Large-sample hydrology: a need to balance depth with breadth

H. V. Gupta¹, C. Perrin², R. Kumar³, G. Blöschl⁴, M. Clark⁵, A. Montanari⁶, and V. Andréassian²

¹Department of Hydrology and Water Resources, The University of Arizona, Tucson AZ, USA
²Iritea, Hydrosystems Research Unit (HBAN), Antony, France
³UFZ-Helmholtz Centre for Environmental Research, Leipzig, Germany
⁴Institute of Hydraulic Engineering and Water Resources Management, Vienna University of Technology, Vienna, Austria
⁵Hydrometeorological Applications Program, Research Applications Laboratory, Boulder, CO, USA
⁶Department DICAM, University of Bologna, Bologna, Italy

Received: 26 June 2013 – Accepted: 27 June 2013 – Published: 12 July 2013

Correspondence to: H. V Gupta (hoshin.gupta@hwr.arizona.edu)

Published by Copernicus Publications on behalf of the European Geosciences Union.
Abstract

A “Holy Grail” of hydrology is to understand catchment processes well enough that models can provide detailed simulations across a variety of hydrologic settings at multiple spatio-temporal scales, and under changing environmental conditions. Clearly, this cannot be achieved only through intensive place-based investigation at a small number of heavily instrumented catchments, or by regionalization methods that do not fully exploit our understanding of hydrology. Here, we discuss the need to actively promote and pursue the use of a “large catchment sample” approach to modeling the rainfall-runoff process, thereby balancing depth with breadth. We examine the history of such investigations, discuss the benefits (improved process understanding resulting in robustness of prediction at ungaged locations and under change), examine some practical challenges to implementation and, finally, provide perspectives on issues that need to be taken into account as we move forward. Ultimately, our objective is to provoke further discussion and participation, and to promote a potentially important theme for the upcoming IAHS Scientific Decade entitled “Panta Rhei”.

1 Introduction

Because almost any model with sufficient free parameters can yield good results when applied to a short sample from a single catchment, effective testing requires that models be tried on many catchments of widely differing characteristics, and that each trial cover a period of many years (Linsley, 1982).

1.1 Motivations for developing large sample hydrology

A “Holy Grail” of hydrological science is to achieve a degree of process understanding that enables construction of models that are capable of providing detailed and physically realistic simulations across a variety of different hydrologic environments (charac-
Large-sample hydrology: a need to balance depth with breadth

H. V Gupta et al.

1. that achieve the three R’s (Reliability, Robustness and Realism),

2. that have greater generality and transposability, and,

3. for which the parameters can be more easily specified from data.

But clearly, this cannot be achieved only through detailed studies at a (relatively) small number of heavily instrumented catchments – although such studies are of critical importance. Nor can it be achieved by simple regionalization methods based primarily in statistical approaches rather than improved understanding of hydrologic behavior. What is needed is to begin taking advantage of the extensive data sets now available (and becoming available) to develop a “large-sample” approach to hydrological investigation (Andréassian et al., 2006).

1.2 The context of current practice

The context of much current hydrological practice is a focus on depth rather than breadth (with notable exceptions, as mentioned later), wherein detailed process investigation and model development/refinement are classically conducted at only one or a limited number of catchments. The typical goal is to either (a) learn more about a
specific catchment by improving upon some prior concept, or (b) establish a basis for prediction and decision-making at that specific location. This might be called “place-based learning”. By contrast, the scientific aspiration is to “generalize” from the study of specific cases (by comparing what does and does not work across locations), so that we can discover and establish general hydrological principles (and models that embody them), thereby advancing hydrological understanding. This drive towards generalization is a key motivation underlying the scientific approach.

Over the past several decades, the hydrology community has supported this approach through development of a number of “generic” model codes (specific realizations of perceptual-conceptual/mathematical/computational development), or model development frameworks (multi-hypothesis environments), that can be applied to a given catchment study (see discussion in Clark et al., 2011; Gupta et al., 2012). With these tools, once a specific “off-the shelf” model structure has been selected, it remains necessary only to specify values for the parameters. If the model performance is deemed inadequate, attempts can be made to “diagnose” model deficiencies (Fig. 1) and find ways to “correct/improve” the model/hypothesis (Gupta et al., 2008; McMillan et al., 2011; Euser et al., 2013).

However, attempts at such generalization are fraught with difficulties. Faced with the tremendous geo-eco-hydro-climatic variability of environments across the world, attempts based on generic models are typically complicated by considerable noise in the individual results (e.g., Oudin et al., 2010; Savenije, 2009) – these being due to unresolved model identification issues arising from a combination of inadequate data (insufficient information), data noise, model structural inadequacy, and weak model identification techniques (e.g., see Gupta et al., 1998, 2008, 2012). This is particularly true as both model realism (process complexity) and spatio-temporal resolution are increased, in search of improved accuracy and detail.

On the one hand, such difficulties have contributed to the counter notion of “uniqueness of place” (Beven, 2000), and that the model structure should adapt to reflect spatial differences in the dominant hydrologic processes. On the other hand, global-scale
hydrological studies typically use only a single conceptual structure to represent all locations around the world, while attempting to account for the place-to-place differences entirely through the specification of the model parameters (representing differences in soil and vegetation properties) while largely ignoring the spatial variability in dominant hydrological processes (Dirmeyer et al., 2006).

1.3 Purpose of the paper

The purpose of this paper is to discuss the need for a greater focus on “Large Catchment Sample” type studies in hydrology, to complement the approach of intensive place-based investigation. We do this by providing some historical perspective, motivating the need for such studies, illuminating some of the challenges, and examining issues related to the design and implementation of such studies. Of course, we do not intend to imply that there should be a reduction in efforts to study individual catchments in detail; both kinds of investigations (as well as ones in the middle ground) are necessary. Ultimately, our objective is to provoke further discussion and participation, and to promote what could be an important theme for the upcoming IAHS Scientific decade (“Panta Rhei”; Montanari et al., 2013)

2 Previous studies: what have we learned?

2.1 Early attempts at large sample studies

The issue of using data sets having large numbers of catchments for hydrological investigation is not new. Early attempts to apply models to large data sets go back more than thirty years, although the common practice at that time was to develop models for a single catchment only. There were practical reasons for the latter – including limitations in data access, computing requirements or the ability to collaborate efficiently – but there was also a common belief that models could not be readily applied outside of
the study area for which they were initially developed. Consequently, it was difficult to really know the respective merits and usefulness of any existing model.

In 1967, the World Meteorological Organization (WMO) launched an initiative to develop an inventory of models, along with advice to users regarding their accuracy under various hydro-climatic conditions. Not surprisingly, it was quickly deemed useful to carry out an actual model inter-comparison study. Subsequently, in 1973, ten simulation models from seven countries were applied to a set of six catchments (from the USA, USSR, Australia, Japan, Cameroon and Thailand) that represent a variety of hydro-climatic conditions (WMO, 1975; Sittner, 1976). Although the investigation did not arrive at definitive conclusions regarding the merits of the models tested, it drew attention to the need for a deeper and wider evaluation of models. At that time, a small number of other model inter-comparison studies involving more than one catchment were also carried out; for example see Mein and Brown (1978) who compared three models on four catchments in Australia, and James (1972); Egbuniwe and Todd (1976); and Magette et al. (1976) for their work on the Stanford model using 2 to 16 catchments.

A few years later Linsley (1982) listed “generality” as one of four main properties desirable in hydrological models (along with accuracy, applicability and ease-of-use). Linsley argued that it was necessary to break out of the habitual practice of developing a different model for each catchment, because that “eliminates the opportunity for learning what comes with repeated applications of the same model”. He also suggested the necessity for “extensive testing” of new models so that models that do not prove to be sufficiently general can be eliminated, saying:

because almost any model with sufficient free parameters can yield good results when applied to a short sample from a single catchment, effective testing requires that models be tried on many catchments of widely differing characteristics, and that each trial cover a period of many years.

Linsley also stressed the usefulness of such large scale applications from the perspective of parameter estimation, saying:
the application of a model to many catchments results in many sets of parameters which can conceivably serve as a basis for objective determination of parameters from physical characteristics of the catchments.

Along the same line of thought, Klemeš (1986) proposed a formal four-level testing scheme to evaluate the transposability of a model in time and space. Despite the demanding nature of such testing, Klemeš regarded this scheme as a minimum requirement, and stated that the use of more test basins [...] would increase the credibility standing of a model, and an accumulation of test results may lead to meaningful generalizations.

Bergström (1991) agreed, stating growing confidence in hydrological modeling can be obtained by applying the model under a span of different geographical, climatological and geological conditions.

These ideas, expressed by leading hydrologists, set the basis for studies involving large catchments samples, that have now begun to be more common.

2.2 Brief review of the relevant literature

While several modeling studies during the 1980’s used more than one catchment (e.g., Weeks and Hebbert, 1980; Naef, 1981; Pirt and Bramley, 1985; Loague and Freeze, 1985; Weeks and Ashkanasy, 1985; WMO, 1986; Srikanthan and Goodspeed, 1988), the actual emergence of large sample studies arguably occurred in the early 1990s, coinciding with a progressive increase in availability of computing power, and using time series that are sufficiently long to enable robust assessments.

To illustrate the growing interest in large sample studies during and since that time, a list of 84 published rainfall-runoff modeling studies that have used more than 30
catchments is presented in the Supplement. The sample sizes ranged from 30 to 1508, with a median of 140 catchments. The earliest hydrologists to do so were mainly from Australia, France and Belgium, followed later by others from the UK, Austria and USA, etc. (see also the comprehensive PUB synthesis report by Blöschl et al., 2013).

The list indicates that samples generally included catchments from a variety of physical, climatic and hydrological conditions. While some studies were limited to national data sets, others included catchments from several countries (although typically less than five). The studies focused on a range of spatial and temporal scales: catchment areas ranged from 1 to 130 000 km$^2$ and models were run at hourly, daily, monthly, annual and inter-annual time steps. The study goals included a variety of purposes, most commonly being related to:

1. model development, application and comparison,
2. estimation of model parameters (calibration and regionalization techniques),
3. evaluation procedures and criteria,
4. sensitivity and uncertainty analysis,
5. impact studies.

Not surprisingly, most studies were carried out using “conceptual-type rainfall-runoff” (CRR) models, probably due to their being easier to implement on large samples than so called “physically-based” models because of lower data and computing requirements. In general, these studies used large catchment samples for three reasons:

1. to achieve conclusions that were more general than could be achieved using a single catchment (e.g., about the relative merits of various methods),
2. to define the range of applicability, or expected level of efficiency, of methods/models, or,
3. to ensure sufficient information to enable statistically significant relationships to be established (e.g., between catchment descriptors and model parameters in regionalization studies).

In addition to catchment samples constructed by individual teams, several groups and institutions collaborated to compile large sample catchment data sets with the goal of facilitating and supporting collective efforts involving several teams. As mentioned above, the World Meteorological Organization (WMO, 1975, 1986, 1992) played a pioneering role by sponsoring several inter-comparison studies in which various teams were invited to work on the same sets of catchments. Subsequently, the Model Parameter Experiment (MOPEX; see e.g., Schaake et al., 2001; Chahinian et al., 2006; Duan et al., 2006; Schaake and Duan, 2006) organized modeling experiments using a large sample of catchment data sets from the USA and France. Most recently, the Distributed Model Intercomparison Projects (DMIP-I and DMIP-II, see Smith et al., 2004, 2012) sponsored by the US National Weather Service facilitated a comparative assessment of spatially-distributed hydrological models (and associated parameter estimation strategies) by several teams, and provided a comprehensive data set consisting of a significant number of catchments in the US.

Likewise other large-scale comparative studies such as the North American Land Data Assimilation System (NLDAS-I and NLDAS-II; http://ldas.gsfc.nasa.gov/index.php), the Project for Intercomparison of Land-surface Parameterization Schemes (PILPS; http://pilps.mq.edu.au/fileadmin/pilps/Elsevier.html), the Rhone-Aggregation Land Surface Scheme Intercomparison Project (http://www.cnrm.meteo.fr/isbadoc/projects/rhoneagg/), and The Global Energy and Water Cycle Exchanges Project (GEWEX; http://www.gewex.org/), have compiled a large array of hydro-meteorological datasets. Of course these datasets have been compiled for running regional scale land surface models at a relatively large spatial resolution (1/8° or larger) compared to that typically used in catchment hydrology, but such datasets could provide a useful starting point for hydrological investigations over large and diverse spatial domains.
2.3 Promising directions

In the context of the recent PUB decade, one major achievement has been the use of large data sets to facilitate comparative studies that could arrive at more general conclusions than those provided by studies based on single or limited numbers of catchments (Blöschl et al., 2013). However, it is interesting that comparative studies of regionalization approaches have resulted in strikingly different conclusions. One reason for such disagreement, though not the only one (see Oudin et al., 2008) is the likely use of insufficiently large catchment samples, which would cause the conclusions to be overly dependent on differences in the characteristics represented by each data set.

Nonetheless, the large sample rainfall-runoff modeling studies mentioned above have pointed towards some promising directions for the further development of large sample hydrology, as discussed (for example) by Merz and Blöschl (2006) and Andréassian et al. (2006, 2009). Importantly, as the practice of routinely conducting such investigations becomes commonplace, it will open the way towards defining benchmarking procedures (Seibert, 2001; Schaefli and Gupta 2007; Parajka et al., 2013), and thereby enabling the routine use of reference datasets from a wide variety of catchments for testing new models and/or methods. In this regard, it will be helpful to shift the focus of model evaluation away from “data fitting” towards an emphasis on reproduction of diagnostic signatures (Fig. 2) – a move away from hydrograph mimicry towards model fidelity (Vogel and Sankarasubramanian, 2003; Gupta et al., 2008; Blöschl et al., 2008; Martinez and Gupta, 2011; Clark et al., 2011; Koster and Mahanama, 2012; etc.). The idea here is to make it harder to “win the game through calibration” (in the sense of model tuning) and to instead seek answers that are correct for the right reasons (Kirchner, 2006).
3 General benefits of large sample hydrology

There are at least four clear benefits to modeling studies that attempt to work with data from large numbers of catchments (see Fig. 3).

3.1 Improved understanding

First, large sample studies provide better opportunities for “learning” (improving hydrological science; see Ehret et al., 2013) by facilitating more rigorous testing of competing model structures and their component hypothesis (see Clark et al., 2011), and enabling better diagnosis of their limitations, range of applicability and capabilities for extrapolation (see e.g., Gupta et al., 2008; Martinez and Gupta, 2011; Coron et al., 2012). In this regard, it is important to note that, because the meaning/function of a model parameter is intimately determined by the choice of model ‘parameterizations’ (the hypotheses about functional forms of catchment sub processes, and about how different sub-processes combine to provide the system-scale responses), progress towards the successful identification of model “structures” will be a necessary pre-condition to the successful identification of associated model “parameters”. Such progress will, then, help us move beyond the primary focus of PUB on “performance and predictability”, towards a better understanding the underlying causes, and ultimately towards a better understanding of how best to represent and tackle “predictions under change” (Montanari et al., 2013).

In this regard, an important step will be to shift the focus away from site-specific parameter estimation towards the development of regionalization methods; e.g., theoretical or empirical transfer functions that establish relationships between the observable hydro-geo-climatic characteristics of catchments and the model structures and parameters (Abdulla and Lettenmaier, 1997; Koren et al., 2000; Hundecha and Bardossy, 2004; Pokhrel et al., 2008; Samaniego et al., 2010; Kumar et al., 2013).
3.2 Robustness of generalizations

The second, obvious, benefit is the possibility of bringing methods of statistical analysis to bear, so that statistical robustness can be achieved, and degrees of confidence in the results/findings can be established (Mathevet et al., 2006). To be statistically robust, scientific results generally require sufficiently large samples so that conclusions can be drawn about central tendencies and degree of variability, based on which “outliers” from the norm (relatively unusual cases) can be more easily identified and targeted for special attention to understand why they differ (Andréassian et al., 2010). With robust parameter estimates, comparisons of catchments with a view to making predictions at ungaged, or under changed, conditions are much more meaningful (e.g., Parajka et al., 2005; Merz et al., 2011).

In the process of doing this, large samples can also help to reduce the impact of data errors. Given the large number of error sources in hydrological data, it has become common for modelers to invoke data errors as a primary cause for “lack of fit” between the modeled results and observed data. However, unless a way can be found to deal with the inevitable problem of data errors, the modeling exercise unavoidably leads to a dead-end, since if the data-reference cannot be trusted, there is no chance of being able to establish confidence in our models.

And perhaps worse than arriving at a dead-end is the problem of circular reasoning that can arise from the practice of “discarding” catchments that are deemed to be “doubtful” when viewed through the perspective of the failure of a model to satisfactorily simulate their responses (Le Moine et al., 2007; Boldetti et al., 2010). Taken to its logical extreme, such a filtering approach will necessarily tend towards the conclusion that the current model structure is adequate. One important way to counter such tendencies is to evaluate models in the context of sufficiently large catchment sets so that the hypothesis that data are error-free becomes unnecessary.
3.3 Classification, regionalization and model transfer

A third advantage of working with large numbers of catchments is that it can support and facilitate the development of a catchment classification system that provides insights into hydrological behavior – for this, it must be stable, clear, unique and simple - and, as mentioned above, make possible the regionalization of (a) foundational model structures, (b) model structural variations, and (c) associated parameter values. These goals have been the key motivation of the PUB initiative – that of transferring understanding to “ungauged” locations (Sivapalan, 2003; Götzinger and Bardossy, 2007; Oudin et al., 2010). To date, the primary focus of regionalization has been on the parameter values associated with a particular, pre-selected, model structure (Blöschl et al., 2013). For example, an extreme case of this is the use of a single dominant-process representation for the land-surface water-energy-carbon response component in Regional and General Circulation Models used for weather forecasting and climate studies (Dirmeyer et al., 2006), wherein parameters are universally specified based on soil and vegetation type.

Arguably, the dominance of hydrological processes varies significantly with climatic region, geology and other factors (such as spatio-temporal scale). Therefore, a satisfactory classification system of hydrological process dominance is needed for development of hydrological science (see e.g., work by Winter, 2001; McDonnell and Woods, 2004; Wagener et al., 2007; among others). Clearly, this cannot be achieved without recourse to studies involving spatially extensive data sets that are representative of the different kinds of catchment types worldwide. With such data sets, it becomes possible to develop model structures that provide realistic simulations of hydrologic behavior across a range of hydro-climate regimes, based perhaps on (a) a progressive modification of universal model types to better explain the variability of hydrological responses seen in data sets, or (b) developing a new hydrologic theory/model capable of reliable predictions everywhere on the planet.
Ultimately, beyond improvements in hydrological understanding, this also paves the way for improved model transfer into operational use by engineers and water managers in practical applications. Reflecting the tendency to universal models in catchment science, it is common for operational agencies (such as the US National Weather Service) to employ universal models for applications such as flood forecasting. Prior model testing on large and diverse catchment sets is, therefore, a logical prerequisite to ease model transfer to operational applications. Expressing this point of view, Bergström (1991) mentioned that

the large number of applications [of the HBV and PULSE models] have gradually built up our confidence in the use of these models to a degree where we can continue our operational applications and accept the models as the foundation for further model development.

3.4 Estimation of uncertainty

A fourth, and particularly important, advantage of working with large numbers of catchments is that it supports and facilitates a better understanding of how much uncertainty can be expected in model predictions, given available knowledge, particularly when addressing the problem of prediction in ungauged basins (Andréassian et al., 2007). It does this by making it possible to achieve a statistical regionalization of uncertainty estimates (Bourgin et al., 2013). By establishing model performance as a function of catchment classification, model structures, and model parameter values, it should be possible to estimate, a priori, how much prediction uncertainty can be expected at an arbitrary location. Alternatively, such uncertainty can be estimated using methods of spatial statistics (eg. Skøien and Blöschl, 2007) or by regional blind testing (Blöschl et al., 2013). Ultimately, this implements a process-based approach to prediction, while exploiting the power of statistics.
4 Practical implementation, challenges & recommendations

4.1 Availability of data sets having large numbers of catchments

Obviously, widespread availability of large catchment sample data sets is a key requirement for further progress in large sample hydrology. However, the attempt to make (what are currently “local”) data sets widely available to the global hydrological community may run into some critical issues, such as related to economics and ownership, depending partly on differences in the legal status of data in various originating countries. Fortunately, hydrology-related data in the USA are mainly in the public domain and easily accessible to scientists worldwide. However, in Europe and other parts of the world, there are economic barriers to the exchange of data (e.g., precipitation and geospatial data; see Viglione et al., 2010), which inhibits the wider spread of existing “national” data sets (see Fig. 4). While discharge data are increasingly more freely available in many countries, climate data are often less easily accessible for hydrologists.

Consequently, it will likely be necessary for the community-at-large to develop, and vigorously promote, specific policies designed to make hydrology-related data more easily and widely available (Beniston et al., 2012); this will require deliberate efforts by a spectrum of organizations, including the International Association of Hydrologic Sciences, and the Predictions Under Change initiative).

Nonetheless, there do exist a few large catchment sample data sets that are freely available to the community. These include a set of 438 catchments in the US, made available by NOAA (http://www.nws.noaa.gov/oh/mopex/) within the context of the MOPEX initiative. A related, and more detailed data set (although for a smaller number of catchments) was made available in the context of the DMIP experiments (http://www.nws.noaa.gov/oh/hrl/dmip/), and is particularly useful for evaluating spatially distributed models. Within PUB, the Top-Down Modeling Working Group (http://tdwg.catchment.org/) has made available a data set consisting of 60 catchments...
in the UK. Another data set consisting of 278 Australian catchments is also freely accessible (Peel et al., 2000).

In addition to these, a number of related data sets may be useful to large sample hydrological investigations, including various global-scale atmospheric data (Van der Ent and Savenije, 2011), regional flux tower data (Williams et al., 2012), and data sets compiled by multi-institutional collaborative projects like PILPS, NLDAS-I & II, and the Global Land-Atmosphere Coupling Experiment (GLACE-I & II; http://gmao.gsfc.nasa.gov/research/GLACE-2/), which have been used to improve our understanding of patterns in land atmosphere interactions.

There is however an outstanding problem. It is well known that while the catchment hydrology datasets described above are extensive in terms of the number and diversity of basins, they are usually limited primarily to observations of system inputs (precipitation, and temperature or potential evapotranspiration) and the system-scale response (streamflow, usually only at the catchment outlet). In short, typical catchment data sets do not have sufficient multivariate information to enable the evaluation of sub-components in a model.

This is important because, as pointed out by Kuczera and Franks (2002), evaluating model sub-components is a major challenge [...] that must be vigorously pursued if conceptual catchment modeling is to avoid degenerating into a sterile curve-fitting exercise.

Similarly, Gupta et al. (2008), among others, have discussed the limitations of traditional aggregate metrics (e.g., the sum of squared differences between model simulations and observations), and highlighted the need for incisive model diagnostics that can scrutinize different sub-components of a model. Coordinated community efforts to assemble and unify the currently diverse array of highly heterogeneous research datasets are, therefore, critically necessary. Efforts such as the CUAHSI Hydrologic Information System in the USA (http://his.cuahsi.org/) are an important step in the right direction, and will ultimately support the pursuit of process-based land surface hydrology model development and evaluation endeavors.
4.2 Data quality

Activities to promote increased availability of large catchment data sets must, of course, take into consideration the important issue of data quality, and insist upon protocols to ensure that adequate meta-data are provided. Certainly, it will be generally more difficult to apply the same kind of rigorous quality checks that can be imposed when compiling smaller catchment samples and, for example, even something as basic as visual inspection could be too time-consuming when compiling samples of hundreds of catchments. In time, however, it will be necessary to develop progressively more rigorous procedures for improved automatic data screening so as to facilitate more confident use of large data sets. From a practical point of view, this will initially encounter the “chicken and egg” problem wherein the use of hydrological and/or statistical models (based in hydrological principles such as mass balance, etc.) to provide automated testing of data sets results in a somewhat circular kind of reasoning – models used to evaluate data sets, which are then used to make inferences about model structural hypotheses!

However, with time and perseverance, not to mention a healthy dose of careful attention to detail, it should ultimately be possible to make progress on this front. In this regard, Chapter 3 of the PUB assessment report (Blöschl et al., 2013) has proposed a “hierarchical data collection approach” that could exploit the trade-offs between scale, data availability and costs. In this approach, global data sets are understood to provide more generalised information at lower cost to the individual user, whereas dedicated local measurements are understood to provide detailed information at high cost over small spatial scales. Accordingly, one can begin at the global scale and, depending on resources available, zoom in to different levels of details at consecutively finer spatial and temporal scales, during which a hierarchy of controls from climate to local catchment and anthropogenic effects will become evident, and can therefore be deciphered.
4.3 Reporting & sharing protocols for data and models

A longstanding issue that continues to retard progress in hydrological science is the need for coherence in the way data and models are reported, stored and shared; this is currently done in a number of different ways depending on their nature and purpose for which they were collected or developed. Data deemed to be of wider interest to the hydrologic community are now typically “published” in journals through “Data and Analysis Notes”, along with meta-data (details regarding how the data were collected, and what can be done with them). Data collected by Hydro-meteorological Services, and other public agencies, are increasingly made easily accessible via the Internet, although typically with much less meta-information (Viglione et al., 2010).

While considerable attention has been given to protocols for documenting and sharing data during the past several decades (Jones et al., 1979; Goodall et al., 2008; Viglione et al., 2010; see also https://www.wmo.int/pages/prog/gcos/Publications/gcos-96.pdf), the procedures for documenting and sharing models (computer codes) continue to remain extremely ad-hoc. At the same time, protocols for the reporting of model “performance” are largely non-existent as noted by Gupta et al. (2008),

As a community, we have fallen into reliance on measures and procedures for model performance evaluation that say little more than how good or bad the model to-data comparison is in some “average” sense.

Consistent reporting of sets of more informative (than Mean Squared Error or Nash-Sutcliffe Efficiency) and properly benchmarked (see Mathevet et al., 2006; Perrin et al., 2006; Schaefli and Gupta, 2007) measures of model performance are necessary to better facilitate the generalization of findings from individual case studies.

The important thing to keep in mind is that the primary purpose of reporting is to make the information useful to the recipient (reader, user). In the comparative assessment of Blöschl et al. (2013), the investigators reported that inconsistency in use and reporting of model performance was a serious problem that hampered their investiga-
4.4 Identification of model structures and parameters

The growing availability of geo-spatial datasets (including both meteorological variables and land surface characteristics), and increases in computational power, have contributed to significant progress in the development and application of local, regional and continental-scale spatially distributed hydrologic models that simulate the distribution and evolution of hydrological processes at relatively high spatial resolution. Consequently, such models are able to simulate hydrological fluxes (such as evapotranspiration, aquifer recharge, overland flow, and streamflow, etc.) and state variables (such as soil moisture) at internal points within the catchment, at which locations observational data about those variables will not typically be available (e.g., see Carpenter and Georgakakos, 2004; Ivanov et al., 2004; Reed et al., 2004; Koren et al., 2004, Smith et al., 2012); i.e., such points are effectively “ungaged”.

While such models offer the promise of detailed support for water resource management, their reliable application is limited by several problems including poor parameter identifiability and consequent inability to transfer parameter values across locations (e.g., see Beven, 1989, 2002; Grayson et al., 1992; Kirchner, 2006; Samaniego et al., 2010; Andréassian et al., 2012). Because observational data about the spatially distributed values of the model parameters is not typically available, parameter estimation must rely on indirect procedures (calibration). However, the high dimensionality of the parameter search space increases the chance of over-fitting during optimization, and contributes to increased predictive uncertainty (Beven and Freer, 2001; Doherty and Johnston, 2003; Pokhrel et al., 2008). Furthermore, such an approach to parameter inference becomes impossible for poorly gauged or ungauged basins.

As is true for the regionalization of parameter values for catchment scale spatially lumped models, such problems can be tackled by recognizing that the parameter values at specific locations within spatial fields are not independent quantities, being that...
they are somehow linked to observable basin physical characteristics (e.g., soil texture, vegetation, topography, etc.) that themselves display strong patterns of correlation and organization in space (e.g., see Abdulla and Lettenmaier, 1997; Fernandez et al., 2000; Hundecha and Bardossy, 2004; Pokhrel et al., 2008; Hundecha et al., 2008; Samaniego et al., 2010; Kumar et al., 2013; among others).

A number of studies have, therefore, pursued the approach of developing regional relationships, based either on hydrological reasoning or empirical assumptions about the functional forms, that can map from observable catchment characteristics (at the sub-basin scale) to model parameters; these are sometimes called a priori parameter estimates (e.g., Koren, 2000; Leavesley et al., 2003; Blöschl, 2005). Implicitly, such regional relationships are based on the idea of similarity, as they assume that sub-catchment scale model units that share similar catchment characteristics must be represented by similar values for the model parameters (this approach is also the primary strategy used in other fields of Earth Science, such as for specifying spatially distributed parameter values for land surface modules used in regional and global atmospheric circulation models). An important benefit of this approach is that it “regularizes” the optimization problem, providing constraints that greatly reduce the degrees of freedom (number of unknowns to be inferred) to a relatively small number of regional transfer function coefficients (referred to in the literature as global-, super-, or hyper-parameters; see Pokhrel et al., 2008; Samaniego et al., 2010).

To properly calibrate the values of the regional transfer function coefficients it is, of course, necessary to provide data for (and then run the spatially distributed hydrological models at) a large number of catchments, and across diverse hydro-climatic conditions, so that hypothesis tests about the structures of the regional transfer functions can be conducted and statistically robust results can be obtained. Note that the nature of the hydrological modeling problem is modified in an interesting way, as one must now evaluate an augmented model hypothesis consisting of both (a) the catchment model structure relating system inputs to state variables and outputs, and (b) the regional transfer function structure and parameter values relating catchment properties
to model parameters (for the assumed catchment model structure). Overall, although increasing the complexity of the hypothesis under investigation, this approach provides the statistical advantage of: (i) accounting for random variability in catchment properties (Kling and Gupta, 2009) that tends to diminish their correlation with lumped catchment scale properties; (ii) reducing the biasing effects of random noise in the data due to the damping effects of the larger sample sizes; and (iii) improving identifiability due to the increased diversity of hydro-climatic conditions represented. In fact, recent work on a “multi-scale parameter regionalization approach” (Samaniego et al., 2010; Kumar et al., 2013) strongly suggests that the approach is robust, and can facilitate the transfer of model hypotheses across spatial domains. The approach can be further strengthened by constraining the model using information about the hydrological dynamics such as provided by soil moisture and snow cover (e.g., Parajka and Blöschl, 2008; Parajka et al., 2009).

Ultimately, for models to be demonstrably robust, they must be able to pass the kind of “crash testing” proposed by Linsley (1982), Klemeš (1986), Andréassian et al. (2009), and Coron et al. (2012), among others (see Fig. 5). This can only be properly done using large sample catchment data sets.

5 Perspectives

5.1 Overcoming barriers to sharing data

To state the obvious, large sample hydrology requires large samples of relevant data sets. Most current studies of large sample catchment hydrology have been focused on regional or national scales (e.g., Parajka et al., 2005, 2007; Oudin et al., 2008, 2010; Kumar et al., 2013) and there is a need to extend these to global scale. As computer power and data storage capabilities continue to increase, one might expect that data exchange will also increase. However, anecdotal evidence suggests that barriers to the free exchange of hydrological data have been growing. To better understand the rea-
sons for barriers to data exchange, Viglione et al. (2010) conducted a survey in which data providers and users from 32 European countries were polled. They reported that the main reasons for such barriers are economic, with the public institutions typically responsible for data collection and administration facing growing financial pressure due to reductions in government funding (Freebairn and Zillman, 2002), and therefore attempting to partially recover costs by selling their data. Other barriers include conflicts of interest (as when the providers sell products derived from the data), and the desire to reduce misuse of data due to inappropriate redistribution by users.

Nonetheless, the PUB Synthesis effort (Blöschl et al., 2013), driven largely by the desire to conduct comparative studies, demonstrated clearly that it is indeed possible to compile and jointly analyse large data sets. By pointing beyond benchmarking, and to the possibility of actually learning new things, we hope that hydrologists will be convinced of the need to find ways to overcome the economic and legal constraints to building and sharing reference data sets.

5.2 Linking large sample studies with process hydrology

Traditionally, large sample studies have focused mainly on statistical analysis; e.g., to develop regression relationships for flood regionalisation or to estimate and transfer model parameters to ungaged basins. There is a need to move beyond this, and use large sample data sets to better understand local scale processes. The tendency to universally apply a fixed set of assumptions regarding driving mechanisms and process structure can sometimes miss the more important processes in a particular catchment (Savenije, 2009). For example, the DMIP project (Reed et al., 2004) found that all the models tested on the Blue River performed poorly due to not taking into consideration the unique, complex, hydrogeology of the basin (Halihan et al., 2009). Similarly, Blöschl et al. (2013) report that the Elkhart catchment in Indiana is different from other catchments in the region, having much higher baseflow contributions due to large wetlands and lakes, thick glacial sediments, and complex topography (USACE, 2010), which detain and release water slowly over long periods of time.
Clearly, any large sample investigation of catchment hydrology should investigate the diversity of hydrological processes represented therein, so that the models structures used are representative and appropriate, and any good performance achieved is “for the right reasons” (Kirchner, 2006; Martinez and Gupta, 2010, 2011). It is therefore important, that emphasis be placed on investigating and demonstrating causal processes (Merz and Blöschl, 2008a, b), by taking advantage of formal methods being developed to capture the diversity of information at the regional scale (e.g., Viglione et al., 2013).

One obvious approach would be to begin with a highly complex representation and progressively simplify the structure applied to each catchment to achieve a justifiably parsimonious representation. However, as expressed by Bergström (1991),

Going from complex to simpler model structures requires an open mind, because it is frustrating to have to abandon seemingly elegant concepts and theories. It is normally much more stimulating, from an academic point of view, to show significant improvement of the model performance by increasing complexity.

Following the latter approach, Nash and Sutcliffe (1970) had presented a strategy for model development from the simple to the complex “which may help to avoid this frustration” (Bergström, 1991). This allows one to begin with simple assumptions regarding process structure, identify catchments where these assumptions result in poor performance, and then progressively introduce appropriate complexity as seems to be justified by diagnostic tests and other evidence. In this regard, the repeated use of a “universal” model structure as a starting point can aid in the development of the ‘experience’ necessary to diagnose what does, and does not, work at a specific location.

Ultimately, the goal of working with large catchment sets is to better understand the hydrological cycle. Through a process of designing improved local models, and by understanding the relationships between functional behavior (expressed through model structure) and observable characteristics of the catchment, we should progress towards better comprehension of catchments as systems.
5.3 Developing a useful system for classifying catchments

This suggests, of course, a link with the use of “classification” methods to establish a basis for grouping together catchments having similar structural form and functional behavior. Classification approaches have been successfully exploited by many disciplines, such as biology or sociology, to provide a vehicle for progress when the underlying processes were not well understood at the scales of interest (Sivapalan, 2003). While catchment classification has traditionally sought to group catchments by behavior to facilitate estimation of quantities such as flood frequency (e.g., Dalrymple, 1960; Laaha and Blöschl, 2006), by a synoptic comparison of catchments with contrasting characteristics, one can seek to understand the process controls and macro-scale signatures that arise from the action and interaction of underlying processes (Wagen and Montanari, 2011; Grauso et al., 2008). In this way, “comparative hydrology” (Falkenmark and Chapman, 1989) can exploit knowledge regarding a much wider array of conditions and processes than can ever be possible with a single model structure, and thereby provide a valuable complement to the detailed investigation of specific catchments. For example, the PUB Synthesis report (Blöschl et al., 2013) revealed significant trends of decreasing performance with increasing aridity, and increasing performance with increasing catchment size and data availability (see Parajka et al., 2013; Salinas et al., 2013), patterns than would have been impossible to detect in any other way.

Of course, several key issues will need deeper investigation. No clarity has yet emerged regarding which (observable) catchment characteristics are hydrologically relevant to catchment classification. Further, how process dominance (and complexity) varies with environment and scale (Skøien et al., 2003), and the role of “thresholds”, are not well understood. These make it difficult to properly link catchment types to the selection of appropriate process representation (thereby determining model structures). Importantly, classification could help in identifying “typical” (i.e., normal or representative) catchments to be targeted for intensive investigation, thereby providing a
stronger basis for “catchment observatory” initiatives such as CUAHSI. Meanwhile, the contrast with “typical” or “normal” ranges for catchments in a given class will facilitate the detection of “outlier” catchments similarly deserving of targeted special attention.

5.4 Testing hypotheses and assessing reliability

As mentioned above, large sample studies offer a unique opportunity to test hypotheses that are based in process understanding. However, the aggregate performance indices traditionally used in hydrology for performance assessment are poorly benchmarked (Legates and McCabe, 1999; Schaefli and Gupta, 2007) and not easily related to catchment properties and functioning (Gupta et al., 2008), and are therefore of limited usefulness in large sample assessments. A key to better use of large sample studies is to find ways to relate model performance and predictive uncertainty (and therefore model adequacy) to catchment structure and function, thereby providing insight into which processes the model is incapable of describing well (Fig. 1).

In this regard, further progress in diagnostic analysis (Yilmaz et al., 2008; Martinez and Gupta, 2010, 2011; McMillan et al., 2011), improved understanding of physical land surface constraints (Koster and Mahanama, 2012), and improved assessments of model structural adequacy (Gupta et al., 2012) and uncertainty (Montanari et al., 2009; Montanari, 2011; Montanari and Koutsoyiannis, 2012) are required. All of these areas depend on increased analytical sophistication, and will undoubtedly benefit from the breadth of information contained in large sample data sets.

5.5 Coping with variability and change

The non-stationarities underlying changing climate and catchment conditions make the problem of estimating hydrological fluxes and predicting catchment response more difficult. In such cases, it becomes important to distinguish between situations that are predictable from those that are not (Blöschl and Montanari, 2010; Kumar, 2011). It is now common to use scenario analysis as a way of assessing the impacts of future
change on hydrological response (Mahmoud et al., 2009, 2011; among many others). However, investigation using large sample data sets can facilitate a much wider range of analyses.

For example, large sample data set were used by Merz et al. (2011) and Coron et al. (2012) to investigate changes in estimated model parameters associated with changes in climate, and by Ter Braak and Prentice (1988) to use spatial gradients as surrogate for temporal gradients in investigations of change. Similarly, Peel and Blöschl (2011) suggested that changing climate could cause hydrological processes in one catchment to become similar to those in other catchments currently experiencing conditions similar to the target climate, thereby providing a basis for understanding the effects of change. Ultimately, these analyses must exploit the information provided by data other than runoff, including hydrogeological information, soil moisture from remote sensing products (Parajka et al., 2006), and snow characteristics (e.g., Blöschl and Kirnbauer, 1991, 1992) etc.

5.6 Trading depth of analysis for sample size

Finally, while the benefits of large sample hydrology are many and important, they come at a cost. When large numbers of catchments are analysed, all procedures for selecting model structures and estimating parameters must be automatic. This makes it difficult to attend to clues that might otherwise be provided by the assessment of local knowledge (whether hard or soft data), or by process understanding that can be gained from field trips. In this regard, soft information on catchment functioning can be as valuable as hard data for improving understanding about catchment functioning (Seibert and McDonnell, 2002). Ultimately, there is no “free lunch”, and investigations that exploit the benefits of large catchment sample sizes (providing breadth) must complement the detailed investigations of specifically targeted catchments (providing depth).
6 Conclusions

This paper has discussed the need to actively promote and pursue the use of a “large catchment sample” approach to modeling the rainfall-runoff process, thereby balancing depth with breadth. In either case, the need to understand hydrological change will require that large sample investigations be guided by process understanding, with statistical analysis playing the important supporting roles of (a) capturing the summary effect of various controls, including feedbacks across processes and scales, and (b) detecting spatial and temporal patterns that will need to be properly explained. Only then can we expect to achieve improved predictability and the ability to extrapolate to new situations (Kumar, 2011). We hope that this paper serves the function of provoking further discussion, and will promote what could be an important theme for “Panta Rhei”, the upcoming IAHS Scientific Decade.

Supplementary material related to this article is available online at: http://www.hydrol-earth-syst-sci-discuss.net/10/9147/2013/hessd-10-9147-2013-supplement.pdf.

Acknowledgements. We wish to recognize and gratefully acknowledge the considerable efforts of John C. Schaake (recently retired from NOAA) who played a major and active role in promoting “large catchment sample hydrology” during the past decade through activities such as the Model Parameter Experiment (MOPEX). The first author was partially supported by a grant from the Co-operative Research Programme (Trade and Agriculture) of the Organization for European Co-Operation and Development (OECD). The fourth author acknowledges funding from the Austrian Science Fund (projects W1219-N22 and P 23723-N21).
Large-sample hydrology: a need to balance depth with breadth

H. V Gupta et al.

References


Large-sample hydrology: a need to balance depth with breadth

H. V Gupta et al.
Large-sample hydrology: a need to balance depth with breadth

H. V Gupta et al.


9176


Large-sample hydrology: a need to balance depth with breadth

H. V Gupta et al.


Large-sample hydrology: a need to balance depth with breadth

H. V Gupta et al.


Fig. 1. How an approach based in evaluation of “signature properties” can be used to “detect and diagnose” model deficiencies and develop appropriate ways to “improve” the model/hypothesis (Figure based on ideas presented in Gupta et al., 2008).
Fig. 2. “Classical” versus “Diagnostic” model evaluation. The classical approach compares model simulations directly with the collected data. In the diagnostic approach, patterns in the model simulations are compared with corresponding patterns in the data. (Figure based on ideas from Gupta et al., 2008).
Fig. 3. Benefits of large-sample hydrology.
Fig. 4. Shows perceptions regarding economic barriers to the availability of hydrometeorological data in Europe. Survey results are stratified by: data providers (dark grey) and data users (light grey); Western Europe (left subplots) and Eastern Europe (right subplots); and type of data (bars indicating streamflow, precipitation, radar, geospatial, and others). ‘Yes’ responses indicate that economic barriers are perceived to exist, whereas ‘No’ indicates that economic barriers are perceived not to exist. (Reproduced from Viglione et al., 2011).
Fig. 5. Crash testing a rainfall-runoff model (reproduced from Andréassian et al., 2009).