Hydrogeology Journal
https://doi.org/10.1007/s10040-020-02227-6

PAPER

A model ensemble generator to explore structural uncertainty in karst systems with unmapped conduits

Chloé Fandel[1] · Ty Ferré[1] · Zhao Chen[2] · Philippe Renard[3] · Nico Goldscheider[4]

Received: 28 April 2020 / Accepted: 10 August 2020
© The Author(s) 2020

Abstract

Karst aquifers are characterized by high-conductivity conduits embedded in a low-conductivity fractured matrix, resulting in extreme heterogeneity and variable groundwater flow behavior. The conduit network controls groundwater flow, but is often unmapped, making it difficult to apply numerical models to predict system behavior. This paper presents a multi-model ensemble method to represent structural and conceptual uncertainty inherent in simulation of systems with limited spatial information, and to guide data collection. The study tests the new method by applying it to a well-mapped, geologically complex long-term study site: the Gottesacker alpine karst system (Austria/Germany). The ensemble generation process, linking existing tools, consists of three steps: creating 3D geologic models using GemPy (a Python package), generating multiple conduit networks constrained by the geology using the Stochastic Karst Simulator (a MATLAB script), and, finally, running multiple flow simulations through each network using the Storm Water Management Model (C-based software) to reject nonbehavioral models based on the fit of the simulated spring discharge to the observed discharge. This approach captures a diversity of plausible system configurations and behaviors using minimal initial data. The ensemble can then be used to explore the importance of hydraulic flow parameters, and to guide additional data collection. For the ensemble generated in this study, the network structure was more determinant of flow behavior than the hydraulic parameters, but multiple different structures yielded similar fits to the observed flow behavior. This suggests that while modeling multiple network structures is important, additional types of data are needed to discriminate between networks.

Keywords Multi-model ensemble · Structural uncertainty · Alpine hydrogeology · Karst · Groundwater flow

Introduction

Approximately 16.5% of the global population lives on karst (Goldscheider et al. 2020). Karst systems form in carbonate rock when water containing CO_2 gradually dissolves a network of conduits through a fractured rock matrix. Conduits are major pathways for groundwater flow in karst aquifers (Worthington et al. 2012), and conduit flow is often rapid

Published in the special issue "Five decades of advances in karst hydrogeology".

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10040-020-02227-6) contains supplementary material, which is available to authorized users.

✉ Nico Goldscheider
nico.goldscheider@kit.edu

Chloé Fandel
cfandel@email.arizona.edu

Ty Ferré
tyferre@email.arizona.edu

Zhao Chen
chen.zhao@gmx.de

[1] Department of Hydrology & Atmospheric Sciences, University of Arizona, 1133 E. James E. Rogers Way, Rm. 122, Tucson, AZ 85721, USA

[2] Environmental Resources Management, Siemensstr. 9, 63263 Neu-Isenburg, Germany

[3] Centre for Hydrogeology and Geothermics, University of Neuchâtel, Rue Emile-Argand 11, 2000 Neuchâtel, Switzerland

[4] Karlsruhe Institute of Technology, Institute of Applied Geosciences, Kaiserstr. 12, 76131 Karlsruhe, Germany

Published online: 02 October 2020

and turbulent, resulting in complex, heterogeneous behavior very different from porous media (Ford and Williams 2007). These characteristics make karst aquifers vulnerable to impacts from human activity (Drew and Hötzl 1999) and challenging to manage (Fleury et al. 2007).

Numerical models are powerful, widely used predictive tools for groundwater resource management, but the uncertainty associated with model predictions must be taken into account if they are to be used to support decision-making (Doherty and Simmons 2013; Ferré 2017). Primary sources of uncertainty in model predictions are input data, model parameter values, and model structure (i.e. conceptualization; Refsgaard et al. 2006). Structural uncertainty is widely thought to be the primary contributor to prediction uncertainty (Refsgaard et al. 2006; Gupta et al. 2012; Neuman and Wierenga 2003). However, traditional modeling approaches have focused on parameter calibration for a single conceptualization of model structure, which may underestimate uncertainty by inadequately representing the range of plausible structures (Refsgaard et al. 2006; Bredehoeft 2005; Enemark et al. 2019). A multi-model approach, using an ensemble of competing model structures to generate a range of predictions, has been proposed to address some of these issues (Neuman and Wierenga 2003; Enemark et al. 2019; Clark et al. 2011), but has yet to be widely used in karst systems.

Applying numerical models in karst aquifers is challenging (Scanlon et al. 2003), exacerbating the difficulties associated with conceptualizing model structure. Existing approaches to karst modeling can be generally classified as lumped or distributed (Kovács and Sauter 2007). Lumped models conceptualize physical processes at the scale of the whole karst system without explicitly modeling spatial variability, based on the relationship between inflow and outflow time series (Hartmann et al. 2014). Such models can represent the overall water balance and dynamics of the system, but not the spatial variability in hydraulic head, or the directions and rates of groundwater flow (Scanlon et al. 2003).

Distributed models discretize the model domain and apply spatially variable hydraulic parameters to each cell. These models are capable of representing the spatial dimensions of groundwater flow, but require far more data: spatial information about matrix and conduit hydraulic properties, aquifer geometry, and/or conduit geometry (Hartmann et al. 2014). This often-prohibitive limitation has restricted the use of spatially distributed models primarily to either synthetic cases, or well-understood, previously studied systems (Chen and Goldscheider 2014). However, even in extensively studied systems, the conduit network is impossible to map fully, because small-diameter conduits (<0.3 m) are inaccessible and difficult to detect by geophysical methods (Jaquet and Jeannin 1994); thus, distributed models of real systems are therefore inherently always incomplete. One relatively new approach to resolving this difficulty is the development of stochastic conduit evolution models, which generate probable network maps of real-world karst systems, based on the geologic setting. These networks can then be used as the basis for distributed flow and transport models (Borghi et al. 2012). The ability to quickly generate many networks constrained by the same geologic context makes it possible to use this technique for multi-model approaches to simulating karst system behavior, though it has only been used in a limited number of studies to date (Borghi et al. 2016; Sivelle et al. 2020).

This study builds on the stochastic conduit evolution modeling approach for distributed karst models. The goals of the study are:

1. To develop a multi-model ensemble approach for karst systems, capable of generating many mutually exclusive and collectively exhaustive plausible models of the same system, using minimal initial data.
2. To compare the influence of uncertainty in the structure versus uncertainty in the parameters on the uncertainty of model predictions.
3. To identify additional data needs to reduce prediction uncertainty.

Study area: Gottesacker karst system

Location, climate, geologic and hydrogeologic setting

The study area is a 35 km^2 catchment located in the Northern Alps, on the border between Germany and Austria (Fig. 1a), with elevations ranging from 1,000 to 2,230 m above sea level (asl). The climate is cool, temperate, and humid, with a mean annual temperature of 5.7 °C and mean annual precipitation of 1,836 mm. Maximum rainfall occurs in June–August, while snow accumulates between November and May; Water Authority Vorarlberg data, 1961–1990).

Hydrogeologically, the catchment consists of a large karstified zone to the north, with primarily subsurface flowpaths, and a smaller, nonkarst zone of flysch to the south, generating surface runoff. The flysch zone consists of several sandstone, mudstone, and marl formations. Both zones drain to the Schwarzwasser Valley, which runs northeast and marks the boundary between karst and nonkarst (Fig. 1b). Two parallel flow systems run along the valley axis: a surface stream that collects water from the flysch zone, and an underground stream (i.e. a series of karst conduits) that collects water from the karst zone and infiltration from the surface stream. The two streams are connected by an estavelle (which is an orifice that may act either as an outlet or an inlet depending on groundwater conditions; QE), which demarcates the boundary between the upper and lower segments of the valley. The upper segment of the Schwarzwasser Stream receives surface

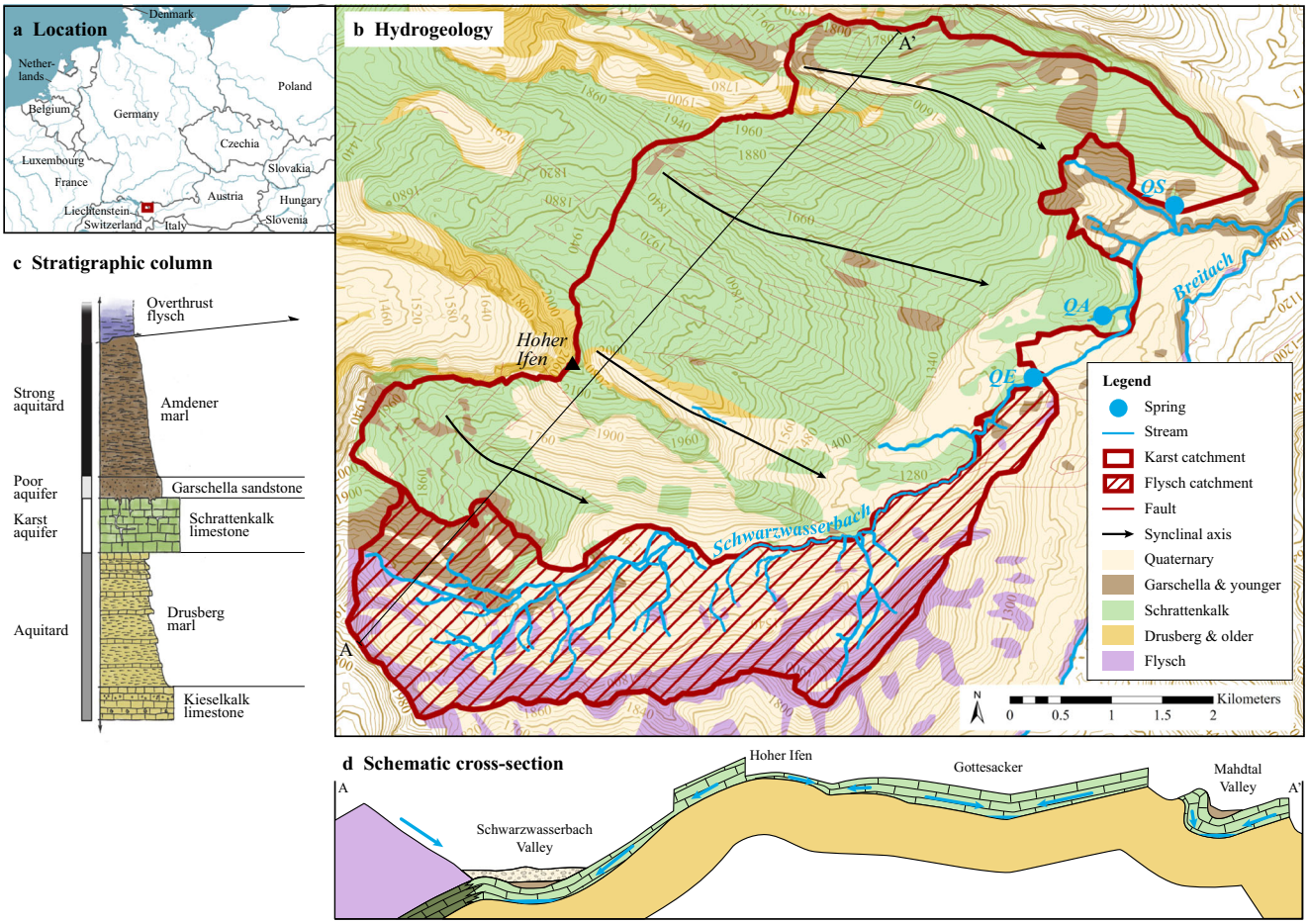

Fig. 1 Hydrogeology of the Gottesacker karst system. **a** Location (modified from European Environment Agency Large Rivers and Large Lakes basemap 2017). **b** Simplified hydrogeology. The area is folded into a series of SE-plunging synclines directing water flow towards the Schwarzwasser Valley (compiled by C. Fandel based on Goldscheider 2005 and Chen et al. 2018). **c** Simplified stratigraphic column. The major karst unit is the Schrattenkalk limestone, bounded above and below by low-permeability formations (compiled by C. Fandel based on Goldscheider 2002 and Goldscheider and Neukum 2010). **d** Simplified geologic cross-section (modified from Goldscheider 2005)

inflow generated from the flysch zone and from a rockfall mass in the uppermost part of the valley, but no contribution from the karst zone. During low-flow conditions, the estavelle acts as a swallow hole, so that the streambed is dry until it receives new inflow from several downstream karst springs, described in the following. During high-flow conditions, the estavelle acts as a karst spring, so that the stream below this point includes a mixture of surface water and karst groundwater.

The karst aquifer developed in the pure, highly fractured and karstifiable Cretaceous Schrattenkalk limestone formation (approximately 100 m thick). The karst aquifer is highly permeable. The Schrattenkalk is underlain by the less permeable Drusberg marl (approximately 250 m thick) acting as an aquitard (Fig. 1c). The karst zone is heavily folded and fractured, resulting in strong structural geologic control on underground flowpaths. Major flowpaths in the karst aquifer follow the axes of plunging synclines, draining towards the southeast, before joining a series of conduits that parallel the Schwarzwasser stream flowing northeast (Goldscheider 2005). The aquifer discharges at three primary springs flowing into the Schwarzwasser stream: the estavelle (QE), described previously, at 1,120 m asl; Aubach Spring (QA), a large, intermittent overflow at 1,080 m asl; and Sägebach Spring (QS), the lowest permanent outlet of the system at 1,035 m asl (Table 1; Goldscheider 2005).

Previous work: spatially distributed numerical model of the study site

To test new modeling approaches, a well-accepted reference model is needed for comparison. Previous work by Chen and Goldscheider (2014) simulated the study catchment using a spatially distributed numerical conduit flow model: the Storm Water Management Model (SWMM; Rossman 2015). SWMM is described in more detail in section 'Flow model: SWMM (EPA)'.

Table 1 Major springs included in this study (after Goldscheider 2005, Goeppert and Goldscheider 2008, and Chen and Goldscheider 2014)

Spring	Abbreviation	Elevation (m asl)	Observed discharge (m^3/s)	
			Minimum	Maximum
The estavelle	QE	1,120	–0.5	4
Aubach	QA	1,080	0	8
Sägebach	QS	1,035	0.17	3.5

The conduit network map and the subcatchment boundaries used for the reference model were drawn based on extensive pre-existing geologic (Wagner 1950), speleological (Höhlenverein Sonthofen 2006), and hydrogeologic mapping, including numerous tracer tests (Goldscheider 2005; Goeppert and Goldscheider 2008). Flow inputs to the conduit model were generated by coupling it to a reservoir model representing recharge, storage and transfer of water in the epikarst and unsaturated zone, based on meteorological data from nearby weather stations. Spring discharge during the study period was recorded by four temporary monitoring stations (Chen and Goldscheider 2014). Extensive parameter estimation, sensitivity analyses, and calibration improved this model (Chen et al. 2017, 2018), and the final version is used in this study as a reference (Fig. 2).

Model ensemble generation

The first goal of this study is to generate an ensemble of models that are structurally diverse, representing multiple possible geologic interpretations and conduit networks, yet still plausible based on the hydrogeologic setting. Flow modeling is used to reject nonbehavioral structures. This approach allows the user to choose which aspects of the conceptual model of the system should be varied in generating the structural

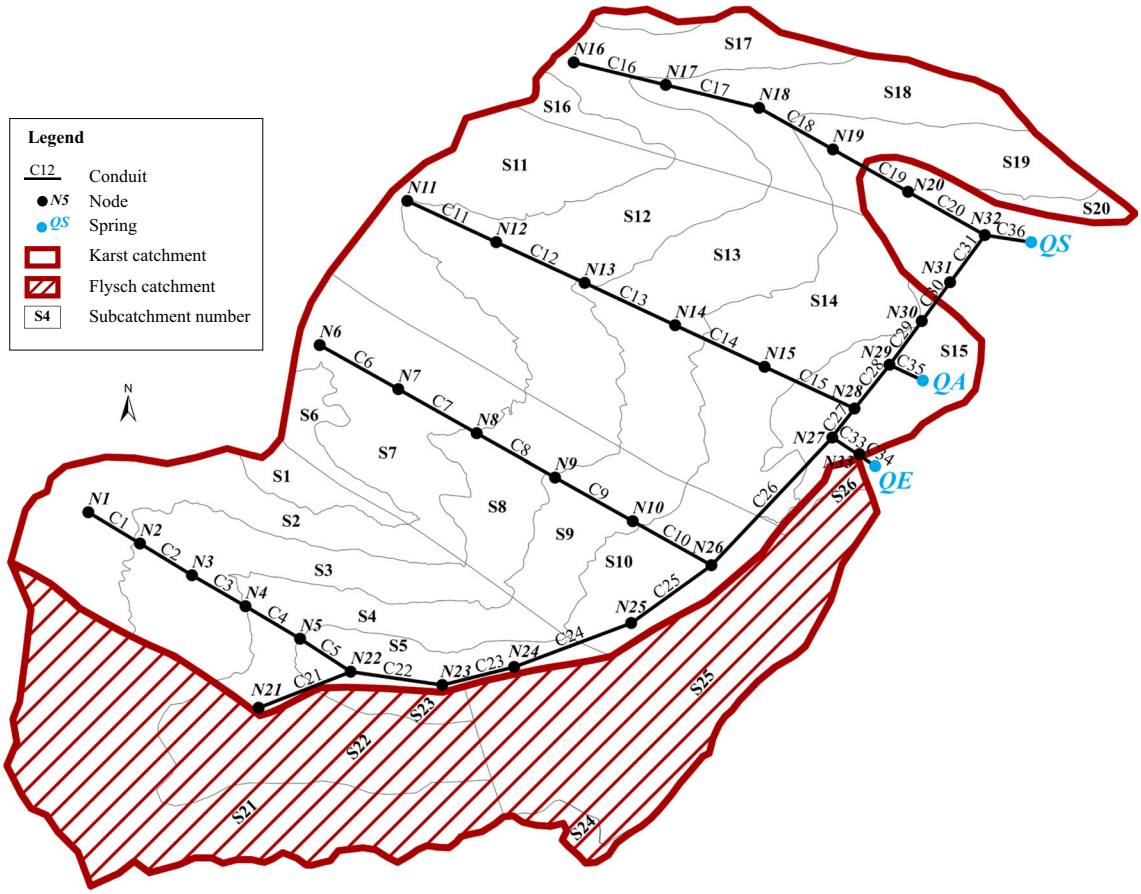

Fig. 2 Reference model developed by Chen et al. 2018. The conduit network was drawn based on a combination of tracer test data, geologic constraints, and hydrogeologic reasoning. Subcatchments were delineated along elevation bands

model ensemble. The total runtime for the entire ensemble, including rejection of nonbehavioral models and parameter exploration, was approximately 2 weeks, run serially on a Lenovo 910 laptop. Each geologic model run took approximately 10 min, conduit network evolution runs took between 1 and 5 min each, and flow model runs took between 2 and 10 min each. Model runtimes could be significantly decreased by parallelization.

Data

For this study, the bulk of data used in previous modeling work was set aside to mimic the more common scenario of a minimally studied catchment. These data (tracer tests, cave maps, and approximate conduit maps) were reserved for future analysis and validation of the methodology. Datasets used in this study were limited to existing geologic and topographic maps, meteorological data, and spring discharge time series. Five overlapping geologic maps (including representative point strike and dip measurements, surface fracture traces, and representative cross-sections) were digitized and compiled in ArcMap (released by Environmental Systems Research Institute 2017) into one detailed map covering the entire catchment (Goldscheider 2005 and related maps; Chen et al. 2018; see the electronic supplementary material (ESM)) Topographic data were obtained as digital elevation model (DEM) raster files, with 50 m × 50 m pixels, from the State of Vorarlberg Digital Atlas. Meteorological data (hourly precipitation, air temperature, relative humidity) were recorded by Chen et al. (2018) from November 2013 to October 2014 at nine weather stations across the catchment, and were interpolated at a 100 × 100 m grid resolution. Discharge was recorded hourly for the same time period at the three major outlets of the system (QE, QA, QS; Chen et al. 2018).

Inflows to the karst system were calculated according to Chen et al. (2018). The meteorological data were fed to a lumped linear reservoir model representing water storage and drainage through the epikarst (Hartmann et al. 2012), paired with the HBV snow routine to account for snow accumulation and melt (Hock 1999). The output is a 100 × 100 m grid of fast and slow recharge time series across the entire catchment, which is then used as input for the groundwater flow model.

Geologic model

The first step in generating a structurally diverse model ensemble is to create a set of plausible three-dimensional (3D) geologic models within which the karst networks will evolve. In mountain catchments, it is generally accepted that the often-complex geologic setting can largely control the system's flow behavior (Goldscheider 2011; Rogger et al. 2013), yet the exact shape of the contact surfaces between units is difficult to fully map, particularly in the subsurface. Representing the geologic structure of the site is therefore extremely important, particularly the shape of the lower boundary of the karstifiable unit, as this is where conduits often form.

The ensemble generator presented in this paper models the geology using an implicit approach (Lajaunie et al. 1997), with the open-source Python package GemPy (de la Varga et al. 2019; see ESM).

First, the stratigraphic sequence of the site was grouped into four major hydrogeologic units, in ascending order from oldest to youngest: a group of underlying low-permeability units, the low-permeability Drusberg marl, the Schrattenkalk limestone aquifer, and a low-permeability cover consisting of the Garschella sandstone and younger units (Fig. 1c). Next, point strike and dip orientation measurements for these units were located on the geologic map and the coordinates (X, Y, Z) were extracted from the DEM. Additional points were hand-picked from the map and from cross-sections to constrain the contact interfaces between units. Then, the model grid was defined based on the DEM, with a resolution of 251 columns × 200 rows × 52 layers, for a total of 2,610,400 cells. The horizontal cell length and width were taken from the 50 m × 50 m DEM, while the vertical cell height was manually selected to be approximately one-quarter of the thickness of the aquifer, or 27.5 m. The vertical resolution can be refined for more accuracy, but this significantly increases the computation time. The geologic model intentionally extends beyond the catchment boundaries to avoid edge effects, and was subsequently cropped.

GemPy interpolates the shape of the contact surfaces between units based on the stratigraphic sequence, and the orientation and interface data points. The resulting 3D model is exported as a GSLIB file recording which unit is present in each grid cell. To capture uncertainty with respect to the contacts, the elevation of the input data points could be perturbed, yielding multiple realizations of the geologic setting. However, this paper focuses on variations in the karst conduit network in a single realization of the geologic model (Fig. 3).

Conduit network model: SKS

Generating plausible conduit network maps within the aquifer requires a computationally efficient karst evolution model. The ensemble generator in this study uses the Stochastic Karst Simulator (SKS), which models conduit evolution at a watershed scale, constrained by hydrogeologic knowledge, without solving the physical and chemical equations driving speleogenesis (Borghi et al. 2012). Conduit evolution is instead simulated based on the assumption that water will follow a minimum-effort path within the boundaries of the karstifiable units, computed using a fast-marching algorithm (Sethian 1996). This approach is sufficiently fast to allow the generation of many equiprobable, hydrogeologically plausible

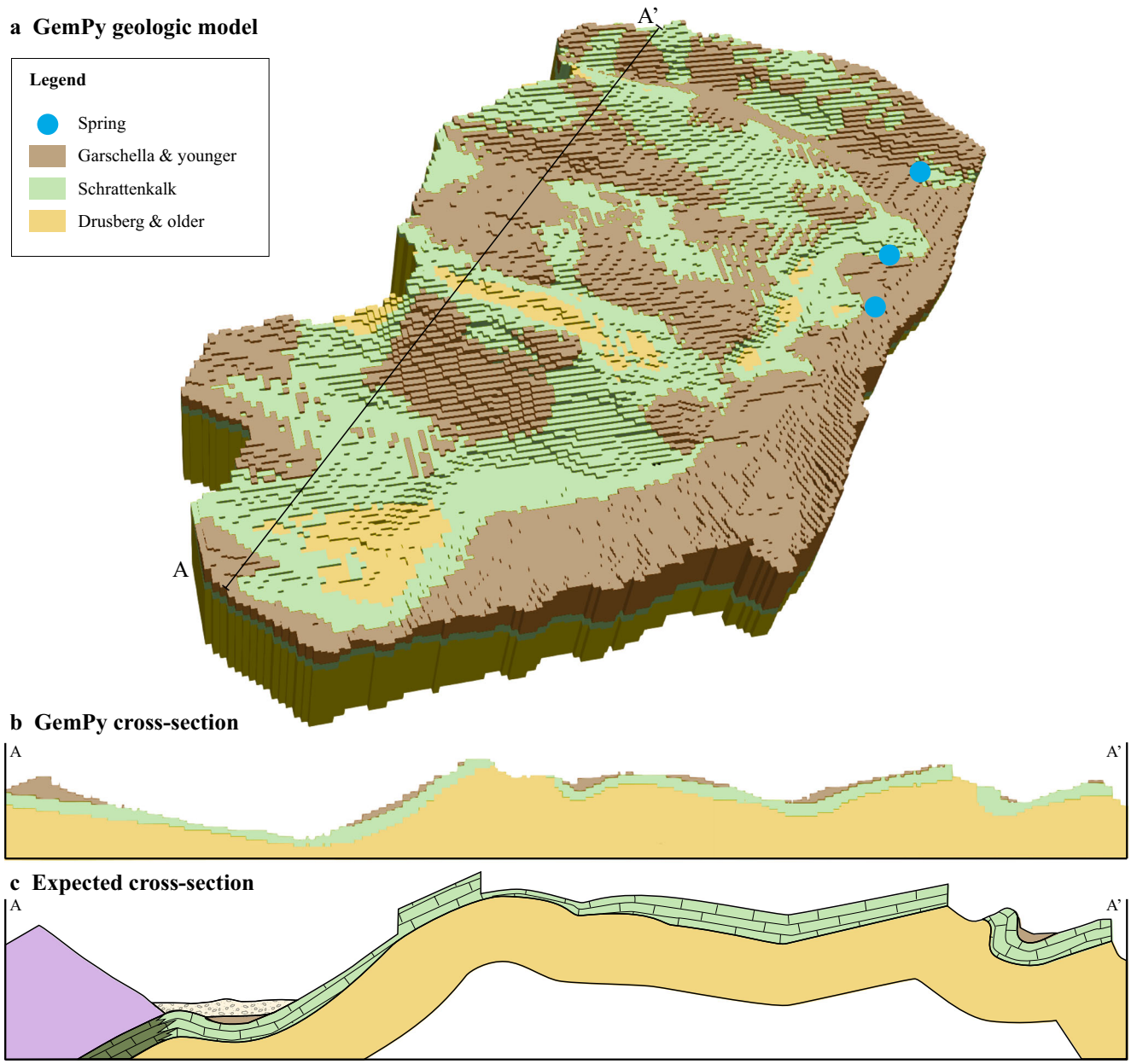

Fig. 3 **a** Simplified three-layer geologic model created with GemPy. **b** Cross-section of the GemPy model. **c** Expected schematic cross-section based on field mapping (Fig. 1d). Cross-sections are not to scale (the expected cross-section is schematic and the GemPy cross-section is in units of model cell indices)

realizations (Borghi et al. 2012). This tool has been combined with parameter estimation to identify properties of the conduit network such as hydraulic conductivity of the matrix, number of major conduits, and conduit radius, for a small number of synthetic and real systems (Borghi et al. 2016, Sivelle et al. 2020).

The SKS includes three stochastic components: a discrete fracture network simulator (Borghi et al. 2015), an inlet point generator, and a randomization of the karstification impacting the hierarchy of the resulting cave conduit systems. Conduit network realizations within the same geologic setting differ from one another as a function of hydrogeologic parameters in SKS such as the ratio between matrix and fracture conductivity (Table 1). It is therefore possible to generate many different conduit networks for the same geologic setting. For this study, a subset of the available parameters was chosen to vary, and 100 SKS realizations were run, each using a different combination of parameter values selected randomly from a range (Fig. 4a,b). The parameter ranges were defined based on a combination of the recommended range provided in the SKS documentation and published fracture maps and orientations for the study site (Goldscheider 2002; Cramer 1959), but the

ranges were also expanded to add variability to the models. The inlets to the system can be assigned as fixed locations (based on hydrogeologic observations such as known dolines or vertical shafts), random points within a defined spatial extent (if the actual inlet locations cannot be observed), or a combination of fixed and random locations (to account for both observed inlets and potential unknown inlets). For this site, lines of dolines and vertical shafts were observed along major synclinal axes. Based on these observations, in the reference model, every node along the branches of the network (which follow the synclinal axes) is treated as an inlet capturing runoff and infiltration from the area drained by that node, for a total of 20 inlet nodes. During the conduit network generation process for this study, to achieve a diverse ensemble of networks and avoid overly imprinting the SKS-generated networks with the reference network, the number of fixed inlet locations known from the reference model was varied in each run, but no random inlets were added. The runs with fewest inlets included only the fixed inlets corresponding to the endpoints of the reference network's branches ("top5" in Table 2), while the runs with the most inlets included the uppermost 3 nodes of the reference network's branches ("top13" in Table 2).

For each realization, SKS outputs a list of nodes and conduits making up the network, with one node in each model grid cell containing a conduit. This high-resolution network is then simplified by removing intermediate nodes along the conduits, using karstnet, a Python package for the statistical analysis of karst networks based on graph theory (Collon et al. 2017; Fig. 5; see ESM). The outlets of the system are then identified as the nodes located closest to the true (X, Y, Z) coordinates of the springs. To enable flow modeling, a new node is added at the true spring coordinates and a new conduit is created connecting the spring to the closest SKS-generated node.

Flow model: SWMM (EPA)

Groundwater flow through each proposed conduit network was simulated using the Storm Water Management Model (SWMM) from the United States Environmental Protection Agency (Rossman 2015). SWMM has been used in several previous studies to simulate karst systems, with good results (Campbell and Sullivan 2002, Peterson and Wicks 2006, and Vuilleumier et al. 2019, among others), and performed well in previous efforts to model the site used in this study (Chen and Goldscheider 2014; Chen et al. 2017, 2018). Choosing the same flow model used in the reference model also enables easier comparison between model predictions from this study and from the reference model.

he SWMM is a dynamic rainfall-runoff and conduit flow model, that routes surface runoff into and through a series of linear subsurface conduit segments. However, it should be noted that SWMM is not designed to simulate contaminant transport, and it does not consider conduit–matrix exchanges, so it would not be suitable for systems with high conduit–

Table 2 Parameters being varied across iterations of the SKS (see Borghi et al. 2016 for complete documentation)

Parameter	Description	Range of values
FMAfra	Multiplier indicating how much faster the fast-marching algorithm can travel through fractures than through matrix	1.2, 1.5, 2, 3, 5, 8, 10, 12, 15, 20
fixedInlets	Set of fixed inlets to start walkers from for the fast-marching algorithm, selected from the upper nodes of the base model	top 5, top 9, top 13
nfrac	Total number of fractures to generate	50, 100, 200, 400, 600, 700, 900, 1,000, 1,500, 2,000
minLen	Minimum fracture length (m)	1, 2, 5, 10, 15, 25, 50, 100, 150, 300
maxLen	Maximum fracture length (m)	350, 400, 500, 600, 700, 800, 900, 1,000, 2,000, 5,000
minStrike	Minimum fracture strike (azimuth) for each family. The number of items in each set indicates the number of fracture families	(35,105), (35,105,70), (120,110,145,35)
maxStrike	Maximum fracture strike for each family. The number of fracture families must match the number of families in minStrike	(45,115), (45,115,80), (130,120,155,45)
minDip	Minimum fracture dip (degrees below horizontal) for each family. The number of families must match minStrike.	(80,80), (80,80,80), (80,80,80,80)
maxDip	Maximum fracture dip (degrees below horizontal) for each family. The number of families must match minStrike	(90,90), (90,90,90), (100,100,100,100)
fracFactor	Relative proportion of total fractures from each family. The number of families must match minStrike	(8,2), (6,3,1), (5,3,1.5,0.5)

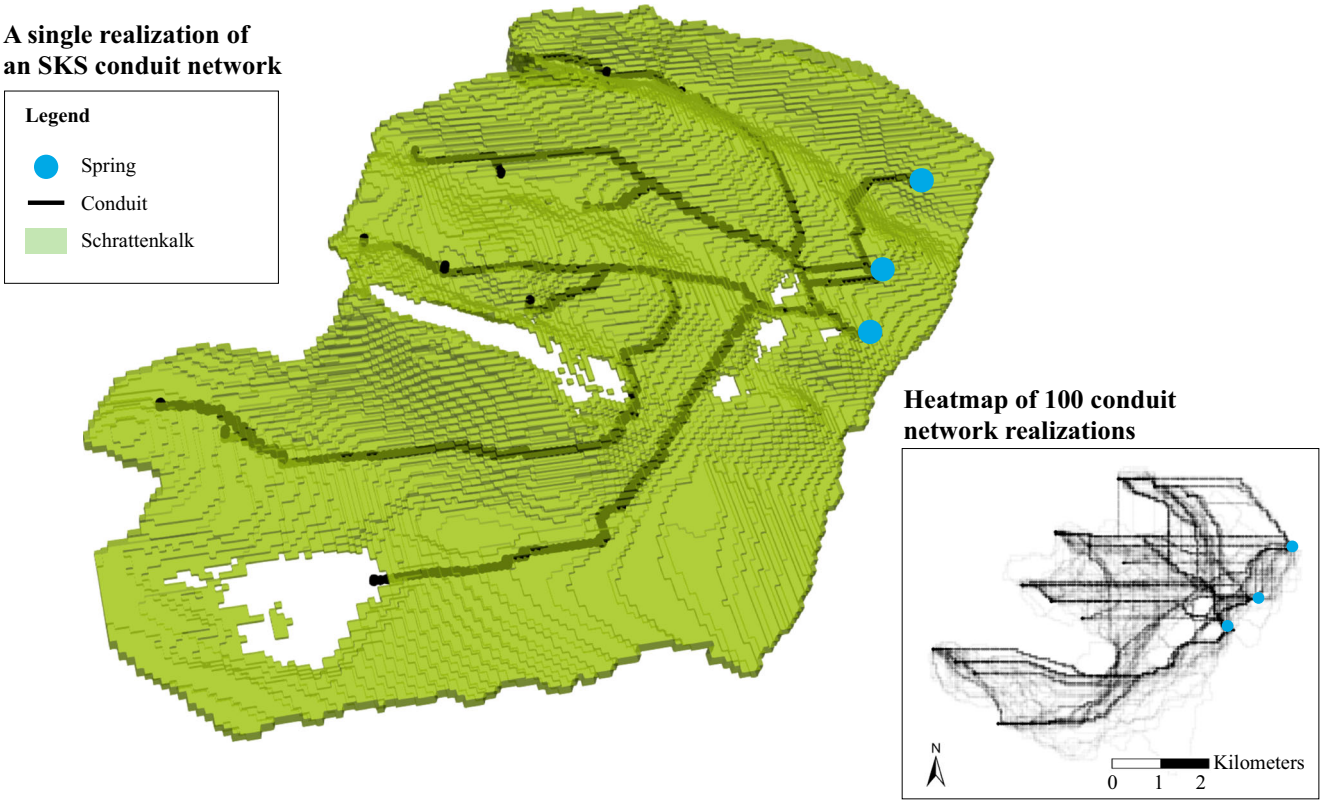

Fig. 4 A single realization of a conduit network generated by SKS given the geologic constraints as defined by the GemPy model. Overlying and underlying layers have been removed to show only the karstifiable limestone. Inset: A heatmap of 100 proposed conduit networks generated by SKS. Darker lines indicate areas where many different realizations placed a conduit in the same location. Lighter lines indicate areas where fewer realizations placed a conduit

matrix exchange rates (Peterson and Wicks 2006). For the Gottesacker karst system, although gradient inversion has been observed during high-flow events (Goldscheider 2005), the conduits generally drain the matrix. This is particularly true in the elevated parts of the system, where most of the karstified rock volume belongs to the unsaturated zone. In this setting, the conduits follow the troughs of the synclines, and lateral inflow can be observed from the "matrix" to the "conduits", mostly with open-channel flow conditions in the conduits. This is different than the classical conceptual model of "conduit–matrix exchange", where both zones are fully saturated, and the exchange can be described by means of hydraulic head difference between matrix and conduits and an empirical exchange coefficient.

For this study, although conventional conduit–matrix exchanges are not represented, inflow from the matrix to the conduits is inherently implemented in the model structure of SWMM, where subcatchments contribute flow towards a network of conduits. This approach is well-adapted to the hydraulic setting and karst aquifer configuration for this study site.

Flow through the conduits is represented by solving the Saint Venant equations (Eqs. 1 and 2), governing conservation of mass and momentum for the unsteady flow of water through a conduit network (Rossman 2015). When the conduit is unpressurized, Manning's equation is used to compute the friction slope, but when the conduit becomes pressurized, the Darcy-Weisbach equation is used instead (Rossman 2015). The Saint Venant equations are solved in the flow

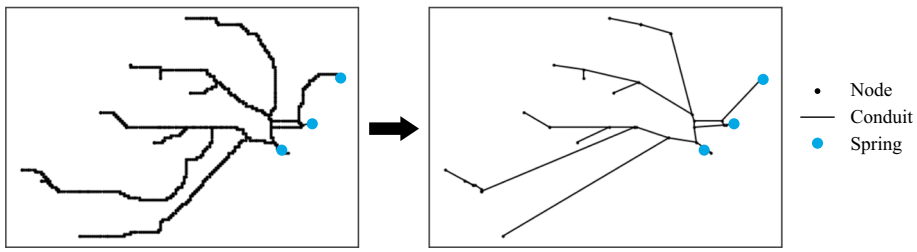

Fig. 5 Example of an SKS-generated conduit network before (left) and after (right) simplification

routing model by converting them into an explicit set of finite difference formulas that compute the flow in each conduit and the head at each junction for each time step. The parameters required for each conduit are the length, diameter, elevations for entering and exiting conduits, and roughness.

$$\frac{\partial A}{\partial t} + \frac{\partial Q}{\partial x} = 0 \tag{1}$$

$$\frac{\partial Q}{\partial t} + \frac{\partial \left(\frac{Q^2}{A}\right)}{\partial x} + gA\frac{\partial H}{\partial x} + gAS_f + gAh_L = 0 \tag{2}$$

Equations (1) and (2) represent Saint Venant equations for flow through a conduit, where x is the distance along the conduit (m), A is the cross-sectional area (m^2), Q is the flow rate (m^3/s), H is the hydraulic head of water in the conduit (m asl), S_f is the friction slope (calculated from Eqs. 3 or 4), h_L is the local energy loss per unit length, and g is the acceleration due to gravity (m/s^2).

For this study, a Python module was developed to automate the process of converting SKS-generated conduit networks to a SWMM format, distributing recharge inflows across subcatchments within the system, assigning hydraulic parameters to the conduits, running the SWMM model, and retrieving and visualizing model outputs (see ESM).

For each SKS-generated conduit network in the ensemble, the node coordinates and the conduit connections between nodes are imported from text files to Python objects. Then, because the node locations are different for each network, the boundaries of the surface subcatchment drained by each node must be redrawn. This is done using a pre-existing Python package, pysheds, which computes the surface flow direction of each grid cell based on the DEM, and thus determines which grid cells drain to specified pour points (i.e., conduit network nodes; see ESM). The slow and fast distributed recharge timeseries data for each cell are then summed by subcatchment and converted to node "inflow" and "baseflow" timeseries (in the terminology of SWMM). Conceptually, this is a simplified representation of the epikarst collecting and funneling precipitation inputs into vertical shafts leading to the subsurface system.

Hydraulic parameters can then be assigned to each node and conduit in Python. The module developed for this study uses all of the preceding information to write a SWMM input file, runs SWMM, and then extracts the spring discharge timeseries at all three system outlets from the SWMM output file. All parameters and outputs for each run are stored as both Python objects and csv files. An additional function calculates basic statistics for each spring's outflow timeseries: Nash-Sutcliffe efficiency, root mean square error, volume error, maximum, minimum, and mean discharge, and the percentage of time that each spring's discharge is below a fixed threshold.

Parameter exploration

Full parameter optimization of all hydraulic parameters for each conduit network is too computationally expensive in the context of an ensemble modeling approach. Therefore, for this study, a minimalist parameter exploration strategy was adopted. All parameters except the conduit diameter were held constant, with values selected based on the upper and lower bounds used in the reference model, which themselves were selected based on common ranges for this type of karst system (Chen et al. 2018). A simple sensitivity analysis was first done on the reference model, to confirm that it was not highly sensitive to parameters other than conduit diameter. Diameter was found to be far more determinant of the spring discharge pattern than roughness, the only other parameter not already provided by SKS. The exact value of the conduit diameters was found to have far less importance than the location of conduit restrictions: conduits closer to the system outlets were more sensitive to restriction than conduits farther from the system outlets. Therefore, only two diameter values were considered: large (4.0 m) and small (1.0 m).

To limit the computation time for flow modeling in SWMM, a subset of ten conduit networks was first selected among the 100 SKS realizations, with the goal of representing the maximum diversity of possible networks. Diversity was based on how different each network was from the others, according to its outflow behavior.

To select the subensemble, each of the 100 networks was allocated four initial flow simulations, each with conduit diameter restrictions in different locations. To determine where to restrict the diameters, a hierarchical conduit order (similar to Horton ordering for surface streams) was calculated as the subcatchment area drained by that conduit over the total catchment area (Borghi et al. 2016). Since higher-order conduits, closer to the system outlets, are more sensitive to restrictions, the four initial flow simulations explore only the effect of restrictions in high-order conduits. For each network, the following four restriction scenarios were applied: (1) the highest-order 25% of conduits were restricted, (2) only the conduit immediately connected to QE was restricted, (3) only the conduit connected to QA was restricted, and (4) only the conduit connected to QS was restricted.

The worst-performing 10% of conduit networks were rejected based on the combined results of this first round of flow simulations. For each spring, the root mean square error (RMSE) was summed across all four initial flow simulations. Then, the conduit networks with the highest summed RMSE were dropped from the ensemble: 4% were dropped for QE, 3% for QA and QS (to prevent the behavior at any one spring from dominating the outcome), totaling a 10% rejection rate and resulting in a remaining ensemble of 90 conduit networks.

k-means clustering (Kriegel et al. 2017) was then used to select a reduced ensemble of the ten most-different networks, using spring discharge timeseries statistics as parameters:

maximum, minimum, and mean discharge, and percentage of time that each spring's discharge is below a fixed threshold (see ESM for a full explanation of k-means clustering methods).

For this study, each conduit network in the reduced ensemble of ten most-different SKS-generated networks was allocated an initial set of 20 SWMM runs, with different conduit diameter sets for each run. These 20 SWMM runs were then reduced to a subset of ten runs for analysis, first by rejecting the worst-performing 10% as nonbehavioral, then by selecting the ten runs with the lowest flow continuity errors, to form the final ensemble of 100 models (10 SKS networks × 10 SWMM runs).

All conduits were initially assigned large diameters (4.0 m), and a probability-weighted random set of conduits to restrict to 1.0 m diameter was sampled, without replacement, for each flow simulation. The probability of a conduit being selected for restriction was based proportionally on its order. The conduit diameters were sampled only from a binary distribution, and not from a Gaussian distribution. The goal of this strategy is not to optimize the conduit hydraulic parameters or to perform an in-depth analysis of the parameter uncertainty, but simply to represent a rough estimate of the uncertainty in spring discharge originating from parameter uncertainty, as opposed to uncertainty in the structure of the conduit network. Future work will explore alternative strategies for efficiently representing and minimizing parameter uncertainty, such as Monte Carlo Tree Search (Silver et al. 2016), or genetic algorithms (Karpouzos et al. 2001).

After rejecting nonbehavioral models and reducing the size of the ensemble, 100 models remain for analysis, each of which yields timeseries of predicted discharge at the three major springs in the karst system (Fig. 6). For comparison, the reference network was also allocated 20 SWMM runs with different parameter sets, using the same process as for the SKS-generated networks.

To enable further comparisons across the ensemble, a global error metric was calculated for each model, summarizing the overall goodness of fit between the model-predicted spring discharge and the observed spring discharge. The global error was calculated as the sum of the RMSE for each of the three springs, divided by the mean observed discharge at that spring (Eq. 3). This metric is only meaningful in terms of relative value, but it is useful to compare overall model fit across the ensemble.

$$\sum \frac{\mathrm{RMSE_{QE}}}{\mathrm{Qobs_{QE}}} + \frac{\mathrm{RMSE_{QA}}}{\mathrm{Qobs_{QA}}} + \frac{\mathrm{RMSE_{QS}}}{\mathrm{Qobs_{QS}}} \quad (3)$$

Equation (3) shows the global error for each model in the ensemble. RMSE is the root mean square error at each of the three springs (QE: the estavelle, QA: Aubach Spring, QS: Sägebach Spring). Qobs is the mean observed discharge at each spring.

Results

Geologic model

The geologic model generated by GemPy generally resembles the expected geologic structure as determined by previous field mapping (Figs. 1 and 3). The most apparent difference is that the upper contact between the Schrattenkalk limestone and the overlying Garschella sandstone, which in reality coincides with the land surface over much of the model extent, often appears several meters below the land surface in the GemPy model. However, the general shape of the Schrattenkalk unit, and particularly its lower boundary, resemble the expected structure as visualized in cross-section (Fig. 3c). Because conduits tend to form along the lower boundary of a karst unit, this boundary is more important than the upper boundary. The small vertical underestimation of the location of the Garschella-Schrattenkalk contact is therefore not concerning for the purposes of this study.

Conduit network models

The initial ensemble of 100 SKS-generated conduit networks (Fig. 7) include networks that, subjectively, appear visually diverse and different from one another with respect to their overall configuration, their degree of branching, and their total number of conduits. However, almost none of these proposed networks visually resemble the reference network, such that they are not collectively exhaustive of conceptual space. The general structure of three or four branching conduits connecting almost perpendicularly to a main collector conduit along the axis of the Schwarzwasser Valley does exist in many of the SKS-generated networks, but the upper conduits rarely follow the synclinal fold axes as had been expected. Many of the proposed networks also bypass the Mahdtal Valley drainage axis (the northmost syncline) almost entirely, although there is clear and repeated tracer evidence for this connection (Goeppert and Goldscheider 2008). The only SKS network that closely resembles the reference network is No. 90 (Fig. 7, rightmost group of columns, row 90); however, this network was not selected by the clustering algorithm for inclusion in the subensemble.

Flow predictions

Many networks that are visually dissimilar from each other performed equally well in predicting discharge timeseries with a good overall fit to the observed discharge. Additionally,

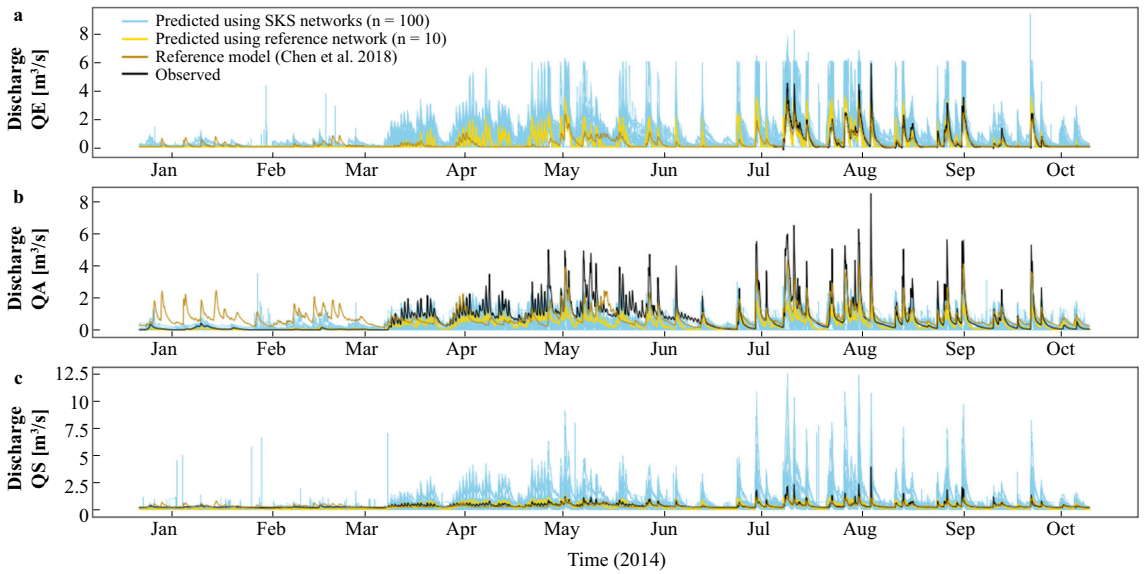

Fig. 6 Predicted spring discharge timeseries for 100 different models at each of the three springs in the system: **a** QE (the estavelle), **b** QA (Aubach spring), **c** QS (Sägebach spring). There are ten SWMM runs with different parameters for ten different SKS-generated conduit networks (light blue). For comparison, the reference network was subjected to the same SWMM parameter assignment process as the SKS networks, yielding ten runs (gold). The fully calibrated reference model (brown) and the observed discharge timeseries (black) are also shown

many networks that were visually dissimilar from the reference network perform equally well or better (Figs. 7 and 8a).

The networks in the initial ensemble appear sensitive to structure but less sensitive to differences in parameter values. This can be observed visually in Fig. 7: rows where one of the SWMM runs in that row are highlighted as having a high global error (darker colors) generally include multiple high-error runs, rather than the high-error runs being scattered across different networks. However, even on networks where most of the parameter sets perform poorly, the parameter set where all the lower conduits are restricted results in much lower error (note the absence of any highlighted runs in the "lower" columns). These results suggest that for this initial ensemble, while the network structure contributes significantly to the differences in performance across models, some combinations of hydraulic parameters can also consistently affect model performance.

With only four parameter sets per network in the initial ensemble, a more quantitative sensitivity assessment would not be reliable; however, for the subensemble of ten SKS networks, with ten behavioral SWMM runs per network, a more detailed analysis is possible. The range of error values in the first round of four SWMM runs generally bracketed the range of error values in the ten runs on the subensemble (Fig. 8b). This validates the assumption that the first-round SWMM runs were representative of a wide enough range of behaviors to use in selecting the subensemble. The subensemble, though it does not include any networks that visually resemble the reference network, does indeed represent a diversity of means and variances in global error values, as well as appearing subjectively visually diverse in terms of general network configuration, number of conduits, and degree of branching (Fig. 8c).

Across the 100 models in the subensemble, plus the 10 uncalibrated SWMM runs on the reference network, the original, calibrated reference model had the lowest global error, as expected. However, all of the uncalibrated runs on the reference network had global error values within 15% of the reference model, suggesting that extensive calibration is not necessary to achieve an acceptable performance. Unexpectedly, for six out of the ten SKS-generated networks, at least one parameter set performed almost as well (global error values within 15%) as the calibrated reference model, and many performed as well or better than the uncalibrated SWMM runs on the reference network (Figs. 8b and 9). Even conduit network structures that are far from the reference (such as No. 91 or No. 86, which, respectively, lack connections known to exist from tracer tests, or include a large number of conduits that cross over structural anticlines to connect in straight east–west lines to the springs) were able, with minimal parameter selection, to yield passably good predictions of spring discharge behavior that were quite similar to predictions yielded by the same process on the reference network (Fig. 9). These structures are therefore not mutually exclusive in terms of flow behavior. Based on the performance of the SKS-generated conduit network that yielded the lowest-error prediction of spring discharge after testing only ten essentially randomly selected parameter sets, if that same network were subjected to further parameter optimization and calibration, it seems likely that it would be able to predict spring discharge behavior nearly as well as the reference model, despite having a very

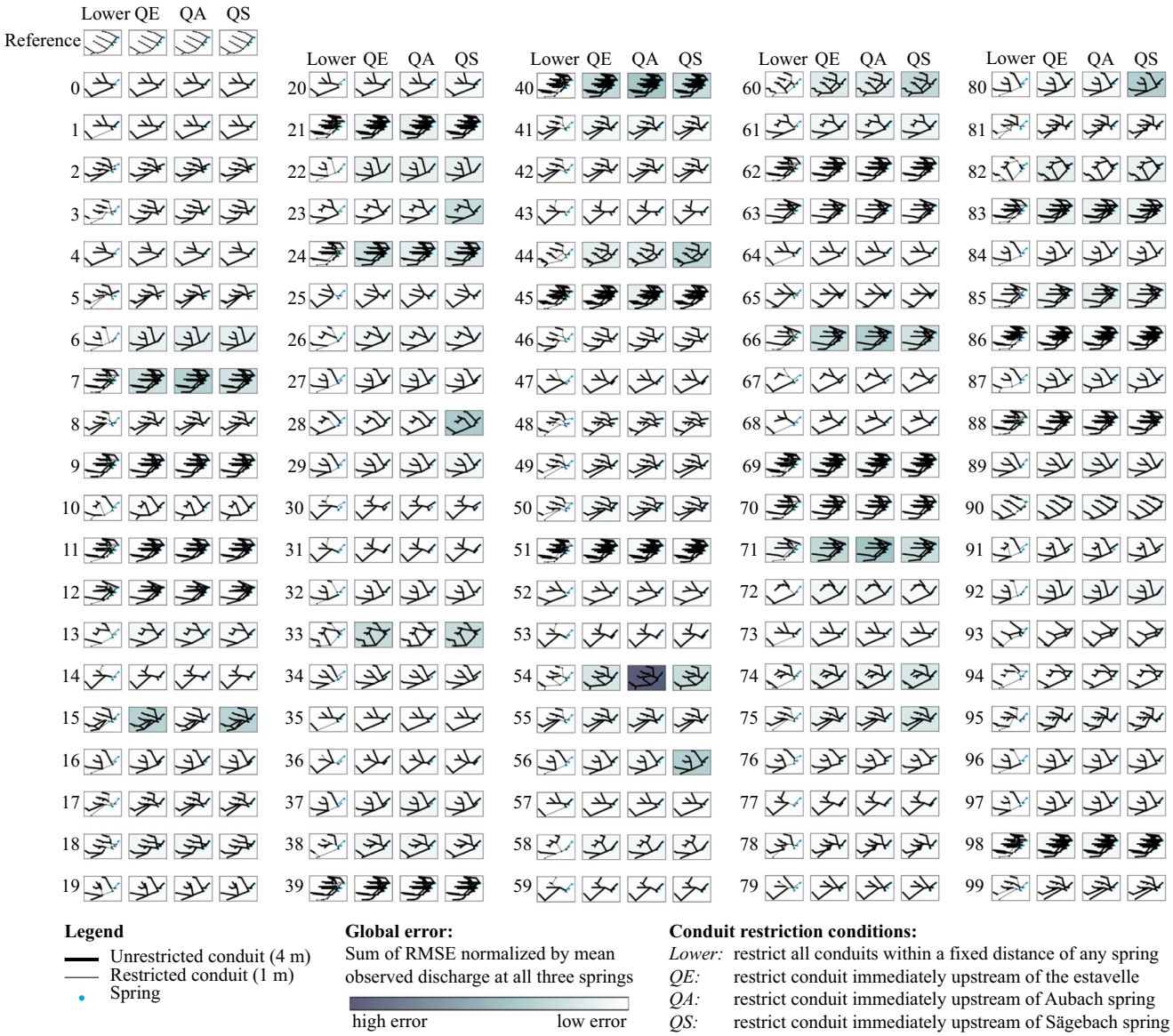

Fig. 7 Model performance across initial ensemble of 100 conduit network realizations (rows) for the same geologic setting generated by SKS, with four iterations of different flow scenarios (columns) for each. The ensemble is broken into five groups for better visualization, but the column names repeat across the top of each of these

different structure. This suggests that the global spring discharge behavior is insufficient to discriminate between conduit network structures.

Returning to the model ensemble as a whole, there do not appear to be any detectable patterns governing which parameter sets or which networks perform better than others (Fig. 10). There do not appear to be categories of conduits more or less sensitive to restrictions, nor do there appear to be particular network configurations that yield higher or lower global errors. Certain networks are insensitive to parameter values, performing similarly under all conduit restriction conditions, while others are highly sensitive to parameter values. Some SWMM runs with restrictions in the lower conduits perform well, while other perform poorly, and vice-versa.

The only visible trend is that models with a higher percentage of restricted conduit segments tend to perform better (Fig. 11). This may be related to the binary nature of the diameter values used in this study, where conduits were assigned to be either wide or narrow, rather than being assigned diameter values selected from a distribution. Chen and Goldscheider (2014) indicate that the Gottesacker system also has variable conduit diameters along the principal drainage axis, and is uniquely sensitive to restrictions in these sections of the network, which may also contribute to the impact of conduit diameter in spring discharge predictions in this study. However, many models with few

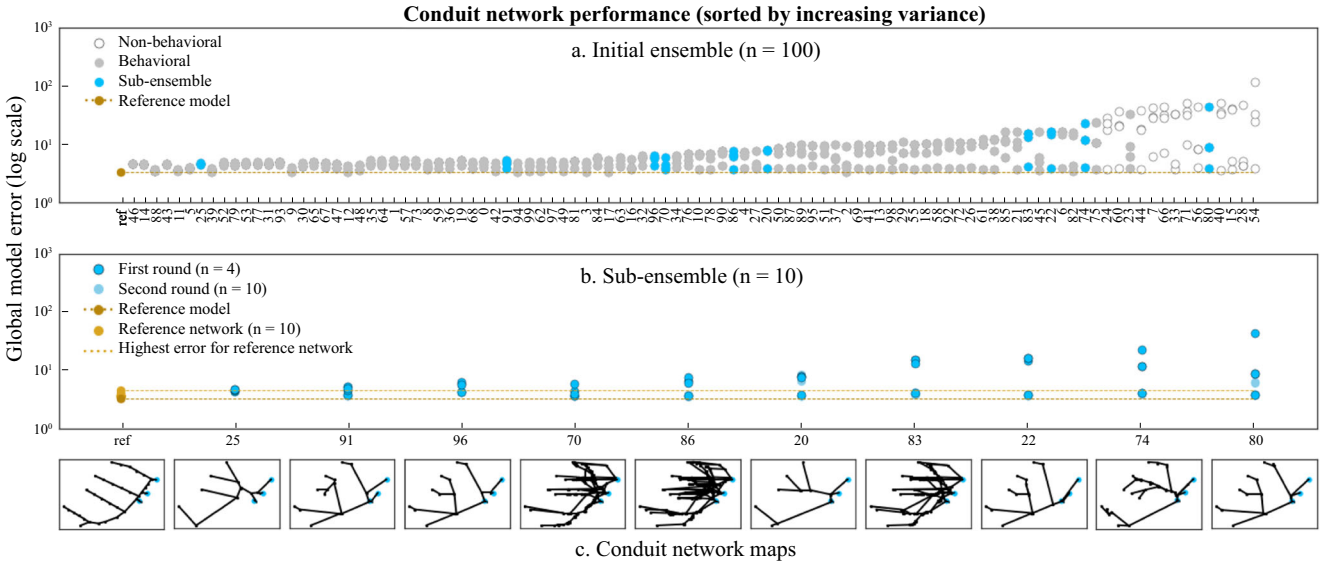

Fig. 8 **a** Initial conduit network ensemble performance, sorted by increasing variance. The worst-performing 10% of models were rejected as nonbehavioral (empty circles). A subensemble (blue) of ten behaviorally diverse conduit networks was selected by k-means clustering based on spring discharge timeseries metrics for four first-round SWMM runs. Many of the networks in the initial ensemble performed similarly to the reference model (gold). **b** Subensemble performance. Ten behavioral SWMM runs for each of the ten networks selected for the subensemble (light blue circles), as well as the initial four SWMM runs from the first round (dark blue circles). Several of the networks in the subensemble yielded one or more SWMM runs with global error values within the same range as the reference model (gold; i.e. between the minimum- and maximum-error SWMM runs on the reference model). **c** Conduit network maps for each of the networks in the subensemble

restricted conduits still perform as well as models with many restricted conduits. A high percentage of restricted conduits is therefore not a requirement for good performance. It should also be noted that Fig. 11 does not indicate *which* conduits are restricted. It is therefore possible that in the SKS-generated networks, spring discharge is only sensitive to restrictions in certain conduits, such that if only those conduits are restricted, the percentage of restricted conduits is low, but the model performance is good. Increasing the percentage of restricted conduits might then result in better performance only in that it increases the likelihood that the sensitive conduits are selected for restriction. However, even within this subset of more-sensitive conduits, there may not be a single unique combination of diameters that yields a best fit. Additionally, it is difficult to compare restriction locations across networks since conduit indexing is unique to each network. Because the conduit configuration is different in each network, parameter sets cannot easily be compared across networks, and the specific conduits that are most sensitive to restriction will be different in each network.

The lack of discernible patterns and the good performance of networks very different from the reference network support earlier findings by Borghi et al. (2016) that flow predictions are not sufficient to identify a unique network structure. However, although flow predictions are not discriminatory between network structures, they are more sensitive to the network structure than to the hydraulic parameters (in this study, the distribution of conduit diameters). This can be seen by comparing the mean distance between flow models within a network to the mean distance between flow models across networks. The distance between any pair of flow models (*i, j*) is calculated as the sum of squared differences (SSD) between the flow values predicted at each timestep (Eq. 4). The mean distance between models within a network is then the mean of the SSDs for every possible combination of pairs (n choose k) of flow runs within that network. The mean distance between models across different networks is the mean over all networks of the means of the SSDs for every possible pair of flow runs across every possible pair of networks (see ESM for a more complete explanation of SSD calculations). These calculations can be performed for each of the springs in the system, as well as for the sum of the SSDs at all three springs.

$$\text{Sum of Squared Differences (SSD)} = \sum_{t} \left(Q_i(t) - Q_j(t) \right)^2 \quad (4)$$

Equation 4 finds the distance between two models, calculated as the sum of squared differences (SSD), where $Q_i(t)$ is the spring discharge predicted by model i at time t, and $Qj(t)$ is the spring discharge predicted by model j at time t.

The mean distance between all flow runs on a single network can be thought of as the overall dissimilarity from one another of predictions generated by different parameter sets on that network, i.e., the importance of the parameters in determining spring discharge behavior. The mean distance between networks can be thought of as the importance of the network

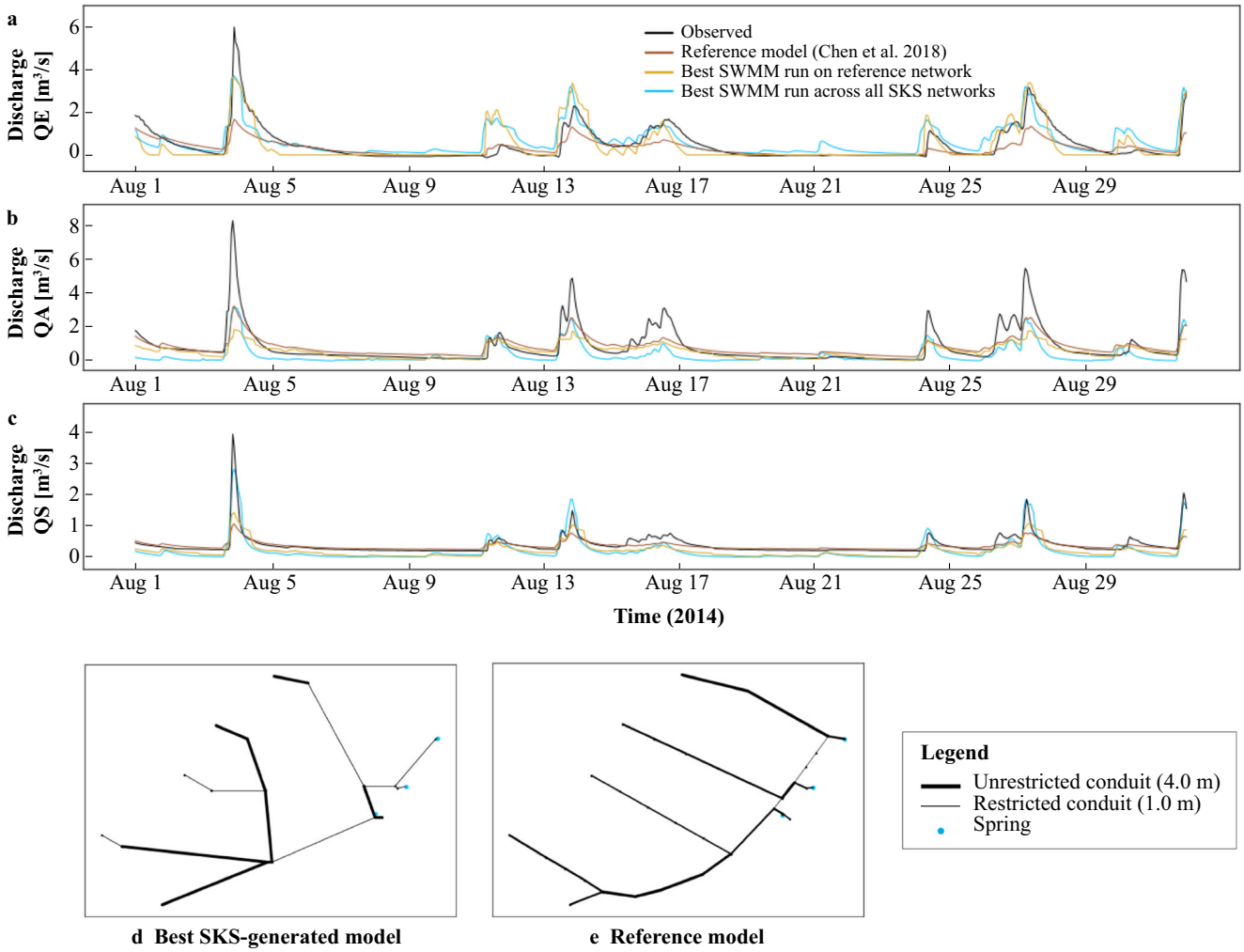

Fig. 9 Predicted and observed spring discharge timeseries at the three major springs in the system (**a**–**c**), focused on the month of August to show detailed comparison. The reference model matches the observed spring discharge patterns closely. The best model in the entire ensemble of 100 is also able to follow the general pattern of peaks and recessions, and is similar to the best flow model on the reference network, although it tends to underpredict baseflow, particularly at the lower two springs. However, the network configuration of the SKS-generated model (**d**) is very different from that of the reference model (**e**)

structure in determining spring discharge behavior. If the mean distance between networks is larger than the mean distances within each network, that would suggest that the network structure has a larger influence on spring discharge predictions than the parameter values. For the subensemble of networks considered in this study, the distance between networks is larger than the distance within networks in every case except for one network (No. 83), where the distance between models within the network is larger at the Sägebach spring only (Fig. 12). This suggests that, as expected based on previous work highlighting the impact of model structure on prediction uncertainty (Refsgaard et al. 2006; Enemark et al. 2019), the conduit network structure is more determinant of spring discharge behavior than the hydraulic parameters (in this case, the conduit diameter).

Discussion

Limitations

The ensemble approach to karst modeling demonstrated in this study provides a way to represent structural and conceptual uncertainty when the location of the conduits is unknown, by generating multiple competing hypotheses as to the network configuration. It is flexible in the choice of which conceptual aspects to vary, and how many computational resources to devote to parameter estimation. The bulk of the model code is open-source (with the exception of the 3D version of SKS; an open-source 2D version, pyKasso, is in progress (see ESM).

However, this approach can still be improved. For the flow models, epikarst routing is not considered, despite its

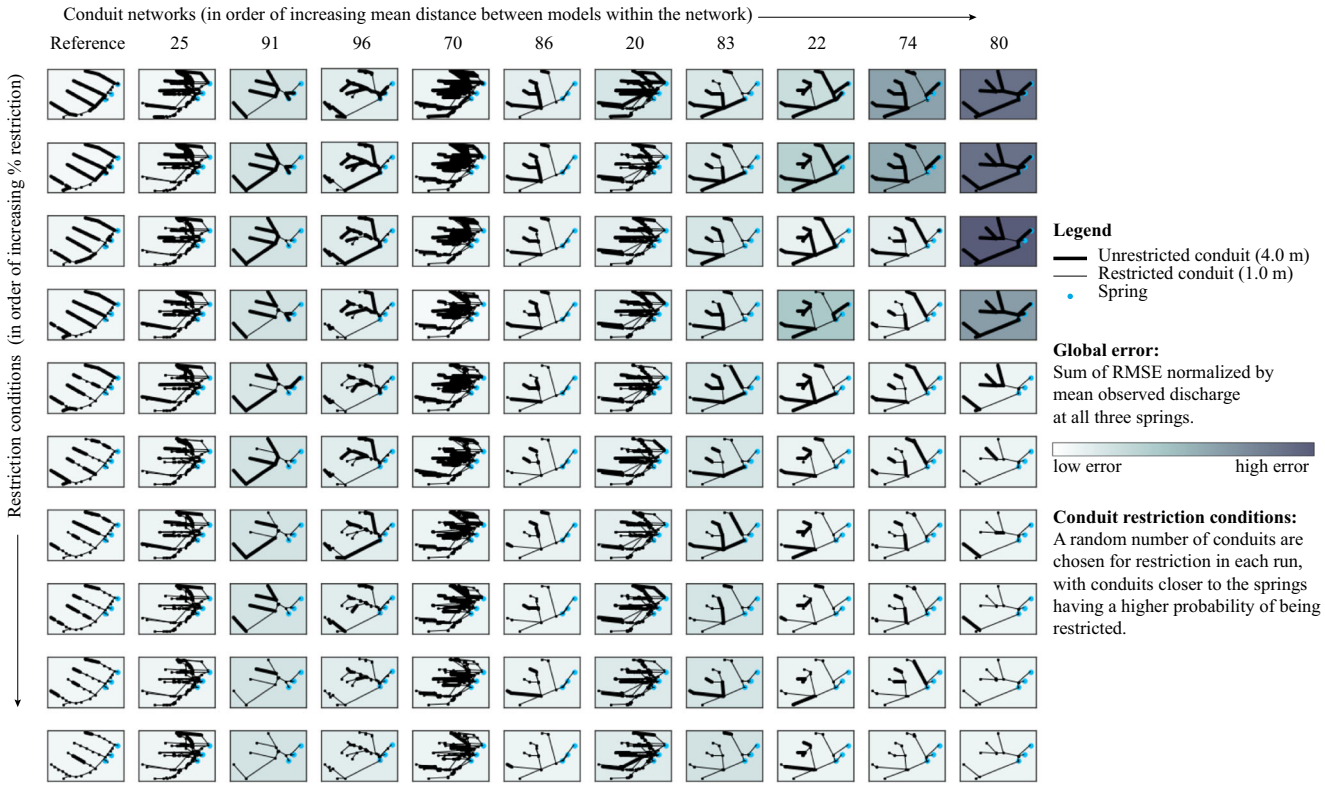

Fig. 10 Model performance across ten different conduit networks in the subensemble, with ten different parameter sets (conduit restriction conditions) for each network. Each column represents a different network, sorted in order of increasing mean distance between models in that network (mean sum of squared differences, Eq. 2). Each row is made up of SWMM runs with different conduit diameter sets for that network, sorted by increasing percentage of conduits that are restricted. Thick lines indicate wide conduits, and thin lines indicate restricted conduits. The first column is the reference network. The background color of each panel represents a global error metric for model performance: the sum of the RMSE at each spring, normalized by the mean observed discharge at each spring

importance for recharge processes and spring discharge curves. The groundwater flow processes modeled using SWMM are based on the assumption of perfectly circular conduit cross-sections, with only two possible diameter values: wide (4 m) and narrow (1 m). Conduit–matrix exchanges are not taken into account, limiting this approach to conduit-dominated systems. Additionally, the system inflow is subject to error because it is calculated based on automatically delineated subcatchments for each network, determined by a combination of surface topography and network node locations. This automatic process occasionally results in gaps where cells on the border between two subcatchments are not assigned to either catchment, and any precipitation falling on those cells is then not included in the system inflow.

For the conduit models, the conduit network subensemble used for the bulk of the analysis in this study is not exhaustive: it does not fully sample the conceptual space, as evidenced by the lack of any proposed networks resembling the reference network. This is due in part to the nature of the SKS conduit generation algorithm, which uses only the distance through the medium to the spring, computed using a fast-marching algorithm, and assuming a certain base level. Because it is assumed that water can flow anywhere when the reservoir is saturated, this distance neither accounts for gravity nor for real flow conditions, especially in the unsaturated zone. Finally, the network simplification (Fig. 5), which is necessary to reduce computation time, may oversimplify the conduit network configuration and obscure the original network proposed by SKS. Despite these conceptual gaps, many of the models in the ensemble were able to predict spring discharge behavior quite well.

Interpretation of results

The model ensemble generated for this study did not contain any conduit networks resembling the reference network. It was therefore not exhaustive. Many conduit networks in the ensemble that did not resemble each other or the reference network predicted the spring discharge behavior almost as well as the calibrated reference model, and as well or better than uncalibrated SWMM runs using the reference network. The networks in the ensemble were therefore also not mutually exclusive: spring discharge behavior was insufficient to discriminate between conduit network structures; however, the flow behavior is still useful as a low-cost first-pass filter to reject nonbehavioral models. Additionally, the conduit

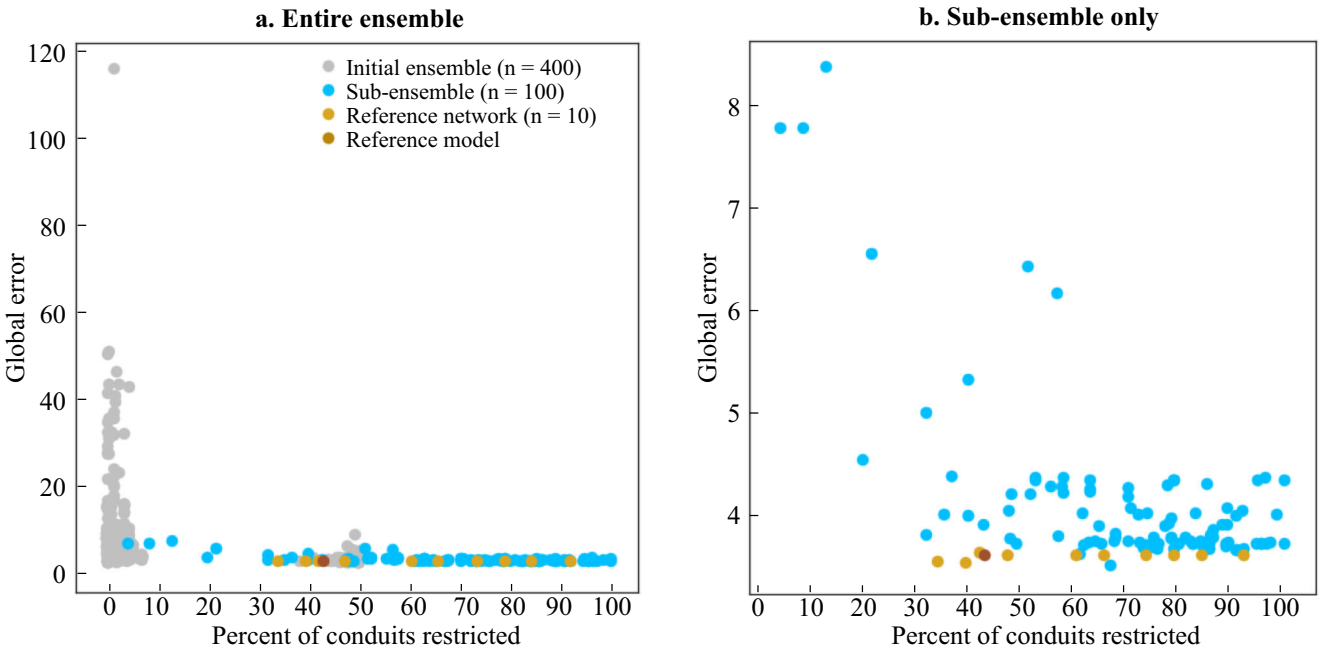

Fig. 11 Effect of conduit restrictions on model performance: **a** entire ensemble, **b** subensemble only. The models with the highest global error values all have a low percentage of conduits that are restricted. Additionally, when the subensemble alone is considered, the higher the percentage of restricted conduits, the lower the global error is. However, many models with few conduit restrictions also have low error values. Having 50% or more of conduits restricted increased the likelihood of fitting the data, but was neither sufficient nor necessary to do so (some models with few restrictions perform well and vice-versa)

network structure did control model predictions of spring discharge behavior more than the flow parameters (i.e. the conduit radius).

These results support previous work indicating that, because model structure is a major contributor to prediction uncertainty (Bredehoeft 2005), approaches using multiple competing model structures are particularly desirable when a good understanding of prediction uncertainty is desired (Neuman and Wierenga 2003). It is not necessarily required to identify a single "best" network structure in order to benefit from a multi-model approach: simply acknowledging and incorporating structural uncertainty into the generation and interpretation of model predictions is useful in and of itself when making decisions based on those predictions.

Three main questions for further reflection arise from these results:

1. What information and/or processes were used in creating the reference network structure that were not available to SKS as it generated possible network structures (resulting in networks that did not resemble the reference network)?
2. What, if any, different types of additional data might be able to discriminate between model structures where flow data could not?
3. What lessons can be drawn for water resource management from this ensemble modeling approach?

One of the initial goals of this study was to generate an ensemble of models that were collectively exhaustive,

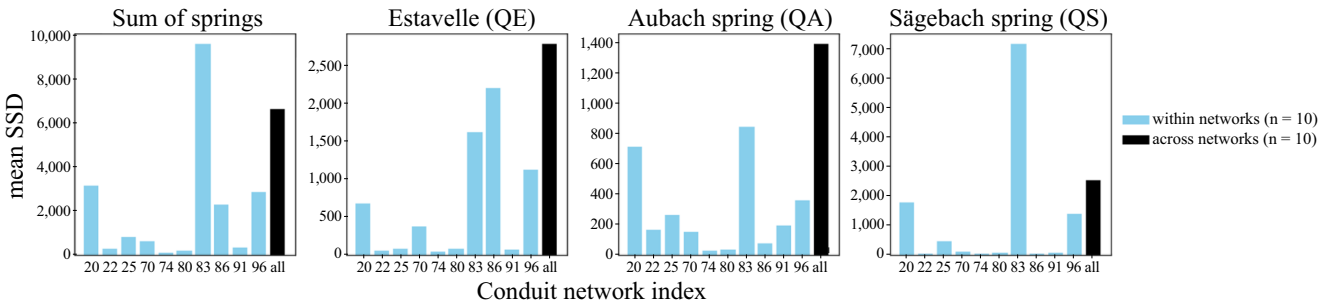

Fig. 12 Influence of network structure and conduit diameter on flow predictions. The mean distance between models across conduit networks (due to structure, in black) is compared to the mean distance between models within conduit networks (due to parameters, in blue)

resulting in an ensemble that includes at least one structure that is close to the "true" structure. If this is not the case, as occurred for the model ensemble in this study, the ensemble can be narrowed down to a subset of more probable networks, and then repopulated using those networks as "seeds," perhaps over the course of multiple iterations. In the initial ensemble of 100 in this study, only a single network (No. 90) resembled the reference network, and it was not selected as part of the subensemble of ten that was used for the bulk of the analysis. Two reasons for this appear possible: either the GemPy-SKS network generation process is lacking key information that went into creating the reference network, or the reference network is lacking information that is reflected in the networks proposed by SKS. The reference network is the result of the integration of large amounts of quantitative and qualitative information after many years of fieldwork and research on this particular karst system, so that although the precise location and properties of the conduits are uncertain, there is a very high degree of certainty that the general configuration of the network is correct. What, then, is missing from the GemPy-SKS network generation process, resulting in an ensemble with so little resemblance to the reference network?

The GemPy geologic model used as an input to SKS indicates only which formation is present in each model cell. It does not include any information as to the orientation of that formation in each cell, or the presence/absence of faults. Other data that was not included but is available includes tracer test results indicating known connections, speleological maps for known portions of the conduit network, and more qualitative information such as the presence of dolines aligned along synclinal fold axes in the upper part of the catchment. All of this information was used in the creation of the reference network. A first approach to integrating this information into the SKS-GemPy conduit network generation process could be to test adding each type of data individually, in order of increasing costliness of acquisition, regenerating an ensemble of networks at each step. This would allow for the identification of not only which types of data result in the most noticeable changes in the composition of the ensemble, but also which types of data provide the most information per unit cost. A first step could be to replace the SKS fast-marching algorithm with a flow simulation accounting for gravity, which is more computationally intensive and requires additional boundary condition information, but would provide more control over the conduit generation (Borghi et al. 2012).

Another goal of this study was to identify what additional data would be most useful in discriminating between models in the ensemble. Previous work by Borghi et al. (2016) suggests that solute transport predictions are significantly more discriminatory of conduit network structure than flow predictions. This is likely because solute transport is much more dependent on particle flow paths, whereas similar flow behaviors can result from many different flow paths. However, it is likely that multiple different structures will fit tracer data as well as flow data (Borghi et al. 2016; Sivelle et al. 2020). Although tracer test data will not identify a single "best" structure, integrating tracer test data will still *reduce* prediction uncertainty, particularly uncertainty in predictions of solute transport, by enabling the rejection of some structures in the ensemble. Tracer test results are therefore the logical next data to include if contaminant transport is a prediction of interest to the model end users. Because SWMM is not designed to model solute transport, simulating tracer tests on the networks in the existing ensemble would require translating them into a different modeling tool such as MODFLOW-CFPv2 (Reimann and Hill 2009; Reimann et al. 2013), which can be coupled to UMT3D to solve solute transport (Hu and Xu 2016), or FEFLOW (Trefry and Muffels 2007). However, a more qualitative use of existing tracer test data would be possible without running a complete transport simulation. Because much of the available tracer test data does not include continuous discharge measurements, the calculation of tracer recoveries involves some degree of uncertainty, but underground connections can clearly be identified and transit times, flow velocities and dispersion can be calculated. Many networks could be rejected by considering the networks as directed graphs and testing whether the points with connections demonstrated by tracer test data are also connected in those networks. Networks where demonstrated connections are missing, or where connections that have been demonstrated to be absent are present, would be rejected. Additionally, the available flow data could be considered more carefully. Rather than simply comparing the global fit across the entire available timeseries at all three springs, a more detailed analysis could focus on each individual spring, and on flow behavior under specific conditions (high, low, or intermediate flows). If the predictions of interest are the behavior of a specific spring under specific conditions, then such an analysis could be more discriminatory, without the need for any additional data collection.

The model ensemble simulations in this study, although they are based on minimal input data, can still provide useful information for practical water resource management applications in karst systems. The primary lessons to be learned are that:

1. The modeling approach and type of data collection chosen for any system should be dependent on the predictions of interest to the users and managers of that system. If the prediction of interest is the spring discharge only, attempting to define the conduit network structure in detail may not be necessary, and can in fact lead to a misleading sense of certainty about the structure, because many different conduit networks can yield similar discharge predictions. In this case, a simple lumped model

or an artificial neural network model, neither of which consider physical processes, may be more appropriate because they are much easier to calibrate and can reproduce rainfall–discharge relationships quite well (Hartmann et al. 2014) Depending on the purpose of the model, the value of the insights as to the system's internal processes and functioning that are achieved by a distributed model that represents the conduit network configuration may not be proportional to the increase in effort/cost to create such a model, and may simultaneously falsely increase users' confidence in model predictions. However, if the prediction of interest is where the conduits are located, or how the system functions internally, expending the effort to build a distributed model exploring the range of possible conduit network structures is necessary. The ensemble modeling approach in this study is also modular. If the prediction of interest is how the system will respond to contamination events, a flow model capable of simulating transport could be substituted for SWMM without changing the overall approach. Additional data of types other than spring discharge and geologic maps would then also be needed.

2. To adequately represent uncertainty in model predictions when the network structure is unknown, devoting computational resources to testing multiple model structures may be as important as or more important than testing a large number of parameter sets.

Conclusions

This paper proposed a multi-model approach to explore the contributions of structural uncertainty to prediction uncertainty in unmapped conduit networks, and to guide further data collection for maximum informativeness. The new approach was tested by applying it to a well-understood study site.

In this approach, projections of flow behavior are based on an ensemble of many competing conduit network maps, each of which is minimally calibrated for hydraulic parameters, using a simple binary distribution (wide conduit diameter or narrow conduit diameter). This stands in contrast to the more common tactic of creating a single, thoroughly calibrated "best estimate" of the conduit network map.

Flow predictions were then used to reject nonbehavioral structures and to discriminate between different proposed structures. The range of simulated flow behaviors represents the range of uncertainty associated with the model ensemble, providing multiple possible narratives for the system functioning that can be used in planning by decision-makers. The ensemble can also guide further data collection to ensure that a wide range of conceptual space is represented, to discriminate between models that disagree with one another, and to reduce uncertainty with respect to predictions of interest to users and managers of the specific system being modeled.

In this study, none of the conduit networks in the ensemble resembled the "true" network (as represented by a reference network developed based on extensive field observations). However, many networks, which were also all different from one another, yielded similar predictions of flow behavior. This supports the conclusion that proposing a single network structure may misleadingly minimize the structural uncertainty associated with model predictions of the system behavior. Additionally, the differences between multiple flow simulations on the same network were almost always smaller than the mean difference between flow simulations across networks, suggesting that the conduit network structure has a larger influence than the hydraulic parameters in controlling flow behavior. The results of this initial modeling effort can now guide future work to sample structural space more broadly and discriminate between different network structures. This will require integrating additional information that is available but was intentionally withheld such as tracer test data, orientation data for geologic units, a more nuanced conceptualization of the conduit evolution process, and discharge behavior at individual springs in response to specific high, intermediate, or low flow conditions.

Finally, the ability of multiple different structures to reproduce observed discharge patterns suggests that modelers and water resource managers should be in conversation about the *purpose* of the modeling during the model construction process, *before* allocating data collection and modeling resources. Constructing an ensemble from minimal available data, while computationally intensive, remains much faster than traditional model building techniques, particularly when allocating only minimal resources to parameter estimation. This may help modelers keep multiple competing conceptualizations of the system in mind at once, rather than becoming attached to a single "best" conceptualization (since multiple other conceptualizations may be equally able to reproduce the system's behavior). This approach then enables a more complete examination of sources of uncertainty, factoring in both structure and parameters. Analyses similar to those performed in this study can be repeated for the same ensemble, but with different predictions of interest (in this case, perhaps the system's response to a contamination event), to determine whether some of the structures in the ensemble project system behaviors that would be concerning to end users. Additional data collection efforts can then be targeted towards data that is more discriminatory between the different proposed model conceptualizations, and that specifically discriminates between models that predict concerning behaviors and models that do not.

Acknowledgements The authors would like to thank Dr. C.L. Winter for advice on model selection.

Funding information Open Access funding provided by Projekt DEAL. The authors acknowledge the National Science Foundation, Deutscher Akademischer Austauschdienst. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant DGE-1143953. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Borghi A, Renard P, Jenni S (2012) A pseudo-genetic stochastic model to generate karstic networks. J Hydrol 414–415:516–529. https://doi.org/10.1016/j.jhydrol.2011.11.032

Borghi A, Renard P, Fournier L, Negro F (2015) Stochastic fracture generation accounting for the stratification orientation in a folded environment based on an implicit geological model. Eng Geol 187: 135–142. https://doi.org/10.1016/j.enggeo.2014.12.019

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. https://doi.org/10.1016/j.advwatres.2016.02.009

Bredehoeft J (2005) The conceptualization model problem—surprise. Hydrogeol J 13:37–46. https://doi.org/10.1007/s10040-004-0430-5

Campbell CW, Sullivan SM (2002) Simulating time-varying cave flow and water levels using the Storm Water Management Model. Eng Geol 65:133–139

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. https://doi.org/10.1016/j.jhydrol.2014.04.005

Chen Z, Hartmann A, Goldscheider N (2017) A new approach to evaluate spatiotemporal dynamics of controlling parameters in distributed environmental models. Environ Model Softw 87:1–16. https://doi.org/10.1016/j.envsoft.2016.10.005

Chen Z, Hartmann A, Wagener T, Goldscheider N (2018) Dynamics of water fluxes and storages in an Alpine karst catchment under current and potential future climate conditions. Hydrol Earth Syst Sci 22: 3807–3823. https://doi.org/10.5194/hess-22-3807-2018

Clark MP, Kavetski D, Fenicia F (2011) Pursuing the method of multiple working hypotheses for hydrological modeling. Water Resour Res 47. https://doi.org/10.1029/2010WR009827

Collon P, Bernasconi D, Vuilleumier C, Renard P (2017) Statistical metrics for the characterization of karst network geometry and topology. Geomorphology 283:122–142. https://doi.org/10.1016/j.geomorph.2017.01.034

Cramer K (1959) Die Geologie des Mahdtales und der Karst des Gottesackergebietes [The geology of the Mahdtal and the karst of the Gottesacker area]. MSc Thesis, Technische Hochschule Munich, Munich, Germany

de la Varga M, Schaaf A, Wellmann F (2019) GemPy 1.0: open-source stochastic geological modeling and inversion. Geosci Model Dev 12:1–32. https://doi.org/10.5194/gmd-12-1-2019

Doherty J, Simmons CT (2013) Groundwater modelling in decision support: reflections on a unified conceptual framework. Hydrogeol J 21: 1531–1537. https://doi.org/10.1007/s10040-013-1027-7

Drew D, Hötzl H (1999) Karst hydrogeology and human activities: impacts, consequences and implications. IAH International Contributions to Hydrogeology, Routledge, Abingdon, UK

Enemark T, Peeters LJM, 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

Ferré TPA (2017) Revisiting the relationship between data, models, and decision-making. Groundwater 55:604–614. https://doi.org/10.1111/gwat.12574

Fleury P, Plagnes V, Bakalowicz M (2007) Modelling of the functioning of karst aquifers with a reservoir model: application to Fontaine de Vaucluse (South of France). J Hydrol 345:38–49. https://doi.org/10.1016/j.jhydrol.2007.07.014

Ford D, Williams P (2007) Karst hydrogeology and geomorphology. Wiley, Hoboken, NJ

Goeppert N, Goldscheider N (2008) Solute and colloid transport in karst conduits under low- and high-flow conditions. Ground Water 46(1): 61–68. https://doi.org/10.1111/j.1745-6584.2007.00373.x

Goldscheider N (2002) Hydrogeology and vulnerability of karst systems: examples from the Northern Alps and the Swabian Alb. PhD Thesis, Universität Karlsruhe, Karlsruhe, Germany

Goldscheider N (2005) Fold structure and underground drainage pattern in the alpine karst system Hochifen-Gottesacker. Eclogae Geol Helv 98:1–17. https://doi.org/10.1007/s00015-005-1143-z

Goldscheider N (2011) Alpine Hydrogeologie. Grundwasser 16(1):1. https://doi.org/10.1007/s00767-010-0157-2

Goldscheider N, Neukum C (2010) Fold and fault control on the drainage pattern of a double-karst-aquifer system, Winterstaude, Austrian Alps. Acta Carsolog 39. https://doi.org/10.3986/ac.v39i2.91

Goldscheider N, Chen Z, Auler AS, Bakalowicz M, Broda S, Drew D, Hartmann J, Jiang G, Moosdorf N, Stevanovic Z, Veni G (2020) Global distribution of carbonate rocks and karst water resources. Hydrogeol J 28(5):1661–1677. https://doi.org/10.1007/s10040-020-02139-5

Gupta HV, Clark MP, Vrugt JA, Abramowitz G, Ye M (2012) Towards a comprehensive assessment of model structural adequacy. Water Resour Res 48. https://doi.org/10.1029/2011WR011044

Hartmann A, Kralik M, Humer F, Lange J, Weiler M (2012) Identification of a karst system's intrinsic hydrodynamic parameters: upscaling from single springs to the whole aquifer. Environ Earth Sci 65:2377–2389. https://doi.org/10.1007/s12665-011-1033-9

Hartmann A, Goldscheider N, Wagener T, Lange J, Weiler M (2014) Karst water resources in a changing world: review of hydrological modeling approaches. Rev Geophys 52:218–242. https://doi.org/10.1002/2013RG000443

Hock R (1999) A distributed temperature-index ice- and snowmelt model including potential direct solar radiation. J Glaciol 45:101–111

Höhlenverein Sonthofen EV (2006) Das Hölloch im Mahdtal: 100 Jahre Höhlenforschung im Kleinwalsertal [The Hölloch in the Mahdtal: 100 years of caving in the Kleinwalsertal]. Höhlenverein Sonthofen, Sonthofen, Germany

Hu BX, Xu Z 2016 Numerical simulation of groundwater flow and solute transport in a karst aquifer with conduits. In: Javaid MS (ed) Groundwater: contaminant and resource management. InTech. https://doi.org/10.5772/63766

Jaquet O, Jeannin PY (1994) Modelling the karstic medium: a geostatistical approach. In: Armstrong M, Dowd PA (eds) Geostatistical simulations: quantitative geology and geostatistics.

Springer, Dordrecht, The Netherlands, pp 185–195. https://doi.org/10.1007/978-94-015-8267-4_15

Karpouzos DK, Delay F, Katsifarakis KL, de Marsily G (2001) A multipopulation genetic algorithm to solve the inverse problem in hydrogeology. Water Resour Res 37:2291–2302. https://doi.org/10.1029/2000WR900411

Kovács A, Sauter M (2007) Modelling karst hydrodynamics. In: Methods in karst hydrogeology. Taylor and Francis, London, pp 201–222

Kriegel H-P, Schubert E, Zimek A (2017) The (black) art of runtime evaluation: are we comparing algorithms or implementations? Knowl Inf Syst 52:341–378. https://doi.org/10.1007/s10115-016-1004-2

Lajaunie C, Courrioux G, Manuel L (1997) Foliation fields and 3D cartography in geology: principles of a method based on potential interpolation. Math Geol 29:571–584. https://doi.org/10.1007/BF02775087

Large Rivers and Large Lakes. *European Environment Agency*, D7925F3C-AFF7-4256-8162-513A2C1C69E3, DAT-1-en (2017).

Trefry MG, Muffels C (2007) FEFLOW: a finite-element ground water flow and transport modeling tool. Groundwater 45:525–528. https://doi.org/10.1111/j.1745-6584.2007.00358.x

Neuman SP, Wierenga PJ (2003) A comprehensive strategy of hydrogeologic modeling and uncertainty analysis for nuclear facilities and sites. NUREG/CR-6805, US Nuclear Regulatory Commission, Washington, DC

Peterson EW, Wicks CM (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. https://doi.org/10.1016/j.jhydrol.2006.02.017

Refsgaard JC, van der Sluijs JP, Brown J, van der Keur P (2006) A framework for dealing with uncertainty due to model structure error. Adv Water Resour 29:1586–1597. https://doi.org/10.1016/j.advwatres.2005.11.013

Reimann T, Hill ME (2009) MODFLOW-CFP: a new conduit flow process for MODFLOW-2005. Ground Water 47:321–325. https://doi.org/10.1111/j.1745-6584.2009.00561.x

Reimann T, Liedl R, Giese M, Geyer T, Maréchal J-C, Doerfliger N, Bauer S, Birk S (2013) Addition and enhancement of flow and transport processes to the MODFLOW-2005 conduit flow process. TU Dresden, Dresden, Germany

Rogger M, Viglione A, Derx J, Blöschl G (2013) Quantifying effects of catchment storage thresholds on step changes in the flood frequency curve: step changes in the flood frequency curve. Water Resour Res 49:6946–6958. https://doi.org/10.1002/wrcr.20553

Rossman LA (2015) Storm Water Management Model user's manual Version 5.1. No. EPA-600/R-14/413b, US EPA, Cincinnati, OH

Scanlon BR, Mace RE, Barrett ME, Smith B (2003) Can we simulate regional groundwater flow in a karst system using equivalent porous media models? Case study, Barton Springs Edwards aquifer, USA. J Hydrol 276:137–158. https://doi.org/10.1016/S0022-1694(03)00064-7

Sethian JA (1996) A fast marching level set method for monotonically advancing fronts. Proc Natl Acad Sci 93:1591–1595. https://doi.org/10.1073/pnas.93.4.1591

Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, Dieleman S, Grewe D, Nham J, Kalchbrenner N, Sutskever I, Lillicrap T, Leach M, Kavukcuoglu K, Graepel T, Hassabis D (2016) Mastering the game of Go with deep neural networks and tree search. Nature 529:484–489. https://doi.org/10.1038/nature16961

Sivelle V, Renard P, Labat D (2020) Coupling SKS and SWMM to solve the inverse problem based on artificial tracer tests in karstic aquifers. Water 12:1139. https://doi.org/10.3390/w12041139

Vuilleumier C, Jeannin P-Y, Perrochet P (2019) Physics-based fine-scale numerical model of a karst system (Milandre Cave, Switzerland). Hydrogeol J 27: 2347–2363. https://doi.org/10.1007/s10040-019-02006-y

Wagner G (1950) Rund um Hochifen und Gottesackergebiet [Around Hochifen and Gottesacker area]. Rau, Öhringen, Germany

Worthington SRH, Smart CC, Ruland W (2012) Effective porosity of a carbonate aquifer with bacterial contamination: Walkerton, Ontario, Canada. J Hydrol 464–465:517–527. https://doi.org/10.1016/j.jhydrol.2012.07.046