HYDROLOGICAL PROCESSES
Hydrol. Process. **30**, 5027–5038 (2016)
Published online 30 August 2016 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/hyp.10955

Diagnostic calibration and cross-catchment transferability of a simple process-consistent hydrologic model

Tyler Smith,[1]* Kaitlin Hayes,[1] Lucy Marshall,[2] Brian McGlynn[3] and Kelsey Jencso[4]

[1] *Department of Civil and Environmental Engineering, Clarkson University, Potsdam, NY, 13699, USA*
[2] *School of Civil and Environmental Engineering, University of New South Wales, Sydney, New South Wales, 2052, Australia*
[3] *Division of Earth and Ocean Sciences, Duke University, Durham, NC, 27708, USA*
[4] *Department of Forest Management, University of Montana, Missoula, MT, USA*

Abstract:

The transferability of hydrologic models is of ever increasing importance for making improved hydrologic predictions and testing hypothesized hydrologic drivers. Here, we present an investigation into the variability and transferability of the recently introduced catchment connectivity model (Smith *et al.*, 2013). The catchment connectivity model was developed following extensive experimental observations identifying the key drivers of streamflow in the Tenderfoot Creek Experimental Forest (Jencso *et al.*, 2009; Jencso *et al.*, 2010), with the goal of creating a simple model consistent with internal observations of catchment hydrologic connectivity patterns. The model was applied across seven catchments located within Tenderfoot Creek Experimental Forest to investigate spatial variability and transferability of model performance and parameterization. The results demonstrated that the model resulted in historically good fits (based on previous studies at the sites) to both the hydrograph and internal water table dynamics (corroborated with experimental observations). The impact of a priori parameter limits was also examined. It was observed that enforcing field-based limits on model parameters resulted in slight reductions to streamflow hydrograph fits, but significant improvements to model process fidelity (as hydrologic connectivity), as well as moderate improvement in the transferability of model parameterizations from one catchment to the next. Copyright © 2016 John Wiley & Sons, Ltd.

KEY WORDS calibration; model; regional; transferability; catchment; connectivity

Received 30 December 2015; Accepted 24 June 2016

INTRODUCTION

Hydrologic models are often assessed in terms of their ability to reproduce observed streamflow hydrographs for a given catchment, following the calibration of its unknown parameters (e.g. Nash and Sutcliffe, 1970; Beven and Binley, 1992; Duan *et al.*, 1992; Gupta *et al.*, 1998; Wagener, 2003). However, perhaps a more rigorous test of the true utility of a model is its transferability across hydrologically similar catchments (Wagener and Gupta, 2005). Transferability can reflect the appropriateness of the model structure (i.e. the model is actually representing the acting streamflow generation processes), and by extension, can be considered a surrogate for the verification of internal model consistency (i.e. dominant processes or states are reliably simulated by the model).

Model assessment is traditionally performed following an a priori selection of a model structure. For example, model X, Y, or Z is chosen by the modeller, with an underlying assumption regarding its appropriateness relative to the study site and on the basis of its ability to recreate a past streamflow hydrograph, following calibration of the model parameters (Klemeš, 1983). Recently, there has been a push to utilize flexible model structures that consist of a library of common hydrologic process components that can be mixed and matched to create an 'appropriate' model structure for the site of interest (e.g. Clark *et al.*, 2008; Clark *et al.*, 2011; Fenicia *et al.*, 2011; Kavetski and Fenicia, 2011; Staudinger *et al.*, 2011). This, of course, requires the testing of multiple model structures for the site of interest. However, as many model comparison studies have shown, alternate model structures can perform similarly to one another (in terms of hydrograph fit) despite differences in complexity and process representation (e.g. Reed *et al.*, 2004; Duan *et al.*, 2006). Thus, there is a need to assess hydrologic models beyond hydrograph fit, given the relative lack of information contained in streamflow

*Correspondence to: Tyler Smith, Department of Civil and Environmental Engineering, Clarkson University, Potsdam, NY 13699, USA.
E-mail: tsmith@clarkson.edu

alone. Additional or alternate indicators such as internal process consistency can be useful in this regard (e.g. Son and Sivapalan, 2007; McMillan *et al.*, 2011; Euser *et al.*, 2013; Smith *et al.*, 2013).

The utility of process consistent and transferable hydrologic models can be immense. Such models are not only critical to improving water resource management (particularly for periods outside the calibration period) but could also provide improved opportunities to reconcile models and data (Klemeš, 1983; Klemeš, 1986; Sivapalan *et al.*, 2003). However, for models to be considered process consistent, they must agree with observations or expert information about key catchment processes and patterns (e.g. Son and Sivapalan, 2007; Gupta *et al.*, 2008; Martinez and Gupta, 2011; Hrachowitz *et al.*, 2014).

Here, we explore the recently introduced catchment connectivity model (CCM) (Smith *et al.*, 2013) in terms of its transferability and fidelity, across multiple neighbouring catchments and through the lens of catchment classification and model regionalization. While regionalization studies may typically be carried out across hundreds of catchments (e.g. McIntyre *et al.*, 2005; Parajka *et al.*, 2005; Oudin *et al.*, 2008; Zhang and Chiew, 2009; Smith *et al.*, 2014), in this study, we advocate an approach that more intently focuses on model appropriateness in all test catchments as opposed to the sheer number of catchments considered. Indeed, by considering adjacent catchments, many of the difficulties in assessing model transferability are reduced; forcing climate, dominant processes, and general physiography are largely equivalent, although there can be differences across such catchments (e.g. geology, vegetation, slope, aspect, etc.) that lead to variability in hydrologic response (Jencso and McGlynn, 2011; Nippgen *et al.*, 2011). This paper addresses three questions: (1) How much variation is there in CCM parameterization and performance across hydrologically similar catchments? (2) How sensitive is CCM output to a priori parameter limits? and (3) How (to what degree) should expert information be used to enforce/ensure CCM process fidelity?

MATERIALS AND METHODS

Hydrologic model

We applied the CCM (Smith *et al.*, 2013) to seven catchments located within Tenderfoot Creek Experimental Forest (TCEF). In this study, the CCM was forced by measured liquid precipitation (rain plus snowmelt) and thus was not coupled with a snowmelt model. The CCM, conceptualized in Figure 1, was developed based on extensive empirical observations (Jencso *et al.*, 2009; Jencso *et al.*, 2010) at TCEF that showed strong correlations between upslope accumulated area (UAA; the lateral area draining to a particular point along the stream) and shallow groundwater hydrologic connectivity – defined by a continuous water table across the hillslope–riparian–stream continuum. The foundational concept of the model structure is that streamflow processes are driven by the frequency of hydrologic connections of the hillslopes to the stream rather than the magnitude of any single hydrologic connection (Jencso *et al.*, 2009).

Modelled hillslope storage (c, equation 1 of Figure 1) is computed via a mass balance considering precipitation (p), evapotranspiration (aet), and shallow groundwater (gw). Hydrologic connectivity of the hillslope to the stream is achieved when hillslope storage exceeds a threshold (c^*, equation 2 of Figure 1) related to hillslope UAA. During an active hydrologic connection ($h_c = 1$, equation 3 of Figure 1), water is released from the hillslope as shallow groundwater at a rate equal to q^* (a calibrated parameter), with an upper limit equal to the current storage (equation 4 of Figure 1). Each connected hillslope contributes water to catchment streamflow (q) via an exponential filter (equation 5 of Figure 1). Modelled hillslope storage is updated at the end of the time step (indexed by i) to account for the release of water as streamflow. The model was implemented using a distributed approach, where each catchment was discretized into hillslopes (indexed by j) based on discrete 10-m units of stream length (referred to as stream cells; for both the left-hand and right-hand sides of the stream). Despite its distributed nature, the CCM is a parsimonious, three parameter model (Table I), where streamflow is driven by the frequency of hydrologic connections and the upslope accumulated area-dependent threshold function. Refer to Smith *et al.* (2013) for additional details regarding model development and implementation.

Study site description

Tenderfoot Creek Experimental Forest, located in the Little Belt Mountains in Montana, USA, was used for this study. The location for numerous hydrologic studies (e.g. Smith and Marshall, 2008, 2010; Jencso *et al.*, 2009; Payn *et al.*, 2009, 2012; Jencso *et al.*, 2010; Jencso and McGlynn, 2011; Nippgen *et al.*, 2011; Smith *et al.*, 2013), TCEF (lat. 46°55′N, long. 1110°52′W) encompasses an area of 22.8 km^2 and is drained by Tenderfoot Creek, a tributary to the Smith River. It contains six principal gauged sub-catchments nested within the gauged lower Tenderfoot Creek (LTC) – Bubbling Creek (BUB), lower Stringer Creek (LSC; studied in Smith *et al.*, 2013), middle Stringer Creek (MSC), Spring Park Creek (SPC), Sun Creek (SUN), and upper Tenderfoot Creek (UTC). The six nested sub-catchments range in size

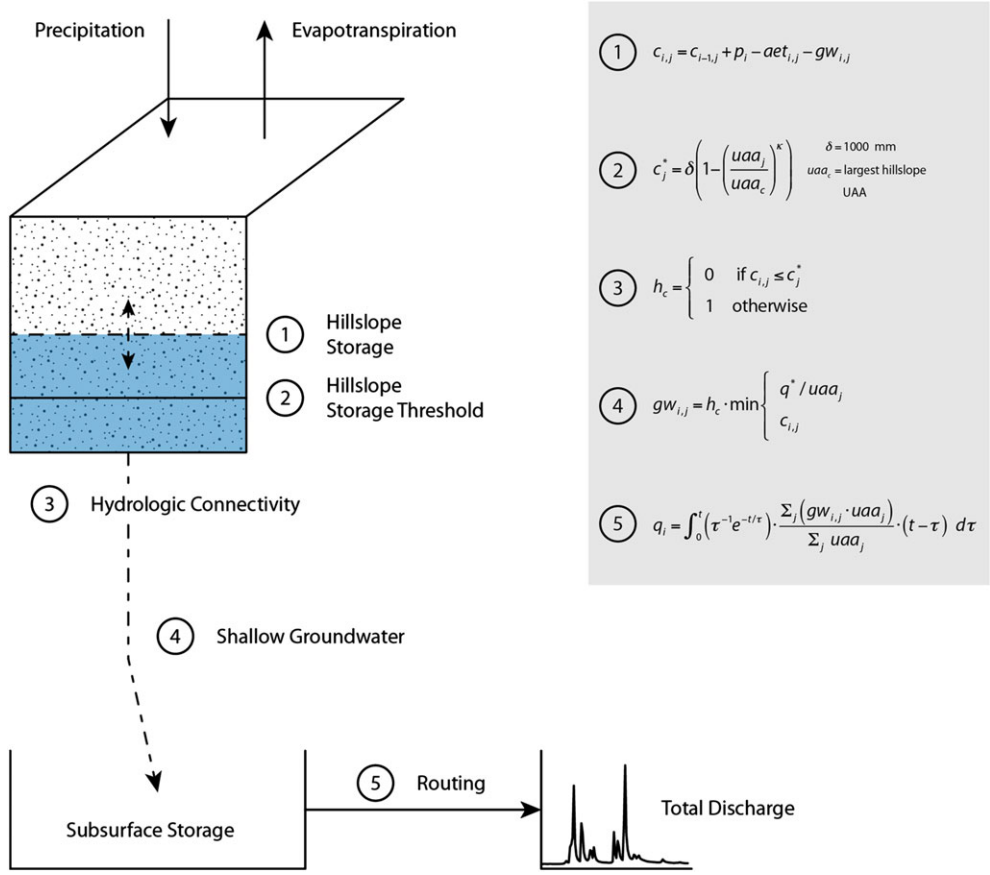

Figure 1. A schematic of the catchment connectivity model, highlighting the computational implementation for a single hillslope unit (reprint from Smith *et al.*, 2013)

Table I. Names and descriptions of the CCM parameters requiring calibration.

Parameter	Description	Limits	Units
q^*	The volumetric discharge from each hillslope per unit time	0–200	m^3/day
k	The Pareto parameter describing the shape of the UAA-connectivity relationship	0–1	—
τ	The residence time parameter of the exponential transfer function	0–7.5 [0–2][a]	Day

[a] Revised parameter limit based on field observations.

from approximately 3 to 5 km^2, while elevations range from 1840 to 2420 m. TCEF, whose mean annual temperature is 0 ° C, receives approximately 75% of its 880-mm average annual precipitation as snow. Vegetation at TCEF generally consists of lodgepole pines in the upland areas and willows, sedges, and rushes in the riparian areas (depending on water table and soil depths). Partial clearcuts were performed on two of the sub-catchments (SPC and SUN) between 1999 and 2001 via thinning and patch cutting treatments, resulting in the removal of approximately 50% of the tree basal area within the treatment areas (32% of SPC total area; 45% of SUN total area) (Nippgen *et al.*, 2011). The geology of higher elevation locations is typified by shale and sandstone that transitions to gneiss at lower elevations, corresponding to a transition from milder slopes (≈8°) near the catchment headwaters to steeper slopes (≈20°) near the catchment outlets (Jencso and McGlynn, 2011). For more detailed descriptions of the study area, refer to Jencso *et al.* (2009, 2010), Jencso and McGlynn (2011), and Nippgen *et al.* (2011).

Data

Model forcing data (precipitation and evapotranspiration), a digital elevation model, and model calibration data (streamflow) were obtained for each of the seven TCEF catchments for the analysis period of 1 October

2005 through 30 September 2009, on a 6-hourly time step. Shallow groundwater connectivity data were collected from 24 transects located across Stringer Creek and Tenderfoot Creek (Jencso *et al.*, 2009) and extrapolated across each sub-catchment located within TCEF (Jencso and McGlynn, 2011). Evapotranspiration data were obtained from a 40-m eddy-covariance tower located within MSC at TCEF (Emanuel *et al.*, 2010) and is assumed to be representative of the TCEF sub-catchments. Precipitation data were collected from the Stringer Creek (1996 m) and Onion Park (2259 m) SNOTEL stations located within TCEF (LSC and UTC catchments respectively). The model precipitation inputs include rain and snowmelt, with snowmelt being calculated using a simple algorithm (refer to Nippgen *et al.*, 2011) that is based on SNOTEL snow water equivalent, precipitation, and temperature data. Precipitation input series were based on an elevation-weighted input that utilized the data from both SNOTEL stations. For each catchment, 10-m digital elevation models (derived from 1-m LiDAR bare earth topography data) were used for landscape analysis, including the calculation of upslope accumulated areas following Grabs *et al.* (2010). Stream discharge was measured (as stage) on 30-min intervals, with ±1-mm resolution at the flumes located at the outlet of each catchment (refer to Nippgen *et al.*, 2011). Discharge data were aggregated to the 6-hourly time step used here.

Parameter estimation

Given the low dimensionality of the model (and thus relative simplicity of the parameter surface), we employed a Monte Carlo sampling approach to investigate the model parameters. Monte Carlo sampling allows for diagnostic assessment of the model response surface, parameter variability (within-catchment and cross-catchment), and parameter constraints based on expert knowledge. We performed 100 000 model simulations. The initial domain of each parameter is given in Table I and was set as wide to allow full and complete exploration of the parameter space. The model calibration was subsequently diagnostically sampled to restrict the value of the routing parameter (τ) from exceeding the expectation that τ should be no greater than 2 days based on extensive field observations at TCEF and estimates of in-stream travel times (Ward *et al.*, 2013). The results will be presented for both the unrestricted ($\tau < 7.5$ days) and restricted ($\tau < 2$ days) calibrations. Nash–Sutcliffe efficiency (NSE) is selected to analyse model fit, given its previous usage in the study catchments (Smith and Marshall, 2010; Smith *et al.*, 2013), ease of interpretation, and general applicability in snow-dominated catchments due to the pronounced snowmelt run-off peak. Model process fidelity is diagnosed based on the absolute bias between empirical and modelled (independent to parameter estimation) connectivity duration curves. The connectivity duration curve represents the binary hydrologic connectivity relationship (connected or disconnected) in terms of the exceedence probability for a given % of the stream network, with saturated connectivity to adjacent hillslopes.

PARAMETER VARIABILITY AND TRANSFERABILITY

In this demonstration, we sought to explore the parameterization of the CCM, with a focus on better understanding the cross-catchment variability in model parameterization and performance and ultimately in the direct transferability of parameter sets across catchments in terms of both hydrograph fit and process fidelity. Model (parameter) transferability was first benchmarked by performing local parameter estimation for each of the catchments.

A comparison of the optimal model parameterization for both the unrestricted and restricted calibration approaches at each of the test catchments (Table II) highlights the impact of enforcing field-based estimates of in-stream travel times. In each catchment, when the routing parameter (τ) is constrained to field-based limits,

Table II. Optimal CCM parameter values for site-specific model calibration.

			Unrestricted calibration ($\tau < 7.5$)			Restricted calibration ($\tau < 2$)		
Catchment	Area (ha)	uaa_c (ha)	q^* (m^3/day)	τ (day)	k (unitless)	q^* (m^3/day)	τ (day)	k (unitless)
Lower Tenderfoot (LTC)	2275.6	28.4	117.4	4.506	0.0481	92.3	1.981	0.0568
Bubbling (BUB)	305.9	21.8	79.0	4.280	0.0854	68.4	1.908	0.0997
Lower Stringer (LSC)	543.6	28.4	197.0	4.457	0.0744	149.9	1.996	0.0881
Middle Stringer (MSC)	407.3	28.4	199.9	3.136	0.0760	188.4	1.967	0.0899
Spring Park (SPC)	394.1	21.4	101.9	5.174	0.1035	67.1	1.962	0.1098
Sun (SUN)	352.8	20.1	66.0	1.927	0.0740	66.0	1.927	0.0740
Upper Tenderfoot (UTC)	467.0	22.5	73.2	2.750	0.1163	71.6	1.934	0.1217

the optimal values of the parameter representing the volumetric discharge from any connected hillslope (q^*) decrease, and the optimal values of the parameter defining the shape of the UAA-connectivity relationship (k) increase. This trade-off is a direct result of the effect of the exponential routing model component, where larger values of τ result in greater smoothing of the streamflow hydrograph and decreased magnitude for any given time (assuming the same input to the routing function). Reduced values of τ will result in less smoothing and an increased routed streamflow magnitude that must be balanced by a reduction in modelled flux (i.e. q^*).

Model fits to the hydrograph (as NSE) averaged 0.84 (Table III; site-specific column) for the unrestricted calibration – good values for TCEF, historically (Smith and Marshall, 2010; Smith *et al.*, 2013). However, despite the 'good' overall performance of the model (as NSE), the average optimal calibrated value of τ across the TCEF catchments was 3.75 days – nearly double the upper limit based on field knowledge. This should necessarily impact model process fidelity (if the system behaves as hypothesized) due to the mismatch between observation and model parameterization, despite providing a suitable match (as NSE) to the observed hydrograph.

To better understand the impact of the streamflow routing parameter on model performance, we diagnostically sampled the model calibration at each catchment by excluding parameter sets that exceeded the field-based upper constraint ($\tau = 2$ days). The restricting of model flexibility resulted in a small decrease in model fits to the hydrograph (as NSE) across all TCEF catchments (performance at SUN was unchanged), with the TCEF-average NSE reducing slightly from 0.84 to 0.81 (Table III; site-specific column). In all TCEF catchments except SUN, the optimal calibrated value of τ decreased (Table II), with the TCEF-average τ shifting from 3.75 to 1.95 days.

Notably, despite the marginally reduced fit to external catchment dynamics (streamflow), enforcing the field-based parameter limits on τ had a positive impact on the simulation of internal catchment dynamics (hydrologic connectivity). Modelled catchment connectivity was compared with catchment connectivity measured in LTC and LSC (and extrapolated to all of TCEF) by means of the connectivity duration curve. The average absolute bias of the connectivity duration curves (Table IV) across each of the seven TCEF catchments decreased from 0.27 to 0.23 (an approximately 15% improvement), with LTC and LSC experiencing the largest improvements in CDC fit. This is not overly surprising because loose constraints on τ allow the CCM to behave more like a transfer function model, thereby subsuming internal catchment processes into an exponential filter that no longer simply reflects in-channel travel times.

Table III. Comparison of site-specific and regionalized hydrograph fit as Nash–Sutcliffe efficiency.

	Unrestricted calibration ($\tau < 7.5$)		Restricted calibration ($\tau < 2$)	
Catchment	Site-specific NSE	Regionalized NSE	Site-specific NSE	Regionalized NSE
Lower Tenderfoot (LTC)	0.903	—	0.877	—
Bubbling (BUB)	0.852	0.785	0.825	0.752
Lower Stringer (LSC)	0.865	0.836	0.836	0.817
Middle Stringer (MSC)	0.856	0.791	0.841	0.769
Spring Park (SPC)	0.891	0.831	0.856	0.784
Sun (SUN)	0.792	0.646	0.792	0.725
Upper Tenderfoot (UTC)	0.697	0.543	0.692	0.576

Table IV. Comparison of site-specific and regionalized connectivity duration curve fit as absolute bias.

	Unrestricted calibration ($\tau < 7.5$)		Restricted calibration ($\tau < 2$)	
Catchment	Site-specific ABIAS	Regionalized ABIAS	Site-specific ABIAS	Regionalized ABIAS
Lower Tenderfoot (LTC)	0.309	—	0.155	—
Bubbling (BUB)	0.082	0.313	0.143	0.164
Lower Stringer (LSC)	0.530	0.291	0.417	0.151
Middle Stringer (MSC)	0.529	0.286	0.507	0.144
Spring Park (SPC)	0.214	0.304	0.174	0.150
Sun (SUN)	0.158	0.303	0.158	0.144
Upper Tenderfoot (UTC)	0.055	0.322	0.064	0.155

Examining the site-by-site results arising from the field-constrained parameterization of τ, model fit to observed streamflow across each of the seven test catchments was comparable in terms of the NSE (Table III), with the exception of the UTC catchment (Figure 2). The model tended to underpredict the hydrograph peaks for each catchment. Flow duration curves for each of the catchments were plotted (Figure 3) to consider model performance cumulatively without respect to timing. The underprediction of high streamflows was evident, along with the ability of the model to accurately simulate streamflows of approximately 10% exceedence and greater in all catchments, except SUN and UTC. Considering the ensemble bounds (based on the top 100 and top 1000 parameter sets), it was clear that the worst performing catchments (SUN, UTC) fail to capture the observed streamflow at intermediate exceedences (Figure 3).

Fidelity to hydrologic connectivity observations is presented in each catchment using connectivity duration curves (Figure 4). Note that the empirical connectivity duration curve was constructed by applying the regression relationship ($R^2 = 0.91$) between upslope accumulated area and connectivity developed by Jensco *et al.* (2009; refer to section 3.4) to the local inflows of UAA for each stream cell (on the left and right stream sides) across each of the test catchments. The results reveal significant variation across the best and worst performing catchments, with BUB, SPC, and UTC very closely matching the empirical curve and LSC and MSC deviating strongly

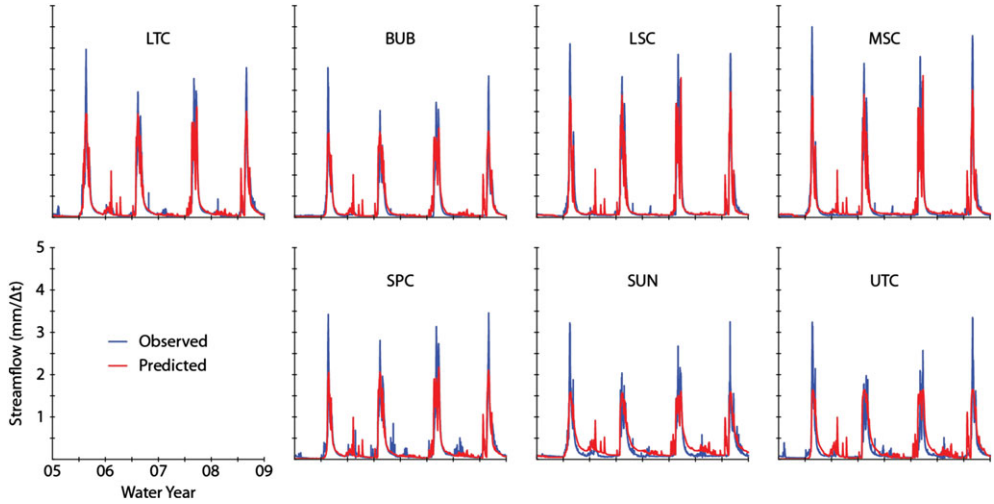

Figure 2. A comparison of the modelled and observed streamflow hydrographs for each catchment based on the restricted calibration

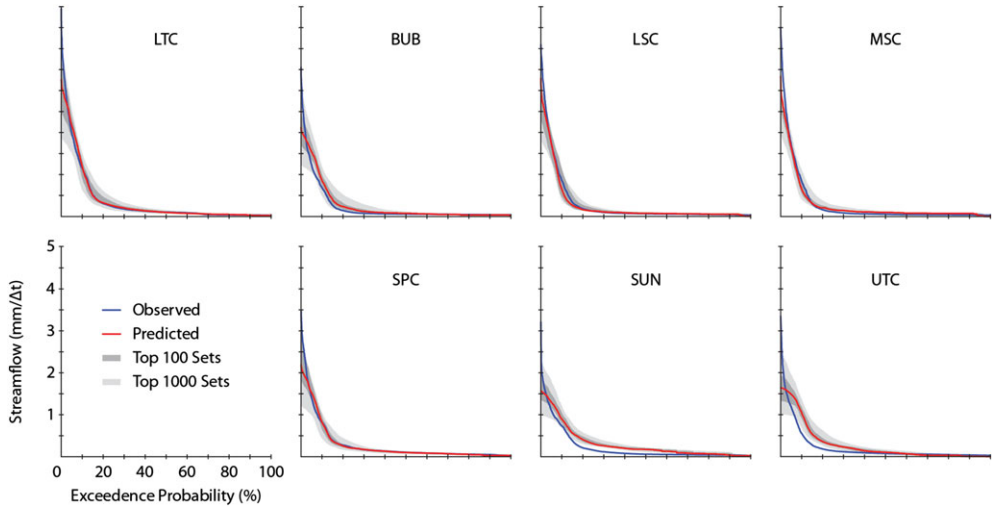

Figure 3. Flow duration curves (both observed and predicted) for each of the catchments included in this study, along with uncertainty envelopes for the top 100/1000 parameter sets based on the restricted calibration

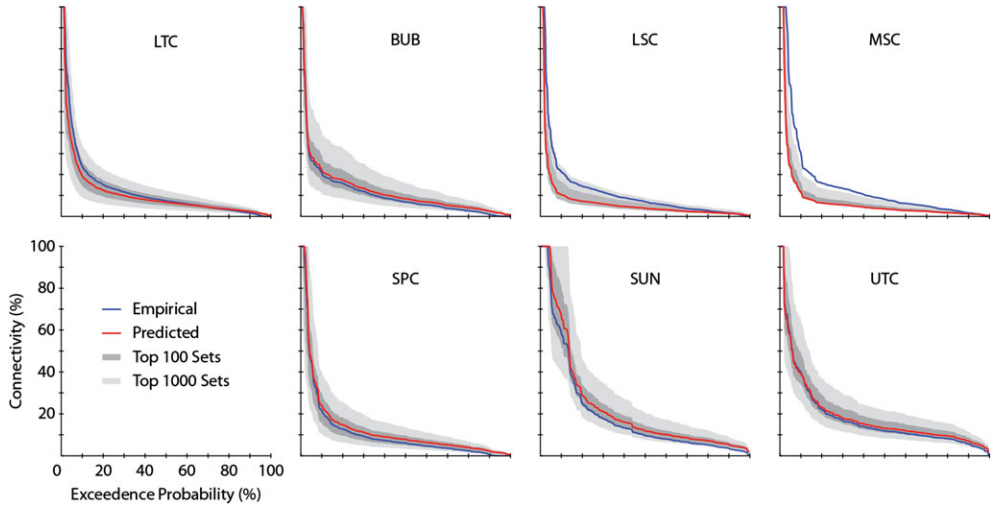

Figure 4. Connectivity duration curve for each of the catchments included in this study, along with uncertainty envelopes for the top 100/1000 parameter sets based on the restricted calibration. Note that the 'empirical' curve is based on an experimentally derived relationship that has been extrapolated to each of the individual catchments

from the empirical duration curve (Table IV). Considering the ensemble bounds constructed using the simulations of the top 100 and top 1000 parameter sets, the catchments with better correspondence to the CDC tended to have greater parameter sensitivity (i.e. fewer parameter sets that provide good simulations) that leads to wider ensemble bounds as a result.

The parameter distributions across each of the catchments show the most variation in the q^* model parameter, while the other model parameters (τ and k) show a high degree of consistency (Figure 5). Note that although the consistency in τ is partially attributable to it being restricted to satisfy field-based evidence, the unconstrained calibration of τ also resulted in largely consistent values (not shown). The variability in parameters found across catchments is not unexpected, as the model structure relies on catchment-specific upslope accumulated areas to drive the frequency of hydrologic connectivity that subsequently determines streamflow generation; however, trends in the parameter variation

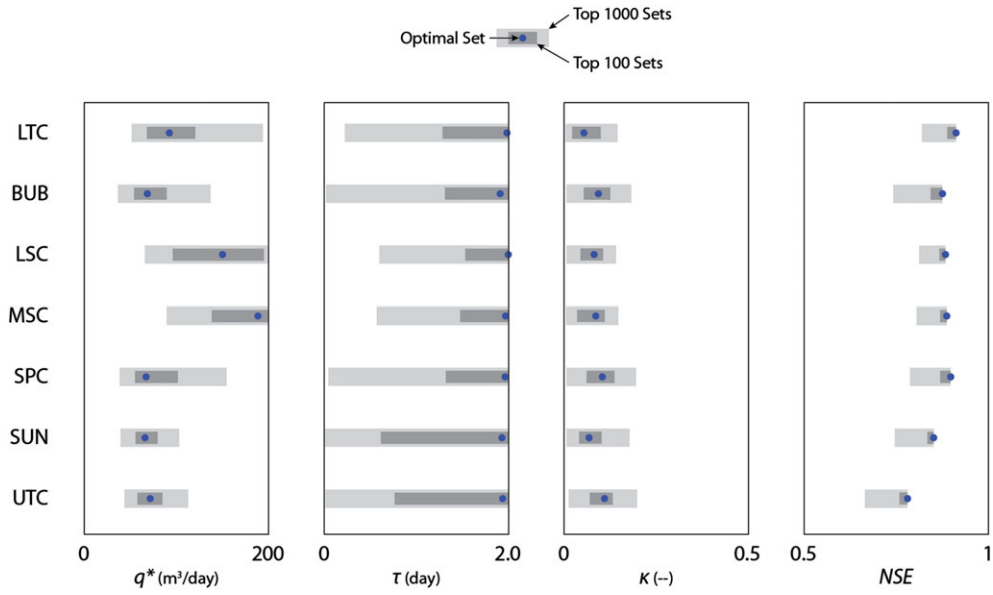

Figure 5. Plots of the posterior distributions for each of the calibrated catchment connectivity model parameters (based on restricted calibration) and each of the Tenderfoot Creek Experimental Forest catchments investigated

indicate potential differences in other physical catchment characteristics, model suitability, and/or model reliability outside the calibration period.

Model transferability was assessed with regard to the ability of parameters optimized for LTC to be successfully transferred to its nested sub-catchments (BUB, LSC, MSC, SPC, SUN, and UTC), relative to the expected optimal performance based on site-specific calibration (Table III). Establishing model transferability in TCEF using the model fit to LTC (i.e. using LTC as the donor catchment) was motivated by an interest in the scalability of the CCM parameters from catchment to nested sub-catchment. Model performance was generally well maintained, with an average reduction in NSE of 8.9% (ranging from 2.3% at LSC to 16.8% at UTC) for the restricted calibration. Model performance was only slightly less transferrable using the unrestricted calibration, with an average reduction in NSE of 11% (ranging from 3.3% at LSC to 22.2% at UTC). The connectivity duration curves, however, indicated a deviation in model transferability between the unrestricted and restricted calibrations. While the unrestricted calibration resulted in a reduction in the average performance (i.e. an increase in the absolute bias) of 13% (from 0.27 to 0.30), the restricted calibration resulted in an increase in the average performance (i.e. a reduction in the absolute bias) of 35% (from 0.23 to 0.15). The best performing catchments from the restricted calibration site-specific results (BUB, UTC) experienced reductions in overall fit (increases in absolute bias), while the worst performing catchments (LSC, MSC) saw improvements to overall fit (Table IV). For the unrestricted calibration, model transferability was less successful in terms of process fidelity (Table IV), where only the two worst performing catchments (LSC, MSC) saw improvements.

DISCUSSION

The CCM (Smith *et al.*, 2013) was developed based on extensive empirical observations at the TCEF that related upslope accumulated area to shallow groundwater connectivity (Jencso *et al.*, 2009; Jencso *et al.*, 2010). However, unlike other semi-distributed, topographically driven models (such as TOPMODEL (Beven and Kirkby, 1979), the manner in which the catchment is discretized based on stream cells is unique (refer to Smith *et al.* (2013) and the Section on Hydrologic Model in the preceding texts). In this formulation, catchment structure is a critical determinant of catchment run-off, just as the type, number, and/or configuration of conceptual storage components are central to many simple hydrologic models. Our study resulted in reasonable hydrograph fits (Table III; Figure 2), flow duration curves (Figure 3), and internal model process consistency (in terms of connectivity duration curves, Table IV; Figure 4) across all catchments when utilizing field-based estimates of acceptable parameter bounds. The CCM streamflow simulations had an average NSE of 0.81, despite the comparatively poor performance for the UTC catchment (NSE of 0.69) when utilizing field-informed parameter bounds.

Despite the apparent similarity (relief, elevation, average slope, annual climatology, etc.) of the catchments, there was notable variation in the optimal parameter values of q^* and, to a lesser extent, k (Figure 5). What controls such differences and can we identify a priori model parameter similarity? A comparison of the distributions of upslope accumulated area for each catchment (Figure 6) provides some insight into potential differences in run-off and connectivity modelling performance across the outwardly similar test catchments. Catchment structure (Figure 6) appears to influence parameter sensitivity (Figure 5). Streamflow generation is a function of the frequency of hillslope connections and the rate of flux for each hillslope connection (q^* is constant across all hillslope UAAs). As the distribution of catchment hillslope UAA (Figure 6) becomes flatter (i.e. having more hillslopes of different sizes), the hydrologic response becomes more sensitive to changes in q^*. At TCEF, catchments with flatter distributions of upslope accumulated area (e.g. UTC and SUN) were more sensitive to variations in model parameter q^* (hillslope volumetric discharge) than more peaked distributions (e.g. SPC and LSC). As a result, these catchments have a narrower range of acceptable values for q^* (Figure 5).

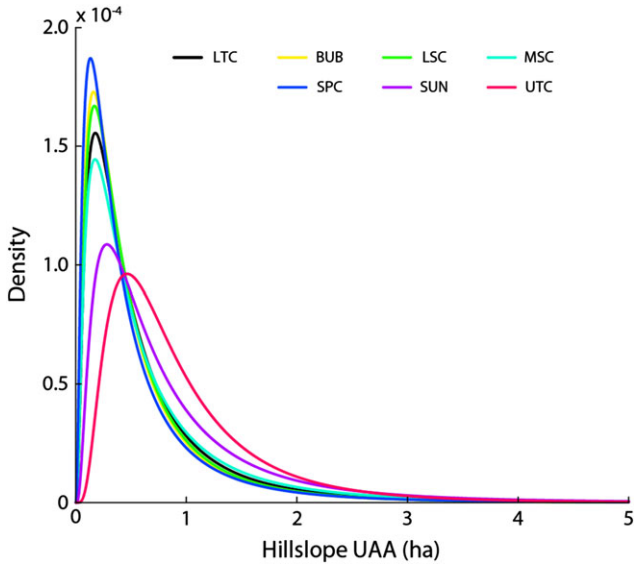

Figure 6. Probability density functions of catchment upslope accumulated area for each of the Tenderfoot Creek Experimental Forest catchments investigated

In our approach, we promote the value of empirical evidence in constraining model calibration. We diagnostically calibrated the model parameters by allowing wide bounds on all parameters initially and then excluding parameter sets that exceeded field-based limits ($\tau < 2$ days) subsequently – analogous to defining such parameter sets as non-behavioural in a generalized likelihood uncertainty estimation approach (Beven and Binley, 1992). While this may be uncomfortable to modellers accustomed to ticking off the boxes of a 'good' model calibration – namely a full exploration of the parameter space – the results of this study demonstrate that objective function optimization is not necessarily the path to optimal *whole* model performance. Using field knowledge to constrain the routing parameter τ resulted in a nearly 50% reduction in its average optimal calibrated value (Table II). This change in parameterization caused only a 3.5% reduction in TCEF-average model fit as NSE (0.84 to 0.81; Table III); however, in doing so, model process fidelity improved significantly for the two worst performing catchments (LSC, MSC), while only changing slightly in the other catchments where performance was already satisfactory (TCEF-average model process fidelity improved by approximately 15% as absolute bias of the CDC; Table IV). Given these results, the question then becomes: Would we rather allow a model to perform less realistically and reliably in deference to the idea of model optimization and the assumption that the training data (streamflow at the catchment outlet) and model structure are error free *or* would we rather enforce the observed reality in the system? We would argue that this is essentially the 'right answers for the right reasons' paradigm advocated by Kirchner (2006) and should be the defining goal of any modelling application.

A strength of the CCM structure is its spatial explicitness. The topographic template of the catchment serves as the foundation of the model structure, providing a correspondence between perceptual model and conceptual model. This correspondence affords the ability to both model and evaluate spatially explicit hillslope dynamics that are not part of the calibration of the model to streamflow at the catchment outlet. The relationship between catchment structure (Figure 6) and model process consistency (Table IV, Figure 4) is imperfect but evident, with catchments with flatter UAA distributions producing better fits to hydrologic connectivity. Potential factors including data (input and/or output) uncertainty, model structural uncertainty, or uncertainty in the empirical CDC may explain the variation in this relationship. Although this relationship was observed for a relatively small sample size (seven test catchments), hydrologic modelling is regularly confronted with the difficultly of reconciling the objectives of having a large enough sample size to make inferences about the population and having a model structure that is appropriate for the population (Gupta *et al.*, 2014). Often, in the context of predictions in ungauged basins, models are applied/transferred/regionalized to hundreds of catchments covering large variations in catchment characteristics. While this allows for a broad interpretation of parameter transferability, it disregards model applicability (i.e. process fidelity). Model applicability is (should be) a critical component to such approaches; if a model is not applicable to a catchment, successful parameter transferability should not be expected to begin with. In this study, we strongly advocate for a model process consistency first approach, and we suggest that the demonstrated model performance and process consistency is a function of the physical template upon which it is founded (Jencso *et al.*, 2009; Smith *et al.*, 2013).

The ensemble bounds were narrower for streamflow (Figure 3) than for hydrologic connectivity (Figure 4). This is to be expected, given that the model was calibrated only to streamflow and emphasizes that calibration and optimum parameter set (or ensemble) selection based solely on streamflow are problematic – not all simulations of streamflow are equivalent, with respect to model process fidelity. If one seeks to identify the best (most useful, reliable, transferable, consistent, etc.) model parameterization, model realism is a necessary assessment component (Seibert and McDonnell, 2002; Wagener, 2003).

The calibrated parameter values (Table II; Figure 5) for each catchment highlight both consistency and variation across the catchments. Optimal CCM parameters were largely consistent across all the catchments, with the exception of q^* in Stringer Creek (both LSC and MSC; Figure 5 and Table II). The q^* parameter represents the volumetric discharge from each hillslope per unit time. As this parameter increases, the potential for increased hydrologic flux also increases (equation 4 of Figure 1). The degree to which this effect is realized is modulated by catchment structure (i.e. the distribution of hillslope UAA; Figure 6), allowing more or less water to be released from the catchment. As a consequence, the Stringer Creek catchments (LSC, MSC) experienced the poorest internal fit (absolute bias of the CDC). In these catchments, connectivity was underpredicted – the model calibration to streamflow was achieved with fewer than (empirically) expected connected hillslopes. Variations in the value of uaa_c (the upslope accumulated area at which connectivity is assumed to be continuous) arise as a result of it being fixed to the value of the largest hillslope within the model domain (catchment). Potential improvements to model performance and/or process accuracy could be achieved through calibration of this parameter or the use of a TCEF-wide (regional) value.

Uncertainty in model input data is a major concern in hydrologic models (e.g. Kavetski *et al.*, 2006). Constrained by availability, our study assumed evapotranspiration to be constant across TCEF and employed site-specific, elevation-weighted precipitation inputs derived from the two SNOTEL stations located within TCEF. Additionally, snowmelt was determined directly from snowpack dynamic data from the SNOTEL stations (Nippgen *et al.*, 2011). Although the inputs used here represented the best available data, there is the potential that the systematic under prediction of streamflow during peak run-off was in part attributable to input error.

In addition to the errors in the model input data, model structural errors arising from process omission and/or numerical implementation are also of concern. The CCM structure is implemented using the traditional explicit Euler approach. Although explicit Euler is known to have potential drawbacks (Clark and Kavetski, 2010; Kavetski and Clark, 2010), the use of an implicit Euler approach would require an iterative solution that is particularly troublesome in distributed models due to the number of model evaluations necessary to apply the model over the study period and catchments. As Kavetski and Clark (2011) note, the use of uncontrolled time-stepping schemes (e.g. explicit Euler) should be expected to negatively impact simulation of internal model processes if the method is not suitable for the specific problem. At the same time, Kavetski and Clark (2011) demonstrated that hourly results (averaged to daily) obtained using the explicit Euler approach performed similarly to the daily results using the implicit Euler scheme, suggesting a potential inherent time scale-dependency related to the usage of explicit *versus* implicit Euler schemes. In this study, we have demonstrated consistent internal (hydrologic connectivity) and external (streamflow) model predictions, suggesting that for our model structure and data (on a 6-h time step), the use of the explicit Euler scheme is not problematic.

Regarding model structural uncertainty, flexible model structures have become increasingly popular (e.g. Clark *et al.*, 2011; Fenicia *et al.*, 2011; Kavetski and Fenicia, 2011; Staudinger *et al.*, 2011) and have the ability to incorporate new/alternate structures to increase flexibility and address perceived model structural shortcomings (Clark *et al.*, 2008). However, although the CCM currently underpredicts the highest flows, a previous analysis (Smith *et al.*, 2013) suggested that while alterations to the model structure could be effective at improving hydrograph fits (particularly during peak run-off), such structural modifications resulted in a significant reduction in process fidelity. This fact provides reinforcement to the reality that a focus on hydrograph representation alone is not sufficient for accurately representing hydrologic process behaviour.

CONCLUSIONS

This study presents an examination of the transferability and variability of parameterizations of the CCM, a conceptual hydrologic model developed following detailed experimental observations made within the TCEF. The intent of the model development was to create a simple, yet process-consistent (and verifiable) structure with the potential for improved model transferability to catchments with similar physical controls. Such a model provides a unique opportunity to undertake a diagnostic calibration approach. We utilized field-based knowledge of hydrologic processes to constrain model parameters by enforcing physically realistic boundaries (e.g. τ and q^*; Table I) and employed catchment observations of shallow subsurface hydrologic connectivity to assess model performance separate to the calibration process (CDC; Table IV and Figure 4).

An exploration of the model, individually calibrated across seven adjacent catchments, highlighted the consistency of the model simulations in terms of both external (i.e. streamflow) and internal (i.e. hydrologic connectivity) catchment processes. Despite the low dimensionality of the model, the CCM performed well in most of the catchments in terms of the streamflow hydrograph (Table III; Figure 2), flow duration (Figure 3), and connectivity duration (Table IV; Figure 4), illustrating the appropriateness of the model structure in representing the system processes and dynamics when field-based parameter limits are enforced. In a majority of the sites (e.g. LSC, MSC, SPC, and SUN), improvements to process fidelity (as hydrologic connectivity) could be achieved through a trade-off with model fit (as NSE). Despite the structural similarity of the test catchments, differences were noted in the model calibration in terms of both optimal model parameters (Figure 5) and the relative fit to observations (in terms of both streamflow and hydrologic connectivity) under site-specific and regionalized (Table III) applications. These differences are related to the differences in catchment structure (Figure 6) and/or uncertainties in the model structure or input data.

Simulations of model output (i.e. streamflow) with the CCM were shown to be marginally sensitive to a priori parameter constraints. However, physically constraining the routing model parameter resulted in improved fits to observed hydrologic connectivity and enhanced transferability of parameters to hydrologically similar catchments across all the catchments tested. These results reinforce past calls for increased cooperation between experimentalists and modellers (Seibert and McDonnell, 2002) as a path towards improved model process representation (i.e. the right answers for the right reasons; Kirchner, 2006).

ACKNOWLEDGEMENTS

This study was supported by NSF grant EAR-1356340 awarded to Marshall and McGlynn and EAR-0943640 awarded to McGlynn. The authors thank the US Forest Service, Rocky Mountain Research Station for site access and logistical support. Data used in this study can be obtained, at request, from the US Forest Service (http://www.fs.fed.us/rm/tenderfoot-creek/data/) for streamflow, from the National Resources Conservation Service (http://www.wcc.nrcs.usda.gov/snow/snotel-data.html) for meteorological variables, and from the National Center for Airborne Laser Mapping (10.5069/G92F7KCN) for topography. We are thankful to Wouter Knoben and one anonymous reviewer for their valuable contributions towards the improvement of this manuscript.

REFERENCES

Beven K, Binley A. 1992. The future of distributed models: model calibration and uncertainty prediction. *Hydrological Processes* **6**: 279–298.

Beven KJJ, Kirkby MJJ. 1979. A physically based, variable contributing area model of basin hydrology. *Hydrological Science Bulletin* **24**(1): 43–69. DOI:10.1080/02626667909491834.

Clark MP, Slater AG, Rupp DE, Woods RA, Vrugt JA, Gupta HV, Wagener T, Hay LE. 2008. Framework for understanding structural rrrors (FUSE): a modular framework to diagnose differences between hydrological models. *Water Resources Research* **44**: W00B02: DOI:10.1029/2007WR006735

Clark MP, Kavetski D, Fenicia F. 2011. Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resources Research* **47**: W09301: DOI:10.1029/2010wr009827

Clark MP, Kavetski D. 2010. Ancient numerical daemons of conceptual hydrological modeling: 1. Fidelity and efficiency of time stepping schemes. *Water Resources Research* **46**(10): W10510: DOI:10.1029/2009WR008894

Duan Q, Sorooshian S, Gupta V. 1992. Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resources Research* **28**: 1015–1031.

Duan Q, Schaake J, Andréassian V, Franks S, Goteti G, Gupta HV, Gusev YM, Habets F, Hall A, Hay L, Hogue T, Huang M, Leavesley G, Liang X, Nasonova ON, Noilhan J, Oudin L, Sorooshian S, Wagener T, Wood EF. 2006. Model parameter estimation experiment (MOPEX): an overview of science strategy and major results from the second and third workshops. *Journal of Hydrology* **320**: 3–17.

Emanuel RE, Epstein HE, McGlynn BL, Welsch DL, Muth DJ, D'Odorico P. 2010. Spatial and temporal controls on watershed ecohydrology in the northern Rocky Mountains. *Water Resources Research* **46**: W11553: DOI:10.1029/2009WR008890

Euser T, Winsemius HC, Hrachowitz M, Fenicia F, Uhlenbrook S, Savenije HHG. 2013. A framework to assess the realism of model structures using hydrological signatures. *Hydrology and Earth System Sciences* **17**: 1893–1912.

Fenicia F, Kavetski D, Savenije HHG. 2011. Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development. *Water Resources Research* **47**: W11510: DOI:10.1029/2010wr010174

Grabs TJ, Jencso KG, McGlynn BL, Seibert J. 2010. Calculating terrain indices along streams: a new method for separating stream sides. *Water Resources Research* **46**(12): W12536: DOI:10.1029/2010WR009296

Gupta HV, Perrin C, Blöschl G, Montanari A, Kumar R, Clark M, Andréassian V. 2014. Large-sample hydrology: a need to balance depth with breadth. *Hydrology and Earth System Sciences* **18**(2): 463–477. DOI:10.5194/hess-18-463-2014.

Gupta HV, Sorooshian S, Yapo PO. 1998. Toward improved calibration of hydrologic models: multiple and noncommensurable measures of information. *Water Resources Research* **34**: 751–763.

Gupta HV, Wagener T, Liu Y. 2008. Reconciling theory with observations: elements of a diagnostic approach to model evaluation. *Hydrological Processes* **22**(18): 3802–3813. DOI:10.1002/hyp.6989.

Hrachowitz M, Fovet O, Ruiz L, Euser T, Gharari S, Nijzink R, Freer J, Savenije HHG, Gascuel-Odoux C. 2014. Process consistency in models: The importance of system signatures, expert knowledge, and process complexity, Water Resources. *Research* **50**(9): 7445–7469. DOI:10.1002/2014WR015484.

Jencso KG, McGlynn BL, Gooseff MN, Wondzell SM, Bencala KE, Marshall LA. 2009. Hydrologic connectivity between landscapes and streams: transferring reach- and plot-scale understanding to the catchment scale. *Water Resources Research* **45**: W04428: DOI:10.1029/2008WR007225

Jencso KG, McGlynn BL, Gooseff MN, Bencala KE, Wondzell SM. 2010. Hillslope hydrologic connectivity controls riparian groundwater turnover: implications of catchment structure for riparian buffering and stream water sources. *Water Resources Research* **46**: W10524: DOI:10.1029/2009WR008818

Jencso KG, McGlynn BL. 2011. Hierarchical controls on runoff generation: topographically driven hydrologic connectivity, geology, and vegetation. *Water Resources Research* **47**: W11527: DOI:10.1029/2011wr010666

Kavetski D, Kuczera G, Franks SW. 2006. Bayesian analysis of input uncertainty in hydrological modeling: theory. *Water Resources Research* **42**: W03407: DOI:10.1029/2005WR004368

Kavetski D, Clark MP. 2010. Ancient numerical daemons of conceptual hydrological modeling: 2. Impact of time stepping schemes on model analysis and prediction. *Water Resources Research* **46**(10): W10511: DOI:10.1029/2009WR008896

Kavetski D, Clark MP. 2011. Numerical troubles in conceptual hydrology: approximations, absurdities and impact on hypothesis testing. *Hydrological Processes* **25**(4): 661–670.

Kavetski D, Fenicia F. 2011. Elements of a flexible approach for conceptual hydrological modeling: 2. Application and experimental insights. *Water Resources Research* **47**: W11511: DOI:10.1029/2011wr010748

Kirchner JW. 2006. Getting the right answers for the right reasons: linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research* **42**: W03S04: DOI:10.1029/2005WR004362

Klemeš V. 1983. Conceptualization and scale in hydrology. *Journal of Hydrology* **65**: 1–23. DOI:10.1016/0022-1694(83)90208-1

Klemeš V. 1986. Operational testing of hydrological simulation models. *Hydrological Sciences Journal* **31**: 13–24.

Martinez GF, Gupta HV. 2011. Hydrologic consistency as a basis for assessing complexity of monthly water balance models for the continental United States. *Water Resources Research* **47**(12): W12540: DOI:10.1029/2011wr011229.

McIntyre N, Lee H, Wheater H, Young A, Wagener T. 2005. Ensemble predictions of runoff in ungauged catchments. *Water Resources Research* **41**: W12434: DOI:10.1029/2005WR004289.

McMillan HK, Clark MP, Bowden WB, Duncan M, Woods RA. 2011. Hydrological field data from a modeller's perspective: part 1. Diagnostic tests for model structure. *Hydrological Processes* **25**: 511–522.

Nash JE, Sutcliffe JV. 1970. River flow forecasting through conceptual models part I — a discussion of principles. *Journal of Hydrology* **10**: 282–290.

Nippgen F, McGlynn BL, Marshall LA, Emanuel RE. 2011. Landscape structure and climate influences on hydrologic response. *Water Resources Research* **47**: W12528: DOI:10.1029/2011wr011161

Oudin L, Andráassian V, Perrin C, Michel C, Le Moine N. 2008. Spatial proximity, physical similarity, regression and ungaged catchments: A comparison of regionalization approaches based on 913 French catchments. *Water Resources Research* **44**: W03413: DOI:10.1029/2007WR006240.

Parajka J, Merz R, Blöschl G. 2005. A comparison of regionalisation methods for catchment model parameters, *Hydrology and Earth System Sciences* **9**(3): 157–171. DOI: 10.5194/hess-9-157-2005.

Payn RA, Gooseff MN, McGlynn BL, Bencala KE, Wondzell SM. 2009. Channel water balance and exchange with subsurface flow along a

mountain headwater stream in Montana, United States. *Water Resources Research* **45**: W11427: DOI:10.1029/2008WR007644

Payn RA, Gooseff MN, McGlynn BL, Bencala KE, Wondzell SM. 2012. Exploring changes in the spatial distribution of stream baseflow generation during a seasonal recession. *Water Resources Research* **48**: W04519: DOI:10.1029/2011wr011552

Reed S, Koren V, Smith M, Zhang Z, Moreda F, Seo D-J. 2004. Overall distributed model intercomparison project results. *Journal of Hydrology* **298**: 27–60.

Seibert J, McDonnell JJ. 2002. On the dialog between experimentalist and modeler in catchment hydrology: use of soft data for multicriteria model calibration. *Water Resources Research* **38**: 1241. DOI:10.1029/2001WR000978

Sivapalan M, Blöschl G, Zhang L, Vertessy R. 2003. Downward approach to hydrological prediction. *Hydrological Processes* **17**: 2101–2111.

Smith T, Marshall L, McGlynn B, Jencso K. 2013. Using field data to inform and evaluate a new model of catchment hydrologic connectivity. *Water Resources Research* **49**: 6834–6846.

Smith T, Marshall L, McGlynn B. 2014. Calibrating hydrologic models in flow-corrected time. *Water Resources Research* **50**: 748–753.

Smith TJ, Marshall LA. 2008. Bayesian methods in hydrology: a study of recent advancements in Markov chain Monte Carlo techniques. *Water Resources Research* **44**: W00B05: DOI:10.1029/2007WR006705

Smith TJ, Marshall LA. 2010. Exploring uncertainty and model predictive performance concepts via a modular snowmelt-runoff modeling framework. *Environmental Modelling & Software* **25**: 691–701.

Son K, Sivapalan M. 2007. Improving model structure and reducing parameter uncertainty in conceptual water balance models through the use of auxiliary data. *Water Resources Research* **43**: W01415: DOI:10.1029/2006WR005032

Staudinger M, Stahl K, Seibert J, Clark MP, Tallaksen LM. 2011. Comparison of hydrological model structures based on recession and low flow simulations. *Hydrology and Earth System Sciences* **15**: 3447–3459.

Wagener T. 2003. Evaluation of catchment models. *Hydrological Processes* **17**: 3375–3378.

Wagener T, Gupta HV. 2005. Model identification for hydrological forecasting under uncertainty. *Stochastic Environmental Research and Risk Assessment* **19**: 378–387.

Ward AS, Payn RA, Gooseff MN, McGlynn BL, Bencala KE, Kelleher CA, Wondzell SM, Wagener T. 2013. Variations in surface water-ground water interactions along a headwater mountain stream: Comparisons between transient storage and water balance analyses. *Water Resources Research* **49**(6): 3359–3374. DOI:10.1002/wrcr.20148.

Zhang Y, Chiew FHS. 2009. Relative merits of different methods for runoff predictions in ungauged catchments, *Water Resources Research* **45**: W07412: DOI:10.1029/2008WR007504.

