Characterizing and reducing equifinality by constraining a distributed catchment model with regional signatures, local observations, and process understanding

Christa Kelleher1,2, Brian McGlynn2, and Thorsten Wagener3,4

1Department of Earth Sciences, Syracuse University, Syracuse, NY, 13244, USA
2Department of Earth and Ocean Sciences, Nicholas School of the Environment, Duke University, Durham, NC, 27706, USA
3Department of Civil Engineering, University of Bristol, Bristol, BS8 1TR, UK
4Cabot Institute, University of Bristol, Bristol, BS8 1TR, UK

Correspondence to: Christa Kelleher (ckellehe@syr.edu)

Received: 1 December 2016 – Discussion started: 22 December 2016
Revised: 17 April 2017 – Accepted: 26 May 2017 – Published: 5 July 2017

Abstract. Distributed catchment models are widely used tools for predicting hydrologic behavior. While distributed models require many parameters to describe a system, they are expected to simulate behavior that is more consistent with observed processes. However, obtaining a single set of acceptable parameters can be problematic, as parameter equifinality often results in several “behavioral” sets that fit observations (typically streamflow). In this study, we investigate the extent to which equifinality impacts a typical distributed modeling application. We outline a hierarchical approach to reduce the number of behavioral sets based on regional, observation-driven, and expert-knowledge-based constraints. For our application, we explore how each of these constraint classes reduced the number of “behavioral” parameter sets and altered distributions of spatiotemporal simulations, simulating a well-studied headwater catchment, Stringer Creek, Montana, using the distributed hydrology–soil–vegetation model (DHSVM). As a demonstrative exercise, we investigated model performance across 10,000 parameter sets. Constraints on regional signatures, the hydrograph, and two internal measurements of snow water equivalent time series reduced the number of behavioral parameter sets but still left a small number with similar goodness of fit. This subset was ultimately further reduced by incorporating pattern expectations of groundwater table depth across the catchment. Our results suggest that utilizing a hierarchical approach based on regional datasets, observations, and expert knowledge to identify behavioral parameter sets can reduce equifinality and bolster more careful application and simulation of spatiotemporal processes via distributed modeling at the catchment scale.

1 Introduction

The field of hydrology has been built upon the combination of field measurements and computational modeling to observe, understand, and predict hydrologic behavior (Crawford and Linsley, 1966; Beven and Kirkby, 1976, 1979; Ponce and Shetty, 1995a, b). Towards this end, distributed, physically based models were first developed as tools to represent spatially discretized processes with physically meaningful parametric relationships (Freeze and Harlan, 1969; Beven and Kirkby, 1979; Band et al., 1991, 1993; Wigmosta et al., 1994; Refsgaard and Storm, 1995; Bixio et al., 2002; Qu, 2004; Qu and Duffy, 2007; Camporese et al., 2010; Fatichi et al., 2016). Distributed models should represent the characteristics of the catchment environment in space, given that this variability impacts how water is stored, partitioned, and released across the landscape (O’Loughlin, 1981; Beven, 1989; Kampf and Burges, 2007; Wagener et al., 2007; Nippgen et al., 2015). Many physically based, distributed models use similar equations to predict water movement, have comparably large data input requirements, and feature numerous
model parameters (see Singh and Woolhiser, 2002; Kampf and Burges, 2007; Paniconi and Putti, 2015, for a review of distributed models).

Alongside variations in the structural equations, distributed models may take a number of different forms, including those that use only distributed inputs (e.g., Ajami et al., 2004; Das et al., 2008; Fencic et al., 2008; Kling and Gupta, 2009; Euser et al., 2015), those that may lump together portions of a watershed with similar characteristics in a semi-distributed fashion (e.g., Levesley et al., 1983; Flugel, 1995; Ajami et al., 2004; Das et al., 2008; Zehe et al., 2014), those that may upscale sub-grid variability to parameterize models (e.g., Samaniego et al., 2010; Kumar et al., 2013), and those that use physically based parameter values (e.g., Qu and Duffy, 2007; Wigmosta et al., 1994) versus conceptual values or transfer functions (e.g., Samaniego et al., 2010; Kumar et al., 2013). These differences can be separated into two primary categories (1) the nature of parameter values (physically based or conceptual) and (2) whether and how parameter values are distributed (cell-by-cell, grouped or lumped in some meaningful way, or undistributed). It is also worth noting that an important feedback on these decisions is the scale of the application alongside the scale at which inputs and parameters are distributed. In this study, we focus specifically on physically based, fully distributed (cell-by-cell basis) models applied at the small-scale watershed scale (<25 km²).

Despite the widespread application and advancement of physically based, fully distributed (cell-by-cell basis) models, one of the foremost challenges in distributed model application continues to be parameter estimation (Beven and Binley, 1992; Gupta et al., 1998; Wagener and Gupta, 2005; Beven, 2006; Gharari et al., 2014; Chen et al., 2017). Model simulations and predictions require specification of parameter set(s) or ranges; selecting these set(s) and appropriate ranges is especially challenging given that fully distributed models, where inputs such as soil or vegetation are distributed across the watershed, require a larger number of model parameters (~50–100 or more) and longer model run times than conceptual or lumped models. In particular, a large number of model parameters, corresponding to a large number of degrees of freedom, may lead to equifinality (Hornberger and Spear, 1980, 1981; Spear and Hornberger, 1980; Beven and Binley, 2014), a well-documented problem with respect to all models (Beven, 1989, 1993, 2001). Many studies have documented the presence of parameter equifinality in terms of parameter sensitivity, concluding that the potential for equifinality in distributed model applications is high in both time (Franks et al., 1998; van Werkhoven et al., 2008; Zhang et al., 2013; Kelleher et al., 2015; Ghasemizade et al., 2017; Guse et al., 2016) and space (Wagener et al., 2009; Herman et al., 2013; Moreau et al., 2013).

A great deal of research has been devoted to developing approaches that either seek to reduce or to explore equifinality across multiple parameter sets, but these approaches are typically demonstrated for lumped or conceptual models (e.g., Keesman, 1990; van Straten and Keesman, 1991; Beven and Binley, 1992; Gupta et al., 1998). Currently, researchers approach distributed modeling and parameter set selection in a number of different ways, all of which have implications for reducing equifinality. Many studies avoid the topics of uncertainty or equifinality by assigning parameter values from measurements (e.g., Du et al., 2014), though the sheer number of parameters, heterogeneity of the catchment environment, uncertainty in model structure and inputs, and problems translating measurements from the point to the grid scale can make this difficult (Grayson et al., 1992; Surfleet et al., 2010; Paniconi et al., 2015). Other approaches involve fixing some parameters, while letting others vary, often through manual calibration. Decisions regarding how to constrain ranges and distributions for priors on model parameters often depend on the availability of field measurements, awareness of the catchment or model, and may involve sensitivity analysis to identify which parameters are most influential to simulations (Tang et al., 2007; Saltelli et al., 2008; Cuo et al., 2011). A subset of distributed modeling studies has directly incorporated parameter uncertainty into final simulations, sampling across ranges of values obtained either from literature or measurements (e.g., Surfleet et al., 2010; Shields and Tague, 2015; Gharari et al., 2014; Silvester et al., 2015). Very few studies have shown the implications of this equifinality during model calibration, an important consideration given that selection of a parameter set or sets will influence conclusions made for validation and scenario analysis. One noted exception is the work that has been done to parameterize models via regularization (Hundecha and Bardossy, 2004; Hundecha et al., 2008; Pohkrel et al., 2008; Samaniego et al., 2010; Rakovec et al., 2016). Regularization creates global functional relationships describing transfer functions that link model parameters and catchment characteristics (e.g., Samaniego et al., 2010; Kumar et al., 2013). As the number of parameters used to describe global functional relationships is far fewer than the number of possible total model parameters using cell-by-cell parameterization, regularization is able to limit parameter equifinality.

Researchers must also determine which observations to incorporate and compare to model simulations (as well as which criteria to measure “goodness of fit” by, and how to judge whether or not this “goodness of fit” is good enough) in order to justify the selection of a parameter set or sets (Bennett et al., 2013). While there are examples where different observations, including snow accumulation and melt (Thyer et al., 2004; Whitaker et al., 2003), soil moisture (Koren et al., 2008; Graeff et al., 2012), and catchment chemistry (Birikel et al., 2014), have been incorporated into the calibration and parameter set selection process, many catchments lack measurements beyond streamflow, suggesting that other sources of information are needed (Yapo et al., 1998; Grayson et al., 2002; Paniconi et al., 2015). Finally, while examples exist where model simulations are
compared to internal measurements of different hydrologic processes at a few points, there are fewer examples where researchers holistically evaluate the patterns of these simulated processes (Franks et al., 1998; Lamb et al., 1998; Grayson et al., 2002; Wealands et al., 2005; Koch et al., 2016).

Given many of the outlined challenges discussed above, we need to improve our understanding of equifinality in distributed model applications and to better constrain this equifinality towards predictive use of distributed models. In this study, we demonstrate an approach to characterizing and reducing equifinality for distributed models. We apply this approach as a case study in the Stringer Creek headwater catchment, located in Tenderfoot Creek Experimental Forest in central Montana, modeled with the widely applied distributed hydrology–soil–vegetation model (DHSVM). We investigate how the paradigm for identifying “behavioral” model simulations, defined as those that meet a certain criterion or multiple criteria for error with respect to observations, may impact the presence of equifinality in terms of parameter estimation and constraining simulations and predictions of different hydrologic processes. Within our proposed framework, we test many of the subjective choices a modeler must make during calibration with respect to impacts on parameter set selection, parameter values, model performance, and model simulations. We explore model performance for a shorter period of time but for a large number of model runs. Our goal is to support distributed models’ utilization to their full potential and ensure that parameter set(s) used for prediction or scenario analysis match hydrologic observations and perceptions in both space and time. Secondly, we also explore whether this type of approach, using observations of a subset of hydrologic processes, may inform simulations of other unmeasured spatially distributed hydrologic processes. Thus, we seek to test whether temporal observations may contain information regarding simulation of spatially distributed hydrologic patterns. While our approach is demonstrated for a single catchment for a relatively short time period, it has broader implications for the use of alternative data sources in the parameter estimation process as well as the application of distributed models to simulate internal catchment behavior.

2 Case study: Stringer Creek, Tenderfoot Creek Experimental Forest

2.1 Study site

We present a case study applying this methodology to a headwater catchment (5.5 km²) located in the Tenderfoot Creek Experimental Forest in central Montana. The case study focuses on Stringer Creek, though we simulate the entire Tenderfoot Creek catchment (22.5 km²). The Stringer Creek headwater catchment is located in Tenderfoot Creek Experimental Forest in central Montana. Tenderfoot Creek experiences a continental climate with the majority of its 880 mm of precipitation falling as snow from November through May (Farnes et al., 1995). Snowmelt, occurring in May or June, is the primary driver of the hydrologic cycle. Bedrock under Stringer Creek includes biotite hornblende quartz monzonite overlaying a layer of shale in the upper parts of the catchment and flathead sandstone underlain by granite gneiss in the lower portions of the catchment (Reynolds, 1995). Lodgepole pine dominates the forest vegetation types, with shrubs and grasses occurring in riparian areas (Farnes et al., 1995; Mincemoyer and Birdsall, 2006). Stringer Creek, the most well-studied catchment in Tenderfoot Creek Experimental Forest, is a second-order catchment that drains an area of 5.5 km².

2.2 The distributed hydrology soil vegetation model

The entire Tenderfoot Creek catchment was modeled using the distributed hydrology–soil–vegetation model (Wigmosta et al., 1994, 2002). DHSVM is a physically based, spatially distributed catchment model typically used to simulate mountainous catchments at small to intermediate scales in the Pacific Northwest and the Mountain West. The model includes a two-layer snow accumulation and melt model with a full energy balance and a Penman–Monteith approach for simulating evapotranspiration. Unsaturated soil water movement is simulated via Darcy’s law with hydraulic conductivity calculated via the Brooks–Corey equation. Saturated subsurface flow is routed cell by cell using either a kinematic or diffusion approximation. Streamflow is routed through a user-defined stream network via linear channel reservoirs. The model framework and forcing data used to simulate Tenderfoot Creek were previously employed and described in Kelleher et al. (2015). Key details for spatial and meteorological forcing data are described below.

2.2.1 Spatial forcing data

Tenderfoot Creek was simulated within DHSVM at a resolution of 10 m and a time step of 3 h using spatially distributed information for topography, soil depth, soil type, and vegetation (Fig. 1). Spatially distributed information for topography and vegetation height were obtained via airborne laser swath mapping (ALSM) at 1 m resolution and resampled to 10 m resolution (Fig. 1a, b). Soil depth was held at a constant 1 m, as spatially distributed soil data are limited. However, >160 well and piezometer installations indicate that 1 m is a reasonable average with limited variance (Jencso et al., 2009). DHSVM distributes parameter values based on vegetation and soil “types”, with one vegetation and soil type assigned to each cell within the catchment. To minimize potential for equifinality, we employed a model framework to distribute soil and vegetation parameters using as few different classes as possible while still representing functional differences that we expected to impact hydrology across the catchment.
Vegetation was grouped into three classes based on vegetation height, as vegetation species are relatively homogenous across the catchment (Farnes et al., 1995; Mincemoyer and Birdsall, 2007) and height is an important determinant of aerodynamic resistance, used within both the snow and evapotranspiration modules to estimate catchment response on a cell-by-cell basis (Wigmosta et al., 2002). Vegetation classes included tall trees (>10 m), trees (2–10 m), and undergrowth (<2 m; Fig. 1b). Vegetation parameters were distributed by vegetation type, with individual parameters partitioned between an understory (13 parameters) and an overstory (17 parameters). Cells with only undergrowth include only understory parameters; cells with a canopy (trees or tall trees) have both an understory and an overstory. Parameter values associated with understory vegetation were varied concurrently across the catchment. Overstory parameters were varied in

Figure 1. Tenderfoot Creek Experimental Forest (a) digital elevation model (DEM) and instrumentation used in the study and (b) vegetation type used by DHSVM to distribute model parameters. (a) Instrumentation includes a streamflow calibration station at the outlet of Stringer Creek (LSC) as well as streamflow gaging stations at the outlet of Tenderfoot Creek (LTC) and middle Stringer Creek (MSC). Most forcing data are measured at SNOTEL sites Stringer Creek and Onion Park. Water balance time series are extracted at both SNOTEL locations, as well as streamflow behavior at the Stringer Creek outlet. (b) Vegetation type is distributed based on vegetation height, with cells with undergrowth vegetation shown in white, cells with canopy vegetation shown in light green, and cells with tall tree vegetation shown in dark green. (c) The model is forced with air temperature and precipitation data collected at the two SNOTEL locations. Catchment observations include snow water equivalent at two SNOTEL sites and streamflow at three gauges.
Table 1. Parameter types, names, numbers (with reference to Fig. 8), and ranges for all 53 parameters included in the analysis. Sources for ranges are detailed in Kelleher et al. (2015).

<table>
<thead>
<tr>
<th>No.</th>
<th>Name</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Snow/climate</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Snow water capacity (–)</td>
<td>0.01</td>
<td>0.08</td>
</tr>
<tr>
<td>2</td>
<td>Rain LAI multiplier (–)</td>
<td>0.0002</td>
<td>0.001</td>
</tr>
<tr>
<td>3</td>
<td>Snow LAI multiplier (–)</td>
<td>0.0002</td>
<td>0.0025</td>
</tr>
<tr>
<td>4</td>
<td>Minimum intercepted snow (–)</td>
<td>0.001</td>
<td>0.01</td>
</tr>
<tr>
<td>5</td>
<td>Snow threshold (°C)</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>Rain threshold (°C)</td>
<td>–2</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>Lateral conductivity (KLAT) (m s⁻¹)</td>
<td>1 × 10⁻⁵</td>
<td>1 × 10⁻²</td>
</tr>
<tr>
<td>8</td>
<td>Exponential decrease in KLAT with depth (–)</td>
<td>0.5</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>Maximum infiltration (m s⁻¹)</td>
<td>3.6 × 10⁻⁵</td>
<td>5.38 × 10⁻⁴</td>
</tr>
<tr>
<td>10</td>
<td>Capillary drive (–)</td>
<td>0.03</td>
<td>0.6</td>
</tr>
<tr>
<td>11</td>
<td>Surface albedo (–)</td>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>12</td>
<td>Porosity (–)</td>
<td>0.38</td>
<td>0.47</td>
</tr>
<tr>
<td>13</td>
<td>Pore size distribution index (–)</td>
<td>0.07</td>
<td>0.559</td>
</tr>
<tr>
<td>14</td>
<td>Bubbling pressure (m)</td>
<td>0</td>
<td>1.24</td>
</tr>
<tr>
<td>15</td>
<td>Field capacity (–)</td>
<td>0.15</td>
<td>0.25</td>
</tr>
<tr>
<td>16</td>
<td>Wilting point (–)</td>
<td>0.07</td>
<td>0.15</td>
</tr>
<tr>
<td>17</td>
<td>Bulk density (kg m⁻³)</td>
<td>1390</td>
<td>1650</td>
</tr>
<tr>
<td>18</td>
<td>Vertical conductivity (m s⁻¹)</td>
<td>1.42 × 10⁻¹</td>
<td>2.02 × 10⁻⁴</td>
</tr>
<tr>
<td>19</td>
<td>Thermal conductivity (W m⁻¹ °C⁻¹)</td>
<td>0.3</td>
<td>0.8</td>
</tr>
<tr>
<td>20</td>
<td>Thermal capacity (J m⁻³ °C⁻¹)</td>
<td>1 × 10⁶</td>
<td>3 × 10⁶</td>
</tr>
<tr>
<td>21</td>
<td>Manning’s n (–)</td>
<td>0.11</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td><strong>Soils</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Maximum snow interception capacity (–)</td>
<td>0.05</td>
<td>0.2</td>
</tr>
<tr>
<td>23</td>
<td>Mass release drip ratio (–)</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>24</td>
<td>Snow interception efficiency (–)</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>25</td>
<td>Height (m)</td>
<td>0</td>
<td>1.2</td>
</tr>
<tr>
<td>26</td>
<td>Maximum resistance (s m⁻¹)</td>
<td>500</td>
<td>10000</td>
</tr>
<tr>
<td></td>
<td><strong>Overstory – vegetation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>Minimum resistance (s m⁻¹)</td>
<td>100</td>
<td>175</td>
</tr>
<tr>
<td>28</td>
<td>Moisture threshold (–)</td>
<td>0.115</td>
<td>0.165</td>
</tr>
<tr>
<td>29</td>
<td>Vapor pressure deficit (Pa)</td>
<td>200</td>
<td>4000</td>
</tr>
<tr>
<td>30</td>
<td>Fraction of photosynthetically active radiation (–)</td>
<td>0.1</td>
<td>1</td>
</tr>
<tr>
<td>31</td>
<td>Root fraction, layer 1 (–)</td>
<td>0.25</td>
<td>0.45</td>
</tr>
<tr>
<td>32</td>
<td>Root fraction, layer 2 (–)</td>
<td>0.45</td>
<td>0.65</td>
</tr>
<tr>
<td>33</td>
<td>Monthly LAI (–)</td>
<td>0.65</td>
<td>1.35</td>
</tr>
<tr>
<td>34</td>
<td>Monthly albedo (–)</td>
<td>0.1</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td><strong>Overstory – trees</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>Fractional coverage (–)</td>
<td>0.287</td>
<td>0.575</td>
</tr>
<tr>
<td>36</td>
<td>Trunk space (–)</td>
<td>0.37</td>
<td>0.79</td>
</tr>
<tr>
<td>37</td>
<td>Aerodynamic attenuation (–)</td>
<td>0.3</td>
<td>4.25</td>
</tr>
<tr>
<td>38</td>
<td>Radiation attenuation (–)</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>39</td>
<td>Maximum snow interception capacity (–)</td>
<td>0.05</td>
<td>0.2</td>
</tr>
<tr>
<td>40</td>
<td>Mass release drip ratio (–)</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>41</td>
<td>Snow interception efficiency (–)</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>42</td>
<td>Height (m)</td>
<td>2</td>
<td>7.5</td>
</tr>
<tr>
<td>43</td>
<td>Maximum resistance (s m⁻¹)</td>
<td>500</td>
<td>3000</td>
</tr>
<tr>
<td>44</td>
<td>Minimum resistance (s m⁻¹)</td>
<td>150</td>
<td>300</td>
</tr>
<tr>
<td>45</td>
<td>Moisture threshold (–)</td>
<td>0.115</td>
<td>0.165</td>
</tr>
<tr>
<td>46</td>
<td>Vapor pressure deficit (Pa)</td>
<td>200</td>
<td>4000</td>
</tr>
<tr>
<td>47</td>
<td>Rpc (–)</td>
<td>0.1</td>
<td>1</td>
</tr>
<tr>
<td>48</td>
<td>Root fraction, layer 1 (–)</td>
<td>0.65</td>
<td>0.15</td>
</tr>
<tr>
<td>49</td>
<td>Root fraction, layer 2 (–)</td>
<td>0.75</td>
<td>0.25</td>
</tr>
<tr>
<td>50</td>
<td>Monthly LAI (–)</td>
<td>0.5</td>
<td>1.5</td>
</tr>
<tr>
<td>51</td>
<td>Monthly albedo (–)</td>
<td>0.06</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td><strong>Overstory – tall trees</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>Fractional coverage (–)</td>
<td>0.322</td>
<td>0.594</td>
</tr>
<tr>
<td>53</td>
<td>Height (m)</td>
<td>10</td>
<td>14.7</td>
</tr>
</tbody>
</table>

2.2.2 Meteorological forcing data

The model was executed at a time step of 3 h to effectively simulate snowmelt while balancing computational cost. Model forcing data include air temperature, precipitation (which is partitioned between rain and snow using two temperature thresholds set by parameter values within the model framework), relative humidity, wind speed, and solar and longwave radiation (Fig. 1c). All but solar and longwave radiation were measured continuously at two SNOTEL sites within Tenderfoot Creek Experimental Forest, with one site located at low elevation and another site located at high elevation. Air temperature was distributed across the basin using a lapse rate calculated at each model time step between the two SNOTEL sites; all other meteorological information from SNOTEL sites was distributed using an inverse distance weighting approach contained within the model. Solar radiation data at the SNOTEL sites were sparse and discontinuous; instead, solar radiation measured at a FLUX tower was scaled to SNOTEL site locations using topographic position (Kelleher et al., 2015). Within the model, solar radiation is distributed across the catchment using monthly averaged shading maps for each model time step across a 24 h
period. These shading maps incorporate both the effects of topographic shading, diel variability, and monthly variability in sun position and angle. Uniform shading maps were averaged for the months of November, December, and January, as spatially distributed shading maps produced simulations of climate and hydrology that exceeded realistic ranges for the model (Kelleher et al., 2015). As these months are primarily periods of accumulation and very little to no melt, we do not expect this to affect model results. Longwave radiation was calculated using the Stefan–Boltzmann equation (Dingman, 2001).

2.2.3 Model initialization and setup

For each parameter sample, the model was run for a 6-month warm-up period from 1 April 2006 to 30 September 2006 and an analysis period of 1 October 2006 to 1 October 2008. We calibrate to the 2008 water year (1 October 2007–1 October 2008) but demonstrate results for the 2007 water year (1 October 2006–1 October 2007) to show the value of our approach across a longer period. While these periods are short, they encompass both a wetter (2008) and a drier (2007) year across the known climatic record at this site. The model was initialized using observations from the two SNOTEL stations. To test that the warm-up period was sufficiently long for the impact of initial conditions on simulations to dissipate, we initialized nine parameter sets from varying initial conditions (results shown in Appendix A). Similar to other distributed model studies (e.g., Melsen et al., 2016), we found that model simulations diverged quickly, suggesting that 6 months should be sufficiently long for the influences of parameter values on model simulations and predictions to emerge. This 6-month period of initialization was not used in any of the calibration or validation processes of the model.

3 A framework for constraining environmental simulations and predictions

3.1 Assessing model performance

There is a wealth of information and approaches researchers use to assess model performance and presence of equifinality, often by quantifying how well simulations match observations, aggregated catchment or regional data, and/or perceptions of system functioning (Bennett et al., 2013). Across the modeling and prediction in ungauged catchment literature, we have found that constraints on model-derived hydrologic behavior generally fall into three general categories:

1. regional signatures of similarity (e.g., Yadav et al., 2007; Bloeschl et al., 2013; Hrachowitz et al., 2014), typically applied to regionalize hydrologic models for streamflow prediction in ungauged catchments;
2. objective functions or error metrics (e.g., Wagener et al., 2001; Gupta et al., 2008; van Werkhoven et al., 2008; Pfannerstill et al., 2014; Shafii and Tolson, 2015), which measure how well simulations match observations; and
3. measures of internal catchment behavior, which describe how well simulations match either observations or perceptions of the spatiotemporal variability of hydrologic behavior (e.g., Franks et al., 1998; Lamb et al., 1998; Grayson et al., 2002; Wealands et al., 2005; Kurás et al., 2011; Koch et al., 2016).

Whether within an uncertainty framework or applied to calibrating a single parameter set, most studies typically use one of the three types of constraints outlined above to obtain a “best” parameter set; few have made use of more than one constraint type, though many studies have shown the value of multi-objective approaches (Yapo et al., 1998; Gupta et al., 1998; van Werkhoven et al., 2009). Recently, a number of studies have advanced approaches for assessing alternative information sources (i.e., model realism via expert knowledge) and constraining both parameter values and model simulations in pursuit of reducing equifinality (Seibert and McDonnell, 2002; Hrachowitz et al., 2014; Gharari et al., 2014; Silvestro et al., 2015). In spite of these many examples, a general framework for how to systematically constrain environmental simulations and parameter inference is still needed.

Building on the uncertainty literature across distributed model applications, we recommend constraining environmental simulations following a hierarchy of metrics/signatures and corresponding constraints (McGlynn et al., 2013). This framework builds from sources of information that should be widely available and have low cost to obtain, to information that may not be available everywhere and that may be more expensive to obtain. We present this framework in Fig. 2, outlining information sources that range from regional signatures, to error metrics, to spatial patterns informed by observations and expert judgment. In this study, we use this framework as a path to evaluate distributed model equifinality during model calibration.

These signatures/error metrics are summarized as follows:

**Regional Signatures** Regional signatures, increasingly used in hydrological model applications (e.g., Yadav et al., 2007; Bloeschl et al., 2013; Hrachowitz et al., 2014), constrain key annual dynamics. Ranges for a region may be informed by global or regional data sources and publications (e.g., Krug et al., 1990; Milly, 1994; Church et al., 1995; Zomer et al., 2007, 2008).

**Local signatures** In the absence of measurements, field researchers often have a broad sense of feasible and infeasible ranges for different types (e.g., evapotranspiration, water table height/soil saturation, snow water equivalent) of annual or seasonal hydrologic behavior. These constraints are best applied to average catchment conditions, with a goal to remove simulations that are too
In general, error metrics are calculated by comparing time series of observational records (most commonly, streamflow, but can include snow water equivalent or snow depth (e.g., Whitaker et al., 2003), road ditch flow (e.g., Surfleet et al., 2010), and soil moisture (e.g., Cuo et al., 2006) to model simulations. We recommend comparing observed and simulated time series data in two ways (e.g., van Werkhoven et al., 2009; Hrachowitz et al., 2014): as statistical metrics, which measure model performance with respect to the entire time series, e.g., root mean squared error (RMSE), Nash–Sutcliffe efficiency coefficient (NSE; Nash and Sutcliffe, 1970), and dynamic metrics, which measure model performance with respect to different periods or types of hydrologic behavior, e.g., the baseflow index, the slope of the flow duration curve (Wagen et al., 2001; Gupta et al., 2008; Pfannerstill et al., 2014; Shafii and Tolson, 2015). Performance across statistical metrics is typically judged with respect to a threshold value, e.g., NSE greater than 0.8, or some threshold percentage, e.g., top 10% of RMSE values (e.g., Moriasi et al., 2007; Harmel et al., 2014). Dynamic metrics may expand assessment of hydrologic behavior, as existing work has shown that there is information contained not only in different types of data but also in different periods for an observational time series (Wagener et al., 2001; Gupta et al., 2008; Pfannerstill et al., 2014; Shafii and Tolson, 2015).

Internal catchment behavior Aggregated spatial predictions (if values are unknown) and simulations (if values are known) may be used to assess the realism of predicted internal catchment behavior (Grayson et al., 2002; Wealands et al., 2005; Kuraš et al., 2011; Koch et al., 2016; Fang et al., 2016). There are many examples of distributed models that have evaluated internal catchment behavior using what Grayson et al. (2002) refer to as the “many points” approach – comparing model simulations to (often, time series) observations at several locations throughout the catchment for a given process of interest (e.g., Thyer et al., 2004; Kuraš et al., 2011). However, in the absence of internal observations, we suggest that internal simulations may still be incorporated into evaluations of distributed model behavior and parameter estimation. First, we suggest researchers apply a simple “reality check”: are the predictions possible? This step may help to identify runs that predict behavior outside of the realm of possibility. Secondarily, we suggest that researchers should consider the parameter influences that generate differences in the spatial representations, to evaluate and either accept or reject parameter sets as more realistic or less realistic representations of the system. This type of evaluation will benefit from involving experimentalists that have experience in a particular catchment. While this evaluation is often qualitative, it will rely on expert knowledge, previous experimental catchment studies, as well as a perceptual catchment model.

3.2 Approach

In this study, we apply the framework demonstrated in Fig. 2 within a Monte Carlo based uncertainty analysis, interrogating relationships between model parameter values, model simulations and predictions, and model metrics that summarize model performance and behavior as catchment-wide signatures and error. We applied Monte Carlo sampling to a priori constrained parameter ranges for 53 independent parameters associated with basin-wide properties, soil properties, and vegetation properties to produce 10 000 parameter samples (Table 1; Appendix A). The model was executed for each parameter sample, and model simulations were used to calculate 11 different metrics for each model run. We acknowledge that 10 000 parameter sets will not fully explore the entire parameter space for a 53-parameter model. However, our approach here serves as a demonstrative exercise to provide an example of how a modeler could approach this type of reduction via multiple constraints. As such, having a greater number of parameter sets should not influence our general conclusions.
Table 2. Signatures and error metrics used in the analysis. The table contains abbreviations for each signature/error metric, units, the general equation used for calculation, and the constraints chosen to separate behavioral and non-behavioral runs. Abbreviations in equations refer to time step $t$, total number of time steps $m$, observed values $O$, and simulated value $S$.

<table>
<thead>
<tr>
<th>Signature/metric</th>
<th>Abbrev.</th>
<th>Data</th>
<th>Units</th>
<th>Equation</th>
<th>Constraints</th>
</tr>
</thead>
<tbody>
<tr>
<td>Runoff ratio</td>
<td>RR</td>
<td>$Q$, $P$</td>
<td>(%)</td>
<td>$100 \cdot \frac{\sum_{t=1}^{m} Q_{t}}{\sum_{t=1}^{m} P_{t}}$</td>
<td>RR &lt; 20%</td>
</tr>
<tr>
<td>Aridity index</td>
<td>AI</td>
<td>PE, $P$</td>
<td>(%)</td>
<td>$100 \cdot \frac{\sum_{t=1}^{m} \frac{P_{t}}{PE}}{\sum_{t=1}^{m} \frac{P_{t}}{PE}}$</td>
<td>AI &lt; 1.206</td>
</tr>
<tr>
<td>Annual evapotranspiration</td>
<td>ET</td>
<td>ET</td>
<td>(mm)</td>
<td>$\sum_{t=1}^{m} ET_{t}$</td>
<td>ET &lt; 300 mm</td>
</tr>
<tr>
<td>Average annual water table depth</td>
<td>WT</td>
<td>WT</td>
<td>(m)</td>
<td>$\frac{1}{m} \sum_{t=1}^{m} WT_{t}$</td>
<td>WT &lt; 0.5 m</td>
</tr>
<tr>
<td>Nash Sutcliffe efficiency coefficient</td>
<td>NSE$_Q$</td>
<td>$Q$</td>
<td>(%)</td>
<td>$1 - \frac{\sum_{t=1}^{m} (O_{t} - S_{t})^{2}}{\sum_{t=1}^{m} (O_{t} - \frac{1}{m} \sum_{t=1}^{m} O_{t})^{2}}$</td>
<td>NSE$_Q$ &gt; 0.6</td>
</tr>
<tr>
<td>SWE (%)</td>
<td>NSE$_S$</td>
<td>SWE</td>
<td>(%)</td>
<td>$100 \cdot \frac{\sum_{t=1}^{m} S_{t}}{\sum_{t=1}^{m} O_{t}}$</td>
<td>NSE$_S$ &gt; 0.8</td>
</tr>
</tbody>
</table>

Following the framework in Fig. 2, Table 2 lists all metrics used to assess catchment behavior, the equations used to calculate each metric, the associated constraints we applied in this study, and the source of information for defining each constraint. Due to the relatively short time period for which all meteorological measurements were present within the catchment and at both SNOTEL sites, we assessed model calibration across a single year of data corresponding to the 2008 water year (WY) for one discharge location (Stringer Creek, LSC). All metrics are calculated with respect to this period and are annual metrics unless otherwise stated. Metrics related to discharge are calculated using observed Stringer Creek streamflow, and metrics using data beyond streamflow represent average conditions for the Tenderfoot Creek watershed. To further demonstrate the value of this approach, we also include model performance for two other discharge locations, a smaller catchment (middle Stringer Creek, MSC) and a larger catchment (Tenderfoot Creek, LTC) for the 2008 water year. We additionally display simulations for the 2007 water year.

The framework begins with an evaluation of whether model simulations match signatures of regional behavior, evaluated via the runoff ratio and aridity index. While the latter is calculated from potential evapotranspiration (PET), which may be supplied as input for some models, PET varies with parameters in DHSVM. Signatures were constrained using existing datasets and regional reports/publications (e.g., Sankarasubramanian and Vogel, 2003), including the CGIAR-CSI Global Aridity and Global-PET database (Zomer et al., 2007, 2008), which was used to limit the aridity index (AI) to values obtained within a 50 km radius of Tenderfoot Creek. Constraints on AI from this dataset represent long-term average conditions for the region, and are used to broadly eliminate simulations outside of the realm of regionally realistic estimates.

We additionally incorporated local signatures of cumulative annual evapotranspiration and average annual water table depth (WTD), applied to catchment-wide averages. Experimentalists can often recommend basic ranges for these values from fieldwork or from numerous visits or observations of other hydrologic processes across the catchment. ET was constrained based on calculations from Mitchell et al. (2015), from observations at a meteorological tower located within the catchment. We limited annual water table depths based on previous findings that the water table is only active for a small portion of the year during snowmelt, with most of the catchment experiencing very little response (Jensco et al., 2009; Jensco and McGlynn, 2011).

Error metrics were calculated for observed streamflow measurements as well as snow water equivalent (SWE) measurements at two SNOTEL stations within the larger Tenderfoot Creek catchment. While there are more observational data present within the catchment than are used to constrain the model, our goal in this application was to make use of observations that are likely to be available in many experimental catchments. Thus, we focus on metrics that assess streamflow and SWE. Continuous observations of snow water equivalent are available at both SNOTEL sites. Statistical fits for simulated and observed streamflow and SWE were assessed with the NSE. We additionally selected five metrics to describe dynamic streamflow and SWE behavior:

- runoff ratio error ($RRE_{Q}$), to capture the annual water balance;
4.1 How do single constraints influence model performance?

As can be seen in Fig. 3, constraining with a single metric yields good performance for matching observations of SWE (with respect to NSE and error in SWE volume). However, the ranges and first and third quantiles for all other metrics are well outside of constraints. For example, error in peak timing may exceed upwards of a month (30 days) and error in peak magnitude ±100 % when the model is constrained with just one metric.

4.2 How do constraints on model simulations influence the number of acceptable parameter sets?

The number of parameter sets that met constraints applied individually, hierarchically, and as a group are shown in Fig. 4. When applied hierarchically, a total of nine parameter sets met all metric constraints (Fig. 4a). Interestingly, statistical and regional constraints identified a similar number of behavioral sets. While combining regional and local metrics reduced the number of behavioral sets, combining statistical and dynamic constraints did not reduce the number of behavioral sets.

4.3 Are all metrics needed?

To evaluate whether all 11 metrics and corresponding constraints are needed to ensure behavioral consistency within this framework, we examined model behavior and performance across subsets of metrics (n = 2 to 10) as well as the level of redundancy across metrics (Appendix B). Employing only 2 or 3 metrics removes 69.7 and 74 %, respectively, of the sets removed when all 11 metrics are used, suggesting that 25 to 30 % of parameter sets inconsistent with model behavior across our framework would remain. By comparison, including eight metrics generally removes an average of 92.6 % of all non-behavioral parameter sets. Of the 165 combinations of eight different metrics evaluated in this analysis, 83 % included at least one metric in each of the different framework groups (regional, local, statistical, dynamic). The redundancy of each metric was also evaluated by comparing the non-behavioral sets identified via constraining one metric that would not be found by another (Appendix B). While we found that redundancy did occur in this type of framework, more than half of the simulations found to be non-behavioral by constraining one metric were considered behavioral by another metric across a majority of metric combinations.

4.4 How do single and multiple constraints impact the distributions of metrics of catchment behavior?

Initial distributions (I) of metric values across all 10 000 parameter sets include a wide range of behavior (Fig. 5). While these distributions can be narrowed by the addition
of one constraint, many distributions of behavioral performance were also narrowed by the addition of all constraints. The addition of all constraints favored higher values (~0.5 to 0.7) for the runoff ratio and lower values (~0.35 to 0.5) for the aridity index and average water table height (~−0.75 to −0.95). Performance improved slightly for both statistical metrics (NSE_Q and NSE_S) and narrowed ranges for both error in the runoff ratio (RRE_Q) and slope of the flow duration curve (SFDCE_Q). While behavioral sets both over- and underestimated RRE_Q and SFDCE_Q, the final behavioral sets only included simulations that underestimated peak streamflow alongside earlier-than-observed peak timing, while overpredicting SWE.

4.5 How do constraints impact predictions and simulations of catchment behavior?

Figure 6 displays the impact of different constraints on the range of model simulations for streamflow (LSC), snow water equivalent (Stringer Creek SNOTEL), and average water table depth for the 2007 and 2008 water years. Streamflow simulations for the calibration period (WY 2008) were best narrowed by the addition of statistical and dynamic constraints and poorly narrowed by the addition of regional and local constraints. Constraining behavior via regional and local metrics resulted in behavioral simulations that underestimated peaks and poorly matched timing of streamflow initiation and maximum response. While statistical and dynamic metrics both predict an earlier streamflow response on the rising limb of the hydrograph, simulations were well matched with behavior on the falling limb of the hydrograph. Dynamic metrics also better constrained streamflow simulations during baseflow, which tended to be overestimated by behavioral sets selected by statistical metrics.

Predictive ranges for SWE were similar across different constraints. All metrics had well constrained SWE behavior (shown for the Stringer Creek SNOTEL station), with consistent differences between simulations and observations regardless of metrics used to constrain behavior. SWE magnitude was overestimated for all behavioral sets, though the initiation of melt timing was similar, producing simulations that...
overestimated the time that the last of the snowpack melted during late May and early June.

In contrast to simulations of SWE, behavioral simulations of average water table depth varied widely depending on which metrics were used to constrain behavior. Behavioral sets identified via statistical metrics included a wide range of average water table behavior, suggesting that statistical metrics may poorly inform water table dynamics. Behavior constrained by local and regional metrics produced a pattern of average water table depth consistent with general understanding of water table behavior (Jensco et al., 2009; Jensco and McGlynn, 2011). As a metric for average water table depth was directly incorporated into the assessment of regional and local behavior, this result is expected. Dynamic metrics had the largest impact on narrowing WTD simulations, identifying behavioral simulations that were most consistent with our understanding of water table response across Tenderfoot Creek.

4.6 How well do model simulations match observations?

Calibration to the 2008 water year utilizing all metrics and corresponding constraints identified behavioral sets corresponding to hydrographs and SWE time series that reasonably matched observations (Fig. 7). Simulations of streamflow at the calibration site (LSC) for the 2008 water year simulated the double streamflow peaks and match periods of wetting up and drying down that occurred in the observational record, with some differences between simulations and observations. However, the timing of streamflow on the ris-
ing limb of the hydrograph and the first streamflow peak was poorly matched. SWE at both SNOTEL sites was overestimated in terms of magnitude and timing, though simulations generally matched observations. Melt timing generally approximated observed behavior at Onion Park but was slightly early at Stringer Creek, though again dynamics in both cases are largely matched, especially the timing of accumulation and melt from April through June.

To demonstrate the value of this approach with respect to other streamflow observation locations and for years beyond the calibration period, we also display fits to the 2008 water year for middle Stringer Creek and lower Tenderfoot Creek. Streamflow NSE at these two locations varied between 0.63 and 0.83 for LTC and between 0.47 and 0.74 for MSC. Like fits to Stringer Creek, timing and peak magnitude for streamflow at MSC and LTC were underpredicted with respect to observations, though simulations approximate behavior of observations on the rising and falling limbs, and accurately predict the presence of observed double streamflow peaks for the 2008 water year. Performance for other metrics for these two discharge locations is reported in Table B1.

Outside of the calibration period, fits for the 2007 water year for all locations were high (Fig. 8). NSE for the calibration site (LSC) varied between 0.65 and 0.82, while NSE for the other two streamflow locations varied between 0.64 and 0.91 at LTC and between 0.51 and 0.78 at MSC. While the water balance (RRE\textsubscript{D}) was well approximated for LSC, estimates were higher than observations for both MSC and LTC, though timing and magnitude of peak streamflow were well approximated. The values for all metrics are reported in Appendix B.

4.7 How do constraints impact parameter uncertainty?

To test whether multiple model constraints on hydrologic behavior were able to reduce parameter equifinality, we in-
vestigated the extent to which constraining model-predicted and simulated behavior narrowed ranges for parameter values, with results displayed in Fig. 8. While ranges remained wide for many parameters, ranges were greatly narrowed for several of the 53 model parameters. Narrowed ranges were observed especially for lateral conductivity (7) and its exponential decrease with depth (8), maximum (26) and minimum (27) understory stomatal resistance, and a scalar on overstory leaf area index (LAI) (50). We applied a two-sample Kolmogorov–Smirnov test to assess whether there was a statistically significant difference (p < 0.1) between the uninform prior distribution of parameter values (Table 1) and the distribution of parameter values for sets that met framework criteria. We found that all but five parameters exhibited statistically significant differences between their original and final value distributions. However, ranges were still wide for many parameter values.

4.8 Do assessments of catchment averages and observations produce consistent internal predictions of catchment behavior?

In this study, we do not directly compare simulations of water table depth to well observations. Instead, we sought to assess the variability across these simulations for three different time periods, to determine whether behavioral parameter sets yielded similar simulations, and to ask whether spatial diagnostics beyond time series observations should be incorporated into assessments of distributed model behavior. All predictions of water table depth are included in Fig. B3, with three predictions of water table depth displayed in Fig. 9. These three encompass the range of predictions from low to high water table depth for annual, snowmelt (1 May–31 July), and late summer dry-down (1 August–30 September) periods. Seasonally, water tables were simulated closer to the surface during snowmelt and closer to bedrock during late summer, with average behavior somewhere between these two extremes.

4.8.1 Are simulations of water table depth consistent in space and time?

Across simulated water table depths for nine behavioral parameter sets, differences in simulations were large in both space and time (Figs. 9, B3). At annual timescales, many parts of the catchment were predicted to have similar annual average behavior, with differences below 0.1 m for 38 % of all cells. Simulations across 13.3 % of the catchment differed by more than 0.2 m, 20 % of the modeled soil depth. These numbers were comparable for simulations during late summer, when the majority of the catchment was simulated to have similar behavior. Differences were greatest during the snowmelt period, exceeding simulated ranges of 0.2 m over more than 64 % of the catchment. The locations of the largest differences also varied across seasonal and annual periods.

Our results show that equifinality can produce vastly different simulations of internal catchment behavior.

4.8.2 Can simulations be used to further reduce equifinality?

Evaluating whether simulations of internal behavior of water table depth match perceptions and observations of catchment functioning may be done using simple metrics of spatial behavior. One such example is the presence of the water table at the land surface. Tenderfoot Creek field researchers suggest that there are few locations and few times where the water table should be at the land surface across the catchment (Jensco et al., 2009; Jensco and McGlynn, 2011). Based on these recommendations, we expect high water tables to be present across minimal areas (<5 %) of the catchment. To test the ability of a spatially distributed high water table metric to discern differences between behavioral sets, we calculated the percentage of cells across Stringer Creek where the water table was simulated to be within 0.05 m of the surface for the entire snowmelt period (May–July). Four different behavioral sets simulated high water tables over more than 5 % of catchment; at the extreme end, one set simulated high water tables over 24 % of the catchment. Five sets simulated high water tables over less than 5 % of the catchment, with simulations below this threshold across the entire catchment for four of these sets. We would conclude from this evaluation that these four sets produce spatial behavior that is consistent with experimental insight across Stringer Creek.

5 Discussion

Within this study, we develop a conceptual uncertainty analysis and framework for parameter uncertainty reduction with application to distributed modeling of headwater systems. We demonstrate how this framework can be applied with respect to a case study in Stringer Creek, a headwater catchment located in Montana. While there are a few examples of frameworks for model calibration (e.g., Refsgaard, 1997) as well as several studies that have evaluated the trade-offs between model complexity and predictive uncertainty, there remain few guidelines for specific use in distributed hydrologic models, which have their own challenges. In this study, we sought to create and test a framework that considered many of the above-discussed limitations and capitalized on the strengths of such complex models.

5.1 The value of suites of metrics in distributed model applications

Many distributed model applications use only one constraint to select behavioral parameter sets or to justify model performance. As we show in Fig. 3, model simulations that meet just one set of criteria may poorly match other catchment-wide basin behavior. Of particular interest is the impact of
The largest number of non-behavioral sets was identified by errors in peak streamflow magnitude. While there were several sets that had very small errors in peak streamflow magnitude (Fig. 5), many overestimated the annual water balance, and were therefore identified as non-behavioral. While peak timing was underestimated by the addition of all 11 criteria, errors in the runoff ratio and the slope of the flow duration curve were narrowed to ranges of lower errors, and average water table depth was constrained to more representative ranges when all 11 criteria were added (Fig. 6, 7).

In the absence of time series observations, our results show that regional and local metrics could be used to narrow predictions but may poorly match hydrologic behavior for some years (Fig. 6). While statistical or dynamic metrics may inform predictive uncertainty for streamflow in similar ways, we found that they impacted simulations of other key catchment behaviors differently. Statistical constraints poorly constrained simulations of water table depth (Fig. 6). In contrast, dynamic constraints, applied only to errors in streamflow and SWE time series, yielded behavioral simulations that closely matched expectations of average water table behavior across the catchment (Jencso et al., 2009; Jencso and McGlynn, 2011). Our results demonstrate that suites of metrics related to hydrologic behavior may inform simulations of other hydrologic processes (i.e., average water table depth). In contrast, we found statistical metrics to carry little infor-
mation regarding other types of simulated or predicted behavior (Figs. 3, 6). To our knowledge, most uncertainty analysis studies with semi-distributed or distributed models typically favor statistical constraints over dynamic constraints (e.g., Shields and Tague, 2012; Safeeq and Fares, 2012). It is unclear whether this result is generalizable to other catchment applications, though we expect the impact of dynamic metrics with regards to discerning catchment behavior to only increase for a greater number of hydrologic events.

5.2 Benchmarking simulations against other model applications to Stringer Creek

Altogether, application of a framework for distributed modeling yields model simulations that match streamflow at the calibration site and that generally match patterns of SWE accumulation and melt (Fig. 7). In general, while results for the 2008 water year bracket observations, results for the 2008 water year are underpredicted with respect to peak behavior and timing. Given that snowmelt drives both rising limb and peak hydrograph response, uncertainty related to solar radiation or air temperature forcings may be driving the difference in rising limb behavior that we found in our simulations.

Simulations compare favorably to other conceptual and lumped model applications to Stringer Creek. Nippgen et al. (2015) developed a parsimonious distributed model, with application to Stringer Creek yielding NSE values of 0.8 for WY 2007 and 0.94 for WY 2008. Smith et al. (2013) developed the catchment connectivity model (CCM), a conceptual hydrologic model that predicts streamflow based on relationships between terrain, connectivity between the hillslope and stream found in Jensco et al. (2009), and the duration of flow. Simulation of Stringer Creek with CCM achieved similar levels of fit to our study (NSE of 0.81 for Box–Cox transformed streamflow; Smith et al., 2013), as did model simulations of LTC (NSE of 0.903) and MSC (NSE of 0.856; Smith et al., 2016). Simulations of streamflow by Nippgen et al. (2015) and Smith et al. (2016) also underestimated peak streamflow for the 2008 water year, suggesting that uncertainty in the meteorological forcing data may be responsible for poor performance during the calibration period for snowmelt. Ahl et al. (2008) modeled Tenderfoot Creek streamflow using the soil and water assessment tool (SWAT) for the period 1995–2002 and achieved an average NSE of 0.86 during calibration and 0.76 during validation. Our best fit by NSE to Stringer Creek was 0.79 for the 2008 water year and 0.82 for the 2007 water year, while best fits for Tenderfoot Creek were 0.83 for the 2008 water year and 0.91 for the 2007 water year. Thus, the level of fit we achieved through this investigation was similar to other model applications to this catchment. Moreover, we ensure that the selected model runs also match key catchment-wide behavior as well as dynamical streamflow behavior through the use of additional metrics. We acknowledge, however, that sources of uncertainty beyond those considered by our study may still be driving differences between simulations and observations. Possible sources of error in simulations of both SWE and peak magnitude and timing errors may be related to uncertainty in the model framework, model forcing data, and observations used to judge performance. Future work modeling these catchments will seek to address these other sources of uncertainty alongside uncertainty in parameter values.

The three models to which we compare our results demonstrate a range of model frameworks that can be used to evaluate model behavior: conceptual (Smith et al., 2013), lumped (Ahl et al., 2008), and distributed without physically based parameters (Nippgen et al., 2015). As is shown in this study, all of these models are able to accurately simulate the hydrograph for this catchment. The primary trade-offs across these models include requirements for inputs and parameters alongside computational requirements, which are inversely related to the complexity of simulated behavior that can be produced from each of these models. While any of these approaches may be used to simulate streamflow, each will enable researchers to answer different questions related to hypotheses about catchment functioning, the use of field information to inform model parameter constraints, and predictions of spatiotemporal hydrologic processes. Finally, these contrasting models also illustrate the differences between a model like DHSVM that may be applied to many different catchments versus the models introduced by Smith et al. (2013) and Nippgen et al. (2015), in which the model framework and structural equations were developed only for this catchment. In this study, we specifically evaluate the application of physically based, distributed models to simulate experimental catchments, though we encourage researchers to select the right tool, and therefore the appropriate model, for a given study objective.

5.3 Parameter uniqueness and equifinality

We are often interested in how constraints influence predictive uncertainty of hydrologic behavior. Thus, the logical follow-up question is whether these constraints help to narrow values for model parameters. Across the 53 model parameters included in our analysis, constraints had a minimal impact on narrowing parameter ranges for most parameters. However, we did detect a subset of soil and vegetation parameters that were reasonably narrowed from their original ranges. This suggests that equifinality is still present but that uncertainty analysis may also narrow parameter uncertainty. In a similar application of DHSVM to the Oak Creek catchment near Corvallis, OR, Surfleet et al. (2010) also found significant equifinality when using a similar approach to characterize uncertainty with respect to streamflow and road-ditch flow prediction. While we cannot formulate any strong conclusions regarding predictive uncertainty in parameter values given our limited sampling of the parameter space, we assert that equifinality may overwhelm the ability to extract much information regarding parameter values in
complex distributed model applications. However, equifinal-
ity with respect to catchment average conditions may man-
ifest in variable predictions of internal catchment behavior 
(e.g., Fig. 9). Thus, evaluating spatial predictions may be one 
of the few practical approaches to reducing this equifinality.

5.4 Interpretation of internal behavior

Both field investigations and modeling at Tenderfoot Creek 
have focused on observation and prediction of water table 
response with the goal of improving our interpretation of 
streamflow generation and the connection between rainfall 
and runoff (Jencso et al., 2009; Jencso and McGlynn, 2011; 
Nippgen et al., 2015). Thus, we chose to evaluate simul-
tions of water table depth for our case study. We found that 
parameter sets that were equifinal with respect to streamflow 
and SWE (Figs. 7, 8) produced vastly different annual and 
seasonal simulations of water table depth (Fig. 9). Moreover, 
the locations where simulated differences were largest and 
smallest across the catchment also varied with time. By as-
sessing the fraction of the catchment simulated to be at or 
near the surface across the snowmelt period, we found that 
four of the nine sets produced near-surface water tables over 
a larger area than previous work suggests is likely. In this 
sense, internal behavior can be used to identify simulations 
that do not match perceptions or direct observations of catch-
ment behavior beyond comparison with time series. This ap-
proach enables one to move beyond just matching to a few 
points with observations, encouraging modelers to holisti-
cally evaluate performance and simulation of multiple hy-
drologic processes.

5.5 The case for adding spatial predictions

Although constraints on the hydrograph and other hydrologic 
behavior may ultimately match observations, we are still 
faced with the likelihood that parameter equifinality may not 
be eliminated by matching predictions to a few time series 
observations. Aggregating model predictions of spatial pat-
terns is time intensive but can be highly informative. There-
fore, we recommend mapping these patterns once regional 
values and/or observations have already been used to narrow 
the parameter space.

Distributed model predictions of internal behavior are 
often performed by matching model predictions of differ-
ent catchment variables (e.g., snowmelt, Thyer et al., 2004; 
groundwater table dynamics, Kurás et al., 2011) to multi-
ple observations at discrete points across a given catchment, 
though several studies have also shown how patterns of hy-
drologic processes may be directly incorporated into evalu-
ating and applying complex environmental models (Grayson 
et al., 2002; Wealands et al., 2005; Fang et al., 2016; Koch 
et al., 2016). In the absence of time series observations of 
other hydrologic processes, simulated patterns may be evalu-
ated using expert knowledge, understanding of process con-
trols across landscapes, and other information knowledge 
and experience about model output realism (e.g., Franks et 
al., 1998). In our example, nine parameter sets that matched 
the hydrograph and associated SNOTEL station SWE ob-
servations equally well (Figs. 5, 6, and 7) predicted very 
different patterns of internal catchment behavior (Fig. 9). 
However, these same sets also exhibited divergent parame-
ter values for several different soil and vegetation properties 
(Fig. 8). This result exemplifies the additional information 
available to constrain behavioral sets as well as the potential 
for catchment-scale predictive uncertainty driven by param-
eter equifinality. Sets that met framework criteria simulated 
very different patterns of water table depth across both an-
nual and seasonal periods (Fig. 9). Differences typically were 
between 0 and 0.2 across all simulations and all catchment 
cells, but were especially large (> 0.5 m) over small fractions 
of the catchment. If spatial predictions were not used to limit 
the parameter sets, any one of the nine sets has potential to 
propagate predictions for future land use or climate change 
scenarios that would lead to vastly different expectations of 
water table presence across the landscape. We conclude that 
equifinal sets may generate very different simulations of in-
ternal catchment behavior and therefore recommend that spa-
tial simulations should be incorporated into assessments of 
distributed catchment behavior (Fig. 2). While we chose to 
highlight only one spatial diagnostic in this paper, we ad-
vocate the inclusion of multiple diagnostics related to key 
catchment storages (e.g., water table depth) and fluxes (e.g., 
evapotranspiration).

5.6 Future uncertainty analysis of distributed 
catchment models

Applying this framework to other catchments will require re-
searchers to select a set of metrics for assessing model perfor-
mance, emphasizing both temporal and spatial metrics that 
may help to ensure appropriate representation of key catch-
ment processes (e.g., Yilmaz et al., 2008). Evaluating any of 
these aggregated signatures, metrics, and spatial diagnostics 
should be based on a strong conceptual understanding of the 
catchment and how processes that govern the water balance 
change temporally and spatially at multiple scales. Such eval-
uations may especially benefit from joint evaluation by mod-
elers and experimentalists (Seibert and McDonnell, 2002).

The approach we recommend in this paper builds on and 
compliments several recent studies that have sought to im-
prove process consistency across models of varying com-
plexity and within distributed hydrologic models. Many re-
cent studies have shown that, despite their weaknesses, dis-
tributed models typically outperform conceptual models with 
respect to reproducing signatures (e.g., Hrachowitz et al., 
2014) and matching hydrograph dynamics (e.g., Euser et al., 
2015). Thus, the new objective we face is how to improve our 
approaches to distributed modeling, ensuring model realism 
while minimizing uncertainty. Work by several researchers
has evaluated methods for the spatial distribution of parameter values, to ensure process consistency across catchment scales (Euser et al., 2015; Fenicia et al., 2016; Nijzink et al., 2016). Others have sought to incorporate expert knowledge to limit the feasible parameter space (Gharari et al., 2014; Nijzink et al., 2016). In this vein, the choice of model structure may also offer another opportunity to reduce equifinality (Clark et al., 2008; Pokhrel et al., 2008; Samaniego et al., 2010; Rakovec et al., 2016). In particular, the extensive body of literature on parameter regularization may offer a pathway for maintaining spatial complexity and consistency while reducing the number of free model parameters (Hundecha and Bardossy, 2004; Hundecha et al., 2008; Pohkrel et al., 2008; Samaniego et al., 2010; Rakovec et al., 2016). Alternatively, there is also a body of work that treats the model framework itself as a form of uncertainty, testing different model structures as hypotheses for how a catchment may function (Clark et al., 2008, 2011; Fenicia et al., 2011; Hrachowitz et al., 2014). This approach may also provide an alternative to predicting hydrology via a model with fewer parameters than the distributed application shown here, with a model structure that incorporates the level of detail mandated by the complexities of the catchment (e.g., Zehe et al., 2014; Euser et al., 2015). As encouraged by Beven (2002), to best represent catchment behavior, we may need to not only focus on model parameters but also the model structure in terms of how this reflects the physical landscape. However, as shown by Surfleet et al. (2010) in an application of DHSVM to a series of small catchment areas in Oregon, equifinality may still overwhelm our ability to draw meaningful conclusions from distributed data, especially at small headwater scales. Taken together, these studies suggest that equifinality is likely to still limit application of distributed models, but that prudently evaluating how this equifinality may impact uncertainty in predictions and simulations alongside parameter values may enable more careful use of distributed models. Similarly, these studies also suggest that incorporating constraints within distributed model frameworks based on expert knowledge and alternative data sources, whether applied to model output, model setup, or model parameter values, may ensure more holistic process representation across a given catchment.

There are few studies that have sought to characterize equifinality and uncertainty for physically based, distributed model applications, but the number of distributed model applications that incorporate uncertainty is growing. Of those that exist, most have focused on characterizing predictive uncertainty in terms of uncertainty in parameter values (e.g., Cuo et al., 2011; Shields and Tague, 2012; Tague et al., 2013), or in terms of model framework uncertainty, by modifying the model formulation to match multiple experimental observations throughout the critical zone (e.g., Thyer et al., 2004). These studies exemplify the common need to consider uncertainty when predicting environmental behavior with complex, multi-parameter models. Our work suggests this will only provide the modeler with a better understanding of the catchment but also of the model in question. Altogether, research with distributed models and our own analysis does critique some of the challenges associated with distributed model application but also highlights the value of distributed models for hydrological predictions (Hrachowitz et al., 2014; Faticchi et al., 2016).

Ultimately, our ability to resolve issues with equifinality and identify appropriate parameter sets in space and time is challenged, as it was in this study, by the computational demand of complex models. Executing model predictions for the relatively short period of time investigated in this study across 10000 parameter samples required thousands of computing hours (and even longer periods if the modeler retains or “saves” spatial predictions across the catchment). While distributed, physically based models like DHSVM may have the ability to resolve predictions of hydrologic processes through space and time, we do not yet have effective, computationally inexpensive approaches for evaluating and representing uncertainties in these types of applications. In order to put these types of models to the test, we need better parameter sampling strategies (e.g., Rakovec et al., 2014; Jefferson et al., 2015) and alternative approaches to those we use for conceptual models, where executing a model many times is not a challenge or limit on analysis. This may come in the form of new methods, or alternatively, approaches that evaluate model adequacy via frameworks for computationally frugal analysis (Hill et al., 2016). While quantifying or limiting equifinality may always be a challenge for physically based, distributed catchment models, we likely will need to reframe our approaches for evaluating the uncertainties associated with complex model applications. This challenge may be best addressed by encouraging interaction across the conceptual modeling community and the fully, distributed, physically based modeling community, in order to address broad issues related to uncertainty and equifinality that, it can be argued, plague all models of any complexity (Hrachowitz and Clark, 2017).

6 Conclusions

Distributed, complex models are powerful tools that enable exploration of spatial and temporal simulations and future scenarios of alteration. While beneficial, their complex nature and subsequent potential for equifinality calls into question the typical process researchers use to achieve a representative set of parameters for a given application. We performed a modeling study for a headwater catchment, Stringer Creek, located in Tenderfoot Creek Experimental Forest in central Montana, and evaluated simulations using observational records, expert insight from Tenderfoot Creek researchers and existing publications, and regional datasets. In this application, we demonstrate a method to evaluate how constraining model predictions via hydrologic signatures, model error,
and process insight impacts predictive and parameter uncertainty, the size of the parameter set space, and potential for equifinality. Constraints include those that have potential to be available everywhere, based on both regional datasets and local knowledge, and those that are based on observational records of hydrologic behavior. We also include evaluation of spatial patterns of model predictions, to evaluate how parameter sets that match point observations predict storages and fluxes across the landscape.

Across all types of metrics and constraints, applied either hierarchically or in smaller subsets, we found dynamic constraints on annual, seasonal, and event behavior to be most important for reducing predictive uncertainty and selecting behavioral parameter sets. This suggests that researchers should use care when utilizing only statistical metrics to judge model performance or to select behavioral parameter sets, as for our application we found many model runs that had high statistical metric performance poorly matched dynamical hydrologic behavior. Despite the large reduction resulting from applying all constraints hierarchically, nine parameter sets met all criteria. Thus, we expect that there is likely a point at which observational records and regional datasets may no longer be able to reduce parameter sets and subsequent equifinality.

It is worth noting that parameter set selection for distributed catchment models is often done without considering whether the predicted internal behavior of the catchment is even within the realm of reality for a site. Here, we recommend the evaluation of internal catchment behavior as a final diagnostic to arrive at a subset of parameter sets that represent time series observations at a few locations as well as internal catchment behavior. While evaluating average catchment behavior and time series observations can be helpful, these types of behavior can often mask spatial variability of simulations across the year. Our evaluation of spatially distributed annual and seasonal water table depth revealed somewhat consistent average behavior but considerably variable spatial behavior.

Overall, this approach relies on a fundamental understanding of the hydrology that governs a given area, and is only improved by adding qualitative experimental insight. Transferring our approach to other locations creates the opportunity for a close interaction between experimentalists and modelers (e.g., Seibert and McDonnell, 2002), given that the value of a specific observations or insights to condition hydrologic models will vary widely. More than any single approach, we are advocating increased evaluation of distributed catchment models as a step towards improved representation and informed use, to ensure that spatiotemporal questions are resolved with spatiotemporally vetted answers.

Code availability. DHSVM model code is freely available at http://www.hydro.washington.edu/Lettenmaier/Models/DHSVM/code.shtml.

Appendix A: Model initialization and parameterization

Figure A1 and Table A1 display information used to inform the model framework and setup. Figure A1 displays the impact of three different hypothetical initial conditions on model predictions and simulations across nine different parameter sets. The nine parameter sets shown in Fig. A1 are the same sets that meet all criteria across the imposed framework. This analysis was performed to determine when the impact of initial conditions dissipated, to justify that a 6-month warm-up period may be sufficiently long enough for the paper application.

Figure A1.
Figure A1. Predictions of several watershed-averaged water balance components during model initialization (1 April 2006 through 30 September 2006) for nine equifinal parameter sets. Color indicates three different values for initialization, with initial values for water balance components corresponding to the value shown on 1 April. (a) contains parameter sets 1 through 5 and (b) contains parameter sets 6 through 9.
Appendix B: Additional reporting of model results and spatial pattern predictions

To demonstrate additional analyses that were undertaken within the uncertainty analysis framework, we have included an assessment of redundancy, additional reporting of errors for all gauge locations, and predictions of water table depth for all equifinal parameter sets. Our analysis included 11 different metrics. To assess whether these metrics provide unique information to the uncertainty analysis, we tested the information contained in different combinations of metrics (Fig. B1). Figure B1 displays the number of parameter sets that would be removed by all different combinations of metrics, comparing these combinations for different subsets of metrics (two metrics, three metrics, etc.). The number of non-behavioral model runs identified by these different combinations is shown as distributions for each subset of metrics (two metrics, three metrics, etc.), and visualized as a percentage of non-behavioral runs as compared to the number of non-behavioral model runs identified by the full 11 criteria.

We also benchmarked the redundancy of metrics and their corresponding constraints within the framework and application to Stringer Creek (Fig. B2). Figure B2 shows the percentage of sets removed by one metric (shown in the x axis) that would not be removed by another metric (shown on the y axis). We found that one metric, error in peak flow magnitude, tended to remove non-behavioral sets identified by other metrics. Similarly, we found that our statistical metric (NSE) may be more informative than our dynamic metric (VOL, error in SWE volume) for matching SWE time series and identifying non-behavioral simulations. In spite of these redundancies, many metrics do provide unique information, demonstrating the value of a multi-objective approach to uncertainty reduction.

Our analysis was applied to streamflow observed at the outlet of Stringer Creek (LSC) during the 2008 water year, but metrics were extracted for two additional gauge locations (lower Tenderfoot Creek, LTC; middle Stringer Creek; MSC) and for 2 years of complete simulation. Table B1 reports error metrics for the three gaging sites for both the 2007 and 2008 water years for nine behavioral parameter sets. Finally, we also include all water table depth predictions for nine equifinal parameter sets included in our analysis in Fig. B3. Figure 9 displays a subset of predictions.

Figure B1. This figure compares the number of non-behavioral sets removed by subsets of different metrics benchmarked by the number of non-behavioral sets removed by all 11 criteria. Distributions for a given subset of metrics (n = 2 to n = 10) are shown for all possible combinations of metrics.
Figure B2. Redundancy across different metrics. The percentage of non-behavioral sets removed by criteria are applied to one metric (x axis) that is not removed by another (y axis).

Figure B3. The impact of removing a single constraint on the number of final behavioral parameter sets, organized by the constraint that was removed from the analysis.
C. Kelleher et al.: Characterizing and reducing distributed model equnaility

Figure B4. Predictions of annual, May–July, and August–September water table depths. Color indicates average depth across a given period between bedrock (−1 m) and the surface (0 m). Columns display predictions across nine parameter sets identified by the framework applied to Stringer Creek hydrology, with a subset of these (sets 1, 5, and 9) shown in Fig. 9.

Table B1. Error metrics for the nine behavioral sets. Metric abbreviations correspond to metrics shown in Table 2 and Figs. 3 and 4.

<table>
<thead>
<tr>
<th>Water year 2008</th>
<th>LTC</th>
<th>MSC</th>
<th>LSC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NSE</td>
<td>RRE</td>
<td>SFDC</td>
</tr>
<tr>
<td>Run 1</td>
<td>0.80</td>
<td>14.36</td>
<td>32.20</td>
</tr>
<tr>
<td>Run 2</td>
<td>0.83</td>
<td>18.76</td>
<td>16.19</td>
</tr>
<tr>
<td>Run 3</td>
<td>0.80</td>
<td>34.63</td>
<td>17.85</td>
</tr>
<tr>
<td>Run 4</td>
<td>0.63</td>
<td>-7.38</td>
<td>-6.19</td>
</tr>
<tr>
<td>Run 5</td>
<td>0.76</td>
<td>3.76</td>
<td>-20.0</td>
</tr>
<tr>
<td>Run 6</td>
<td>0.72</td>
<td>0.87</td>
<td>-7.65</td>
</tr>
<tr>
<td>Run 7</td>
<td>0.79</td>
<td>-4.48</td>
<td>-18.41</td>
</tr>
<tr>
<td>Run 8</td>
<td>0.74</td>
<td>-6.86</td>
<td>-3.72</td>
</tr>
<tr>
<td>Run 9</td>
<td>0.78</td>
<td>-14.41</td>
<td>3.47</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.63</td>
<td>-14.41</td>
<td>-18.41</td>
</tr>
<tr>
<td>Average</td>
<td>0.76</td>
<td>2.48</td>
<td>5.50</td>
</tr>
<tr>
<td>Maximum</td>
<td>0.83</td>
<td>34.63</td>
<td>32.20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Water year 2007</th>
<th>LTC</th>
<th>MSC</th>
<th>LSC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NSE</td>
<td>RRE</td>
<td>SFDC</td>
</tr>
<tr>
<td>Run 1</td>
<td>0.80</td>
<td>23.95</td>
<td>-0.55</td>
</tr>
<tr>
<td>Run 2</td>
<td>0.91</td>
<td>10.79</td>
<td>13.06</td>
</tr>
<tr>
<td>Run 3</td>
<td>0.64</td>
<td>48.94</td>
<td>-6.01</td>
</tr>
<tr>
<td>Run 4</td>
<td>0.85</td>
<td>18.21</td>
<td>-27.78</td>
</tr>
<tr>
<td>Run 5</td>
<td>0.78</td>
<td>48.37</td>
<td>-9.65</td>
</tr>
<tr>
<td>Run 6</td>
<td>0.75</td>
<td>30.18</td>
<td>14.63</td>
</tr>
<tr>
<td>Run 7</td>
<td>0.80</td>
<td>24.85</td>
<td>11.06</td>
</tr>
<tr>
<td>Run 8</td>
<td>0.75</td>
<td>21.94</td>
<td>39.75</td>
</tr>
<tr>
<td>Run 9</td>
<td>0.83</td>
<td>11.05</td>
<td>41.99</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.64</td>
<td>10.79</td>
<td>-27.78</td>
</tr>
<tr>
<td>Average</td>
<td>0.79</td>
<td>26.48</td>
<td>8.50</td>
</tr>
<tr>
<td>Maximum</td>
<td>0.91</td>
<td>48.94</td>
<td>41.99</td>
</tr>
</tbody>
</table>

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This work was supported by NSF EAR 0943640, 0837937, 0404130, and 0337650 awards to Brian McGlynn and in part through instrumentation funded by the National Science Foundation (grants OCI-0821527 and NSF EAR 8943640, 0943640, and 1356340). This work was additionally partially supported by the Natural Environment Research Council (Consortium on Risk in the Environment: Diagnostics, Integration, Benchmarking, Learning and Elicitation (CREDIBLE); grant no. NE/J017450/1).

Edited by: Ralf Merz
Reviewed by: Uwe Ehret and Markus Hrachowitz

References


Herman, J. D., Kollat, J. B., Reed, P. M., and Wagener, T.: From maps to movies: high-resolution time-varying sensitivity analysis for spatially distributed watershed models, Hydrol. Earth Syst.
Kelleher, C., Wagener, T., and McGlynn, B.: Characterizing and reducing distributed model equifinality

Hydrol. Earth Syst. Sci., 21, 3325–3352, 2017

3350

C. Kelleher et al.: Characterizing and reducing distributed model equifinality


Silvestro, F., Gabellani, S., Rudari, R., Delogu, F., Laiolo, P., and Boni, G.: Uncertainty reduction and parameter estima-


