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

Christa Kelleher[1,2], Brian McGlynn[2], Thorsten Wagener[3,4]

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

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

Abstract. Distributed catchment models are widely used tools for predicting hydrologic behaviour. While distributed models require many parameters to describe a system, they are expected to simulate behaviour 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 'behavioural' sets that fit observations (typically streamflow). In this study, we investigate the extent to which equifinality impacts a typical distributed modelling application. We outline a hierarchical approach to reduce the number of behavioural 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 'behavioural' parameter sets altered distributions of spatio-temporal simulations, simulating a well-studied headwater catchment, Stringer Creek, MT 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 further reduced the number of behavioural 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 behavioural parameter sets can reduce equifinality and bolster more careful application and simulation of spatio-temporal processes via distributed modelling at the catchment scale.

1 Introduction

The field of hydrology has been built upon the combination of field measurements and computational modelling to observe, understand, and predict hydrologic behaviour (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), or Paniconi et al. (2015) for a review of distributed models).

Alongside variations in the structural equations, 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; Fenicia 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., Leavesley 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 km^2).

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., 1997; van Werkhoven et al., 2008; Zhang et al., 2013, Kelleher et al., 2015; Ghasemizade et al., 2016; 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 modelling 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., 2012), 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., 2011; 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 modelling studies have 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, 2012; Gharari et al., 2014; Silvestro 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 (Birkel 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, modelled with the widely applied Distributed Hydrology-Soil-Vegetation Model (DHSVM). We

investigate how the paradigm for identifying 'behavioural' 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 modeller 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. Secondarily, 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 behaviour.

2.0 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 km^2) 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 km^2). 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 880mm 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 biototite 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 km^2.

2.2 The Distributed Hydrology Soil Vegetation Model

The entire Tenderfoot Creek catchment was modelled using the Distributed Hydrology-Soil-Vegetation Model (Wigmosta et al., 1994; Wigmosta et al., 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 was 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 10m and a time step of 3 hours using spatially distributed information for topography, soil depth, soil type, and vegetation (Figure 1). Spatially distributed information for topography and vegetation height were obtained via airborne laser swath mapping (ALSM) at 1m resolution and resampled to 10m resolution (Figure 1a, b). Soil depth was held at a constant one meter, as spatially distributed soil data is limited. However, >160 well and piezometer installations indicate that 1m 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 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 (>10m), trees (2m-10m), and undergrowth (<2m; Figure 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 tandem for cells with both canopy and tall canopy vegetation, with the exception of vegetation height and overstory fractional coverage (aerial percentage of each cell covered by canopy vegetation) which we expected to differ between these two vegetation classes. Parameter values for vegetation were constrained, when possible, based on the plant species at different heights (lodgepole pine when species information was available, more generally 'evergreen conifer' for 2 – 10m and 'grasses and shrubs' for <2m). All vegetation parameters are listed in Table 1, with the minimum and maximum bounds on uninformed (uniform) prior parameter distributions specified in Appendix A and sources for these ranges provided in Kelleher et al. (2015).

A single soil type was used to distribute soil parameters across the catchment, as there is limited small-scale information about soil properties across Stringer Creek. Parameter values for soil information were obtained, when possible, from the CONUS soils database as well as from texture estimates from CONUS combined with the soil water characteristics

hydraulic properties calculator following Saxton and Rawls (2006). Within DHSVM, a total of 15 parameters are used to describe soil characteristics for each soil type.

2.2.2 Meteorological Forcing Data

The model was executed at a time step of three hours to effectively simulate snowmelt while balancing computational cost. Model forcing data includes 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 (Figure 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 was 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-hour 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, 2002).

2.2.3 Model Initialization and Setup

For each parameter sample, the model was run for a six-month warm-up period from April 1, 2006 through September 30, 2006 and an analysis period of October 1, 2006 through October 1, 2008. We calibrate to the 2008 water year (Oct 1, 2007 – Oct 1, 2008), but demonstrate results for the 2007 water year (Oct 1, 2006 – Oct 1, 2007) to show the value of our approach across a longer period. While these periods are short, they encompass both a wetter (2008) and 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 six months should be sufficiently long for the influences of parameter values on model simulations and predictions to emerge. This six month period of initialization was not used in any of the calibration or validation of the model.

3.0 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 modelling and prediction in ungauged catchment literature, we have found that constraints on model-derived hydrologic behaviour 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.
3. Measures of internal catchment behaviour, which describe how well simulations match either observations or perceptions of the spatio-temporal variability of hydrologic behaviour (e.g. Franks et al., 1998; Lamb et al., 1998; Grayson et al., 2002; Wealands et al., 2005; Kuras 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 Figure 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 et al., 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 behaviour. These constraints are best applied to average catchment conditions, with a goal to remove simulations that are too wet or too dry at an annual or seasonal timescale. This information could be considered "soft information" as described by Seibert and McDonnell (2002), because it may not directly based on measurements within the catchment.
- **Error Metrics:** 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), 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 behaviour, e.g. the baseflow index, the slope of the flow duration curve (Wagener 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 behaviour, 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 behaviour:** Aggregated spatial predictions (if values are known) and simulations (if values are unknown) may be used to assess the realism of predicted internal catchment behaviour (Grayson et al., 2002; Wealands et al., 2005; Kuraś et al., 2011; Koch et al., 2016; Fang et al., 2016). There are many examples of distributed models that have evaluated internal catchment behaviour 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; Kuraś et al., 2011). However, in the absence of internal observations, we suggest that internal simulations may still be incorporated into evaluations of distributed model behaviour 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 behaviour 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 Figure 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 behaviour 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 eleven 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 modeller could approach this type of reduction via multiple constraints. As such, having a greater number of parameter sets should not influence our general conclusions.

Following the framework in Figure 2, Table 2 lists all metrics used to assess catchment behaviour, 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 behaviour, 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., Sankarasubranmanian and Vogel, 2003), including the CGIAR-CSI Global Aridity and Global-PET database (Zomer et al., 2007; 2008), which was used to limit 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, 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 is more observational data present within the catchment than is 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 behaviour:

- Runoff ratio error (RRE_Q), to capture the annual water balance,
- Error in the slope of the flow duration curve ($SFDCE_Q$), assessed between the 10th (p_l) and 30th (p_h) percentile exceedance flows, to capture the variability in transition flows,
- Error in the magnitude of the streamflow peak (P_Q),
- Timing of the peak magnitude (T_Q), both selected to capture the interaction between snowmelt and streamflow, and
- Total storage error (VOL_S), similar to the RR_E but representative of total annual SWE storage behaviour, to match overall SWE volume.

As individual daily flow events are the most difficult to match and tend to be influenced by a range of different model parameters, we left constraints on these behaviours to be widest (Table 2). Additionally, as peak flow is likely to be most influenced by epistemic errors, we left bounds on peak flow to be reasonably wide, so as to avoid incorrectly constraining a potentially uncertain measurement. Recognizing that many of these metrics may be conflicting, our goal is to end with parameter sets that perform reasonably well for many different types of streamflow and SWE behaviour, as opposed to performing well for only one or two metrics, yielding an unrealistic hydrograph or SWE time series.

Error metrics for the SNOTEL sites were computed at each site, but presented in the manuscript as a single measure across both sites using an 80% (Onion Park) – 20% (Stringer Creek) weighting. We chose this weighting because Onion Park elevation is more representative of Stringer Creek catchment topography, with more than 90% of the catchment at an elevation closer to the Onion Park (2259m) than to Stringer Creek (1983m) site elevations. When weighting both sites into a single metric, the absolute values of the storage volumes were averaged, as the model tended to underpredict at Stringer Creek and overpredict at Onion Park. Last, we extracted spatial predictions of water table depth for Stringer Creek at noon each day from Oct 1, 2007 through Oct 1, 2008 across the entire catchment. These maps were averaged across different periods to create maps of average annual predictions, average predictions during snowmelt (May 1 – July 30), and average predictions during the dry period during the late summer (August 1 through September 30).

4 Results

4.1 How Do Single Constraints Influence Model Performance?

As can be seen in Figure 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 plus or minus 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 Figure 4. When applied hierarchically, a total of 9 parameter sets met all metric constraints (Figure 4a). Individually, a large number of parameter sets met behavioural constraints on SWE NSE, removing only 489 of the 10,000 sets. The opposite was true for behavioural sets constrained based on peak streamflow, which identified only 626 behavioural parameter sets.

Figure 4(b) summarizes the number of acceptable parameter sets that meet groups of constraints, as introduced in Figure 2. Across the different types of signatures and metrics, dynamic constraints overwhelmingly had the single greatest impact on reducing acceptable parameter sets, with just 26 sets meeting all criteria for dynamic metrics. Interestingly, statistical and regional constraints identified a similar number of behavioural sets. While combining regional and local metrics reduced the number of behavioural sets, combining statistical and dynamic constraints did not reduce the number of behavioural sets.

4.3 Are All Metrics Needed?

To evaluate whether all eleven metrics and corresponding constraints are needed to ensure behavioural consistency within this framework, we examined model behaviour and performance across subsets of metrics (n = 2 to 10) as well as the level of redundancy across metrics (Appendix B). Employing only two or three 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 behaviour across our framework would remain. By comparison, including eight metrics generally removes an average of 92.6% of all nonbehavioural 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 nonbehavioural 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 nonbehavioural by constraining one metric were considered behavioural by another metric across a majority of metric combinations.

4.4 How Do Single and Multiple Constraints Impact the Distributions of Metrics of Catchment Behaviour?

Initial distributions (I) of metric values across all 10,000 parameter sets include a wide range of behaviour (Figure 5). While these distributions can be narrowed by the addition of one constraint, many distributions of behavioural performance were also narrowed by the addition of all constraints. The addition of all constraints favoured 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 behavioural sets both over- and underestimated RRE_Q and $SFDCE_Q$, the final behavioural 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 Behaviour?

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 behaviour via regional and local metrics resulted in behavioural 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 well-matched behaviour on the falling limb of the hydrograph. Dynamic metrics also better constrained streamflow simulations during baseflow, which tended to be overestimated by behavioural sets selected by statistical metrics.

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

In contrast to simulations of SWE, behavioural simulations of average water table depth varied widely depending on which metrics were used to constrain behaviour. Behavioural sets identified via statistical metrics included a wide range of average water table behaviour, suggesting that statistical metrics may poorly inform water table dynamics. Behaviour constrained by local and regional metrics produced a pattern of average water table depth consistent with general understanding of water table behaviour (Jensco et al., 2009; Jencso and McGlynn, 2011). As a metric for average water table depth was directly incorporated into the assessment of regional and local behaviour, this result is expected. Dynamic metrics had the largest impact on narrowing WTD simulations, identifying behavioural 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 behavioural sets corresponding to hydrographs and SWE time series that reasonably matched observations (Figure 8). 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 rising limb of the hydrograph and the first streamflow peak were 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 behaviour 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 behaviour 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 (Figure 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_Q) 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 behaviour were able to reduce parameter equifinality, we investigated the extent to which constraining model-predicted and simulated behaviour narrowed ranges for parameter values, with results displayed in Figure 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 LAI (50). We applied a two-sample Komolgov-Smirnov test to assess whether there was a statistically significant difference ($p < 0.1$) between the uninformed 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 Behaviour?

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 behavioural parameter sets yielded similar simulations, and to ask whether spatial diagnostics beyond time series observations should be incorporated into assessments of distributed model behaviour. All predictions of water table depth are included in Figure B3, with three predictions of water table depth displayed in Figure 9. These three encompass the range of predictions from low to high water table depth for annual, snowmelt (May 1 – July 31), and late summer dry-down (Aug 1 – Sept 30) periods. Seasonally, water tables were simulated closer to the surface during snowmelt and closer to bedrock during late summer, with average behaviour 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 behavioural parameter sets, differences in simulations were large in both space and time (Figure 9; Figure B3). At annual timescales, many parts of the catchment were predicted to have similar annual average behaviour, with differences below 0.1m for 38% of all cells. Simulations across 13.3% of the catchment differed by more than 0.2m, 20% of the modelled soil depth. These numbers were comparable for simulations during late summer, when the majority of the catchment was simulated to have similar behaviour. Differences were greatest during the snowmelt period, exceeding simulated ranges of 0.2m 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 behaviour.

4.8.2 Can Simulations Be Used to Further Reduce Equifinality?

Evaluating whether simulations of internal behaviour of water table depth match perceptions and observations of catchment functioning may be done using simple metrics of spatial behaviour. 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 behavioural sets, we calculated the percentage of cells across Stringer Creek where the water table was simulated to be within 0.05m of the surface for the entire snowmelt period (May – July). Four different behavioural 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 behaviour 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 modelling 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 behavioural parameter sets or to justify model performance. As we show in Figure 3, model simulations that meet just one set of criteria may poorly match other catchment-wide basin behaviour. Of particular interest is the impact of other constraints on catchment-wide signatures, including the runoff ratio, aridity index, annual evapotranspiration, and annual average water table depth. For our application, applying a single statistical or dynamic metric did not narrow any signature to ranges deemed to be representative of regional or catchment behaviour.

Interestingly, despite the snow-driven nature of Tenderfoot Creek, many parameter sets were equifinal with respect to the NSE for SWE (Figure 5). This suggests, for this particular application, that a priori parameter ranges did not generate wide variability in SWE time series at the two observation sites, leading us to believe that within DHSVM the meteorological forcing data is a greater driver of modelled SWE. Since this forcing data is also uncertain, there is potential for this uncertainty to propagate into model simulations, and may be driving the overestimation of SWE observed at both SNOTEL sites. Previous model analysis in this catchment found that many snow accumulation and melt parameters were insensitive and highly interactive (Kelleher et al., 2015). Together, these results broadly suggest that not all observations will reduce equifinality. Other studies have questioned the empirical equations used to calculate SWE accumulation and melt, and have found improvements in the prediction of SWE accumulation and melt by altering hard-coded parameters within the model framework (Thyer et al., 2004; Jost et al., 2009).

The largest number of nonbehavioural sets was identified by error in peak streamflow magnitude. While there were several sets that had very small errors in peak streamflow magnitude (Figure 5), many overestimated the annual water balance, and were therefore identified as nonbehavioural. While peak timing was underestimated by the addition of all eleven criteria, error in the runoff ratio and the slope of the flow duration curve were narrowed to ranges of lower error and average water table depth was constrained to more representative ranges when all eleven criteria were added (Figure 6).

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 behaviour for some years (Figure 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 behaviour differently. Statistical constraints poorly constrained simulations of water table depth (Figure 6). In contrast, dynamic constraints, applied only to errors in streamflow and SWE time series, yielded behavioural simulations that closely matched expectations of average water table behaviour across the catchment (Jencso et al., 2009; Jensco and McGlynn, 2011). Our results demonstrate that suites of metrics related to hydrologic behaviour may inform simulations of other hydrologic processes (i.e., average water table depth). In contrast, we found statistical metrics to carry little information regarding other types of simulated or predicted behaviour (Figure 3, Figure 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 behaviour 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 modelling yields model simulations that match streamflow at the calibration site and that generally match patterns of SWE accumulation and melt (Figure 7). In general, while results for the 2008 water year bracket observations, results for the 2008 water year are underpredicted with expect to peak behaviour 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 behaviour 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 calibration period snowmelt. Ahl et al. (2008) modelled 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 behaviour as well as dynamical streamflow behaviour 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 modelling 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 behaviour: 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 behaviour 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 spatio-temporal 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 behaviour. 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, equifinality with respect to catchment average conditions may manifest in variable predictions of internal catchment behaviour (e.g., Figure 9). Thus, evaluating spatial predictions may be one of the few practical approaches to reducing this equifinality.

5.4 Interpretation of Internal Behaviour

Both field investigations and modelling 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; Jensco and McGlynn, 2011; Nippgen et al., 2015). Thus, we chose to evaluate simulations of water table depth for our case study. We found that parameter sets that were equifinal with respect to streamflow and SWE (Figure 7, 8) produced vastly different annual and seasonal simulations of water table depth (Figure 9). Moreover, the locations where simulated differences were largest and smallest across the catchment also varied with time. By assessing 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 behaviour can be used to identify simulations that do not match perceptions or direct observations of catchment behaviour beyond comparison with time series. This approach enables one to move beyond just matching to a few points with observations, encouraging modellers to holistically evaluate performance and simulation of multiple hydrologic processes.

5.5 The Case for Adding Spatial Predictions

Although constraints on the hydrograph and other hydrologic behaviour 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 patterns is time intensive but can be highly informative. Therefore, 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 behaviour are often performed by matching model predictions of different catchment variables (e.g., snowmelt, Thyer et al., 2004; groundwater table dynamics, Kuras et al., 2011) to multiple observations at discrete points across a given catchment, though several studies have also shown how patterns of hydrologic processes may be directly incorporated into evaluating 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 evaluated using expert knowledge, understanding of process controls 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 observations equally well (Figures 5, 6, and 7) predicted very different patterns of internal catchment behaviour (Figure 9). However, these same sets also exhibited divergent parameter values for several different soil and vegetation properties (Figure 8). This result exemplifies the additional information available to constrain behavioural sets as well as the potential for catchment-scale predictive uncertainty driven by parameter equifinality. Sets that met framework criteria simulated very different patterns of water table depth across both annual and seasonal periods (Figure 9). Differences typically were between 0 and 0.2 across all simulations and all catchment cells, but were especially large (> 0.5m) 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 internal catchment behaviour, and therefore recommend that spatial simulations should be incorporated into assessments of distributed catchment behaviour (Figure 2). While we chose to highlight only one spatial diagnostic in this paper, we advocate 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 researchers to select a set of metrics for assessing model performance, emphasizing both temporal and spatial metrics that may help to ensure appropriate representation of key catchment 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 evaluations may especially benefit from joint evaluation by modellers and experimentalists (Seibert and McDonnell, 2002).

The approach we recommend in this manuscript builds on and compliments several recent studies that have sought to improve process consistency across models of varying complexity and within distributed hydrologic models. Many recent studies have shown that, despite their weaknesses, distributed 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 modelling, 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; Clark et al., 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 behaviour, 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 either 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 behaviour with complex, many parameter models. Our work suggests this will only provide the modeller with a better understanding of the catchment but also of the model in question. Altogether, research with distributed models and our own analysis do critique some of the challenges associated with distribute model application, but also highlight the value of distributed models for hydrological predictions (Hrachowitz et al., 2014; Fitachi 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 10,000 parameter samples required thousands of computing hours (and even longer periods if the modeller 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., 2015). 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 modelling community and the fully, distributed, physically-based modelling community, 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 modelling 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 behaviour. 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 behaviour to be most important for reducing predictive uncertainty and selecting behavioural parameter sets. This suggests that researchers should use care when utilizing only statistical metrics to judge model performance or to select behavioural parameter sets, as for our application we found many model runs that had high statistical metric performance poorly matched dynamical hydrologic behaviour. 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 behaviour of the catchment is even within the realm of reality for a site. Here, we recommend the evaluation of internal catchment behaviour 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 behaviour. While evaluating average catchment behaviour and time series observations can be helpful, these types of behaviour 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 behaviour but considerably variable spatial behaviour.

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 modellers (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 spatio-temporal questions are resolved with spatio-temporally-vetted answers.

Code Availability

DHSVM model code is freely available at <http://www.hydro.washington.edu/Lettenmaier/Models/DHSVM/>.

Data Availability

Data sets were provided by the United States Forest Service and Tenderfoot Creek Experimental Forest researchers and by NSF support to Brian McGlynn. Stringer Creek streamflow data may be obtained from the US Forest Service at a 15-minute resolution from (http://www.fs.usda.gov/rds/archive/Product/RDS-2010-0003.2/). Aridity data was obtained from the CGIAR-CSI Global-aridity and Global-PET database, available at (http://www.cgiar-csi.org/). SNOTEL datasets for stations Stringer Creek (Site # 1009) and Onion Park (Site # 1008) are available at (http://www.wcc.nrcs.usda.gov/snow/snow_map.html). ALSM data, including topography and vegetation height, for Tenderfoot Creek are maintained by OpenTopography (http://www.opentopography.org/).

Appendices.

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 Figure 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 six month warm-up period may be sufficiently long enough for the manuscript application.

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 eleven different metrics. To assess whether these metrics provide unique information to the uncertainty analysis, we tested the information contained in different combinations of metrics (Figure B1). Figure B1 displays number of parameter sets that would be removed by all different combinations of metrics, comparing these combinations for different subsets of metrics (2 metrics, 3 metrics, etc.). The number of nonbehavioural model runs identified by these different combinations are shown as distributions for each subset of metrics (2 metrics, 3 metrics, etc.),

and visualized as a percentage of nonbehavioural runs as compared to the number of nonbehavioural model runs identified by the full eleven criteria.

We also benchmarked the redundancy of metrics and their corresponding constraints within the framework and application to Stringer Creek (Figure 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 nonbehavioural 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 nonbehavioural 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 two years of complete simulation. Table B1 reports error metrics for the three gaging sites for both the 2007 and 2008 water years for nine behavioural parameter sets. Finally, we also include all water table depth predictions for nine equifinal parameter sets included in our analysis in Figure B3. Figure 9 displays a subset of predictions.

Competing Interests.

The authors declare 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 number NE/J017450/1).

References

Ahl, R. S., Woods, S. W., and Zuuring, H. R.: Hydrologic calibration and validation of swat in a snow-dominated rocky mountain watershed, Montana, USA, J. Am. Water Resour. As., 44, 1411-1430, doi: 10.1111/j.1752-1688.2008.00233.x, 2008.

Ajami, N. K., Gupta, H., Wagener. T., and Sorooshian, S.: Calibration of a semi-distributed hydrological model for streamflow estimation along a river system, J Hydrol., 298, 112-135, doi: 10.1016/j.jhydrol/2004.03.033, 2004.

Band, L., Peterson, D. , Running, S., Coughlan, J., Lammers, R., Dungan, J. and Nemani, R.: Forest ecosystem processes at the watershed scale: Basis for distributed simulation. Ecol. Modell., 56, 171–196, doi:10.1016/0304-3800(91)90199-B, 1991.

Band, L., Patterson, J., Nemani, R., and Running, S.: Forest ecosystem processes at the watershed scale: Incorporating hillslope hydrology Agric. For. Meteor., 63, 93–126, doi:10.1016/0168-1923(93)90024-C, 1993.

Bennett, N.D., Croke, B.F., Guariso, G., Guillaume, J.H., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T., Norton, J.P., Perrin, C. and Pierce, S.A.: Characterising performance of environmental models, Environ. Modell. Softw., 40, 1-20, doi:10.1016/j.envsoft.2012.09.011, 2013.

Beven K. J.: Changing ideas in hydrology: the case of physically based models, J. Hydrol., 105, 157–172, doi:10.1016/0022-1694(89)90101-7, 1989.

Beven, K. J.: Prophecy, reality and uncertainty in distributed hydrological modelling, Adv. Water Resourc., 16, 41–51, doi:10.1016/0309-1708(93)90028-E, 1993.

Beven, K. J.: How far can we go in distributed hydrological modelling?, Hydrol. Earth Syst. Sci., 5, 1-12, doi:10.5194/hess-5-1-2001, 2001.

Beven, K. J.: Towards an alternative blueprint for a physically based digitally simulated hydrologic response modelling system, Hydrol. Proces., 16, 189-206, doi: 10.1002/hyp.343, 2002.

Beven, K. J.: A manifesto for the equifinality thesis, J. Hydrol., 320(1–2), 18-36, doi:10.1016/j.jhydrol.2005.07.007, 2006.

Beven, K. J. and Binley, A. M.: The future of distributed models: model calibration and uncertainty prediction, Hydrol. Proces., 6, 279-298, doi:10.1002/hyp.3360060305, 1992.

Beven, K. J. and Binley, A. M.: GLUE: 20 years on, Hydrol. Process., 28, 5897–5918, doi:10.1002/hyp.10082, 2014.

Beven, K. J. and Kirkby, M. J.: Towards a simple physically based variable contributing model of catchment hydrology, Working Paper 154, School of Geography, University of Leeds, 1976.

Beven, K. J. and Kirkby, M. J.: A physically based, variable contributing area model of basin hydrology, Hydrological Sciences Bulletin, 24:1, 43-69, doi:10.1080/02626667909491834, 1979.

Birkel, C., Soulsby, C., and Tetzlaff, D.: Developing a consistent process-based conceptualization of catchment functioning using measurements of internal state variables, Water Resour. Res., 50, 3481–3501, doi:10.1002/2013WR014925, 2014.

Bixio, A. C., Gambolati, G. , Paniconi, C., Putti, M., Shestopalov, V. M., Bublias, V. N., Bohuslavsky, A. S., Kastelteseva, N. B. and Rudenko, Y. F.: Modelling groundwater–surface water interactions including effects of morphogenetic depressions in the Chernobyl exclusion zone, Environ. Geol., 42, 162–177, doi:10.1007/s00254-001-0486-7, 2002.

Bloschl, G., Sivapalan, M., Wagener, T., Viglione, A., and Savenije, H.: Runoff prediction in ungauged basins: Synthesis across processes, places, and scales, Cambridge University Press, Cambridge, UK, 2013.

Camporese, M., Paniconi, C., Putti, M. and Orlandini, S.: Surface-subsurface flow modelling with path-based runoff routing, boundary condition-based coupling, and assimilation of multisource observation data, Water Resour. Res., 46, W02512, doi:10.1029/2008WR007536, 2010.

Chen, Z., Hartmann, A., and Goldscheider, N.: A new approach to evaluate spatiotemporal dynamics of controlling parameters in distributed environmental models, Environ. Modell. Softw., 87, 1-16, doi:10.1016/j.envsoft.2016.10.005, 2017.

Church, M. R., Bishop, G. D., and Cassell, D. L.: Maps of regional evapotranspiration and runoff/precipitation ratios in the Northeast United States, J. Hydrol., 168, 283-298, doi:10.1016/0022-1694(94)02640-W, 1995.

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., Wagener, T., and Hay, L. E.: Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models, Water Resour. Res., 44, W00B02, doi:10.1029/2007WR006735, 2008.

Clark, M. P., Kavetski, D., and Fenicia, F.: Pursuing the method of multiple working hypotheses for hydrological modelling, Water Resour. Res., 47, W09301, doi:10.1029/2010WR009827, 2011.

Crawford, N. H. and Linsley, R. K.: Digital Simulation in Hydrology: Stanford Watershed Model IV, Technical Report No. 39, Department of Civil Engineering, Stanford University, Palo Alto, CA, 1966.

Cuo, L., Lettenmaier, D. P., Alberti, M., and Richey, J. E.: Effects of a century of land cover and climate change on the hydrology of Puget Sound basin, Hydrol. Process., 23, 907-933, doi:10.1002/hyp.7228, 2009.

Cuo, L., Giambelluca, T. W., and Ziegler, A. D.: Lumped parameter sensitivity analysis of a distributed hydrological model within tropical and temperate catchments, Hydrol. Process., 25, 2405–2421, doi:10.1002/hyp.8017, 2011.

Das., T., Bardossy, A., Zehe., E, and He., Y.: Comparison of conceptual model performance using different representations of spatial variability, J. Hydrol., 356, 106-118, doi:10.1016/j.jhydrol.2008.04.008, 2008.

Dingman, L.: Physical Hydrology, Waveland Press, Long Grove, IL, 2001.

Du, E., Link, T. E., Gravelle, J. A., and Hubbart, J. A.: Validation and sensitivity test of the distributed hydrology soil-vegetation model (DHSVM) in a forested mountain watershed, Hydrol. Process., 28, 6196–6210, doi:10.1002/hyp.10110, 2014.

Euser, T., Winsemius, H. C., Hrachowitz, M., Fenicia, F., Uhlenbrook, S., and Savenije, H. H. G.: A framework to assess the realism of model structures using hydrological signatures, Hydrol. Earth Syst. Sci., 17, 1893-1912, doi:10.5194/hess-17-1893-2013, 2013.

Euser, T. Hrachowitz, M., Winsemius, H.C., and Savenije, H. H.: The effect of forcing and landscape distribution on performance and consistency of model structures, Hydrol. Process., 29, 3727-3743, doi: 10.1002/hyp.10445, 2015.

Fang, Z., Bogena, H., Kollet, S., and Vereecken, H.: Scale dependent parameterization of soil hydraulic conductivity in 3D simulation of hydrological processes in a forested headwater catchment, J. Hydrol., 536, 365-375, doi: 10.1016/j.jhydrol.2016.03.020, 2016.

Farnes, P. E., Shearer, R. C., McCaughey, W. W., and Hansen, K. J.: Comparisons of Hydrology, Geology, and Physical Characteristics Between Tenderfoot Creek Experimental Forest (East Side) Montana, and Coram Experimental Forest (West Side) Montana, Final Report RJVA-INT-92734, USDA Forest Service, Intermountain Research Station, Forestry Sciences Laboratory. Bozeman, Mont., 19 pp, 1995.

Fatichi, S., Vivoni, E.R., Ogden, F.L., Ivanov, V.Y., Mirus, B., Gochis, D., Downer, C.W., Camporese, M., Davison, J.H., Ebel, B. and Jones, N: An overview of current applications, challenges, and future trends in distributed process-based models in hydrology, J. Hydrol., 537, 45-60, doi: 10.1016/j.jhydrol.2016.03.026, 2016.

Fenicia., F., Savenije, H. H., Matgen. P., and Pfister, L.: Understanding catchment behaviour through stepwise model concept improvement, Water Resour. Res., 44, W01402, doi: 10.1029/2006WR005563, 2008.

Fenicia, F., Kavetski, D., and Savenije, H. H. G.: Elements of a flexible approach for conceptual hydrological modelling: 1. Motivation and theoretical development, Water Resour. Res., 47, W11510, doi:10.1029/2010WR010174, 2011.

Fenicia, F., Kavetski, D., Savenije, H. H. G., and Pfister, L.: From spatially variable streamflow to distributed hydrological models: Analysis of key modelling decisions, Water Resour. Res., 52, 954–989, doi:10.1002/2015WR017398, 2016.

Flugel, W.-A.: Delineating hydrological response units by geographical information system analyses for regional hydrological modelling using PRMS/MMS in the drainage basin of the River Brol, Germany, Hydrol. Process., 9, 423-436, doi: 10.1002/hyp.2260090313, 1995.

Franks, S. W., Gineste, P., Beven, K. J., and Merot, P.: On constraining the predictions of a distributed model: The incorporation of fuzzy estimates of saturated areas into the calibration process, Water Resour. Res., 34, 787–797, doi:10.1029/97WR03041, 1998.

Freeze, R. A. and Harlan, R.L.: Blueprint for a physically-based, digitally-simulated hydrologic response model, J. Hydrol., 9, 237-258, doi:10.1016/0022-1694(69)90020-1, 1969.

Gharari, S., Hrachowitz, M., Fenicia, F., Gao, H., and Savenije, H. H. G.: Using expert knowledge to increase realism in environmental system models can dramatically reduce the need for calibration, Hydrol. Earth Syst. Sci., 18, 4839-4859, doi:10.5194/hess-18-4839-2014, 2014.

Ghasemizade, M., Baroni, G., Abbaspour, K., and Schirmer, M.: Combined analysis of time-varying sensitivity and identifiability indices to diagnose the response of a complex environmental model. Environmental Modelling & Software, 88, 22-34, doi: 10.1016/j.envsoft.2016.10.011, 2017.

Graeff, T., Zehe, E., Blume, T., Francke, T. and Schröder, B.: Predicting event response in a nested catchment with generalized linear models and a distributed watershed model. Hydrol. Process., 26: 3749–3769. doi:10.1002/hyp.8463, 2012.

Grayson, R. B., Moore, I. D., and McMahon, T. A.: Physically based hydrologic modelling: 2. Is the concept realistic?, Water Resour. Res., 28(10), 2659–2666, doi:10.1029/92WR01259, 1992.

Grayson, R. B., Blöschl, G., Western, A. W., and McMahon, T. A.: Advances in the use of observed spatial patterns of catchment hydrological response, Adv. Water Resour., 25, 1313-1334, doi:10.1016/S0309-1708(02)00060-X, 2002.

Gupta, H.V., Sorooshian, S., and Yapo, P. O.: Towards improved calibration of hydrologic models: multiple and noncommensurable measures of information, Water Resour. Res., 34, 751-763, doi:10.1029/97WR03495, 1998.

Gupta, H. V., Wagener, T., and Liu, Y.: Reconciling theory with observations: elements of a diagnostic approach to model evaluation, Hydrol. Process., 22, 3802–3813, doi: 10.1002/hyp.6989, 2008.

Guse, B., Pfannerstill, M., Strauch, M., Reusser, D. E., Lüdtke, S., Volk, M., Gupta, H., and Fohrer, N.: On characterizing the temporal dominance patterns of model parameters and processes, Hydrol. Process., 30: 2255–2270. doi: 10.1002/hyp.10764, 2016.

Harmel, R. D., Smith, P. K., Migliaccio. K. L., Chaubey, I., Douglas-Mankin, K., Benham, B., Shukla, S., Muñoz-Carpena, R., Robson, B. J.: Evaluating, interpreting, and communicating performance of hydrologic/water quality models considering intended use: a review and recommendations, Environ. Model. Softw., 57, 40–51. doi:10.1016/j.envsoft.2014.02.013, 2014.

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. Sci., 17, 5109-5125, doi:10.5194/hess-17-5109-2013, 2013.

Hill, M. C., Kavetski, D., Clark, M., Ye, M., Arabi, M., Lu, D., Foglia, L. and Mehl, S.: Practical Use of Computationally Frugal Model Analysis Methods, Groundwater, 54, 159–170, doi:10.1111/gwat.12330, 2016.

Hornberger, G. M. and Spear, R. C.: Eutrophication in Peel Inlet – I. The problem-defining behaviour and a mathematical model for the phosphorus scenario, Water Res., 14, 29–42, doi:10.1016/0043-1354(80)90039-1, 1980.

Hornberger G. M. and Spear, R. C.: An approach to the preliminary analysis of environmental systems, J. Environ. Manage, 12, 7–18, 1981.

Hrachowitz, M., Fovet, O., Ruiz, L., Euser, T., Gharari, S., Nijzink, R., Freer, J., Savenije, H. H. G., and C. Gascuel-Odoux: Process consistency in models: The importance of system signatures, expert knowledge, and process complexity, Water Resour. Res., 50, 7445–7469, doi:10.1002/2014WR015484, 2014.

Hrachowitz, M. and Clark, M.: HESS Opinions: The complementary merits of top-down and bottom-up modelling philosophies in hydrology, Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-36, in review, 2017.

Hundecha, Y., and Bardossy, A.: Modeling effect of land use changes on runoff generation of a river basin through parameter regionalization of a watershed model, J. Hydrol., 292, 281–295, doi: 10.1016/j.jhydrol.2004.01.002, 2004.

Hundecha, Y., Ouarda, T. B. M. J., and Bardossy, A.: Regional estimation of parameters of a rainfall-runoff model at ungauged watersheds using the spatial structures of the parameters within a canonical physiographic-climatic space, Water Resour. Res., 44, W01427, doi:10.1029/2006WR005439, 2008.

Jefferson, J. L., Gilbert, J. M., Constantine, P. G., and Maxwell, R. M.: Active subspaces for sensitivity analysis and dimension reduction of an integrated hydrologic model, Comput. Geosci., 90, 78-89, doi:10.1016/j.cageo.2015.11.002, 2016.

Jencso, K. J., McGlynn, B. L., Gooseff, M. N., Wondzell, S. M., Bencala, K. E., and Marshall, L. A.: Hydrologic connectivity between landscapes and streams: Transferring reach and plot scale understanding to the catchment scale, Water Resour. Res., W04428, doi:10.1029/2008WR007225, 2009.

Jencso, K. G., and McGlynn, B. L.: Hierarchical controls on runoff generation: Topographically driven hydrologic connectivity, geology, and vegetation, Water Resour. Res., 47, W11527, doi:10.1029/2011WR010666, 2011.

Jost, G., Moore, R. D., Weiler, M., Gluns, D. R., and Alila, Y.: Use of distributed snow measurements to test and improve a snowmelt model for predicting the effect of forest clear-cutting, J. Hydrol., 376, 94-106, 2009.

Kampf, S. K., and Burges, S. J.: A framework for classifying and comparing distributed hillslope and catchment hydrologic models, Water Resour. Res., 43, W05423, doi:10.1029/2006WR005370, 2007.

Keesman, K. J.: Set-theoretic parameter estimation using random scanning and principal component analysis, Math. Comput. Simul., 32, 535-543, doi:10.1016/0378-4754(90)90009-8, 1990.

Kelleher, C., Wagener, T., and McGlynn, B.: Model-based analysis of the influence of catchment properties on hydrologic partitioning across five mountain headwater subcatchments, Water Resour. Res., 51, 4109–4136, doi:10.1002/2014WR016147, 2015.

Kling, H., and Gupta, H.: On the development of regionalization relationships for lumped watershed models: The impact of ignoring sub-basin scale variability, J Hydrol., 373(3), 337-351, doi: 10.1016/j.jhydrol.2009.04.031, 2009.

Koch, J., Cornelissen, T., Fang, Z., Bogena, H., Diekkrüger, B., Kollet, S. and Stisen, S.: Inter-comparison of three distributed hydrological models with respect to seasonal variability of soil moisture patterns at a small forested catchment, J. Hydrol., 533, 234-249, doi: 10.1016/j.jhydrol.2015.12.002, 2016.

Koren, V., Moreda, F., and Smith, M.: Use of soil moisture observations to improve parameter consistency in watershed calibration, Phys. Chem. Earth, 33, 1068–1080, doi:10.1016/j.pce.2008.01.003, 2008.

Krug, W.R., Gebert, W.A., Graczyk, D.J., Stevens, D.L., Rochelle, B.P., and Church, M.R.: Map of mean annual runoff for the northeastern, southeastern, and mid-Atlantic United States Water Years 1951- 80, U.S. Geological Survey Water Resources Investigations Report 88-4094, Denver, CO, 1990.

Kumar, R., Livneh, B., and Samaniego, L.: Toward computationally efficient large-scale hydrologic predictions with a multiscale regionalization scheme, Water Resour. Res., 49, doi:10.1002/wrcr.20431, 2013.

Kuraś, P. K., Alila, Y., Weiler, M., Spittlehouse, D. and Winkler, R.: Internal catchment process simulation in a snow-dominated basin: Performance evaluation with spatiotemporally variable runoff generation and groundwater dynamics, Hydrol. Process., 25, 3187–3203, doi:10.1002/hyp.8037, 2011.

Kuraś, P. K., Alila, Y., and Weiler, M.: Forest harvesting effects on the magnitude and frequency of peak flows can increase with return period, Water Resour. Res., 48, W01544, doi:10.1029/2011WR010705, 2012.

Lamb, R., Beven, K., and Myrabø, S.: Use of spatially distributed water table observations to constrain uncertainty in a rainfall–runoff model, Adv. Water Resour., 22, 305-317, doi: 10.1016/S0309-1708(98)00020-7, 1998.

Leavesley, G. H., Lichty, R. W., Troutman, B. M., and Saindon, L. G.: Precipitation-Runoff Modeling System: User's manual, Water Resources Investigations Report 83-4238, United States Department of the Interior, Denver, Colorado, USA, 1983.

McGlynn, B.L., Bloschl, G., Borga, M., Borman, H., Hurkmans, R., Nandagiri, L., Uijlenhoet, R., and Wagener, T.: A data acquisition framework for runoff prediction in ungauged basins, in: Runoff Prediction in Ungauged Basins: Synthesis across Processes, Places and Scales, Cambridge University Press, Cambridge, UK, 2013.

Melsen, L., Teuling, A., Torfs, P., Zappa, M., Mizukami, N., Clark, M., and Uijlenhoet, R.: Representation of spatial and temporal variability in large-domain hydrological models: case study for a mesoscale pre-Alpine basin, Hydrol. Earth Syst. Sci., 20, 2207-2226, doi:10.5194/hess-20-2207-2016, 2016.

Milly, P. C. D: Climate, soil water storage, and the average water balance, Water Resour. Res., 30, 2143–2156, doi:10.1029/94WR00586, 1994.

Mincemoyer, S. A. and Birdsall, J. L.: Vascular flora of the Tenderfoot Creek Experimental Forest, Little Belt Mountains, Montana, Madrono, 53, 211–222, doi:10.3120/0024-9637(2006)53(211:VFOTTC)2.0.CO;2, 2006.

Mitchell, S. R., Emanuel, R., McGlynn, B. L.: Land–atmosphere carbon and water flux relationships to vapor pressure deficit, soil moisture, and stream flow, Agric. For. Meteorol., 208, 108-117, doi:10.1016/j.agrformet.2015.04.003, 2015.

Moreau, P., Viaud, V., Parnaudeau, V., Salmon-Monviola, J. Durand, P.: An approach for global sensitivity analysis of a complex environmental model to spatial inputs and parameters: A case study of an agro-hydrological model, Environ. Modell. Softw., 47, 74-87, doi:10.1016/j.envsoft.2013.04.006, 2013.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., and Veith, T. L.: Model evaluation guidelines for systematic quantification of accuracy in watershed simulations, Trans ASABE, 50, 885 – 900, doi:10.13031/2013.23153., 2007.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models, Part I - A discussion of principles, J. Hydrol., 10, 282–290, doi:10.1016/0022-1694(70)90255-6, 1970.

Nijzink, R. C., Samaniego, L., Mai, J., Kumar, R., Thober, S., Zink, M., Schäfer, D., Savenije, H. H. G., and Hrachowitz, M.: The importance of topography-controlled sub-grid process heterogeneity and semi-quantitative prior constraints in distributed hydrological models, Hydrol. Earth Syst. Sci., 20, 1151-1176, doi:10.5194/hess-20-1151-2016, 2016.

Nippgen, F., McGlynn, B. L., and Emanuel, R. E.: The spatial and temporal evolution of contributing areas, Water Resour. Res., 51, 4550– 4573, doi:10.1002/2014WR016719, 2015.

O'Loughlin, E. M.: Saturation regions in catchments and their relations to soil and topographic properties, J. Hydrol., 83, 307-335, doi:10.1016/0022-1694(81)90003-2, 1981.

Paniconi, C., and Putti, M.: Physically based modelling in catchment hydrology at 50: Survey and outlook, Water Resour. Res., 51, 7090–7129, doi:10.1002/2015WR017780, 2015.

Pfannerstill, M., Guse, B., Fohrer, N.: Smart low flow signature metrics for an improved overall performance evaluation of hydrological models, J. Hydrol., 510, 447-458, doi: 10.1016/j.jhydrol.2013.12.044, 2014.

Pokhrel, P., Gupta, H. V., and Wagener, T.: A spatial regularization approach to parameter estimation for a distributed watershed model, Water Resour. Res., 44, W12419, doi:10.1029/2007WR006615, 2008.

Ponce, V. M., and Shetty, A. V.: A conceptual model of catchment water balance. 1. Formulation and calibration, J. Hydrol., 173, 27–40, 1995a.

Ponce, V. M., and Shetty, A. V.: A conceptual model of catchment water balance. 2. Application to runoff and baseflow modelling, J. Hydrol., 173, 41–50, 1995b.

Qu, Y.: An integrated hydrologic model for multi-process simulation using semi-discrete finite volume approach. Ph.D. thesis, The Pennsylvania State University, 136 pp, 2004.

Qu, Y., and Duffy, C. J.: A semidiscrete finite volume formulation for multiprocess catchment simulation. Water Resour. Res., 43, W08419, doi:10.1029/2006WR005752, 2007.

Rakovec, O., Hill, M. C., Clark, M. P., Weerts, A. H., Teuling, A. J., and Uijlenhoet, R.: Distributed evaluation of local sensitivity analysis (DELSA), with application to hydrologic models, Water Resour. Res., 50, 409-426, doi: 10.1002/2013WR01463, 2014.

Rakovec, O., Kumar, R., Attinger, S., and Samaniego, L.: Improving the realism of hydrologic model functioning through multivariate parameter estimation, Water Resour. Res., 52, 7779-7792, doi: 10.1002/2016WR019430, 2016.

Refsgaard, J. C.: Parameterisation, calibration and validation of distributed hydrological models, J. Hydrol., 198, 69–9, doi:10.1016/S0022-1694(96)03329-X, 1997.

Refsgaard, J.C. and Storm, B.: MIKE SHE, in Computer Models of Watershed Hydrology, Singh, V.P., Ed., Water Resources Publications, Colorado, USA, 809-846, 1995.

Reynolds, M.: Geology of Tenderfoot Creek Experimental Forest, Little Belt Mountains, Meagher County, Montana, in Hydrologic and Geologic Characteristics of Tenderfoot Creek Experimental Forest, Montana, Final Rep. RJVA-INT-92734, edited by P. Farnes, pp. 21–32, Intermt. Res. Stn., For. Serv., U.S. Dep. of Agric., Bozeman, Mont, 1995.

Safeeq, M. and Fares, A.: Hydrologic response of a Hawaiian watershed to future climate change scenarios, Hydrol. Process., 26, 2745–2764, doi:10.1002/hyp.8328, 2012.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S.: Global Sensitivity Analysis: The Primer, John Wiley and Sons, Hoboken, NJ, 2008.

Samaniego, L., Kumar. R., and Attinger, S.: Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale, Water Resour. Res., 46, W05523, doi:10.1029/2008WR007327, 2010.

Saxton, K. E. and Rawls, W. J.: Soil Water Characteristic Estimates by Texture and Organic Matter for Hydrologic Solutions, Soil Sci. Soc. Am. J., 70, 1569–1578, doi:10.2136/sssaj2005.0117, 2006.

Seibert, J. and McDonnell, J. J.: On the dialog between experimentalist and modeller in catchment hydrology: Use of soft data for multicriteria model calibration, Water Resour. Res., 38, 1241, doi:10.1029/2001WR000978, 2002.

Shafii, M.H. and Tolson, B. A.: Optimizing hydrological consistency by incorporating hydrological signatures into model calibration objectives, Water Resour. Res, 51, 3796-3814, doi:10.1002/2014WR016520, 2015.

Shields, C. and Tague, C.: Assessing the Role of Parameter and Input Uncertainty in Ecohydrologic Modelling: Implications for a Semi-arid and Urbanizing Coastal California Catchment, Ecosystems, 15, 775-791, doi:10.1007/s10021-012-9545-z, 2015.

Silvestro, F., Gabellani, S., Rudari, R., Delogu, F., Laiolo, P., and Boni, G.: Uncertainty reduction and parameter estimation of a distributed hydrological model with ground and remote-sensing data, Hydrol. Earth Syst. Sci., 19, 1727-1751, doi:10.5194/hess-19-1727-2015, 2015.

Singh, V. P. and Woolhiser, D. A.: Mathematical modelling of watershed hydrology, J. Hydrol. Eng., 7, 270-292, doi:10.1061/(ASCE)1084-0699(2002)7:4(270), 2002.

Smith, T., Marshall, L., McGlynn, B., and Jencso, K.: Using field data to inform and evaluate a new model of catchment hydrologic connectivity, Water Resour. Res., 49, 6834–6846, doi:10.1002/wrcr.20546, 2013.

Smith, T., Hayes, K., Marshall, L., McGlynn, B., and Jencso, K.: Diagnostic calibration and cross-catchment transferability of a simple process-consistent hydrologic model. Hydrol. Process., doi: 10.1002/hyp.10955, 2016.

Spear R. C. and Hornberger, G. M.: Eutrophication in Peel Inlet – II. Identification of critical uncertainties via generalized sensitivity analysis, Water Res, 14, 43–49, doi:10.1016/0043-1354(80)90040-8, 1980.

Surfleet, C. G., Skaugset, A. E., and McDonnell, J. J.: Uncertainty assessment of forest road modelling with the Distributed Hydrology Soil Vegetation Model (DHSVM), Can. J. For. Res., 40, 1397-1409, doi:10.1139/X10-079, 2010.

Tague, C. L., Choate, J. S., and Grant, G.: Parameterizing sub-surface drainage with geology to improve modelling streamflow responses to climate in data limited environments, Hydrol. Earth Syst. Sci., 17, 341-354, doi:10.5194/hess-17-341-2013, 2013.

Tang, Y., Reed, P., Wagener, T., and van Werkhoven, K.: Comparing sensitivity analysis methods to advance lumped watershed model identification and evaluation, Hydrol. Earth Syst. Sci., 11, 793-817, doi:10.5194/hess-11-793-2007, 2007.

Thyer, M., Beckers, J., Spittlehouse, D., Alila, Y., and Winkler, R.: Diagnosing a distributed hydrologic model for two high-elevation forested catchments based on detailed stand- and basin-scale data, Water Resour. Res., 40, 1029–1049, doi:10.1029/2003WR002414, 2004.

van Straten, G. and Keesrnan, K. J.: Uncertainty propagation and speculation in projective forecasts of environmental change: A lakeeutrophication example, J. Forecasting, 10, 163-190, doi:10.1002/for.3980100110, 1991.

van Werkhoven, K., Wagener, T., Reed, P., and Tang, Y.: Characterization of watershed model behaviour across a hydroclimatic gradient, Water Resour. Res., 44, W01429, doi:10.1029/2007WR006271, 2008.

Wagener, T., Boyle, D. P., Lees, M. J., Wheater, H. S., Gupta, H. V., and Sorooshian, S.: A framework for development and application of hydrological models, Hydrol. Earth Syst. Sci., 5, 13-26, doi:10.5194/hess-5-13-2001, 2001.

Wagener, T. and Gupta, H. V.: Model identification for hydrological forecasting under uncertainty, Stoch. Environ. Res. Risk Assess., 19, 378-387, doi:10.1007/s00577-005-0006-5, 2005.

Wagener, T., Sivapalan, M., Troch, P., and Woods, R.: Catchment classification and hydrologic similarity, Geography Compass, 1, 901, doi:10.1111/j.1749-8198.2007.00039., 2007.

Wealands, S.R., Grayson, R. B., and Walker, J. P.: Quantitative comparison of spatial fields for hydrological model assessment—some promising approaches, Adv. in Water Resour., 28, 15-32, doi:10.1016/j.advwatres.2004.10.001, 2005.

Westerberg, I. K. and McMillan, H. K.: Uncertainty in hydrological signatures, Hydrol. Earth Syst. Sci., 19, 3951-3968, doi:10.5194/hess-19-3951-2015, 2015.

Whitaker A., Alila, Y., Beckers, J., Toews, D.: Application of the distributed hydrology soil vegetation model to Redfish Creek, British Columbia: model evaluation using internal catchment data, Hydrol. Process., 17, 199–224, doi:10.1002/hyp.1119, 2003.

Wigmosta, M.S., Vail, L., and Lettenmaier, D. P.: A distributed hydrology-vegetation model for complex terrain, Water Resour. Res., 30, 1665-1679, doi:10.1029/94WR00436, 1994.

Wigmosta, M.S., Nijssen, B., Storck, P., and Lettenmaier, D.P.: The Distributed Hydrology Soil Vegetation Model, in Mathematical Models of Small Watershed Hydrology and Applications, V.P. Singh, D.K. Frevert, Water Resource Publications, Littleton, CO, 2002.

Yadav, M., Wagener, T. W., and Gupta, H.: Regionalization of constraints on expected watershed response behaviour for improved predictions in ungauged basins, Adv. Water Resour., 30(8), 1756-1774, doi:10.1016/j.advwatres.2007.01.005, 2007.

Yapo, P.O., Gupta, H. V., and Sorooshian, S.: Multi-objective global optimization for hydrologic models, J. Hydrol., 204, 83- 97, doi:10.1016/S0022-1694(97)00107-8, 1998.

Yi, X., Zou, R., and Guo, H.: Global sensitivity analysis of a three-dimensional nutrients-algae dynamic model for a large shallow lake, Ecol. Model., 327, 74-84, 2016.

Yilmaz, K. K., Gupta, H. V., and Wagener, T.: A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, Water Resour. Res., 44, W09417, doi:10.1029/2007WR006716, 2008.

Zehe, E., Ehret, U., Pfister, L., Blume, T., Schröder, B., Westhoff, M., Jackisch, C., Schymanski, S. J., Weiler, M., Schulz, K., Allroggen, N., Tronicke, J., van Schaik, L., Dietrich, P., Scherer, U., Eccard, J., Wulfmeyer, V., and Kleidon, A.: HESS Opinions: From response units to functional units: a thermodynamic reinterpretation of the HRU concept to link spatial organization and functioning of intermediate scale catchments, Hydrol. Earth Syst. Sci., 18, 4635-4655, doi:10.5194/hess-18-4635-2014, 2014.

Zhang, C., Chu, J., and Fu, G.: Sobol's sensitivity analysis for a distributed hydrological model of Yichun River Basin, China, 480, 58-68, doi:10.1016/j.jhydrol.2012.12.005, 2013.

Zhang, X., Srinivasan, R., Zhao, K., and Liew, M. V.: Evaluation of global optimization algorithms for parameter calibration of a computationally intensive hydrologic model, Hydrol. Process., 23, 430–441, doi:10.1002/hyp.7152, 2009.

Zomer R. J., Bossio, D. A., Trabucco, A., Yuanjie, L., Gupta D. C., and Singh, V. P.: Trees and Water: Smallholder Agroforestry on Irrigated Lands in Northern India, IWMI Research Report 122, International Water Management Institute, Colombo, Sri Lanka, 2007.

Zomer R. J., Trabucco, A., Bossio, D. A., van Straaten, O., and Verchot, L. V.: Climate Change Mitigation: A Spatial Analysis of Global Land Suitability for Clean Development Mechanism Afforestation and Reforestation, Agric. Ecosystems and Envir., 126, 67-80, doi:10.1016/j.agee.2008.01.014, 2008.

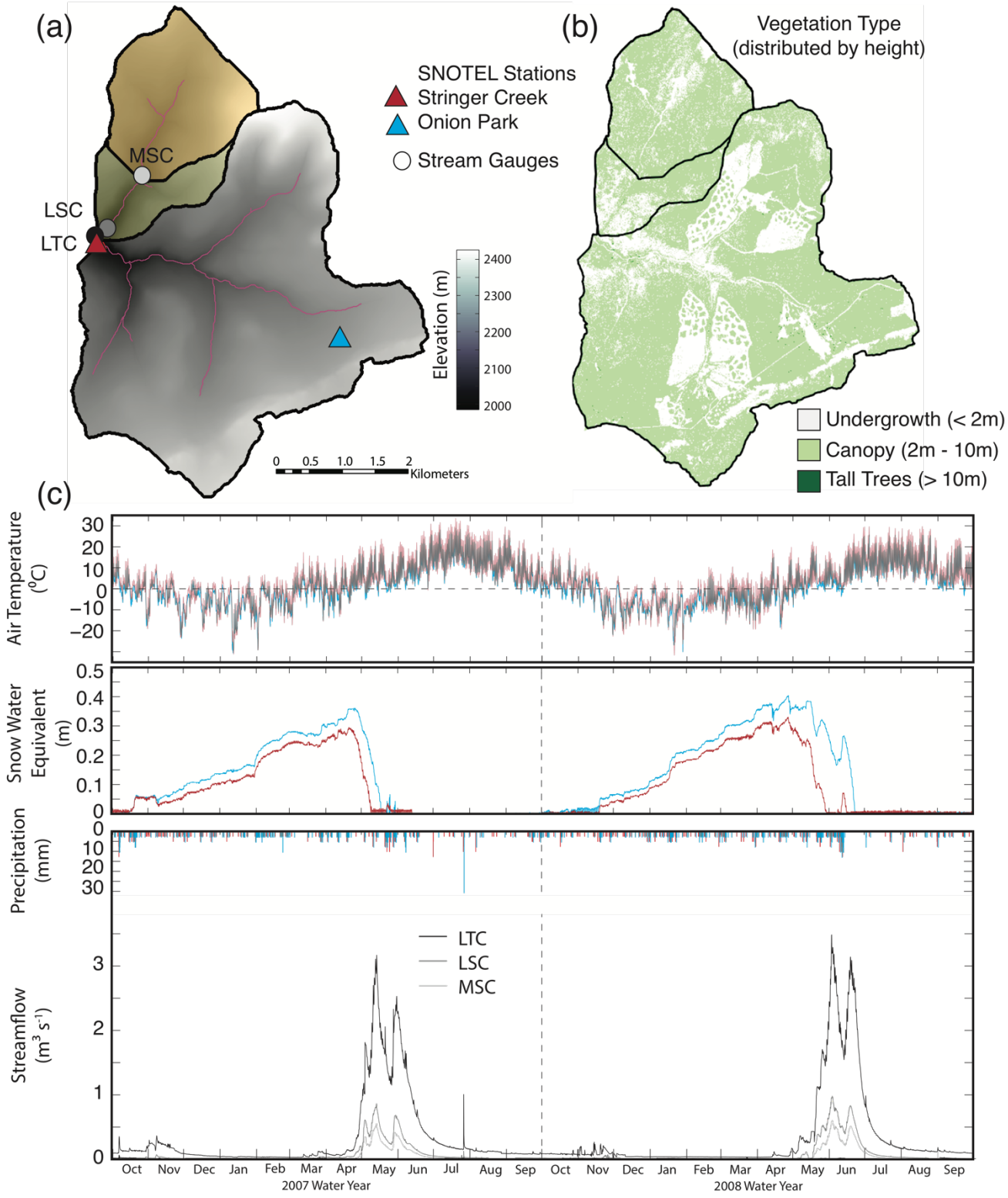
(a)
MSC
LSC
LTC
SNOTEL Stations
Stringer Creek
Onion Park
Stream Gauges
Elevation (m)
2400
2300
2200
2100
2000
0 0.5 1.0 1.5 2
Kilometers
(b)
Vegetation Type
(distributed by height)
Undergrowth (< 2m)
Canopy (2m - 10m)
Tall Trees (> 10m)
(c)
Air Temperature (°C)
Snow Water Equivalent (m)
Precipitation (mm)
Streamflow (m³ s⁻¹)
LTC
LSC
MSC
Oct Nov Dec Jan Feb Mar Apr May Jun Jul Aug Sep
2007 Water Year
Oct Nov Dec Jan Feb Mar Apr May Jun Jul Aug Sep
2008 Water Year

Figure 1: Tenderfoot Creek Experimental Forest (a) 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 behaviour 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.

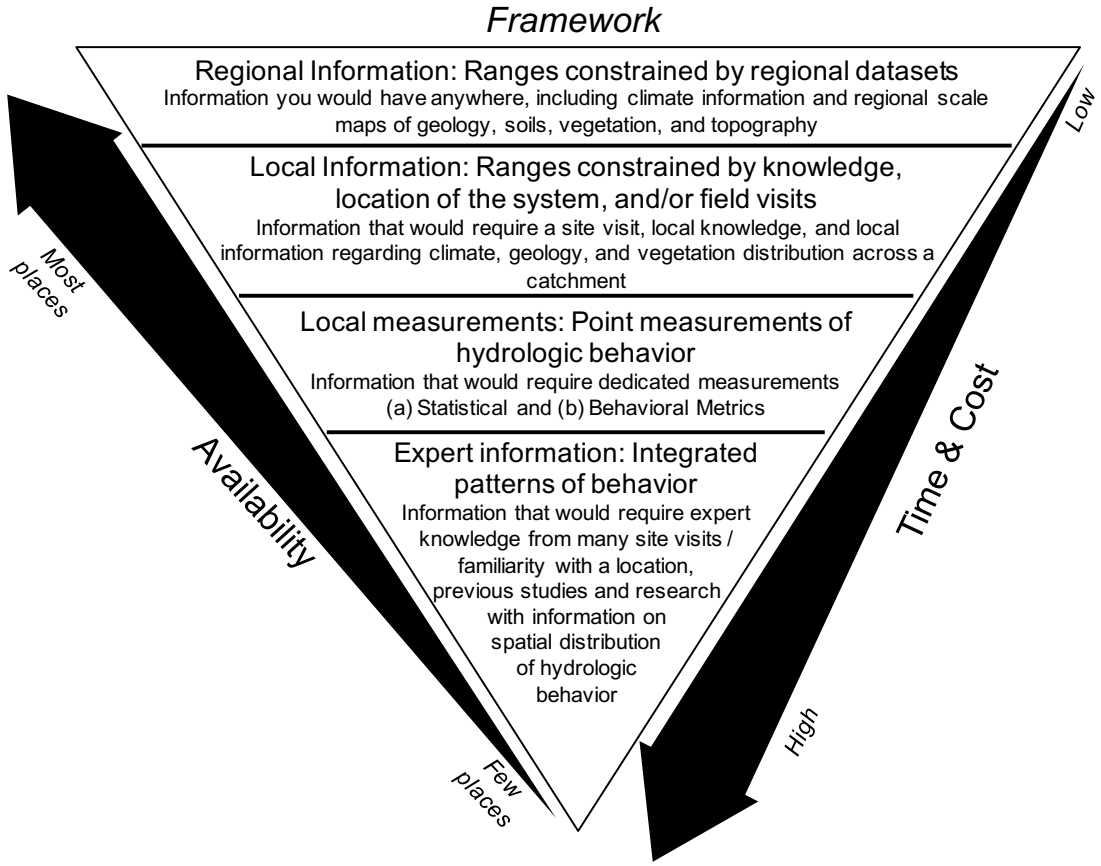

Figure 2: Conceptual framework for constraining environmental predictions for application to a distributed hydrologic model. The framework begins with criteria based on information that has potential to be widely available and relatively inexpensive to acquire, to criteria based on observations, ending with criteria based expert knowledge that may not be available or achievable everywhere.

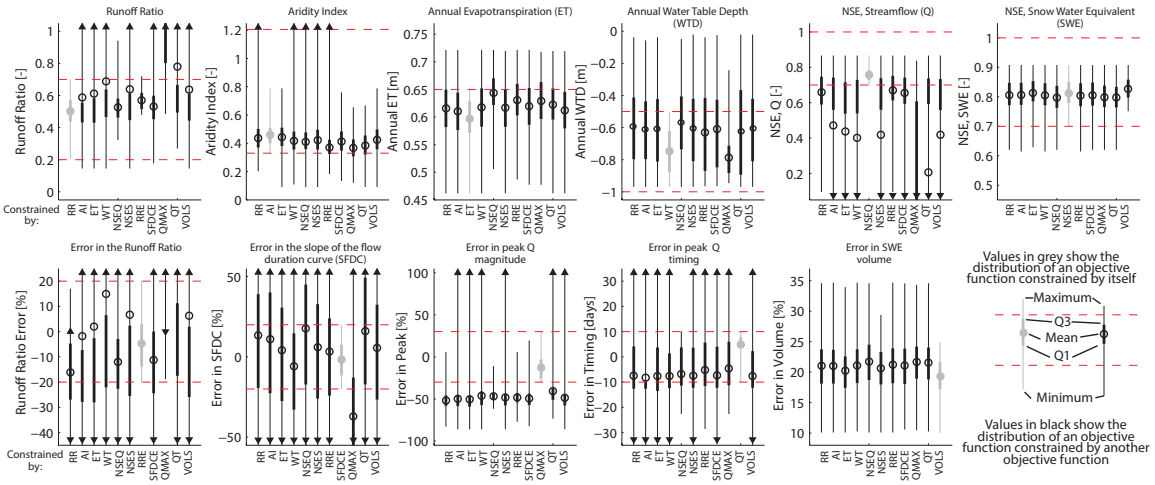

Figure 3: Ranges, interquartile ranges, and mean values for distributions of regional and local signatures, statistical metrics, and dynamic metrics used in the proposed framework. Distributions are shown when a given metric is constrained to behavioural ranges and when a metric is constrained by other metrics, compared to behavioural ranges set in this study. Arrows indicate that ranges exceed the visualized axis limits.

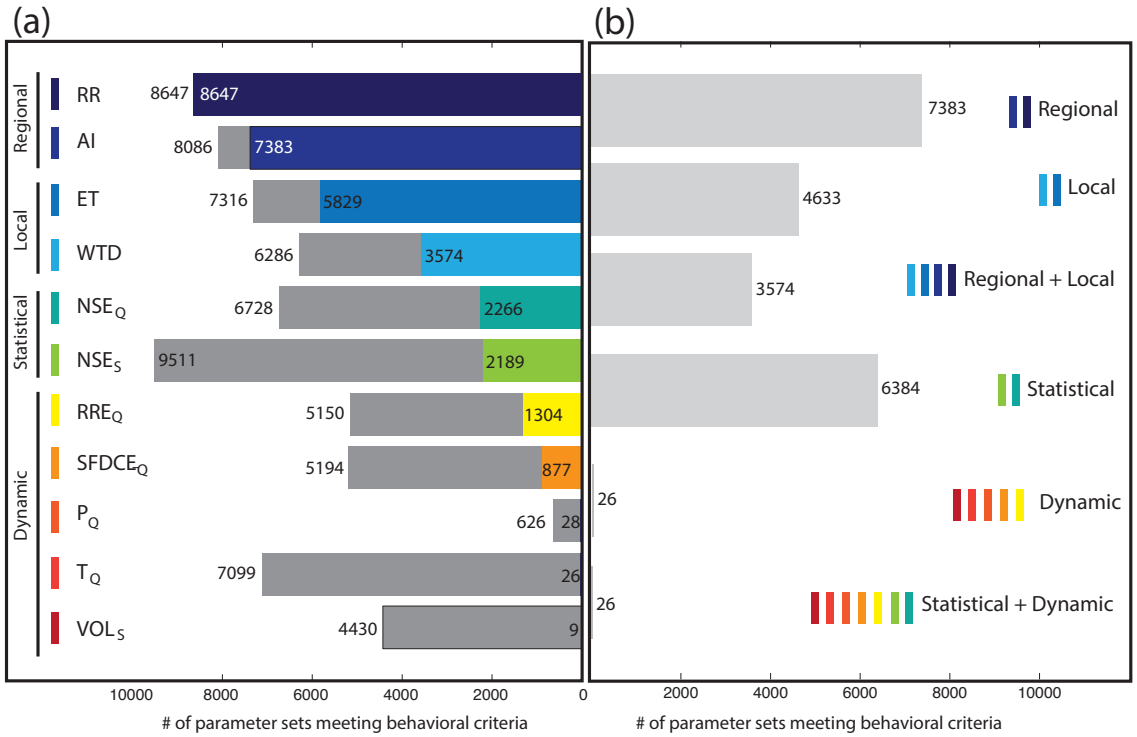

Figure 4: Number of parameter sets retained by each stage of the conceptual framework, shown for (a) individual metrics and (b) groups of metrics specified in Figure 1. For individual metrics (a), bars in color indicate the number of parameter sets retained by sequential application of each of the steps, moving from top to bottom. Bars in grey indicate the number of parameter sets removed if each constraint is applied independent of the others. For groups of constrains (b), bars indicate the number of parameter sets that meet all constraints in a given 'group'.

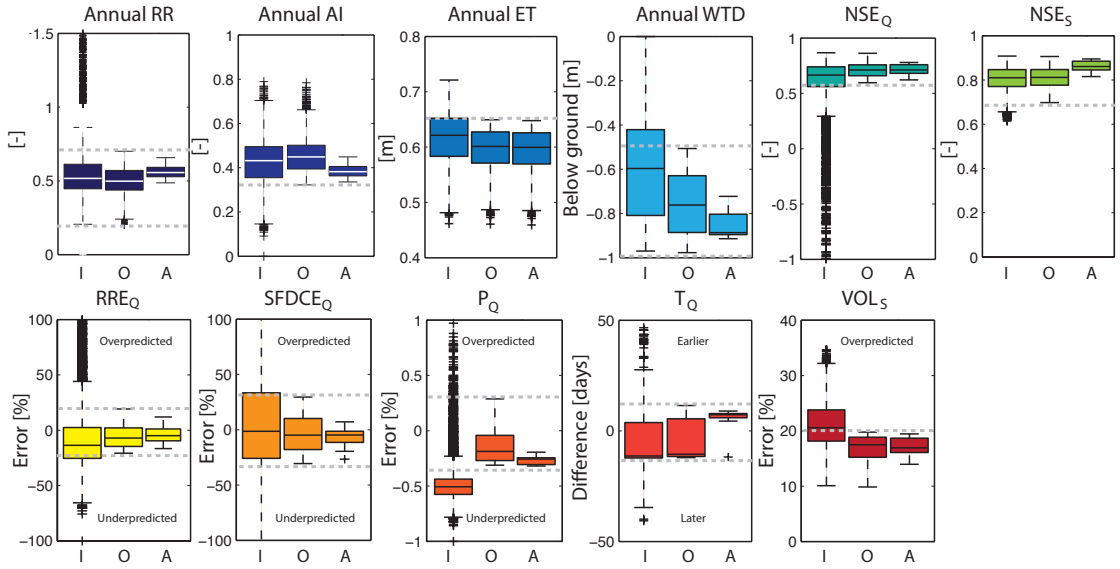

Figure 5: The distribution of model-predicted behaviour across initial distributions of all behaviour (I), distribution of values that meet a given behavioural constraint (O), and distribution of values that meet all eleven constraints (A). Constraints are indicated by grey dashed lines. Colors correspond to different metrics outlined in Figure 3.

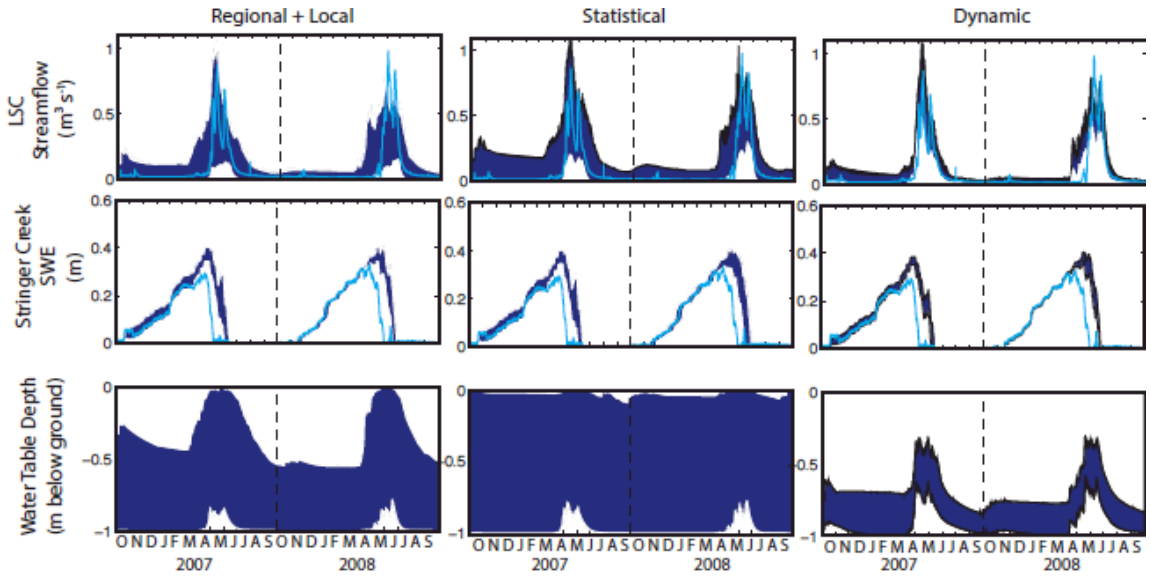

Figure 6: Ranges and distributions for predictions of streamflow, snow water equivalent (Stringer Creek SNOTEL site), and average water table depth for grouped constraints, which are outlined in Figure 4. Predictions that meet a given set of constraints are shown as ranges in blue as a function of time. Results are shown for the calibration period (2008 WY) as well as a period of validation (2007 WY).

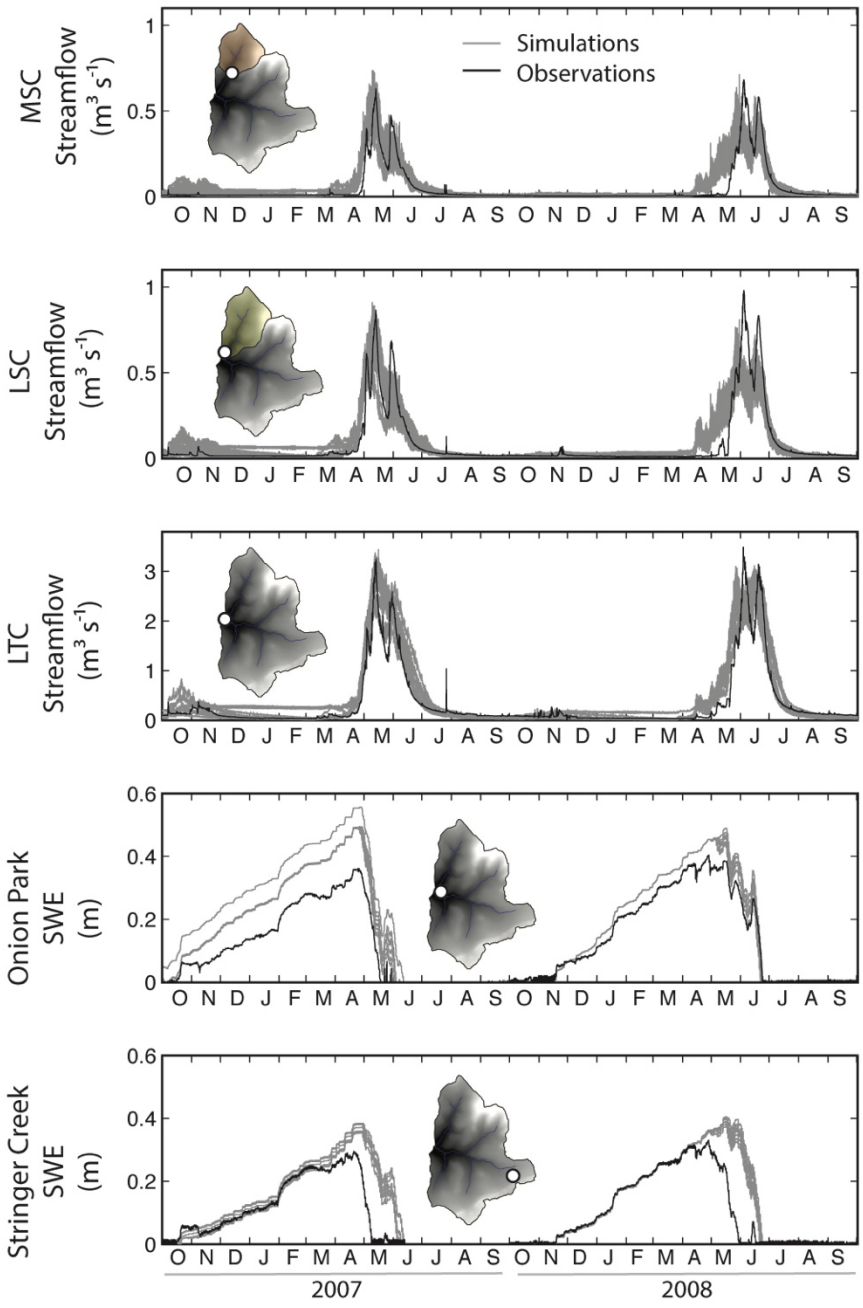

Figure 7: Plots of model-predicted streamflow at the outlets for Middle Stringer Creek, Stringer Creek (calibration site), and Tenderfoot Creek alongside predictions and observations of snow water equivalent at the Onion Park and Stringer Creek SNOTEL that satisfy all hierarchically applied constraints. Each grey line corresponds to the simulation from one of nine behavioural parameter sets.

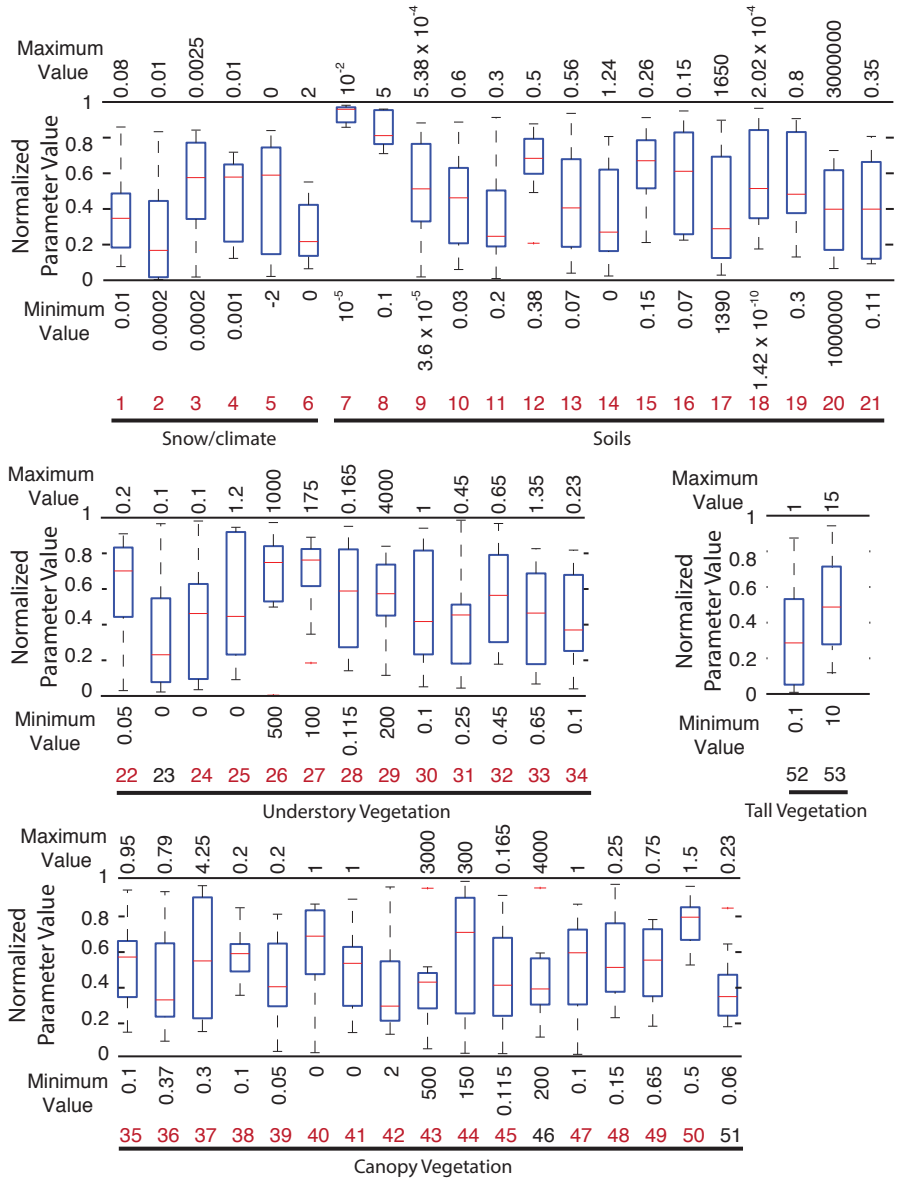

Figure 7: Distributions of behavioural parameter values meeting framework criteria for the 53 parameters varied within the analysis (nine final sets). The ranges for each parameter are listed above (maximum) and below (minimum) each distribution. Distributions are normalized to values between zero and one (y-axis) to enable easier comparison (linearly scaled for all parameters except KLAT, which was $\log_{10}$ scaled). Parameters are grouped by type, with numbers referring to parameter names listed in Table 2. A two tailed Komolgov-Smirnov test was used to assess whether there was a statistically significant difference ($p < 0.1$) between the original parameter sample and the parameter sets that met all framework criteria (nine sets).

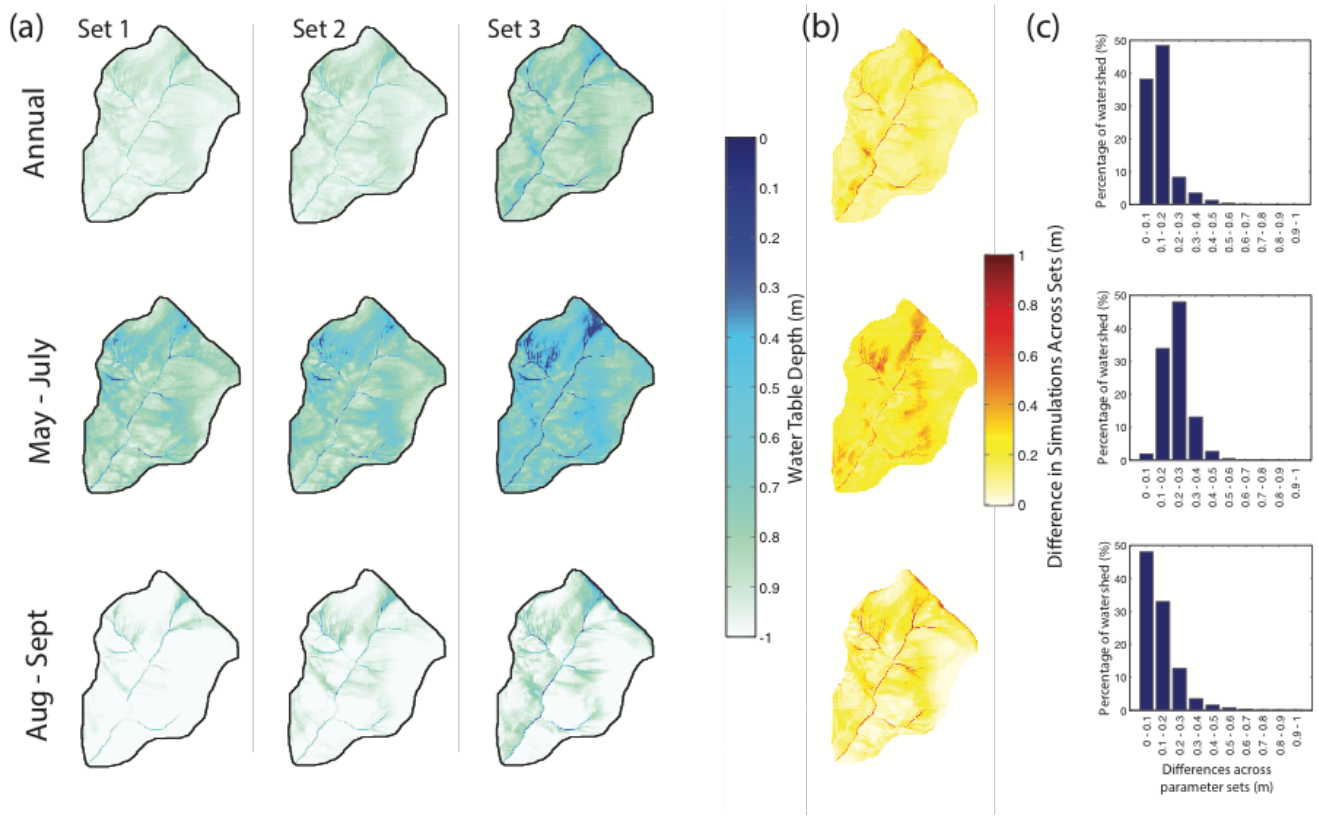

Figure 8: Predictions of annual, May through July, and August through September water table depths (a) in space, across the watershed (for a subset of three parameter sets, 1, 2, and 3 corresponding to sets 1, 5, and 9 in Figure B3), (b) shown as differences in space across nine equifinal parameter sets, and (c) with differences on a cell-by-cell basis summarized as histograms per time period. Color on maps indicates (a) the average depth across a given period between bedrock (-1 m) and the surface (0 m) and (b) the largest difference between all nine predictions at each cell. Predictions are shown for three parameter sets that span the range of behaviour from lower (Set 1) to higher (Set 9) water table conditions across all nine parameter sets. Predictions of water table depth across all behavioural sets are included in Figure B3.

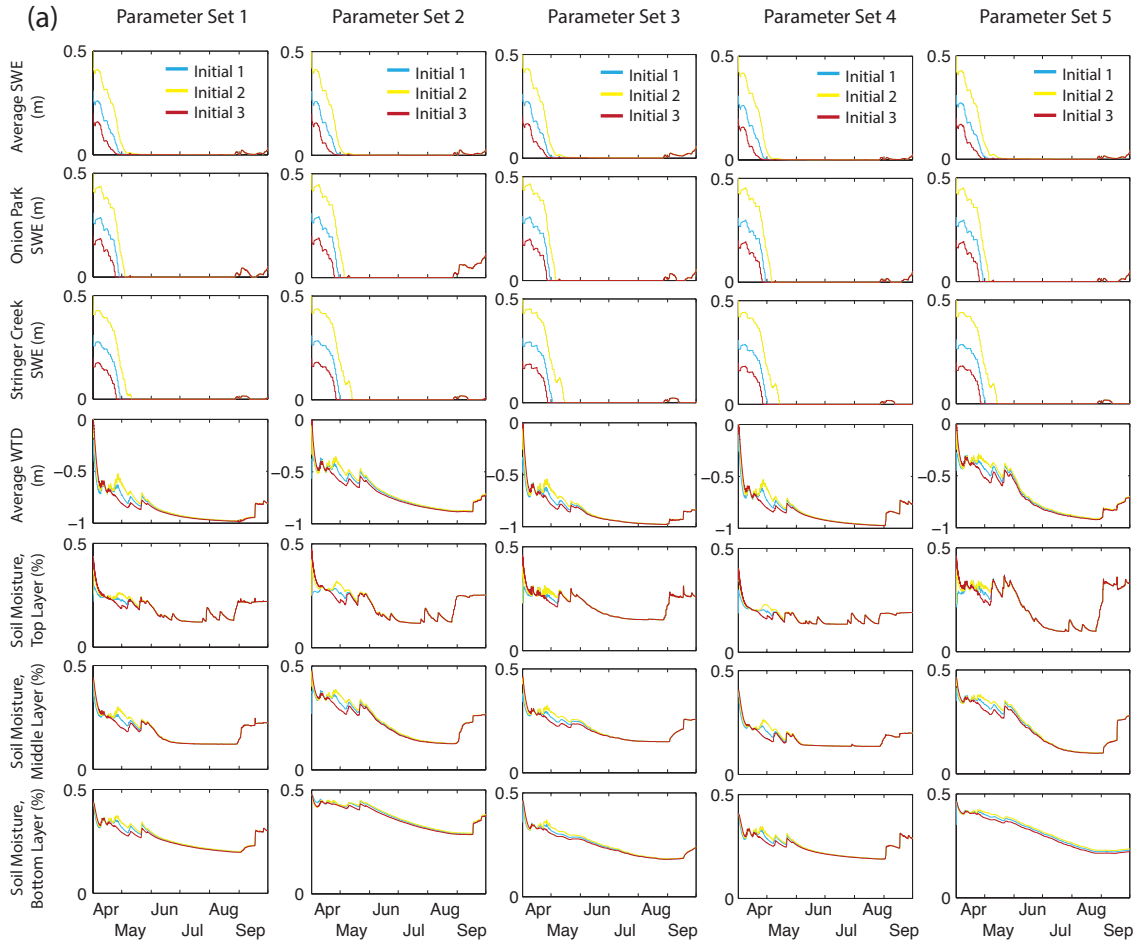

Figure A1: Predictions of annual, May through July, and August through September water table depths (a) in space, across the watershed, (b) shown as differences in space across nine equifinal parameter sets, and (c) with differences on a cell-by-cell basis summarized as histograms per time period. Color on maps indicates (a) the average depth across a given period between bedrock (-1 m) and the surface (0 m) and (b) the largest difference between all nine predictions at each cell. Predictions are shown for three parameter sets that span the range of behaviour from drier (set 1) to average (set 5) to wetter (set 9) conditions across all nine parameter sets.

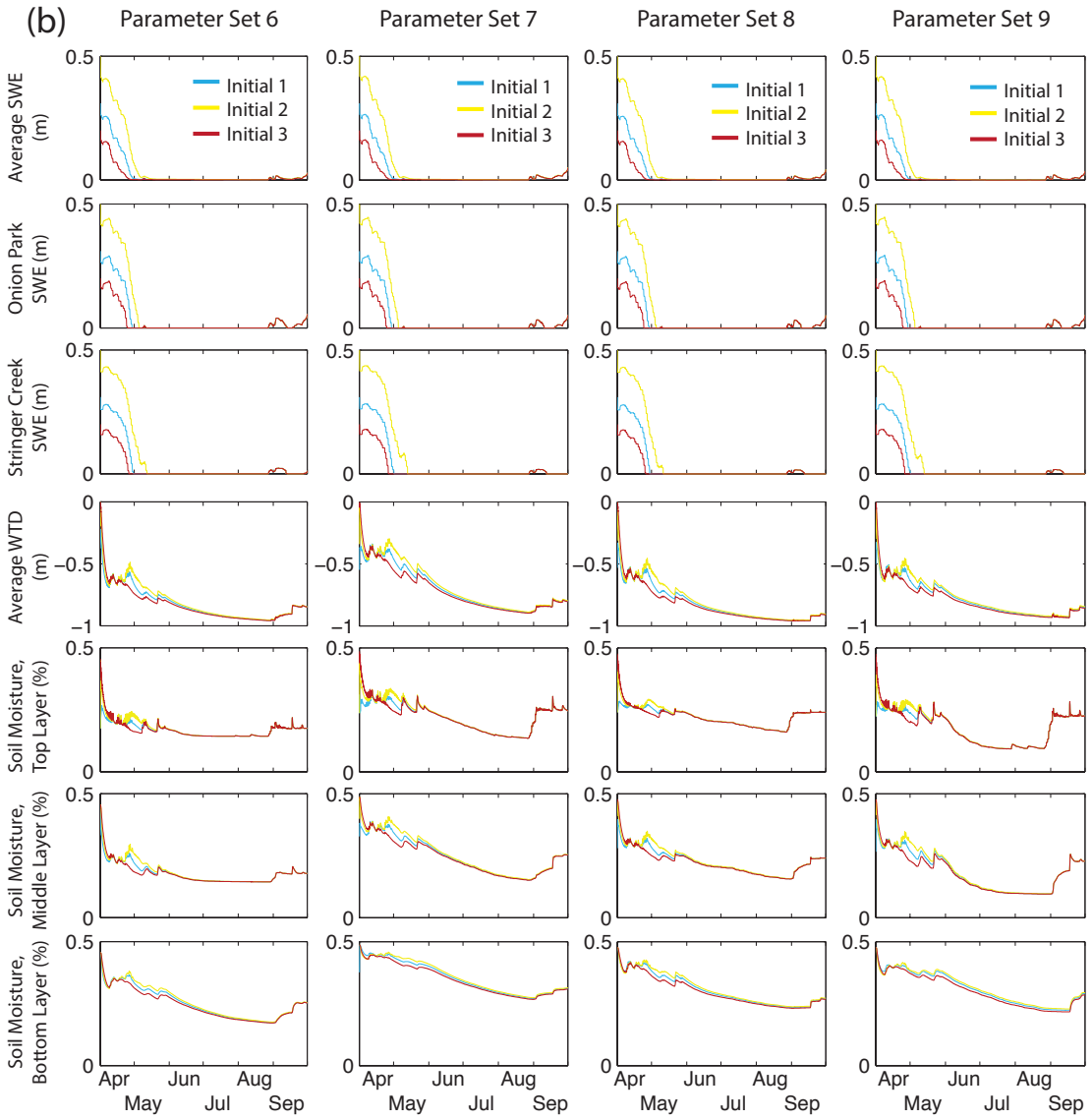

Figure A2: Predictions of annual, May through July, and August through September water table depths (a) in space, across the watershed, (b) shown as differences in space across nine equifinal parameter sets, and (c) with differences on a cell-by-cell basis summarized as histograms per time period. Color on maps indicates (a) the average depth across a given period between bedrock (-1 m) and the surface (0 m) and (b) the largest difference between all nine predictions at each cell. Predictions are shown for three parameter sets that span the range of behaviour from drier (set 1) to average (set 5) to wetter (set 9) conditions across all nine parameter sets.

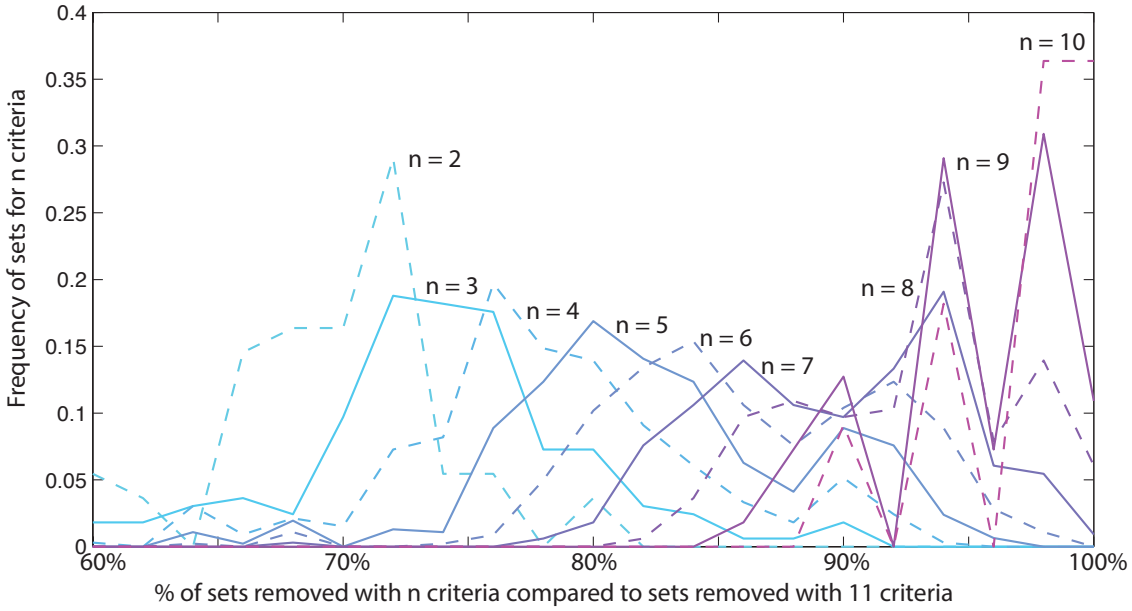

Figure B1: This figure compares the number of nonbehavioural sets removed by subsets of different metrics benchmarked by the number of nonbehavioural sets removed by all eleven criteria. Distributions for a given subset of metrics (n = 2 to n = 10) are shown for all possible combinations of metrics.

5

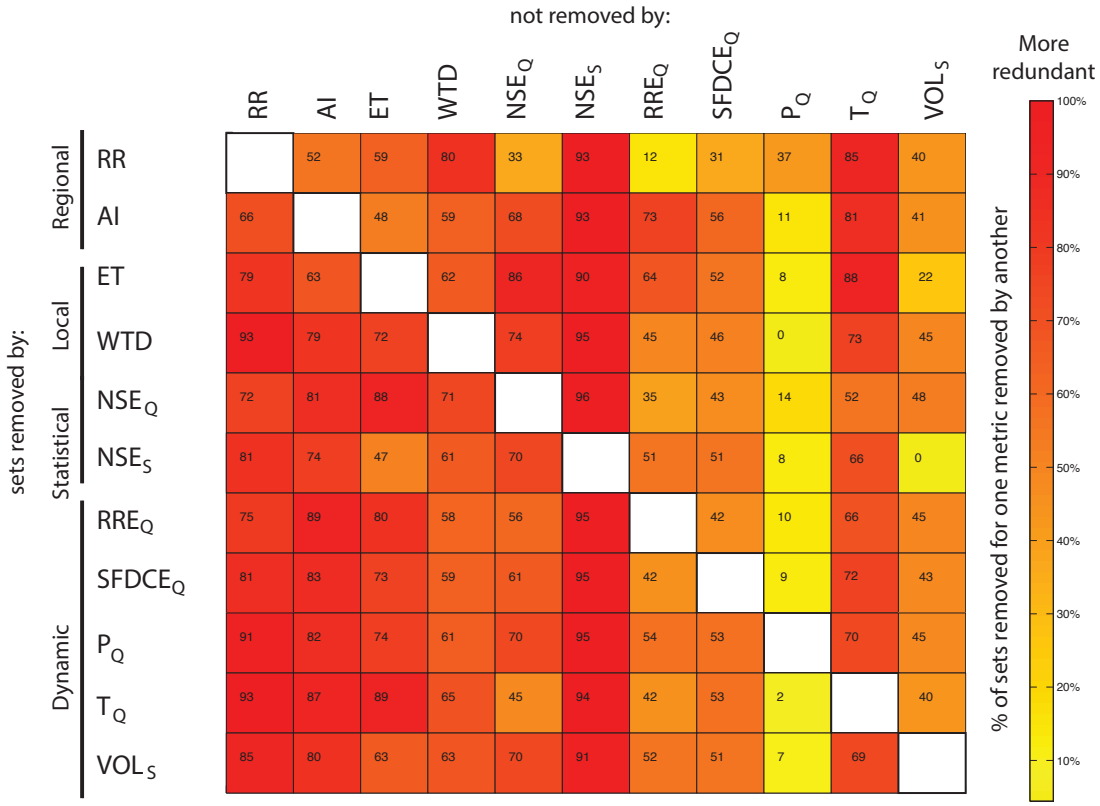

Figure B2: Redundancy across different metrics. The percentage of nonbehavioural sets removed by criteria applied to one metric (x-axis) that is not removed by another (y-axis).

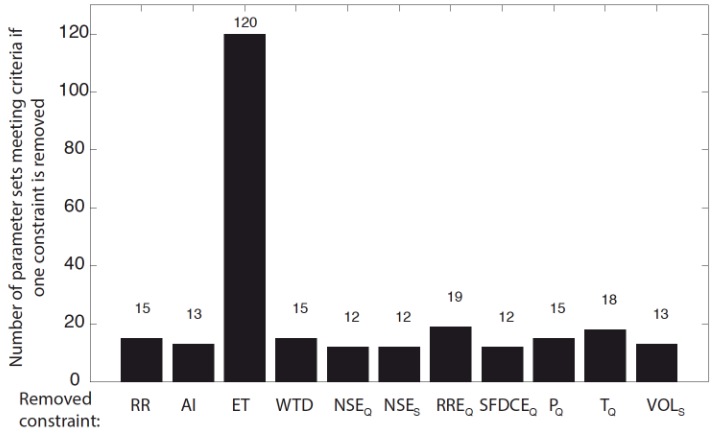

Figure B3: The impact of removing a single constraint on the number of final behavioural parameter sets, organized by the constraint that was removed from the analysis.

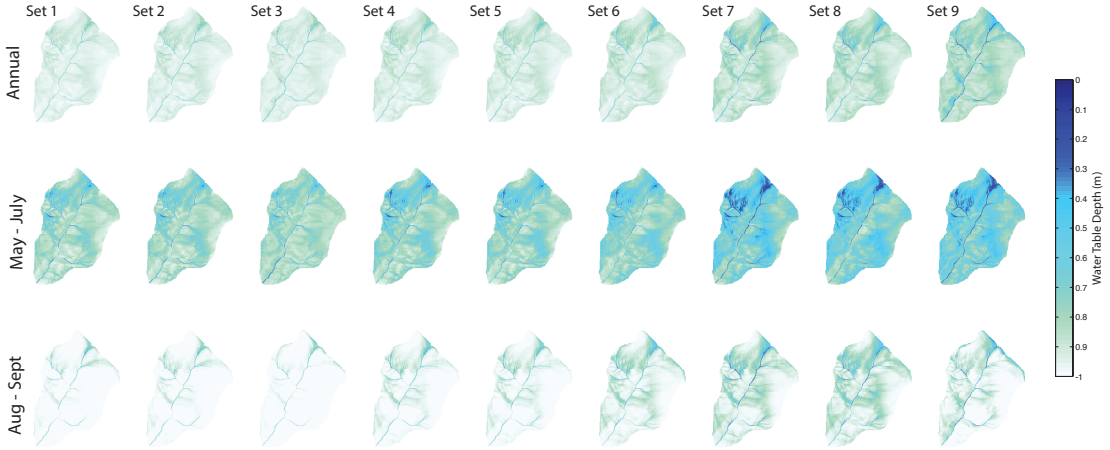

Figure B4: Predictions of annual, May through July, and August through September water table depths. Color indicates average depth across a given period between bedrock (-1 m) and the surface (0 m). Columns display predictions across nine parameter sets identified by the framework applied to Stringer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

Table 1: Parameter types, names, numbers (with reference to Figure 8) and ranges for all 53 parameters included in the analysis. Sources for ranges are detailed in Kelleher et al. (2015).

#	Name	Minimum	Maximum	#	Name	Minimum	Maximum
Snow/climate				***Understory Vegetation***			
1	Snow water capacity [-]	0.01	0.08	27	Minimum resistance [s m^{-1}]	100	175
2	Rain LAI multiplier [-]	0.0002	0.001	28	Moisture threshold [-]	0.115	0.165
3	Snow LAI multiplier [-]	0.0002	0.0025	29	Vapor pressure deficit [Pa]	200	4000
4	Minimum intercepted snow [-]	0.001	0.01	30	Rpc [-]	0.1	1
5	Snow threshold [°C]	0	2	31	Root fraction, layer 1 [-]	0.25	0.45
6	Rain threshold [°C]	−2	0	32	Root fraction, layer 2 [-]	0.45	0.65
Soils				33	Monthly LAI [-]	0.65	1.35
7	Lateral conductivity (KLAT) [m s^{-1}]	1×10−5	1×10−2	34	Monthly albedo [-]	0.1	0.23
8	Exponential decrease in KLAT with depth [-]	0.5	5	***Overstory - Trees***			
				35	Fractional coverage [-]	0.287	0.575
9	Maximum infiltration [m s^{-1}]	3.6×10−5	5.38 × 10−4	36	Trunk space [-]	0.37	0.79
10	Capillary drive [-]	0.03	0.6	37	Aerodynamic attenuation [-]	0.3	4.25
11	Surface albedo [-]	0.2	0.3	38	Radiation attenuation [-]	0.1	0.2
12	Porosity [-]	0.38	0.47	39	Maximum snow interception capacity [-]	0.05	0.2
13	Pore size distribution index [-]	0.07	0.559	40	Mass release drip ratio [-]	0	1
14	Bubbling pressure [m]	0	1.24	41	Snow interception efficiency [-]	0	1
15	Field capacity [-]	0.15	0.25	42	Height [m]	2	7.5
16	Wilting point [-]	0.07	0.15	43	Maximum resistance [s m^{-1}]	500	3000
17	Bulk density [kg m-3]	1390	1650	44	Minimum resistance [s m^{-1}]	150	300
18	Vertical conductivity [m s^{-1}]	1.42×10−10	2.02×10−4	45	Moisture threshold [-]	0.115	0.165
19	Thermal conductivity [W m^{-1} °C^{-1}]	0.3	0.8	46	Vapor pressure deficit [Pa]	200	4000
20	Thermal capacity [J m^{-3} °C^{-1}]	1 × 106	3 × 106	47	Rpc [-]	0.1	1
21	Mannings n [-]	0.11	0.35	48	Root fraction, layer 1 [-]	0.65	0.15
Understory Vegetation				49	Root fraction, layer 2 [-]	0.75	0.25
22	Maximum snow interception capacity [-]	0.05	0.2	50	Monthly LAI [-]	0.5	1.5
23	Mass release drip ratio [-]	0	1	51	Monthly albedo [-]	0.06	0.23
24	Snow interception efficiency [-]	0	1	***Overstory - Tall trees***			
25	Height [m]	0	1.2	52	Fractional coverage [-]	0.322	0.594
26	Maximum resistance [s m^{-1}]	500	1000	53	Height [m]	10	14.7

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 behavioural and nonbehavioural runs. Abbreviations in equations refer to time step *t*, total number of time steps *m*, observed values *O*, and simulated value *S*.

		Signature/ Metric	Abbrev.	Data	Units	Equation	Constraints
Signatures	Regional	Runoff Ratio	RR	Q, P	[-]	$\sum_{t=1}^{m} Q_t / \sum_{t=1}^{m} P_t$	0.2 < RR < 0.7
		Aridity Index	AI	PE, P	[-]	$\sum_{t=1}^{m} PE_t / \sum_{t=1}^{m} P_t$	0.33 < AI < 1.206
	Local	Annual Evapotranspiration	ET	ET	[mm]	$\sum_{t=1}^{m} ET_t$	300mm < ET < 650 mm
		Average Annual Water Table Depth	WT	WT	[m]	$\frac{1}{m}\sum_{t=1}^{m} WT_t$	WT < 0.5m
Error Metrics	Statistical	Nash Sutcliffe Efficiency Coefficient	NSE_Q	Q	[-]	$1 - \sum_{t=1}^{m} (O_t - S_t)^2 / \sum_{t=1}^{m} \left(O_t - \frac{1}{m}\sum_{t=1}^{m} O_t\right)^2$	NSE_Q > 0.6
			NSE_S	SWE	[-]		NSE_S > 0.8
	Dynamic	Runoff Ratio Error	RR_E	Q	[%]	$100 \cdot \left\lvert \frac{\sum_{t=1}^{m} Q_{t,S}}{\sum_{t=1}^{m} P_{t,S}} - \frac{\sum_{t=1}^{m} Q_{t,O}}{\sum_{t=1}^{m} P_{t,O}} \right\rvert \cdot \left\lvert \frac{\sum_{t=1}^{m} P_{t,O}}{\sum_{t=1}^{m} Q_{t,O}} \right\rvert$	$\lvert RR_E \rvert$< 20%
		Error in the Slope of the Flow Duration Curve	$SFDC_E$	Q	[%]	$100 \cdot \left\lvert \frac{Q_{10\%,S} - Q_{30\%,S}}{30-10} - \frac{Q_{10\%,O} - Q_{30\%,O}}{30-10} \right\rvert \cdot \left\lvert \frac{30-10}{Q_{10\%,O} - Q_{30\%,O}} \right\rvert$	$\lvert SFDC_E \rvert$< 30%
		Error in Peak Q Magnitude	P_Q	Q	[%]	$100 \cdot \lvert P_{Q,S} - P_{Q,O} \rvert / P_{Q,O}$	$\lvert P_Q \rvert$ < 35%
		Error in Peak Q Timing	T_Q	Q	[days]	$\lvert T_{Q,S} - T_{Q,O} \rvert$	$\lvert P_T \rvert$ < 12 days
		Error in SWE Volume	VOL_S	SWE	[%]	$100 \cdot \left(\int_{t=1}^{m} S(t)dt - \int_{t=1}^{m} O(t)dt\right) / \int_{t=1}^{m} O(t)dt$	$\lvert VOL_S \rvert$ < 20

Table B1: Error metrics for the nine behavioural sets. Metric abbreviations correspond to metrics shown in Figures 3 and 4.

		LTC					MSC					LSC				
		NSE_Q	RRE_Q	$SFDCE_Q$	P_Q	T_Q	NSE_Q	RRE_Q	$SFDCE_Q$	P_Q	T_Q	NSE_Q	RRE_Q	$SFDCE_Q$	P_Q	T_Q
		[-]	[%]	[%]	[%]	[days]	[-]	[%]	[%]	[%]	[days]	[-]	[%]	[%]	[%]	[days]
Water Year 2008	Run 1	0.80	14.36	32.20	-14.8	-12.25	0.54	31.32	2.57	-16.35	7.50	0.68	-3.02	8.19	-29.02	7.50
	Run 2	0.83	1.87	26.19	-19.3	-13.00	0.60	13.53	-5.19	5.73	7.50	0.70	-16.12	-2.43	-22.46	7.50
	Run 3	0.80	34.63	17.85	-10.7	-11.50	0.56	44.15	0.95	-3.52	6.63	0.72	12.42	-6.20	-22.45	6.63
	Run 4	0.63	-7.38	-0.19	-159.7	8.75	0.47	25.17	-7.88	0.54	8.75	0.63	-7.38	-1.60	-21.84	8.75
	Run 5	0.76	3.76	-0.20	-25.6	-11.63	0.72	29.12	-13.96	-9.62	-11.63	0.76	3.76	-25.56	-26.61	-11.63
	Run 6	0.72	0.87	-7.65	-22.1	7.75	0.56	36.60	-11.54	6.94	4.63	0.72	0.87	-7.65	-23.49	7.75
	Run 7	0.79	-4.48	-18.41	-21.2	7.75	0.74	29.55	-18.67	-8.34	-12.00	0.79	-4.48	-18.41	-29.69	7.75
	Run 8	0.74	-6.86	-3.72	-20.3	9.13	0.63	26.97	-7.80	-3.71	8.63	0.74	-6.86	-3.72	-28.19	9.13
	Run 9	0.78	-14.41	3.47	-19.7	4.63	0.67	16.90	3.61	18.25	4.63	0.78	-14.41	3.47	-17.41	4.63
	Minimum	0.63	-14.41	-18.41	-159.7	-13.00	0.47	13.53	-18.67	-16.35	-12.00	0.63	-16.12	-25.56	-29.69	-11.63
	Average	0.76	2.48	5.50	-34.8	-1.15	0.61	28.15	-6.43	-1.12	2.74	0.72	-3.91	-5.99	-24.58	5.33
	Maximum	0.83	34.63	32.20	-10.7	9.13	0.74	44.15	3.61	18.25	8.75	0.79	12.42	8.19	-17.41	9.13
Water Year 2007	Run 1	0.80	23.95	-0.55	-1.18	-0.50	0.60	42.90	-33.27	17.89	1.88	0.77	9.28	-22.17	-2.34	1.88
	Run 2	0.91	10.79	13.06	-21.14	-1.50	0.69	25.21	-30.73	-3.69	4.63	0.76	-3.92	-17.57	-19.90	8.38
	Run 3	0.64	48.94	-6.01	-5.04	-4.13	0.58	59.92	-35.76	14.30	6.38	0.78	34.54	-31.08	-8.78	-1.75
	Run 4	0.85	18.21	-27.78	-26.55	-1.00	0.61	34.31	-64.39	-19.75	9.00	0.65	3.32	-62.07	-36.58	9.00
	Run 5	0.78	48.37	-9.65	-18.38	-18.00	0.78	54.61	-35.83	-6.32	2.00	0.82	35.42	-30.08	-26.02	0.38
	Run 6	0.75	30.18	14.63	8.49	-2.38	0.51	49.05	-28.69	28.23	2.38	0.75	14.29	-15.48	-1.14	0.25
	Run 7	0.80	24.85	11.06	-5.44	-3.50	0.71	39.99	-14.98	7.89	-3.00	0.82	7.45	0.06	-13.26	1.38
	Run 8	0.75	21.94	39.75	5.14	0.50	0.54	38.43	-4.29	32.23	3.75	0.75	5.48	14.82	5.35	3.75
	Run 9	0.83	11.05	41.99	6.45	0.88	0.61	25.69	-10.32	22.52	2.63	0.75	-4.33	9.20	2.95	2.63
	Minimum	0.64	10.79	-27.78	-26.55	-18.00	0.51	25.21	-64.39	-19.75	-3.00	0.65	-4.33	-62.07	-36.58	-1.75
	Average	0.79	26.48	8.50	-6.40	-3.29	0.63	41.12	-28.69	10.37	3.29	0.76	11.28	-17.15	-11.08	2.88
	Maximum	0.91	48.94	41.99	8.49	0.88	0.78	59.92	-4.29	32.23	9.00	0.82	35.42	14.82	5.35	9.00