Scientific Library Services and Information Systems –
e-Research Technologies (implementation)

InfraStructure for dAta-BasEd Learning in environmental sciences (ISABEL)

Achim Streit, Steinbuch Centre for Computing (SCC), Karlsruhe Institute of Technology
Erwin Zehe, Institute of Water and River Basin Management (IWG), Karlsruhe Institute of Technology

1 Starting point

1.1 State of the art and preliminary work

1.1.1 State of the art

Hydrological research and modelling require enormous data e.g. for characterization of initial and boundary conditions, system states, geometries and characteristics and for model calibration and validation (Gupta et al. 1998; Beven and Freer 2001). Due to the heterogeneity of hydrological landscapes, such observations are incomplete and non-exhaustive (Blöschl and Sivapalan 1995). This in turn implies that hydrological analysis and predictions are inherently uncertain (Kavetski et al., 2006). The amount and diversity of environmental data is steadily growing due to the advent of remote sensing techniques, new observation methods and automated monitoring networks, and the zoo of methods for data-based learning is more diverse than ever. All these recent developments open up exciting opportunities for environmental data science, model improvement and uncertainty reduction, and ultimately even for new discoveries.

The realization of this vision requires, however, a new generation of virtual research environments (VREs), which merge traditional services (storage, management, visualization) with a suite of advanced algorithms for rapid preprocessing, high-level data analysis and knowledge extraction, within the same software architecture. Useful VREs should allow for a flexible updating of the toolbox, to tap the collective creativity of the community, as well as a straightforward inclusion of emerging data and related updating of the metadata catalogue. They also should fuse data from academia and environmental agencies, because only the latter provide consistent long-term monitoring, which is the key to assess global environmental change.

The request to share data according to the FAIR principles has become an essential part of the policy of funding agencies and scientific societies (American and European Geoscience Unions, AGU, EGU) to foster reproducible science and publishing. Data sharing requires common standards for geo-(spatial) data, as e.g. defined by the Open Geospatial Consortium (OGC, founded in 1994 by the EU), the Infrastructure for Spatial Information in the European Community (INSPIRE) directive 2007/2/EC and the commercial standard from the International Organization for Standardization, ISO19115 Geographic Information-Metadata. While data storage and standardization are important steps, data sharing means to provide those in user-friendly formats via appropriate portals. Key to facilitate their scientific use is to provide generic preprocessing tools, which allow the combination of heterogeneous, multiscale and multi-sensor data and harmonization of their spatial and temporal resolution. This step is, however, by itself scientific when using methods beyond inverse distance interpolation. Assessment of data quality (errors and their propagation) and consistency of causally related data is of similar importance.

Several web portals provide access to hydrological and environmental data, and they differ greatly with respect to their scope and included datasets. Data repositories such as GFZ Data Services[1], PANGAEA[2] archive published datasets for download and make them citable via their persistent object identifier (e.g. DOI/Digital Object Identifier). Project-specific data portals, as e.g. for the TERENO[3] project, provide access to project data and include basic tools for analysis or visualization. Federal and national state offices often provide their own data portals (e.g. LUBW[4], USGS[5], NASA[6]). However, most of these data portals do not allow for proper preprocessing of heterogeneous data to prepare scientific analyses.

At present, only a handful existing systems provide a working environment for hydrological and environmental data. The commercial KISTERS[7] system has been tailored for water resources management practice, which is too narrow for advancing science. In line with this, the Swedish Water Evaluation and Planning System (WEAP[8]) focuses on water resources management and planning, while it includes scenario calculations for predictions. Within the set of open non-profit portals, the CUAHSI HydroShare[9] offers a data management system to access a large range of datasets (mainly from the US) and an environment for collaborative work. However, a toolbox with readily implemented preprocessing tools is missing. Within the interdisciplinary research unit CAOS (FOR 1598, Zehe et al. 2014) we evaluated the CUAHSI system for the project data management, but for this diverse data collection the CUAHSI metadata model was too inflexible.

Finally, the EU-funded project SWITCH-ON[10] identified the need to provide access to a wide range of different data in Europe and to connect those to tools to make their analyses reproducible (cf. Hutton et al. 2016). The SWITCH-ON platforms were built with a strong focus on collaboration within projects. Their so-called virtual laboratories enabled large research consortia to define common standards, discuss and work cooperatively on the same data on a project or experiment (Ceola et al. 2015). However, more than three years after project end and despite its ambitious goals, the portal seemed to be hardly used by the hydrological community. This is probably due to a) the lack of a sufficiently large database of open data and b) a missing toolbox for preprocessing or further analyses of the included datasets.

ISABEL aims to address the demand for VREs to accelerate data-based learning in hydrological sciences by expanding the functionalities of the existing system “V-FOR-WaTer”. Particular emphasis will be put on making the system attractive for PhD students and PostDocs.

1.1.2 Preliminary work

V-FOR-WaTer in a nutshell

V-FOR-WaTer has been developed within the last five years, starting within the E-Science initiative of the Ministry for Science, Research and Arts Baden-Württemberg (MWK), to foster professional management of diverse hydro-meteorological data and offer tools for preprocessing, standard hydrological procedures and some more sophisticated analyses. During a kickoff workshop in 2016 with representatives of universities, the federal environmental state office in Baden-Württemberg (LUBW) and the MWK, we discussed common demands and specific requirements accentuating the need for an open data portal that is appealing for a broad range

[1] https://dataservices.gfz-potsdam.de/

[2] https://www.pangaea.de/

[3] https://ddp.tereno.net/ddp/

[4] https://www.lubw.baden-wuerttemberg.de/

[5] https://waterdata.usgs.gov/nwis

[6] https://disc.gsfc.nasa.gov/

[7] https://water.kisters.de/en/

[8] https://www.weap21.org/

[9] https://www.hydroshare.org/

[10] https://cordis.europa.eu/project/id/603587/reporting

of scientific use cases. This interest group subsequently extended to include members of the Water Research Network Baden-Württemberg[11] and the members of the CAOS research unit FOR 1598. The product of the first development phase is a B-prototype of a web portal, and ISABEL will transform it into an excellent tool to explore the diverse CAOS dataset, as one example. The CAOS community forms, hence, a user base nucleus, cutting across six science fields and several institutions:

- Hydrology: Prof. Weiler, Univ. Freiburg; Dr. Blume, GFZ Potsdam; Prof. Pfister, LIST Luxembourg; Prof. van Schaik, Univ. Wageningen NL; Prof. Schulz, BOKU Austria, Prof. Zehe, KIT;
- Geoecology and Geochemistry: Prof. Schröder, Univ. Braunschweig; Jun.-Prof. Jackisch, Univ. Freiberg; Dr. Christophe Hissler, LIST, Luxembourg;
- Meteorology and Climate Research: Prof. Wulfmeyer, Univ. Hohenheim; PD. Dr. Kleidon, MPI Jena;
- Geophysics and Remote Sensing: Prof. Tronicke, Univ. Potsdam; Dr. Scherf, LIST Luxembourg; Prof. Hinz, KIT.

V-FOR-WaTer is already suited to manage a wide range of hydro-meteorological data on rainfall, wind velocity, air pressure, air temperature and humidity, stream flow and stream temperature, soil moisture, soil temperature and matric potentials, piezometric head and sap flow velocity. The system provides an advanced metadata catalogue[12], a fine-grained user and authentication management, workflows and tools for data visualization and data analysis. The V-FOR-WaTer web portal includes map-based operations, sophisticated data filters, workflows and visualization of time series. Moreover, the Python equivalent of the V-FOR-WaTer toolbox already contains a set of example tools and packages, from simple hydrological signatures to comprehensive variogram analyses.

Within ISABEL we aim to further develop the VRE to i) considerably expand its scientific scope, the toolbox and its user-friendliness, ii) broaden the spectrum of hosted data to include data from state offices, complex data structures and important remote sensing products and iii) provide access to data and tools in a modern, secure, and responsive web portal with GIS functions and drag&drop tools for workflows. We implement and test these functionalities along three use cases (UC, detailed in Section 2): UC 1 covers many generic research aspects of catchment hydrology and related model evaluation. UC 2 addresses improved precipitation estimates using rainfall radar and new data sources, while UC 3 aims at facilitating comparison and harmonization of cross-disciplinary and cross-scale measurements of evapotranspiration (ET).

V-FOR-WaTer is a joint development of the Steinbuch Centre for Computing (SCC) and the Institute for Water and River Basin Management (IWG) at the Karlsruhe Institute of Technology (KIT). The close cooperation with the CAOS research unit, the Water Research Network Baden-Württemberg and the LUBW has guaranteed valuable user feedback and more importantly allowed the successful incorporation of the official discharge dataset of the LUBW. To meet standards for data publication we are in close exchange with the GFZ Data Services repository, both for accessing their published data and for enabling publication of data with DOI from V-FOR-WaTer in their repository. We started to cooperate with Digital Earth and the TERENO community to develop the BRIDGET package. Our joint efforts to achieve a sustainable research data management have gained even more inertia since KIT regained the status of a DFG Excellence University (EXU). Within the ongoing EXU project SmaRD-AI, we collaborate with players of

[11] https://www.wassernetzwerk-bw.de/english/index.php

[12] https://github.com/VForWaTer/metacatalog

Digital Earth and NFID4Earth to establish research data management with V-FOR-WaTer as strategic backbone and enabler for data science at the KIT Climate and Environment Center.

Technical details of V-FOR-WaTer

The design of the web portal follows well-known and intuitive Geographic Information Systems (GIS), like ArcGIS or QGIS. Special emphasis was on a modular design, which can easily be extended. V-FOR-WaTer provides a collection of tools that are based on Web Processing Services (WPS), a standard of the Open Geospatial Consortium (OGC)[13] for web based applications. Using WPS ensures that the portal can be easily expanded with new tools and also enables external access to these tools. The V-FOR-WaTer web portal is composed of open source projects (Fig. 1) and is itself open source. The web portal is coded in the secure and scalable Python web framework django, which is well documented and actively supported by a large community. Interaction with the map is handled with the JavaScript library OpenLayers. Authentication is accomplished by B2ACCESS, the authentication and authorization system of EUDAT[14]. We are working on a model builder to facilitate combination of WPS tools (WPS*flow*), where data and tools can be connected with simple drag and drop tasks. While V-FOR-WaTer is intended to provide easy and open access to data and tools, in some cases restricted data access is necessary, either because the data contain sensitive information or because the data owner places a maximum 2-year embargo period to finish data analyses and publication. Consequently, access to the WPS tools and data is not directly provided, but requests are verified and redirected in django instead. The first datasets included in the portal originate from the CAOS project, the LUBW and the IWG. These datasets are time series with a point geometry, stored locally in a PostgreSQL database. Extension of the database to multidimensional data is work in progress. Access to data and metadata happens through django and Geoserver. The latter is used to visualize the position of datasets on the map via WFS[15]. A first instance is running and used for demonstrations[16].

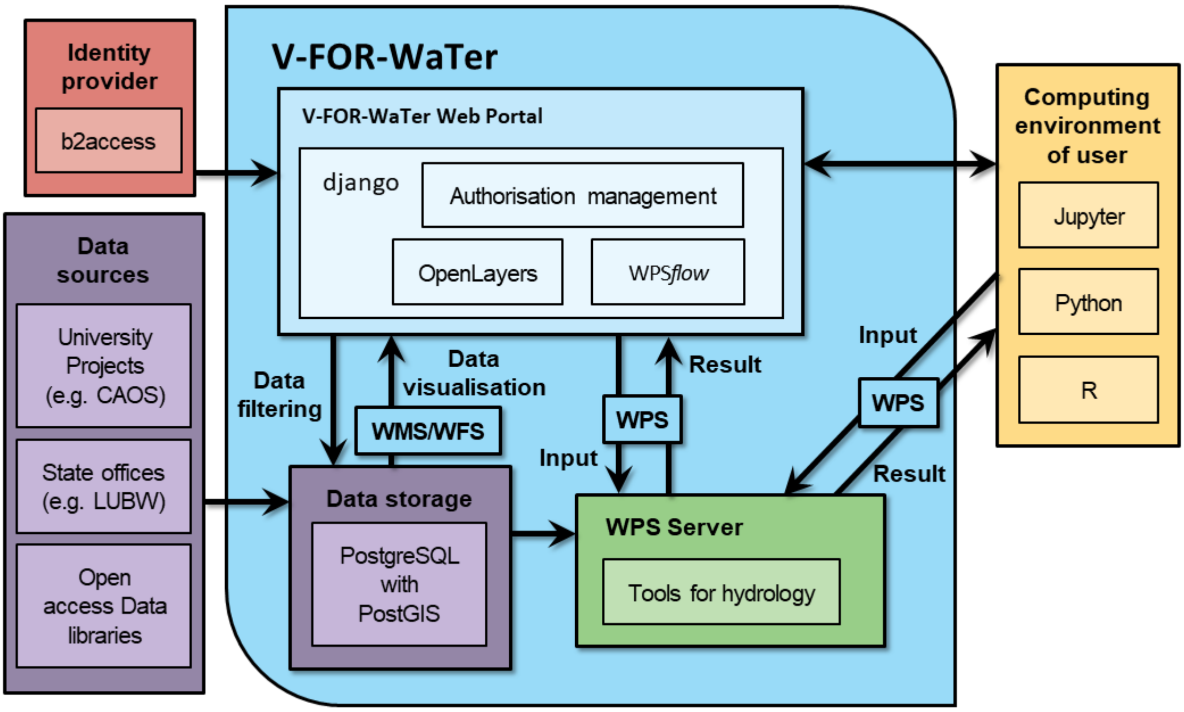

Figure 1: Architecture of the V-FOR-WaTer portal.

[13] https://www.opengeospatial.org/standards/wps
[14] https://www.eudat.eu/services/b2access
[15] https://www.opengeospatial.org/standards/wfs
[16] https://portal.vforwater.de

Environmental science meets computer science

The project consortium has the right mix of competences and developed a common language to address the aforementioned challenges. The group of Erwin Zehe (IWG) has strong expertise in the fields of hydrological modelling (Zehe et al. 2005) and data-based learning using geostatistical methods (Zehe et al., 2010, Mälicke et al., 2020). This includes (generalized) linear models (Hassler et al, 2018; Graeff et al. 2012) and machine learning methods (self- organizing maps, clustering) for grouping of model errors (Reusser et al. 2009), soil moisture patterns (Mälicke et al., 2020) or redundant model components (Ehret et al. 2020). A related research focus is on metrics for uncertainty assessment and model errors (Series Distance, Ehret and Zehe 2011; Seibert et al., 2016), information-based similarity indices (Loritz et al. 2018; Mälicke et al., 2020) and model diagnostics (TIGRE Reusser et al, 2009; FAST Reusser and Zehe, 2011). Erwin Zehe was the scientific speaker of the DFG research unit CAOS FOR 1598. CAOS explored new ways to collect data on functioning and organization of hydrological landscapes, for their synoptic analysis and the related development of process-adequate models; and it was in fact this “CAOS data experience”, which motivated the V-FOR-WaTer research environment. Sibylle Haßler coordinated the preparation of the huge and diverse CAOS dataset for the CUAHSI database system and has been consequently coordinating and promoting V-FOR-WaTer as well. She is also an editor of the data journal Earth System Science Data ESSD. Within the Helmholtz Initiative Digital Earth she conducts the Bridging PostDoc project BRIDGET to facilitate comparison and scaling of ET data. The IWG has been cooperating since 2000 with the LUBW in order to regionalize stream flow characteristics within the interactive LUBW service UDO (Umwelt-Daten und -Karten Online[17]/ Environmental Data and Maps Online).

The group of Achim Streit (The Steinbuch Centre for Computing, SCC) has ample experience in developing and providing data services. The close interaction of research and service provision ensures that new scientific findings are quickly incorporated into the design and development of IT services and infrastructure. V-FOR-WaTer cooperates with the Simulation and Data Lifecycle Lab (SDL) “Earth System Science” at SCC, a sub-topic of the Helmholtz Program “Engineering Digital Futures”. SDL enables joint research and development between data scientists and environmental researchers. Today, SCC is member of the European EUDAT Common Data Infrastructure (CDI) and many European projects including the building of the European Open Science Cloud[18] as well as national projects such as NFDI[19]. As part of the worldwide LHC Computing Grid (WLCG[20]) SCC operates the national Tier-1 center GridKa for the high-energy physics experiments at the Large Hadron Collider at CERN. In the Helmholtz Association, SCC is partner of HelmholtzAI[21], the Helmholtz Federated IT Services (HIFIS[22]), the Helmholtz Metadata Collaboration (HMC[23]) and Exascale Earth System Modeling (ExaESM[24]). SCC coordinates the Helmholtz Data Federation (HDF[25]). In addition, SCC provides services for users in the state of Baden-Württemberg like bwSync&Share and bwCloud and is a partner of the MWK projects Science Data Center MoMaF[26]. Within SCC the activities of most EU-Projects are centralized in the department Data Analytics, Access and Applications (D3A), headed by Jörg Meyer. Due to his research background in geoscience, Marcus Strobl is experienced in data

[17] https://udo.lubw.baden-wuerttemberg.de/public/

[18] https://eoscsecretariat.eu, https://eosc-pillar.eu, https://eosc-synergy.eu, https://www.dice-eosc.eu/, EOSC FUTURE

[19] https://www.nfdi.de/en-gb

[20] https://wlcg.web.cern.ch

[21] https://www.helmholtz.ai/

[22] https://www.hifis.net

[23] https://helmholtz-metadaten.de/en

[24] https://www.exaesm.de/

[25] https://www.helmholtz.de/forschung/information-data-science/helmholtz-data-federation-hdf/

[26] https://momaf.scc.kit.edu/

analyses with GIS systems and the development and implementation of domain-specific software tools.

1.2 Project-related publications

1.2.1 Articles published by outlets with scientific quality assurance, book publications

Ehret, U., R. van Pruijssen, M. Bortoli, R. Loritz, E. Azmi, and E. Zehe (2020). “Adaptive clustering: reducing the computational costs of distributed (hydrological) modelling by exploiting time-variable similarity among model elements”. Hydrol. Earth Syst. Sci., 24(9), 4389-4411.

Graeff, T., E. Zehe, T. Blume, T. Francke, and B. Schröder (2012). “Predicting event response in a nested catchment with generalized linear models and a distributed watershed model”. Hydrol. Process., 26(24), 3749–3769.

Loritz, R., H. Gupta, C. Jackisch, M. Westhoff, A. Kleidon, U. Ehret, and E. Zehe (2018). “On the dynamic nature of hydrological similarity”. Hydrol. Earth Syst. Sci., 22(7), 3663–3684.

Mälicke, M., S. K. Hassler, T. Blume, M. Weiler, and E. Zehe (2020). “Soil moisture: variable in space but redundant in time”. Hydrol. Earth Syst. Sci., 24(5), 2633–2653.

Reusser, D., T. Blume, B. Schaefli, and E. Zehe (2009). “Analysing the temporal dynamics of model performance for hydrological models”. Hydrol. Earth Syst. Sci., 13(7), 953–1297. (TIGRE Package)

Reusser, D. E., and E. Zehe (2011). “Inferring model structural deficits by analyzing temporal dynamics of model performance and parameter sensitivity”. Water Resour. Res., 47(7), W07550. (FAST Package)

Seibert, S., U. Ehret, and E. Zehe (2016). “Disentangling timing and amplitude errors in streamflow simulations”. Hydrol. Earth Syst. Sci., 20(9), 3745–3763. (Series Distance Package)

Streit, A., J. van Wezel (2021). "1.2 Deutschland in der European Open Science Cloud". Praxishandbuch Forschungsdatenmanagement, edited by Markus Putnings, Heike Neuroth and Janna Neumann, Berlin, Boston: De Gruyter Saur, 31–52.

1.2.2 Other publications, both peer-reviewed and non-peer-reviewed

Hassler, S. K., P. Dietrich, R. Kiese, M. Mälicke, M. Mauder, J. Meyer, C. Rebmann, M. Strobl, and E. Zehe (2021). Integration of evapotranspiration estimates from scaled sap flow values and eddy covariance measurements in the BRIDGET toolbox. EGU General Assembly 2021, 19–30 Apr 2021, EGU21-3889.

Strobl, M., E. Azmi, S. K. Hassler, M. Mälicke, J. Meyer, and E. Zehe (2021). V-FOR-WaTer: A Virtual Research Environment for Environmental Research, EGU General Assembly 2021, 19–30 Apr 2021, EGU21-3356.

2 Objectives and work program

2.1 Anticipated total duration of the project

We request funding for three years.

2.2 Objectives

The objective of ISABEL is to expand the V-FOR-WaTer prototype to provide high-end data storage and management facilities in combination with a comprehensive toolbox for rapid and reproducible research. We will implement new datasets, according to FAIR data principles, and various tools to address three scientific use cases. The tools comprise standard preprocessing and scaling methods (e.g. inverse distance and spline interpolations), tools for more advanced scientific analyses (e.g. multivariate statistics, geostatistics, machine learning), existing user-developed packages (e.g. TIGRE, FAST, Series Distance, BRIDGET) and new tools for assessing data quality and uncertainty propagation. Technically, this requires an expansion of the metadata model, connecting to already existing databases via their API and implement functionalities to prepare the consolidation phase. We will adhere to the established international standards for interoperability of our system with initiatives such as NFDI4Earth, the Helmholtz DataHubs and ENVRIplus.

We will implement three important, generic and connected use cases ensure that V-FOR-WaTer covers a wide range of hydrological research and practice applications. These use cases require integration of a variety of functions, data types and several user-developed packages as described in the following:

- **UC 1: Catchment hydrology and model evaluation**. Catchment hydrology developed largely as an engineering discipline around traditional tasks of designing and operating reservoirs, flood risk assessment and water resources management. Hence we will implement many generic procedures to characterize the catchment water balance, runoff behavior and related evaluation of hydrological models. Challenges arise for example (i) in the automated and reproducible detection of rainfall-runoff events, and (ii) in the multivariate consistency check among rainfall, streamflow, soil moisture and groundwater storage data when attempting to close the water balance. (iii) A related key generic problem is the regionalization of rain-gauge data, soil moisture data or ET data to catchment-scale estimates. While inverse distance interpolation is the technical minimum standard, geostatistical interpolation or simulation methods (and the use of rainfall radar, see UC 2) are scientifically more appealing. (iv) Finally, UC 1 will allow evaluation of hydrological model simulations, because the quest for the appropriate model structure and the evaluation of the underlying uncertainty remain key research issues in hydrology. We want to offer several automated tools and user-developed packages (for both model diagnostics and uncertainty assessment) and support these procedures by creating a new data type for model outputs in V-FOR-WaTer. Besides research datasets from the CAOS community, we will provide operational data of the LUBW (and reach out to other state offices), to create a suitable data pool for analyses and for users who would like to test and familiarize themselves with the tools before applying them to their own datasets.
- **UC 2: (Near real-time) distributed precipitation estimates.** While precipitation is the most important source of fresh water, the retrieval of reliable highly resolved spatially distributed rainfall is still a major challenge in hydrology and meteorology. We will address this challenge by developing tools for scientific handling of rainfall radar data (e.g. from the operational DWD network) and combining them with standard measurements such as rain-gauge data. Radar-based characterization of the space-time pattern of rainfall requires a translation of the observed radar reflectivity Z into a rainfall rate R by means of a Z-R-relation (Heistermann et al. 2008), and a merging of the radar image with the more precise ground-based observations. The latter can be achieved by means of geostatistical interpolations using rain gauge data (Sinclair and Pegram 2005). Even more appropriate is to use disdrometer data (Neuper and Ehret, 2019), because they directly measure the drop size distribution which largely determines the Z-R-relation (Neuper and Ehret 2019). V-FOR-WaTer will provide an interface to the most recent DWD RADKLIM[27] product, and it will also permit the integration of radar data, to facilitate research on the improvement of merging procedures. This implies the challenge to store and rapidly process big 4D remote sensing datasets and provides a playground for a widely applicable geostatistical interpolations and conditional simulations (Berndt et al., 2014; Zehe et. al, 2005) and machine learning. Additionally we will evaluate and incorporate new data products such as rainfall estimates from the attenuation of commercial microwave links (e.g. using first open datasets[28] and communicating with mobile phone providers). We will include an approach, which, after calibration, contains rainfall information along the transmission line (Leijnse et al., 2007). Given the high density of microwave links, these data are particularly valuable in the otherwise underrepresented small catchments and developing countries.

[27] https://www.dwd.de/DE/leistungen/radarklimatologie/radarklimatologie.html

[28] https://zenodo.org/record/4923125#.YO7paExCSUk

- **UC 3: Multi-sensor and multiscale evapotranspiration data and formulas**. Here we will cooperate with the BRIDGET project, which is part of the Helmholtz initiative Digital Earth and aims at developing a toolbox to facilitate comparison and scaling of various evapotranspiration (ET) estimates; ET, the sum of evaporation and plant transpiration (T), amounts in many catchments to about 60% of the terrestrial precipitation (Oki and Kanae 2006). As water vapor fluxes are difficult to observe directly, a variety of methods exist to estimate ET, e.g. from eddy covariance measurements, xylem sap flow (T) or the residual water balance of lysimeters. While these methods are prone to specific uncertainties, have different footprints and require a specific post-processing, this information is rarely shared across the related disciplines (meteorology, hydrology, plant physiology). There is a large amount of available ET formulas based on meteorological variables (e.g. temperature-based, radiation-based, energy-balance methods), with different complexity, which even increase this almost Babylonian confusion about the probably most important hydrological flux. To address this challenge, we will incorporate the BRIDGET package, a Python package that will support handling of these diverse data, into V-FOR-WaTer and expand it by the most common ET estimation formulas to enable comparisons.

The key to keep V-FOR-WaTer alive and growing after the developments within ISABEL is to support accelerated research at the PhD student and PostDoc level. When these science individuals experience the functionalities of data access (and publication), the range of tools within the toolbox and intuitive handling of the portal as satisfactory and beneficial to their research, they will continue to use the VRE and contribute their own tools. We will reach out to this user group and include them in the developments within ISABEL by organizing hands-on workshops, in cooperation with EGU's young hydrological society.

The joint development between hydrologists and computer scientists within ISABEL is structured along five interrelated work packages: Project management (WP 1), Data management and data quality (WP 2), Web portal implementation (WP 3), Scientific use cases and tools (WP 4), Outreach and dissemination (WP 5).

2.3 Work program and proposed research methods

While the implementation of the computing environment and the web portal will be addressed by SCC, IWG will focus on state-of-the-art tools for the use cases, the exchange with the hydrological community and providers of repositories to gather data and subsequently adapt the metadata model of the portal. The system release to the community will occur within three stages, which are also the anticipated project milestones (**M**):

- **M1: Alpha Release.** Users will be able to find and access open source data, their own data and request restricted data from other users in the portal. Data will be available for download and scientific analysis with the tools for UC 1 and UC 3, related results can be further analyzed, visualized or downloaded. The portal will be usable, yet features will be missing, and the stability and user experience will likely need improvement.
- **M2**: **Beta Release.** At the end of this stage, after two years, we expect most features to be implemented, major bugs, identified by our alpha testers, will be removed.
- **M3: Release to Web and start of regular operations.** Bugs identified by our beta testers will have been removed with the help of the monitoring of user behavior. The portal will be ready for community testing. Interfaces to simplify data exchange with data portals will be working, users can upload their own data by themselves and basic user support will be available. Monitoring of portal services is running to ensure high uptime Services of V-FOR-WaTer are accessible via an API and Jupyter. Documentation is satisfying and

supportive. The portal will be integrated in the KIT Climate and Environment Center to foster research data management.

Figure 2 summarizes the time plan, milestones **(M)**, work packages **(WP)** and tasks **(T)**.

2.3.1 WP 1: Project management (Lead IWG, 12 PM)

T 1.1 Project management (6 PM): The project management ensures a close cooperation between IWG and SCC, sustains the communication with external partners (e.g. GFZ Potsdam, LUBW), the wider community, and evaluates and establishes new contacts (e.g. PANGAEA), and reaches out to national and federal state offices (German Weather Service (DWD), Federal Institute of Hydrology (BFG), Bavarian Federal Environmental Office (LfU)).

T 1.2 Terms of Use (3 PM): Develops a concept for handling of user data according to the General Data Protection Regulation including a register of processing operations; prepares and updates Terms of Use and conditions according to the versions Alpha, Beta, Release to Web.

T 1.3 Operations model (3 PM): Here we design an operations model, including possibilities for long-term funding and possibilities to charge commercial projects for resources and support for the projects data management plans.

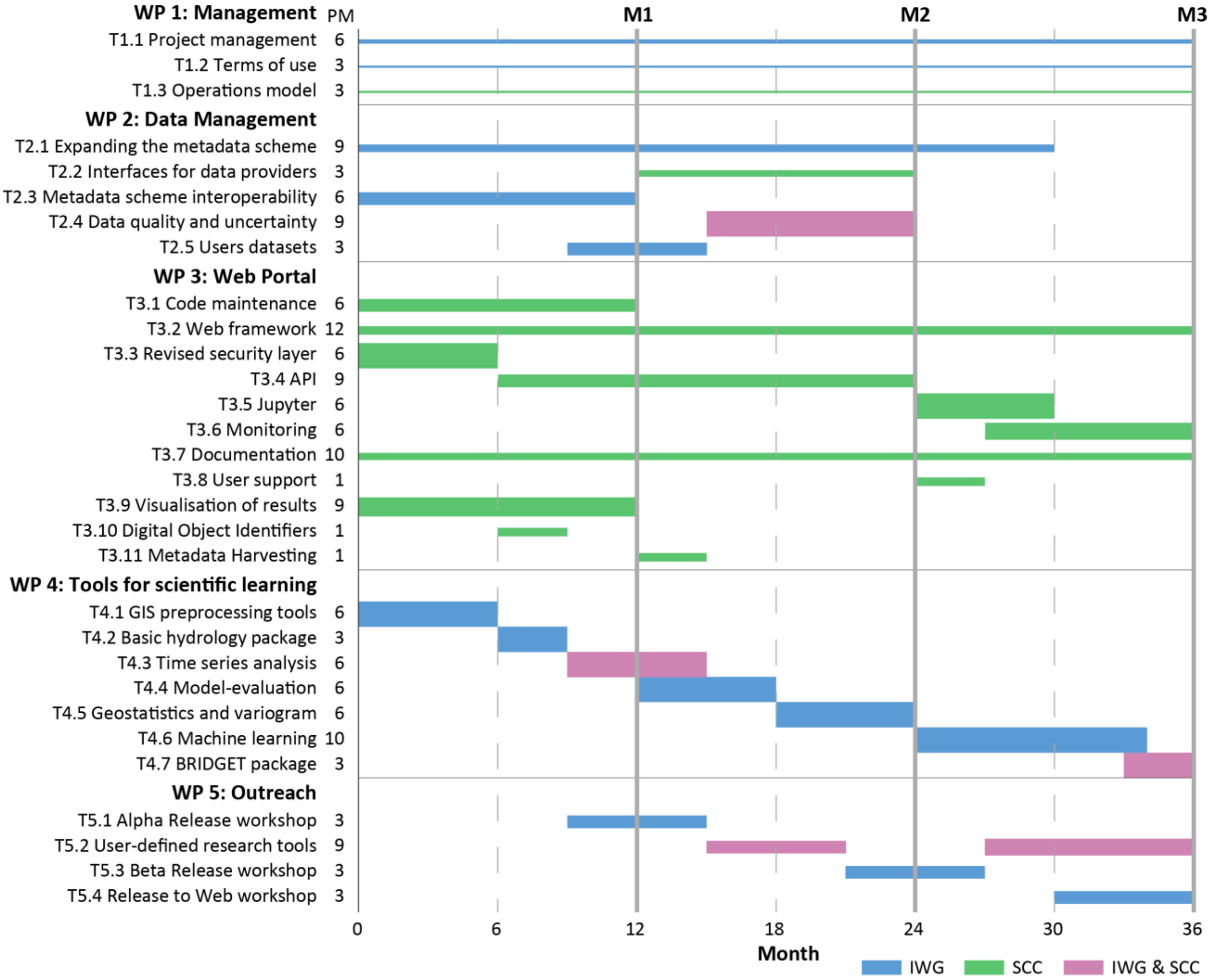

Figure 2: Work schedule, anticipated milestones **(M)**, work package **(WP)** and tasks **(T)**: The area enclosed in the colored bars represents the actual workload (PM) of a task.

2.3.2 WP 2: Data Management and data quality (Lead IWG, 30 PM)

Here we will organize the collection of data, implementation of standards for storage and easy data exchange, and a basic quality control especially for the data provided from users. Generally, we will put major emphasis on data quality. This includes quantification of observation errors and the related error propagation, as well as hydrological consistency of datasets and if possible the use of Monte Carlo Methods to quantify uncertainty in up-scaled/regionalized products.

T 2.1 Expanding the metadata scheme (9 PM): We will broaden the existing metadata scheme to support hosting of 2D time-invariant data (spatial maps), complex 3D and 4D data and virtual data types from model simulations:

- Precipitation estimates from rainfall radar and microwave links (UC 2): Rainfall radar measurements yield 4D data cubes, which need proper referencing to a rain gauge on the ground e.g. to explore the relationship between radar reflectance and rainfall rate, and to merge both data sources. We will expand our metadata scheme to accommodate these data, in in cooperation with the DWD. Commercial microwave links within mobile phone networks open up new avenues for distributed rainfall estimates, especially in regions with poor coverage of meteorological measurement stations. These estimates and data products are neither standardized nor available from open databases. Here we will evaluate how to include this information alongside other precipitation data, necessary metadata and data schemes, in exchange with data providers and scientists. We will include some of the first openly accessible datasets as blueprints for future expansion.
- Metadata for BRIDGET (UC 3): The BRIDGET evapotranspiration (ET) package aims to synthesize and compare various ET data from different sensors, measurement methods, and with varying spatial reference and footprints. We plan to expand the data spectrum to remotely sensed surface temperature data from drones and satellites, as this broadens the spectrum of approaches to derive ET (e.g. Brenner, et al. 2018), which requires appropriate adaptation of the metadata and data model. BRIDGET will also include uncertainty estimates that will be propagated through the various methods, including ET estimates from formulas.
- Virtual data from simulations (UC 1): We will include model outputs as a new data type to support model comparisons. The starting point is discharge simulations. We will develop a metadata accounting for the model specification (HBV, LARSIM, CATFLOW), the processer, input data, numerical schemes etc., and extend this to soil moisture and simulated ET from water balance models.

T 2.2 Interfaces for data providers (3 PM): The collaboration with GFZ Data Services was a good starting point to establish interoperability between various data providers. In order to connect to more data sources, appropriate interfaces for the import and export of datasets will be defined and implemented. To this end, we will intensify our collaboration with the repository PANGAEA and LUBW and start new co-operations with federal and national offices, which provide hydro-meteorological data particularly the Bavarian Federal Environmental Office[29] and the Bundesanstalt für Gewässerkunde BfG[30] (UC 1), as well as the German Weather Service (RADKLIM, UC 2). We also follow the new “Environmental Data Retrieval API Standards Working Group” of OGC to incorporate their outcome.

T 2.3 Extend metadata scheme for interoperability (6 PM): Although the current metadata scheme was developed using the diverse CAOS dataset, it needed to be expanded when adding datasets from the LUBW. The current state is a flexible ISO19115-compatible version, which also allows the exchange of datasets with GFZ Data Services. To increase the interoperability,

[29] https://www.lfu.bayern.de/index.htm

[30] https://www.bafg.de/GRDC/

controlled vocabularies and commonly used keyword lists, we will consider NASA GCMD Science Keywords[31], the GEMET Thesaurus[32], INSPIRE keywords[33] or GeoSciML[34] for implementation.

T 2.4 Data quality and uncertainty (9 PM): Generally, we will put major emphasis on data quality. Here we will combine several approaches and measures, that will developed in close cooperation with users:

- Estimation of data quality by data providers according to an expert classification scheme,
- Estimation of measurement errors and their propagation (Gaußian error propagation, Monte Carlo methods),
- Estimation of statistical uncertainties in regionalized data products (conditional geostatistical simulations) for Monte Carlo studies,
- Rating system for users (e.g. stars, comments),
- Tools for automated plausibility checks and for consistency of different data (no runoff without rainfall or snow melt, soil moisture should remain within meaning full ranges of porosity and residual soil water content, runoff smaller than rainfall input).
- We will also evaluate the suitability of other quality control approaches developed in related projects, for example within Digital Earth.

T 2.5 Datasets provided by users (3 PM): Users might naturally combine the data provided from V-FOR-WaTer with their own data. The required data upload into V-FOR-WaTer portal, needs the metadata scheme to be flexible enough to deal with emerging new data types, to perform automatic checks to avoid mistakes in the upload process and security check mechanisms to detect and prevent attacks during the upload. Data upload shall be straightforward to encourage users to work with V-FOR-WaTer, so a minimal set of metadata should be required when including own datasets. To win users for a comprehensive description of their data with metadata, we will create a reward system grating for example be a quick publication of the datasets in a connected repository and a highlighting of the best-documented datasets. These are likely to be used (and cited) more frequently because they enable better understanding of the data. For the front end of the metadata upload, we will consider the tool developed at GFZ[35].

2.3.3 WP 3: Web portal (Lead SCC, 67 PM)

This work package aims at the further implementation of the portal software to provide a system that is ready for operation. This requires several tasks on front and back end.

T 3.1 Code maintenance (6 PM): While the existing code has been designed with a modular and scalable structure it was implemented in a bottom-up approach. Since then, new features have been added and parts of the code have been re-implemented. Here we will revise the existing code structure to identify and re-implement weak designs, overall to improve the maintainability, extendibility, and security in a top-down fashion. The documentation of the code will be a continuous task. Additionally, we encourage external developers to contribute to this open-source project.

T 3.2 Web framework (12 PM): The web front end is subject to continuous changes. The look and feel for users is essential for the acceptance of the portal. The user friendliness depends on an easy and intuitive way to navigate through the portal. The use of modern JavaScript libraries supports this but requires adaptations and enhancements of the existing front end. Furthermore,

[31] https://earthdata.nasa.gov/earth-observation-data/find-data/gcmd/gcmd-keywords

[32] https://www.eionet.europa.eu/gemet/en/themes/

[33] https://inspire.ec.europa.eu/glossary/MetadataElement-Keyword

[34] http://www.geosciml.org

[35] https://github.com/ulbricht/pmdmeta

inclusion of new datasets implies the integration of the specific features and tools into both front and back end.

T 3.3 Revised security layer (6 PM): A well-designed security architecture is crucial to restrict access to data to authorized users. The authentication in the existing V-FOR-WaTer portal prototype is currently realized with B2ACCESS, while authorization is done locally with user permissions stored in our database. We will simplify our authorization module to increase maintainability. In addition, we will introduce a token mechanism to allow direct access to resources (like the WPS server, see T3.4) in addition to the indirect access via the portal.

T 3.4 Application Programming Interface (API) (9 PM): There are many ways users like to analyse data. We will develop an API that allows users to process data from Jupyter Notebooks or their local computer using or to execute WPS. This requires tokens introduced in T3.3. For this purpose, we will evaluate the APIs of comparable systems, e.g. from HydroShare, to identify core functionalities for the V-FOR-WaTer API and implement an alpha version thereof. Simple R- and Python example scripts using the API will be provided.

T 3.5 Jupyter (6 PM): Jupyter[36] Notebooks became very popular among scientists to analyze scientific data with the programming languages R, Python and Octave, as they support interactive data science and scientific computing. This task aims at providing Jupyter Notebooks for authorized users with access to the data stored in V-FOR-WaTer. In this way, users can run their individual data analysis online without the need to download the data to their local machines. The main challenge is to integrate Jupyter Notebooks in the V-FOR-WaTer authorization infrastructure and to manage and provide the computing resources.

T 3.6 Monitoring (6 PM): In this task we will set up a monitoring system for the operations of V-FOR-WaTer services using existing monitoring and notification tools at SCC (e.g. Icinga[37], Telegraf[38]). The monitoring of specific V-FOR-WaTer services might need the implementation of appropriate probes. In order to ensure a high uptime of the portal services, a proper monitoring of the availability of services and an automated notification of problems is required. Another reason for monitoring the usage of the portal is to study user demands and to identify computer resources bottlenecks. We will only collect anonymous statistics, i.e., we will not track individual users (as will be guaranteed in our Terms of Use).

T 3.7 Documentation (10 PM): Acceptance and use of a software product depends crucially on the quality of its documentation, which ensures to get started smoothly. Here we will ensure a complete, up-to-date and comprehensible user guide of V-FOR-WaTer The text documentation will be supplemented by user tutorials and/or short videos demonstrating typical analysis steps using the virtual research environment.

T 3.8 User support (1 PM): In case of questions or problems, users need a straightforward way to communicate with the operators or developers of the portal. As first step, we will set up a generic mailing list and organize a support team. We will consider the usage and integration of a ticket system. The actual user support is beyond this task, as it will start in the operations phase.

T 3.9 Visualization of results (9 PM): Evaluation of plots, maps or tables is a common way to evaluate results of a scientific analysis. To support this, we integrate widely-adopted tools for the visualization of results (Matplotlib[39], Bokeh[40]), and provide downloads of the styled images in common data formats. This allows for zooming and rescaling of plots, changes in color bars or

[36] https://jupyter.org/

[37] https://icinga.com/

[38] https://www.influxdata.com/time-series-platform/telegraf/

[39] https://matplotlib.org/

[40] https://bokeh.org

hiding of certain elements. In this way, we simplify and facilitate the preparation of results computed on the V-FOR-WaTer web portal for publication.

T 3.10 Digital Object Identifiers (1 PM): In the production phase we want to offer users the possibility to assign DOIs to their datasets stored in V-FOR-WaTer. In this task, we prepare the organizational and technical steps. KIT is a member of DataCite.

T 3.11 Metadata Harvesting (1 PM): Given our support of the metadata standard ISO19115 (see T2.3) we will setup a OAI-PMH service that allows automated harvesting of our metadata or to connect to external tools like B2FIND.

2.3.4 WP 4: Scientific use cases and tools (Lead IWG, 40 PM)

In this work package we will design and implement various tools that will make the VRE a useful exploration and analysis instrument for a wide user group from hydrology and environmental sciences. We structure the developments in this work package along the use cases where the tools will be used primarily.

UC 1: Catchment hydrology and model evaluation

This use case requires basic preprocessing steps such as GIS tools, a wide set of generic hydrological analysis methods, comprehensive time series analysis tools and model evaluation procedures.

T 4.1: The Geographic Information System (GIS) preprocessing tools (6 PM) will allow for elementary GIS functionalities such as delineation of water shed boundaries, flow directions and accumulation and river networks in the web portal. Those will be taken from existing open source libraries and GIS (GDAL, QGIS) and adapted for a user-friendly and trouble-free use in the portal.

T 4.2: The basic hydrology package (3 PM) will provide several methods allowing for i) an automated detection and characterization of rainfall runoff events (peak, runoff coefficient, timing, base flow separation), ii) estimation of different measures of antecedent catchment wetness (antecedent precipitation, dynamic storage), iii) to calculate functional fingerprints of the water balance (discharge regimes, flow duration curves, double mass curves) and iv) an analysis of floods (potting positions, fitting of probability density functions).

T 4.3: The time series analysis and metrics package (6 PM) will provide established methods of time series analysis such as trend detection, autoregressive models, Fourier analysis and possibly wavelet analysis. We will integrate existing tools and functionalities, e.g. provided by Python, R and Octave.

T 4.4: Model evaluation, diagnostics and uncertainty package (6 PM) will provide established goodness-of-fit criteria (Nash-Sutcliffe-efficiency, Kling-Gupta-efficiency) and our own Series Distance package, which allows disentangling errors in timing and magnitude (Seibert at al., 2016). We will also support model uncertainty assessment by providing the opportunity to create uncertain precipitation input as different realizations of geostatistically simulated rainfall fields for ensemble modelling in UC 2 (Cloke and Pappenberger, 2008; Zehe et al., 2005). As the system will allow hosting of the related simulated streamflow or soil moisture ensembles, one can quantify their uncertainty either using statistics, methods from information theory such as the Shannon entropy or mutual information content (Loritz et al. 2018). To give an entry point to information theory, we will provide a tutorial using a simulation ensemble by Loritz et al. (2018). For model diagnostics, we will also implement two widely used R packages. The first allows clustering of recurrent model errors in typical groups by means of self-organizing maps (TIGRE Reusser et al., 2009), while the second quantifies temporal variability of parameter sensitivity to stream flow simulations by means of the Fast Fourier Sensitivity Test (FAST Reusser and Zehe, 2011).

UC 2: (Near real-time) distributed precipitation estimates

This use case requires tools for the handling of rainfall radar and CML data and associated data products, processing, regionalization of point information to spatially distributed estimates and the combination of various methods. While inverse distance interpolation is the technical minimum standard, including geostatistical interpolation methods will provide further options.

T 4.5: The geostatistics and variogram package (6 PM) will provide a wide range of kriging (Ordinary, External Drift, Simple Updating, Universal) and conditional geostatistical simulation methods using the Python package Scikit-Gstat (KIT) and GSTools (UFZ). This will also provide an additional option for a Monte Carlo-based uncertainty assessment. The meaningful application of geostatistics depends essentially on the quality of the variogram. Estimating and assessing variograms is the main focus of Scikit-Gstat.

4.6: The multivariate statistical model and machine learning package (10 PM) will include both standard multivariate statistical methods and machine learning tools. Some approaches from multivariate statistics we intend to add are (generalized) linear models and principal component analysis. Machine learning tools comprise both methods for classification and regression, e.g. random forests, support vector machines or AdaBoost. Additionally, we intend to include artificial neural networks to fully exploit the data science options, i.e. to support hydrological modelling, that the data scheme and metadata model of V-FOR-WaTer provides.

UC 3: Multi-sensor and multiscale evapotranspiration data and formulas

This use case requires the incorporation of the BRIDGET package and the expansion of its functionalities by various ET estimation formulas based on meteorological variables.

T 4.7: The BRIDGET package (3 PM) will be combined with various formulas of different complexity to calculate potential ET. This will allow for better understanding of ET estimates, their inconsistencies, performance and uncertainty across scales, landscape and climate settings. We will sustain the cooperation with the TERENO community, because most of the addressed problems are generic for other estimates of fluxes between the land surface and the atmosphere.

2.3.5 WP 5: Outreach and dissemination (Lead IWG, 18 PM)

The goal of WP 5 is to continuously advertise V-FOR-WaTer, broaden the user community and receive valuable user feedback to improve developments and evaluate the project progress. This will be achieved within three release steps, coinciding with the milestones in Figure 2:

T 5.1: The Alpha Release workshop (3 PM) with selected users from universities and state offices after the first year and **M1**. This workshop will be the first test of the portal, data management, tools, functionalities and general handling.

T 5.2: User-defined research tools (9 PM). While a simple standardized interface for this is task already exists with WPS, this will be considerably expanded in cooperation with the alpha and beta testers group, to evaluate and update the workflow for including new tools.

T 5.3: The Beta Release workshop (3 PM) at EGU will encourage the participating PhD students and PostDocs of the young hydrological society to upload their data in advance and explore the tools and visualization possibilities during the workshop. Encouraging data publication and communicating the licensing and metadata requirements, we will most likely gain additional datasets for V-FOR-WaTer. Moreover, the workshop will support the participants to add their own packages to the system and attract potential new users.

T 5.4 The Release to Web workshop (3 PM). All features of the work packages will be implemented and tested within a second EGU workshop, which will also serve as a roll-out for V-FOR-WaTer.

3 Bibliography

Berndt, C., E. Rabiei, and U. Haberlandt (2014). “Geostatistical merging of rain gauge and radar data for high temporal resolutions and various station density scenarios”. J. Hydrol., 508, pp. 88–101.

Beven, K. and J. Freer (2001). “Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology”. J. Hydrol., 249(1-4), 11–29.

Blöschl, G. and M. Sivapalan (1995). “Scale issues in hydrological modelling: a review”. Hydrol. Process, 9, 251–290.

Brenner, C., Zeeman, M., Bernhardt, M. and K. Schulz (2018). “Estimation of Evapotranspiration of Temperate Grassland Based on High-Resolution Thermal and Visible Range Imagery from Unmanned Aerial Systems“. Int. J. Remote Sens., 39(15–16), 5141–5174.

Ceola, S. et al. (2015). “Virtual laboratories: new opportunities for collaborative water science”. Hydrol. Earth Syst. Sci., 19(4), 2101–2117.

Cloke, H. L. and F. Pappenberger (2008). “Evaluating forecasts of extreme events for hydrological applications: An approach for screening unfamiliar performance measures”. Meteorol. Appl., 15(1), 181–197.

Ehret, U. and E. Zehe (2011). “Series distance - An intuitive metric to quantify hydrograph similarity in terms of occurrence, amplitude and timing of hydrological events”. Hydrol. Earth Syst. Sci., 15(3), 877–896.

Ehret, U., R. van Pruijssen, M. Bortoli, R. Loritz, E. Azmi, and E. Zehe (2020). “Adaptive clustering: reducing the computational costs of distributed (hydrological) modelling by exploiting time-variable similarity among model elements”. Hydrol. Earth Syst. Sci., 24(9), 4389-4411.

Graeff, T., E. Zehe, T. Blume, T. Francke, and B. Schröder (2012). “Predicting event response in a nested catchment with generalized linear models and a distributed watershed model”. Hydrol. Process., 26(24), 3749–3769.

Gupta, H. V., S. Sorooshian, and P. O. Yapo (1998). “Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information”. Water Resour. Res., 34(4), 751–763.

Hassler, S., M. Weiler, and T. Blume (2018). “Tree-, stand- and site-specific controls on landscape-scale patterns of transpiration”. Hydrol. Earth Syst. Sci., 22(1), 13–30.

Heistermann, M., U. Ehret, A. Bronstert, and E. Zehe (2008). “Merging radar and ground observations for operational rainfall estimation in a mountainous head catchment in Germany”. Internat. Symp. Weather Radar and Hydrology.

Hutton, C., T. Wagener, J. Freer, D. Han, C. Duffy, and B. Arheimer (2016). “Most computational hydrology is not reproducible, so is it really science?” Water Resour. Res., 52.10, 7548–7555.

Kavetski, D., G. Kuczera, and S. W. Franks (2006). “Bayesian analysis of input uncertainty in hydrological modeling: 2. application”. Water Resour. Res., 42(3), W03408.

Leijnse, H., R. Uijlenhoet, and J. Stricker (2007). “Rainfall measurement using radio links from cellular communication networks”. Water Resour. Res., 43(3), W03201.

Loritz, R., H. Gupta, C. Jackisch, M. Westhoff, A. Kleidon, U. Ehret, and E. Zehe (2018). “On the dynamic nature of hydrological similarity”. Hydrol. Earth Syst. Sci., 22(7), 3663–3684.

Mälicke, M., S. K. Hassler, T. Blume, M. Weiler, and E. Zehe (2020). “Soil moisture: variable in space but redundant in time”. Hydrol. Earth Syst. Sci., 24(5), 2633–2653.

Neuper, M. and U. Ehret (2019). “Quantitative precipitation estimation with weather radar using a data-and information-based approach”. Hydrol. Earth Syst. Sci., 23, 3711–3733.

Oki, T. and S. Kanae (2006). “Global hydrological cycles and world water resources”. Science, 313(5790), 1068–1072.

Reusser, D., T. Blume, B. Schaefli, and E. Zehe (2009). “Analysing the temporal dynamics of model performance for hydrological models”. Hydrol. Earth Syst. Sci., 13(7), 953–1297.

Reusser, D. E., and E. Zehe (2011). “Inferring model structural deficits by analyzing temporal dynamics of model performance and parameter sensitivity”. Water Resour. Res., 47(7), W07550.

Seibert, S., U. Ehret, and E. Zehe (2016). “Disentangling timing and amplitude errors in streamflow simulations”. Hydrol. Earth Syst. Sci., 20(9), 3745–3763.

Sinclair, S. and G. Pegram (2005). “Combining radar and rain gauge rainfall estimates using conditional merging”. Atmos. Sci. Lett., 6(1), 19–22.

Zehe, E., R. Becker, A. Bardossy, and E. Plate (2005). “Uncertainty of simulated catchment runoff response in the presence of threshold processes: Role of initial soil moisture and precipitation”. J. Hydrol., 315(1–4), 183–202.

Zehe, E., T. Graeff, M. Morgner, A. Bauer, and A. Bronstert (2010). “Plot and field scale soil moisture dynamics and subsurface wetness control on runoff generation in a headwater in the Ore Mountains”. Hydrol. Earth Syst. Sci., 18, 873–889.

Zehe, E., U. Ehret, L. Pfister, T. Blume, B. Schröder, M. Westhoff, C. Jackisch, S.J. Schymanski, M. Weiler, K. Schulz, N. Allroggen, J. Tronicke, L. van Schaik, P. Dietrich, U. Scherer, J. Eccard, V. Wulfmeyer, and A. Kleidon (2014). “HESS Opinions: From response units to functional units: a thermodynamic reinterpretation of the HRU concept to link spatial organization and functioning of intermediate scale catchments”. Hydrol. Earth Syst. Sci., 18, 4635–4655.

4 Supplementary information on the research context

4.1 General ethical aspects

Not applicable.

4.2 Measures to meet funding requirements and handle project results

4.2.1 Project requirements

At the end of the implementation phase proposed in ISABEL, we will establish a long-term operating research data infrastructure. This is in line with the KIT strategy to foster research data management due to the regained excellence status of KIT, and the KIT Climate and Environment Center is envisioned as the institutional host. We expect that we can operate the V-FOR-WaTer VRE at least in the silver operation mode:

- **Bronze operations mode:** The portal will be operated as developed in ISABEL. The software will be maintained in terms of security updates. Issues will be handled as soon as possible on a best effort basis, i.e., without guaranteed availability times.
- **Silver operations mode:** This mode allows actively supporting users of the infrastructure, making new data sources accessible, and integrating external software tools as WPS.
- **Gold operations mode:** In addition to the silver mode, developers enhance the functionality of the research infrastructure based on community demands and feature requests.

The necessary IT infrastructure to run the virtual research environment will be provided by SCC. This consists of servers running the web front end software and connected components (e.g. WPS server and Geoserver). In addition, storage capacities up to one petabyte for the database back end will be provided. Communities or projects that are interested in uploading many large datasets need to negotiate possible payment models. In order to minimize software maintenance efforts, to facilitate replacement of outdated software components, and to assure interoperability with external tools, V-FOR-WaTer is based on commonly used and actively maintained software frameworks (e.g. django). It also implements or supports international standards (e.g. geotools: WPS, WMS; metadata: ISO19115, INSPIRE; authentication: OpenID connect). Although there will be no special requirements to run the V-FOR-WaTer services, we will integrate those in the computing environment of SCC as much as possible. Specifically, we will choose a commonly used Linux distribution, make use of existing monitoring tools, and will make use of SCCs configuration management tools. In this way, all non-V-FOR-WaTer specific maintenance (e.g. operating system or web server updates) can be handled by a large group of system administrators without special knowledge on the V-FOR-WaTer services.

The realization of the two operation modes depends on the funding situation at the end of ISABEL. The department Data Analytics, Access and Applications (D3A) at SCC headed by Dr. Jörg Meyer (permanent contract) will offer the bronze operations mode as part of the Simulation and Data Lifecycle Lab (SDL) "Earth System Science" funded by the Program-Oriented Funding of the Helmholtz Association. A higher level of service (silver, gold) is desired. In T 1.3 we will explore long-term funding opportunities and operation models. The vision of V-FOR-WaTer is to offer data and resources free of charge, still for long-term operations we will consider charging projects for resources and support for the projects data management plans. In addition to foster long term funding V-FOR-WaTer is already in contact with existing initiatives for science data collaborations. The team contributes to the Helmholtz project Digital Earth a Data Science project integrating SMART Monitoring and Data Exploration for a holistic understanding of the Earth System[41]. V-

[41] https://www.digitalearth-hgf.de/en

FOR-WaTer aims at becoming a member of the consortium for the National Research Data Infrastructure for earth sciences[42].

4.2.2 Project results

“Source code for the software developed under the project will be documented in accordance with the principles of open source and made available for use by third parties.” The Code of the V-FOR-WaTer Portal is already published[43] under the MIT License and also all new developments of V-FOR-WaTer will be published under the MIT License as open source software project that may be used free of charge. It will include documentation and a detailed user guide. The software will implement and support relevant international standards. The research infrastructure will continue operations after this phase.

4.3 Formal assurances

- Research data management is a strategic focus of KIT within the current funding period of the German Excellence Initiative. The institutions to host and foster a digital data center for earth and environmental research data are the KIT Climate and Environment Center in cooperation with the SCC.
- Publications resulting from the project and any relevant documentation will be available via open access, making them widely accessible for use by third parties.
- The source code for the software developed under the project will be documented in accordance with the principles of open source and made available for use by third parties.
- We hereby confirm complete compliance with all commitments made in the aforementioned project proposal with regard to implementing the project, in particular to making the necessary financial contributions and, if applicable, to providing the project results on a permanent basis and/or ensuring them in the long term, even after the project is complete. Retroactive amendments may only be made in coordination with the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation).

5 People/collaborations/Funding

5.1 Employment status information

Streit, Achim, Professor (applicant)

Zehe, Erwin, Professor (applicant)

5.2 Composition of the project group

Mälicke, Mirko, M.Sc., doctoral researcher, budget funds of IWG, employed until 2024.

Meyer, Jörg, Dr., researcher, permanent, budget funds of SCC.

5.3 Institutions or researchers in Germany with which/whom you have agreed to cooperate on this project

Not applicable.

5.4 Institutions or researchers abroad with which/whom you have agreed to cooperate on this project

Not applicable.

[42] https://www.nfdi4earth.de

[43] https://github.com/VForWaTer/vforwater-portal

5.5 Institutions and researchers with which/whom you have collaborated scientifically within the past three years

Achim Streit: A. Brinkmann, V. Gülzow, V. Heuveline, V. Lindenstruth, T. Lippert, W. Nagel, H. Pfeiffenberger, M. Resch, G. Schneider, S. Wesner, R. Yahyapour.

Erwin Zehe: B. Berkowitz, T. Blume, U. Ehret, J. Klaus, A. Kleidon, J. Lange, L. Pfister, H. Savenije, B. Schröder, K. Schulz, J. Tronicke, L. van Schaik, W. Wilcke, M. Weiler, M. Westhoff, V. Wulfmeyer.

5.6 Project-relevant cooperation with commercial enterprises

Not applicable.

5.7 Project-relevant participation in commercial enterprises

Not applicable.

5.8 Other submissions

Not applicable.

5.9 Financial contributions

Achim Streit (SCC) and Erwin Zehe (IWG) respectively will cover the workspace, computing needs, and consumables for its staff as part of the basic support.

5.9.1 Funding for staff

5.9.1.1 Research staff

5.9.1.1.1 Postdoctoral researcher or comparable

Streit, Achim	5% head of group (Jörg Meyer) for administrative tasks for the web portal for 3 years
Zehe, Erwin	50% PostDoc (Mirko Mälicke) position for data management and development of tools (WP 2/4)

5.9.1.1.2 Doctoral researcher or comparable

Not applicable.

5.9.1.1.3 Other research assistant

Not applicable.

5.9.1.2 Non-academic staff member

Not applicable.

5.9.1.3 Miscellaneous staff

Not applicable.

5.9.2 Funding for Direct Project Costs

5.9.2.1 Equipment up to €10,000, software and consumables

Not applicable.

5.9.2.2 Visiting Researchers

Not applicable.

5.9.2.3 Other

Streit, Achim	Expenses for storage and servers at SCC for development and hosting of the web portal €500/month.	€18,000

5.9.2.4 Project-related publication expenses

Not applicable.

5.9.3 Funding for instrumentation

Not applicable.

6 Requested modules/funding

For each applicant institution, we apply for funding within the Basic Module.

6.1 Funding for staff

6.1.1 Research staff

6.1.1.1 Postdoctoral researcher or comparable

We apply for the following positions for three years:

Streit, Achim	1 PostDoc position (TVL E13) for Dr. Marcus Strobl for front end development, data management and collaboration with domain scientists (WP 2/3) 1 PostDoc position (TVL E13) for back end development and communication with infrastructure experts (WP 3/4)
Zehe, Erwin	1 PostDoc position (TVL E14) for Dr. Sibylle Hassler for project coordination, data management and quality, evaluation of scientific assets (WP 1, WP 2, & WP 5), tool development in WP 4 (UC 3) 1 PhD position (TVL E13) for WP 4, design and implementation of scientific use cases, tool development in WP 4 (UC 1, 2, 3) WP 5

In addition, we apply for one more web developer for the first year:

Streit, Achim	1 PostDoc position (TVL E13) for an experienced web developer (WP 3.1/3.2)

6.1.2 Non-academic staff member

Not applicable.

6.1.3 Miscellaneous staff

6.1.3.1 Support staff (research support staff and student assistants)

Streit, Achim	1 student assistant to support cleaning and documentation of code, and testing of new frameworks (3 years)
Zehe, Erwin	1 student assistant for reviewing data and tools before they are imported to database and portal, assistance in evaluation of data quality, testing of toolbox (3 years)

6.2 Funding for Direct Project Costs

6.2.1 Equipment up to €10,000, software and consumables

Not applicable.

6.2.2 Travel

To get a wide awareness of the V-FOR-WaTer Portal and to connect to projects in the domain and for the infrastructure as well, travel to conferences for domain sciences and infrastructure hence from scientists from SCC and IWG as well are necessary.

National conferences of the hydrology domain e.g. two-day-conference Tag der Hydrologie: conference fee including social event €200, train ticket (conference cities change every year and

	are not set yet) ~€100 one way, 2x daily benefits €24/day, hotel costs ~€60/night sums up to ~**€600**.	
Streit, Achim	2 x for infrastructure sessions	€1,200
Zehe, Erwin	3 x for domain sessions	€1,800
	International conferences of domain with presentations in domain and infrastructure sessions. Most important in Europe is the General Assembly of the European Geoscience Union (EGU) in Vienna. Globally most recognized earth science conference is the Fall Meeting of the American Geophysical Union (AGU). We will attend both, to advertise and connect and exchange with possible partners and colleagues in Europe at EGU and in a later stage of the project to present the portal and for global contacts at AGU. Costs for EGU: abstract + conference fee €625, train ticket to Vienna ~€150 one way, 6x daily benefits for Austria €33/day, 6x hotel costs ~€100/night sums up to ~**€1,700**. Costs for AGU: conference fee €810, travel to San Francisco ~€500 one way, 6x daily benefits for San Francisco €42/day, 6x hotel costs ~€108/night sums up to ~**€2,700**.	
Streit, Achim	3 x EGU, 1 x AGU Fall meeting	€7,800
Zehe, Erwin	3 x EGU, 1 x AGU Fall meeting	€7,800
	National conferences for infrastructure e.g. three-day-conference E-Science-Tage, conference fee €50, train ticket to Heidelberg ~€10 one way sums up to ~**€120**.	
Streit, Achim	3 x for infrastructure sessions	€360
	International infrastructure conference e.g. International Workshop on Science Gateways (3 days), venue changes every year, estimated costs for 2019: conference fee €205, travel to Ljubljana, Slovenia ~€200 one way, 3x daily benefits for Slovenia €27/day, 3x hotel costs ~€85/night sums up to ~**€940**.	
Streit, Achim	1 x for infrastructure sessions	€940
	Coordination meetings with external data providers e.g. GFZ, UFZ, AWI/MARUM (PANGAEA) train ticket~€100 one way, one day daily benefit €24, hotel cost~€60/night sums up to ~**€285**.	
Streit, Achim	3 x for developer	€850
Zehe, Erwin	3 x for domain scientist	€850
Streit, Achim	**Total travel costs in basic module for infrastructure science**	**€11,150**
Zehe, Erwin	**Total travel costs in basic module for domain science**	**€10,450**

6.2.3 Visiting Researchers

Not applicable.

6.2.4 Experimental animals

Not applicable.

6.2.5 Other

Funding for workshops

We request for the workshops a funding of	€5,000

6.2.6 Project-related publication expenses

Streit, Achim	3 x Publication fees €750/year	€2,250
Zehe, Erwin	3 x Publication fees €750/year	€2,250

6.3 Funding for instrumentation

Not applicable.

7 Additional information

Appendices:

CV_PubList_Streit

CV_PubList_Zehe

ISABEL_Letters_of_intent_and_support

Confirmation of compliance