimiert. Hierbei wurde zunächst die sogenannte LASSO-Technik eingesetzt, um relevante Eingabegrößen für die Korrektur von Niederschlagsvorhersagen mit beiden Methoden zu identifizieren. Als Eingangsparameter verwendeten wir eine 5x5-Gitterpunkte große Datenmaske aus den COSMO-EPS-Vorhersagen um den Stationsstandort, damit das GLM Informationen aus der lokalen Umgebung der Station extrahieren kann. Die punktbasierten Ansätze erwiesen sich im Vergleich zu den COSMO-DE-EPS-Vorhersagen als effizient, schnitten aber etwas schlechter ab als das oben beschriebene ANN-Netzwerk (vergleiche Abbildung 3). Ausführlichere Informationen über den Algorithmus und die Ergebnisse sind in der oben genannten Veröffentlichung von Rojas-Campos et al., 2022 enthalten, die derzeit bei der Fachzeitschrift Geophysical Model Development begutachtet wird.

Für die räumlichen Niederschlagsvorhersagen haben wir zunächst die Analog-Ensemble Technik (AnEn) zur Vorhersage von 2-dimensionalen Feldern verwendet. Das AnEn verwendet eine vordefinierte Metrik, um alle Datenpunkte im Trainingsdatensatz hinsichtlich ihrer Ähnlichkeit mit der aktuellen Situation zu bewerten. Dann werden für die Fälle mit der besten Übereinstimmung (d.h. die Fälle mit den ähnlichsten räumlichen Mustern in den Wetterdaten) die entsprechenden Beobachtungen (hier die Radarniederschlagsschätzungen aus dem RADKLIM-Datensatz) als Mitglieder in das Analog-Ensemble aufgenommen. Um eine größere Variabilität in den räumlichen Ausgabefeldern zu ermöglichen, wurde ein so genannter Kachel-Ansatz verwendet, bei das gesamte Datenfeld in kleinere Unterfelder aufgeteilt wird. Während das AnEn in der Lage ist, räumlich höher aufgelöste Korrekturen als andere klassische Methoden zu erstellen, erwies sich der verfügbare Trainingsdatensatz als zu klein für zuverlässige Niederschlagsvorhersagen. Trotz erheblicher Anstrengungen, das AnEn zu optimieren (verschiedene Einstellungen für die Größe der Kacheln, die Metrik usw.), und trotz der Verwendung von Wavelets (siehe unten) lieferte das AnEn im Durchschnitt ähnliche oder nur geringfügig bessere Ergebnisse als die Rohdaten des COSMO-DE EPS und schnitt damit schlechter ab als das oben beschriebene CGAN-Modell.

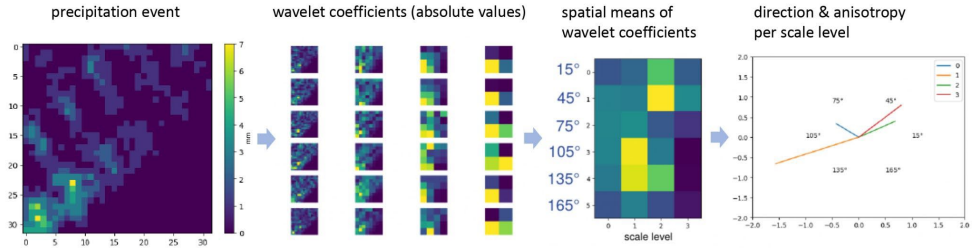


Abbildung 5: Exemplarische Darstellung der Ableitung der räumlichen Strukturen mithilfe der Wavelet-Transform Methode für den Analog-Ensemble Ansatz.

Um einen Referenzdatensatz für ein räumliches DL-Downscaling bereitzustellen, wurde dann ein punktbasierter GLM-Ansatz verwendet, der dem oben für Zeitreihen beschriebenen Konzept ähnelt. Der oben skizzierte Ansatz wurde erweitert, indem Wavelet-Koeffizienten aus den COSMO-DE-EPS-Daten als Prädiktoren im GLM abgeleitet wurden, um räumliche Korrelationen in den Daten zu berücksichtigen. Die Wavelet-Koeffizienten werden aus dem Niederschlagsfeld (genauer: der entsprechenden Kachel) mittels einer komplexen Dual-Tree-Wavelet-Transformation bestimmt. Anschließend werden die räumlichen Mittelwerte für die sechs verschiedenen Orientierungswinkel und vier verschiedenen räumlichen Skalenniveaus berechnet. Um die Komplexität weiter zu reduzieren, wurden die Werte für Richtung und Anisotropie pro Skalenniveau bestimmt (vgl. Abbildung 5).

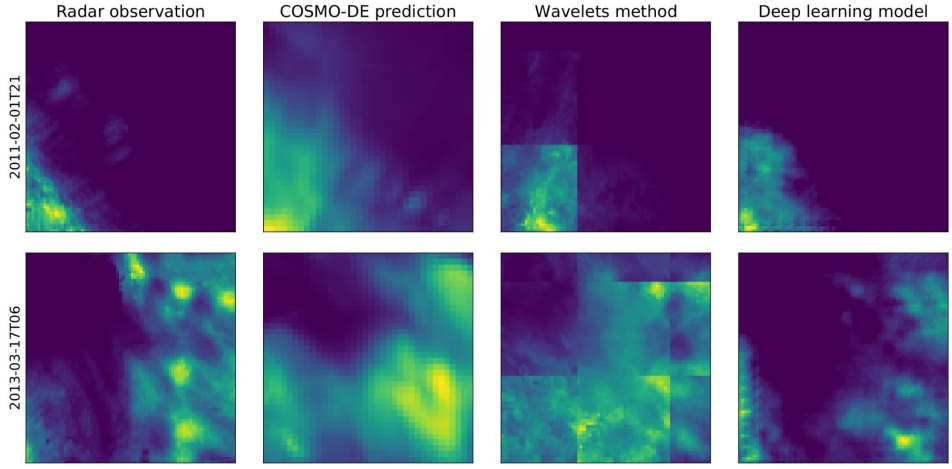


Abbildung 6: Vergleich zwischen den Prognosen des Analog Ensemble Ansatzes und des CGAN-Modells für zwei Beispielfälle. Erste Spalte: RADKLIM Beobachtungen, zweite Spalte: unbearbeitete COSMO-DE Modellvorhersage, dritte Spalte: Wavelet-basierte GLM Vorhersage, und vierte Spalte: CLGAN-Vorhersage.

Ein exemplarischer Vergleich der Beobachtungsdaten (RADKLIM-Daten), des numerischen Modells (COSMO-DE-EPS), der beschriebenen GLM-Wavelet Methode und der ML-Modelle ist in Abb. 6 dargestellt. Die Ergebnisse zeigen zwar gewisse Verbesserungen durch das AnEn im Vergleich zu den originalen COSMO-Daten, insbesondere im Hinblick auf kleinräumige Strukturen, aber die Kacheln und ihre Ränder sind immer noch sichtbar,

wohingegen das in AP3 beschriebene CLGAN-Modell ein kontinuierliches Feld liefert, das dem Niederschlagsfeld aus den RADKLIM-Beobachtungen recht gut ähnelt.

Arbeitspaket 5: Evaluierung und Vergleich der Ergebnisse (U Bonn, DWD, FZ Jülich)

Die Erfassung des gesamten COSMO-EPS-Datensatzes in Jülich ermöglichte eine umfassende Bewertung der Datenqualität und -konsistenz (siehe auch AP4). Da die Analysetools von AP5 auf die Arbeit mit den ursprünglichen GRIB-Dateien des DWD zugeschnitten waren, wurde in der frühen Projektphase ein Katalog fehlender Daten, vertauschter Zeitreihen und unvollständiger GRIB-Zeitstempel erstellt. Gemeinsam mit AP4 konnte dann ein bereinigter Datensatz erzeugt werden, der von allen Projektpartnern genutzt werden konnte.

Im Rahmen von AP5 wurde der bereinigte COSMO-DE EPS-Datensatz dann verwendet, um ein erstes Nachbereitungsverfahren zu entwickeln, das aus einem logistischen Regressionsmodell (GLM, dessen Prädiktand einer Bernoulli-Verteilung folgt) für Schwellenwert-Überschreitungen des 3-Stunden-Niederschlags (z.B. 0,5 mm/3h) an ausgewählten Beobachtungsstandorten besteht. Außerdem wurde eine Reihe geeigneter Prädiktoren aus dem COSMO-EPS abgeleitet, die in Tabelle 3 aufgeführt sind.

Tabelle 3: Prädiktoren der GLM Regression.

Grib code	Variable	Dimensionalität	Einheit
113	Convective Available Potential Energy	2D, instantan	J kg^{-1}
129	Convective Inhibition Energy	2D, instantan	J kg^{-1}
137	Cloud Cover (0 - 400 hPa)	2D, instantan	%
161	Total Cloud Cover	2D, instantan	%
170	Geopotential	3D, instantan	m^2 s^{-2}
273	Vertical velocity	3D, instantan	Pa s^{-1}
361	Surface pressure	2D, instantan	Pa
385	Relative humidity	2D, instantan	%
458	2 metre temperature	2D, instantan	K
540	Model-level temperature	3D, instantan	K
561	Total Precipitation	2D, akkumuliert	kg m^{-2}
665	10 metre U wind component	2D, instantan	m s^{-1}
721	10 metre V wind component	2D, instantan	m s^{-1}

Die in DeepRain entwickelten Nachbereitungs-Modelle wurden systematisch mit Hilfe von partiellen Korrelationen auf der Grundlage von Entropiewerten evaluiert. Diese Evaluationstechnik ermöglicht eine Zerlegung der Zielinformation (d.h. die Vorhersage des Nachbearbeitungs-Modelle) in ihre korrekt-redundanten, ihre falsch-redundanten und ihre Mehrwert-Anteile zu zerlegen. Die Zerlegung wurde in der Studie von Glowienka-Hense et al. (2020) in Anlehnung an Williams und Beer (2010) um den falsch-redundanten Teil erweitert. Eine nähere Beschreibung ist der Legende unter Abbildung 7 zu entnehmen.

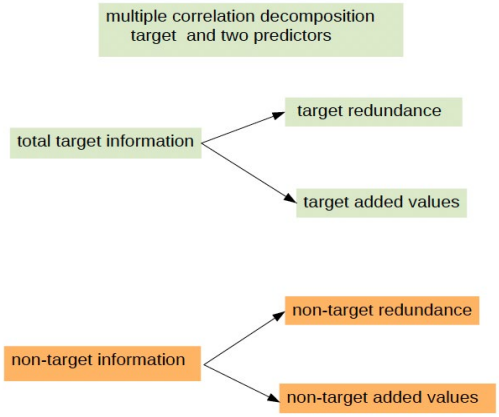


Abbildung 7: Diagramm der neuartigen Mehrfachzerlegungstechnik für den Vergleich von zwei Vorhersagesystemen auf Basis von verifizierenden Beobachtungen (=Ziel) als Erweiterung von Glowienka-Hense et al. (2020). Im Allgemeinen wird die Information durch die negative Entropie gemessen, im Falle von Gaußschen Zufallsvariablen kann sie durch die Varianz ersetzt werden. Die Zielredundanz ist die beobachtete Varianz, die von beiden Vorhersagesystemen gemeinsam vorhergesagt wird, während der Mehrwert eines Vorhersagesystems sich aus der Zunahme der Varianz gegenüber dem anderen ergibt. Die Nicht-Zielredundanz ist die Varianz in beiden Vorhersagesystemen, die in den Beobachtungen nicht dargestellt wird ("gemeinsamer Vorhersagefehler"), der Nicht-Zielmehrwert („individueller Vorhersagefehler“) ist die Varianz in einem Vorhersagesystem gegenüber dem anderen, die in den Beobachtungen nicht dargestellt wird.

Weitere Arbeiten in AP5 betrafen die Verfeinerung der zugrundeliegenden statistischen Regressionsansätze und der getesteten ML-Verfahren.

Die Analyse einzelner Stationen zeigte, dass in den Fällen hoher Vorhersage-*wahrscheinlichkeiten* (>50%) des logistischen Modells das Verhältnis von richtigen zu falsch-positiven Ereignissen viel höher ist als bei einem COSMO-Gitterpunkt in der Nachbarschaft dieser Station. Ein *Misfit* wurde analog zum Brier Skill Score (BSS) berechnet, wobei die Wahrscheinlichkeiten aus der logistischen Regression auf 0 und 1 gesetzt wurden, wenn sie kleiner bzw. größer als 0,5 waren. Diese Zahl ist somit strenger als der BSS und ist ein Maß für die relative Anzahl der falschen Ja/Nein-Entscheidungen auf Grundlage des logistischen Modells im Vergleich zum Referenzfall. Ein negativer Wert bedeutet, dass unter der Annahme, dass kein Niederschlag fällt, häufiger richtige Vorhersagen getroffen werden. Die Gesamtzahl der vom Modell korrekt vorhergesagten Niederschlagsereignisse im Verhältnis zur Gesamtzahl aller Niederschlagsereignisse liegt in der Regel bei 0,4, so dass das logistische Regressionsmodell nur einen relativ kleinen Teil der Niederschlagsereignisse erfasst. Um diesen Anteil zu erhöhen, wurden weitere logistische Modelle nach der Methode der Hauptkomponentenanalyse (Anderson, 1984, S.454) untersucht. Zu diesem Zweck wurden die Zeitpunkte, die vom ersten Modell als Niederschlagsereignisse eingestuft wurden, aus den Daten entfernt. Diese Methode wird im Bereich des maschinellen Lernens als Boosting bezeichnet (Friedman et al. 2000). Folglich sollten die Ergebnisse des logistischen Modells nur im Falle der Vorhersage eines Niederschlagsereignisses verwendet werden.

In den Fällen, in denen Niederschlagsereignisse vorhergesagt werden, wird die Eintrittswahrscheinlichkeit im Allgemeinen durch das Boosten erhöht. Ein dritter Boosting-Schritt der logistischen Regression wurde ebenfalls getestet, nachdem die Ereignisse eliminiert wurden, für die die zweite Regression eine Niederschlagswahrscheinlichkeit von

mehr als 50 % vorhersagte. Auch dieser Boosting-Schritt führte zu einem ähnlich informativen Modell. Daraus ergibt sich die so genannte bedingte Bewertung: "Da eine Schwellenwertüberschreitung vorhergesagt wird", ist die Wahrscheinlichkeit des beobachteten Niederschlags sehr hoch.

Die Regressionsmodelle auf Basis der getesten neuralen Netzwerke (NN) können jedoch nicht in der gleichen Weise behandelt werden wie die logistischen Regressionsmodelle mit dem Boosting-Verfahren. Durch Training von mehreren NN unter zufälliger Variation der initialen Parameter des Optimierungsprozesses wurde eine Ensemble-Interpretation ermöglicht, deren Ergebnisse dem geboosteten logistischen Regressionsmodell ähneln. Vergleicht man die logistische Regression mit Boosting und die Ensemble-Interpretation des NN, lässt sich eine Lücke in der Vorhersagbarkeit identifizieren: Es scheint, dass mehrere Lösungen (Nandwani et al., 2020; Holzinger et al., 2021) für das Regressionsproblem in dem Sinne existieren könnten, dass derselbe Schwellenwert des Niederschlags von verschiedenen meteorologischen Strömungskonfigurationen überschritten wird, ein Merkmal, das jedem operationellen Wettervorhersager wohl bekannt ist.

Weitere Arbeiten in AP5 waren der Definition der Kostenfunktion gewidmet, insbesondere im Fall von binären Ereignissen wie Schwellenwert-Überschreitungen des Niederschlags. Der entwickelte Ansatz basiert auf einer strikten Anwendung der Bayes'schen Statistik, bei der die posteriore (bedingte) Wahrscheinlichkeit eines Vorhersagesystems zu einer festen Vorhersagezeit auf Basis von binären Beobachtungen evaluiert wird. Hierbei wird ein Gauß'sches Prior-Modell zur Beschreibung des Vorhersage-Ensembles mit einer Wahrscheinlichkeits-Massenfunktion vom Bernoulli-Typ kombiniert, um die Wahrscheinlichkeit für die binären Ereignisse abzuschätzen. Der Posterior wird dann durch ein hochdimensionales Integral über den Phasenraum der Vorhersagen bestimmt. Um eine kostspielige Markov-Kette Monte-Carlo Auswertung solcher Integrale zu vermeiden, wird die Sattelpunkt/Laplace-Näherung (Reich und Cotter, 2015) angewendet, die semi-analytisch berechnet werden kann. Sie definiert als Kostenfunktion die freie Energie der posterioren Dichte. Der Ansatz kann auf jede Wahrscheinlichkeitsfunktion verallgemeinert werden, die als Wahrscheinlichkeitsdichtefunktion der Exponentialfamilie geschrieben werden kann (Wainwright und Jordan, 2008).

Arbeitspaket 6: System-Design und Workflow-Analyse (FZ Jülich and Jacobs U Bremen)

Die anfängliche Arbeit von AP6 konzentrierte sich auf die Übertragung der DWD COSMO-EPS- und RADKLIM-Datensätze auf das JSC-Speichersystem (siehe Beschreibung in AP2 oben).

Um den Umgang mit großen Datenmengen, zu erleichtern, entwickelten wir eine effiziente und einfach zu implementierende, skalierbare Software in Python für parallele Berechnungen auf HPC-Systemen (PyStager, siehe github). PyStager wird nun im Rahmen des EuroHPC-JU-Projekts MAELSTROM weiterentwickelt. Nach dem Datenvorfall im Frühjahr 2021 wurde ein auf PyStager basierender Überwachungsprozess entwickelt, um eine mögliche Beschädigung von Datenblöcken zu erkennen. Dies trug dazu bei, Datenfehler während des zweiten Datentransfers im Mai 2021 zu vermeiden (siehe AP2 oben). PyStager wurde auch zur Vorverarbeitung und Reprojektion des RADKLIM-Datensatzes verwendet, um die Zusammenführung mit den COSMO-DE EPS-Daten zu ermöglichen.

Das Hauptziel des Arbeitspakets 6 war es, eine vollständige Prozesskette für die Verarbeitung von Modelldaten im Terabyte-Bereich in ML-Anwendungen für Wetter und Klima zu entwerfen und zusammenzustellen. Es wurden mehrere Konzepte entwickelt, die die Arbeit in AP2 und

AP3 komplementierten, und es wurden Software-Tools entwickelt, um die Projektpartner bei der effizienten Datenverarbeitung zu unterstützen (Abbildung 8). Zusammen decken die in DeepRain entwickelten Workflow-Komponenten die meisten Aspekte der im Antrag vorgesehenen Nachprozessierungs-Anwendung für Niederschlagsvorhersagen ab. Die Integration dieser Komponenten in ein kohärentes Software-Framework konnte jedoch aufgrund von Verzögerungen in anderen Arbeitspaketen und eventuellem Zeitmangel nicht erreicht werden.

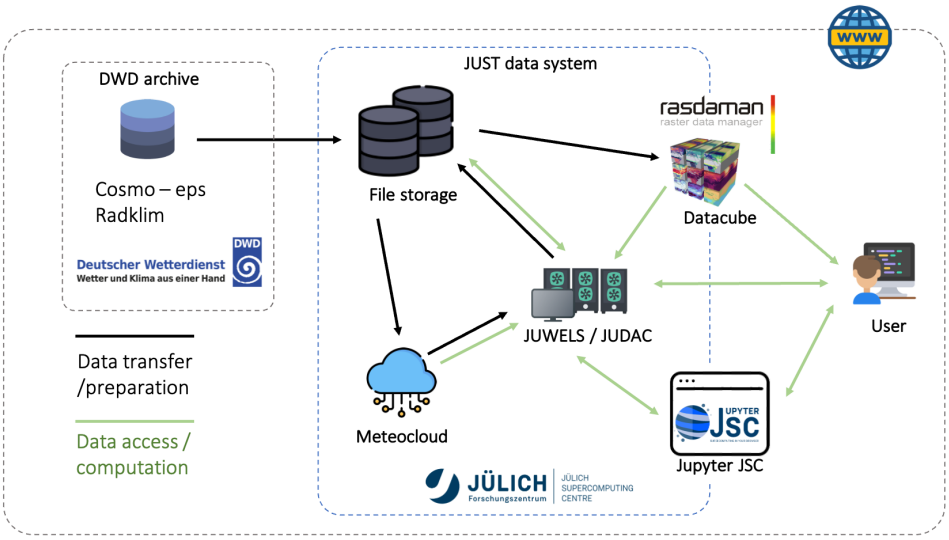


Abbildung 8: Schaubild des Datentransfers, der Datenverarbeitung, sowie der unterschiedlichen Zugriffswege und Weiterverarbeitungswege in DeepRain.

Um über die klassische Workflow-Entwicklung hinauszugehen, untersuchten wir das Potenzial moderner Softwarekonzepte zur Verbesserung der FAIRness von Prozessketten im Bereich meteorologischer Datenverarbeitung. FAIRness zielt darauf ab, Daten und Metadaten auffindbar, zugänglich, interoperabel und wiederverwendbar zu machen (Wilkinson et al., 2016). Im Zuge des Projekts haben ein neuartiges Konzept vorgeschlagen, das auf FAIR digitalen Objekten (FDO) basiert (De Smed et al., 2020). Dieses Konzept nutzt die hochmoderne Jupyter-JSC-Infrastruktur in Verbindung mit dem gitlab-Server (Abb. 9; Mozaffari et al., 2022).

Wie in Mozaffari et al. (2022) beschrieben, sind Jupyter-Notebooks als solche nicht gut für die reproduzierbare Gestaltung von Arbeitsabläufen geeignet, da sich ihr Zustand und Inhalt bei jeder Benutzerinteraktion ändert. Um dieses Problem zu umgehen, wird die Versionskontroll-Funktion von gitlab in das Jupyter-Notebook integriert, wodurch eine Dokumentation von Änderungen gewährleistet ist. Ein Dashboard für ML-Experimente, das modernen Tools wie *Weights and biases* oder *MLFlow* ähnelt, fungiert als FDO. Spezielle Programme (z. B. mit der Python-Bibliothek papermill) ermöglichen zudem die Initiierung von mehreren ML-Experimenten mit vordefinierten Parametersätzen, die über das Dashboard konfiguriert werden. Auf diese Weise können einfach zu verwendende und reproduzierbare Arbeitsabläufe erstellt und mit anderen Forschern geteilt werden.

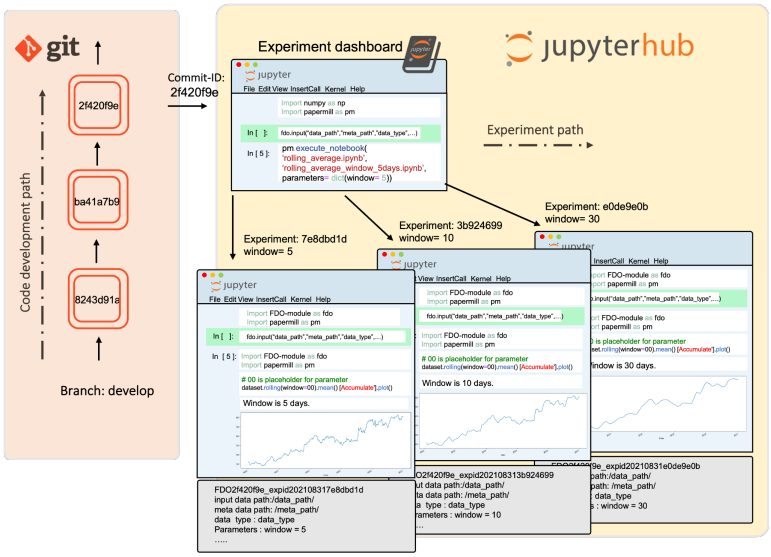


Abbildung 9: Veranschaulichte Integration von git in das HPC-fähige Jupyter-System am JSC über ein einfaches Experiment-Dashboard. Individuelle Experimente mit benutzerdefinierten Parametern werden mittels papermill über eine individuelle Jupyter-Notebook Instanzen ausgeführt, während die entsprechende FDO automatisch erzeugt wird (aus Mozaffari et al., 2022).

3) Voraussichtlicher Nutzen und Verwertbarkeit der Projektergebnisse

Das DeepRain-Projekt untermauerte das Potenzial moderner Deep-Learning-Methoden zur Verbesserung von lokaler Niederschlagsvorhersagen auf Basis von Ensemble-Vorhersagen eines numerischen Wettervorhersagemodells sowie von Radar- und Stationsbeobachtungen. Es zeigte auch, wie wichtig die Nutzung modernster IT-Technologien für die Verwaltung von Big Data und deren parallelisierte Verarbeitung ist. Entgegen der ursprünglichen Planung war es nicht möglich, alle Komponenten zu einer kohärenten Servicearchitektur zu entwickeln, die als Prototyp für einen operationellen Betrieb beim DWD dienen könnte. Dennoch wurden im Projekt die Voraussetzungen dafür geschaffen, dass bei ausreichender Finanzierung in naher Zukunft moderne Konzepte des maschinellen Lernens Teil der operationellen Wettervorhersage in Deutschland werden können.

Notwendige Schritte zur Erreichung dieses Ziels sind unter anderem:

- Anpassung der DeepRain-Datenverarbeitungskette an das aktuelle operationelle ICON-Modell des DWD,
- Anpassung der DeepRain-Datenverarbeitungskette an die Hardware-Architektur und Speichersysteme des DWD,
- Training der in DeepRain entwickelten ML-Modelle mit ICON-Daten -- ggfls. Genügt ein weniger aufwendiges Nachtrainieren (*Finetuning*),

- Weitere Verbesserungen und Evaluierung der in DeepRain eingesetzten ML-Modelle,
- Entwicklung von robusten, automatisierten Output-Verarbeitungsroutinen,
- Systemintegration und -test.

Neben dem Potenzial zur Operationalisierung von DeepRain-Komponenten werden viele der DeepRain-Ergebnisse, Softwarepakete und Datenverarbeitungsstrategien in der kommenden akademischen Forschung und Lehre eingesetzt werden. Die DeepRain-Ergebnisse werden bereits im EuroHPC-JU-Projekt MAELSTROM und im ERC Advanced Grant IntelliAQ eingesetzt. Die in DeepRain begonnene Verknüpfung von Rasdaman mit Prozessierungsketten des maschinellen Lernens wird bereits in den Folgeprojekten EU H2020 CENTURION, BMWI AI-Cube und EU H2020 FAIRiCUBE fortgesetzt. Die im Rahmen von DeepRain installierte Rasdaman-Instanz und die Integration in den EarthServer-Verbundknoten werden inzwischen auch im rheinischen Projekt BioökonomieREVIER eingesetzt.

Im Einzelnen planen die DeepRain-Projektpartner die Nutzung der entwickelten Werkzeuge und -Ergebnisse wie folgt:

Universität Bonn und DWD:

Beide Institutionen planen eine Weiterentwicklung der konventionellen Vorhersagemethoden durch eine Kombination von Ansätzen, die den Vorschlägen von Buschow und Friederichs (2021) und Brune et al. (2021) zur Verwendung von Wavelets folgen. Ziel ist, die räumliche Struktur von multi-skaligen Feldern wie Niederschlag effektiv zu beschreiben.

Alle Methoden und Ergebnisse des DeepRain-Projekts werden in zukünftige Lehrveranstaltungen des Masterstudiengangs "Physik der Erde und Atmosphäre" an der Universität Bonn integriert. Dabei wird nicht nur der notwendige theoretische Hintergrund vermittelt, sondern es können auch Übungen am Computer mit COSMO- und RADKLIM-Datensätzen sowie mir Synop-Stationsbeobachtungen des DWD durchgeführt werden. Jupyter-Notebooks mit R oder Python sind bereits ein Standardlehrmittel in den Studiengängen der Universität Bonn und erleichtern den Übergang vom Forschungscharakter der DeepRain-Ergebnisse in die Lehre.

Universität Osnabrück und FZ Jülich:

Die Entwicklung und Anwendung von tiefen neuronalen Netzen für die Niederschlagsvorhersage und das Downscaling wird im Rahmen von MAELSTROM und anderen Projekten fortgeführt. Neben verschiedenen Varianten von GAN-Modellen werden in Zukunft auch neue, Transformer-basierte Deep Learning-Architekturen wie der Swin-Transformer (siehe z.B. Liang et al., 2021 und Zhang et al., 2022) erforscht. Die Methoden werden zudem für die Anwendung auf weitere meteorologische Variablen wie Temperatur oder Wind erweitert.

Darüberhinaus wird eine Überführung der GAN-Modelle in eine probablistische Anwendung exploriert, um der chaotischen und nicht-linearen Natur von Niederschlagsprozessen gerecht zu werden (Gilleland et al., 2009). Das FZ Jülich plant auch zu untersuchen, ob die Einbeziehung physikalischer Randbedingungen in das Optimierungsverfahren dazu beitragen kann, dass die generierten Niederschlagsvorhersageprodukte realistisch bleiben.

Die in DeepRain entwickelten HPC-fähigen ML-Prozessierungsketten und Datenverarbeitungsabläufe werden in verwandten Forschungsprojekten sowie in Trainingskursen eingesetzt.

Jacobs Universität Bremen

Die Jacobs University wird das Open-Source Community-Projekt Rasdaman weiterhin leiten und dessen Entwicklung vorantreiben, um die Nutzung der Datenwürfeltechnologie zu verbessern und ihre Akzeptanz zu fördern. Darüber hinaus wird die von der Jacobs University kuratierte EarthServer-Föderation verbessert und erweitert werden. Insbesondere die Integration von Datenwürfeln in KI-Prozessketten wird weiter erforscht.

Jacobs Universität Bremen und FZ Jülich:

Beide Institutionen arbeiten im Rahmen des Projekts Digitales Geosystem Rheinisches Revier (DG-RR) weiter an großskaligen Klimadatendiensten. Die ersten Datensätze des DG-RR-Projekts sind bereits in den föderierten Rasdaman-Knoten integriert und über den Earth Server-Verbund verfügbar. Diese können mit Hilfe des Standard Web Coverage Service (WCS) und des Web Map Service (WMS) leicht abgerufen und mit dem Standard Web Coverage Processing Service (WCPS) verarbeitet werde. Hierdurch wird die Datenextraktions- und Verarbeitungszeiten erheblich verkürzt und damit die Qualifikations-hürde für den Nutzer gesenkt.

FZ Jülich:
Über die Entwicklung des Workflow-Design, der HPC-Systemfähigkeit, der Entwicklung und der Implementierung des DeepRain-Projekts in die JSC-Infrastruktur wurden die Möglichkeiten einer big data ML-Pipeline aufgezeigt. Das kürzlich vom BMBF geförderte Projekt WarmWorld sieht die Implementierung und Untersuchung eines operationelles Datenspeichersystems vor, das sich am vom ECMWF betriebenen MARS-System orientiert. Darüber hinaus wird die Integration der FREVA-Software das Konzept des föderierten Datenzentrums auf eine neue Ebene heben, bei dem die Datenzentren des ECMWF, des DKRZ und des FZ Jülich miteinander verbunden werden.

Neuartige Konzepte für reproduzierbare, großskalige Datenanalyse-Prozessierungsketten, die im Rahmen von DeepRain konzipiert wurden, sollen in Folgeprojekten weiterentwickelt werden. Insbesondere ist geplant, funktionale Software zu entwickeln, um das FAIR Digital Object Konzept in realen Arbeitsabläufen mit Erdsystemdaten zu testen (Mozaffari et al., 2022b).

4) Relevante wissenschaftliche Entwicklungen außerhalb des DeepRain Projektes

Während der Laufzeit des DeepRain-Projekts gab es viele Fortschritte bei ML-Anwendungen für Wetter und Klima, bei der parallelen Verarbeitung von Big Data sowie bei der Bereitstellung von Erdsystem-Datendiensten.

Verschiedene meteorologische Agenturen und Forschungseinrichtungen auf der ganzen Welt haben frei zugängliche, groß angelegte Erdsystem-Datendienste eingerichtet. Nennenswerte Beispiele in diesem Zusammenhang sind u.a.: die Klimadatenspeicher des ECMWF, der ESA, der NASA und der NOAA. Auch der Deutsche Wetterdienst hat einen großen Teil seines Archivs geöffnet und über das DWD-Geoportal zugänglich gemacht. Während diese Websites im Allgemeinen gute Funktionen für die Suche, die Visualisierung und das Herunterladen großer Datensätze bieten, sind die Möglichkeiten der On-Demand-

Verarbeitung minimal, vor allem wenn es um die Ableitung neuer Informationen aus mehreren Eingabevariablen geht. Dies ist die besondere Stärke von Datenwürfeln, wie sie im EarthServer-Verbunddienst zu finden sind. Speziell zu Array-Datenbanken wurde von der Jacobs University eine umfassende Analyse einer Vielzahl von Ansätzen veröffentlicht (Baumann et al, 2021).

Im Zusammenhang mit der parallelen Datenverarbeitung von Wetter- und Klimadaten ergaben sich erhebliche technische Verbesserungen. Das EZMW produziert täglich ca. 120 TeraByte Rohdaten aus hochauflösenden Modellen und Ensemble-Vorhersagen. Eine maßgeschneiderte Datenverarbeitungslösung des EZMW, das so genannte Meteorological Archival and Retrieval System (MARS) ermöglicht es den Nutzern, meteorologische Daten im GRIB- oder NetCDF-Format aus dem riesigen Archiv zu erkunden und abzurufen. Die Rohdaten werden ebenfalls in MARS gespeichert (dem weltweit größten meteorologischen Archiv, das derzeit über 300 PetaBytes an Primärdaten enthält). In Deutschland zielt das vom Bundesministerium für Bildung und Forschung (BMBF) geförderte Forschungsprojekt RegIKlim darauf ab, entscheidungsrelevantes Wissen über den Klimawandel in Kommunen und Regionen zu entwickeln, um eine solide Grundlage für regionalspezifische Informations- und Bewertungsdienste zu schaffen. ReglKlim hat mit dem Free Evaluation System Framework (FREVA) ein standardisiertes Daten- und Bewertungssystem geschaffen - entwickelt am DKRZ und der FUB in Deutschland. FREVA bietet einen effizienten und umfassenden Zugriff auf die Modelldatenbank und die Auswertungsdatensätze. Das Anwendungssystem ist als einfach zu bedienende Anwendung konzipiert, die die technischen Anforderungen an Nutzer und Tool-Entwickler minimiert.

Im Bereich Deep Learning hat sich das exponentielle Wachstum fortgesetzt, und insbesondere ML-Anwendungen im Wetter- und Klimabereich haben bei führenden Entwicklern von Deep-Learning-Methoden, darunter Großunternehmen wie Google, Amazon, Microsoft oder NVidia, erhöhte Aufmerksamkeit erlangt. Während zu Beginn des DeepRain-Projekts Convolutional Neural Networks (CNNs) als Stand der Technik galten, haben inzwischen Generative Adversarial Networks (GANs) eine zentrale Rolle eingenommen (siehe z. B. Price und Rasp, 2022; Harris et al., 2022, und Jeong und Yi, 2022). Die Prozessierung von Ensemble-Vorhersagen mit tiefen neuronalen Netzen führt zu erheblichen Verbesserungen der Vorhersagequalität (Grönquist et al., 2020). Darüber hinaus versprechen auch graphische neuronale Netze (Graph Neural Networks) einen bedeutenden Durchbruch bei der (globalen) Wettervorhersage (Keisler, 2022).

Eine neuere Entwicklung stellt die Nutzung von sogenannten Transformer-Netzwerken dar (Vaswani et al., 2017). Unter der Bedingung von hinreichend großen Trainingsdatensätze können diese Modelle eine generalisierte Abstraktion des Atmosphärenzustands erlernen, die dann in einer Vielzahl von sogenannten Downstream-Anwendungen genutzt werden kann. Niederschlags-Downscaling ist ein mögliches Beispiel hierfür. FZ Jülich ist an der atmorep-Initiative beteiligt, die einen Prototyp eines Transformer-Netzwerks auf der Grundlage von ERA5-Daten entwickelt (Hoffmann und Lessig, 2022).

Im Rahmen des DeepRain-Projekts konnten verschiedene Verbesserungen der konventionellen Wetter- und Klimavorhersage erreicht werden. Das EZMW hat eine klassische Nachbearbeitungsmethode für Niederschlagsvorhersagen (EcPoint) eingeführt, die ähnliche Ansätze wie die Analog-Ensemble Technik verwendet (Hewson et al., 2021). Weitere Verbesserungen bei der Verifizierung von Niederschlagsprognosen und beim Fehlervergleich (Stein, J., & Stoop, F., 2019; Buschow und Friederichs, 2021), sowie Verbesserungen beim statistischen Postprocessing für Wettervorhersagen wurden publiziert (Vannitsem et al., 2021). Der aktuellen Arbeit von Sha et al. (2022) folgend könnte sich in Zukunft auch ein hybrider Ansatz aus Analog-Ensemble Technik und ML-Methoden für die Niederschlagsvorhersage durchsetzen.

Im Bereich des Forschungsdaten-Managements und speziell der FAIR-Praktiken in der Forschung stellen Fair Digital Objects (FDO; De Smed et al., 2020) eine neue Methode dar, um die Reproduzierbarkeit wissenschaftlicher Studien zu gewährleisten. In FDOs werden Daten, Code und Dokumentation zu einem einzigen, unabhängigen und selbsterklärenden Informationspaket zusammengefasst. Durch persistente Objektreferenzen (persistent identifiers = PID) werden Daten, Metadaten und maschinenlesbare Beschreibungen von Analyseketten verknüpft, sodass archivierte Auswertungen vollständig reproduziert werden können. Es wird allgemein erwartet, dass FDOs zu einem neuen Standard für Open-Data-Repositories werden. So plant beispielsweise die niederländische Regierung, FDOs im Oktober 2022 als neuen Datenstandard anzuerkennen. Andere in Wissenschaftskreisen betriebene Bemühungen um granulare FAIR-Segmente, wie RO-CRATE, wurden in letzter Zeit ausgeweitet (Soiland-Reyes et al., 2021).

5) Veröffentlichungen der DeepRain Projektergebnisse

Arbeitspaket 2:

- Baumann P., (2021), "Towards a Model-Driven Datacube Analytics Language," IEEE International Conference on Big Data (Big Data), pp. 3740-3746, https://doi.org/10.1109/BigData52589.2021.9672038. [**veröffentlicht**]
- Baumann, P. (2021). A General Conceptual Framework for Multi-Dimensional Spatio-Temporal Data Sets. Environmental Modelling & Software. 143. 105096. https//doi.org/10.1016/j.envsoft.2021.105096. [**veröffentlicht**]
- Baumann, P., Misev, D., Merticariu, V. et al. (2021), Array databases: concepts, standards, implementations. J Big Data 8, 28. https://doi.org/10.1186/s40537-020-00399-2 [**veröffentlicht**]
- Campos Escobar O. J. , Misev D. and Baumann P., (2020), "Making an Array Database Language Server-Side Extensible," IEEE International Conference on Big Data (Big Data), pp. 2743-2750, https://10.1109/BigData50022.2020.9378108. [**veröffentlicht**]
- Villarroya S. and Baumann P. (2020), "On the Integration of Machine Learning and Array Databases," IEEE 36th International Conference on Data Engineering (ICDE), pp. 1786-1789, https://10.1109/ICDE48307.2020.00170. [**veröffentlicht**]
- Villarroya, S., Baumann, P. A survey on machine learning in array databases. Appl Intell (2022). https://doi.org/10.1007/s10489-022-03979-2. [**veröffentlicht**]

Arbeitspaket 3:

- Kesselheim, S., Herten, A., Krajsek, K., Ebert, J., Jitsev, J., Cherti, M., ... & Lippert, T. (2021, June). JUWELS Booster–A Supercomputer for Large-Scale AI Research. In *International Conference on High Performance Computing* (pp. 453-468). Springer, Cham. https://doi.org/10.1007/978-3-030-90539-2_31 [**veröffentlicht**]
- Schultz, M. G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L. H., ... & Stadtler, S. (2021). Can deep learning beat numerical weather prediction?. *Philosophical Transactions of the Royal Society A*, *379*(2194), 20200097. https://doi.org/10.1098/rsta.2020.0097. [**veröffentlicht**]
- Gong, B., Langguth, M., Ji, Y., Mozaffari, A., Stadtler, S., Mache, K., & Schultz, M. G. (2022). Temperature forecasting by deep learning methods. *Geoscientific Model Development Discussions*, 1-35. https://doi.org/10.5194/gmd-2021-430. [**zur Veröffentlichung angenommen**]

- Ji, Y., Gong, B., Langguth, M., GAN-based video prediction models for precipitation nowcasting **[eingereicht]**
- Rojas-Campos, A., Langguth, M., Wittenbrink, M. & Pipa, G. (2022). Deep learning model for generation of precipitation maps based on Numerical Weather Prediction. EGUsphere. https://doi.org/10.5194/egusphere-2022-648 **[in Begutachtung, Preprint verfügbar]**

Arbeitspakete 3 und 4:

- Rojas-Campos, A., Wittenbrink, M., Nieters, P., Schaffernicht, E., Keller, J. D. & Pipa, G. (2021). Post-processing of NWP precipitation forecasts using deep learning. *Weather and Forecast*. [**in Begutachtung**]

Arbeitspaket 4:

- Wittenbrink, M., Keller, J. D. (2022). A two-dimensional analog ensemble approach for precipitation forecasting based on wavelet transforms [**geplant**]

Arbeitspaket 5:

- Glowienka-Hense, R., Hense, A., Brune, S., and Baehr, J (2020): Comparing forecast systems with multiple correlation decomposition based on partial correlation, Adv. Stat. Clim. Meteorol. Oceanogr., 6, 103–113, https://doi.org/10.5194/ascmo-6-103-2020 [**veröffentlicht**]
- Glowienka-Hense, R., Hense A. (2022): Evaluating models sensitivities with partial multiple correlation decomposition, [**geplant**]

Arbeitspaket 6:

- Mozaffari, A., Langguth, M., Gong, B., Ahring, J., Rojas-Campos, A., Nieters, P., Campos Escobar, O.J., Wittenbrink, M., Baumann, P., Schultz, M. (2022) ; HPC-oriented Canonical Workflows for Machine Learning Applications in Climate and Weather Prediction. *Data Intelligence*; https://doi.org/10.1162/dint_a_00131 [**veröffentlicht**]

6) Mittelverwendung

Die für das DeepRain Projekt bewilligten Mittel i.H.v. knapp über 2 Mio. € wurden hauptsächlich zur Finanzierung der Gehälter von Postdoktoranden und Doktoranden verwendet, die den Hauptteil der Projektarbeit leisteten. Dabei wurden 96 % der geplanten Mittel verausgabt. 40.000 € wurden in die Beschaffung eines Rechenknotens mit GPU Prozessoren am JSC investiert. Dieser wurde in ein größeres Rechnersystem integriert, wodurch den Projektteilnehmern mehr Rechenleistung für ihre Anwendungen zur Verfügung stand, als dies mit einem Einzelsystem erreichbar gewesen wäre. Die Rechenzeiten wurden genau protokolliert und anteilig für die Abschreibung des Systems abgerechnet. Von den bewilligten Reisekosten i.H.v. insgesamt 36.673 € wurden lediglich 16.226 € (= 44 %) ausgegeben, da es wegen der Covid-19 Pandemie zu erheblichen Einschränkungen der Reisetätigkeit kam. Sonstige Ausgaben umfassen 74.630 €, von denen 50.397 € (= 68 %) ausgegeben wurden. Die Aufteilung der Projektausgaben nach Kostenart ist in Abbildung 10 dargestellt.

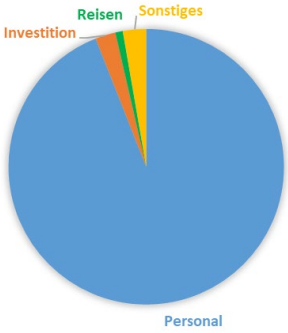


Abbildung 10: Übersicht über die Projektausgaben nach Kategorien

Abbildung 11 zeigt die Aufschlüsselung des Arbeitseinsatzes nach Arbeitspaket. Hierbei ist zu beachten, dass die aus dem Projekt finanzierten Mitarbeiter:innen vor allem in den Arbeitspaketen 1 und 6 durch vorhandenes Personal unterstützt wurden, so dass der reale Anteil etwas gleichmäßiger zwischen den Arbeitspaketen verteilt ist.

Wie aus Abbildung 11 ersichtlich, wurde mehr als die Hälfte der projektfinanzierten 246 Personenmonate in den beiden Kernthemen des Projekts verausgabt. Ohne diese Förderung wären die Entwicklung des Daten-Managements, der maschinellen Lernverfahren und der Evaluierungsmethoden nicht durchführbar gewesen. Auch die gut funktionierende transdisziplinäre Zusammenarbeit zwischen den Partnern hätte es ohne die Projektfinanzierung nicht gegeben.

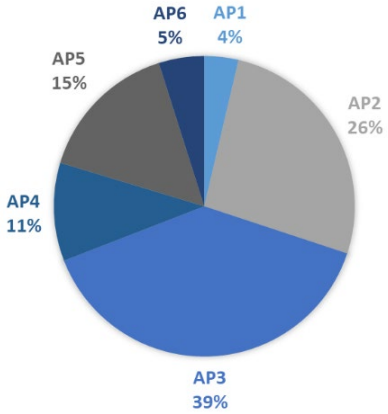


Abbildung 11: Verteilung der durch DeepRain finanzierten Personenmonate auf die Arbeitspakete. Insgesamt wurden 246 Personenmonate veranschlagt.

7) Literaturverzeichnis

Abadi, M., et al. (2015) *TensorFlow: Large-scale machine learning on heterogeneous systems*. https://www.tensorflow.org/

Anderson, T.W. 1984: An introduction to multivariate statistical analysis (second edition). Wiley series in probability and mathematical statistics.John Wiley and Sons, New York, 675pp

Bach, Liselotte; Schraff, Christoph; Keller, Jan D. and Hense, Andreas (2016): Towards a probabilistic regional reanalysis system for Europe: evaluation of precipitation from experiments - Tellus A, Vol. 68, No. 1, https://doi.org/10.3402/tellusa.v68.32209

Buschow, S, Friederichs, P. (2021) SAD: Verifying the scale, anisotropy and direction of precipitation forecasts. *QJR Meteorol Soc*.; https://doi.org/10.1002/qj.3964

Brune, S., S. Buschow and P. Friederichs (2021): The Local Wavelet-based Organization Index - Quantification, Localization and Classification of Convective Organization from Radar and Satellite Data. *Q. J. R. Meteorol. Soc*. **147**, 1853-1872, https://doi.org/10.1002/qj.3998

De Smedt K, Koureas D, Wittenburg P (2020) FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units. Publications 8: 21. https://doi.org/10.3390/publications8020021

Dmitrienko, V. D., Zakovorotnyi, A. Y., Leonov, S. Y., & Khavina, I. P. (2014). Neural Networks Art: Solving problems with multiple solutions and new teaching algorithm, The Open Neurology Journal, 8, 15

Dorninger, M., P. Friederichs, S. Wahl, M. P. Mittermaier, C. Marsigli, and B. G. Brown (2018): Editorial: Forecast verification methods across time and space scales – Part I. - Meteorologische Zeitschrift 27, 433 - 434, https://doi.org/10.1127/metz/2018/0955

Feng, Y., Zhou, M. & Tong, X. (2020). Imbalanced classification: a paradigm-based review. *arXiv*. https://doi.org/10.48550/arxiv.2002.04592

Friedman, J., Hastie, T. and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*, *28*, 337-407.

Gilleland, E., Ahijevych, D., Brown, B. G., Casati, B., & Ebert, E. E. (2009). "*Intercomparison of spatial forecast verification methods*". Weather and forecasting, 24(5), 1416-1430.

Grönquist, P. Yao, C., Ben-Nun, T., Dryden, N., Dueben, P., Li, S., Hoefler, T, (2021), "Deep learning for post-processing ensemble weather forecasts",Philosophical Transactions of the Royal Society, https://doi.org/10.1098/rsta.2020.0092

Harris, L., McRae, A. T., Chantry, M., Dueben, P. D., & Palmer, T. N. (2022). "*A Generative Deep Learning Approach to Stochastic Downscaling of Precipitation Forecasts*". *arXiv preprint arXiv:2204.02028*.

Hewson, T. D., Pillosu, F. M. (2021) *"A low-cost postprocessing technique improves weather forecasts around the world"* Communications Earth & Environment https://doi.org/10.1038/s43247-021-00185-9.

Hoffmann, S., and Lessig, C. (2021) Towards representation learning for atmospheric

dynamics, arXiv:2109.09076

Holzinger, A., Malle, B., Saranti, A., & Pfeifer, B. (2021), Towards multi-modal causability with Graph Neural Networks enabling information fusion for explainable AI. Information Fusion, 71, 28-37.

Jeong, C.-H. and Yi, M. Y. (2022) *Correcting rainfall forecasts of a numerical weather prediction model using generative adversarial networks*, The Journal of Supercomputing, https://doi.org/10.1007/s11227-022-04686-y .

Keisler, R. (2022). Forecasting Global Weather with Graph Neural Networks. *arXiv preprint arXiv:2202.07575*.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). "*Swin transformer: Hierarchical vision transformer using shifted windows*". Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 10012-10022).

Martinez-Villalobos, Cristian, and J. David Neelin. "Why do precipitation intensities tend to follow gamma distributions?." *Journal of the Atmospheric Sciences* 76.11 (2019): 3611-3631.

Mozaffari, A., Selke, N., Schultz M., (2022b) Advancing caching and automation with FDO (in press)

Price, I., and Rasp, S. (2022) *"Increasing the accuracy and resolution of precipitation forecasts using deep generative models*." International Conference on Artificial Intelligence and Statistics. PMLR, 2022. Available at https://doi.org/10.48550/arXiv.2203.12297.

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., ... & Mohamed, S. (2021). "*Skilful precipitation nowcasting using deep generative models of radar"*. Nature, 597(7878), 672-677.

Raissi, M., Yazdani, A. and Karniadakis, G. E. (2020). Hidden fluid mechanics: Learning velocity and pressure fields from flow visualizations. *Science*, *367* (6481), 1026-1030

Reich, S. and Cotter, C. (2015). Probabilistic forecasting and Bayesian data assimilation. Cambridge University Press. 296pp

Stein, J., and Stoop, F. (2019). Neighborhood-Based Contingency Tables Including Errors Compensation, *Monthly Weather Review;* https://doi.org/10.1175/MWR-D-17-0288.1

Soiland-Reyes, S., et al.: (2021) Packaging research artefacts with RO-Crate. arXiv preprint arXiv: 2108.06503

Nandwani, Y., Jindal, D., & Singla, P. (2020), Neural Learning of One-of-Many Solutions for Combinatorial Problems in Structured Output Spaces. arXiv preprint arXiv:2008.11990.

Vannitsem et al. (2021). Statistical Postprocessing for Weather Forecasts: Review, Challenges, and Avenues in a Big Data World, *Bulletin of the American Meteorological Society;* https://doi.org/10.1175/BAMS-D-19-0308.1

Sabrina Wahl (2015): Uncertainty in mesoscale numerical weather prediction: probabilistic forecasting of precipitation, 2015, Bonner Meteorologische Abhandlung, 108 S, https://hdl.handle.net/20.500.11811/6560

Shorten, C., & Khoshgoftaar, T. M. (2019). "*A survey on image data augmentation for deep learning.*" Journal of big data, 6(1), 1-48.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., Polosukhin, I. (2017) *Attention is all you need*, Proceedings of the 31st International Conference on Neural Information Processing Systems, Red Hook, NY, USA, ISBN 978-1-5108-6096-4.

Wainwright, M. J., and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, *1*(1–2), 1-305.

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.1

Williams, P. L. and Beer, R. D. (2010): Nonnegative Decomposition of Multivariate Information, arXiv [preprint], arXiv:1004.2515,14 April 2010

Zhang, B., Gu, S., Zhang, B., Bao, J., Chen, D., Wen, F., ... & Guo, B. (2022). "*Styleswin: Transformer-based gan for high-resolution image generation*". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11304-11314).

Zolina, O., Kapala, A., Simmer, C. and Gulev, S. K. (2004). Analysis of extreme precipitation over Europe from different reanalyses: a comparative assessment. Global and Planetary Change, 44(1-4), 129-161.

Schriften des Forschungszentrums Jülich
IAS Series

Band / Volume 40
Extreme Data Workshop 2018
Forschungszentrum Jülich, 18-19 September 2018
Proceedings
M. Schultz, D. Pleiter, P. Bauer (Eds.) (2019), 64 pp
ISBN: 978-3-95806-392-1
URN: urn:nbn:de:0001-2019032102

Band / Volume 41
A lattice QCD study of nucleon structure with physical quark masses
N. Hasan (2020), xiii, 157 pp
ISBN: 978-3-95806-456-0
URN: urn:nbn:de:0001-2020012307

Band / Volume 42
Mikroskopische Fundamentaldiagramme der Fußgängerdynamik – Empirische Untersuchung von Experimenten eindimensionaler Bewegung sowie quantitative Beschreibung von Stau-Charakteristika
V. Ziemer (2020), XVIII, 155 pp
ISBN: 978-3-95806-470-6
URN: urn:nbn:de:0001-2020051000

Band / Volume 43
Algorithms for massively parallel generic *hp*-adaptive finite element methods
M. Fehling (2020), vii, 78 pp
ISBN: 978-3-95806-486-7
URN: urn:nbn:de:0001-2020071402

Band / Volume 44
The method of fundamental solutions for computing interior transmission eigenvalues
L. Pieronek (2020), 115 pp
ISBN: 978-3-95806-504-8

Band / Volume 45
Supercomputer simulations of transmon quantum computers
D. Willsch (2020), IX, 237 pp
ISBN: 978-3-95806-505-5

Band / Volume 46
The Influence of Individual Characteristics on Crowd Dynamics
P. Geoerg (2021), xiv, 212 pp
ISBN: 978-3-95806-561-1

Schriften des Forschungszentrums Jülich
IAS Series

Band / Volume 47
Structural plasticity as a connectivity generation and optimization algorithm in neural networks
S. Diaz Pier (2021), 167 pp
ISBN: 978-3-95806-577-2

Band / Volume 48
Porting applications to a Modular Supercomputer
Experiences from the DEEP-EST project
A. Kreuzer, E. Suarez, N. Eicker, Th. Lippert (Eds.) (2021), 209 pp
ISBN: 978-3-95806-590-1

Band / Volume 49
Operational Navigation of Agents and Self-organization Phenomena in Velocity-based Models for Pedestrian Dynamics
Q. Xu (2022), xii, 112 pp
ISBN: 978-3-95806-620-5

Band / Volume 50
Utilizing Inertial Sensors as an Extension of a Camera Tracking System for Gathering Movement Data in Dense Crowds
J. Schumann (2022), xii, 155 pp
ISBN: 978-3-95806-624-3

Band / Volume 51
Final report of the DeepRain project
Abschlußbericht des DeepRain Projektes
(2022), ca. 70 pp
ISBN: 978-3-95806-675-5

Weitere ***Schriften des Verlags im Forschungszentrum Jülich*** unter
http://wwwzb1.fz-juelich.de/verlagextern1/index.asp

IAS Series
Band / Volume 51
ISBN 978-3-95806-675-5

Mitglied der Helmholtz-Gemeinschaft

JÜLICH
Forschungszentrum