










A new distributed data analysis framework for better scientific collaborations

 **Philipp S. Sommer**⁴,  **Dr. Viktoria Wichert**⁴,  **Daniel Eggert**³,  **Dr. Tilman Dinter**¹,  **Klaus Getzlaff**²,  **Andreas Lehmann**²,  **Christian Werner**⁶, **Brenner Silva**¹, **Lennart Schmidt**⁵, and  **Angela Schäfer**¹

1 Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung (AWI), Germany **2** GEOMAR Helmholtz Centre for Ocean Research Kiel, Germany **3** German Research Centre for Geosciences (GFZ), Germany **4** Helmholtz Zentrum Geesthacht (HZG), Institute for Coastal Research, Germany **5** Helmholtz-Zentrum für Umweltforschung GmbH - UFZ, Germany **6** Karlsruhe Institute of Technology, Institute of Meteorology and Climate Research - Atmospheric Environmental Research (IMK-IFU), Germany

DOI: [10.5072/zenodo.733344](https://doi.org/10.5072/zenodo.733344)

Software

- [SORSE](#) 
- [Event Website](#) 

Category: talks

Published: February 19, 2021

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC-BY](#)).

A common challenge for projects with multiple involved research institutes is a well-defined and productive collaboration. All parties measure and analyze different aspects, depend on each other, share common methods, and exchange the latest results, findings, and data. Today this exchange is often impeded by a lack of ready access to shared computing and storage resources. In our talk, we present a new and innovative remote procedure call (RPC) framework. We focus on a distributed setup, where project partners do not necessarily work at the same institute, and do not have access to each others resources. We present an application programming interface (API) developed in Python that enables scientists to collaboratively explore and analyze sets of distributed data. It offers the functionality to request remote data through a comfortable interface, and to share and invoke single computational methods or even entire analytical workflows and their results. The prototype enables researchers to make their methods accessible as a backend module running on their own infrastructure. Hence researchers from other institutes may apply the available methods through a lightweight python or Javascript API. In the end, the overhead for both, the backend developer and the remote user, is very low. The effort of implementing the necessary workflow and API usage equalizes the writing of code in a non-distributed setup. Besides that, data do not have to be downloaded locally, the analysis can be executed “close to the data” while using the institutional infrastructure where the eligible data set is stored. With our prototype, we demonstrate distributed data access and analysis workflows across institutional borders to enable effective scientific collaboration. This framework has been developed in a joint effort of the DataHub and Digital Earth initiatives within the Research Centers of the Helmholtz Association of German Research Centres, HGF.