Manuscript prepared for Hydrol. Earth Syst. Sci.
with version 2014/09/16 7.15 Copernicus papers of the LaTeX class copernicus.cls.
Date: 16 March 2015

Virtual laboratories: New opportunities for collaborative water science

S. Ceola[1], B. Arheimer[2], E. Baratti[1], G. Blöschl[3], R. Capell[2], A. Castellarin[1], J. Freer[4], D. Han[5], M. Hrachowitz[6], Y. Hundecha[2], C. Hutton[4,5], G. Lindström[2], A. Montanari[1], R. Nijzink[6], J. Parajka[3], E. Toth[1], A. Viglione[3], and T. Wagener[5,7]

[1]Department DICAM, University of Bologna, Bologna, Italy
[2]Hydrology Research Section, Swedish Meteorological and Hydrological Institute (SMHI), Norrköping, Sweden
[3]Institute of Hydraulic Engineering and Water Resources Management, Vienna University of Technology, Vienna, Austria
[4]School of Geographical Sciences, University of Bristol, Bristol, UK
[5]Department of Civil Engineering, University of Bristol, Bristol, UK
[6]Water Resources Section, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, The Netherlands
[7]Cabot Institute, University of Bristol, Bristol, UK

Correspondence to: S. Ceola (serena.ceola@unibo.it)

Abstract. Reproducibility and repeatability of experiments are the fundamental prerequisites that allow researchers to validate results and share hydrological knowledge, experience and expertise in the light of global water management problems. Virtual laboratories offer new opportunities to enable these prerequisites since they allow experimenters to share data, tools and pre-defined experimental procedures (i.e., protocols). Here we present the outcomes of a first collaborative numerical experiment undertaken by five different international research groups in a virtual laboratory to address the key issues of reproducibility and repeatability. Moving from the definition of accurate and detailed experimental protocols, a rainfall-runoff model was independently applied to 15 European catchments by the research groups and model results were collectively examined through a web-based discussion. We found that a detailed modelling protocol was crucial to ensure the comparability and reproducibility of the proposed experiment across groups. Our results suggest that sharing comprehensive and precise protocols and running the experiments within a controlled environment (e.g., virtual laboratory) is as fundamental as sharing data and tools for ensuring experiment repeatability and reproducibility across the broad scientific community and thus advancing hydrology in a more coherent way.

1 Introduction

Global water resources are increasingly recognised to be a major concern for a sustainable development of the society (e.g., Haddeland et al., 2014; Schewe et al., 2014; Berghuijs et al., 2014). Ongo-

ing changes in demography, land use and climate will likely exacerbate the current circumstances (Montanari et al., 2013). Water availability and distribution support both ecosystem (Ceola et al., 2013, 2014a) and human demand for drinking water, food, sanitation, energy, industrial production, transport and recreation. Water is also recognised as the most important environmental hazard: floods (Ceola et al., 2014), droughts and water-borne diseases (Rinaldo et al., 2012) cause thousands of casualties, famine, significant disruption and damage worth billions every year (e.g., Jongman et al., 2012; UNISDR, 2013; Ward et al., 2013). Efficient water management is thus crucial for sustainable development of human society. As a consequence, a sound coherent science underpinning decision making is urgently needed. Many studies have already acknowledged the needs for a scientific advancement in water resources management and improved computational models for decision support, which should be capable of predicting the implications of a changing world (Milly et al., 2008; Montanari and Koutsoyiannis, 2012, 2014a, b; Montanari et al., 2013; Koutsoyiannis and Montanari, 2014; Wagener et al., 2010; Gao et al., 2014; Ceola et al., 2014b). Unfortunately, the large diversity of hydrological systems (i.e., catchments) makes it very difficult to identify overarching, scale independent organizing principles of hydrological functions that are required for sustainable and systematic global water management (Beven, 2000; Wagener et al., 2007; Hrachowitz et al., 2013). Blöschl et al. (2013, p. 4) noted that, as hydrologists, we do not have a single object of study. Many hydrological research groups around the world are studying different objects, i.e., different catchments with different response characteristics, thus contributing to the fragmentation of hydrology at various levels. In addition, environmental data are often not easily accessible for hydrological comparisons to enable universal principles to be identified (Viglione et al., 2010). Data are often not provided in appropriate formats, quality checked and/or adequately documented. The hydrological community has therefore recently started to urge for more collaboration between different research groups, to establish large data samples, improve interoperability and comparative hydrology (Duan et al., 2006; Arheimer et al., 2011; Blöschl et al., 2013; Gupta et al., 2014). Sharing data and tools, embedded within virtual observatories, may be a way forward to advance hydrological sciences in a coherent way. In Europe, a major recent development has been the implementation of the INSPIRE Directive (2007/2/EC) in 2007, which provides a general framework for Spatial Data Infrastructure (SDI) in Europe. This directive requires that common implementing rules are adopted in all member states for a number of specific areas (e.g., metadata, data specifications, network services, data and service sharing, monitoring and reporting) by 2020. Worldwide, similar initiatives can be found by the World Meteorological Organisation, WMO (http://www.whycos.org/whycos/), the Earth Observation Communities, GEOSS (http://www.earthobservations.org/geoss.php), and the World Water Assessment Programme by UNESCO (2012). However, sharing of open data and source codes does not automatically lead to good research and scientific advancement.

Reproducibility and repeatability of experiments are the core of scientific theory for ensuring scientific progress. Reproducibility is the ability to perform and reproduce results from an exper-

iment conducted under near identical conditions by different observers in order to independently test findings. Repeatability refers to the degree of agreement of tests or measurements on replicate specimens by the same observer under the same control conditions. Thus, only providing data through open online platforms (or any other way) is not enough to ensure that reproducibility objectives can be met. In fact, the inference previously drawn may be ambiguous to different observers if insufficient knowledge of the experimental design is available. Holländer et al. (2009, 2014) highlighted the impact of modellers' decisions on hydrological predictions. Hydrology is therefore likely to be similar to other sciences that have not yet converged to a common approach to modelling their entities of study. In such cases, meaningful interpretations of comparisons are problematic, as illustrated by many catchment – or model – inter-comparison studies in the past. Model inter-comparison studies at a global scale, including social interactions with the natural system, like e.g. ISLSCP (http://daac.ornl.gov/ISLSCP_II/islscpii.shtml), EU-WATCH (http://www.eu-watch.org/) and ISI-MIP (https://www.pik-potsdam.de/research/climate-impacts-and-vulnerabilities/research/rd2-cross-cutting-activities/isi-mip), but also comparative model inter-comparison experiments in hydrology (i.e. performed by different and independent research groups) such as MOPEX (Duan et al., 2006; Andreassian et al., 2006), DMIP (Reed et al., 2004) or LUCHEM (Breuer et al., 2009), though successful with respect to data sharing, have contributed little to disentangle the causes of performance differences between different models and to increase our understanding of underlying hydrological processes. This was ultimately often rooted in the problems that (see e.g., Clark et al., 2011; Gudmundsson et al., 2012): (i) there are considerable differences in model structures which hinder the identification of particular features that make it perform better or worse; (ii) different research groups make various different decisions for pre-processing data and calibrating models (although often thought to be negligible, this may, cumulatively, prevent a valid comparison of differences in the results); and (iii) comparing model outputs without analysis of model states and internal fluxes provides limited insight into the workings of a model. Hence, greater acknowledgement is required of the dependency of scientific experiments on the applied procedure and choices made in observation and modelling to identify causal relationships (e.g., setting up of boundary conditions, forcing conditions, narrowing of degrees of freedom), both in empirical field work (Parsons et al., 1994) and modelling studies (Duan et al., 2006; Gudmundsson et al., 2012). This would ensure more transparency in the data and methods used in experiments. In particular, hydrology suffers from the perceived difficulty of reporting detailed experiment protocols in the research literature, largely underexploiting the convenient option to provide supplementary information in scientific journals. Thus, in the presence of open data platforms, setting up strategies to guarantee experiment reproducibility and thereby a means for meaningful inter-experiment comparison is a challenging target. It requires a concerted and interdisciplinary effort, involving information technology, environmental sciences and dissemination policy in developing and communicating strict, detailed, coherent and generally unambiguous experiment protocols.

In this paper we explore the potential of a virtual water-science laboratory to overcome the aforementioned problems. A virtual laboratory provides a platform to share data, tools and experimental protocols (Ramasundaram et al., 2005). In particular, experimental protocols constitute an essential part of a scientific experiment, as they guarantee quality assurance and good practice (e.g., Refsgaard et al., 2005; Jakeman et al., 2006) and, we argue, are at the core of repeatability and reproducibility of the scientific experiments. More specifically, a protocol is a detailed plan of a scientific experiment that describes its design and implementation. Protocols usually include detailed procedures and lists of required equipment and instruments, information on data, experimenting methods and standards for reporting the results through post-processing of model outputs. By including a collection of research facilities, such as e-infrastructure and protocols, virtual laboratories have the potential to stimulate entirely new forms of scientific research through improved collaboration. Pilot studies, such as the Environmental Virtual Observatory (EVO - http://www.evo-uk.org), have already explored a number of these issues and, additionally, the legal and security challenges to overcome. Other example projects related to hydrology, which are exploring community data sharing and interoperability, include DRIHM (http://www.drihm.eu), NEON in the USA (http://www.neoninc.org), and the Organic Data Science Framework (http://www.organicdatascience.org/). To sum up, virtual laboratories aim at (i) facilitating repetition of numerical experiments undertaken by other researchers for quality assurance, and (ii) contributing to collaborative research. Virtual laboratories therefore provide an opportunity to make hydrology a more rigorous science. However, virtual laboratories are relatively novel in environmental research and their essential requirements to ensure the repeatability and reproducibility of experiments are still unclear. Therefore, we have undertaken a collaborative experiment, among five universities and research institutes, to explore the possible critical issues that may arise in the development of virtual laboratories. This paper presents a collaborative simulation experiment on reproducibility in hydrology, using the Virtual Water-Science Laboratory, established within the context of the EU funded research project "Sharing Water-related Information to Tackle Changes in the Hydrosphere – for Operational Needs (SWITCH-ON)", (http://www.water-switch-on.eu/), which is currently under development. The paper aims to address the following questions:

1. What factors control reproducibility in computational scientific experiments in hydrology?

2. What is the way forward to ensure reproducibility in hydrology?

After presenting the structure of the Virtual Water-Science Laboratory (VWSL), we describe in detail the collaborative experiment, carried out by the research groups in the VWSL. We deliberately decided to design the experiment as a relatively traditional exercise in hydrology in order to better identify critical issues that may arise in virtual laboratories development and dissemination and that are not associated with the complexity of the considered experiment. This experiment therefore

supports subsequent research within the VWSL, and provides an initial guidance to design protocols and share evaluation within virtual laboratories by the broad scientific community.

2 The SWITCH-ON Virtual Water-Science Laboratory

The purpose of the SWITCH-ON VWSL is to provide a common workspace for collaborative and meaningful comparative hydrology. The laboratory aims to facilitate, through the development of detailed protocols, the sharing of data tools, models and any other relevant supporting information, thus allowing experiments on a common basis of open data and well defined procedures. This will not only enhance the general comparability of different experiments on specific topics carried out by different research groups, but the available data and tools will also facilitate researchers to more easily exploit the advantages of comparative hydrology and collaboration, which is widely regarded as a prerequisite for scientific advance in the discipline (Falkenmark and Chapman, 1989; Duan et al., 2006; Wagener et al., 2007; Arheimer et al., 2011; Blöschl et al., 2013; Gupta et al., 2014). In addition, the VWSL aims to foster cooperative work by actively supporting discussions and collaborative work. Although the VWSL is currently used only by researchers who are part of the EU FP7-project SWITCH-ON, it is also open to external research groups to obtain feedback and to establish a sustainable infrastructure that will remain after the end of the project. Any experiment formulated within the VWSL needs to comply with specific stages, shown as an 8-point workflow described in detail below, which outline the scientific process and the structure for using the facilitating tools in the VWSL.

STAGE 1: Define science questions. This stage allows researchers to discuss through a dedicated on-line forum (available at https://groups.google.com/forum/#!forum/virtual-water-science-laboratory-forum) specific hydrological topics to be elaborated upon by different research groups in a collaborative context. Templates are available to formulate new experiments.

STAGE 2: Set up experiment protocols. In this step a recommended protocol for collaborative experiments needs to be developed. This protocol formalises the main interactions between project partners and acts as a guideline for the experiment outline in order to ensure experiment reproducibility and thus controlling the degree of freedom of single modellers.

STAGE 3: Collect input data. The VWSL contains a catalogue of relevant external data available as open data from any source on the Internet in a format that can be directly used in experiments. Stored data are organised in Level A (pan-European scale covering the whole of Europe) and Level B (local data covering limited or regional domains). Currently Level A includes input data to the E-HYPE model (Donnelly et al., 2014) with some 35,000 sub-basins covering Europe such as precipitation, evaporation, soil and land use, river discharge and nutrients data, while Level B includes hydrological data (i.e. precipitation, temperature and river

discharge) for 15-20 selected catchments across Europe. In addition, a Spatial Information Platform (SIP) has been created. This platform includes a catalogue with a user interface for browsing among metadata from many data providers. So far, the data catalogue has been filled with 6990 items of files for download, data viewers and web pages. The SIP also includes functionalities for linking more metadata, and visualisation of datasets. Therefore, through stored data and the SIP, researchers can easily find and explore data deemed to be relevant for a hydrological experiment.

STAGE 4: Repurpose data to input files. In this step, raw original data from STAGE 2 can be processed (i.e., transformed, merged, etc.) to create suitable input files for hydrological experiments or models. For example, the World Hydrological Input Set-up Tool (WHIST) can tailor data to specific models or resolutions. An alternative example, planned to be used for future activities in the VWSL, is provided by land use data, which can be aggregated to relevant classes and adjusted to specific spatial discretisations (e.g., model grid or sub-basin areas across Europe). Both raw original and repurposed data (STAGES 2 and 3) should be accompanied by detailed metadata (i.e., a protocol), which specify e.g., data origin, spatial and temporal resolution, observation period, description of the observing instrument, information on data collection, measures of data quality, coherency of the measured method and instrument, and any other relevant information. Data should be provided to international open source data standards (i.e., http://www.opengeospatial.org) and, for water related research in particular, it should be compliant with the WaterML2 international initiatives (see above site for more information).

STAGE 5: Compute model outputs. By employing open source model codes, freely available via the VWSL, or through links to model providers, researchers can perform hydrological model calculations using the same tools. Results can then be compared, evaluated, reused and/or repurposed for new experiments. In addition, templates for protocols are available to ensure the reproducibility and repeatability of model analysis and results. The protocol may include, for instance, a description of the hydrological experiment, and information on the model, input data and metadata, employed algorithms and temporal scales. Protocols for model experiments will thus create a framework for a generally accepted, scientifically valid and identical environment for specific types of numerical experiments within the VWSL, and will promote transparency and data sharing, therefore allowing other researchers to download and reproduce the experiment on their own computer.

STAGE 6: Share results. Links to model results are uploaded to the VWSL in order to ensure the post audit analyses and transparency of the performed experiments, which can be reproduced by other research groups.

STAGE 7: Explore the findings. Here, researchers can extract, evaluate and visualise experiment results gathered at STAGE 5. A separate space for discussion and comparisons of results, through the on-line forum, additionally facilitates direct and open knowledge exchange between researchers and research teams.

STAGE 8: Publish and access papers. Links to scientific papers and technical reports on comparative research resulting from collaboration and experiments based on data in the VWSL will be found in the VWSL.

3 The first collaborative experiment in the SWITCH-ON Virtual Water-Science Laboratory

3.1 Description and purpose of the experiment

The first pilot experiment of the SWITCH-ON VWSL aims to assess the reproducibility of the calibration and validation of a lumped rainfall-runoff model over 15 European catchments (Fig. 1) by different research groups using open software and open data (STAGE 1). Calibration and validation of rainfall-runoff models is a fundamental step for many hydrological analyses (Blöschl et al., 2013), including drought and flood frequency estimation (see, for instance, Moretti and Montanari, 2008). The rainfall-runoff model adopted in the experiment is a HBV-like model (Bergström, 1976) called TUWmodel (Parajka et al., 2007; Parajka and Viglione, 2012), which is designed to estimate daily streamflow time series from daily rainfall, air temperature and potential evaporation data (STAGE 5). The TUWmodel code (see Supplementary Material for further information), written as a script in the R programming environment (R Core Team, 2014), is run for each of the selected catchments by five research groups, based at the Swedish Meteorological and Hydrological Institute (SMHI), University of Bologna (UNIBO), Technical University Wien (TUW), Technical University Delft (TUD), and University of Bristol (BRISTOL). The R script is run by the five research groups using different operating systems (i.e., Linux by UNIBO, TUW and TUD; Windows 7 by SMHI and BRISTOL). The groups a priori agreed on a rigorous protocol for the experiment (STAGE 2), which is described in detail below, conducted the experiment (STAGES 3, 4, 5), and subsequently engaged in a collective discussion of the results (STAGES 6, 7). Despite the relatively simple hydrologic exercise, this experiment is expected to benefit from a comparison of model outcomes, an exchange of views and modelling strategies among the research partners in order to identify and assess potential sources of violations of the condition of reproducibility. Indeed the experiment has the purpose of bringing scientists to work together collaboratively in a well-defined and controlled hydrological study for result comparison. By exploring reproducibility, this experiment places itself as a base-line for comparative hydrology.

3.2 Study catchment and hydrological data

European catchments characterised by a drainage area larger than 100 km^2 with at least 10 years of daily hydro-meteorological data, as lumped information on rainfall, air temperature, potential evaporation and runoff are considered (STAGE 3). The selected 15 catchments are located in Sweden, Germany, Austria, Switzerland and Italy (Fig. 1). Daily time series of rainfall, temperature and streamflow, gathered from national environmental agencies and public authorities (see Acknowledgements for more details), are pre-processed by the partner who contributed the data set to the experiment (e.g., to homogenise units of measurement) to be employed in the TUWmodel (STAGE 4). Potential evaporation data are derived, as repurposed data (STAGE 4), from hourly temperature and daily potential sunshine duration by a modified Blaney-Criddle equation (for further details, see Parajka et al., 2003). Table 1 reports the foremost features of the 15 study catchments investigated.

3.3 Experiment protocols

As detailed above, the objective of this experiment is to test the reproducibility of the TUWmodel results on the 15 study catchments when implemented and run independently by different research groups. Consequently, the experiment provides an indication of the experimental implementation uncertainty (see e.g., Montanari et al., 2009) due to combined effects of insufficiently developed protocols, human error or computational architecture. To this aim, identical implementations (the R code) of the TUWmodel are distributed to the research groups, and two different protocols (i.e., Protocol 1 and Protocol 2) establishing how to perform the experiment are defined (STAGES 2, 5). Protocol 1 is characterised by a rigid setting, such that the researchers are required to strictly follow pre-defined rules for model calibration and validation, as specified in the distributed R script. By following Protocol 1, all research groups are expected to obtain the same results in terms of comparable model performance. The alternative Protocol 2 allows researchers more flexibility in order to explore and compare several different model calibration options. In this case, research groups have the opportunity to add their personal experience to assess model performance. This will likely provide less comparable results among research groups, but the expected added value of Protocol 2 would be a more extended exploration of different modelling options, which could be synthesized and used for future hydrological experiments in the VWSL. In both protocols the observation period (n years) is divided into two equal-length sub-periods (n/2 years): the first period is used for calibration, and the second for validation as in a classical split-sample test. In Protocol 1, we also switched the two periods (i.e., first period for validation and second period for calibration). Detailed model specifications for the two protocols are described in what follows and their main settings are summarised in Tables 2 and 3.

3.3.1 Protocol 1

For Protocol 1, the calibration of the TUWmodel is based on the Differential Evolution optimisation algorithm (DEoptim, Mullen et al., 2011). This global optimisation tool with differential evolution is readily embedded in the R package that was used to run the entire experiment. Protocol 1 pre-defines the uniform prior model parameter distributions (Table 2). 10 calibration runs, each of them based on different random seeds, are performed in order to identify the best calibration run. The objective function used to determine the optimal model parameters is the mean square error (MSE). Model parameters estimated during the calibration phase are then used to test the TUWmodel in the validation period. For the validation period, Protocol 1 further requires the computation of MSE; root mean square error, RMSE; Nash-Sutcliffe efficiency, NSE; NSE of logarithmic discharges, log(NSE); bias; mean absolute error, MAE; MAE of logarithmic discharges, MALE; and volume error, VE. A model warm-up period of 1 year for both calibration and validation (i.e., model calibration and validation are applied on n/2-1 years), was adopted in order to minimise the influence of initial conditions. The model realisations of the individual research groups were then compared based on the performance metrics and the obtained optimal parameter values. The R-script describing Protocol 1 is presented as Supplementary Material.

3.3.2 Protocol 2

In Protocol 2, the different research groups could make individual choices in an attempt to improve model performances. More specifically, during model calibration on the first half of the observation period, users could (i) shorten the calibration period by excluding what they believe are potentially unreliable pieces of data and providing detailed justifications, (ii) modify the prior parameter distributions, (iii) change the optimisation algorithm and its settings, (iv) select alternative objective functions, and (v) freely choose the model warm-up period (see Table 3 and Supplementary Material for a detailed description). Similarly to Protocol 1, the calibrated parameter values are used as inputs for the evaluation of the simulated discharge during the validation period, and the same goodness-of-fit statistics evaluated in Protocol 1 are also computed.

4 Results

A web-based discussion (STAGES 6, 7) was engaged among the researchers to collectively assess the results, by comparing the experiment outcomes and benefiting from their personal knowledge and experience. The results revealed that reproducibility is ensured when:

- experiment and modelling purpose are outlined in detail, which requires a preliminary agreement on semantics and definitions,

– a standardised format of input data (e.g., file format, data presentation, and units of measurement) and pre-defined variable names are proposed,

– the same model tools (i.e., code and software) are used.

Within a collaborative context, this can be achieved only if the involved research groups completely agreed on the detailed protocol of the experiment. In what follows we report the experiences gained from the experiment, and we finally suggest a process that enables research groups to improve the set-up of protocols.

4.1 Protocol 1

The variability in the optimal calibration performance obtained from all research groups for Protocol 1, ordered by catchments, is shown in Fig. 2. For some catchments, notably the Gadera (ITA) and Großarler Ache (AUT), optimal calibration performance is very similar between groups, indicating that the Protocol has been executed properly by each research group. However, for some other catchments including the Vils (AUT), Broye (SUI), Hoan (SWE) and Juktån (SWE), more variability in optimal performance between groups was obtained. Given that Protocol 1 is not deterministic, as the optimisation algorithm contains a random component, variability in optimal performance will be expected even if the protocol were repeated by a given research group. Thus, in order to make proper comparison between research groups – e.g., assess the reproducibility of an experiment – an understanding of this within-group variability, or repeatability, is required. The range in optimal performance obtained by one research group (BRISTOL) when the optimisation algorithm was run 100 times, instead of 10 times as per Protocol 1, is also plotted in Fig. 2 to give an indication of the within-group variability. With the exception of the second calibration period for the Vils (AUT) catchment, where UNIBO found a lower RMSE, the between-group variability in calibration performance falls within the bounds of the within-group variability, which indicates a successful execution of the Protocol across all catchments. Of the 100 optimisation runs conducted for the Vils (AUT) catchment during the second calibration, 99 were at the upper end of the range in Fig. 2, alongside the results of all groups except UNIBO, and only one result at the lower end of the range. In this case, and in the case of the poorer performance of the BRISTOL calibration for the Broye (SUI), where early stopping of the optimisation algorithm consistently occurred, the results suggest the algorithm became trapped in a local minimum and struggled to converge to a global minimum – or at least to an improved solution, as identified by other groups/runs. In addition to convergence issues causing differences in the results of each group, differences in the identified optimal parameter sets suggest that divergence in performance may also result from parameter insensitivity and equifinality (Fig. 3). Furthermore, performance is also affected by the presence of more complex catchment processes which are not fully captured by the chosen hydrological model (e.g., snowmelt or soil moisture routines in catchments with large altitude range or diverse land covers). Thus, from

a hydrological viewpoint, the results were not completely satisfactory, and detailed analysis at each location is required. However, given that in the majority of cases the between-group variability in performance (reproducibility) was within the range of within-group variability (repeatability) identified, it can be concluded that Protocol 1 ensured reproducibility between groups for the proposed model calibration.

4.2 Protocol 2

To overcome the problems arising from Protocol 1 and possibly improve model performances, the effects of personal knowledge and experience of research groups were explored in Protocol 2. Here, researchers were allowed to more flexibly change model settings, which may introduce a more pronounced variability in the results among the individual research groups, due to different decisions in the modelling processes. Given that flexibility allows a more proficient use of expert knowledge and experience, one may expect an improvement of model performances. Flexibility indeed enables modellers to introduce new choices in order to improve model performance in terms of process representation and consequently correct automatic calibration artefacts for model parameter value selection (as in Protocol 1), which could lead to unexpected model behaviour. The increase in flexibility in Protocol 2 led to a significant divergence in model performance between groups, as exemplified in Fig. 4 for the NSE performance metric. Such changes reflect the different approaches taken in an attempt to improve model performance in terms of process representation, and to correct problems from Protocol 1. In turn, these changes delineate the effects of different personal knowledge and experience of the different research groups. More specifically, BRISTOL and UNIBO both chose to exclude potentially unreliable data from the calibration data. In the case of BRISTOL, following visual inspection of the data, it was felt that a more thorough data evaluation procedure prior to calibration was required. Based on the calculation of event runoff coefficients, a subset of the time-series in 9 catchments was excluded. Researchers from UNIBO decided to exclude nearly one quarter of available data for each study watershed. Data were removed by looking for the highest MSE for each separate year by using the parameter set that allowed the best results on the calibration set in the Protocol 1 experiment. Data removal appeared to lead to improved calibration performance, and to a lesser extent, improved validation performance. As per Protocol 2, data were not removed from the validation period. Conversely, researchers from TUW and TUD decided not to remove any data in the calibration period but to adopt alternative optimisation procedures to enhance the robustness of the calibration (see Table 3). The discussion among modellers pointed out that changing the objective function from MSE to different formulations did not lead to an actual decay of the model performances, but only to lower values of the NSE, due to assigning lower priority to the simulation of the peak flows, while other features of the hydrograph were better simulated. For instance, the Kling-Gupta efficiency was used by TUD as it provides a more equally weighted combination of bias, correlation coefficient and relative variation compared to NSE. This led to reduced bias

and volume error compared to the results of the other groups, but in a trade-off, it worsened the performances in terms of the NSE. Similarly, the use of MSE by BRISTOL led to improvements in log(NSE), MAE and MALE for nearly all catchments in calibration and validation, but increased bias and volume errors in some cases. As there was no uniquely defined objective of Protocol 2, such choices reflected attempts by the groups to achieve an appropriate compromise across performance metrics. SMHI adopted a hydrological process-based approach, where the modellers accepted small performance penalties in terms of NSE if the conceptual behaviour of the model variables looked more appropriate during the calibration procedure. This was done to get a good model for the right reasons, and expert knowledge on hydrological processes and model behaviour was then included along with the statistical criteria. The evaluation of the goodness-of-fit by SMHI was performed by visual comparison and an analysis of several (internal) model variables, e.g., soil moisture, evapotranspiration rates, and snow water equivalents, instead of simply using a different objective function. These analyses pointed to conceptual model failures in several catchments (e.g., Loisach (GER) catchment, Fig. 4), leading to the adoption of a calibration approach which considered the structural limitations of the TUWmodel and their implications for model performance (see also Supplementary Material).

4.3 Identified issues in a collaborative experiment

Collaboration implies communication between scientists. During this first experiment, researchers engaged in a frequent and close communication both via e-mail and through the VWSL forum in order to highlight encountered problems, discuss about model results and their interpretation, and also identify challenges for future improvement of the VWSL itself. In particular, during this experiment several incidents showed the importance of well-defined terms to be able to cope with reproducibility between the research groups. These problems pointed out that communication between different groups through the web may be problematic. Indeed, the hydrological community is not well acquainted with inter-group cooperation. Detailed guidelines, including a preliminary rigorous setting of definitions and terminology, are needed to make a virtual laboratory properly working.

4.4 Suggested procedure to establish protocols for collaborative experiments

Based on the experiment results, we were able to identify a recommended workflow sequence for collaborative experiments, to streamline the work among largely disjoint and independent working partners. The workflow covers three distinct phases: Preparation, Execution, and Analysis (Fig. 5). The Preparation phase contains the bulk of processes specific to collaboration between independent partner groups. Starting from an initial experiment idea, partners are brought together and a coordination structure is chosen. A lead partner, who is responsible for coordination of the experiment preparation, needs to be identified. There are two main tasks in the Preparation phase: establishment and clear communication of the experiment protocol as well as the compilation of a project database.

The definition of protocol specifications can be chosen by the partners, but they must provide detailed and exhaustive instructions regarding (i) driving principles of the protocol, which include and reflect the purpose of the experiment; (ii) data requirements and formatting, (iii) experiment execution steps, and (iv) result reporting and formatting. An initial protocol version is prepared and then evaluated by single partners and returned for improvement if ambiguities are found. Personal choices, independently made by partner groups during a test execution of the experiment, might be included. Such choices need to be well defined, and a comparability of results must be ensured through requirements in the protocol. Once the experiment protocol is agreed, partners collect, compile and publish the data necessary for the experiment using formal version-control criteria, following again a release and evaluation cycle. The Execution phase starts immediately after the completion of these tasks, and the protocol is released to all partners, who perform the experiment independently. The protocol execution can include further interaction between partners, which must be well defined in the protocol. During this phase, there should be a formal mechanism to notify partners of unexpected errors that lead to an experiment abort and return to the protocol definition. Errors can then be corrected in a condensed iteration of the Preparation phase. All partners report experiment results to the coordinating partner, who then compiles and releases the overall outcomes to all partners. The Analysis phase requires partners to analyse experiment results with respect to the proposed goals of the experiment. Partners communicate their analyses, leading to (i) rejection of experiment results as inconclusive regarding the original hypothesis, or (ii) publication of the experiment to a wider research community. This formalized workflow can then be filled by the experiment partners with more specific agreements on the infrastructure for a specific experiment. These may include:

- technical agreements, as data documentation standards to adhere to or computational platforms to be used by the partners;
- means of communication between partners, which could range from simple solutions as the establishment of an e-mail group to more complex forms, as an online communication platform with threaded public and private forums as well as online conferencing facilities;
- file exchange between partners, including data, metadata, instructions, and experiment result content. This could be implemented through informal agreements as a deadline-based collection-compilation-release system, or formal solutions as the use of version-controlled file servers with well-defined release cycles.

5 Discussion and Conclusions

Hydrology has always been hindered by the large variability of our environment. This variability makes it difficult for us to derive generalisable knowledge given that no single group can assess many locations in great detail or build up knowledge about a wide range of different systems. Open

environmental data and the possibilities of a connected world offer new ways in which we might overcome these problems.

In this paper, we present an approach for collaborative numerical experiments using a virtual laboratory. The first experiment that was carried out in the SWITCH-ON VWSL suggests that the value of comparative experiments can be improved by specifying detailed protocols. Indeed, in the context of collaborative experiments, we may recognise two alternative experimental conditions: (i) experimenters want to do exactly the same things (i.e., same model with same data) or (ii) researchers decide to accomplish different model implementations and assumptions based on their personal experience. In the first case, the protocol agreed upon by project participants needs to be accurately defined in order to eliminate personal choices from experiment execution. Under this experimental condition, reproducibility of experimental results among different research groups should be consistent with repeatability within a single research group. The experience from using Protocol 1 showed the importance of an accurate definition of experiment design and a detailed selection of appropriate tools, which helped to overcome several incidents during experimental set-up and execution. Problems related to insensitive parameters, local optima, and inappropriate model structure for the study catchments led to variability in performance across research groups. Our experience revealed that quantifying the within-group variability (i.e., repeatability) is necessary to adequately assess reproducibility between-groups. In turn, residual variability may indicate a lack of reproducibility, and aid in the identification of specific issues, as considered above. In the second case, the experiment is similar to traditional model intercomparison projects (e.g., WMO, 1986, 1992; Duan et al., 2006; Breuer et al., 2009; Parajka et al., 2013), where each group is allowed to perform the experiment by making personal choices and using their own model concept. These choices may lead to major differences in the model setup and parameters (Holländer et al., 2009, 2014). Under these more flexible experimental conditions, the main goal of the experiment should be clearly defined. In Protocol 2, all research groups aimed at improving model performances, even though we did not deliberately specify what 'model improvement' meant a priori: this could be either reaching a higher statistical metric, less equifinality among parameter values or a more reliable model in terms of realistic internal variables. In this case, the main goal of the experiment was to profit from researchers personal experience in order to improve model performances. Indeed, each interpretation could be justified and different considerations could be normally taken by the modeller depending on the purpose of the experiment. Through this process, the modellers were able to engage in a collective discussion that pointed out the model limitations and the sensitivity of the results to different modelling options. Even though results from Protocol 2 are less comparable than the outcomes from Protocol 1, the collective numerical experiment allowed comparison between different approaches suggested by individual experience and knowledge.

Multi-basin applications of hydrological models allowed the experimenters to identify links between physical catchment behaviours, efficient model structures and reliable priors for model param-

eters – all based on expertise with different systems by different groups. Even though we engaged in a relatively simple collaborative hydrological exercise, the results discussed here show that it is important to revisit experiments that are seemingly simpler than existing inter-group model comparisons to understand how small differences affect model performance. What is clear is that it is fundamental to control for different factors that may affect the outcomes of more complex experiments, such as modeller choice and calibration strategy. In more complex situations the virtual experiments could be conducted through comparisons at different levels of detail. For example, if models with different structures were to be compared there will be no one-to-one mapping of the state variables and model parameters and the comparison would be applied to a higher level of conceptualizations. There are a number of examples in the literature where comparisons at different levels of conceptualization have been demonstrated to provide useful results. One such example is Chicken Creek model inter comparison (Holländer et al., 2009, 2014) where the modellers were given an increasing amount of information about the catchment in steps, and in each step the model outputs in terms of water fluxes were compared. The Chicken Creek inter comparison involved models of vastly different complexities, yet provided interesting insights in the way models made assumptions about the hydrological processes in the catchment and the associated model parameters. Another example is the Predictions in Ungauged Basins (PUB) comparative assessment (Blöschl et al., 2013) where a two step process was adopted. In a first step (Level 1 assessment), a literature survey was performed and publications in the international refereed literature were scrutinised for results of the predictive performance of runoff, i.e. a meta-analysis of prior studies performed by the hydrological community. In a second step (Level 2 assessment) some of the authors of the publications from Level 1 were approached with a request to provide data on their runoff predictions for individual ungauged basins. At Level 2 the overall number of catchments involved was smaller than in the Level 1 assessment but much more detailed information on individual catchments was available. Level 1 and Level 2 were therefore complementary steps. In a similar fashion, virtual experiments could be conducted using the protocol proposed in this paper at different, complementary levels of complexity. The procedure for protocol development (Figure 5), which notably checks on independent model choices between partners and feedback to earlier stages in protocol development, will help in developing protocols for more complex collaborative experiments, addressing real science questions on floods, droughts, water quality and changing environments. More elaborated experiments are part of ongoing work in the SWITCH-ON project, and the adequacy of the protocol development procedure itself will be evaluated during these experiments. The modelling study presented in this paper therefore represents a relatively simple, yet no less important first step towards collaborative research in the Virtual Water-Science Laboratory.

To sum up, in this study we set out to answer to the following specific scientific questions related to the concepts of reproducibility of experiments in computational hydrology, previously outlined in the Introduction.

1. *What factors control reproducibility in computational scientific experiments in hydrology?*
   The reproducibility is preliminarily governed by shared data and models along with experiment protocols, which define data requirements (metadata, also indicating versions of datasets) and format (for example, units of measurement, identification of no data, significant observation period), experiment execution (e.g., selection of a well-documented hydrological model code), and result analysis (e.g., criteria for judging model performances). These protocols aim at providing a common agreement and understanding among the involved research groups about data and experiment purpose. Human errors (e.g., ambiguity in variable names, small oversights during model execution) and unclear file-exchange procedures can be considered the main cause of a reduced reproducibility in the case researchers want to do the same thing. Conversely, if different model implementations are allowed, reduced reproducibility may depend on the lack of means of communication and clarity of the purpose of the modelling exercise or on the condition of multiple choices at once.

2. *What is the way forward to ensure reproducibility in hydrology?*
   In the case different research groups use the same data input and model code, an essential prerequisite to set up a reliable experiment is to formalise a rigorous protocol that has to be based on an agreed taxonomy along with a technical environment to avoid human mistakes. If, on the other hand, researchers are allowed to perform different model implementations, the main purpose of the modelling exercise needs to be clearly defined. For instance, in Protocol 2, the added value of researchers scientific knowledge was capable of extensively exploring alternative modelling options, which can be helpful for future hydrological experiments in the VWSL. Furthermore, the experiment should be designed such that the relationship between experimental choices (e.g., cause) and the experimental results (e.g., the effects of these choices) can be clearly determined. This is required to avoid a form of equifinality that results from experimental set-up, where the relative benefits of different choices made between research groups cannot be established. Also in this second case, a controlled technical environment will help to produce reproducible experiments. Therefore, version management of databases, code documentation, metadata, preparation of protocols, and feedback mechanisms among the involved partners are all issues that need to be considered in order to establish a virtual laboratory in hydrology. Virtual laboratories provide the opportunity to share data, knowledge and facilitate scientific reproducibility. Therefore they will also open the doors for the synthesis of individual results. This perspective is particularly important to create and disseminate knowledge and data on water science and open the way to more coherence of hydrological research.

Acknowledgements. The SWITCH-ON Virtual Water-Science Laboratory for water science is being developed within the context of the European Commission FP7 funded research project "Sharing Water-related Information to Tackle Changes in the Hydrosphere - for Operational Needs" (grant agreement number 603587). The overall aim of the project is to promote data sharing to exploit open data sources. The study contributes to developing the framework of the "Panta Rhei" Research Initiative of the International Association of Hydrological Sciences (IAHS). The authors acknowledge the Austrian Hydrographic Service (HZB); the Regional Agency for the Protection of the Environment - Piedmont Region, Italy (ARPA-Piemonte); the Regional Hydrologic Service - Tuscany Region, Italy (SIR Toscana); the Hydrographic Service of the Autonomous Province of Bolzano, Italy; the Global Runoff Data Centre (GRDC); European Climate Assessment & Dataset (ECA&D) for providing hydro-meteorological data that were not yet fully open.

References

Andreassian, V., Hall, A., Chahinian, N., and Schaake, J.: Large Sample Basin Experiments for Hydrological Model Parameterization: Results of the Model Parameter Experiment - MOPEX - IAHS Proceedings & Reports No. 307, IAHS Press, 2006.

Arheimer, B., Wallman, P., Donnelly, C., Nyström, K., and Pers, C.: E-HypeWeb: Service for Water and Climate Information - and Future Hydrological Collaboration across Europe?, in: Environmental Software Systems. Frameworks of eEnvironment, edited by Hřebíček, J., Schimak, G., and Denzer, R., vol. 359 of *IFIP Advances in Information and Communication Technology*, pp. 657–666, Springer Berlin Heidelberg, doi:10.1007/978-3-642-22285-6_71, http://dx.doi.org/10.1007/978-3-642-22285-6_71, 2011.

Berghuijs, W. R., Woods, R. A., and Hrachowitz, M.: A precipitation shift from snow towards rain leads to a decrease in streamflow, Nature Climate Change, 4, 583–586, doi:10.1038/nclimate2246, 2014.

Bergström, S.: Development and application of a conceptual runoff model for Scandinavian catchments - SMHI Reports RHO No.7, Tech. rep., SMHI, Norrköping, 1976.

Beven, K. J.: Uniqueness of place and process representations in hydrological modelling, Hydrology and Earth System Sciences, 4, 203–213, doi:10.5194/hess-4-203-2000, http://www.hydrol-earth-syst-sci.net/4/203/2000/, 2000.

Blöschl, G., Sivapalan, M., Wagener, T., Viglione, A., and Savenije, H.: Runoff Predicitons in Ungauged Basins – Synthesis Across Processes, Places and Scales, Cambridge University Press, Cambridge, UK, 2013.

Breuer, L., Huisman, J. A., Willems, P., Bormann, H., Bronstert, A., Croke, B. F. W., Frede, H. G., Graeff, T., Hubrechts, L., Jakeman, A. J., Kite, G., Lanini, J., Leavesley, G., Lettenmaier, D. P., Lindstroem, G., Seibert, J., Sivapalan, M., and Viney, N. R.: Assessing the impact of land use change on hydrology by ensemble modeling (LUCHEM). I: Model intercomparison with current land use, Advances in Water Resources, 32, 129–146, doi:10.1016/j.advwatres.2008.10.003, 2009.

Ceola, S., Laio, F., and Montanari, A.: Satellite nighttime lights reveal increasing human exposure to floods worldwide, Geophysical Research Letters, 41, 7184–7190, doi:10.1002/2014GL061859, http://dx.doi.org/10.1002/2014GL061859, 2014.

Ceola, S., Hoedl, I., Adlboller, M., Singer, G., Bertuzzo, E., Mari, L., Botter, G., Waringer, J., Battin, T. J., and Rinaldo, A.: Hydrologic Variability Affects Invertebrate Grazing on Phototrophic Biofilms in Stream Microcosms, PLoS ONE, 8, doi:10.1371/journal.pone.0060629, 2013.

Ceola, S., Bertuzzo, E., Singer, G., Battin, T. J., Montanari, A., and Rinaldo, A.: Hydrologic controls on basin- scale distribution of benthic invertebrates, Water Resources Research, 50, 2903–2920, doi:10.1002/2013WR015112, 2014a.

Ceola, S., Montanari, A., and Koutsoyiannis, D.: Toward a theoretical framework for integrated modeling of hydrological change, WIREs Water, 41, doi:10.1002/wat2.1038, 2014b.

Clark, M. P., Kavetski, D., and Fenicia, F.: Pursuing the method of multiple working hypotheses for hydrological modeling, Water Resources Research, 47, doi:10.1029/2010WR009827, 2011.

Donnelly, C., Andersson, J., and Arheimer, B.: Using flow signatures and catchment similarities to evaluate the E-HYPE multi-basin model across Europe, Journal of Hydrological Sciences, In Review, 2014.

Duan, Q., Schaake, J., Andréassian, V., Franks, S., Goteti, G., Gupta, H., Gusev, Y., Habets, F., Hall, A., Hay, L., Hogue, T., Huang, M., Leavesley, G., Liang, X., Nasonova, O., Noilhan,

J., Oudin, L., Sorooshian, S., Wagener, T., and Wood, E.: Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops, Journal of Hydrology, 320, 3 – 17, doi:http://dx.doi.org/10.1016/j.jhydrol.2005.07.031, http://www.sciencedirect.com/science/article/pii/S002216940500329X, 2006.

Falkenmark, M. and Chapman, T.: Comparative hydrology : an ecological approach to land and water resources, UNESCO, Paris, 1989.

Gao, H., Hrachowitz, M., Schymanski, S., Fenicia, F., Sriwongsitanon, N., and Savenije, H.: Climate controls how ecosystems size the root zone storage capacity at catchment scale, Geophysical Research Letters, doi:10.1002/2014GL061668, http://dx.doi.org/10.1002/2014GL061668, 2014.

Gudmundsson, L., Wagener, T., Tallaksen, L. M., and Engeland, K.: Evaluation of nine large-scale hydrological models with respect to the seasonal runoff climatology in Europe, Water Resources Research, 48, doi:10.1029/2011WR010911, 2012.

Gupta, H. V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., and Andreassian, V.: Large-sample hydrology: a need to balance depth with breadth, Hydrology and Earth System Sciences, 18, 463–477, doi:10.5194/hess-18-463-2014, 2014.

Haddeland, I., Heinke, J., Biemans, H., Eisner, S., Floerke, M., Hanasaki, N., Konzmann, M., Ludwig, F., Masaki, Y., Schewe, J., Stacke, T., Tessler, Z. D., Wada, Y., and Wisser, D.: Global water resources affected by human interventions and climate change, Proceedings of the National Academy of Sciences of the United States of America, 111, 3251–3256, doi:10.1073/pnas.1222475110, 2014.

Holländer, H. M., Blume, T., Bormann, H., Buytaert, W., Chirico, G., Exbrayat, J.-F., Gustafsson, D., Hölzel, H., Kraft, P., Stamm, C., Stoll, S., Blöschl, G., and Flühler, H.: Comparative predictions of discharge from an artificial catchment (Chicken Creek) using sparse data, Hydrology and Earth System Sciences, 13, 2069–2094, doi:10.5194/hess-13-2069-2009, http://www.hydrol-earth-syst-sci.net/13/2069/2009/, 2009.

Holländer, H. M., Bormann, H., Blume, T., Buytaert, W., Chirico, G. B., Exbrayat, J.-F., Gustafsson, D., Hölzel, H., Kraußе, T., Kraft, P., Stoll, S., Blöschl, G., and Flühler, H.: Impact of modellers' decisions on hydrological a priori predictions, Hydrology and Earth System Sciences, 18, 2065–2085, doi:10.5194/hess-18-2065-2014, http://www.hydrol-earth-syst-sci.net/18/2065/2014/, 2014.

Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., Arheimer, B., Blume, T., Clark, M. P., Ehret, U., Fenicia, F., Freer, J. E., Gelfan, A., Gupta, H. V., Hughes, D. A., Hut, R. W., Montanari, A., Pande, S., Tetzlaff, D., Troch, P. A., Uhlenbrook, S., Wagener, T., Winsemius, H. C., Woods, R. A., Zehe, E., and Cudennec, C.: A decade of Predictions in Ungauged Basins (PUB) a review, Hydrological Sciences Journal – Journal des Sciences Hydrologiques, 58, 1198–1255, doi:10.1080/02626667.2013.803183, 2013.

Jakeman, A., Letcher, R., and Norton, J.: Ten iterative steps in development and evaluation of environmental models, Environmental Modelling & Software, 21, 602–614, doi:10.1016/j.envsoft.2006.01.004, 2006.

Jongman, B., Kreibich, H., Apel, H., Barredo, J. I., Bates, P. D., Feyen, L., Gericke, A., Neal, J., Aerts, J. C. J. H., and Ward, P. J.: Comparative flood damage model assessment: towards a European approach, Natural Hazards and Earth System Sciences, 12, 3733–3752, doi:10.5194/nhess-12-3733-2012, 2012.

Koutsoyiannis, D. and Montanari, A.: Negligent killing of scientific concepts: the stationarity case, Hydrological Sciences Journal – Journal des Sciences Hydrologiques, doi:10.1080/02626667.2014.959959, http://dx.doi.org/10.1080/02626667.2014.959959, 2014.

Merz, R., Parajka, J., and Blöschl, G.: Time stability of catchment model parameters: Implications for climate impact analyses, Water Resources Research, 47, doi:10.1029/2010WR009505, http://dx.doi.org/10.1029/2010WR009505, 2011.

Milly, P. C. D., Betancourt, J., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W., Lettenmaier, D. P., and Stouffer, R. J.: Climate change - Stationarity is dead: Whither water management?, Science, 319, 573–574, doi:10.1126/science.1151915, 2008.

Montanari, A. and Koutsoyiannis, D.: A blueprint for process-based modeling of uncertain hydrological systems, Water Resources Research, 48, doi:10.1029/2011WR011412, 2012.

Montanari, A. and Koutsoyiannis, D.: Reply to comment by Grey Nearing on "A blueprint for process-based modeling of uncertain hydrological systems", Water Resources Research, 50, 6264–6268, doi:10.1002/2013WR014987, 2014a.

Montanari, A. and Koutsoyiannis, D.: Modeling and Mitigating Natural Hazards: Stationarity is Immortal!, Water Resources Research, Accepted Article, doi:10.1002/2014WR016092, 2014b.

Montanari, A., Shoemaker, C. A., and van de Giesen, N.: Introduction to special section on Uncertainty Assessment in Surface and Subsurface Hydrology: An overview of issues and challenges, Water Resources Research, 45, doi:doi:10.1029/2009WR008471, 2009.

Montanari, A., Young, G., Savenije, H. H. G., Hughes, D., Wagener, T., Ren, L. L., Koutsoyiannis, D., Cudennec, C., Toth, E., Grimaldi, S., Blöschl, G., Sivapalan, M., Beven, K., Gupta, H., Hipsey, M., Schaefli, B., Arheimer, B., Boegh, E., Schymanski, S. J., Di Baldassarre, G., Yu, B., Hubert, P., Huang, Y., Schumann, A., Post, D. A., Srinivasan, V., Harman, C., Thompson, S., Rogger, M., Viglione, A., McMillan, H., Characklis, G., Pang, Z., and Belyaev, V.: "Panta Rhei-Everything Flows": Change in hydrology and society-The IAHS Scientific Decade 2013-2022, Hydrological Sciences Journal – Journal des Sciences Hydrologiques, 58, 1256–1275, doi:10.1080/02626667.2013.809088, 2013.

Moretti, G. and Montanari, A.: Inferring the flood frequency distribution for an ungauged basin using a spatially distributed rainfall-runoff model, Hydrology and Earth System Sciences, 12, 1141–1152, doi:10.5194/hess-12-1141-2008, http://www.hydrol-earth-syst-sci.net/12/1141/2008/, 2008.

Mullen, K. M., Ardia, D., Gil, D. L., Windover, D., and Cline, J.: DEoptim: An R Package for Global Optimization by Differential Evolution, Journal of Statistical Software, 40, 1–26, 2011.

Parajka, J. and Viglione, A.: TUWmodel: Lumped hydrological model developed at the Vienna University of Technology for education purposes, R package version 0.1-2., http://CRAN.R-project.org/package=TUWmodel), 2012.

Parajka, J., Merz, R., and Blöschl, G.: Estimation of daily potential evapotranspiration for regional water balance modeling in Austria, in: 11th International Poster Day and Institute of Hydrology Open Day "Transport of Water, Chemicals and Energy in the Soil – Crop Canopy – Atmosphere System", pp. 299–306, Slovak Academy of Sciences, Bratislava, 2003.

Parajka, J., Merz, R., and Blöschl, G.: Uncertainty and multiple objective calibration in regional water balance modelling: case study in 320 Austrian catchments, Hydrological Processes, 21, 435–446, doi:10.1002/hyp.6253, 2007.

Parajka, J., Viglione, A., Rogger, M., Salinas, J. L., Sivapalan, M., and Blöschl, G.: Comparative assessment of predictions in ungauged basins - Part 1: Runoff-hydrograph studies, Hydrology and Earth System Sciences, 17, 1783–1795, doi:10.5194/hess-17-1783-2013, 2013.

Parsons, A. J., Abrahams, A. D., and Wainwright, J.: On determining resistance to interrill overland-flow, Water Resources Research, 30, 3515–3521, doi:10.1029/94WR02176, 1994.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/, 2014.

Ramasundaram, V., Grunwald, S., Mangeot, A., Comerford, N., and Bliss, C.: Development of an environmental virtual field laboratory, Computers & Education, 45, 21–34, doi:10.1016/j.compedu.2004.03.002, 2005.

Reed, S., Koren, V., Smith, M., Zhang, Z., Moreda, F., Seo, D., and Participants, D.: Overall distributed model intercomparison project results, Journal of Hydrology, 298, 27–60, doi:10.1016/j.jhydrol.2004.03.031, 2004.

Refsgaard, J., Henriksen, H., Harrar, W., Scholten, H., and Kassahun, A.: Quality assurance in model based water management - review of existing practice and outline of new approaches, Environmental Modelling & Software, 20, 1201–1215, doi:10.1016/j.envsoft.2004.07.006, 2005.

Rinaldo, A., Bertuzzo, E., Mari, L., Righetto, L., Blokesch, M., Gatto, M., Casagrandi, R., Murray, M., Vesenbeckh, S. M., and Rodriguez-Iturbe, I.: Reassessment of the 2010-2011 Haiti cholera outbreak and rainfall-driven multiseason projections, Proceedings of the National Academy of Sciences of the United States of America, 109, 6602–6607, doi:10.1073/pnas.1203333109, 2012.

Schewe, J., Heinke, J., Gerten, D., Haddeland, I., Arnell, N. W., Clark, D. B., Dankers, R., Eisner, S., Fekete, B. M., Colon-Gonzalez, F. J., Gosling, S. N., Kim, H., Liu, X., Masaki, Y., Portmann, F. T., Satoh, Y., Stacke, T., Tang, Q., Wada, Y., Wisser, D., Albrecht, T., Frieler, K., Piontek, F., Warszawski, L., and Kabat, P.: Multimodel assessment of water scarcity under climate change, Proceedings of the National Academy of Sciences of the United States of America, 111, 3245–3250, doi:10.1073/pnas.1222460110, 2014.

UNISDR: From Shared Risk to Shared Value –The Business Case for Disaster Risk Reduction. Global Assessment Report on Disaster Risk Reduction, Tech. rep., UNISDR, Geneva, 2013.

Viglione, A., Borga, M., Balabanis, P., and Blöschl, G.: Barriers to the exchange of hydrometeorological data in Europe Results from a survey and implications for data policy, Journal of Hydrology, 394, 63–77, doi:10.1016/j.jhydrol.2010.03.023, 2010.

Wagener, T., Sivapalan, M., Troch, P. A., and Woods, R.: Catchment Classification and Hydrologic Similarity, Geography Compass, 1, 901–931, doi:10.1111/j.1749-8198.2007.00039.x, 2007.

Wagener, T., Sivapalan, M., Troch, P. A., McGlynn, B. L., Harman, C. J., Gupta, H. V., Kumar, P., Rao, P. S. C., Basu, N. B., and Wilson, J. S.: The future of hydrology: An evolving science for a changing world, Water Resources Research, 46, doi:10.1029/2009WR008906, 2010.

Ward, P., Jongman, B., Weiland, F., Bouwman, A., van Beek, R., Bierkens, M., Ligtvoet, W., and Winsemius, H.: Assessing flood risk at the global scale: model setup, results, and sensitivity, Environmental Research Letters, 8, doi:10.1088/1748-9326/8/4/044019, 2013.

WMO: Intercomparison of models of snowmelt runoff. Operational Hydrology Report No. 23, WMO-No. 646, Tech. rep., WMO, Geneva, 1986.

WMO: Simulated Real-time Intercomparison of Hydrological Models. Operational Hydrology Report No. 38, Tech. rep., WMO, Geneva, 1992.

World Water Assessment Programme: The United Nations World Water Development Report 4: Managing Water under Uncertainty and Risk, Tech. rep., UNESCO, Paris, 2012.

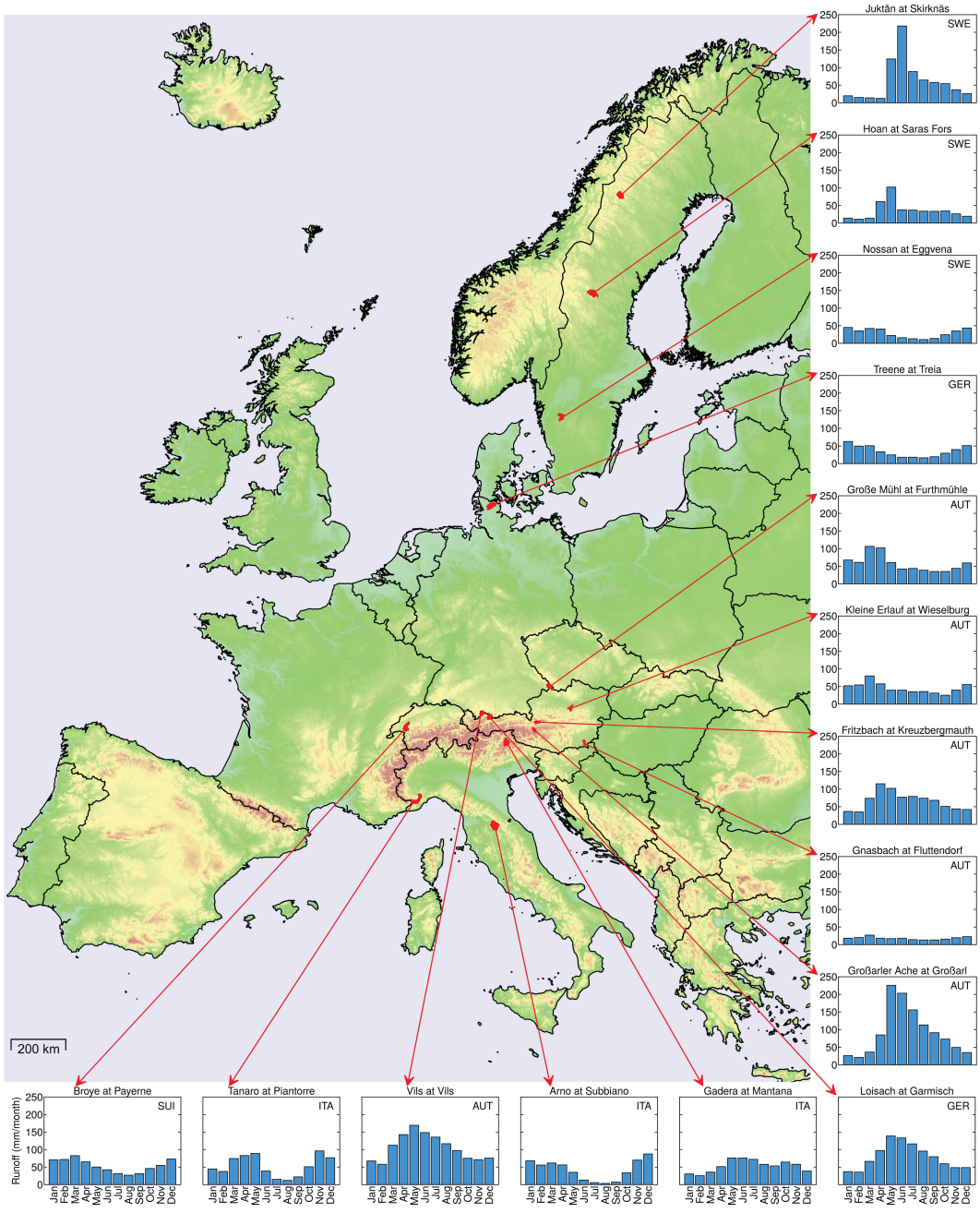

Figure 1. Geographical location and runoff seasonality (average among the observation period listed in Table 1) (mm month^{-1}) for the 15 catchments considered in the first collaborative experiment of the SWITCH-ON Virtual Water-Science Laboratory.

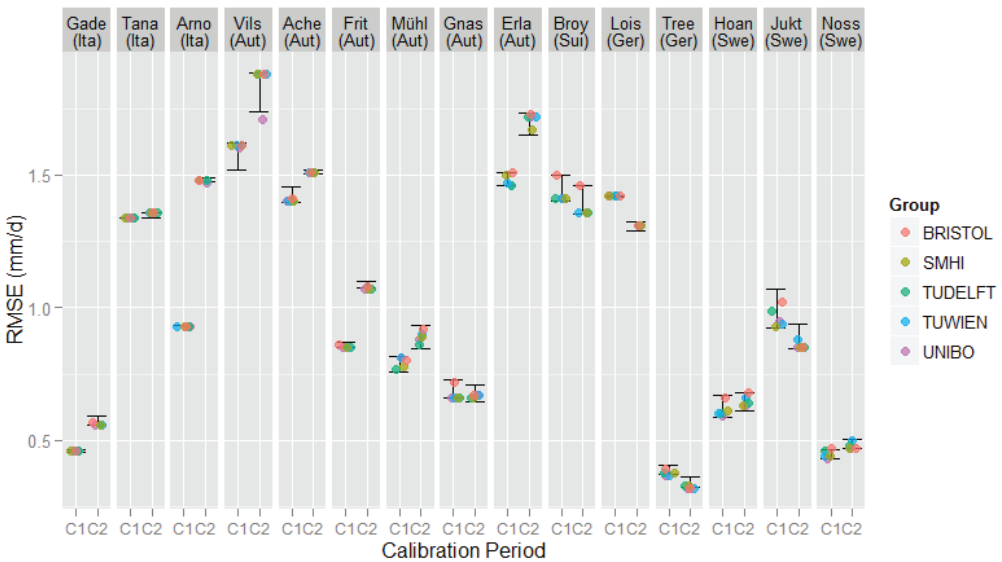

Figure 2. Optimal Root Mean Square Error of runoff (RMSE; square root of the objective function) obtained for calibration period 1 and calibration period 2 by each research group for the 15 catchments. The black bars show the range in optimal performance obtained by a single research group (BRISTOL) from 100 calibration runs initiated from different random seeds.

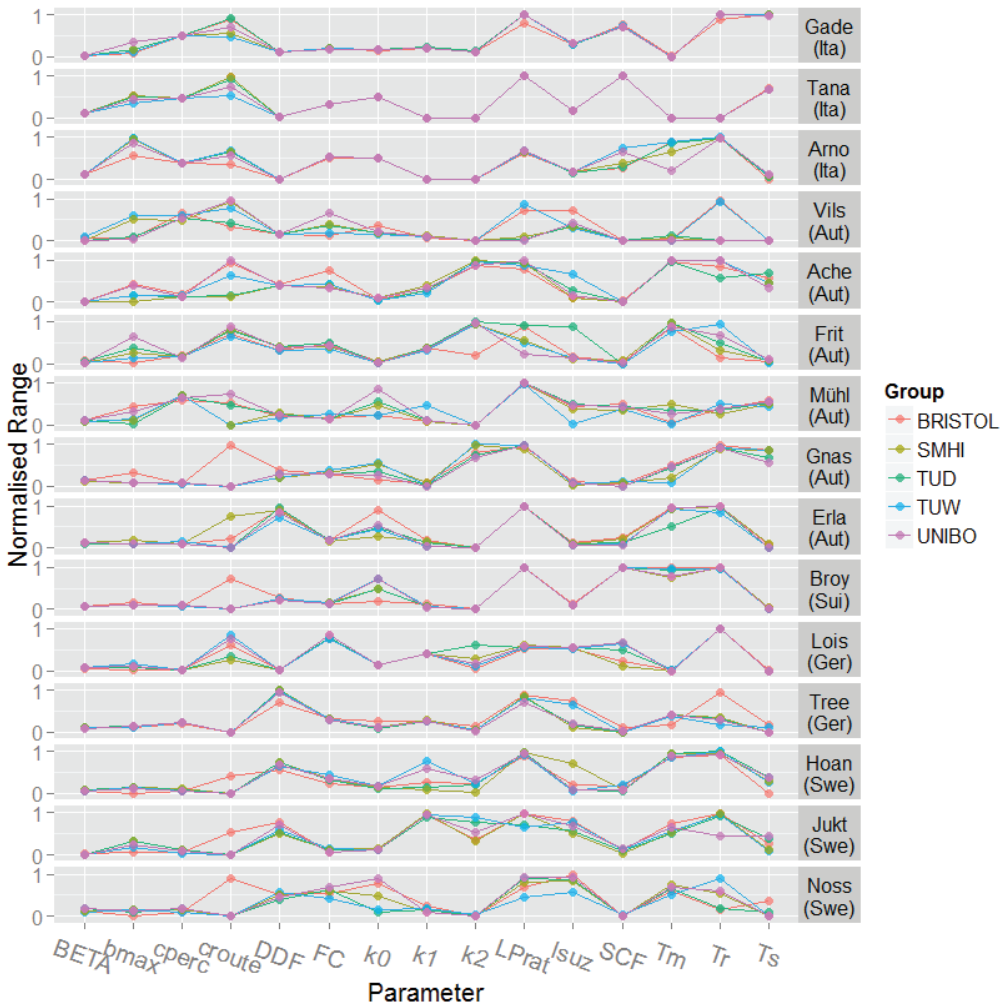

Figure 3. Parallel co-ordinate plots of the optimal parameter set estimates derived from each participant group in each of the 15 catchments for Protocol 1. Model parameters are shown on the x-axis and catchments on the right hand y-axis. The parameters have been scaled to the ranges shown in Table 2.

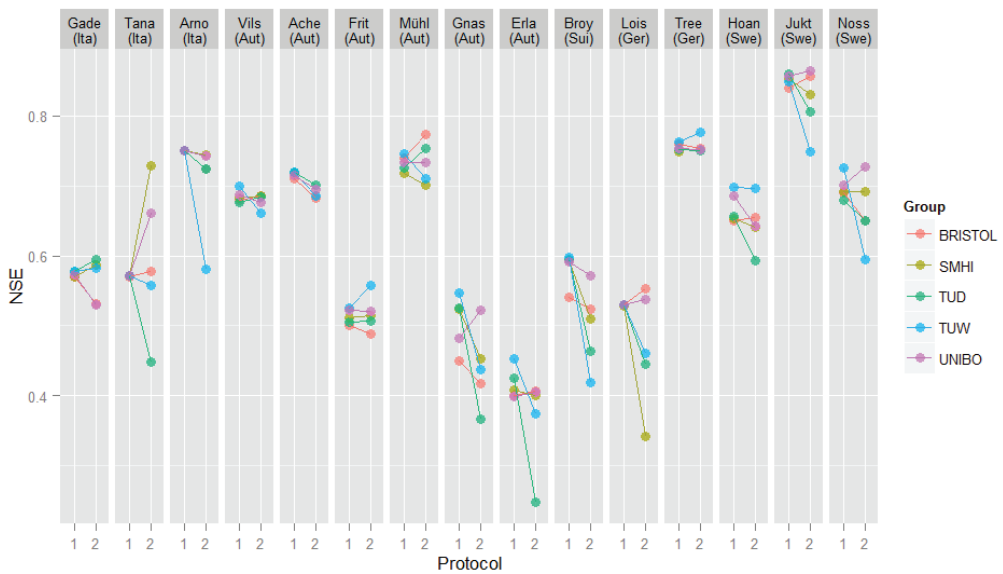

Figure 4. Nash-Sutcliffe Efficiency (NSE) estimated for model validation, obtained by the five research groups, for the 15 catchments, according to Protocol 1 and 2.

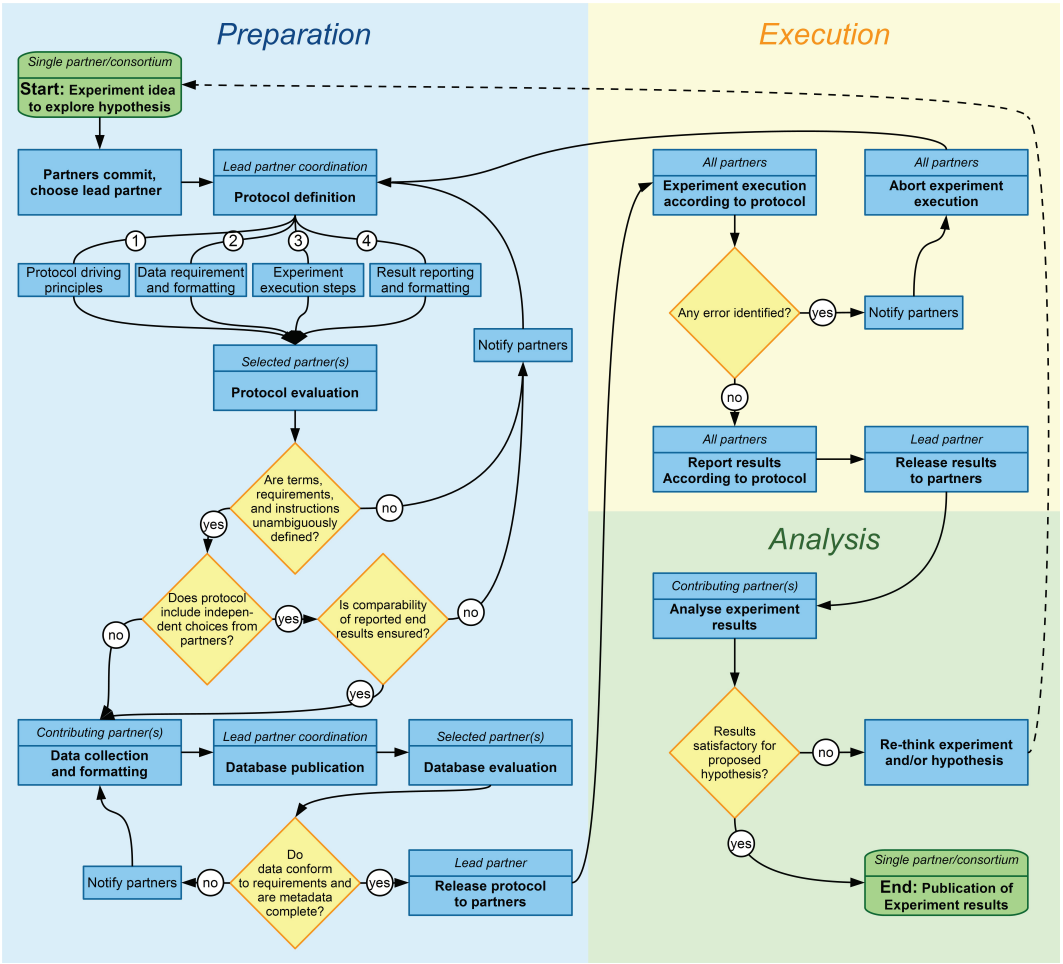

Figure 5. Flowchart of the suggested procedure to establish protocols for collaborative experiments.

Table 1. Summary of the key geographical and hydrological features for the 15 catchments considered in the first collaborative experiment of the SWITCH-ON Virtual Water-Science Laboratory.

Catchment	**Area (km^2)**	**Mean (min, max) elevation (m a.s.l.)**	**Observation period start - end**	**Mean catchment rainfall (mm $year^{-1}$)**	**Mean catchment temperature ($^\circ$ C)**	**Mean observed streamflow per unit area (mm $year^{-1}$)**
Gadera at Mantana (Italy)	394	1844 (811, 3053)	1.1.1990 - 31.12.2009	842	5.2	640
Tanaro at Piantorre (Italy)	500	1067 (340, 2622)	1.1.2000 - 31.12.2012	1022	8.6	692
Arno at Subbiano (Italy)	751	750 (250, 1657)	1.1.1992 - 31.12.2013	1213	11.5	498
Vils at Vils (Austria)	198	1287 (811, 2146)	1.1.1976 - 31.12.2010	1768	5.5	1271
Großarler Ache at Großarl (Austria)	145	1694 (859, 2660)	1.1.1976 - 31.12.2010	1314	3.5	1113
Fritzbach at Kreuzbergmauth (Austria)	155	1169 (615, 2205)	1.1.1976 - 31.12.2010	1263	5.7	799
Große Mühl at Furtmühle (Austria)	253	723 (252, 1099)	1.1.1976 - 31.12.2010	1075	7.2	696
Gnasbach at Fluttendorf (Austria)	119	311 (211, 450)	1.1.1976 - 31.12.2010	746	9.8	218
Kleine Erlauf at Wieselburg (Austria)	168	514 (499, 1391)	1.1.1976 - 31.12.2010	973	8.6	545
Broye at Payerne (Switzerland)	396	714 (391, 1494)	1.1.1965 - 31.12.2009	899	9.1	647
Loisach at Garmisch (Germany)	243	1383 (716, 2783)	1.1.1976 - 31.12.2001	2010	5.8	957
Treene at Treia (Germany)	481	25 (-1, 80)	1.1.1974 - 31.12.2004	905	8.4	413
Hoan at Saras Fors (Sweden)	616	503 (286, 924)	27.4.1988 - 31.12.2012	739	2.3	428
Juktån at Skirknäs (Sweden)	418	756 (483, 1247)	19.5.1980 - 31.12.2012	941	-1.4	739
Nossan at Eggvena (Sweden)	332	168 (91, 277)	10.10.1978 - 31.12.2012	894	6.4	344

Table 2. Main settings of Protocol 1 of the first collaborative experiment of the SWITCH-ON Virtual Water-Science Laboratory.

Component	**Description & Link**		
Model version	TUWmodel, http://cran.r-project.org/web/packages/TUWmodel/index.html		
Input data	Rainfall, temperature and potential evaporation data; catchment area		
Objective function	Mean square error (MSE)		
Optimisation algorithm	DEoptim, http://cran.r-project.org/web/packages/DEoptim/index.html		
Parameter values or ranges		Lower limits	Upper limits
	SCF [-]	0.9	1.5
	DDF [mm $^{\circ}C^{-1}$ day^{-1}]	0.0	5.0
	Tr [$^{\circ}$C]	1.0	3.0
	Ts [$^{\circ}$C]	-3.0	1.0
	Tm [$^{\circ}$C]	-2.0	2.0
	LPrat [-]	0.0	1.0
	FC [mm]	0.0	600.0
	BETA [-]	0.0	20.0
	k0 [day]	0.0	2.0
	k1 [day]	2.0	30.0
	k2 [day]	30.0	250.0
	lsuz [mm]	1.0	100.0
	cperc [mm day^{-1}]	0.0	8.0
	bmax [day]	0.0	30.0
	croute [day^2 mm^{-1}]	0.0	50.0
Calibration and validation periods	Divide the observation period in two subsequent pieces of equal length. First calibrate on the first period and validate on the second and then invert the calibration and validation periods		
Initial warm-up period	365 days for both calibration and validation periods		
Temporal scales of model simulation	Daily		
Additional data used for validation (state variables, other response data)	None		
Uncertainty analysis (Y/N)	None		
Method of uncertainty analysis	None		
Post-calibration evaluation metrics (skills)	MSE, RMSE, NSE, log(NSE), bias, MAE, MALE, VE		

Table 3. Comparison among Protocol 1 and Protocol 2 settings of the first collaborative experiment of the SWITCH-ON Virtual Water-Science Laboratory.

	Protocol 1	**Protocol 2**				
	All research groups	**BRISTOL**	**SMHI**	**TUD**	**TUW**	**UNIBO**
Identification of unreliable data	All data are considered	Runoff coefficient analysis	All data are considered	Visual inspection of unexplained hydrograph peaks	All data are considered	Exclusion of 25% of calibration years with high MSE
Parameter ranges	See Table 2	See Table 2	See Table 2	See Table 2	See Table 2 except for Tr, Ts, Bmax, croute (fixed values)	See Table 2
Optimisation algorithm	Differential Evolution optimisation (DEoptim) – 10 times, 600 iterations	Differential Evolution optimisation (DEoptim) – 10 times, 1000 iterations	Latin hypercube approach	Dynamically Dimensioned Search (DDS) – 10 times, 1000 iterations	Shuffle Complex Evolution (SCE)	Differential Evolution optimisation (DEoptim) – 10 times, 600 iterations
Objective function	Mean Square Error (MSE)	Mean Absolute Error (MAE)	Mean Square Error (MSE)	Kling-Gupta Efficiency (KGE)	Objective function from Merz et al. (2011), Eq. 3	Mean Square Error (MSE)
Warm-up period	1 year for calibration and validation	1 year for calibration and validation	1 year for calibration and validation	1 year for calibration and validation	1 year for calibration and validation	1 year for calibration and validation