Journal of Hydrology 600 (2021) 126508

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Karst modelling challenge 1: Results of hydrological modelling

Pierre-Yves Jeannin [a,q,*], Guillaume Artigue [b], Christoph Butscher [c], Yong Chang [d], Jean-Baptiste Charlier [e,f], Lea Duran [g], Laurence Gill [g], Andreas Hartmann [h], Anne Johannet [b], Hervé Jourde [b], Alireza Kavousi [i], Tanja Liesch [j], Yan Liu [h], Martin Lüthi [k], Arnauld Malard [a], Naomi Mazzilli [l], Eulogio Pardo-Igúzquiza [m], Dominique Thiéry [n], Thomas Reimann [i], Philip Schuler [g], Thomas Wöhling [o,p], Andreas Wunsch [j]

[a] Swiss Institute for Speleology and Karst-Studies, SISKA, Serre 68, CH-2300 La Chaux-de-Fonds, Switzerland
[b] HydroSciences Montpellier (HSM), Univ. Montpellier, IMT Mines Alès, CNRS, IRD, Alès, France
[c] TU Bergakademie Freiberg, Institute of Geotechnics, Gustav-Zeuner-Straße 1, 09599 Freiberg, Germany
[d] Nanjing University, Department of Earth Sciences, China
[e] BRGM, University of Montpellier, Montpellier, France
[f] G-eau, INRAE, CIRAD, IRD, AgroParisTech, Supagro, BRGM, Montpellier, France
[g] Department of Civil, Structural and Environmental Engineering, Trinity College Dublin, Dublin 2, Ireland
[h] University of Freiburg, Chair of Hydrological Modeling and Water Resources, Friedrichstraße 39, 79098 Freiburg, Germany
[i] TU Dresden, Institute for Groundwater Management, D-01069 Dresden, Germany
[j] Institute of Applied Geosciences, Division of Hydrogeology, Karlsruhe Institute of Technology (KIT), Kaiserstr. 12, 76131 Karlsruhe, Germany
[k] University of Zurich, Department of Geography, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland
[l] EMMAH, INRAE, Avignon Université, 84000 Avignon, France
[m] Instituto Geológico y Minero de España (IGME), 28003 Madrid, Spain
[n] BRGM, F-45060 Orléans, France
[o] TU Dresden, Institut for Hydrology und Meteorology, D-01069 Dresden, Germany
[p] Lincoln Agritech Ltd., Ruakura Research Centre, Hamilton 3240, New Zealand
[q] University of Neuchâtel, Centre for Hydrogeology and Geothermics, Emile-Argand 11, CH-2000 Neuchâtel, Switzerland

ARTICLE INFO

This manuscript was handled by Corrado Corradini, Editor-in-Chief, with the assistance of Junbing Pu, Associate Editor

Keywords:
Modelling
Karst
Comparison
Efficiency criteria
Recharge
Baseflow

ABSTRACT

The complexity of karst groundwater flow modelling is reflected by the amount of simulation approaches. The goal of the Karst Modelling Challenge (KMC) is comparing different approaches on one single system using the same data set. Thirteen teams with different computational models for simulating discharge variations at karst springs have applied their respective models on one single data set coming from the Milandre Karst Hydrogeological System (MKHS). The approaches include neural networks, reservoir models, semi-distributed models and fully distributed groundwater models. Four and a half years of hourly or daily meteorological input and hourly discharge data were provided for model calibration. The validation comprised forecasting one year of discharge, without the observed discharge data. The model performance was evaluated using the volume conservation, Nash-Sutcliffe efficiency (NSE) and the Kling-Gupta efficiency (KGE) applied on the total discharge and individual flow components. As a result, the comparison of model performances is a challenging task due to the differences in the model architecture but also required time steps: some of the models require aggregated daily steps while others could be run using hourly data, which provided some interesting differences depending on how the data was transformed. The use of instantaneous data (e.g. value at noon) produces less bias that averaging hourly data over one day. The transformation of hourly into daily data produces a decrease of Nash and KGE of 0.05 to 0.08 (i.e. from 1 to ~0.93). The resulting simulations (forecasted values for year 2016) produced KGEs ranging between 0.83 and 0.37 (0.83 to −0.24 for NSE). Although the simulations matched the monitored flows reasonably well, most models struggled to simulate baseflow conditions accurately. In general, the models that performed the best for this exercise were the global ones (Gardenia and Varkarst), with a limited number of parameters, which can be calibrated using automatic calibration procedures. The neural network models also showed a fair potential, with one providing reasonable results despite the relatively short dataset

* Corresponding author.
E-mail address: Pierre-yves.jeannin@isska.ch (P.-Y. Jeannin).

https://doi.org/10.1016/j.jhydrol.2021.126508
Received 8 March 2021; Received in revised form 8 May 2021; Accepted 27 May 2021
Available online 6 June 2021

available for warming-up (4.5 years). Semi-and fully distributed models also suggested that with some more effort they could perform well. The accuracy of model predictions does not seem to increase by using models with more than 9–12 calibration parameters. An evaluation of the relative errors between the forecasted and the observed values revealed that for most models, 50% of the forecasted values contained more than 50% of difference against the observed discharge rate, with 25% having a difference larger than 100%. A significant part of the poorly forecasted values corresponded to base-flow which was surprising given that as base-flow is generally much easier to predict than peak flow. Hence, this shows that modelling approaches and criteria for the calibration are too oriented towards peak-flow sections of the hydrographs, and that improvements could be gained by more focus on the base-flow.

1. Introduction

1.1. Idea

The aim of the modelling challenge was to invite research teams from all over the world to compare their modelling approaches and tools by applying them on the same data set. A common evaluation method had to be defined and results from each team have been analyzed with the same criteria. In this way all results could be directly compared. Various objective functions (various ways to compute the difference between the forecasted and the observed time series) have been tested, with three subsequently used for the final comparison.

1.2. Problem definition

In order to start with a rather simple question (Step 1) it was decided to focus on the hydrological behavior of karst hydrogeological systems (KHS), i.e. the relationship between parameters controlling the input of water into karst (mainly precipitation and temperature) and the discharge rate of karst springs at the output of the system (Fig. 1). The question to be solved by the respective teams is to forecast as precisely as possible the discharge rates at the karst outlet from the meteorological input data.

It should be noted that further steps are foreseen, with the simulation of the flow and head distribution in space and time (Step 2), and the simulation of transport processes (Step 3). However, the evaluation for this step 1 of the challenge does not take into account whether the models will be able to address the objectives of the further steps or not.

In this paper, the word “Model” refers to approaches and corresponding numerical tools used by the respective teams for their modelling exercise. The word simulation is used to name the results of the modelling work.

1.3. Flow through a karst massif

As outlined in most text-books (e.g. Ford and Williams, 1989), karst massifs are characterized by a specific geomorphology and hydrology (Fig. 1). Karst landforms are due to the dissolution of the rock in precipitation water. Surface landforms such as karrenfields, dolines and swallowholes are produced by water infiltration into the rock. Vertical shafts, caves (vadose or phreatic) and large springs are related to the flowthrough and output of water.

From a hydrological point of view precipitation (rain and snowmelt) infiltrates through vegetation cover and soils into the top layer of the rock, which is more or less weathered (epikarst). One part of precipitation waters returns back to the atmosphere through evapotranspiration. In some cases, surface water may form a small stream or even a river, which is swallowed into concentrated location (swallow-hole). If the stream drains a non-karstic area this part of the catchment will be called “allogenic”.

Epikarst distributes water partially towards vertical shafts leading the liquid directly down to the phreatic zone. Another part of water is trapped at the bottom of the epikarst and seeps slowly downwards through cracks and joints. At some points water is drained again by the network of karst channels, which efficiently drains the rock mass, and leads water at the outlet of the karst system: the karst spring (Mangin,

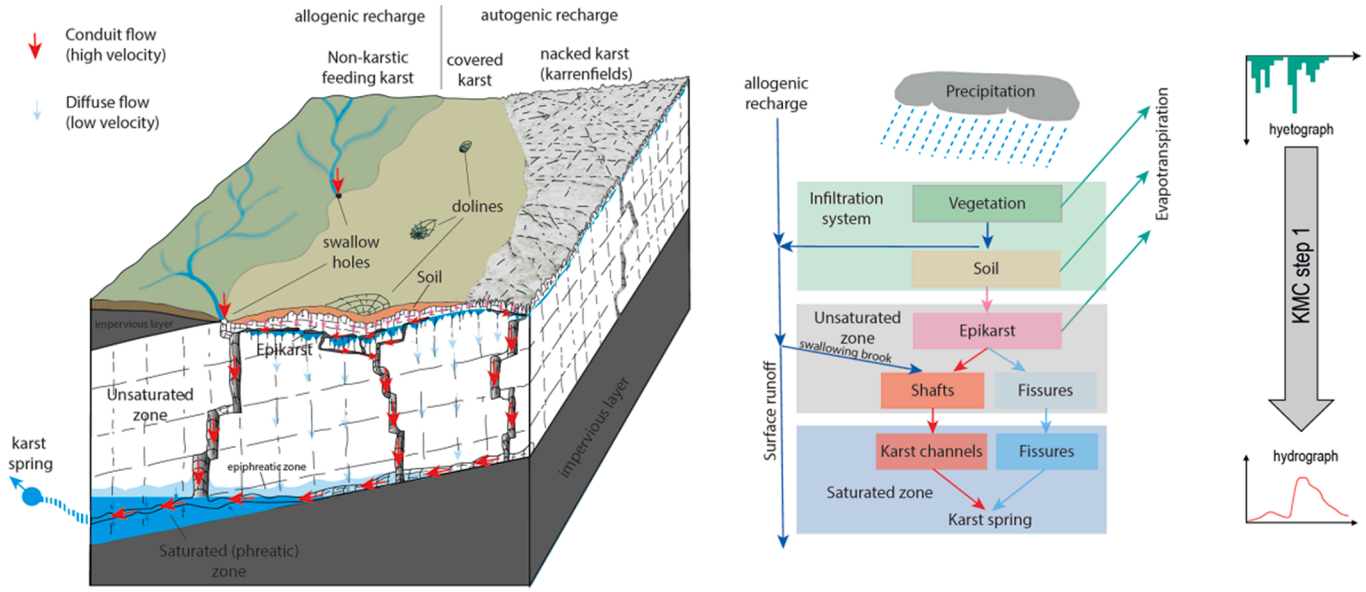

Fig. 1. Conceptual model of a karst hydrological system (left). Precipitation water (allogenic and autogenic) flow through the karst massif along conduits (red arrows) and fissures (blue arrows) with corresponding peak of discharge at the spring after a few hours or days. For KMC step 1 we only consider the relationship between precipitation and discharge. Subsystems may or may not be considered in the applied modelling approaches. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

1975; Williams, 1983; Smart and Friederich, 1986). The volume of water stored in fissures is significant compared to that stored in karst channels (Király, 1975; Drogue, 1980; Smart and Hobbs, 1986).

The relationship between recharge and discharge at springs seems quite direct in most karst systems as rain events are followed by clear peaks of discharge rates at karst springs usually within a few hours to days. The response is therefore fairly similar to that usually observed in surface hydrology.

The degree of diffuse flow in the rock matrix compared to concentrated flow in conduits, as well as the amount of concentrated recharge into swallow-holes vs diffuse recharge through soils and epikarst may be significantly different depending on rock characteristics and on geomorphological and climatological contexts.

Various models have been developed to simulate the functional relationship between precipitation and discharge peaks. Some of them include physical processes taking place in each subsystem traversed by water, other ones simplify the nature into linear subsystems, and other ones just consider the mathematical relationship between an input signal and an output one. The aim of step 1 of the Karst Modelling Challenge is to compare these approaches, and their capability to simulate a hydrograph from a precipitation times series.

1.4. Short outlook on groundwater flow modelling in karst

The analysis of hydrographs started mainly after the publication of Maillet (1905). However, continuous monitoring of karst springs really started only since the 1950s. For the analysis of karst springs, the first models occurred along the 1960s (Schoeller, 1962; Forkasiewicz and Paloc, 1967; Castany, 1968; Brown, 1970, 1973; Mangin, 1970, 1975; Chemin, 1974; Bezes, 1976; Milanovic, 1976; Atkinson, 1977; Dreiss, 1982, 1983; Dodge, 1983; Bonacci, 1987; Ford and Williams, 1989). Early on two components were recognized, one explaining the quick and intense response of spring discharge rates to precipitations, and the other explaining the slow decrease of discharge rates (recession) through long periods of time without recharge events.

Several conceptual ideas were developed through various types of mathematical models, which can be grouped into three large families (Teutsch and Sauter, 1991; Kovács and Sauter, 2007; Ghasemizadeh, 2012; Fiorillo, 2014):

1) Data-driven or "black-box" models, including reservoir & non-parametric transfer functions
2) Hydrogeological models (enclosing matrix flow with or without conduits)
3) Pipe-flow models (based on the hydraulics of flow in conduits)

1.4.1. Data-driven or "black-box" models

In these models, the system is only considered as a black-box or a combination of black-boxes, each box transforming an input signal (inflow) into an output one (outflow).

The transformation is based on a purely mathematical function, without considering physical processes controlling flow in the underground or by strongly simplifying them (Dewandel et al., 2003; Fiorillo, 2014). The simplest model of this type would be a single box model, taking the total precipitation as input and transforming it into a discharge time series. The classical "unitary hydrograph" model, assuming a single linear transfer function (kernel), as first introduced first by Sherman (1932), was tested on karst systems. It quickly turned out that KHS are neither linear nor steady and that such a simple model cannot describe karst aquifers in a meaningful way. Thus, more sophisticated schemes, coupling several kernels, were necessary to fit observed data somehow (Dreiss, 1982). Neural networks are the most recent developments of this category of models.

Another category of lumped models was attempted by various authors, using cascading and parallel reservoirs (Forkasiewicz and Paloc, 1967; Bezes, 1976; Halihan and Wicks, 1998). Usually one reservoir simulates the quick-flow component, and the others the slower component of KHS. A new parameter then occurred, i.e. the distribution of recharge into the respective components. The best results were obtained with a combination of three to five reservoirs, some of them being partially emptied by evapotranspiration. This approach led to rather precise simulation results (e.g. Long, 2009). In most cases, input parameters are not distributed, i.e. each reservoir is supposed to expand over the whole size of the catchment area. In fact, this approach can also be distributed in space, and may take into account some spatial differences of the aquifer characteristics in the respective parts of the catchment area (e.g. Bittner et al., 2018). With three to five reservoirs distributed in tens or hundreds of zones the number of parameters becomes very high, making it difficult to calibrate the models. Thanks to automatic calibration procedures, this task can now be achieved with a reasonable effort (Pianosi et al., 2016; Mazzilli et al., 2017).

1.4.2. Hydrogeological models (fully distributed models)

Darcy (1856) empirically described the mathematical relationship between groundwater gradients (head differences) and groundwater flow quantity. This was in fact the first groundwater flow model, which was one-dimensional and steady-state. Analytical solutions for some typical geometries and boundary conditions were developed. In the 1970s the development of computers enabled equations to be solved numerically, making it possible to compute flow across a much wider range of situations (e.g. Bredehoeft and Pinder, 1970). Application of this approach to karst was first attempted in the mid-1970s (Király and Morel, 1976), but it quickly became obvious that KHS cannot easily be modelled using this approach, as the simulation of quick and intense flood events, as well as a sustained base-flow, as observed in most karst springs proved challenging. Modellers decided to introduce a network of conduits within a matrix with a low hydraulic conductivity. With contrasts of 10^{-4} to 10^{-5} m/s between matrix and conduits, karst hydrographs could be simulated. However, the recharge process also had to be improved which was achieved by adding a layer with a high hydraulic conductivity at the top of the model (epikarst), able to absorb precipitation and distribute it to the conduits and matrix (Király et al., 1995). Given that the high contrast of hydraulic conductivities in spatial models can be quite challenging from a numerical point of view, various solutions have been developed, such as double continuum or double medium models (Teutsch and Sauter, 1991).

Two further improvements occurred later: turbulent flow in conduits and partially saturated flow (Thierrien and Sudicky, 1996; Annable and Sudicky, 1998).

In these fully distributed models the number of parameters is very high (with several values for each cell) making calibration manual very time consuming and practically impossible. Parameter zonation and automatic calibration procedures improved this aspect (Doherty et al., 1994; Borghi et al., 2016). However, as the basic structure of the network needs to be defined first, its geometry and topology is not usually adjusted during the calibration process, though it may play a significant role in some situations (Kovács, 2003) especially in the epiphreatic zone (Jeannin, 2001). Assuming the physics considered in the model is correct, models can however reveal unrealistic structures and, therefore, can act as a very advanced tool that provides further insights (Enemark et al., 2019; Gill et al., 2020; Duran and Gill, 2021).

These models belong to *process-based groundwater flow models* according to the classification proposed by Anderson et al. (2015).

1.4.3. Pipe flow models (semi-distributed models)

Pipe flow models only consider a network of conduits and neglect flow through the matrix, or add it as a non-spatialized entity. The idea is not new as many original authors (Martel, 1921; Lièvre, 1915, 1940; Trombe, 1948), exploring karst with both a hydrogeological and a speleological perspective, suggested flow in karst to be well described by equations of the pipe hydraulics. These models also belong to process-

based models according to Anderson et al., (2015), but the main process considered is the hydraulics of flow through conduits in the phreatic and epihreatic zones (turbulent and partially saturated). Darcy's flow is not or only marginally considered.

A few authors then used pipe models for analyzing their data (White and White, 1970; Atkinson, 1977; Boegli, 1980; Lauritzen et al., 1985; Smart, 1988). All of them demonstrated that flow is turbulent in most natural conduits (caves), at least during medium to high water conditions. Flow may be transitional or even laminar during very low flow conditions. Ford and Williams (1989) and Worthington (1991) have provided suggestions on values of friction factors in caves, but these have only really been based on defined known conduit sections, and none of these authors really attempted to simulate flow in a complex network of pipes.

Jeannin and Maréchal (1995) and Jeannin (2001) assessed head-losses due to bends and changes in cross-sections compared to those related to friction along walls. It turns out that the latter are usually dominating, and typical values could be assigned for natural conduits. They also calculated the hydraulics of a network of natural conduits for modelling flow in the downstream part of the Hölloch cave (Switzerland). The role of epiphreatic conduits on the hydraulics and flow velocity was found to be significant for many KHS. This specific behavior cannot be simulated with the usual hydrogeological software packages. For this reason, pipe-flow modelling (e.g. the SWMM software developed by US EPA and other urban drainage software) have increasingly been used in modeling karst hydrology: Campbell and Sullivan (2002); Peterson and Wicks (2006); Wu et al. (2008); Gill et al. (2013); Chen and Goldscheider (2014); Jeannin et al. (2015); Kaufmann et al. (2016); Malard (2018); Vuilleumier et al. (2019); Morrissey et al. (2020); Schuler et al. (2020). These models enable flow to be simulated in a complex network of pipes with various diameters, roughness, and cross-sectional shapes.

If this approach is efficient to link observed heads and discharge rates, it is not adequate to link precipitation to discharge hydrograph. Indeed, a recharge model is always necessary for doing it, in order to transform meteorological data into groundwater recharge.

1.4.4. Modelling recharge of karst hydrogeological systems (KHS)

Recharge is defined here as the quantity of water infiltrating and flowing through the karst system. It roughly corresponds to the total precipitation minus evapotranspiration and overland runoff bypassing karst in some circumstances. Snow and ice storage are also to be considered in recharge processes.

Recharge strongly depends on evapotranspiration. Several formulae have been developed for assessing the Potential Evapotranspiration (PET) from meteorological parameters (Thornthwait, 1948; Penman, 1948; Turc, 1961). These models provide an assessment of the quantity of water which would be evaporated or transpirated by vegetation if sufficient water were available in soils.

In reality, during periods of low water, soils become strongly unsaturated and plants may suffer a water deficit, reducing their capacity of bringing water from the soil into the atmosphere. Real Evapotranspiration (RET) is thus lower than PET. In humid climates the difference may be negligible, but it becomes significant or even extreme in dryer climates.

Many models can be used for the assessment of PET and RET, from very simple to highly sophisticated ones. Real measurements of RET are difficult, especially at catchment scale.

1.5. Objectives of the modelling challenge

The overall aim of the modelling challenge, of which this paper presents step 1, is to explore models for their ability to reproduce the following aspects: groundwater recharge, groundwater flow velocity in space and time, groundwater heads in space and time, spring discharge hydrographs.

The first step of the challenge is limited to the relationship between meteorological parameters (parameters of groundwater recharge) and spring hydrographs (discharge rate at system's output), in other words it is focused on aspects of groundwater recharge and spring hydrographs.

Any group or person interested in comparing his model results to the other ones was openly invited to participate[1].

2. Modelling approaches compared in KMC

All types of models described above were applied for the challenge: The two categories of lumped parameter models (neural networks and reservoirs), as well as semi-distributed and fully distributed models.

2.1. Data-driven models 1: Neural networks

Two research teams applied neural networks for the simulation of karst hydrographs (KIT-Karlsruhe with three models and IMT Mines Alès with one). The latter team uses deep recurrent Multilayer Perceptron (MLP) and the former team applies NARX model, Convolutional Neural Networks (CNN) as well as Long Short-Term Memory Networks (LSTM).

Artificial neural networks (ANN) form a branch of artificial intelligence. They are based on single neuron operations, which are arranged and connected in a specific architecture using parameters (also called weights). The behaviour of the model is "programmed" by the values assigned to the parameters set. During the training stage, data is presented to the network and special training algorithms fit the target data by modifying the parameters between the neurons. One neuron calculates two values: (i) the weighted sum of its input vector with its parameters, and (ii) its output, which transforms the weighted sum in a scalar value using predefined functions (e.g. linear or sigmoid), known as activation functions. The graded, nonlinear response of sigmoid activation functions enables the ANN to capture nonlinear relationships within the training data.

At this stage, it is important to notice that ANN models have no predefined function before training. The goal of the training is to build both the "function" and to calculate its "parameters". For this purpose, a database representing all the kinds of behaviors is necessary. Usually, to represent the behavior with good quality, the necessary length of the database depends on the behaviors to be considered. When the database is continuous with a small time-step (1 h), as in this study, several behaviors at different time scales take place; thus the database must be sufficiently long to allow the model catching the underlying phenomena: 15 years should be necessary (Artigue et al., 2012; Coutouis et al., 2016). In this study only 4.5 years in two different periods of time were available. It is thus important to notice that the database used in this work is not sufficient for ANN approaches. Results shown hereafter can thus be considered as a proof of feasibility.

The various approaches (models) used in this study are summarized hereafter. As these techniques are diverse and not commonly used in karst modelling further details are given as supplementary material.

All neural network models designed in this work used hourly precipitation and temperature, daily evapotranspiration data resampled at hourly time step, and discharge data at an hourly time step. The data ranges used for training or validation differ slightly depending on the model.

2.1.1. Recurrent deep multilayer perceptron (ANN/rec_MLP)

The chosen multilayer perceptron is a recurrent neural network (RNN) with one hidden layer of n_h hidden neurons and one output neuron. The hidden layer neurons are nonlinear and apply a sigmoid function. The output neuron is linear; its output is equal to the weighted

[1] The challenge is going on: any person interested can use data available in the supplementary material to this paper, and can send her/his result to the corrsponding author for obtaining the criteria values.

sum of its inputs. This specific architecture displays the two properties of universal approximation and parsimony, the latter of which is due to the nonlinearity relative to model inputs and parameters (Barron, 1993). The Universal approximation (Hornik et al., 1989) means that the model is able to identify any differentiable function, provided the existence of a database representing the behavior to identify. For further details on this architecture, the interested reader is referred to Dreyfus (2005). In this study a deep multilayer perceptron is used by adding a deep layer, connected in order to implement a preprocessing of data. Output neuron is a sigmoid one. The model contains 381 parameters.

2.1.2. NARX model

The NARX model (nonlinear autoregressive network with exogenous inputs) used by KIT (Karlsruhe) is quite similar to the previous model used by mining school in Alès (France). The differences are for example the number of hidden layers, the input data being used, and the architecture.

NARX models can also belong to the group of recurrent neural networks (RNN). NARX in particular have a global feedback connection between their output and input layer, enabling information flow in different directions in the network, and making them especially suitable for modelling nonlinear dynamic systems. This also means that each output value is not only regressed on the independent input signals, but also on previous output signals. Depending on the chosen model, RNNs can have difficulties in capturing long-term dependencies greater than 10 time steps due to the problem of vanishing gradients (Bengio et al., 1994); however, NARX can keep information up to three times longer than simple RNNs (Lin et al., 1996a, 1996b). To build and apply NARX models, we use Matlab 2019a (Mathworks Inc.) and its Deep Learning Toolbox. The model contains 371 parameters.

2.1.3. Long short-term memory network

Deep-learning approaches such as Long Short-Term Memory Networks (LSTM) received a lot of attention in recent years, due to significant successes in various disciplines. LSTMs also belong to the group of RNNs and are widely applied to model sequential data like time series or natural language. Unlike other RNNs, LSTMs have been explicitly designed to overcome the problem of vanishing gradients. Besides the hidden state inherent to all types of RNNs, LSTMs have a cell memory (or cell state) to store information and three gates to control the information flow. This enables information to remain in the cell memory, which is why LSTMs can handle long-term dependencies (Hochreiter and Schmidhuber, 1997). The LSTM model that simulates the outflow of the Milandre cave has 11,311 trainable parameters.

2.1.4. Convolutional neural networks

Convolutional Neural Networks (CNNs) are a type of deep learning multi-layer neural network that was originally designed to efficiently handle image data, but which has been also successfully applied to time series forecasting, treating a sequence of observations like a one-dimensional image. The CNN that simulates the outflow of the Milandre cave has 950,921 trainable parameters.

2.2. Data-driven models 2: Reservoirs

Gardenia (BRGM), RCD-Seasonal (TU-Freiberg), CHLEM (Uni-Zuerich) and KarstMod (SNO KARST) models are all very similar tools designed to simulate the main processes of the water balance at catchment scale using simplified physical laws representing flows through a succession of reservoirs. Using climatic time series (precipitation, potential evapotranspiration, air temperature) on a recharge area, these models are able to compute the flow at the outlet of a KHS (spring).

2.2.1. RCD_seasonal

The karst system is represented by three storages, which are connected by links: one storage for representing the recharge system (soil and epikarst), and two others for the conduit and the diffuse flow system. Precipitation and actual evapotranspiration define the model input. Groundwater recharge to the conduit and diffuse flow systems occurs when the water content of the recharge system exceeds a threshold. This threshold is seasonally variable (sine wave with one-year wavelength, minimum in the summer and maximum in the winter). For calibration, maxima (peaks) and minima (base flow before peak) of measured spring discharge from the years 2014–2015 were taken as fit target. Model RCD-seasonal is described in Butscher and Huggenberger (2008).

2.2.2. KarstMod

KarstMod is a modular platform for modeling of the rain-level-flow relationship in karstic basins developed by the French SNO Karst. In its most complete form (Mazzilli et al., 2019) it offers 4 compartments organized on as a two-levels structure. In this study, it was used to test the performance of a two-parameter transfer function with an infinite characteristic time (Guinot et al., 2015). The proposed model has a two-level structure. An upper compartment is meant to account for flow within the soil and epikarst. Above a certain level water overflows to a lower compartment, which stands for the infiltration and saturated zones. The unit response of the lower compartment is an inverse power *function of* time, which allows for infinite characteristic times. Such response may be well suited to reproduce long-term memory effects of KHS.

2.2.3. Gardenia

The GARDENIA numerical code (Thiéry, 2014, 2015, http://gardenia.brgm.fr/) has 4 reservoirs: i) a superficial reservoir S representing the water retention capacity of the soil (quadratic law), ii) an intermediate linear reservoir H roughly representing the unsaturated zone, and two underground linear reservoirs G1 and G2 representing compartments of the saturated zone. The total discharge is the sum of the fast, slow and deeper flows, which is routed to the outlet.

2.2.4. CHLEM (Uni-Zuerich)

Cave Hydrology Lumped Element Model (CHLEM) provides different types of hydraulic elements that can be freely joined. For KMC, the hydraulic system was simulated with for parallel conduits that produce the model discharge, which is compared to the measurements. Each of the four pipes consists of one storage element and one delay element and receives water from an input element with or without evapotranspiration. In the storage elements, the discharge depends on the storage volume following a power-law, and is limited by an adjustable parameter. Each of the elements has one to three adjustable parameters that were adjusted with an optimization algorithm (COBYLA from the nlopt optimization library).

2.3. Semi-distributed models

KRM_1 (SISKA), Varkarst (Uni-Freiburg), and InfoWorks (TCD Dublin) models are all semi-distributed models but with significant differences. The number of parameters for these models is moderate.

KRM_1 is very similar to lumped parameter models with reservoirs but input parameters can be semi or fully distributed. It has been developed as an alternative to models (daily rainfall-runoff models based on two reservoirs and three parameters (Edijatno and Michel, 1989), in order to further refine interception and soil infiltration processes which are assumed to be the main recharge controls for lowland and vegetated karst aquifers. Storage capacities of interception reservoirs vary according to the type and respective proportion of land-uses (forests, cultures, urban areas, denudated areas, etc.) and according to seasons in order to reproduce the dynamic of the vegetation. Interception and ETP parameters are not uniform over the whole catchment. Two slow reservoirs are introduced to mimic processes in the soil/epikarst and in the deep vadose zone. They allow underflow and overflow depending on recharge intensity. Parameters of soil/epikarst reservoirs

vary according to a drought index while those of the deeper reservoirs are unique. In addition, an "exchanger" makes the recharged water pass from the low permeability volumes (LPV) to the karst conduits and vice-et-versa depending on pseudo hydraulic relations. The complete description of KRM_1 is available in Malard [2018]. The spatial distribution of rainfall over the catchment area can be taken into account, but this was not applied in the present application (rain is spatially uniform, but vegetation cover is semi-distributed). The number of parameters for the calibration is 18.

Varkarst (Hartmann et al., 2013) is a semi-distributed model that considers the spatial heterogeneity using a distribution function. The meteorological forcing is assumed homogeneous over the catchment area. With shape parameters describing the distributions of soil and epikarst storages, vertical hydraulic conductivities, the separation into diffuse & concentrated recharge, and the groundwater hydraulic conductivities, the model consists of 15 compartments reproducing the spatial variability of recharge and storage dynamics. This value was tested in Hartmann et al., (2012), and shown to be enough to obtain stable integrated discharge and its variability. It has also been confirmed by several following applications (Hartmann et al., 2013; Hartmann et al., 2021; Liu et al., 2021). Each compartment includes three superposed reservoirs (soil, epikarst and groundwater). The first 14 compartments represent the dynamics of diffuse recharge and matrix flow, while the last compartment represents the dynamics of concentrated recharge and conduit flow (Hartmann et al., 2014). The discharge of the main spring is determined by the addition of flow from the groundwater reservoir of all model compartments, which are controlled by the variable groundwater storage coefficients and conduit storage coefficient, respectively. The number of parameters for the calibration is 8.

TCD-Dublin developed a semi-distributed 1D model of the catchment based on the drainage software InfoWorks ICM (Innovyse), which models the hydraulics of the karst conduit network in both open channel and pressurized pipe flow (Gill et al., 2013). The governing model equations are the Saint-Venant equations of conservation of mass and momentum. Diffuse recharge from rainfall is modelled per sub-catchment via a series of reservoirs: rainfall runoff, soil and groundwater stores in series, all yielding different delayed discharges in parallel into the pipe network. Flows can also discharge into permeable pipes, one connected for each sub-catchment to represent the primary and secondary permeability expressed by Darcy's Law (Morrissey et al., 2020; Schuler et al., 2020). The higher the density of these pipes included in this type of model moves towards becoming a fully distributed.

In the InfoWorks model, the Milandre catchment area is explicitly subdivided into 30 sub-catchments of 0.45 to 0.475 km^2. The sizes of the sub-catchments are the result of modelling experience with respect to balancing the numerical stability of the model along with reasonable diameters of the draining pipes and heads. Every sub-catchment generates recharge and storage through a hierarchical sequence of processes related to a) runoff (quick recharge); b) soil store contribution (intermediate recharge); and c) groundwater store contribution (slow recharge). 19 parameters control these dynamics, and in this model, all sub-catchments have the same parameter settings.

2.4. Fully distributed models

In KarstFLOW (Pardo-Igúzquiza et al., 2018) recharge is split between fast recharge and slow recharge. Fast recharge is a percentage of rainfall that reaches the water level through the conduit network instantaneously (for practical purposes). The slow (diffuse) recharge takes place over the whole recharge domain, while fast (concentrated) recharge is limited to specified regions (Fig. 2). Diffuse recharge is estimated by the method of Thornthwaite applied to the percentage of rainfall that does not undergo quick recharge. The model domain is discretized into voxels with absolute altitudes using a digital elevation model. The diffuse flow along each column of voxels is described by the

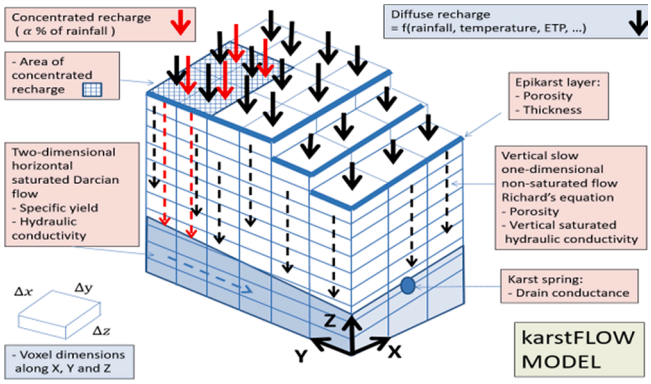

Fig. 2. Parameterization used in KARSTFLOW methodology.

Richards equation (assuming a porous medium). When the non-saturated flow reaches the phreatic zone, the one-dimensional vertical flow couples, as recharge, with a two-dimensional domain where flow through conduits and fractures is simulated by the equations of a two-dimensional Darcian flow in a porous medium. The karst spring discharge is simulated by using a drain. The parameters needed by the model are described in Fig. 2.

MODFLOW-CFPv2, in short CFPv2 (Fig. 3), is a flow and transport discrete-continuum model which combines CAVE heat and solute transport routines (Liedl et al., 2003) with MODFLOW2005-CFP Mode1 flow code (Shoemaker et al., 2008), considering some additional improvements (Reimann et al., 2018). In CFPv2 the aquifer encloses a network of conduits with turbulent/laminar flow embedded within a matrix of lower permeability (porous medium). Recharge can be split between diffuse recharge through the low permeability porous medium, infiltration through the epikarst reservoir (e.g., Chang et al., 2019), and direct infiltration into conduits and their associated fast-response

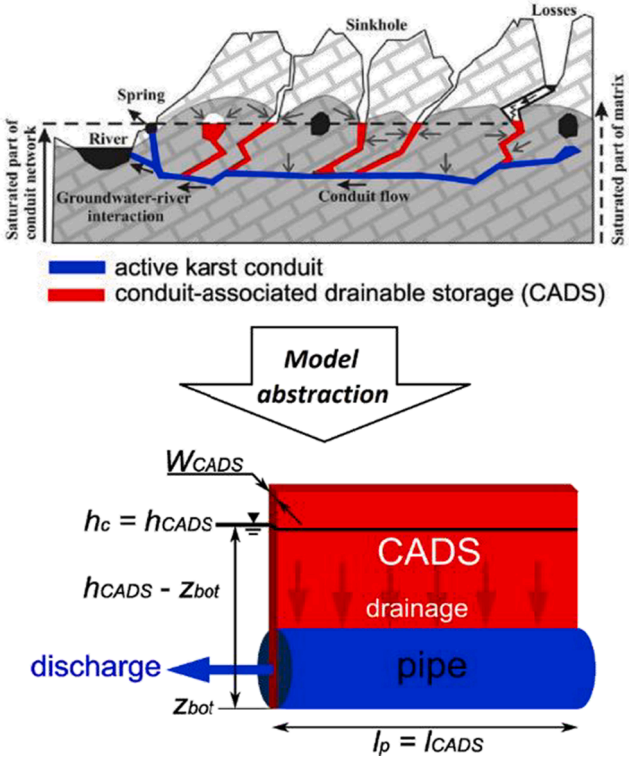

Fig. 3. Conceptualization on karst aquifers by MODFLOW CFPv2 (See Reimann et al., 2014 for detailed explanation); note that sinkholes and losing streams are not present within the Milandre catchment.

reservoirs, namely CADS (see Kavousi et al., 2020). Accordingly, several recharge conditions were conceptualized and initiated as numerical model variants of the catchment. In the presented simulations, recharge is calculated by subtracting potential evapotranspiration (ETP) to total precipitation (P), considering a simple degree day snow-cover-snowmelt approach for estimation of hourly snowmelt, as well (for the methodology of degree-day approach, see Rango and Martinec (1995) among many others). It should be pointed out that, the distributed recharge was given here as a single value all across the modeling domain. Moreover, only six input parameters were homogenously assigned to the model, keeping in mind that the overall aim of karst modeling challenge as a proof of concept, not calibration statistics. It should be noted that spatial variation of recharge and model input parameters (which were not also provided with current state catchment data), could improve the performance of CFPv2 process-based model. The inverse problem was solved using the PEST parameter estimation code (Doherty, 2015).

2.5. Comment on the various approaches

The challenge is focused on the recharge-discharge relationship only (mainly the uncertainty of forecasted hydrographs at the system output). Table 1 provides an outlook of the main characteristics of the respective models. Lumped parameter models are thus basically sufficient to address the question. Although semi- and fully-distributed models have been applied, they neither considered the spatial variability of infiltration nor of the aquifer parameters (parameters were assumed to be the same all over the catchment area). Only the spatial variability of the interception of precipitation and evapotranspiration was considered in some of the models.

Most models include some parameters which are more or less physically-based, and others which aren't. InfoWorks, KarstFLOW and MODFLOW CFPv2 are mostly based on physical descriptions of flow with most of their parameters having a physical meaning. However, the physical concepts used in these models are considerably different and simplified compared to processes really taking place in nature.

3. Study site and data

The Milandre karst hydrogeological system (MKHS) is located in the Tabular Jura, in the front part of the Jura Mountains in Northern Switzerland. The site is known as being the karst laboratory used by SISKA for conducting surveys and experiments since early 1990s. The system outputs are formed by two perennial springs: Saivu (20–200 $L{\cdot}s^{-1}$, 373 m.a.s.L) and Font (12–600 $L{\cdot}s^{-1}$, 369 m.a.s.L). During medium to high water periods water overflows at Bâme spring (0–3000 L s^{-1}, 375 m.a.s.L.). All springs are located on the left (West) side of the Allaine river (Fig. 4). MKHS is fed by a recharge area of ~ 13 km2 (Grasso and Jeannin, 1994; Jeannin, 1996; Perrin et al., 2003) which consists in a limestone plateau at an altitude of ~ 550 m.a.s.L. This area is occupied by forests (~30% of the catchment area), pastures (~30%), cultivated lands (~30%) and urban regions (~5%). The limestone is almost completely covered (95%) by soil with a thickness in the order of 0.3 to 0.5 m in forested parts and up to ~ 2 m in cultivated land. GW recharge is purely autogenic and diffuse (no surface stream and swallowholes).

The mean annual precipitation is 1070 mm as measured by the MeteoSwiss weather station at Fahy, located 7 km away from the center of the catchment area. The mean annual effective precipitation (precipitation minus real evapotranspiration) is 520 mm. An overview of the hydrogeological settings of the area is provided by Kovacs and Jeannin (2003). The aquifer is hosted by Upper Jurassic limestone and is underlain by the Oxfordian marls, which act as a regional aquiclude. As the active conduits lie almost directly above this impervious formation, there is no deep phreatic zone and the system is qualified as a shallow karst system (Perrin, 2003).

The downstream part of the catchment contains a 10.5 km speleological network, the Milandre cave, which hosts a 4 km long perennial cave stream, the Milandrine. This is the main drainage axis of the catchment. The stream flows into a sump around ~ 500 m upstream from the Saivu spring. Two main underground tributaries feed the Milandrine and each of them contributes to 25 to 35% to the cumulated discharge of the springs (Grasso and Jeannin, 1994).

The underground stream is monitored along its course at three places for discharge, temperature and electrical conductivity. Several drillholes located around the main karst conduits, have been monitored showing the groundwater behavior around the conduit network. This data will be important for the next steps of the challenge.

4. Methods

4.1. Workflow of the karst modelling challenge (KMC)

For step 1 of KMC, dedicated to the relationship between meteorological parameters and spring hydrograph, the discharge rates of the three spring of the system (Saivu, Font and Bâme) were simply summed in order to provide one single discharge time series for the whole KHS.

4.1.1. Calibration data

In February 2017, each team received 4 years of meteorological data, from 1.1.1992 to 31.12.1995. For the main meteorological station (Fahy), located 7 km south-west of center of the catchment area, hourly precipitation, hourly temperature and daily PET were given. Hourly precipitation data were also given for the Maira station, which is located in the middle of the catchment area. Daily precipitation were also given for a third station (Mormont) located ~ 2 km south-east of the catchment. Each team also received ~ 2.5 years of hourly discharge data (sum of Saivu, Font and Bâme springs), from 24.9.1992 to 28.3.1995.

This data set was used for a first calibration of the models, as well as for adjusting criteria for the comparison. All teams returned their results, and a first comparison was made in 2018. Based on this first exercise, it was then decided to provide a second dataset for a further evaluation of model performances.

4.1.2. Model evaluation

In June 2019 a new dataset was sent to all teams with two years of data (2014–2015) for the meteorological station of Fahy and for the total combined discharge rates of Saivu, Font and Bâme springs. Data from Maira and Mormont were not provided because they were declared as not useful by the teams after the analysis of the first dataset.

Teams were asked to apply their model to the second dataset with the parameters calibrated with the first dataset, to run the quality criteria, and to write a comment on it. If they wished, they had the possibility to improve the calibration using years 2014–2015.

4.1.3. Test data

Finally, all teams were invited to forecast discharge time series for the year 2016, for which they only received input data. That meant that they could use their calibration data based on the 1992 to 1995 time series data for the simulation (forecast) of 2016, or based on the 2014–2015 time series, or a combination of both. They sent their best model to the first author, who applied the same evaluation criteria to all received time series.

Among the 15 teams originally interested in the challenge, 8 of them sent their results and 5 additional teams joined afterwards, making a total of 13 modelling teams comparing their approaches and results.

4.2. Uncertainty of the data

We briefly discuss here two different types of uncertainties, which are attached to measured data used for the challenge:

Table 1
Overview of the main characteristics of the 13 models applied for the Karst Modelling Challenge step 1.

Model summary		Approach	Main characteristics	Parameter #	Calibration type	Objective function	Computational cost (for 1 run)	Remark
BRGM, France	Gardenia	Lumped, reservoir	4 reservoirs	9	Autom	Nash	Seconds	
Uni-Freiburg, Germany	Varkarst	Semi-distributed	15 compartments	8	Autom	min. residuals	Seconds	Distribution functions are used to reflect spatial variability
IMT Mines Alès, France	ANN/ rec_MLP	Lumped, neural	Recurrent Deep Multilayer Perceptron	381	Autom.	mean quadratic error	Seconds	Not enough hydrologic situations in the database (too short) to design a robust neural-network model.
SISKA-Switzerland	KRM_1	Semi-distributed	3 reservoirs + 1 exchanger	22	Manual	Volume conservation, similar shape of peaks and base flow	Seconds	The upper reservoir (soil) is semi-spatially distributed
IGME Madrid, Spain	KarstFLOW	Fully distributed	2 submodels (recharge + aquifer)	12	Autom.	minimization of residuals	15 min	Aquifer simulated with Darcy's law
TCD Dublin, Ireland	InfoWorks	Semi-distributed	Double recharge reservoirs (diffuse/concentrated) + pipe flow	min. 19	Manual	Volume conservation, similar shape of peaks and base flow	5-10 min.s	Subcatchments linked by a pipe network –> spatial distribution of recharge
KIT-Karlsruhe, Germany	CNN	Lumped, neural	Convolutional Neural Networks	950921	Autom.	MSE	Seconds	
KIT-Karlsruhe, Germany	NARX	Lumped, neural	Nonlinear AutoRegressive networks with eXogenous inputs	371	Autom.	MSE	Seconds	
SNO KARST, France	KarstMod	Lumped, reservoir	2 reservoirs	6 for KMC (up to 12)	Autom.	MSE	Seconds	Infinite characteristic time configuration with fixed ra, emin, kec and calibrated hmax, tau0, alpha
TU-Dresden, Germany	CFP-modified	Fully distributed	Discrete-continuum model	min. 6	Autom.	sum of squared errors, SSE	Minutes	
TU BA-Freiberg, Germany	RCD-Seasonal	Lumped, reservoir	3 reservoirs	9	Autom.	peak max and flow min	Seconds	Calibration needs 40 seconds (10 runs)
Uni-Zürich, Switzerland	CHLEM	Lumped, reservoir	9 Elements (flexible)	min. 30	Autom.	RMS difference	0.01 sec	Tracks flux, temperature and dye, flexible chaining of elements
KIT-Karlsruhe, Germany	LSTM	Lumped, neural	Long Short-Term Memory networks	11311	Autom.	MSE	Seconds	

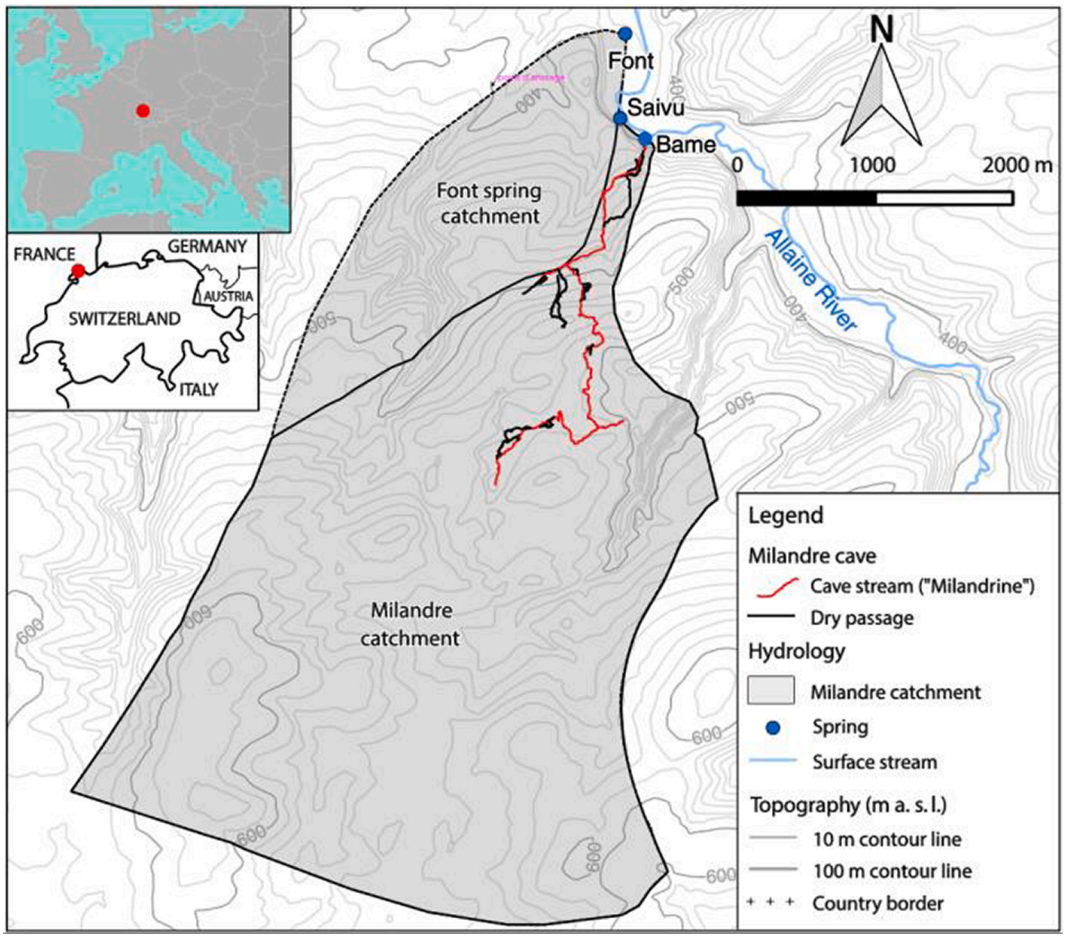

Fig. 4. Map of the Milandre catchment with its main characteristics. The catchment area is about 13 km2 large and encloses the Milandre underground stream.

4.2.1. Measurement uncertainty

Air temperature is measured with a high precision, better than +-0.1 °C. Precipitation is measured with a high-quality rain gage (calibrated tipping bucket of MeteoSwiss), with a precision in the order of 0.2 mm. Both are measured every 10 min and averaged (T) or cumulated (P) over one hour.

For the spring discharges the water level is measured every 15 or 30 min and is transformed into discharge using a rating curve established empirically for the measurement section. The obtained discharge rates are averaged over one hour. For this step 1 of the challenge, the output discharge rate is obtained by summing the three outlet points (springs) of the MKHS. Rating curves were derived by taking discharge rate measurements at the springs across a range of various rates (or water levels). Most measurements were made by salt dilution method. Measurements are repeated every few years in order to look at potential changes in the rating curves.

Uncertainties on discharge rate data are in the order of ±20% in absolute values. However, this uncertainty must be considered as global and not as a noise related to every single value.

4.2.2. Uncertainty related to spatial heterogeneity

Meteorological parameters are measured at an official meteorological station of the Swiss Meteorological Agency (SMA). The station is located 7 km away from the center of the catchment area, at a similar elevation. Summer showers locally can be quite different from what is measured at the rain gage. However, for most major events, the rain distribution is rather uniform, as given by a comparison with other meteorological stations in the area, including the one located in the middle of the catchment area, which is rather small (10–15 km^2) and rather flat. This latter rain gage was compared but not used for the challenge, as its measurements appeared to be less liable than those of the official SMA station.

Land-use is not uniform. As detailed previously the catchment area is covered by forest (30%), pastures and cultivated land (60%), and about 10% of urban area. The effect on recharge of the differences in RET models for these respective regions is not directly measurable, inducing another source of spatial uncertainty.

4.3. Evaluation of model performances

The only aim in the present step of the challenge was to simulate the observed discharge rate at the system output “as well as possible”. The quality criteria (or efficiency criteria) must therefore result from the comparison between the observed and the simulated (forecasted) time series. Six methods or objective functions were considered, providing an evaluation of model performances as a function of different objectives (Madsen et al., 2002; Moriasi et al., 2007; Biondi et al., 2012): Mean Squared Error (MSE), Normalized Mean Squared Error (NMSE), Variance (VAR), Nash criteria (NASH), Kling-Gupta Efficiency criteria (KGE), and Volume Conservations Coefficient (VCC). Each method summarizes the difference between the forecasted and the observed curves into one single value, which is obtained with a different calculation.

During the first exercise using data from the 1990 s, discussions between participants of the challenge showed that only NASH, KGE and VCC were applicable for the comparison. The other criteria are not normalized, and can therefore hardly be compared from one simulation to another. Hence, although they have been calculated, they will not be

further discussed in this paper.

VCC is the ratio between the forecasted volume of flow and the observed one. Ideally the value must be 1. Quality classes are given in Table 2 as well as the range between very good to low.

NASH takes into account the ratio between the mean squared error (MSE) and the variance (VAR) as defined by Nash and Sutcliffe (1970). The NASH coefficient ranges from -∝ to 1 (1 being the ideal value). For NASH less than 0, the mean of the observed values is a better indicator than the calibrated model. For NASH = 0, the results provided by the model are as accurate as the mean of the observed values. Usually, the model may be considered as reliable when NASH coefficient is higher than ~0.75.

The Nash criterion (least squares) is well known for giving a lot of weight to flood periods, but in return being less sensitive with regards to the simulation of low water levels. This motivated Gupta et al. (2009) to introduce an improved criteria (KGE) taking into account a normalized distance between the observed and forecasted curves. KGE ranges between 1 and 0.

These three criteria were applied to the whole time series. However, in order to evaluate more closely the apparent strengths and weaknesses of the respective models, we also split the observed curve into four components: peak rising limb, peak recession, base flow, and undetermined. We calculated the quality criteria for the overall time series, as well as for each of its four components.

In addition, a fourth criteria was used to assess the applicability of the respective models, i.e. the effort necessary to set up, calibrate and run the model. Classes are given in Table 2.

A "final note" was arbitrarily calculated using a weighted mean between KGE (weight 2/5), VCC (weight 2/5) and Nash (weight 1/5).

Relative errors are also used for characterizing the difference between the forecasted and the observed values. For each value (hour) the difference is calculated and divided by the larger value of observed or forecasted discharge rate. Doing so, a forecasted value 3 times lower than the observed one makes a relative error of − 300%, and a forecasted value 3 times larger than the observed one a relative error of + 300%.

4.4. Time step

Hourly Discharge Measurements (HMs) of the Milandre springs show fast fluctuations on a time-scale which is shorter than a day which would suggest that model simulations should be computed at an hourly time-step.

However, some models are designed for daily time-steps and provide daily values (DVs). These may be compared to the reference (observed) time series (HMs) either by taking daily instantaneous values (DIVs), e.g. the measured value at noon, or by averaging values over 24 h (DAVs).

In the framework of KMC, results of the simulations had to be compared to HMs for computing the efficiency criteria. Therefore, simulated DVs were retransformed into hourly values (HVs). A linear decomposition of DVs into HVs was thus necessary.

Conversions from hourly values to daily values and then to hourly values again may significantly affect the quality of the simulation. Two biases were tested: the effect of resampling/decomposing, and the way to resample (from instantaneous or from average values).

Table 2
Quality classes for VCC, NASH and KGE criteria, and for the necessary effort.

		Criteria				
		KGE	VCC	Nash	Effort	Final score
Quality classes	Very good	>0.85	within 2%	>0.75	~2 hours	>0.80
	Good	0.75 - 0.85	within 5%	0.5 - 0.75	< 1day	0.7 - 0.8
	Fair	0.5 - 0.75	within 15%	0.25 - 0.5	~ 1day	0.6 - 0.7
	Medium	0.25 - 0.5	15 to 25%	0.0 - 0.25	2-5 days	0.5 - 0.6
	Low	<0.25	>25%	<0.0	> 5 days	<0.5

One question raised by participants was if it was better to use average or instantaneous values. Therefore, we made a comparison between reference HMs and hourly values issued from the resampling of DIVs and DAVs. The new time-series are called HDIVs and HDAVs respectively (Fig. 5).

Fig. 6 shows that significant differences do exist between HMs, HDIVs and HDAVs. Most of the time, instantaneous values (HDIVs) show a better fit with HMs than average values (HDAVs): HDAVs systematically underestimate peak-flows and often overestimate the rising limb of the floods. HDIVs usually show a better fit with peak-flow except when variations of peak-flow are clearly shorter than one day (e.g. the flood on Dec. 5th, 1992). For this case, the averaged value may provide a better fit. HDIVs better reproduce low-flows than HDAVs.

In summary HDAVs show a more systematic bias than HDIVs.

HDIVs and HDAVs are thus compared to HMs using the provided efficiency criteria (see Table 3). In both cases it appears that VCC (volume conservation coefficient) is well conserved. NASH and KGE are about ~0.9 to ~0.95. meaning that conversion from hourly to daily and then to hourly values again degrades the time series in the same order of magnitude of what is usually targeted (and obtained) with (the best) simulations.

It should be observed that NASH and KGE for HDIVs are slightly better than those computed for HDAVs. VCC is a bit lower for HDAVs, which makes sense as HDVAs systematically underestimate peak flows.

Our advice for KMC was then to work at hourly time-step as far as possible for simulations at the Milandre test site. But if this was not possible, then teams were encouraged to use DIVs from HMs to compare daily values obtained with their models.

Finally, as all teams were required to compute their final efficiency criteria at an hourly time step, their simulated results, if calculated at daily time-step, needed to be resampled into hourly data using a linear decomposition.

5. Results

5.1. Overall performance

As shown on Fig. 7 the different model forecasts generally seem to reproduce the correct order of magnitude of discharge rates for most of the prediction period. Models tend to underestimate discharge during recessions occurring after significant high-water conditions (e.g. March and June on Fig. 7).

Given that most models provided reasonable results, quality criteria were then applied in order to identify strengths and weaknesses of the respective models. The following comparison takes into account only the best prediction for the year 2016 provided by the participating teams. Most are obtained with calibration on years 2014–2015, but some used a mix with 1992–1995 as well.

In Table 4 model results have been classified according to their "final score" as defined in chapter 3. Effort is given as indication. Concerning KGE and VCC all models are acceptable to good, but none is very good.

Forecasted curves for the year 2016 are given for all individual models in Fig. 8. The log scale provides a better view of the relative errors between the forecasted and the observed curve than the linear scale.

All models reproduce the overall shape of the curve of discharge variations, covering periods of time of low flow and other periods with peaks. The amplitude of the peaks is in the right order of a magnitude. Low flow conditions are correctly identified. However, a closer look shows that low flow conditions are rather poorly forecasted by most models whereby often the forecasted discharge rate is too low and too flat.

It appears that curves predicted by reservoir models and semi-distributed models reproduce somehow better the observed behavior with many peaks from January to June, and less peaks from July to October. InfoWorks and KRM1 are semi-distributed models with their

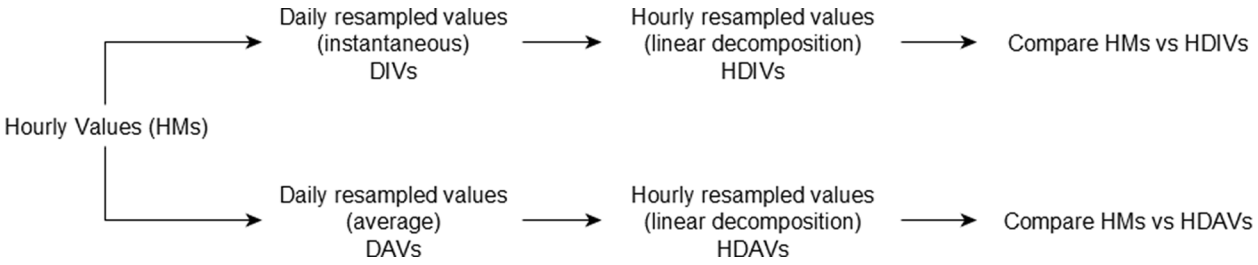

Fig. 5. Workflow applied for testing the effect of resampling and of using daily average values or instantaneous values (e.g. the measure value at noon).

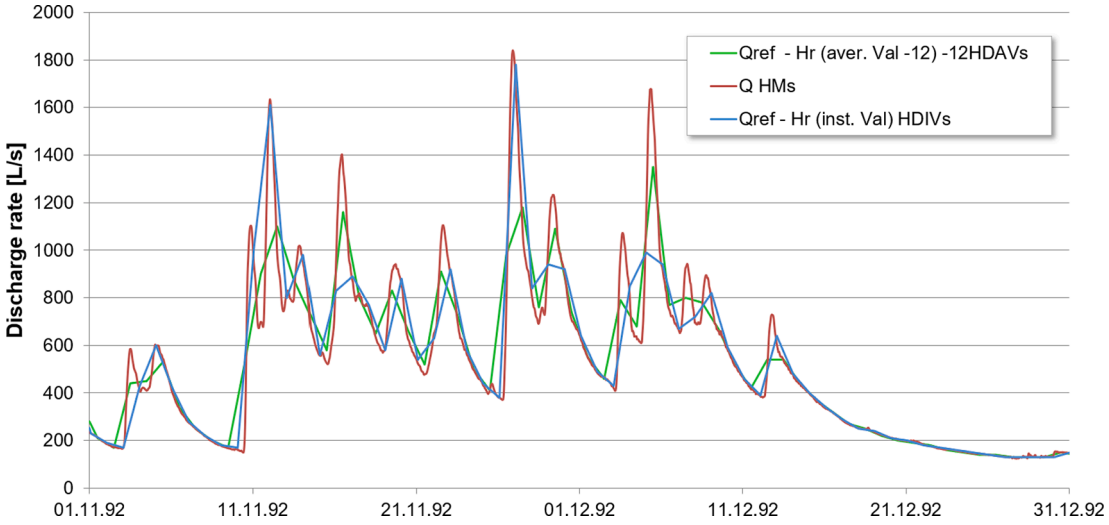

Fig. 6. Zoom on the three time-series HMS, HDIVs and HDAVs (period 1/11/1992 to 31/12/1992).

Table 3
Comparison of the efficiency criteria for HDIVs and HDAVs.

	Flow components				
HDIVs	Rising limb	Base flow	Flood recession	Undet. Flow	Total
VCC	0.95	1.01	1.01	1.04	1.002
NASH	0.816	0.985	0.917	0.927	0.92
KGE	0.86	0.96	0.95	0.93	0.95
	Flow components				
HDAVs	Rising limb	Base flow	Flood recession	Undet. Flow	Total
VCC	0.97	1.01	0.99	1.06	0.995
NASH	0.818	0.972	0.918	0.934	0.91
KGE	0.83	0.96	0.90	0.91	0.92

recharge processes based on reservoir models; their curves are thus quite similar to those of reservoir models. VarKarst is based on a different approach (it explicitly considers spatial variability), and it can be seen that the peaks are pretty well predicted, but the simulation does not reflect the observed behavior well during low flow conditions. RCD-seasonal oversimplifies recession during summer with just a single long recession. The distributed models fail at forecasting low flow conditions as too many little peaks are produced, and the subsequent recessions are much too quick. Peaks are underestimated in winter and overestimated in Summer. CHLEM, which is a reservoir model, produces curves which are similar to those of distributed models. Lumped models are not bad for the first half of the year, but have a significant drift in the second part of the year, with, ANN-2 performing better overall than the three others.

Relative errors are given in Fig. 9. Errors on peaks are of short duration, but may be of considerable intensity (usually larger than 400%, positive or negative, for some events along the year). The flow-rate during the recession in March is underestimated by most models. CNN is the only model providing a prediction close to observations for this event. VarKarst, KarstMOD and CHLEM are particularly low. The main recession in July tends also to be underestimated by most models with Gardenia, Karstflow and KRM1 performing better than the others.

It is noticeable that most models using neural-derived approaches display a significant drift in the last months of the simulations except the MLP model.

By classifying the relative errors, we can display the percentage of the forecasted data (hours/year) for which the error of the model is lower than a value (Fig. 10). For instance, the LSTM model (intense pale blue) has 22% of the forecasted values with an error lower than 50% (e.g. between 67 and 150 L/s instead of 100 L/s). Gardenia has 72%: the higher is the curve, the better the performance of the model.

5.2. *Performance by flow components*

The observed times series was divided into four components: rising limb of the peaks, recession, base flow and undetermined flow. Base-flow was empirically determined when the discharge rate is decreasing over five successive hours with a rate lower than 1 L/s per hour. The latter component includes points which are not clearly related to one of the other three classes.

For the year 2016 the significance of the four components is given in Table 5 (in hours). Variations from one year to the other are in the order of ±5%:

Each value of the observed curve is attributed to one single flow component. For the forecasted curves, the components of the observed curve are used for the separation, meaning that all models are compared on the basis of the same components. As can be seen on Fig. 11, most values belonging to the undetermined flow correspond in fact to base-flow conditions. They could not be attributed to base-flow because

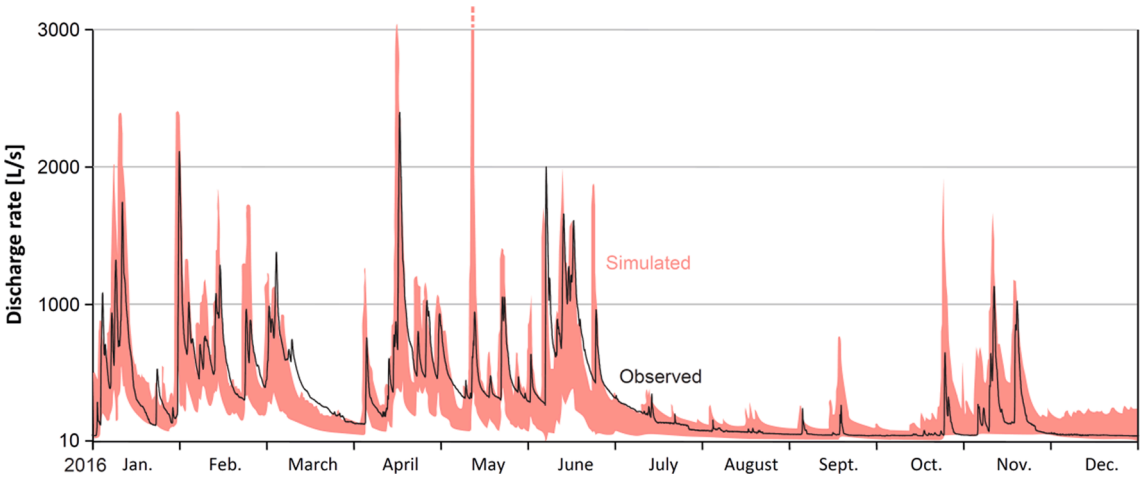

Fig. 7. Forecasted discharge rates in red (envelope of results coming from all models) compared to the observed rates (Black line). Most observed peaks are forecasted. Models tend to underestimate discharge during recessions occurring after significant high-water conditions (e.g. March and June). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
Comparison of the model performances. All models produced acceptable results. Color ratings are given in Table 2.

		KGE	VCC	NASH	Effort	Score
BRGM France	Gardenia	0.83	0.854	0.83	< 1day	0.84
Uni-Freiburg	Varkarst	0.80	0.85	0.79	< 1day	0.82
IMT Mines-Alès	ANN / rec_MLP	0.72	0.84	0.61	~ 1day	0.75
SISKA-Switzerland	KRM_1	0.71	0.78	0.63	~ 1day	0.72
IGME Madrid	KarstFLOW	0.71	0.78	0.63	3 days	0.72
TCD Dublin	InfoWorks	0.70	0.82	0.58	2-5 days	0.72
KIT-Karlsruhe	CNN	0.69	0.87	0.40	~ 1day	0.71
KIT-Karlsruhe	NARX	0.68	0.90	0.35	~ 1day	0.70
SNO KARST	KarstMod	0.68	0.72	0.65	< 1day	0.69
TU-Desden	CFP-modified	0.54	0.72	0.25	~7 days	0.55
TU-Freiberg	RCD-Seasonal	0.47	0.69	0.04	< 1day	0.47
Uni-Zürich	CHLEM	0.45	0.71	-0.10	< 1day	0.44
KIT-Karlsruhe	LSTM	0.37	0.59	-0.24	< 1day	0.34

they include slight increases of the discharge rate, which may be related to some technical problems (e.g. temperature variations on the electronics, or changes of the atmospheric pressure which are not perfectly corrected).

The respective models have specific results for each component. Gardenia has among the best performances for most criteria and components. This is less the case for Varkarst, especially concerning volume conservation (VCC). Most models have one component with KGE close to or below 0.5. Nash criteria is low for base flow in most models. CHLEM, CFP-modified and RCD_seasonal, three models with the lowest scores, are better than most other models for base-flow according to VCC and KGE.

The Nash criteria is quite poor for most models as this criterium cannot be good for a trending time series (as the mean value is taken as reference in the denominator of the Nash-Sutcliff formula).

Other quality criteria have also been calculated for the respective models, mainly Mean squared error and the variance test. However, these criteria do not provide more information than those presented in Table 6.

5.3. Comparison of calibrated parameters

Given that the models are based on different concepts, their parameters are all different and cannot be compared directly. Some aspects, however, can be roughly compared for several models, i.e. catchment size, storage capacity, conduit/diffuse flow components.

5.3.1. Catchment size

In karst regions catchment size is never known exactly, and so this was considered to be a calibration parameter by some of the models. The calibrated size area varies between 11.6 and 16.24 km^2, which ranges between-9 and + 28% of the assessed surface area (12.7 km^2). The main reason for these variations is related to the model applied for assessing evapotranspiration: the size was calibrated to adjust the water balance on the calibration period.

Neural networks models do not provide any catchment size. The role of this parameter is approached by a series of unidentified functions and parameters.

5.3.2. Storage capacity

The attenuation between recharge and discharge is related to the characteristics of the storage in all models. However, the way it is generated is different in the respective models. Most models have quick and slow storages included somehow, which reflects what is commonly accepted in karst aquifers since the 1970s as summarized e.g. by Bakalowicz (2005). Their number and characteristics are, however, different and hardly comparable. It may be only mathematical, or based on the infilling of a reservoir, or on the routing of water, etc. Somehow, all models generate peaks within 0.2–5 days after a rain event, and a base-flow usually corresponding to the emptying of some kind of reservoir (or mathematical function) with slow recession keeping water flowing out of the system over more than 6 months without recharge. The storage volume of this slow component is in the order of 100–250 mm.

5.3.3. Conduit/diffuse flow components

Another long-lasting debate among karst hydrogeologist (see e.g., Bakalowicz, 2005) is about the ratio between the quick-flow component (assumed to be related to conduits) and the slow-flow component (assumed to be related to matrix or diffuse flow). From our results we were hoping to be able to compare the amount of recharge flowing through both components of the respective models. However, as every model has a different architecture and concept, this could hardly be done. The only thing we can compare is the amount of peak flow (rising limb and recession) and base-flow (including undetermined flow) for each model. The comparison shows, however, that differences between models are not very significant, the amount of peak flow ranging between 64 and 78%, and the volume of base flow ranging between 22 and

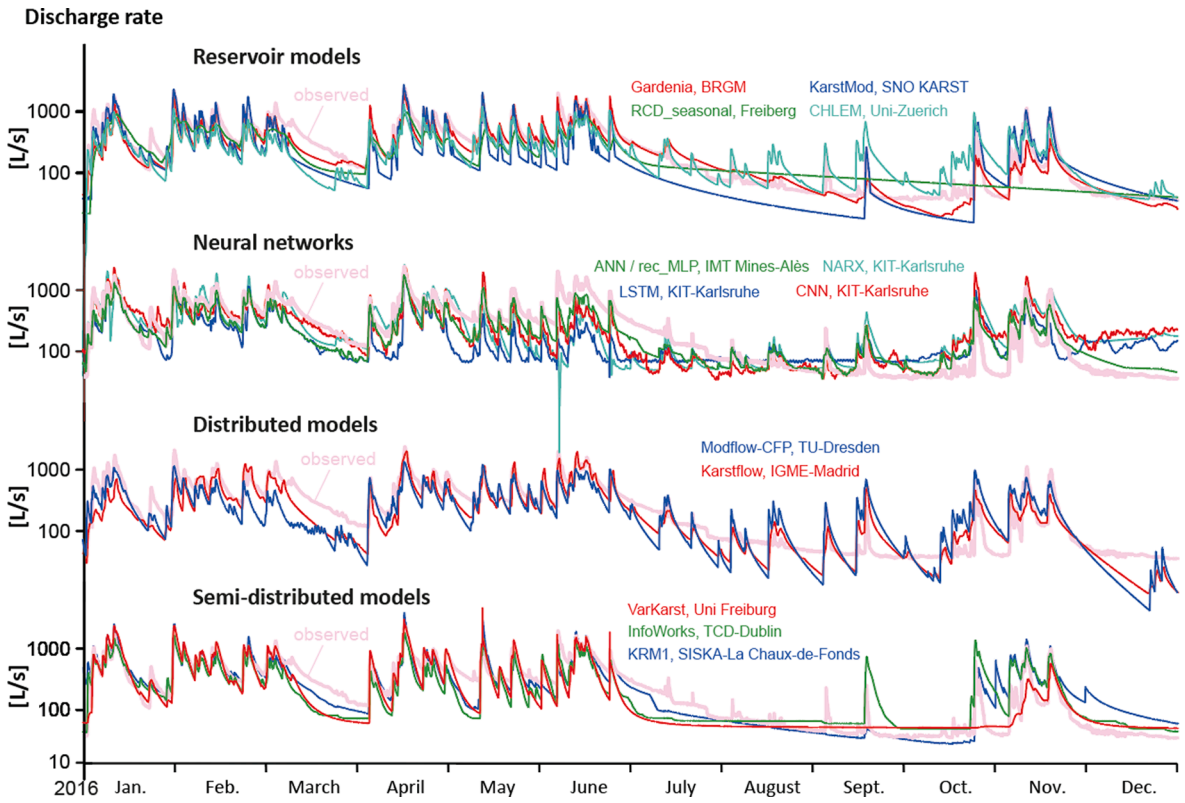

Fig. 8. Forecasted hydrographs of the thirteen models applied for KMC-step 1 compared to the observed curve (pale red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

36%.

6. Discussion and model comparison

The comparison shows that forecasted curves with similar values for their quality criteria may have different shapes. This illustrates the fact that quality criteria for model comparison could still be improved. Our first assessment of the results was that all models produced fair to good simulations. However, this assessment is probably rather qualitative than quantitative. Indeed, when looking at relative errors, we realized that most models produce values with errors larger than 50% for about 50% of the time or more. Results are clearly worse during low-flow conditions. Absolute differences are low during these periods of time, but relative ones are significant.

By looking at the respective curves (Fig. 8) and at the relative errors (Fig. 9), one can say that the three criteria we selected are not sufficient to characterize the overall quality of the results. For instance, the shape of the curve forecasted by the VarKarst model reproduces very well the peaks, be very poorly the values during the dry-season. Gardenia is better for base-flow, but still include significant differences in March and October of the simulated year. The fact that most models obviously tend to underestimate or poorly reproduce low-flow (e.g., LSTM or VarKarst) could be related to the way calibration is implemented. It is usually mainly targeted on minimising the squared errors metric, which gives much weight to the peaks and not much to low-flow values. A criterium minimizing relative errors instead of absolute errors could possibly improve the weighting of base-flow conditions in the overall results. Another option could also be to calibrate the parameters which control base-flow first as they not dependent much on the recharge model, and then to calibrate the other parameters on the residuals of this first calibration step. In the case of RCD_seasonal and possibly VarKarst, the main reason for the poor forecasting of low-flow conditions seems to be related to the recharge model, which completely fails to produce any recharge between June and November or December. The water deficit in the recharge reservoir is obviously too high over this period compared to the reality.

Interestingly the spatial distribution of precipitation appeared not to be significant and was abandoned along the process in agreement with all participants. Obviously, the number of recharge event for which differences are significant (e.g. summer storms) is low. This may be the case for the Milandrine KHS, which is small and located in a temperate and hilly region. This may be different for a larger catchment with more relief and contrasting climate.

KRM_1 model is the only model for which land-use was taken into consideration. However, this does not appear to improve the results in a way to produce better results than the other models.

The number of parameters in the respective models ranges between less than 10 and several thousands. Despite automatic calibration/optimization procedures, models with less parameters tend to be better than the models with many parameters, i.e. distributed models did not provide better results than global ones. Not considering catchment characteristics in the respective models could be a reason for this. Therefore, it seems that the transformation between meteorological parameters and discharge rates is mainly controlled by a few general characteristics of the system (such as size, recharge process and one or two storages) rather than by the detailed spatial distribution of parameters controlling these). Modellers using neural networks, with many parameters, claimed that data delivered for calibration were not sufficient. As no basic structure is provided in their models, compared to the models based upon a more physical function, they have to build this structure using data. They require therefore at least 15 years of data without any gap, which is challenging!

These observations raise the question of uncertainty and overfitting associated with numerical models. In fact, we can probably conclude that the best models are as good as the data we provided.

MODFLOW-CFPv2 and KarstFLOW are fully distributed models.

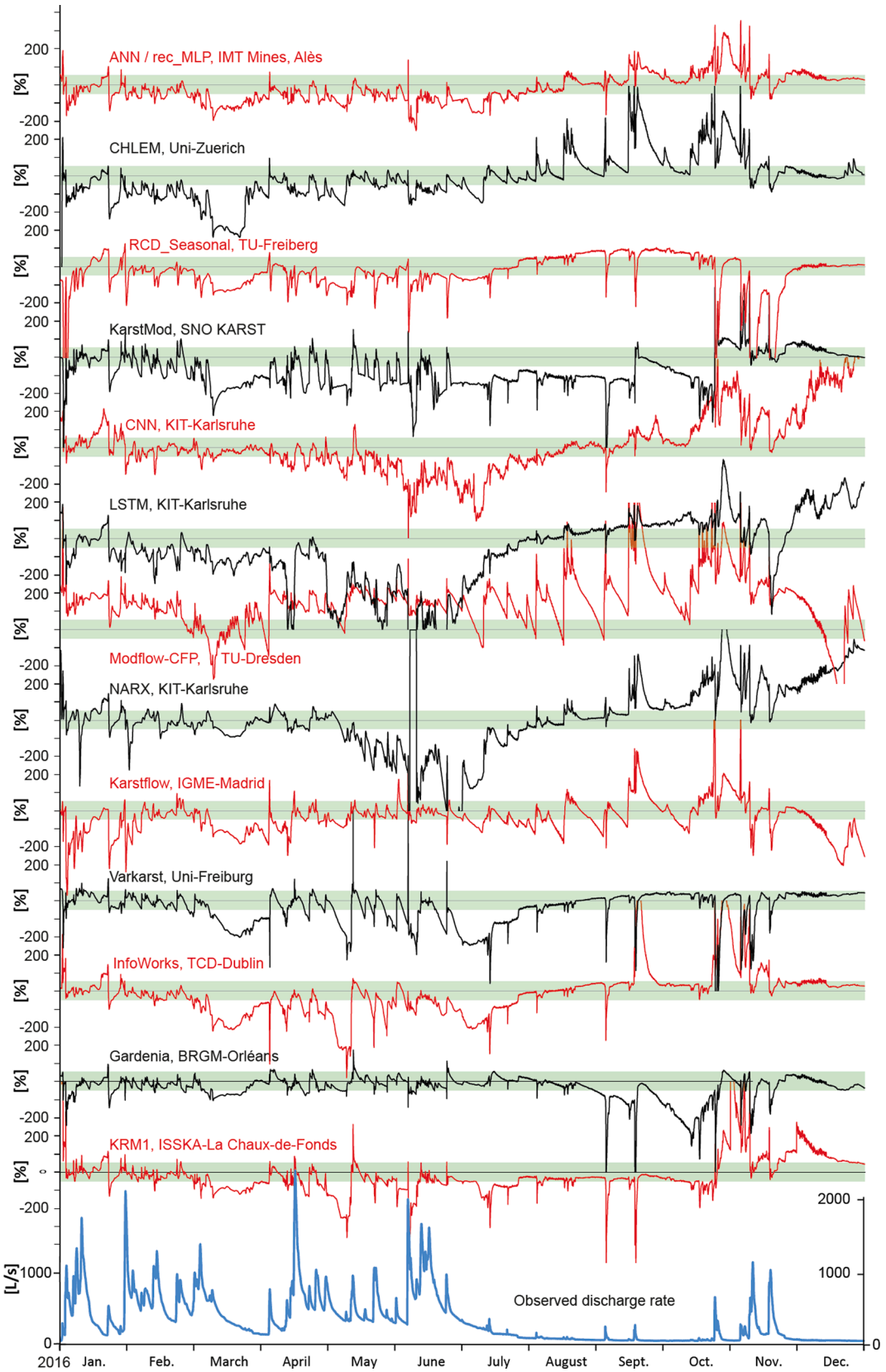

Fig. 9. Relative errors of the respective models for year 2016. All predictions show periods of time with errors larger than 200%. The green band shows the domain with + -50% of relative error. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

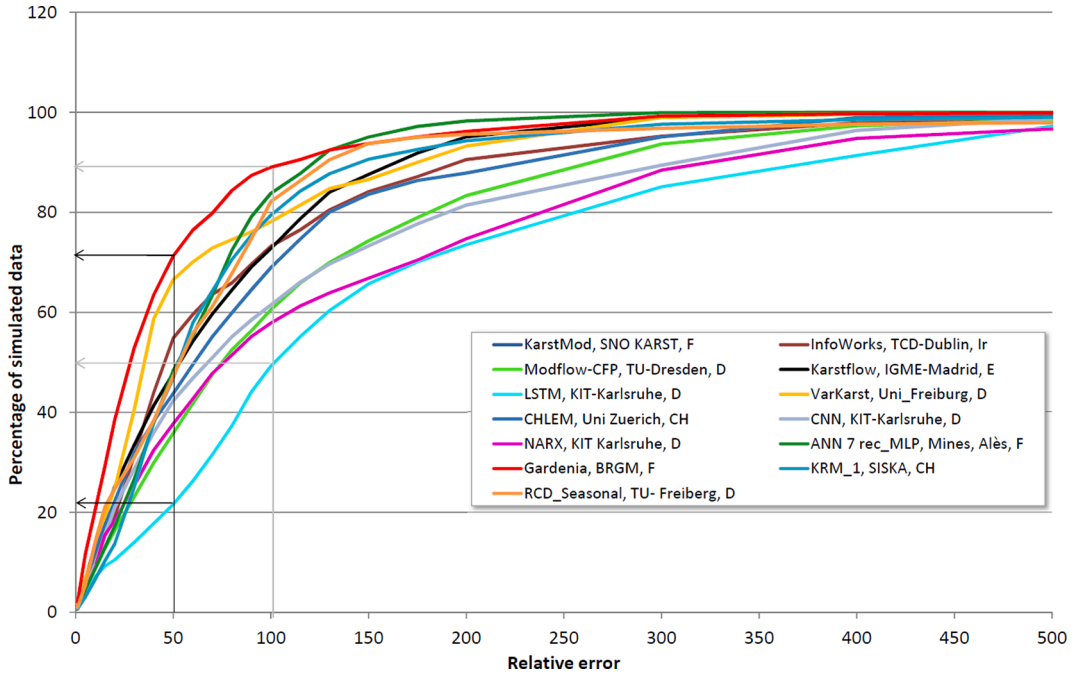

Fig. 10. Percentage of forecasted data (of time) as a function of relative errors. The best model has 72% of forecasted data within a relative error of + -50%. The worse model has only 22% of forecasted data with a relative error lower than 50%. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 5
Part of the total hydrograph in the respective components.

Flow component	Percentage in hours	Percentage in volume	Mean discharge rate
Rising limb	11	21	608 L/s
Recession	33	57	555 L/s
Base flow	9	9	106 L/s
Undetermined	28	13	156 L/s

These models are based on the principle that the transformation of the input signal into the output one is mainly related to the characteristics of the aquifer. This approach is more deductive than for the other models because these models are constructed from the supposed characteristics of the catchment and of the aquifer. A consequence is that the recharge part of the models is simplified. This is the main reason why peaks in the summer season are strongly exaggerated with both models (see Fig. 8). For the given exercise, this approach is also not the most efficient as it requires more work for digitizing a lot of data, but at the end the results include more bias than many other models because recharge was not considered properly. Would a spatial discretization of the model parameters improve the results? Whilst this was not explicitly investigated in this research, we can argue that the main discrepancy of these models is related to (too) large peaks in summer, which seem to be more related to an oversimplification of the recharge model, rather than to the spatialization of the model parameters. Recharge seems therefore really to

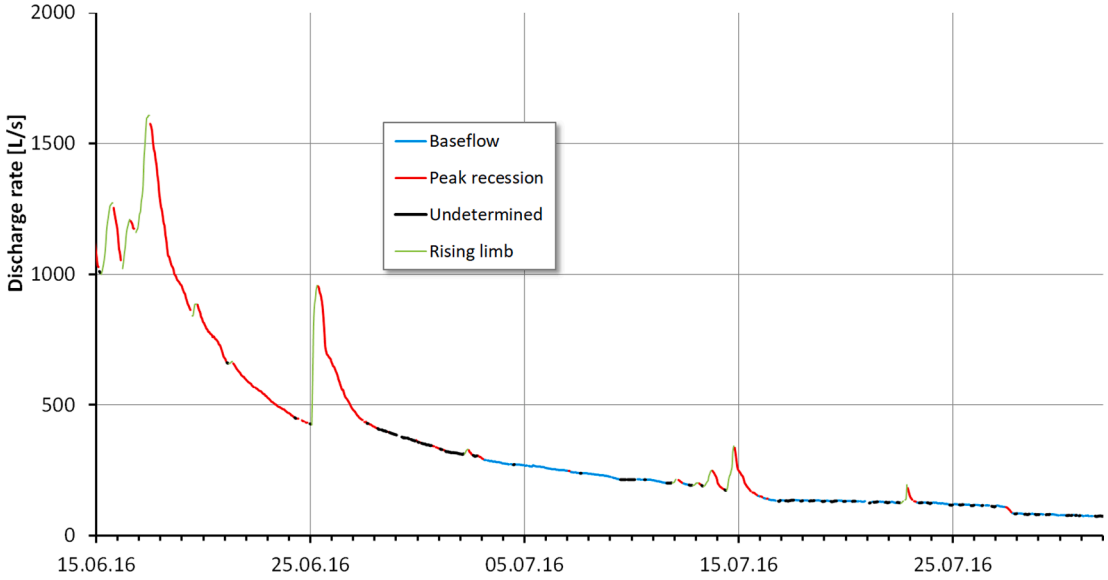

Fig. 11. Decomposition of the time series of observed discharge-rates. The undetermined component mainly corresponds to base flow. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 6
Table showing the three criteria applied to the four components of the 13 models compared for the challenge. Models are sorted according to their global performance (see Table 4).

Flow components		VCC				NASH				KGE			
		Rising limb	Flood recession	Base flow	Undet. Flow	Rising limb	Flood recession	Base flow	Undet. Flow	Rising limb	Flood recession	Base flow	Undet. Flow
BRGM France	Gardenia	0.89	0.83	0.89	0.85	0.80	0.71	0.80	0.83	0.82	0.82	0.83	0.80
Uni-Freiburg	Varkarst	0.95	0.87	0.72	0.73	0.71	0.73	-0.72	0.51	0.70	0.76	0.48	0.66
IMT Mines Alès	ANN / rec_MLP	0.78	0.80	1.06	0.96	0.51	0.44	-0.11	0.52	0.67	0.74	0.32	0.76
SISKA-Switzerland	KRM_1	0.79	0.78	0.78	0.80	0.39	0.50	-0.17	0.74	0.42	0.66	0.49	0.74
IGME Madrid	KarstFLOW	0.79	0.78	0.78	0.80	0.39	0.50	-0.17	0.74	0.42	0.66	0.49	0.74
TCD Dublin	InfoWorks	0.84	0.78	1.01	0.79	0.60	0.35	-0.52	0.23	0.78	0.72	0.67	0.67
KIT-Karlsruhe	CNN	0.83	0.83	1.30	0.98	0.27	0.11	-0.35	0.12	0.72	0.75	0.53	0.83
KIT-Karlsruhe	NARX	0.78	0.82	1.24	1.00	0.24	0.27	-0.21	-0.10	0.65	0.70	0.65	0.71
SNO KARST	KarstMod	0.94	0.68	0.64	0.60	0.71	0.48	-0.42	0.26	0.81	0.67	0.39	0.51
TU-Desden	CFP-modified	0.80	0.67	0.86	0.72	-0.18	-0.04	-0.43	0.15	0.53	0.51	0.73	0.59
TU-Freiberg	RCD-Seasonal	0.57	0.65	0.98	0.86	-0.29	0.02	0.14	0.44	0.26	0.44	0.65	0.63
Uni-Zürich	CHLEM	0.72	0.63	1.04	0.82	-0.29	-0.31	-0.91	-0.12	0.45	0.40	0.63	0.55
KIT-Karlsruhe	LSTM	0.51	0.53	0.96	0.72	-0.19	0.02	-2.39	-2.09	0.36	0.48	0.29	0.35

Legend
within 2%
within 5%
within 15%
15 to 25%
>25%
>0.75
0.5 - 0.75
0.25 - 0.5
0.0 - 0.25
<0.0
>0.85
0.75 - 0.85
0.5 - 0.75
0.25 - 0.5
<0.25

be the key factor in the present case.

On the contrary, lumped parameter models are more inductive because they include less or no constraining hypotheses. The disadvantage is that none of their parameters can be compared somehow with natural characteristics (not even the size of the catchment area!). They provide purely functional results. Another problem is that they require long and continuous data sets for learning the input–output relationship of a natural flow system.

In semi and fully distributed models, the spatial distribution of recharge (land-use, vegetation, precipitation) can be taken into account. However, as the performance of these models turned out to be no better than results of similar models (reservoir) which did not take it into account, one can infer that, at least in the present case-study, the role of spatial distribution of the recharge is low. The effect of spatial distribution was, however, not analysed in any depth and so could be further investigated (e.g., Bittner et al., 2018; Gill et al., 2020). As the Milandre site is rather flat, small and without allogenic recharge, the use of a complimentary catchment area with more contrasted conditions would probably be better for this analysis. It should be noted that the relative significance of the spatial distribution of recharge function in any models will vary according to the characteristics of the karst catchment being simulated: for example, whether there is significant allogenic recharge or not, the size of the catchment area, and/or the steepness of the topography may all be important factors.

An evaluation of the number of parameters hydrologically constrained in a rainfall-runoff model in surface hydrology was investigated by Jakeman and Hornberger (1993). They show that most hydrological systems are well described by models with 4 to 6 parameters. The contribution of any supplementary parameter seems to be within the data uncertainty related to monitoring data. This explains somehow the fact that, in our study, models with many parameters (greater than20) do not provide better results than those with less parameters. In karst, it is difficult to obtain reasonable results with less than 6 parameters: three for the routing and three for the transformation of precipitation to recharge. Because of some visible threshold processes in karst systems, 9 parameters seems to be a more reasonable minimum. Table 1 indicates that some of the models had less than 9 parameters, however they include implicit assumptions fixing some of the parameters (e.g. distribution functions, fixed topology or geometry, etc.). Jakeman and Hornberger (1993) also suggest that a standardized simple model could be applied systematically to any hydrological system. This would help comparing systems' characteristics and identifying drifts in data related to technical or natural reasons (e.g. climate change). The model comparison presented in our paper could therefore be a good basis for defining such a standardized karst model.

7. Conclusion and outlook

7.1. Conclusion

Before making a conclusion, it is important to have in mind that the comparison between the models was only carried out a one-year simulation period. Due to the non-stationary nature of hydrologic phenomena, some models could be very good for one kind of situation and not so good for others. The presented interpretations must thus be considered as "snapshot" interpretations.

Compared to a similar exercise conducted for surface catchment (Holländer et al., 2009) all models performed reasonably well! Although they are based on strongly different modelling approaches all models lead to reasonable and somehow similar results for the proposed exercise (input–output hydrological modelling). Despite considerable simplifications with respect to reality, Gardenia provides excellent results in the present case. It fits the observed data almost within the range of uncertainty on the data themselves. However, despite very good scores, the forecasted hydrograph displays significant differences with the observed one, especially in low water conditions. This is even more obvious for results issued from VarKarst model.

Apart from the different aspects of the modelling approach it can also be concluded from the modelling exercise that recharge is crucial in the modelling of any precipitation-discharge relationship. More specifically it can be observed that the real evapotranspiration must be carefully approached, otherwise models will fail at reproducing discharge rates from precipitation data, at least in temperate regions. It can even be inferred that hydrogeological processes taking place within the aquifer are subsidiary compared to recharge processes taking place in the soil and epikarst.

Lumped parameters models are interesting options if they can include non-linear functions and adequate parameters for taking evapotranspiration into account. Regarding ANN models, their limitation is that they require long and complete datasets for calibration and validation, and that they usually do not provide any physical indication about the real system, even if this "knowledge extraction" has been successfully attempted (as suggested by Kong-A-Siou et al. (2013) and Darras et al. (2015)) these models remain mainly functional. Taking into account the limited duration covered by the available data, several ANNs show an acceptable behavior.

From the comparison, as well as from tests of some models on other sites, it also seems that the order and priority in which the calibration of the models is carried out has a strong effect on the result. Calibrating base-flow first and peak-flow afterwards appears to improve the results. This point was not properly investigated in this study and should be further examined. Some physical reasons could support this idea as storage characteristics of karst aquifer are probably relevant concerning base-flow. As this part of the hydrograph is more independent on recharge processes, it may be meaningful to calibrate the storage components of the models first using one or two storage reservoirs. This is also supported by ideas derived from Kovacs et al. (2005). Then the recharge model must focus on modelling the difference between precipitation and the assessed base-flow.

It should also be highlighted here that the scoring scheme used to evaluate of strengths and weaknesses needs to be adapted for the purpose for the model is being developed for. Models used to provide flood assessments may not be assessed in the same way as those models focussing on the implications of low flow/drought conditions. The scores for the respective flow components (Table 5) shows the differences between the respective models.

It should be noted here that some of the modelling approaches could

be combined in order to use their respective strengths. For instance, semi-distributed models could be used in combination with neural methods in order to improve their recharge assessment component. Such optimizations will need to be adjusted to specific conditions of the site, available data and question to be addressed.

7.2. Outlook

In the case of the Milandre Catchment area, only one model (KRM1) within the challenge attempted to take the spatial distribution of evapotranspiration (land-use) into account. This did not improve the model results in a measurable way. However, as real evapotranspiration is very significant, and as we know that this differentiation is necessary in catchments with stronger altitudinal gradients, it would be meaningful to further investigate this aspect.

This report covers the very first step of a model comparison. It was focused only on the temporal discharge response of the spring to recharge events. We would like to expand the challenge to further aspects of the modelling of karst hydrogeology. The next step is to take into account the spatial distribution of flow and heads within the aquifer. A first question with this respect is the fact that the discharge rate given at the spring for the challenge in reality corresponds to the flow-rate coming out of three different springs: Font, Saivu and Bâme, which are located several hundred meters apart. We also have head measurements at various locations within the aquifer, and it would be interesting to see how various modelling approaches could reproduce those data. For this new exercise, it seems that lumped models will be hardly usable, and that distributed models will be more adequate. It will be therefore interesting to see how the respective teams will adapt their approach for this new challenge.

Once heads and discharge rates will be approached, we will try to simulate flow velocities and some transport processes for which we also acquired data. This may also result in better performance of the models to the overall flows simulated in step 1 as some parameters will be better constrained.

CRediT authorship contribution statement

Pierre-Yves Jeannin: Conceptualization, Methodology, Formal analysis, Investigation, Resources, Data curation, Writing - original draft, Visualization, Supervision, Project administration. **Guillaume Artigue:** Software, Validation, Formal analysis, Investigation. **Christoph Butscher:** Software, Validation, Formal analysis, Writing - review & editing. **Yong Chang:** Software, Validation, Formal analysis. **Jean-Baptiste Charlier:** Methodology, Software, Validation, Formal analysis, Writing - review & editing. **Lea Duran:** Software, Validation, Formal analysis. **Laurence Gill:** Methodology, Software, Validation, Formal analysis, Writing - review & editing. **Andreas Hartmann:** Methodology, Software, Validation, Formal analysis, Writing - review & editing. **Anne Johannet:** Methodology, Software, Validation, Formal analysis, Writing - review & editing. **Hervé Jourde:** Software, Validation, Formal analysis. **Alireza Kavousi:** Software, Validation, Formal analysis, Writing - review & editing. **Tanja Liesch:** Software, Validation, Formal analysis. **Yan Liu:** Software, Validation, Formal analysis. **Martin Lüthi:** Software, Validation, Formal analysis. **Arnauld Malard:** Methodology, Software, Validation, Formal analysis, Writing - review & editing. **Naomi Mazzilli:** Software, Validation, Formal analysis, Writing - review & editing. **Eulogio Pardo-Igúzquiza:** Methodology, Software, Validation, Formal analysis, Writing - review & editing. **Dominique Thiéry:** Software, Validation, Formal analysis. **Thomas Reimann:** Methodology, Software, Validation, Formal analysis, Writing - review & editing. **Philip Schuler:** Software, Validation, Formal analysis. **Thomas Wöhling:** Software, Validation, Formal analysis. **Andreas Wunsch:** Methodology, Software, Validation, Formal analysis, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Swiss National Research Foundation (SNF) for Swisskarst and Thermokarst projects (Grant Numbers 406140-125962 and 200021_188636), the Deutsche Forschungsgemeinschaft, DFG for the iKarst project (Grant Numbers: LI 727/31-1 and RE 4001/2-1). The work of E. Pardo-Igúzquiza was supported by research project PID2019-106435 GB-I00 of the Ministerio de Ciencia e Innovación of Spain. This TCD-Dublin was conducted within the Irish Centre for Research in Applied Geosciences (ICRAG), supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number 13/RC/2092. The authors would like to thank the French Karst National Observatory Service (SNO KARST) initiative at the INSU/CNRS for their diffusion of KarstMod platform. It strengthens dissemination of knowledge and promote cross-disciplinarily research on karst systems at the French national scale.

The data used in the project were acquired along several projects supported by the Federal Roads Office (FEDRO) and by the University of Neuchâtel. Marc Hessenauer (MFR SA, Delémont) and Pierre-Xavier Meury (Géo & Environment, Delémont) also participated to the data collection. The caving club (Spéléo-Club Jura was also very supportive). We also would like to thank Prof. Carol Wicks and an anonymous reviewer as well as Dr. Junbing Pu (associate editor) for their constructive comments.

Appendix. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhydrol.2021.126508.

References

Anderson M. P., Woessner W. W., Hunt R. J., 2015. Applied Groundwater Modelling, Simulaton of Flow and Advective Transport. Elsevier, Amsterdam: 564 p.

Annable, W.L., Sudicky, E.A., 1998. Simulation of karst genesis: hydrodynamic and geochemical rock-water interactions in partially-filled conduits. Bull. Hydrogéol. 16, 211–221.

Atkinson, T.C., 1977. Diffuse flow and conduit flow in limestone terrain in the Mendip Hills, Somerset (Great Britain). J. Hydrol. 35 (1–2), 93–110.

Artigue G., Johannet A., Borrell V., Pistre S., 2012. Flash flood forecasting in poorly gauged basins using neural networks: case study of the Gardon de Mialet basin (southern France). — Nat. Hazards Earth Syst. Sci. 12, 3307–3324. 10/f4gt8k.

Bakalowicz, M., 2005. Karst groundwater: a challenge for new resources. Hydrogeol. J. 13, 148–160.

Barron, A.R., 1993. Universal approximation bounds for superpositions of a sigmoidal function. Inf. Theory IEEE Trans. On 39 (3), 930–945.

Bengio Y., Simard P., Frasconi P., 1994. Learning long-term dependencies with gradient descent is difficult, IEEE Trans. Neural Netw., 5(2): 157–166, 10/bmf3dr, 1994.

Bezes, C., 1976. Contribution à la modélisation des systèmes aquifères karstiques. Université des Sciences et Techniques du Languedoc, Montpellier, Thèse.

Biondi, D., Freni, G., Iacobellis, V., Mascaro, G., Montaniri, A., 2012. Validation of hydrological models: Conceptual basis, methodological approaches and a proposal for a code of practice. Phys. Chem. Earth. 42–44, 70–76.

Bittner D., Narany T. S., Kohl B., Disse M., Chiogna G., 2018. Modeling the hydrological impact of land use change in a dolomite-dominated karst system. Journal of Hydrology, 567, December 2018: 267-279.

Boegli A., 1980. Karst Hydrology and Physical Speleology. Springer Verlag: 284 p.

Bonacci O., 1987. Karst hydrology, with Special Reference to the Dinaric Karst. Springer Verlag: 184 p.

Borghi, A., Renard, P., Cornaton, F., 2016. Can one identify karst conduit networks geometry and properties from hydraulic and tracer test data? Adv. Water Resour. 90, 99–115.

Bredehoeft, J.D., Pinder, G.F., 1970. Digital analysis of areal flow in multi-aquifer groundwater systems. A quasi three-dimensional model. Water Resour. Res. 6 (3), 883–888.

Brown, M.C., 1970. Karst hydrology of the lower Maligne Basin. Jasper, Alberta, Cave Stud. 13, 179–193.

Brown, M.C., 1973. Mass balance and spectral analysis applied to karst hydrologic networks, Water Resour. Res. 9, 749–752.
Butscher, C., Huggenberger, P., 2008. Intrinsic vulnerability assessment in karst areas: a numerical modeling approach. Water Resour. Res. 44, W03408.
Campbell, C.W., Sullivan, S.M., 2002. Simulating time-varying cave flow and water levels using the Storm Water Management Model. Eng. Geol. 65, 133–139.
Castany G., 1968. Prospection et exploitation des eaux souterraines: Dunod P. (Ed.): 718 p.
Chang, Y., Wu, J., Jiang, G., Liu, L., Reimann, T., Sauter, M., 2019. Modelling spring discharge and solute transport in conduits by coupling CFPv2 to an epikarst reservoir for a karst aquifer. —. J. Hydrol. 569, 587–599. https://doi.org/10.1016/j.jhydrol.2018.11.075.
Chemin J., 1974. Essai d'application d'un modèle mathématique conceptuel au calcul du bilan hydrique de l'aquifère karstique de la source du Lez (region N de Montpellier). Thèse, Université de Montpellier: 67 p.
Chen, Z., Goldscheider, N., 2014. Modeling spatially and temporally varied hydraulic behavior of a folded karst system with dominant conduit drainage at catchment scale, Hochifen-Gottesacker, Alps. J. Hydrol. 514, 41–52.
Coutouis A., Johannet A., Pistre S., Ayral P.-A., Cadilhac L. 2016. Towards a neural networks-based prediction tool devoted to low water-levels forecasting. Case study on the Méjannes-le-Clap karst aquifer (France). — 8th International Congress on Environmental Modelling and Software (iEMSs), Jul 2016, Toulouse, France. ⟨hal-02127821⟩.
Darcy H., 1856. Les fontaines publiques de la ville de Dijon : exposition et application des principes à suivre et des formules à employer dans les questions de distribution d'eau, Paris, Victor Dalmont,.
Darras, T., Estupina, V.B., Kong-A-Siou, L., Vayssade, B., Johannet, A., Pistre, S., 2015. Identification of spatial and temporal contributions of rainfalls to flash floods using neural network modelling: case study on the Lez basin (southern France). Hydrol. Earth Syst. Sci. 19 (10), 4397–4410.
Dewandel, B., Lachassagne, P., Bakalowicz, M., Weng, Ph, Al-Malki, A., 2003. Evaluation of 17 aquifer thickness by analysing recession hydrographs. Application to the Oman ophiolite 18 hard-rock aquifer. J. Hydrol. 74 (2), 248–269.
Dodge, E.D., 1983. Hydrogéologie des aquifères karstiques du Causse Comtal (Aveyron, France) Thèse Sci. Univ, Bruxelles.
Doherty, J., 2015. Calibration and Uncertainty Analysis for Complex Environmental Models. Watermark Numerical Computing, Brisbane, Australia.
Doherty, J., Brebber, L., Whyte, P., 1994. Pest: Model-independent parameter estimation. Watermark Computing, Corinda, Australia 122. https://doi.org/10.1016/j.jhydrol.2011.02.004.
Dreiss, S.J., 1982. Linear kernels for karst aquifers. Water Resour. Res. 18 (4), 865–876.
Dreiss, S.J., 1983. Linear unit-response functions as indicators of recharge areas for large karst springs. J. Hydrol. 61, 31–44.
Dreyfus, G., 2005. Neural Networks: Methodology and Applications, Softcover reprint of hardcover, 1st ed, 2005 edition. Springer, Berlin, New York.
Drogue, C., 1980. Essai d'identification d'un type de structure de magasins carbonatés fissurés. Application à l'interprétation de certains aspects du fonctionnement hydrogéologique. Mémoires hors série Société Géologique de France 11, 101–108.
Duran, L., Gill, L.W., 2021. Modeling spring flow of an Irish karst catchment using Modflow-USG with CLN. J. Hydrol. 597 https://doi.org/10.1016/j.jhydrol.2021.125971.
Edijatno, C., Michel, C., 1989. Un modèle pluie-débit à trois paramètres. La Houille Blanche 2, 113–121.
Enemark, T., Peeters, L.J., Mallants, D., Batelaan, O., 2019. Hydrogeological conceptual model building and testing: a review. J. Hydrol. 569, 310–329. https://doi.org/10.1016/j.jhydrol.2018.12.007.
Fiorillo, F., 2014. The recession of spring hydrographs, focused on karst aquifers. Water Resour. Manage. 28, 1781–1805. https://doi.org/10.1007/s11269-014-0597-z.
Ford D., Williams P. W., 1989. Karst Geomorphology and Hydrology. Chapman & Hall: 601 p.
Forkasiewicz, J., Paloc, H., 1967. Le régime de tarissement de la Foux de la Vis. La Houille Blanche (BRGM) 1, 29–36.
Ghasemizadeh, R., et al., 2012. Review: Groundwater flow and transport modeling of karst aquifers, with particular reference to the North Coast Limestone aquifer system of Puerto Rico. Hydrogeol. J. 20, 1441–1461.
Gill, L.W., Naughton, O., Johnston, P.M., 2013. Modelling a network of turloughs in lowland karst. Water Resour. Res. 49 (6), 3487–3503.
Gill, L.W., Schuler, P., Duran, L., Johnston, P.M., Morrissey, P., 2020. An evaluation of semi-distributed-pipe-network and distributed-finite-difference models to simulate karst systems. Hydrogeology Journal 29, 259–279. https://doi.org/10.1007/s10040-020-02241-8.
Grasso, A., Jeannin, P.Y., 1994. Etude critique des méthodes d'analyse de la réponse globale des systèmes karstiques. Application au site de Bure (JU, Suisse). Bulletin d'Hydrogéologie 13, 87–113.
Guinot V., Savéan M., Jourde H., Neppel L., 2015. Conceptual rainfall-runoff model with a two-parameter, infinite characteristic time transfer function. Hydrological Processes, Wiley, 2015, 29 (22): 4756-4778.
Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the Mean Squared Error and NSE Performance Criteria: Implications for Improving Hydrological Modeling. J. Hydrol. 377, 80–91.
Halihan, T., Wicks, C.M., 1998. Modeling of storm responses in conduit flow aquifers with reservoirs. J. Hydrol. 208, 82–91.
Hartmann A., Lange J., Weiler M., Arbel Y., & Greenbaum N., 2012. A new approach to model the spatial and temporal variability of recharge to karst aquifers. Hydrology and Earth System Sciences, 16(7), 2219–2231. 10.5194/hess-16-2219-2012.
Hartmann, A., Barberá, J.A., Lange, J., Andreo, B., Weiler, M., 2013. Progress in the hydrologic simulation of time variant recharge areas of karst systems – Exemplified at a karst spring in Southern Spain. Adv. Water Resour. 54 https://doi.org/10.1016/j.advwatres.2013.01.010.
Hartmann, A., Mudarra, M., Andreo, B., Marín, A., Wagener, T., Lange, J., 2014. Modeling spatiotemporal impacts of hydroclimatic extremes on groundwater recharge at a mediterranean karst aquifer. Water Resour. Res. 50 (8) https://doi.org/10.1002/2014WR015685.
Hartmann, A., Liu, Y., Olarinoye, T., Berthelin, R., Marx, V., 2021. Integrating field work and large-scale modeling to improve assessment of karst water resources. Hydrogeol. J. 29 (1), 315–329. https://doi.org/10.1007/s10040-020-02258-z.
Hochreiter S., Schmidhuber J., 1997. Long Short-Term Memory, Neural Computation, 9 (8): 1735–1780, 10/bxd65w, 1997.
Holländer H. M., Blume T., Bormann H., Buytaert W., Chirico G. B., Exbrayat J.-F., Gustafsson D., Hölzel H., Kraft P., Stamm C., Stoll S., Blöschl G., Flühler H., 2009. Comparative predictions of discharge from an artificial catchment (Chicken Creek) using sparse data, Hydrol. Earth Syst. Sci., 13: 2069–2094, doi 10.5194.
Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. Neural Netw. 2 (5), 359–366.
Jakeman, A.J., Hornberger, G.M., 1993. How Much Complexity Is Warranted in a Rainfall-Runoff Model? –. Water Resour. Res. 29 (8), 2637–2649.
Jeannin, P.Y., Malard, A., Rickerl, D., Weber, E., 2015. Assessing karst-hydraulic hazards in tunneling – the Brunnmühle spring system – Bernese Jura, Switzerland. –. Environ. Earth Sci. 74 (12), 7655–7670.
Jeannin P.-Y., 1996. Structure et comportement hydraulique des aquifères karstiques. – Thèse Sc. Univ. Neuchâtel : 270 p.
Jeannin, P.-Y., 2001. Modelling flow in phreatic and epiphreatic karst conduits in the Hölloch Cave (Muotathal, Switzerland). —. Water Resour. Res. vol. 37 (2), 191–200.
Jeannin P.-Y., Maréchal J.-C., 1995. Lois de Pertes de charge dans les conduits karstiques. Base théorique et observations. – Bulletin d'Hydrogéologie No 14, Neuchâtel : 149-176.
Kaufmann, G., Gabrovsek, F., Turk, J., 2016. Modelling flow of subterranean Pivka river in Postojnska Jama, Slovenia. Acta Carsol. 45 (1), 57–70.
Kavousi, A., Reimann, T., Liedl, R., Raeisi, E., 2020. Karst aquifer characterization by inverse application of MODFLOW-2005 CFPv2 discrete-continuum flow and transport model. —. J. Hydrol. 587 https://doi.org/10.1016/j.jhydrol.2020.124922.
Király, L. 1975. Rapport sur l'état actuel des connaissances dans le domaine des caractères physiques des roches karstiques. Hydrogeology of karstic terrains. In : A. Burger & L. Dubertret (eds), int. Union of geol. Sciences, B, 3 : 53.-67.
Király L., Morel G., 1976. Etude de régularisation de l'Areuse par modèle mathématique. Bulletin d'Hydrogéologie de l'Université de Neuchâtel, 1: 19-36.
Király, L., Perrochet, P., Rossier, Y., 1995. Effect of the epikarst on the hydrograph of karst springs: a numerical approach. Bull. d'hydrogéologie 14, 199–220.
Kong-A-Siou, L., Cros, K., Johannet, A., Borrell, Estupina V., Pistre, S., 2013. KnoX method, or Knowledge eXtraction from neural network model. Case study on the Lez karst aquifer (southern France). J. Hydrol. 507, 19–32.
Kovács, A., 2003. Estimation of conduit network geometry of a karst aquifer by the means of groundwater flow modeling (Bure, Switzerland). Bol. Geol. Min. 114 (2), 183–192.
Kovacs, A., Jeannin, P.-Y., 2003. Hydrogeological overview of the Bure Plateau, Ajoie, Switzerland. — Ecolgae geol. Helv. 96, 367–379.
Kovács, A., Perrochet, P., Kiraly, L., Jeannin, P.-Y., 2005. A quantitative method for the characterisation of karst aquifers based on spring hydrograph analysis. —. J. Hydrol. 303, 152–164.
Kovács A., Sauter M., 2007. Modelling karst hydrodynamics. In: D. Goldscheider N & Drew, ed. Methods in Karst Hydrogeology, Taylor & Francis, London: 201-222.
Lauritzen, S.-E., Abott, J., Arnesen, R., Crossley, G., Grepperud, D., Ive, A., 1985. Morphology and hydraulics of an active phreatic conduit. Cave Sci. 12 (4), 139–146.
Liedl, R., Sauter, M., Hückinghaus, D., Clemens, T., Teutsch, G., 2003. Simulation of the development of karst aquifers using a coupled continuum pipe flow model. Water Resour. Res. 39, 1–11. https://doi.org/10.1029/2001WR001206.
Lièvre L., 1915. Le problème hydrologique de la Haute Ajoie et le Creux-Genaz. Contribution à l'étude de la circulation souterraine en terrains calcaires. Actes de la Société Jurassienne d'émulation, années 1915-1916 : 75-111.
Lièvre L., 1940. Le karst jurassien. Hydrologie de la Haute-Ajoie et découverte d'une rivière souterraine du Jura bernois. — Le Jura S.A. imprimerie, Porrentruy: 159 p.
Lin, T., Horne, B.G., Tiňo, P., Giles, C.L., 1996. Learning long-term dependencies in NARX recurrent neural networks. —. IEEE Trans. Neural Networks 7 (6), 1329–1338.
Lin T., Horne B. G., Tiňo P., Giles C. L., 1996b. Learning long-term dependencies is not as difficult with NARX networks. — Advances in Neural Information Processing Systems: 577–583.
Liu, Y., Wagener, T., Hartmann, A., 2021. Assessing streamflow sensitivity to precipitation variability in karst-influenced catchments with unclosed water balances. Water Resour. Res. 57 (1) https://doi.org/10.1029/2020WR028598.
Long, A.J., 2009. Hydrograph separation for karst watersheds using a two-domain rainfall–discharge model, —. J. Hydrol. 364 (3–4), 249–256.
Madsen, H., Wilson, G., Ammentorp, H.C., 2002. Comparison of different automated strategies for calibration of rainfall-runoff models. —. J. Hydrol. 261, 48–59.
Maillet E. T., 1905. Essais d'hydraulique souterraine et fluviale. — Hermann A. (Ed.) Paris : 218 p.
Malard A., 2018. Hydrogeological characterization of karst aquifers in Switzerland using a pragmatic approach. – PhD-Thesis, University of Neuchâtel, Switzerland: 369 p.
Mangin A., 1970. Contribution à l'étude d'aquifères karstiques à partir de l'analyse de courbes de décrue et de tarissement. — Annales de Spéléologie, Volume 25(3): 581-609.

Mangin A., 1975. Contribution à l'étude hydrodynamique des aquifères karstiques. — Annales de Spéléologie, 19 (3, 29 (4), 30 (1) : 172 p.
Martel E.-A., 1921. Le Nouveau traité des eaux souterraines. — Paris, 0. Doin, G. Doin : 838 p.
Mazzilli, N., Bertin, D., Guinot, V., Jourde, H., Lecoq, N., Labat, D., Arfib, B., Baudement, C., Danquigny, C., Dal Soglio, L., 2017. KarstMod: a modelling platform for rainfall - discharge analysis and modelling dedicated to karst systems. —. Environ. Modell. Software 122, 10.1016.
Mazzilli N., Guinot V., Jourde H., Lecoq N., Labat D., et al., 2019. KarstMod: A modelling platform for rainfall - discharge analysis and modelling dedicated to karst systems. — Environmental Modelling and Software, Elsevier, 2019: 122 p.
Milanovic P., 1976. Water regime in dee karst: case study of Ombla Spring drainage area. — in: Yevjevich V. (ed.). Karst Hydrology and Water Resources, Vol. 1, Karst Hydrology, Fort Collins, Colorado, Water Resources Publications: 165–186.
Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D., et al., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Soil & Water. — Am. Soc. Agric. Biol. Eng. 50 (3), 885–900.
Morrissey, P., McCormack, T., Naughton, O., Johnston, P.M., Gill, L.W., 2020. Modelling groundwater flooding in a lowland karst catchment. —. J. Hydrol. 580.
Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I — A discussion of principles. —. J. Hydrol. 10 (3), 282–290.
Pardo-Igúzquiza, E., Dowd, P., Pulido-Bosch, A., Luque-Espinar, J.A., Heredia, J., Durán-Valsero, J.J., 2018. A parsimonious distributed model for simulating transient water flow in a high-relief karst aquifer. —. Hydrogeol. J. 26 (8), 2617–2627.
Penman H. L., 1948. Natural evaporation from open water, bare soil, and grass. — Proc. Roy. Soc., London, U.K., vol. A193, no 1032, 1948: 120–145.
Perrin J., 2003. A conceptual model of flow and transport in a karst aquifer based on spatial and temporal variations of natural tracer. — PhD-Thesis, University of Neuchâtel: 227 p.
Perrin, J., Jeannin, P.-Y., Zwahlen, F., 2003. Epikarst storage in a karst aquifer : a conceptual model based on isotopic data, Milandre test site, Switzerland. J. Hydrol. 279, 106–124.
Peterson, E.W., Wicks, C.M., 2006. Assessing the importance of conduit geometry and physical parameters in karst systems using the storm water management model (SWMM). —. J. Hydrol. 329, 294–305.
Pianosi, F., Beven, K., Freer, J., Hall, J., Rougier, J., Stephenson, D., Wagener, T., 2016. Sensitivity analysis of environmental models: A systematic review with practical workflow. —. Environ. Modell. Software vol. 79, 214–232. https://doi.org/10.1016/j.envsoft.2016.02.008.
Rango, A., Martinec, J., 1995. Revisiting the degree-day method for snowmelt computations. —. J. Am. Water Resour. Assoc. 31, 657–669. https://doi.org/10.1111/j.1752-1688.1995.tb03392.x.
Reimann, T., Giese, M., Geyer, T., Liedl, R., Maréchal, J.C., Shoemaker, W.B., 2014. Representation of water abstraction from a karst conduit with numerical discrete-continuum models. — Hydrol. Earth Syst. Sci. 18, 227–241. https://doi.org/10.5194/hess-18-227-2014.
Reimann T., Liedl R., Birk S., Bauer S., 2018. Modifications and enhancements to CFPM1 flow subroutines and addition of transport subroutines. — Accessible at: http://tu-dresden.de/die_tu_dresden/fakultaeten/fakultaet_forst_geo_und_hydrowissenschaften/fachrichtung_wasserwesen/igw/forschung/downloads/cfpv2.
Schoeller H., 1962. Les eaux souterraines. — Ed. Masson, Paris : 642 p.
Schuler P., Duran L., Johnston P.M., Gill L.W., 2020. Quantifying and numerically representing recharge and flow components in a karstified carbonate aquifer. — Water Resources Research 56, e2020WR027717.
Sherman, L., 1932. Stream Flow from Rainfall by the Unit Graph Method. — Engineering Newa. Record 108, 501–505.
Shoemaker W.B., Kuniansky E.L., Birk S., Bauer S., Swain E.D., 2008. Documentation of a Conduit Flow Process (CFP) for MODFLOW-2005 A product of the Ground-Water Resources Program. — U.S. Geological Survey. 10.3133/tm6A24.
Smart C. C., 1988. Artificial Tracer Techniques for the Determination of the Structure of Conduit Aquifers. — Groundwater, Vol 26 (4): 445-453.
Smart P. L. & Friederich H. 1986. Water movement and storage in the unsaturated zone of a maturely karstified carbonate aquiter. Mendip Hills, England. – Proceedings of the Conference on Environmental Problems in Karst Terranes and their Solutions, Bowling Green, Kentucky, October 28-30, published by National Water Well Association, Dublin, Ohio : 59-87.
Smart, P.L., Hobbs, S.L., 1986. Characterisation of carbonate aquifers: A conceptual base. – Proceedings of the Conference on Environmental Problems in Karst Terranes and their Solutions, Bowling Green, Kentucky, October 28–30, published by National Water Well Association. 1–14, Dublin, Ohio.
Teutsch G., Sauter M., 1991. Groundwater modeling in karst terranes: Scale effects, data acquisition and field validation. — Third Conference on Hydrogeology, Ecology, Monitoring, and Management of Ground Water in Karst Terranes. National Ground Water Association, Dublin, Ohio: 17-35.
Thierrien, R., Sudicky, E.A., 1996. Three-dimensional analysis of variably saturated flow and solute transport in discretely fractured porous media. J. Contam. Hydrol. 23, 1–44.
Thiéry, D., 2014. Logiciel GARDÉNIA, version 8.2. Guide d'utilisation. — BRGM/RP-62797-FR, 136 p. http://infoterre.brgm.fr/rapports/RP-62797-FR.pdf.
Thiéry D., 2015. Validation du code de calcul GARDÉNIA par modélisations physiques comparatives. — BRGM/RP-64500-FR: 48 p. http://infoterre.brgm.fr/rapports/RP-64500-FR.pdf.
Thornthwaite C. W., 1948. An approach toward a rational classification of climate.— Geographical Review, vol. 38, no 1, 1948: 55–94.
Trombe F., 1948. Traité de Spéléologie. — Paris, Payot: 376 p.
Turc, L., 1961. Evaluation des besoins en eau d'irrigation, évapotranspiration potentielle. Ann. agron 12 (1), 13–49.
Vuilleumier, C., Jeannin, P.-Y., Perrochet, P., 2019. Physics-based fine-scale numerical model of a karst system (Milandre Cave Switzerland). Hydrogeol. J. 27 (7), 2347–2363.
White, W.B., White, E.L., 1970. Channel hydraulics of free-surface streams in caves. Caves Karst 12 (6), 41–48.
Williams, P.W., 1983. The role of the subcutaneous zone in karst hydrology. J. Hydrol. 61, 45–67.
Worthington S. R. H., 1991. Karst hydrogeology of the Canadian Rocky Mountains. — PhD-Thesis, Mc-Master University, Hamilton, Ontario: 370 p.
Wu Y., Jiang Y., Yuan D., Li L., 2008. Modeling hydrological responses of karst spring to storm events: example of the Shuifang spring (Jinfo Mt., Chongqing, China). — Environmental Geology, 55: 1545–1553.