Moving towards a unified threshold-based hydrological theory through inter-comparison and modelling

by

Cody A. Ross

A thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Department of Geological Sciences

University of Manitoba

Winnipeg, Manitoba, Canada

ABSTRACT

Numerous studies have ascertained detailed descriptions of hydrological processes for unique hillslopes and catchments. These efforts have contributed tremendously to our understanding of hydrologic processes and have facilitated hydrological modelling. These individual studies have also revealed considerable spatial and process heterogeneity. Recently, thresholds in runoff response have been identified as a potential foundation for a new unified hydrological theory, as thresholds are emergent properties that integrate both spatial and process heterogeneity. However, our current understanding of threshold behaviour has limitations. Most threshold research in hydrology has been conducted for humid temperate environments with a focus on storage thresholds. The ubiquity of threshold behaviour across a wider range of environments is unknown, and the degree to which thresholds of rainfall intensity or hydrologic abstraction caused by evapotranspiration affect different aspects of hydrologic response has not been thoroughly assessed. Additionally, the potential benefits of incorporating threshold information in rainfall-runoff model evaluation have not been determined. Our understanding of threshold behaviour has also been guided by conceptualizations featuring response as a function of a single meteorological factor, which is at odds with the growing body of work that has demonstrated a range of controls on hydrologic response. The overarching goal of this thesis was, therefore, to address these knowledge gaps and to contribute to a unified threshold-based hydrological theory. First, analyses were completed to assess the spatial and temporal variability in rainfall-runoff event dynamics and the influence of fixed and dynamic controls on hydrologic response for twenty-one sites across seven diverse study areas. As part of this effort, 1,641 rainfall-runoff events were delineated and characterized using an unprecedented suite of response metrics and meteorological factors. This analysis showed considerable temporal variability in rainfall-runoff

event response and illustrated the influence of intensity-driven processes on hydrologic response. Second, the suite of response metrics and meteorological factors was used to evaluate thresholds in runoff response. Threshold behaviour was observed at twenty out of twenty-one sites considered and in addition to rainfall depth, rainfall intensity and hydrologic abstractions caused by evapotranspiration proved to be important controls. Third, the potential benefits related to constraining rainfall-runoff model outputs through model evaluation using multiple hydrologic descriptors, including thresholds, was appraised. This work showed that threshold information can be powerful for identifying model simulations that adequately predict real-world processes. Lastly, the simultaneous influence of multiple, potentially interacting, meteorological factors on response thresholds was assessed using techniques borrowed from other disciplines. The results of this analysis showed that meteorological factor interactions can lead to complex and highly nonlinear hydrological responses that cannot otherwise be observed. These findings challenge hydrologists to consider a wider range of threshold behaviours and to evaluate response nonlinearities as a function of multiple, potentially interacting meteorological factors. This thesis led to many interesting results that are encouraging for the development of a unified threshold-based hydrologic theory. Through the synthesis of this work, remaining knowledge gaps were identified that require insight from the hydrology community to advance this theory.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my thesis advisors Dr. Genevieve Ali (University of Guelph and University of Manitoba) and Dr. Christopher Spence (Environment and Climate Change Canada and University of Manitoba). I am greatly appreciative of the opportunity to work with two accomplished Hydrologists who have personally provided me with indispensable support and autonomy throughout my research. Dr. Ali and Dr. Spence, along with the other members of my advisory committee, Dr. David Lobb (University of Manitoba) and Dr. Ian Ferguson (University of Manitoba), offered me valuable guidance and commentary that contributed greatly to this thesis and my future.

This thesis would not have been possible without the data that was made available by multiple researchers and organizations. Specifically: the New Zealand Institute of Water and Atmospheric Research, Hilary McMillan, Ross Woods, Andrew Western, and Kathy Walters who assisted greatly in the acquisition of data for the Mahurangi River Catchment; Francois Courchesne, André Roy and Marie-Claude Turmel for allowing the use of data for the Hermine catchment, and Claire Oswald for providing data from the IISD Experiment Lakes Area. Data were also provided by the H.J. Andrews Experimental Forest and Long Term Ecological Research program administered cooperatively by the USDA Forest Service Pacific Northwest Research Station, Oregon State University, and the Willamette National Forest.

I would also like to thank my fellow graduate students and collaborators who have contributed to my research and have provided inspiration as they pursue their endeavours. Finally, I am extremely appreciative of my parents, partner, and friends who have provided me with support and encouragement throughout my education. Their involvement made this accomplishment possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

CHAPTER 1. INTRODUCTION

1.1 Threshold behaviour in runoff response

A fundamental objective in hydrology is to reach a better understanding of catchment response to meteorological inputs (Davie, 2008; Dingman, 2015; Gregory & Walling, 1973). This objective is motivated by the possibility of better process conceptualizations and by social implications related to water supply, flood forecasting, and water quality. Hillslopes and catchments exhibit tremendous heterogeneity in physical properties and hydrologic response (McDonnell et al., 2007; Sivapalan, 2006). Hydrologic research focused on catchment response to meteorological inputs has advanced our understanding of catchment function and has led to the development of numerous hydrologic models (Beven, 2011). However, these efforts have focused primarily on fine-scale process descriptions for individual hillslopes and catchments. The uniqueness of place (Beven, 2011) inherent to isolated studies and the absence of a unifying theory to connect or compare disparate process descriptions have led to the development of increasingly parameterized models intended to represent vastly heterogeneous landscapes (Kirchner, 2006; McDonnell et al., 2007; Sivapalan et al., 2002). The limited transferability of some findings from individual studies has motivated a shift in focus towards emergent properties, i.e., properties that cannot be predicted from individual landscape components but are representative of landscape heterogeneity and process complexity (Ali et al., 2013; Sivapalan, 2006; Sivapalan et al., 2002; Spence, 2010). Notably, thresholds in runoff processes have been identified as an emergent property that could support a paradigm shift in how catchment responses are characterized and compared (Ali et al., 2013; Spence, 2010).

1.1.1 Variability and threshold behaviour in hillslope and catchment response

Abundant heterogeneities in hydrologic systems result in hydrologic response characterized by significant spatial and temporal variability (Beven et al., 1988; McDonnell, 2013; McDonnell et al., 2007; Sivapalan, 2006; Sivapalan et al., 2002). Catchment morphology and hydrological processes are coupled with geomorphic processes (Beven et al., 1988) and geomorphic systems are typically nonlinear and threshold-mediated (Phillips, 2006). In hydrology, thresholds in runoff processes are defined as a critical moment in time or point in space at which runoff behaviour rapidly changes (Ali et al., 2013; Phillips, 2006). In the available literature, highly variable and threshold-mediated hydrologic responses have been shown for a variety of processes, scales, and environments. More specifically, these studies have shown that hydrologic response at the hillslope and catchment scales are sensitive to critical values of total precipitation, soil moisture storage, and water table elevations (Ali et al., 2015; Detty & McGuire, 2010; Lehmann et al., 2007; Mosley, 1979).

Rainfall-induced subsurface stormflow is a dominant runoff generation mechanism in many steep, forested headwater catchments and it has been the focus of most threshold-related research (Graham et al., 2010; Graham & McDonnell, 2010; Lehmann et al., 2007; Weiler, 2005). In these studies, critical volumes or depths of precipitation above which runoff generation was initiated were reported: Mosley (1979) and Tani (1997) reported 20 mm rainfall depth thresholds near Reefton, New Zealand and Okayama, Japan, respectively; Tromp-van Meerveld & McDonnell (2006) reported a 55 mm rainfall depth threshold in Georgia, United States, and Whipkey (1965) reported a 50 mm precipitation depth threshold in Ohio, United States. Similarly, Redding & Devito (2008) observed precipitation depth thresholds between 20 and

78 mm in Alberta, Canada and Sidle et al. (2000) observed precipitation depth thresholds ranging between 40 and 150 mm in Chiba, Japan. Also, hillslope response controlled by thresholds of antecedent soil moisture and depth to water table has been observed in steep, forested headwater catchments. For example, in New Hampshire, United States, Detty & McGuire (2010) observed a nonlinear quickflow response when a 316 mm combined antecedent soil moisture index (ASI) and gross precipitation threshold was exceeded. They also observed a significant increase in streamflow once a 200 to 500 mm depth to water table threshold was satisfied (Detty & McGuire, 2010). In Quebec, Canada, James & Roulet (2007) reported threshold behaviour in storm response when a mean antecedent soil moisture value of 0.24 vol/vol was reached, and Ali et al. (2011) reported threshold behaviour between catchment response and perched groundwater levels.

While most research on hydrologic thresholds has been conducted at temperate or humid forested hillslopes and catchments (e.g., Ali et al., 2011; Detty & McGuire, 2010; Graham & McDonnell, 2010, 2010; James & Roulet, 2007; Lehmann et al., 2007; McGlynn & McDonnell, 2003; Mosley, 1979; Sidle et al., 2000; Tani, 1997; Tromp-van Meerveld & McDonnell, 2006; Whipkey, 1965), threshold behaviour has also been observed for a variety of other environments, including arid and semi-arid environments (e.g., Cammeraat, 2002; Reaney et al., 2007). Threshold behaviour in rainfall-runoff relationships has been reported for semi-arid environments in Spain: Cammeraat (2002) found that catchment response required the exceedance of both a rainfall intensity threshold (4.2 mm/10 min) and a total rainfall depth threshold (30 mm). Similarly, Reaney et al. (2007) reported that the exceedance of an infiltration-excess threshold (84 mm/hr) was required to form a connection between surface runoff generating points and the catchment outlet. In prairie pothole landscapes that are

characterized by topographic depressions from historical glaciation, runoff events occur in response to the exceedance of storage thresholds (Shaw et al., 2012; Stichling & Blackwell, 1957). These thresholds are associated with fill-and-spill dynamics resulting from the exceedance of depression storage that leads to runoff generation and the connectivity of adjacent depressions. Measurement of specific storage thresholds at the catchment scale is complicated in these circumstances, as the area contributing to runoff is dynamic and depressions are not always simultaneously connected during runoff events (Shaw et al., 2012, 2013). Nevertheless, runoff processes in these environments are threshold mediated and are heavily influenced by antecedent conditions, hydrologic abstraction, and input event frequency and duration (Shaw et al., 2012). In high relief landscapes, nonlinear relationships have been observed between subsurface flow and water table elevation. When examining the relationship between throughflow response and catchment streamflow response in a temperate, forested mountain catchment, Kim et al. (2004) found that the exceedance of a 620 mm water table elevation threshold was required for throughflow and streamflow response to occur. Threshold behaviour has also been observed in snow and permafrost dominated catchments. In the sub-arctic Canadian shield, the exceedance of lake storage thresholds is often required for outflow to occur (Mielko & Woo, 2006; Spence et al., 2010). Mielko & Woo (2006) showed that lake storage needed to be satisfied before outflow occurred, and that hydrologic abstraction caused by evaporation prolonged the time before the storage threshold was met, indicating a system controlled by dynamic storage and multiple processes.

There is tremendous variability among documented hydrologic thresholds, which has been associated with unique local characteristics such as slope and topography, hydraulic conductivity of the vadose zone, bedrock permeability, soil depth, and soil macropores

(Lehmann et al., 2007). These examples of thresholds for hillslope and catchment response illustrate that threshold behaviour is present not only across a range of environments but also across multiple runoff processes. Thresholds have been observed for surface and subsurface saturation excess flow (Ali et al., 2011; Detty & McGuire, 2010; Graham et al., 2010; Graham & McDonnell, 2010; James & Roulet, 2007; Lehmann et al., 2007; McGlynn & McDonnell, 2003; Mosley, 1979; Sidle et al., 2000; Tani, 1997; Tromp-van Meerveld & McDonnell, 2006a; Whipkey, 1965) and infiltration excess flow (Cammeraat, 2002; Reaney et al., 2007; Shaw et al., 2012). While there are similarities in the conceptualization of these processes (McDonnell, 2013), thresholds related to storage (e.g., rainfall volume or depth, soil moisture, water table depth) and rainfall intensity are typically associated with saturation excess flow and infiltration excess flow processes, respectively (Ali et al., 2015). It is clear that most research on thresholds has focused on storage thresholds, while fewer studies have focused on rainfall intensity thresholds, the influence of hydrologic abstraction on threshold behaviour, or how interactions between different processes may affect or lead to emergent threshold behaviour (Ali et al., 2015). The consideration of thresholds other than storage threshold and potential controls on threshold behaviour may be required to adequately characterize hydrologic response, especially in environments where infiltration is limited (Ali et al., 2015). These considerations are likely context-dependent: some studies conducted in humid, temperate forested catchments with high soil infiltration capacity have suggested that rainfall intensity exerts little influence on runoff generation (Graham & McDonnell, 2010; Tromp-van Meerveld & McDonnell, 2006a). In contrast, rainfall and snowmelt intensity thresholds have a more significant effect on runoff processes in environments with limited infiltration capacity. Findings from the Sleepers River Research Watershed in Vermont, United States (Shanley & Chalmers, 1999) and the boreal

Kryckland catchment in northern Sweden (Laudon et al., 2007) suggest that snowmelt intensity significantly influences runoff processes in soils with reduced infiltration capacity from frost development. Likewise, runoff generation mechanisms like pipeflow appear to be governed by macropore storage thresholds, which are influenced by water table levels and the exceedance of soil matric infiltration capacity (Spence, 2010). To properly characterize threshold behaviour in these instances, when the rate of water delivery to macropores is higher than the surrounding soil matrix, it has been suggested that thresholds in the rate of water delivery become acutely important (McDonnell, 1990). These observations are evidence of dynamic systems of multiple highly variable and threshold-mediated processes. However, **it is unclear when variables that quantify different processes should be considered in the evaluation of response variability and threshold behaviour.**

1.2 Identifying threshold behaviour

A variety of different hydrometric and meteorological data types have been used towards the identification of threshold behaviour in runoff response. Threshold mediated relationships previously reported in the literature have mostly involved variables that quantify rainfall and discharge data, primarily because of their availability in most hydrological studies and since threshold behaviour is often observed incidentally rather than sought after (Ali et al., 2013). Following a targeted literature review (n = 29 papers), roughly 63% of related studies considered precipitation thresholds for catchment discharge. Conversely, water table, soil moisture, and water chemistry data appeared less frequently in these studies: ~30%, 48%, and <5%, respectively, and were typically considered in addition to precipitation-discharge data. Regularly,

scatter plots are constructed from observation data to examine the relationship between hydrologic response and measures of precipitation (e.g., Biron et al., 1999; Cammeraat, 2002; Carey et al., 2010; Detty & McGuire, 2010; Freer et al., 2002; Graham et al., 2010; Graham & McDonnell, 2010; James & Roulet, 2007; Kim et al., 2004; Laudon et al., 2007; Lehmann et al., 2007; Mielko & Woo, 2006; Mosley, 1979; Oswald et al., 2011; Redding & Devito, 2008; Scaife & Band, 2017; Shaw et al., 2012; Sidle et al., 2000; Sivakumar, 2005; Spence, 2007; Tani, 1997; Tromp-van Meerveld & McDonnell, 2006a, 2006b; Wei et al., 2020; Weiler, 2005; Whipkey, 1965), soil moisture (e.g.,Cammeraat, 2002; Detty & McGuire, 2010; Freer et al., 2002; Graham et al., 2010; Graham & McDonnell, 2010; James & Roulet, 2007; Laudon et al., 2007; Lehmann et al., 2007; Mosley, 1979; Redding & Devito, 2008; Scaife & Band, 2017; Tromp-van Meerveld & McDonnell, 2006a, 2006b; Wei et al., 2020; Weiler, 2005; Whipkey, 1965), and water table data (e.g., Cammeraat, 2002; Detty & McGuire, 2010; Freer et al., 2002; James & Roulet, 2007; Kim et al., 2004; Lehmann et al., 2007; Mielko & Woo, 2006; Oswald et al., 2011). Observation data can be used directly to examine these relationships or it can be analyzed or aggregated over a specific period (e.g., rainfall-runoff event) to derive input meteorological factors (e.g., total event rainfall, rainfall intensity, potential evapotranspiration) and output hydrologic response metrics (e.g., peak discharge, runoff ratio) to assess relationships between meteorological inputs and response. Assessing data aggregated to the rainfall-runoff event scale is particularly useful, as event-based response dynamics are of interest for understanding hydrologic behaviour and since such aggregation mitigates challenges associated with interpreting continuous high-frequency data that may be characterized by diurnal patterns and autocorrelation. **There is significant variability in the type and form of observation data used to evaluate the**

presence of hydrologic thresholds and little is known about how interpretations of threshold behaviour might be influenced by the specific type or form of data used.

Observation data are typically assessed for threshold behaviour visually through the evaluation of scatter plots showing independent input (e.g., total precipitation) and dependent output (e.g., discharge) variables (e.g., Ali et al., 2011; Detty & McGuire, 2010; Graham et al., 2010; James & Roulet, 2007; Lehmann et al., 2007; McGlynn & McDonnell, 2003; Mosley, 1979; Sidle et al., 2000; Tani, 1997; Tromp-van Meerveld & McDonnell, 2006a; Whipkey, 1965). Visual threshold identification is associated with methodological difficulties: while studies with few data points or dominant hillslope scale dynamics may show clear threshold behaviour in input-output relationships, the identification process can be complicated by clustering around the point of inflection, which suggests a range of possible threshold values (Oswald et al., 2011; Scaife & Band, 2017). The identification process is further complicated by the fact that nonlinear hydrologic behaviour manifests in a variety of shapes that can be represented by different mathematical functions, including the hockey-stick shape (Detty & McGuire, 2010; Tromp-van Meerveld & McDonnell, 2006a; Weiler, 2005), the Heaviside or step function (James & Roulet, 2007), the Dirac function, and the sigmoid function (Zehe & Blöschl, 2004) (Figure 1-1). While these distinct threshold shapes have been reported in a variety of studies, process-based rationales for different shapes have yet to be offered (Ali et al., 2015). To overcome limitations of visual identification methods (i.e., user bias and clustering) and to account for different threshold shapes, alternative techniques for threshold identification and characterization have been proposed. Identification of hockey-stick thresholds has been automated and made more objective through piecewise linear regression analysis (Oswald et al., 2011; Scaife & Band, 2017) and the approximation of threshold changes using exponential

functions (Oswald et al., 2011). Others have suggested the use of statistical indices that quantify the abruptness of nonlinear response changes for threshold detection and comparison (Lintz et al., 2011). For relationships characterized by multiple nonlinear response changes, domain-dependent mathematical functions have been recommended over different ranges of input values (Ali et al., 2011).

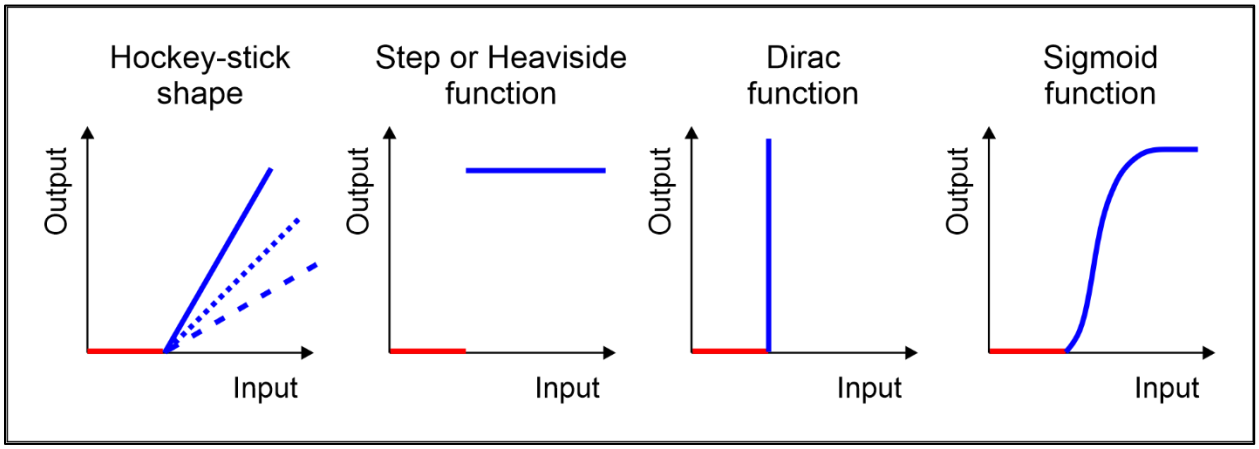


Figure 1-1. Models of nonlinear relationship shapes reported in threshold-related hydrologic literature. Red and blue segments of the curve signify runoff behaviour before and after the threshold respectively. Used with permission of John Wiley & Sons, from Ali *et al.* (2013); permission conveyed through Copyright Clearance Center, Inc.

Besides the type and form of observation data and the method used to assess possible threshold behaviour, other factors like record frequency and duration may influence how nonlinearities in input-output relationships are interpreted. More information can be extracted from high-frequency observation data relative to data averaged over weeks, months, or seasons (Ross et al., 2015). While the existing literature showing threshold behaviour has used data varying in frequency and duration, a benchmark of record length and resolution has yet to be

established for threshold detection. Other studies focused on temporally dynamic hydrologic processes have leaned towards the use of high-frequency (i.e., sub-daily) data, which is characteristic of datasets from iconic research basins like the Panola Mountain Research Watershed (Tromp-van Meerveld et al., 2008), H.J. Andrews Experimental Forest (McKee & Druliner, 1998), and the Tarrawarra catchment (Western & Grayson, 1998).

1.3 Cross-site and cross-scale synthesis

Hillslopes and small catchments are a common unit of hydrologic inquiry (Beven et al., 1988; Dingman, 2015; Kirchner, 2009) and threshold behaviour has mostly been examined at these scales (Tani, 1997). This is primarily because hillslopes are fundamental landscape units for understanding runoff generation processes (Tromp-van Meerveld & McDonnell, 2006b), and because most high-frequency precipitation and hydrometric data are available at these scales (Ali et al., 2015). While most threshold-based research has been conducted at the hillslope and small catchment scales, threshold behaviour has also been identified at larger scales and as a part of hierarchical processes, where the output from one process becomes the input for another (Ali et al., 2013; Lehmann et al., 2007). **Little is known about how threshold behaviour in runoff response might be influenced by scale.**

There have only been a small number of studies that have focused on the identification and comparison of threshold information between different study areas. Research that has evaluated threshold behaviour in multiple environments has focused on the identification of common physical characteristics among study areas that exhibit threshold runoff response (e.g., Carey et al., 2010; Uchida et al., 2005). For example, Ali et al. (2015) compared threshold

behaviour among catchments and identified correlations between precipitation threshold values and catchment physiographic characteristics. These findings reveal an opportunity to develop regression equations with the intent of predicting precipitation threshold values for other catchments. **There remain opportunities for robust comparisons of threshold information between study areas.**

1.4 Physically realistic hydrologic models and threshold behaviour

Hydrologists continue to advance the quest for physically realistic hydrologic models (Clark et al., 2017). Model realism can be assessed in terms of how well model outputs agree with observed behaviour and the processes that are explicitly represented within a model (Beven, 2002, 2011; Clark et al., 2017). The physical realism of unique parameter sets with a comparable performance from a statistical standpoint can be evaluated and constrained based on how well different aspects of hydrologic behaviour are reproduced. Physical realism constraints have been imposed on a variety of rainfall-runoff models using data like soil moisture, water chemistry, isotope compositions, and groundwater levels (Ala-aho et al., 2017; Kelleher et al., 2017; Seibert & McDonnell, 2002; Son & Sivapalan, 2007; Stadnyk et al., 2013). These modelling exercises have allowed for the development of models with increased fidelity (i.e., the degree to which a model simulation reproduces hydrologic processes observed in nature) and have helped constrain equifinality - i.e., multiple unique parameter sets may lead to simulations with a comparable model fit (Beven, 2006, 2011). **Threshold information has not been used in this way towards model evaluation, and potential benefits for parameter selection and model performance are unknown**.

An enduring challenge faced in environmental modelling is accurately predicting response nonlinearities (Beven, 2002). Threshold behaviour has been observed for a variety of processes (e.g., infiltration) and has been associated with a range of physical hillslope and catchment characteristics (e.g., soil hydraulic conductivity and soil depth) (Ali et al., 2013, 2015; Lehmann et al., 2007; Spence, 2007; Spence et al., 2010; Spence, 2010). These processes and characteristics are represented in the equations and parameters of many of the hydrological models that are currently available. However, emergent properties, like threshold behaviour in runoff response, cannot be predicted from individual landscape components (Lehmann et al., 2007; McDonnell et al., 2007). It has been suggested that controls of hydrologic response might interact or be hierarchical in terms of their relative importance for determining response behaviour (Buttle, 2006; Devito et al., 2005; Merz & Blöschl, 2009; Yadav et al., 2007). Relationships and/or interactions between different processes and control factors that may lead to the emergence of thresholds in runoff response are not fully understood and have not been incorporated into hydrological models. Only threshold mediated hydrologic responses involving a single control factor have been considered, except for select studies that qualitatively assess the dual influence of rainfall amount and rainfall intensity on the response (Cammeraat, 2002; Scaife et al., 2020; Scaife & Band, 2017). This differs from some other disciplines, like ecology, where threshold mediated system response has been modelled as a function of multiple, potentially interacting control factors (Andersen et al., 2009; Kinzig et al., 2006; Limburg et al., 2002; Lintz et al., 2011). **A better understanding of the simultaneous effects of multiple control factors on hydrologic thresholds is needed to inform the development of models that more explicitly consider threshold information.**

1.5 Thesis research objectives

A new paradigm in hydrology has been proposed that includes a shift in focus to emergent properties (Ali et al., 2013; Bonell, 1993; Lehmann et al., 2007; Spence, 2010; Weiler, 2005; Zehe et al., 2005). Increasingly, hydrologic thresholds are described as catchment hydrological signatures (Spence, 2007) that are of significant utility for catchment comparison and grouping of similar hydrological responses (Ali et al., 2013). Limitations remain in our theoretical and operational understanding of threshold information. In addition to challenges related to the type and form of data used for threshold identification and the comparison of threshold information across study areas and scales, challenges for incorporating threshold information in hydrologic models are unknown. It is unclear whether the identification of high-fidelity model simulations can be improved through the inclusion of threshold information. Therefore, this thesis intends to address these knowledge gaps and contribute to a unified threshold-based hydrological theory aimed at enhancing understanding of relationships between meteorological inputs and hydrologic response across a range of scales and environments. This research is guided by four research objectives, specifically:

(1) Assess the spatial and temporal variability in rainfall-runoff event dynamics and the influence of fixed (e.g., topography) and dynamic (e.g., climate) controls on hydrologic response across a range of scales and environments.

(2) Evaluate the ubiquity of threshold behaviour in rainfall-runoff event response, with a special focus on rainfall depth thresholds, rainfall intensity thresholds, and thresholds in hydrologic abstraction from evapotranspiration.

(3) Appraise the potential benefits of constraining rainfall-runoff model outputs using multiple hydrologic descriptors, including thresholds.

(4) Characterize the simultaneous influence of multiple, potentially interacting, meteorological factors on threshold mediated hydrologic response.

1.6 Study areas

This study was carried out using existing data for one experimental hillslope and twenty catchments from seven distinct study areas. These study sites vary significantly in scale and are characterized by distinct climate, topography, geology, soil properties, and land cover. Each study site has been the focus of previous hydrologic research and has field data available with several documented variables (e.g., discharge, precipitation, soil moisture, water table position, geochemistry, and elevation) across various observation frequencies and durations. This study focuses on topographic, hydrometric (i.e., discharge), and meteorological (e.g., rainfall, temperature, and evapotranspiration) data. For each site, discharge data are assumed to represent total site discharge and meteorological data are assumed to be representative of conditions across the site. The study areas considered include the Panola Mountain Research Watershed experimental hillslope (Georgia, USA), the Hermine catchment (Quebec, Canada), the IISD-ELA Lake 658 UP1 catchment (Ontario, Canada), the Tarrawarra catchment (Victoria, Australia), the Catfish Creek watershed (Manitoba, Canada), eight sub-catchments of the Mahurangi River catchment (New Zealand) and eight nested catchments of the H.J. Andrews Experimental Forest (Oregon, United States).

1.6.1 The Panola Mountain Research Watershed experimental hillslope

The Panola Mountain Research Watershed experimental hillslope (PMRW) is in the southern Piedmont approximately 25 km southeast of Atlanta, Georgia, United States (Figure 1-2). The PMRW features a 20 m trench at the base of a relatively planar (~14 m relief) 50 m hillslope (Tromp-van Meerveld et al., 2008). The hillslope is covered by predominantly deciduous forest, hillslope soils are sand-loam ranging in depth from 0 to 1.86 m and are underlain by Panola granite with irregular subsurface topography (Freer et al., 2002; Higgins et al., 1988; Tromp-van Meerveld et al., 2008). The climate is humid continental to sub-tropical with a mean annual temperature of 15.2 °C and a mean annual precipitation of 1220 mm (Tromp-van Meerveld et al., 2008). The data used in this study include a 1x1 m digital elevation model (DEM) derived from a total station survey, high-frequency (15-minute) stormflow measured at the hillslope trench, air temperature, and rainfall records combined from tipping-bucket, weighing-bucket, and standard rain gauges (Tromp-van Meerveld et al., 2008).

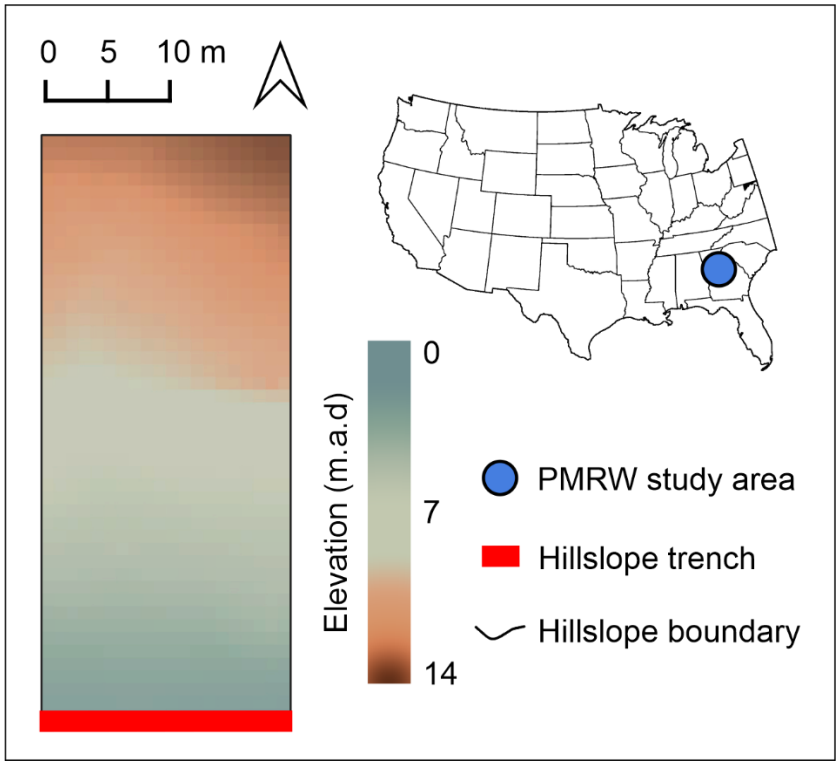


Figure 1-2. Digital elevation model of the Panola Mountain Research Watershed experimental trenched hillslope (PMRW) and the hillslope's location within the United States. The elevation is shown in meters above an arbitrary datum (m.a.d) located at the trench base

1.6.2 The Hermine catchment

The Hermine catchment (HRM) is a 0.05 km^2 headwater forested catchment in the Laurentians, located approximately 80 km north of Montreal, Quebec, Canada (Figure 1-3). The area is characterized by a cool temperate climate with a mean annual temperature of 6.9 °C and a mean annual precipitation of ~1150 mm, 30 % of which falls as snow (Biron et al., 1999). The catchment has 31 m of relief and is drained by an ephemeral stream. Soils of the HRM are typically 1 to 2 m deep bouldery Podzols developed over a bouldery glacial till with a confining layer at a depth of 50 to 75 cm. The forest canopy is dominated by sugar maple and other

deciduous tree species (Ali & Roy, 2010a). Data of the HRM used in the current study include a 2x2 m DEM derived from a manual topographic survey, and high-frequency (15-minute) catchment discharge, precipitation, and other meteorological measurements collected between 2006 and 2008 (Ali & Roy, 2010a, 2010b).

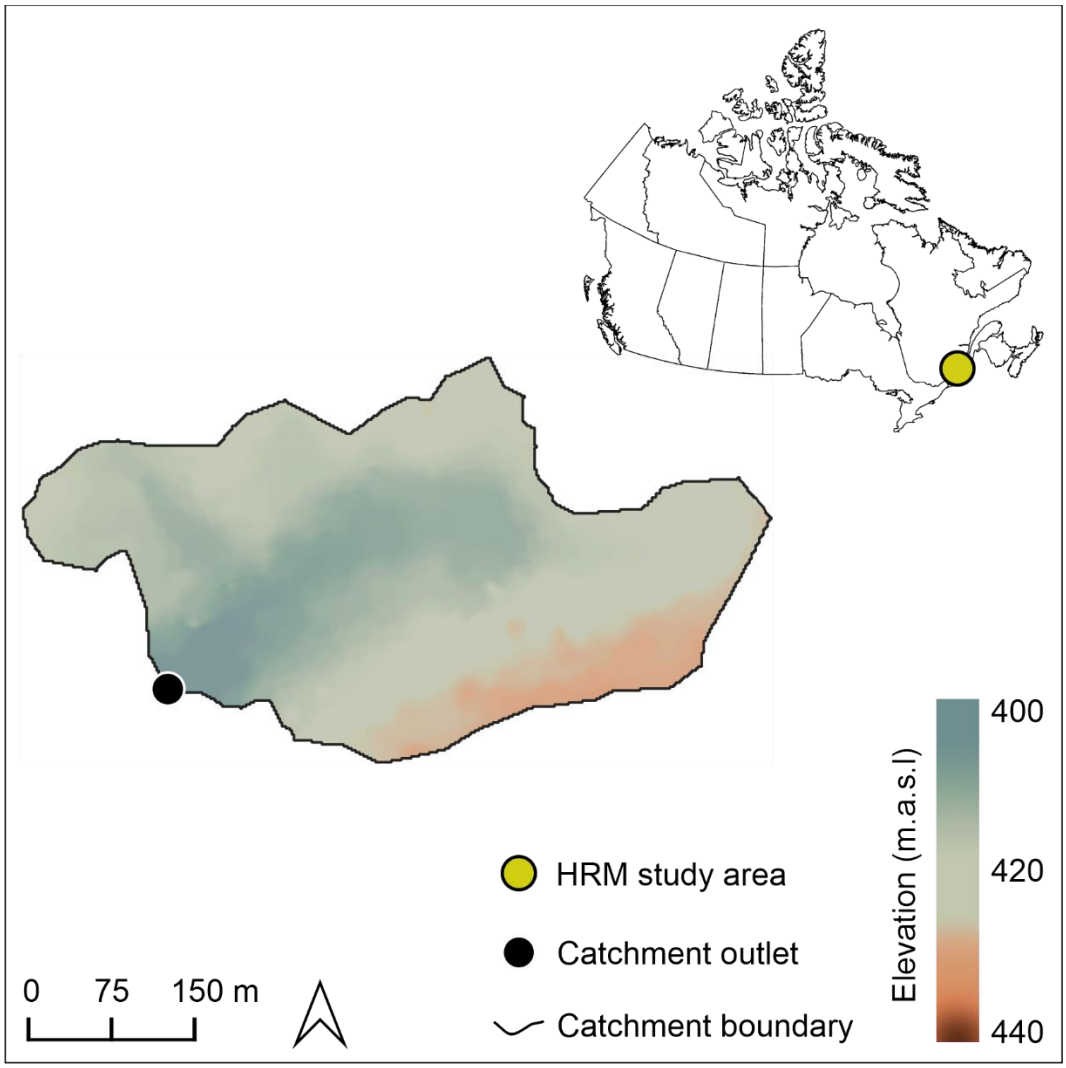


Figure 1-3. Digital elevation model of the Hermine catchment (HRM) and its location within Canada. The elevation is shown in meters above sea level (m.a.s.l).

1.6.3 The IISD-ELA Lake 658 UP1 catchment

The Lake 658 UP1 catchment (UP1) is a 0.08 km^2 sub-catchment of the Lake 658 experimental watershed at the International Institute of Sustainable Development Experimental Lakes Area (IISD-ELA) in north-western Ontario, Canada (Figure 1-4). The climate is boreal cold temperate with a mean annual air temperature of 2.8 ºC and a mean annual precipitation of 708 mm, 75% of which falls as rain (Oswald et al., 2011). The UP1 has a highly variable local topography, with a sequence of alternating soil-filled bedrock depressions and near-vertical bedrock ridges. Overall, the catchment has an average slope of 12º with 63 m of relief. Catchment bedrock is granite, which is exposed or covered by a shallow soil layer (~ 23 cm) for nearly 40% of the catchment, while the rest of the catchment has a mean soil depth of 54 cm (Oswald et al., 2011). Soils of the UP1 are mostly acidic Brunisols that are texturally classified as silt loams. Approximately 14% of the UP1 supports a deciduous forest of red maple and paper birch, while the remainder is dominated by mature black spruce (Oswald et al., 2011). Data from the UP1 used in the current study include a detailed DEM produced from manual surveys as well as high-frequency (15-minute) catchment discharge, precipitation, and other meteorological measurements spanning 2008 and 2009.

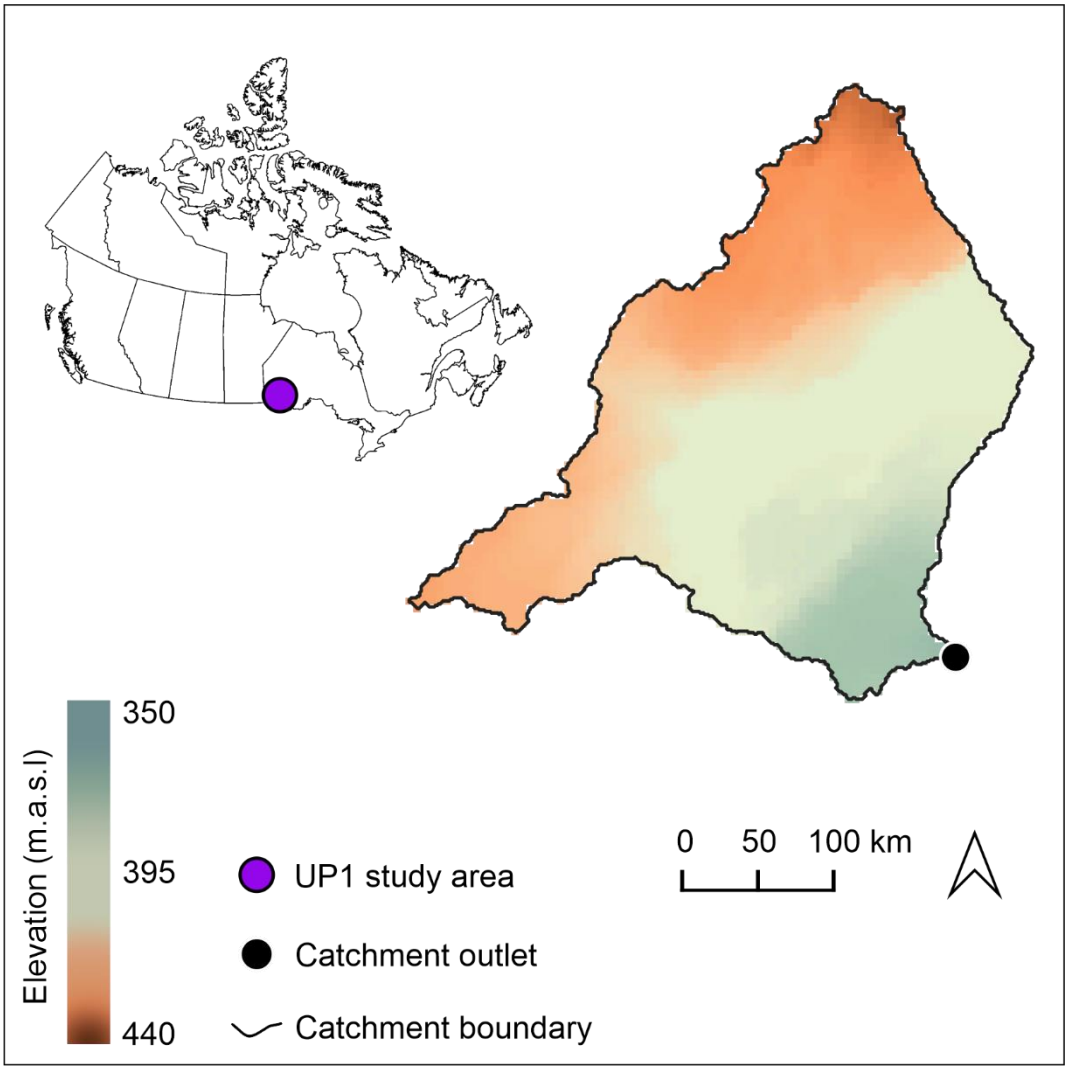


Figure 1-4. Digital elevation model of the IISD-ELA Lake 658 UP1 catchment (UP1) and its location within Canada. The elevation is shown in meters above sea level (m.a.s.l).

1.6.4 The Tarrawarra catchment

The Tarrawarra catchment (TRC) drains 0.108 km^2 and is located near Melbourne, Victoria, Australia (Figure 1-5). The TRC has a temperate climate with a summer rainfall deficit and winter rainfall excess: the mean annual rainfall and potential evapotranspiration are ~820 and ~830 mm, respectively (Western & Grayson, 1998). Soils of the TRC comprise three units:

upper slopes have texture contrast soil with a loam-clay A horizon (15-25 cm) and a heavy yellow-grey B horizon; midslope soils have a loam-clay A horizon (15-30 cm) with a silty B1 horizon and a silty-clay B2 horizon; and depressions have a deeper (25-40 cm) silty A horizon and a silty B horizon. Soils are underlain by the Humevale Formation, comprising siltstone with interbedded thin sandstone and local bedded limestone lenses (Western & Grayson, 1998). The TRC has a smoothly undulating terrain that is primarily used for dryland grazing. There are windbreaks along two-thirds of the northern catchment boundary (10 m Cypress trees), a small part of the southern boundary has 5 m mixed Australian native trees, and there are Eucalyptus trees in the southeastern corner of the catchment. Data from the TRC used in this study include a DEM derived from a 10x10 m topographic survey, local weather station data including precipitation and other meteorological parameters recorded every 6 minutes, and high-frequency (1-minute) catchment discharge measured using a flume (Western & Grayson, 1998).

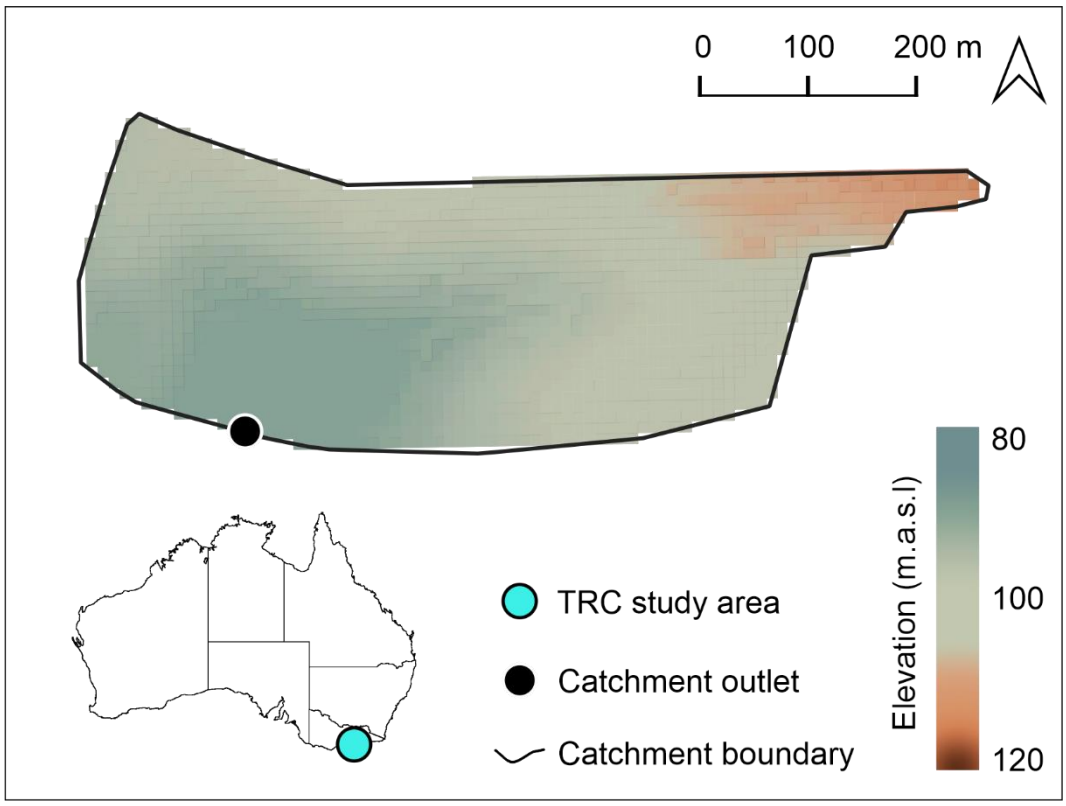


Figure 1-5. Digital elevation model of the Tarrawarra catchment (TRC) and its location within Australia. The elevation is shown in meters above sea level (m.a.s.l).

1.6.5 The Catfish Creek watershed

The Catfish Creek watershed drains 642 km^2 and is located ~100 km north-east of Winnipeg, Manitoba, Canada. The watershed has a dry continental boreal climate that is characterized by short warm summers and long cold winters with a mean annual precipitation of approximately 530 mm, 20% of which falls as snow (Ross et al., 2017). The topographic profile of the catchment is nearly level and the soils are moderate to poorly drained and are developed on crystalline Archean bedrock covered by a discontinuous mantle of sandy glacial till veneer (Smith et al., 1998). Land use across the catchment is equally split between agriculture and

forested lands (Ross et al., 2017). The current study focuses on a 145.36 km^2 sub-catchment (CCW) located in the south-eastern portion of the watershed (Figure 1-6). Data for the CCW used in the current study include a 1x1 m DEM and high-frequency discharge and meteorological data collected from 2013 to 2015. Discharge data is derived from high-frequency (15-minute) stage data using a site-specific rating curve, while high-frequency (1-minute) rainfall and meteorological data are from a HOBO station deployed over the data collection period.

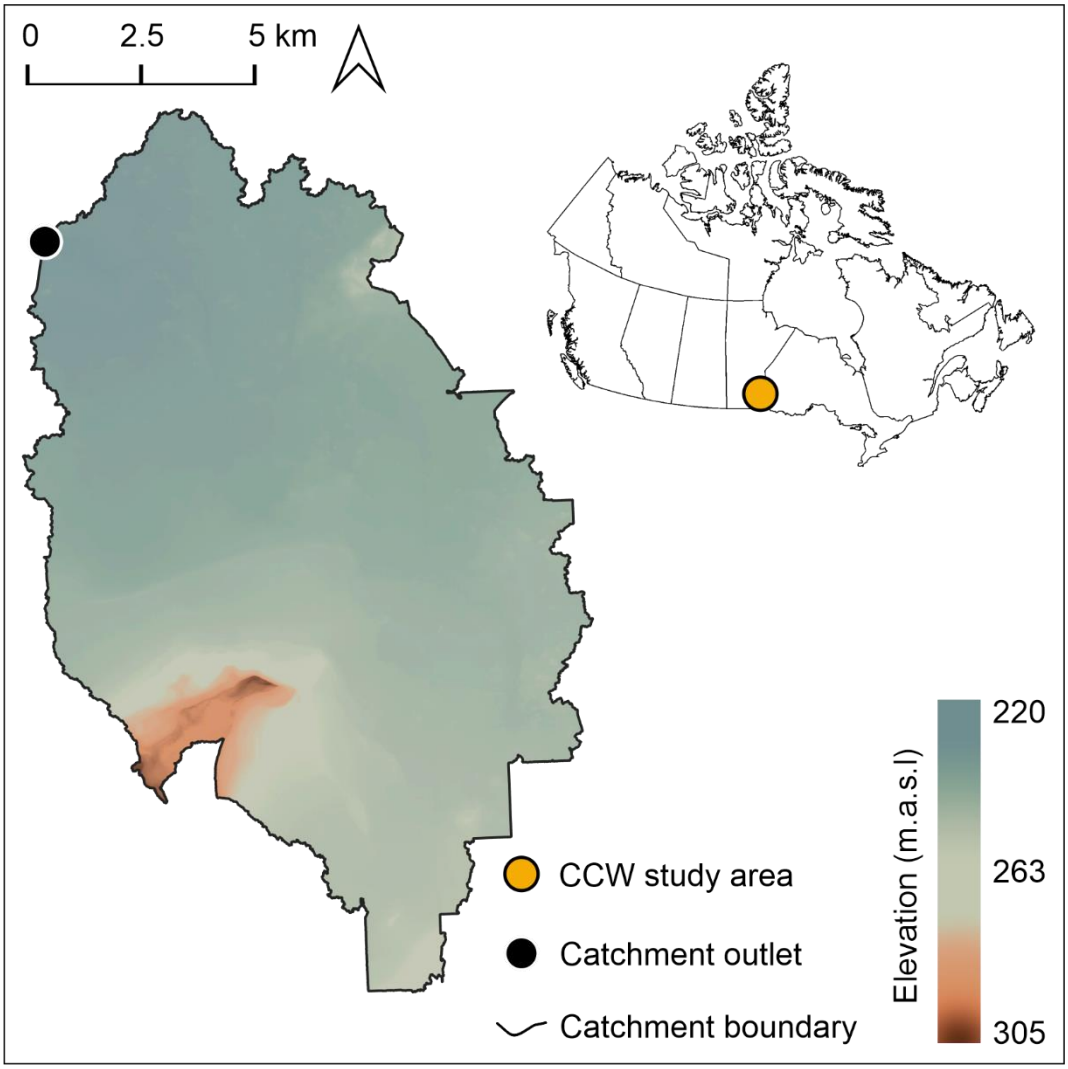


Figure 1-6. Digital elevation model of a sub-catchment of the Catfish Creek watershed (CCW) and its location within Canada. The elevation is shown in meters above sea level (m.a.s.l).

1.6.6 The Mahurangi River catchment

The Mahurangi River catchment drains 50 km^2 of steep to gently rolling lowlands and is located approximately 70 km north of Auckland, New Zealand (Woods et al., 2013). The catchment has ~250 m of relief and its soils are clay loam with a maximum depth of 1 m and are developed on Waitemata sandstones. The regional climate is warm humid maritime with a mean annual temperature of 15.6 ºC and a mean annual rainfall of ~1600 mm. Land use and land cover in the MRC include pasture for grazing, plantation forest (primarily *Pinus radiate*), native forests, and pasture (Woods et al., 2013). The current study focuses on eight sub-catchments (MRC1-MRC8) of the Mahurangi River catchment (Figure 1-7) that range in drainage area from 0.51 km^2 to 24.80 km^2 and were previously part of the Mahurangi River Variability Experiment (MARVEX) (Woods et al., 2013). Data for these catchments used in this study include a 1x1 m DEM, discharge, rainfall, and meteorological data collected from 1997-2002. Discharge and rainfall data were collected at a 2-minute interval using tipping bucket rain gauges and v-notch weirs, outfitted with floats and counterweights, respectively.

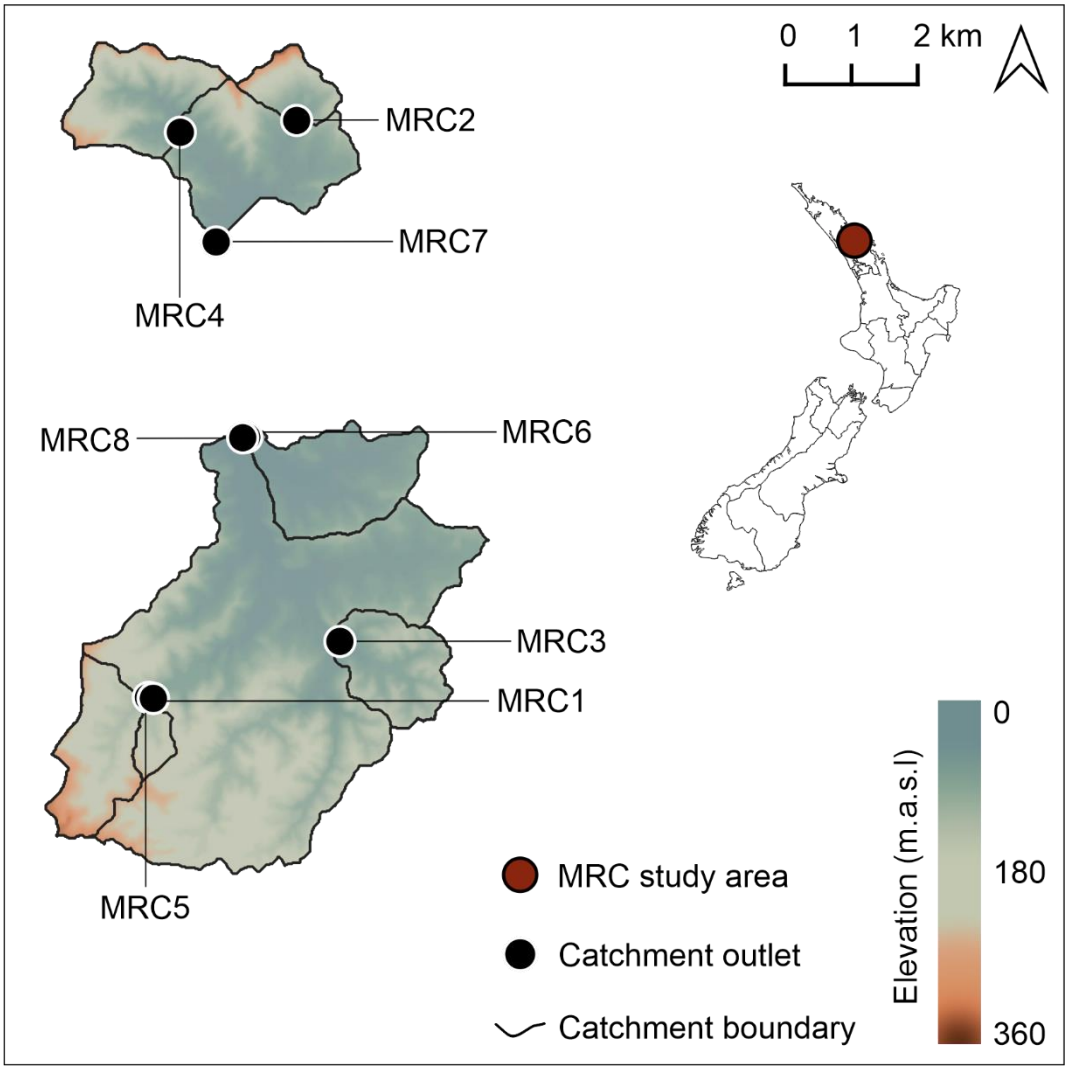


Figure 1-7. Digital elevation model of the eight sub-catchments (MRC1-MRC8) of the Mahurangi River catchment and their location within New Zealand. The elevation is shown in meters above sea level (m.a.s.l).

1.6.7 The H.J. Andrews Experimental Forest

The H.J. Andrews Experimental Forest is located approximately 80 km east of Eugene, Oregon, United States on the western slope of the Cascade Range (McKee & Druliner, 1998). The area is characterized by a maritime climate with wet, mild winters and dry, cool summers.

The mean annual precipitation is approximately 2300 mm at lower elevations and over 3550 mm at the upper elevations of the forest (McKee & Druliner, 1998). Precipitation at low elevations is predominantly rain, while snow is common at higher elevations. The forest is steep, with elevation ranging from 410 to 1630 m; lower elevations comprise Oligocene-lower Miocene volcanic rocks, while higher-area bedrock comprises andesite lava flow of Miocene age and younger High Cascade Rocks. Soils are primarily Inceptisols, with areas of Alfisols and Spodosols. Historical forest composition has been altered by timber cutting and forest fires; lower elevation forests are dominated by Douglas-fir, western hemlock, and western red cedar, while upper elevation forests contain noble fir, Pacific silver fir, Douglas-fir, and western hemlock (McKee & Druliner, 1998). The current project will include data from eight nested catchments (HJA1-HJA8) of the H.J. Andrews Experimental Forest (Figure 1-8). A substantive long-term data record is available for catchments of the forest starting in the 1950s (McKee & Druliner, 1998); however, this project will focus on a five-year subset (2010-2015) of rainfall, discharge, and meteorological data, in addition to terrain data. The terrain data includes a 10x10 m DEM and LiDAR data. High-frequency (15-minute) rainfall and discharge data are available for each of the eight nested catchments.

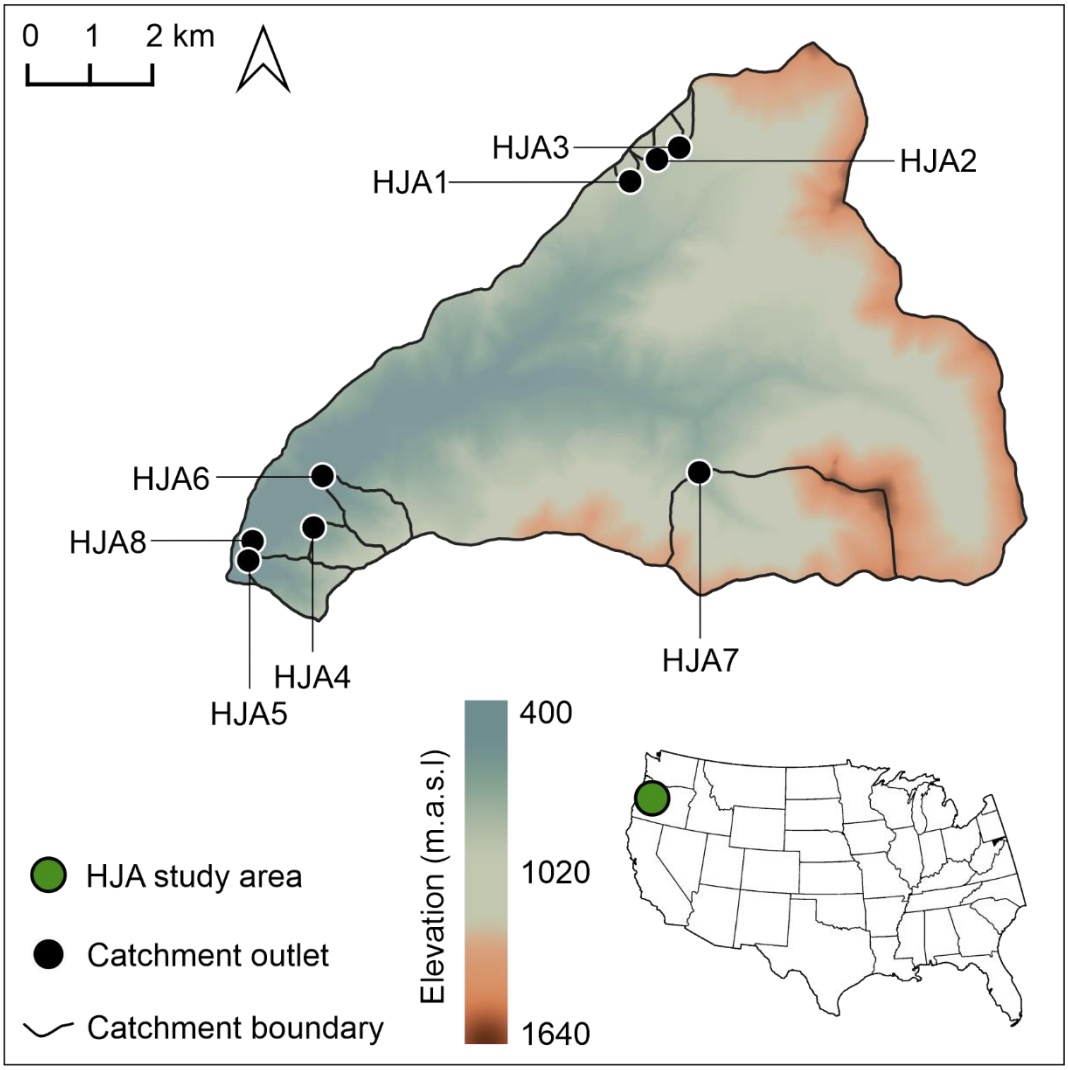


Figure 1-8. Digital elevation model of the eight nested catchments (HJA1-HJA8) of the H.J. Andrews Experimental Forest and their location within the United States. The elevation is shown in meters above sea level (m.a.s.l).

1.7 Study significance

Hydrological thresholds are important catchment functional traits that facilitate the grouping of similar hydrologic responses (Ali et al., 2013; Lehmann et al., 2007; McDonnell et al., 2007). A threshold-based hydrological theory has been identified as a potential way to

enhance our understanding of hillslope and catchment scale hydrological processes (Ali et al., 2013; McDonnell et al., 2007; Spence, 2010). Individual studies in catchment hydrology have led to the characterization of heterogeneous and complex rainfall-runoff processes in an increasing number of environments (Kirchner, 2009; McDonnell, 2013; McDonnell et al., 2007). This is also true for research focused on threshold behaviour in hydrological processes: studies have predominantly been conducted at the hillslope and small catchment scales in humid temperate environments (Ali et al., 2015), and few studies have assessed or compared threshold information obtained from different environments or at different scales (Ali et al., 2013). Additionally, there is no consensus on the type and form of observation data that should be used to characterize hydrologic response, making the relative ubiquity of threshold behaviour in hydrologic response unknown. This research begins to address these knowledge gaps through the consideration of relationships between different aspects of hydrologic response and meteorological inputs that quantify rainfall amount, rainfall intensity, and hydrologic abstraction from evaporation. Evaluating nonlinear hydrologic response as a function of multiple meteorological factors that are associated with different processes (e.g., catchment storage) may reveal new information about controls on emergent threshold behaviour. Furthermore, threshold behaviour observed at the catchment scale has yet to be thoroughly incorporated into modelling activities, including model evaluation. This study aims to evaluate hydrologic thresholds and other hydrologic descriptors as potentially valuable assets for constraining parameter sets with a similar statistical performance for rainfall-runoff models. The inclusion of data from studies previously conducted at one hillslope and twenty catchments, which have distinct climates and physiographic features, will enhance the reach of findings and help address analytic and modelling challenges associated with knowledge transferability across scales and environments.

This study will therefore move toward the development of a unified threshold-based hydrological theory. Additionally, enhancing knowledge of hydrologic thresholds that are representative of storage release processes, flow pathway activation and the establishment of hydrologic connectivity presents an opportunity to better understand water-quality dynamics and to inform flood forecasting and water policy in general.

1.8 Thesis structure

This thesis is structured in the grouped manuscript style and consists of a collection of manuscripts, which are published or are soon to be submitted to peer-reviewed journals. The first chapter introduces the overall theme of the thesis. In Section 1.5 of the introductory chapter, four research objectives are introduced that are the basis for four manuscripts that correspond to Chapters 2, 3, 4, and 5. Chapter 6 provides a synthesis of the major findings from Chapters 2, 3, 4, and 5 and discusses study limitations and recommendations for future work.

Chapter 2 is published in the journal *Hydrological Processes*:

Ross, C. A., Ali, G., Spence, C., Oswald, C., & Casson, N. (2019). Comparison of event-specific rainfall-runoff responses and their controls in contrasting geographic areas. *Hydrological Processes*, 33(14), 1961–1979. https://doi.org/10.1002/hyp.13460

Chapter 3 is published in the journal *Water Resources Research*:

Ross, C., Ali, G., Spence, C., & Courchesne, F. (2021). Evaluating the Ubiquity of Thresholds in Rainfall-Runoff Response Across Contrasting Environments. *Water Resources Research*, *57*(1). https://doi.org/10.1029/2020WR027498

Chapters 4 and 5 will be submitted to peer-reviewed journals at a later date. All four manuscripts are multi-author works. The nature and extent of my contribution to each manuscript are as follows: Chapter 2 was written by me including the completion of all data analyses that support the manuscript. Editorial revisions and conceptual suggestions were provided by co-authors Dr. Genevieve Ali, Dr. Chris Spence, Dr. Claire Oswald, and Dr. Nora Casson. For Chapter 3, writing and analyses were completed by me with suggestions from Dr. Genevieve Ali, Dr. Chris Spence, and Dr. Francois Courchesne. The writing and analyses for Chapters 4 and 5 were completed by me with suggestions from Dr. Genevieve Ali and Dr. Chris Spence. The exact formatting of each manuscript has been altered from the published version for consistency within this thesis.

1.9 References

Ala-aho, P., Tetzlaff, D., McNamara, J. P., Laudon, H., & Soulsby, C. (2017). Using isotopes to constrain water flux and age estimates in snow-influenced catchments using the STARR (Spatially distributed Tracer-Aided Rainfall–Runoff) model. *Hydrology and Earth System Sciences*, *21*(10), 5089–5110. https://doi.org/10.5194/hess-21-5089-2017

Ali, G., L'Heureux, C., Roy, A., Turmel, M.-C., & Courchesne, F. (2011). Linking spatial patterns of perched groundwater storage and stormflow generation processes in a

headwater forested catchment. *Hydrological Processes*, *25*(25), 3843–3857. https://doi.org/10.1002/hyp.8238

Ali, G., Oswald, C., Spence, C., Cammeraat, E., McGuire, K., Meixner, T., & Reaney, S. M. (2013). Towards a unified threshold-based hydrological theory: Necessary components and recurring challenges. *Hydrological Processes*, *27*(2), 313–318. https://doi.org/10.1002/hyp.9560

Ali, G., & Roy, A. (2010a). A case study on the use of appropriate surrogates for antecedent moisture conditions (AMCs). *Hydrology and Earth System Sciences*, *14*(10), 1843–1861.

Ali, G., & Roy, A. (2010b). Shopping for hydrologically representative connectivity metrics in a humid temperate forested catchment. *Water Resources Research*, *46*(12). https://doi.org/10.1029/2010WR009442

Ali, G., Tetzlaff, D., McDonnell, J., Soulsby, C., Carey, S., Laudon, H., McGuire, K., Buttle, J., Seibert, J., & Shanley, J. (2015). Comparison of threshold hydrologic response across northern catchments. *Hydrological Processes*, *29*(16), 3575–3591. https://doi.org/10.1002/hyp.10527

Andersen, T., Carstensen, J., Hernandez-Garcia, E., & Duarte, C. M. (2009). Ecological thresholds and regime shifts: Approaches to identification. Trends in Ecology & Evolution, 24(1), 49–57. https://doi.org/10.1016/j.tree.2008.07.014

Beven, K. (2002). Towards a coherent philosophy for modelling the environment. Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences, 458(2026), 2465–2484. https://doi.org/10.1098/rspa.2002.0986

Beven, K. (2006). A manifesto for the equifinality thesis. Journal of Hydrology, 320(1–2), 18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007

Beven, K. (2011). *Rainfall-runoff modelling: The primer*. John Wiley & Sons.

Beven, K., Wood, E., & Sivapalan, M. (1988). On hydrological heterogeneity—Catchment morphology and catchment response. *Journal of Hydrology*, *100*(1), 353–375. https://doi.org/10.1016/0022-1694(88)90192-8

Biron, P. M., Roy, A. G., Courschesne, F., Hendershot, W. H., Côté, B., & Fyles, J. (1999). The effects of antecedent moisture conditions on the relationship of hydrology to hydrochemistry in a small forested watershed. *Hydrological Processes*, *13*(11), 1541–1555. https://doi.org/10.1002/(SICI)1099-1085(19990815)13:11<1541::AID-HYP832>3.0.CO;2-J

Bonell, M. (1993). Progress in the understanding of runoff generation dynamics in forests. *Journal of Hydrology*, *150*(2), 217–275. https://doi.org/10.1016/0022-1694(93)90112-M

Buttle, J. (2006). Mapping first-order controls on streamflow from drainage basins: The T3 template. *Hydrological Processes*, *20*(15), 3415–3422. https://doi.org/10.1002/hyp.6519

Cammeraat, L. H. (2002). A review of two strongly contrasting geomorphological systems within the context of scale. *Earth Surface Processes and Landforms*, *27*(11), 1201–1222. https://doi.org/10.1002/esp.421

Carey, S. K., Tetzlaff, D., Seibert, J., Soulsby, C., Buttle, J., Laudon, H., McDonnell, J., McGuire, K., Caissie, D., Shanley, J., Kennedy, M., Devito, K., & Pomeroy, J. W. (2010). Inter-comparison of hydro-climatic regimes across northern catchments: Synchronicity, resistance and resilience. *Hydrological Processes*, *24*(24), 3591–3602. https://doi.org/10.1002/hyp.7880

Clark, M. P., Bierkens, M. F. P., Samaniego, L., Woods, R. A., Uijlenhoet, R., Bennett, K. E., Pauwels, V. R. N., Cai, X., Wood, A. W., & Peters-Lidard, C. D. (2017). The evolution

of process-based hydrologic models: Historical challenges and the collective quest for physical realism. *Hydrology and Earth System Sciences*, *21*(7), 3427–3440. https://doi.org/10.5194/hess-21-3427-2017

Davie, T. (2008). *Fundamentals of Hydrology* (2nd ed.). Routledge

Detty, J. M., & McGuire, K. J. (2010). Threshold changes in storm runoff generation at a till-mantled headwater catchment. *Water Resources Research; Washington*, *46*(7). http://dx.doi.org.uml.idm.oclc.org/10.1029/2009WR008102

Devito, K., Creed, I., Gan, T., Mendoza, C., Petrone, R., Silins, U., & Smerdon, B. (2005). A framework for broad-scale classification of hydrologic response units on the Boreal Plain: Is topography the last thing to consider? *Hydrological Processes*, *19*(8), 1705–1714. https://doi.org/10.1002/hyp.5881

Dingman, S. L. (2015). *Physical hydrology*. Waveland press.

Freer, J., McDonnell, J. J., Beven, K. J., Peters, N. E., Burns, D. A., Hooper, R. P., Aulenbach, B., & Kendall, C. (2002). The role of bedrock topography on subsurface storm flow. *Water Resources Research*, *38*(12). https://doi.org/10.1029/2001WR000872

Graham, C. B., & McDonnell, J. J. (2010). Hillslope threshold response to rainfall: (2) Development and use of a macroscale model. *Journal of Hydrology*, *393*(1–2), 77–93. https://doi.org/10.1016/j.jhydrol.2010.03.008

Graham, C. B., Woods, R. A., & McDonnell, J. J. (2010). Hillslope threshold response to rainfall: (1) A field based forensic approach. *Journal of Hydrology*, *393*(1–2), 65–76. https://doi.org/10.1016/j.jhydrol.2009.12.015

Gregory, K. J., & Walling, D. E. (1973). *Drainage basin form and process; a geomorphological approach*. Edward Arnold.

Higgins, M. W., Atkins, R. L., & Crawford, T. J. (1988). The structure, stratigraphy, tectonostratigraphy, and evolution of the southern most part of the Appalachian orogen. *Пущдюыгкмун Зкщаюзфзукжс1475*.

James, A., & Roulet, N. (2007). Investigating hydrologic connectivity and its association with threshold change in runoff response in a temperate forested watershed. *Hydrological Processes*, *21*(25), 3391–3408. https://doi.org/10.1002/hyp.6554

Kelleher, C., McGlynn, B., & Wagener, T. (2017). Characterizing and reducing equifinality by constraining a distributed catchment model with regional signatures, local observations, and process understanding. *Hydrology and Earth System Sciences*, *21*(7), 3325–3352. https://doi.org/10.5194/hess-21-3325-2017

Kim, H. J., Sidle, R. C., Moore, R. D., & Hudson, R. (2004). Throughflow variability during snowmelt in a forested mountain catchment, coastal British Columbia, Canada. *Hydrological Processes*, *18*(7), 1219–1236. https://doi.org/10.1002/hyp.1396

Kinzig, A., Ryan, P., Etienne, M., Allison, H., Elmqvist, T., & Walker, B. (2006). Resilience and Regime Shifts: Assessing Cascading Effects. *Ecology and Society*, *11*(1). https://doi.org/10.5751/ES-01678-110120

Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research*, *42*(3), W03S04. https://doi.org/10.1029/2005WR004362

Kirchner, J. W. (2009). Catchments as simple dynamical systems: Catchment characterization, rainfall-runoff modeling, and doing hydrology backward. *Water Resources Research*, *45*(2). https://doi.org/10.1029/2008WR006912

Laudon, H., Sjöblom, V., Buffam, I., Seibert, J., & Mörth, M. (2007). The role of catchment scale and landscape characteristics for runoff generation of boreal streams. *Journal of Hydrology*, *344*(3), 198–209. https://doi.org/10.1016/j.jhydrol.2007.07.010

Lehmann, P., Hinz, C., McGrath, G., Tromp-van Meerveld, H. J., & McDonnell, J. J. (2007). Rainfall threshold for hillslope outflow: An emergent property of flow pathway connectivity. *Hydrol. Earth Syst. Sci.*, *11*(2), 1047–1063. https://doi.org/10.5194/hess-11-1047-2007

Limburg, K. E., O'Neill, R. V., Costanza, R., & Farber, S. (2002). Complex systems and valuation. *Ecological Economics*, *41*(3), 409–420. https://doi.org/10.1016/S0921-8009(02)00090-3

Lintz, H. E., McCune, B., Gray, A. N., & McCulloh, K. A. (2011). Quantifying ecological thresholds from response surfaces. *Ecological Modelling*, *222*(3), 427–436. https://doi.org/10.1016/j.ecolmodel.2010.10.017

McDonnell, J. (1990). A Rationale for Old Water Discharge Through Macropores in a Steep, Humid Catchment. *Water Resources Research*, *26*(11), 2821–2832. https://doi.org/10.1029/WR026i011p02821

McDonnell, J. (2013). Are all runoff processes the same? *Hydrological Processes*, *27*(26), 4103–4111. https://doi.org/10.1002/hyp.10076

McDonnell, J., Sivapalan, M., Vaché, K., Dunn, S., Grant, G., Haggerty, R., Hinz, C., Hooper, R., Kirchner, J., Roderick, M. L., Selker, J., & Weiler, M. (2007). Moving beyond heterogeneity and process complexity: A new vision for watershed hydrology. *Water Resources Research*, *43*(7), W07301. https://doi.org/10.1029/2006WR005467

McGlynn, B. L., & McDonnell, J. (2003). Quantifying the relative contributions of riparian and hillslope zones to catchment runoff. *Water Resources Research*, *39*(11), 1310. https://doi.org/10.1029/2003WR002091

McKee, A., & Druliner, P. (1998). *HJ Andrews Experimental Forest.* http://andrewsforest.oregonstate.edu/pubs/pdf/pub2415.pdf

Merz, R., & Blöschl, G. (2009). A regional analysis of event runoff coefficients with respect to climate and catchment characteristics in Austria. *Water Resources Research*, *45*(1). https://doi.org/10.1029/2008WR007163

Mielko, C., & Woo, M. (2006). Snowmelt runoff processes in a headwater lake and its catchment, subarctic Canadian Shield. *Hydrological Processes*, *20*(4), 987–1000. https://doi.org/10.1002/hyp.6117

Mosley, M. P. (1979). Streamflow generation in a forested watershed, New Zealand. Water Resources Research, 15(4), 795–806. https://doi.org/10.1029/WR015i004p00795

Oswald, C., Richardson, M. C., & Branfireun, B. A. (2011). Water storage dynamics and runoff response of a boreal Shield headwater catchment. *Hydrological Processes*, *25*(19), 3042–3060. https://doi.org/10.1002/hyp.8036

Phillips, J. D. (2006). Evolutionary geomorphology: Thresholds and nonlinearity in landform response to environmental change. *Hydrol. Earth Syst. Sci.*, *10*(5), 731–742. https://doi.org/10.5194/hess-10-731-2006

Reaney, S. M., Bracken, L. J., & Kirkby, M. J. (2007). Use of the Connectivity of Runoff Model (CRUM) to investigate the influence of storm characteristics on runoff generation and connectivity in semi-arid areas. *Hydrological Processes*, *21*(7), 894–906. https://doi.org/10.1002/hyp.6281

Redding, T. E., & Devito, K. J. (2008). Lateral flow thresholds for aspen forested hillslopes on the Western Boreal Plain, Alberta, Canada. *Hydrological Processes*, *22*(21), 4287–4300. https://doi.org/10.1002/hyp.7038

Ross, C., Ali, G., & Lobb, D. (2017). Inferring soil water movement and streamflow response in Canadian Prairie riparian areas using hydrologic state variables. *Hydrological Processes*, *31*(22), 3765–3782. https://doi.org/10.1002/hyp.11303

Ross, C., Petzold, H., Penner, A., & Ali, G. (2015). Comparison of sampling strategies for monitoring water quality in mesoscale Canadian Prairie watersheds. *Environmental Monitoring and Assessment*, *187*(7), 395. https://doi.org/10.1007/s10661-015-4637-9

Scaife, C. I., & Band, L. E. (2017). Nonstationarity in threshold response of stormflow in southern Appalachian headwater catchments. *Water Resources Research*, *53*(8), 6579–6596. https://doi.org/10.1002/2017WR020376

Scaife, C. I., Singh, N. K., Emanuel, R. E., Miniat, C. F., & Band, L. E. (2020). Non-linear quickflow response as indicators of runoff generation mechanisms. *Hydrological Processes*.

Seibert, J., & McDonnell, J. J. (2002). On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria model calibration. *Water Resources Research*, *38*(11), 23-1-23–14. https://doi.org/10.1029/2001WR000978

Shanley, J. B., & Chalmers, A. (1999). The effect of frozen soil on snowmelt runoff at Sleepers River, Vermont. *Hydrological Processes*, *13*(12-13), 1843–1857. https://doi.org/10.1002/(SICI)1099-1085(199909)13:12/13<1843::AID-HYP879>3.0.CO;2-G

Shaw, D. A., Pietroniro, A., & Martz, L. w. (2013). Topographic analysis for the prairie pothole region of Western Canada. *Hydrological Processes*, *27*(22), 3105–3114. https://doi.org/10.1002/hyp.9409

Shaw, D. A., Vanderkamp, G., Conly, F. M., Pietroniro, A., & Martz, L. (2012). The Fill–Spill Hydrology of Prairie Wetland Complexes during Drought and Deluge. *Hydrological Processes*, *26*(20), 3147–3156. https://doi.org/10.1002/hyp.8390

Sidle, R. C., Tsuboyama, Y., Noguchi, S., Hosoda, I., Fujieda, M., & Shimizu, T. (2000). Stormflow generation in steep forested headwaters: A linked hydrogeomorphic paradigm. *Hydrological Processes*, *14*(3), 369–385. https://doi.org/10.1002/(SICI)1099-1085(20000228)14:3<369::AID-HYP943>3.0.CO;2-P

Sivakumar, B. (2005). Hydrologic modeling and forecasting: Role of thresholds. *Environmental Modelling & Software*, *20*(5), 515–519. https://doi.org/10.1016/j.envsoft.2004.08.006

Sivapalan, M. (2006). Pattern, Process and Function: Elements of a Unified Theory of Hydrology at the Catchment Scale. In *Encyclopedia of Hydrological Sciences*. John Wiley & Sons, Ltd. https://doi.org/10.1002/0470848944.hsa012

Sivapalan, M., Jothityangkoon, C., & Menabde, M. (2002). Linearity and nonlinearity of basin response as a function of scale: Discussion of alternative definitions. *Water Resources Research*, *38*(2). https://doi.org/10.1029/2001WR000482

Smith, R. E., Velhuis, H., Mills, G. F., Eilers, R. G., Fraser, W. R., & Lelyk, G. W. (1998). Terrestrial Ecozones, Ecoregions, and Ecodistricts of Manitoba. *Technical Bulletin*, 9E.

Son, K., & Sivapalan, M. (2007). Improving model structure and reducing parameter uncertainty in conceptual water balance models through the use of auxiliary data. *Water Resources Research*, *43*(1). https://doi.org/10.1029/2006WR005032

Spence, C. (2007). On the relation between dynamic storage and runoff: A discussion on thresholds, efficiency, and function. *Water Resources Research*, *43*(12), W12416. https://doi.org/10.1029/2006WR005645

Spence, C. (2010). A Paradigm Shift in Hydrology: Storage Thresholds Across Scales Influence Catchment Runoff Generation. *Geography Compass*, *4*(7), 819–833. https://doi.org/10.1111/j.1749-8198.2010.00341.x

Spence, C., Phillips, R., Hedstrom, N., Granger, R., & Reid, B. (2010). Storage dynamics and streamflow in a catchment with a variable contributing area. *Hydrological Processes*, *24*(16), 2209–2221. https://doi.org/10.1002/hyp.7492

Stadnyk, T. A., Delavau, C., Kouwen, N., & Edwards, T. W. D. (2013). Towards hydrological model calibration and validation: Simulation of stable water isotopes using the isoWATFLOOD model. *Hydrological Processes*, *27*(25), 3791–3810. https://doi.org/10.1002/hyp.9695

Stichling, W., & Blackwell, S. R. (1957). Drainage area as a hydrologic factor on the glaciated Canadian prairies. *International Association for Scientific Hydrology Publication*, *45*.

Tani, M. (1997). Runoff generation processes estimated from hydrological observations on a steep forested hillslope with a thin soil layer. *Journal of Hydrology*, *200*(1), 84–109. https://doi.org/10.1016/S0022-1694(97)00018-8

Tromp-van Meerveld, H. J., J, H., James, A. L., McDonnell, J. J., & Peters, N. E. (2008). A reference data set of hillslope rainfall-runoff response, Panola Mountain Research Watershed, United States. *Water Resources Research*, *44*(6). https://doi.org/10.1029/2007WR006299

Tromp-van Meerveld, H. J., & McDonnell, J. J. (2006a). Threshold relations in subsurface stormflow: 1. A 147-storm analysis of the Panola hillslope. *Water Resources Research*, *42*(2). https://doi.org/10.1029/2004WR003778

Tromp-van Meerveld, H. J., & McDonnell, J. J. (2006b). Threshold relations in subsurface stormflow: 2. The fill and spill hypothesis. *Water Resources Research*, *42*(2). https://doi.org/10.1029/2004WR003800

Uchida, T., Tromp-van Meerveld, I., & McDonnell, J. J. (2005). The role of lateral pipe flow in hillslope runoff response: An intercomparison of non-linear hillslope response. *Journal of Hydrology*, *311*(1), 117–133. https://doi.org/10.1016/j.jhydrol.2005.01.012

Wei, L., Qiu, Z., Zhou, G., Kinouchi, T., & Liu, Y. (2020). Stormflow threshold behaviour in a subtropical mountainous headwater catchment during forest recovery period. *Hydrological Processes*, *34*(8), 1728–1740. https://doi.org/10.1002/hyp.13658

Weiler, M. (2005). An infiltration model based on flow variability in macropores: Development, sensitivity analysis and applications. *Journal of Hydrology*, *310*(1–4), 294–315. https://doi.org/10.1016/j.jhydrol.2005.01.010

Western, A. W., & Grayson, R. B. (1998). The Tarrawarra Data Set: Soil moisture patterns, soil characteristics, and hydrological flux measurements. *Water Resources Research*, *34*(10), 2765–2768. https://doi.org/10.1029/98WR01833

Whipkey, R. Z. (1965). Subsurface Stormflow from Forested Slopes. *International Association of Scientific Hydrology. Bulletin*, *10*(2), 74–85. https://doi.org/10.1080/02626666509493392

Wood, E. F., Sivapalan, M., Beven, K., & Band, L. (1988). Effects of spatial variability and scale with implications to hydrologic modeling. *Journal of Hydrology*, *102*(1), 29–47. https://doi.org/10.1016/0022-1694(88)90090-X

Woods, R., Grayson, R., Western, A., Duncan, M., Wilson, D., Young, R., Ibbitt, R., Henderson, R., & McMahon, T. (2013). Experimental Design and Initial Results from the Mahurangi River Variability Experiment: MARVEX. In *Land Surface Hydrology, Meteorology, and Climate: Observations and Modeling* (pp. 201–213). American Geophysical Union. http://onlinelibrary-wiley-com.uml.idm.oclc.org/doi/10.1029/WS003p0201/summary

Yadav, M., Wagener, T., & Gupta, H. (2007). Regionalization of constraints on expected watershed response behaviour for improved predictions in ungauged basins. *Advances in Water Resources*, *30*(8), 1756–1774. https://doi.org/10.1016/j.advwatres.2007.01.005

Zehe, E., Becker, R., Bárdossy, A., & Plate, E. (2005). Uncertainty of simulated catchment runoff response in the presence of threshold processes: Role of initial soil moisture and precipitation. *Journal of Hydrology*, *315*(1), 183–202. https://doi.org/10.1016/j.jhydrol.2005.03.038

Zehe, E., & Blöschl, G. (2004). Predictability of hydrologic response at the plot and catchment scales: Role of initial conditions. *Water Resources Research*, *40*(10), W10202. https://doi.org/10.1029/2003WR002869

CHAPTER 2. COMPARISON OF EVENT-SPECIFIC RAINFALL-RUNOFF RESPONSES AND THEIR CONTROLS IN CONTRASTING GEOGRAPHIC AREAS

2.1 Introduction

Rainfall-runoff analyses for individual hillslopes and small catchments have been routinely performed for decades (Bates & Pilgrim, 1983; Betson, 1964; Jones & Swanson, 2001; Laudon et al., 2007), exposing tremendous variability in hydrologic response (McDonnell, 2013; McDonnell et al., 2007) that has been broadly related to a range of event factors (e.g., rainfall duration, total rainfall) and antecedent moisture conditions (Ambroise, 2004; Dingman, 2015). Rainfall-runoff response variability among sites has also been attributed to site-specific characteristics like drainage area, climatic regime, topography, land cover, and subsurface properties (Beven, 2001; McDonnell et al., 2007). Collectively, research efforts dedicated to rainfall-runoff analyses have provided great insights into runoff generation processes, catchment function, model development, and calibration (Beven, 2011; Jones & Swanson, 2001). However, opportunities for broad-scale comparisons of event-specific hydrologic response remain (Jones, 2006), especially when it comes to the better characterization of response magnitude and timing, the relative influence of storage- and intensity-driven meteorological factors on rainfall-runoff relationships, and the predictability of hydrologic response from site characteristics.

During rainfall-runoff analyses, response metrics are commonly derived from event hydrographs since they provide first-order information on hydrologic function (Carey & Woo, 2001; Tang & Carey, 2017) and are thought to be the expression of controls exerted by physiography and climate on hydrological processes (Te Chow, 1959; Te Chow et al., 1988). Those metrics aim to characterize either response magnitude (e.g., peak event discharge or total event runoff) or response timing (e.g., lag to initial hydrograph rise or lag to peak) (Beven, 2011; Hannah et al., 2000). Response magnitude metrics have been particularly useful in flood

mitigation, resource management, and engineering projects (Beven, 2011; Dingman, 2015; Fedora & Beschta, 1989) and their estimation was the goal of early hydrologic models (e.g., Mulvaney, 1851). Response magnitude metrics have also been used to quantify the efficiency of drainage areas in converting rainfall to runoff, and the storage deficit that must be exceeded before the initiation of hydrograph rise (Dingman, 2015). As for response timing metrics, they have been linked to rainfall duration and intensity (Carey & Woo, 2001; Dingman, 2015) and antecedent moisture conditions (Carey & Woo, 2001; Fedora & Beschta, 1989), and they have been used to discriminate base flow and storm flow (Tallaksen, 1995). One source of confusion, however, is the fact that the available literature reports on many single-site studies relying on a broad range of metrics of response magnitude and timing, thus making it difficult to identify which of those metrics might best describe the temporal and spatial variability in hydrologic response.

Beyond uncertainties surrounding which metrics to use to best characterize the variability in hydrologic response, there are knowledge gaps regarding the relative role – and possible joint influence – of storage-driven and intensity-driven meteorological factors. Storage-driven meteorological factors are mostly related to rainfall amount and can take the form of total event rainfall or antecedent rainfall. Conversely, intensity-driven meteorological factors are related to the rate of water delivery or removal, notably through measures of rainfall intensity or potential evapotranspiration. Previous studies have shown that meteorological factors determine dominant runoff generation mechanisms (Dunne & Black, 1970, 1970; Hewlett & Hibbert, 1967; Horton, 1933), with storage-driven factors being typically associated with saturation-excess flow processes and intensity-driven factors being associated with infiltration-excess flow processes (Ali et al., 2015; McDonnell, 2013). For instance, in the case of surface and subsurface flow

generation through saturation excess, hydrologic response typically occurs only after critical values of soil water storage (Ali et al., 2011; Detty & McGuire, 2010; Graham et al., 2010; James & Roulet, 2007; Lehmann et al., 2007; McGlynn & McDonnell, 2003; Mosley, 1979; Oswald et al., 2011; Sidle et al., 2000; Tani, 1997; Tromp-van Meerveld & McDonnell, 2006a; Whipkey, 1965) or total event rainfall (e.g., Mosley, 1979; Oswald et al., 2011; Redding & Devito, 2008; Sidle et al., 2000; Tani, 1997; Tromp-van Meerveld & McDonnell, 2006a; Whipkey, 1965) are exceeded. In the case of infiltration-excess flow generation, nonlinear hydrologic response has been shown, whereby critical values of rainfall intensity and evapotranspiration rate control runoff magnitude and timing (e.g., Cammeraat, 2002; Reaney *et al.*, 2007). It should be noted that while both storage- and intensity-driven meteorological factors are known to individually influence hydrologic response (Betson, 1964; Dingman, 2015; Holtan & Overton, 1963, 1965), they are not mutually exclusive and have the potential to exert a dual influence on rainfall-runoff dynamics (Betson, 1964).

One way to further understand hydrologic behaviour is through site inter-comparisons that leverage existing physiographic, hydrometric, and climate data along with robust statistical approaches. Such comparison exercises maximize the value of existing datasets and answer calls to better define first-order controls on catchment response (Ali et al., 2010; Blume et al., 2007; Jones & Swanson, 2001; Uchida et al., 2006). Paired catchment studies have been performed extensively in forest hydrology (Bosch & Hewlett, 1982; Uchida et al., 2006) but they mostly focused on a limited number of response metrics (e.g., peak event discharge) for neighbouring catchments (Dunne, 1978; Jones, 2000). Few studies have, simultaneously, evaluated storage- and intensity-driven meteorological factors (e.g., Ali et al., 2015), and comparisons rarely include sites from multiple continents and that have a broad range of scales and climate regimes

(Uchida et al., 2006). Further, past comparison exercises have not evaluated differences in the relative importance of site characteristics (often static) and event-specific meteorological controls (dynamic) in driving variability in hydrologic response. The lack of comparative studies is, in part, attributable to the burden associated with synthesizing and standardizing data from various sources, and the limited number of computer tools available for delineating rainfall-runoff events. While rainfall-runoff event analysis has typically been performed manually, recently introduced toolboxes (Tang & Carey, 2017) have automated the separation of base flow and storm flow and the matching of rainfall and runoff events, thus facilitating event-based rainfall-runoff analysis across sites. Therefore, the current paper applies such a toolbox to large, existing datasets, to identify rainfall-runoff events across seven geographic areas with contrasting climate, topography, geology, soil properties and land cover, and analyze these events using a selection of univariate and multivariate statistical techniques. Specifically, four research questions were addressed:

(1) How do event response magnitude and timing metrics vary between geographic areas?
(2) What is the relative influence of storage-driven and intensity-driven meteorological factors on event response metrics?
(3) Can drainage area, long-term climate variables, and site-specific physiographic variables be used to predict typical (i.e., average, median) values of event response metrics and their degree of temporal variability?
(4) Which event metrics are the most important for capturing the variability in site-specific hydrologic response?

2.2 Methods

2.2.1 Study sites

Twenty-one sites spanning four countries and seven geographic areas were selected for this study (Figure 2-1). Three of those sites are in Canada, including the Hermine catchment in the Lower Laurentians (HRM) in Quebec, the Lake 658 UP1 catchment at the IISD Experimental Lakes Area in the Canadian Boreal Shield (UP1) in Ontario, and the Catfish Creek Watershed in the eastern Prairies (CCW) in Manitoba. Nine sites are in the United States, namely the Panola Mountain Research Watershed experimental hillslope (PMRW) in the Piedmont of Georgia and eight nested catchments of the HJ Andrews Experimental Forest on the western slope of the Cascade Range (HJA1-HJA8) in Oregon. The Tarrawarra catchment (TRC) is in south-eastern Australia, while eight nested catchments of the Mahurangi River Catchment (MRC1-MRC8) are located in New Zealand near Auckland. The selected sites vary in drainage area, climate regime, topography, land cover, and subsurface properties. Each site has been the subject of substantial hydrologic research and has been described in detail by others (Ali & Roy, 2010b; McKee & Druliner, 1998; Oswald et al., 2011; Ross et al., 2017; Tromp-van Meerveld et al., 2008; Western & Grayson, 1998). Site characteristics, including drainage area, physiographic variables (i.e., relief, mean slope, and standard deviation of the slope), and long-term climate variables (i.e., mean annual values of temperature, potential evapotranspiration, and total precipitation) were derived from available meteorological and elevation data. The annual PET for each site was determined by summing monthly PET estimated using the Thornthwaite equation (Thornthwaite,

1948). Major regional and site-specific characteristics are summarized in Figure 2-1 and Table 2-1.

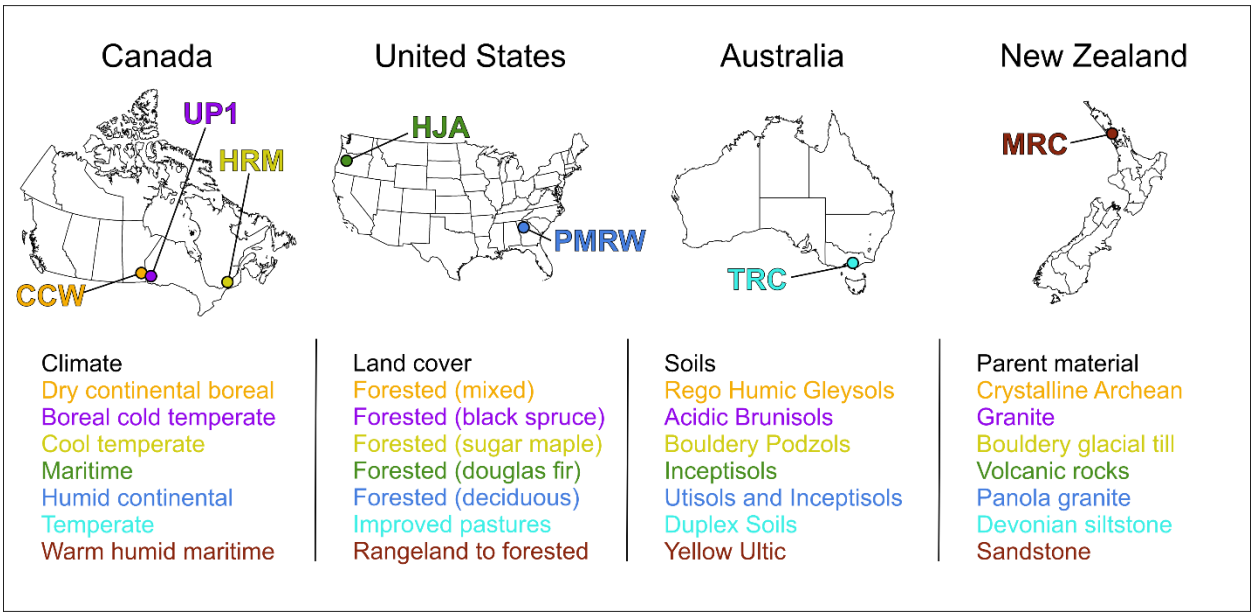


Figure 2-1. Location of the seven geographic areas in four countries. For both the MRC and HJA study areas, eight nested catchments were selected. Abbreviated site names and site-specific dominant climate type, land cover, soils, and parent material are colour-coded by study area. Refer to the text for full, non-abbreviated study area names. For the TRC site, Duplex refers to soils with contrasting textures across soil horizons.

Table 2-1. Site-specific drainage area (DA), topographic characteristics and mean annual values of temperature (T), potential evapotranspiration (PET), precipitation (P), and proportion of P that is rainfall (P_{RAIN}). MRC1 through MRC8, and HJA1 through HJA8, refer to nested catchments in the MRC and HJA study areas. SD: standard deviation.

	DA (km^2)	Relief (m)	Mean Slope (°)	SD Slope (°)	T (°C)	PET (mm)	P (mm) / P_{RAIN}
PMRW	1.0×10^{-3}	12.5	13.0	3.5	17.4	835	1240 / 1.0
HRM	0.05	27.0	0.2	0.2	6.9	508	1150 / 0.7
UP1	0.08	63.1	11.2	10.3	3.2	451	708 / 0.75
TRC	0.11	34.3	3.1	2.2	15.6	724	820 / 1.0
CCW	145.32	77.0	0.4	0.7	2.2	444	530 / 0.8
MRC1	0.51	104.8	14.2	6.6	15.6	716	1600 / 1.0
MRC2	0.71	116.4	17.0	7.6	15.6	716	1600 / 1.0
MRC3	2.30	113.6	13.4	7.1	15.6	716	1600 / 1.0
MRC4	2.63	210.8	15.2	8.1	15.6	716	1600 / 1.0
MRC5	2.65	155.1	14.5	6.4	15.6	716	1600 / 1.0
MRC6	2.96	73.1	5.7	4.9	15.6	716	1600 / 1.0
MRC7	4.61	30.1	8.2	7.7	15.6	716	1600 / 1.0
MRC8	24.80	286.6	11.9	7.4	15.6	716	1600 / 1.0
HJA1	0.13	129.3	23.1	6.8	7.5	489	2300 / 0.80
HJA2	0.15	155.1	21.2	7.2	7.1	475	2300 / 0.80
HJA3	0.21	126.6	19.5	10.1	7.8	492	2200 / 0.80
HJA4	0.60	432.0	24.2	8.8	8.7	525	2500 / 0.75
HJA5	0.96	408.3	23.1	9.3	8.9	528	2600 / 0.75
HJA6	1.01	276.6	52.4	23.9	8.7	523	2400 / 0.75
HJA7	14.36	860.7	21.2	9.9	8.8	527	2600 / 0.60
HJA8	62.42	1206.0	20.7	10.2	7.4	485	2400 / 0.75

2.2.2 Data processing

Rainfall and discharge records were analyzed to identify rainfall-runoff events for each site. Records varied in length (from 6 months to 5 years) and observation frequency (from 1 minute to 1 hour). High-frequency observations were aggregated to 1 hour to ensure consistency

across sites. The MATLAB toolbox HydRun (Tang & Carey, 2017) was used to identify runoff events and match them with associated rainfall events. Runoff events were identified based on hydrograph shape and may include single or multiple runoff peaks (Tang & Carey, 2017). Rainfall events were defined as adjacent rainfall observations separated by a rainless period that exceeded a user-defined duration (Tang & Carey, 2017). For each site, a unique rainless period duration was selected (between 4 and 48 hours) and visually confirmed following event delineation. It should be noted that while snowmelt is hydrologically significant for some of the examined sites, only rainfall-triggered events were considered to maintain consistency among sites and to limit the scope of the current study. Rainfall-runoff events with a runoff ratio value above 1 or with a hydrograph response that preceded rainfall were assumed to be a result of delineation errors or an indication of rain-on-snow events and were, therefore, discarded. Multi-peak runoff events were also checked, manually, to ensure that they were matched with the appropriate rainfall events. In total, 1,641 rainfall-runoff events were retained for further analyses, with the number of events per site ranging from 13 to 134. Examples of rainfall-runoff events for each site are featured in Figure 2-2 to illustrate typical response dynamics, i.e., dynamics prevailing when the total rainfall amount is close to the average total event rainfall. This average total event rainfall (i.e., arithmetic mean) was computed across all of the events delineated for each site. Figure 2-2 shows that typical response dynamics vary widely across sites, not only in terms of absolute peak discharge values but also with respect to the presence/absence of multiple peaks and the flashiness of the hydrograph.

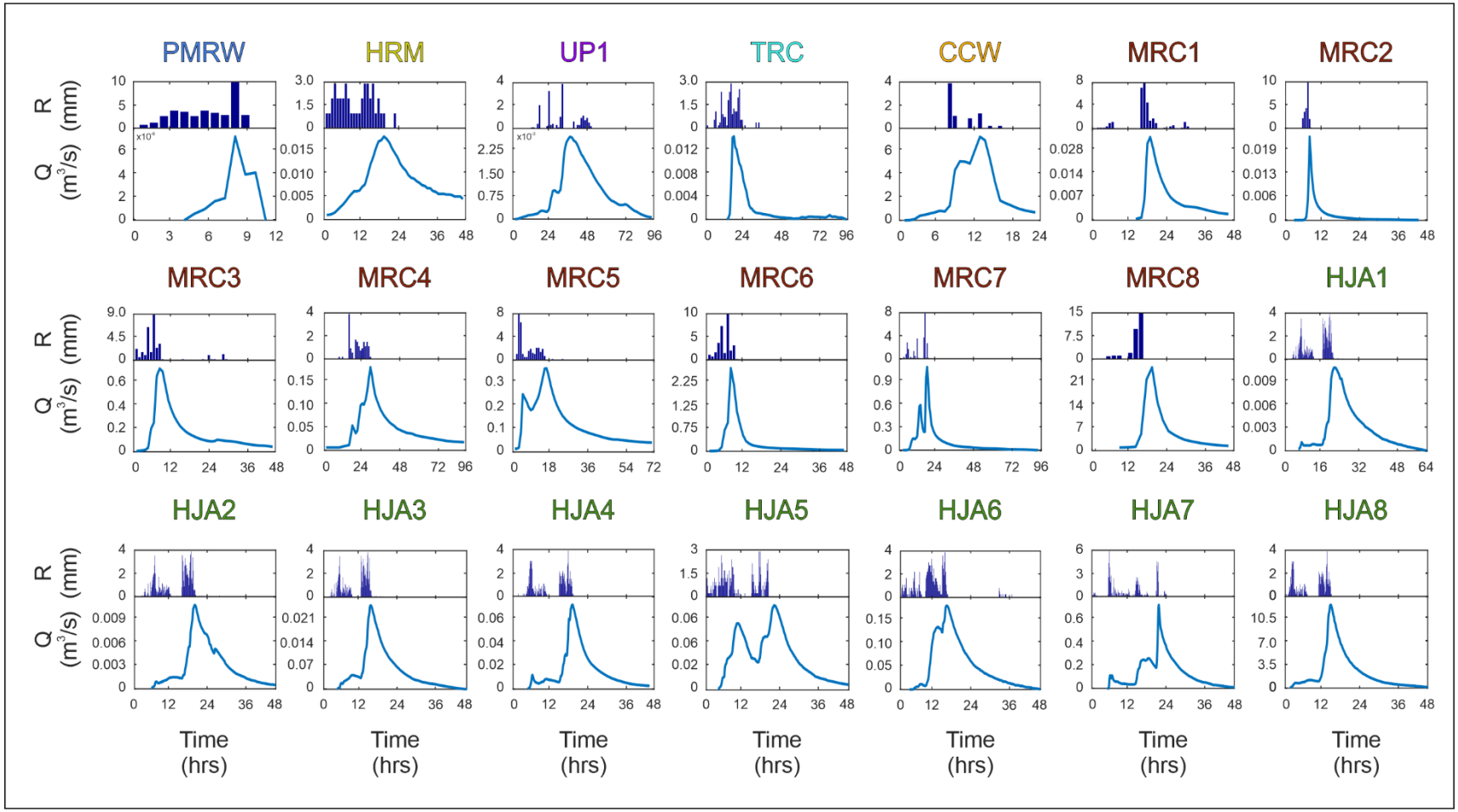


Figure 2-2. Site-specific examples of rainfall-runoff events delineated using HydRun. Event hyetograph and hydrographs are represented using bar and line charts, respectively, and are shown to illustrate typical hydrological dynamics in response to "average" event rainfall amounts (see text). R: rainfall; Q: discharge.

For each of the identified rainfall-runoff events, response magnitude, and response timing metrics were derived (Figure 2-3 – arrow 1). Response magnitude metrics include the runoff ratio (RR), peak discharge (Q_{MAX}), total runoff (Q_{TOT}), and initial abstraction (I_{abs}), while response timing metrics include the response lag (T_{LR}), time of rise (T_r), lag-to-peak (T_{LP}), centroid lag (T_{LC}), and time of concentration (T_c) (Table 2-2). Event-specific storage- and intensity-driven meteorological factors were also derived using site-specific rainfall and temperature records (Figure 2-3 – arrow 2). Storage-driven meteorological factors include the total event rainfall (R_{TOT}) and cumulative rainfall over a range of antecedent temporal windows (AR_x values computed over x days before the event, with x varying between 1 and 30). Compound storage-driven meteorological factors (e.g., R_{TOT}+AR_7) were also estimated to account for both event rainfall and antecedent moisture conditions. For intensity-driven meteorological factors, R_{TOT} was divided by the rainfall event duration to determine the average event rainfall intensity (RI_{AVG}). Hourly rainfall intensities were examined to identify the maximum event rainfall intensity (RI_{MAX}). Temperature data were used to derive potential evapotranspiration over a range of antecedent temporal windows ($APET_x$ – see above for details on antecedent window) using the Hargreaves equation (Hargreaves & Samani, 1982). Notably, while daily temperature data facilitated the estimation of $APET_x$ via the Hargreaves equation for the observation periods, longer-term daily temperature data records were not available for all sites, which is why mean annual PET (Table 2-1) was estimated using the Thornthwaite equation.

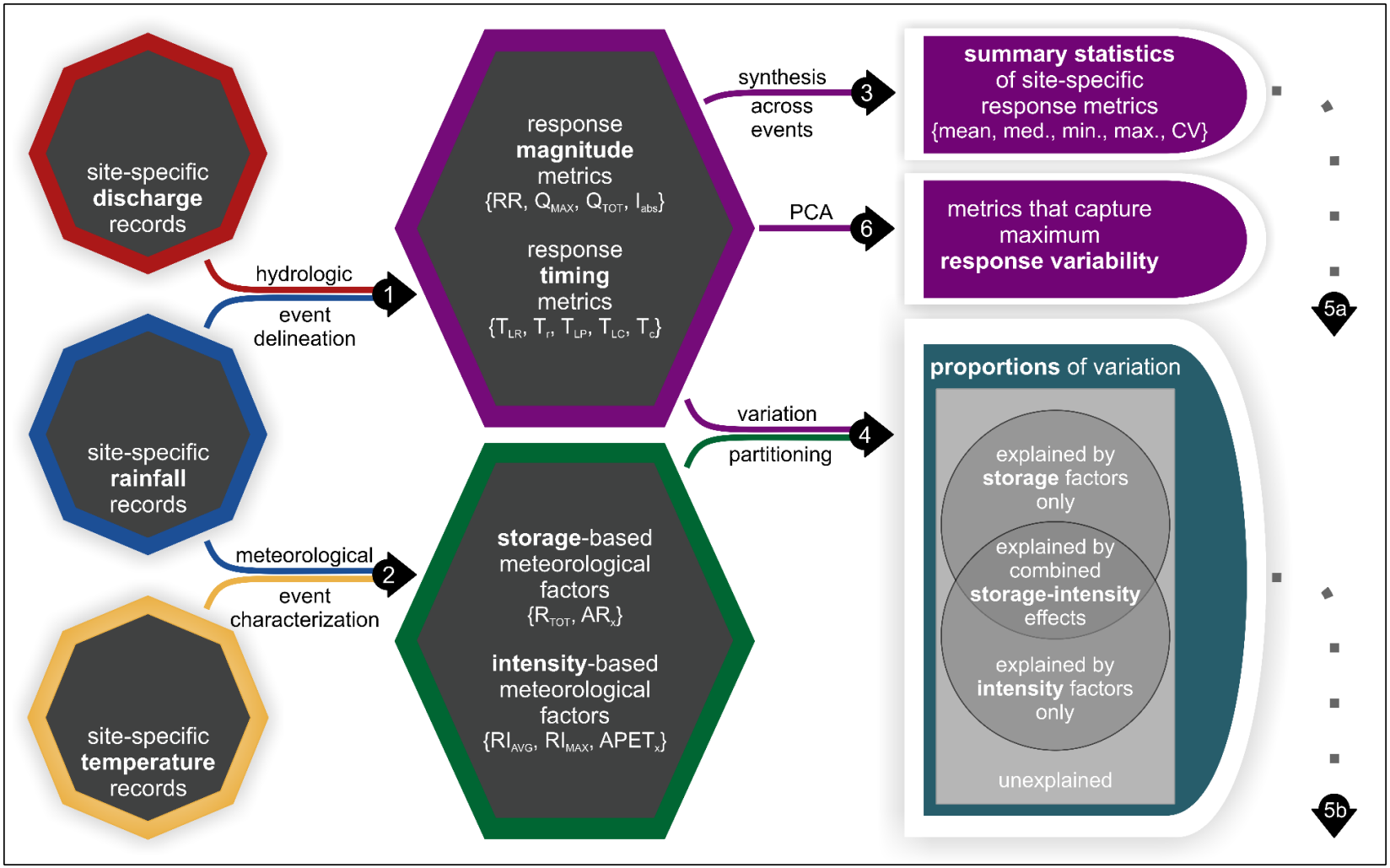


Figure 2-3. The methodological approach that was taken in this study. Numbered arrows represent key steps for cross-referencing with text. Elements connected to Steps 5a and 5b indicate input data for partial correlation analysis. See text and Table 2-2 for full names of response metrics and meteorological factors. CV: coefficient of variation; max.: maximum; med.: median; min.: minimum.

Table 2-2. Metrics used to describe event response.

Response metric	Definition
Response magnitude metrics	
Runoff ratio (RR)	The fraction of event rainfall that becomes runoff
Peak discharge (Q_{MAX})	Area-normalized maximum event discharge
Total runoff (Q_{TOT})	Area-normalized total event discharge
Initial abstraction (I_{abs})	Storage deficit satisfied before hydrograph response, estimated as the amount of event rainfall occurring before the initial hydrograph rise
Response timing metrics	
Response lag (T_{LR})	Time elapsed between the beginning of rainfall and the initial hydrograph rise
Time of rise (T_r)	Time elapsed between the beginning of hydrograph rise and peak discharge
Lag-to-peak (T_{LP})	Time elapsed between the beginning of rainfall and peak discharge
Centroid lag (T_{LC})	Time elapsed between the centroid of the hyetograph and the centroid of the hydrograph
Time of concentration (T_c)	Time elapsed between the end of rainfall and the end of hydrograph response

2.2.3 Statistical analyses

In relation to the first research question, the variability in site-specific response metrics was assessed using boxplots and summary statistics including the mean, median, minimum, maximum, standard deviation, and coefficient of variation across all events (Figure 2-3 – arrow 3). Scatter plots were also built to compare site-specific coefficients of variation for response metrics and coefficients of variation for meteorological factors. This was done to examine whether individual sites tend to have greater variability in response metrics than in meteorological factors, or vice versa.

In relation to the second research question, for each site, the relative influence of storage- and intensity-driven meteorological factors on response metrics was evaluated via variation partitioning. Variation partitioning is typically used to separate the variation of response variables according to the influence of sets of explanatory variables using a series of regressions

or canonical analyses (Ali et al., 2010; Borcard et al., 2011). Here, variation partitioning was used to determine the proportion of variation in response metrics that is: (1) explained uniquely by storage-driven meteorological factors (pure storage effects); (2) explained uniquely by intensity-driven meteorological factors (pure intensity effects); (3) jointly explained by storage- and intensity-driven meteorological factors (combined storage-intensity effects); (4) unexplained by either storage- or intensity-driven meteorological factors (Figure 2-3– arrow 4). Variation partitioning was performed three times to evaluate differences in variability based on the type of response metric considered: first considering all response metrics, then response magnitude metrics only, and then response timing metrics only. Analyses were carried using functions from the *Vegan* package (Dixon & Palmer, 2003; Oksanen et al., 2007) in the R environment (R Core Team, 2017).

In relation to the third research question, partial Spearman correlation coefficients were computed. Partial correlation analyses quantify the strength of a relationship between two variables while removing the influence of a set of other variables (Sokal & Rohlf, 1995). Here, partial correlation analyses were performed between summary statistics of response metrics (i.e., minimum, maximum, mean, median, and coefficient of variation computed across all events at each site) and site characteristics (Figure 2-3– arrow 5a). Each site characteristic (from Table 2-1) was considered while removing the influence of the remaining site characteristics (e.g., a partial correlation coefficient was computed between mean RR and drainage area while controlling for all the other physiographic variables and long-term climate characteristics listed in Table 2-1). Similarly, partial correlation coefficients were computed between variation partitioning proportions and site characteristics (Figure 2-3 – arrow 5b) to evaluate if the influence exerted by storage- and intensity-driven meteorological factors on response metrics is,

itself, dictated by site characteristics. All partial Spearman correlation coefficients (hereafter referred to as rho values) were derived using functions from the MATLAB® statistics toolbox.

Lastly, to address the fourth research question, principal component analysis (PCA) was used to identify the response metrics that best capture response variability (Figure 2-3 – arrow 6). PCA synthesizes data into a set of compound ordination axes, the principal components (PCs), which are uncorrelated and are ordered based on the proportion of variance captured from the original data (Legendre & Legendre, 2012). Here, site-specific analyses were performed, whereby the data fed into each PCA comprised all event response metrics for a given site. The first three PCs were retained so that a high percentage of response variability could be captured (>70%). One of the advantages of PCA is that it produces a PC loading matrix that quantifies the contribution of each original variable to the variance explained by each PC (Legendre & Legendre, 2012). Response metrics with loading values of |0.45| or more on at least one of the first three PCs were selected as the metrics that best capture response variability. Site-specific PCA was done using functions from the *Vegan* package in R (Dixon & Palmer, 2003; Oksanen et al., 2007).

2.3 Results

2.3.1 Variability in event meteorological factors and response metrics

Summary statistics of event-specific storage- and intensity-driven meteorological factors varied significantly across sites (Table 2-3 and Table 2-4). For storage-driven factors, the median R_{TOT} was highest for the HJA sites (> 53.4 mm) and lowest for the CCW (5.6 mm). R_{TOT} was

most variable at the CCW and MRC sites, and least variable at the PMRW, as indicated by the coefficient of variation values (Table 2-3). Medians of AR_7 were greatest for the HJA sites (> 28.0 mm) and lowest for the CCW (5.7 mm). AR_7 was most variable at the MRC sites and least variable at the TRC, as indicated by coefficients of variation (Table 2-3). For intensity-driven meteorological factors, $APET_7$ had relatively low temporal variability for most sites, with the highest median values at the PMRW and CCW. Median values of RI_{AVG} and RI_{MAX} were greatest at the PMRW site (RI_{AVG}: 1.3 mm/hr, RI_{MAX}: 10.2 mm/hr) and lowest at the UP1 site (RI_{AVG}: 0.2 mm/hr, RI_{MAX}: 1.9 mm/hr). Variability in RI_{AVG} and RI_{MAX}, portrayed by coefficients of variation, was greatest at the HRM and MRC4 sites, respectively, and lowest at the HJA sites.

Table 2-3. Site-specific event summary statistics for select storage-driven meteorological factors. med: median, min: minimum, max: maximum, CV: coefficient of variation. Refer to the text for the meaning of other abbreviations.

	R_{TOT} (mm)					AR_7 (mm)					R_{TOT} + AR_7 (mm)				
	Mean	Med.	Min.	Max.	CV	Mean	Med.	Min.	Max.	CV	Mean	Med.	Min.	Max.	CV
PMRW	36.3	39.9	16	66.5	0.5	30.2	18.8	5.3	67.6	0.8	66.5	68.8	26.4	103.1	0.4
HRM	29.7	27	7	95	0.7	22.1	19	0	91	0.9	51.8	43.5	12	142	0.5
UP1	29.4	22.1	1	92.2	0.8	16.8	9.9	0.1	51.1	0.9	46.1	46.6	4.7	143.3	0.7
TRC	30.9	25.2	5.2	85.4	0.8	19.5	15.8	3	62.4	0.7	50.5	33.4	16.4	147.8	0.7
CCW	12.6	5.6	0.6	82.4	1.2	11.4	5.7	0	54.6	1.2	24	14.1	1.2	120.4	1.1
MRC1	33.4	28.2	6	105.4	0.6	39.8	30.7	1.2	161.4	0.9	73.2	61.1	18	266.8	0.6
MRC2	32.2	21.2	4.8	146	1	32.7	20.8	0	156	1.1	65	50.4	11	246.4	0.8
MRC3	38.6	24.7	1.8	230.8	1	34.6	22.4	0	211	1.2	73.2	53.4	13.6	345.6	0.8
MRC4	38.4	25.2	2.2	171.6	1	47.3	24.2	0.2	282.8	1.1	85.8	63.1	9.2	286.2	0.8
MRC5	30.4	21.3	4.4	141	1	33.9	20.7	0.2	152	1	64.3	52.1	6.2	280.8	0.8
MRC6	31	19.8	1.8	278.8	1.3	43.5	22.9	0	279	1.3	74.6	53.3	10	555.6	1
MRC7	36.9	24.4	4.6	249.4	1.1	37.8	21.6	0	208.8	1.1	74.7	57.9	12.4	326	0.9
MRC8	36.3	23	1.2	276.8	1.1	37.2	23.4	0	211.4	1.2	73.6	54.4	7.2	285.4	0.8
HJA1	118.4	94.9	11.1	449.8	0.8	40.8	30.3	0	306.5	1	159.2	131.6	21.4	756.3	0.8
HJA2	115	87.2	3	458.6	0.9	43.9	29.8	0	306.2	1	159	136.6	6.9	756	0.8
HJA3	119.7	93.2	12.8	458.6	0.8	46.2	28	0	306.2	1.1	165.9	135.5	16.3	756	0.8
HJA4	101.3	87.5	18.7	334.4	0.6	57	34.5	0	312.3	1	158.3	142.7	35.1	449.1	0.5
HJA5	86.1	79.5	18.7	250.1	0.6	66.9	49.5	0	312.1	0.9	153	136.2	32.8	443	0.5
HJA6	76.3	53.4	4.5	334.4	0.9	52.8	32.1	0	312.3	1	129.1	114.3	11.9	449.1	0.7
HJA7	95.6	61.1	1.9	449.8	0.9	55.9	34.6	0	306.5	1.1	151.5	115.5	19.2	756.3	0.9
HJA8	93	75.4	7.5	438.7	0.8	47.2	29.9	0	312.3	1.1	140.2	120.4	15.8	474.4	0.6

Table 2-4. Site-specific event summary statistics for select intensity-driven meteorological factors. med: median, min: minimum, max: maximum, CV: coefficient of variation. Refer to the text for the meaning of other abbreviations.

	$APET_7$ (mm)					RI_{AVG} (mm/hr)					RI_{MAX} (mm/hr)				
	Mean	Med.	Min.	Max.	CV	Mean	Med.	Min.	Max.	CV	Mean	Med.	Min.	Max.	CV
PMRW	16.3	14.4	4.2	31.3	0.4	1.5	1.3	0.3	3.6	0.6	11.1	10.2	3.3	35.1	0.8
HRM	5.2	5.2	0.5	10.6	0.6	2.1	1.2	0.1	12	1.3	8.3	6	1	38	0.9
UP1	3.7	3.7	0.6	7.6	0.6	0.3	0.2	0	1.6	1.1	2.9	1.9	0	20.8	1.4
TRC	3.3	2.6	1.6	6.9	0.5	0.3	0.2	0.1	0.8	0.6	4	3.4	1.2	13.4	0.6
CCW	10	10.6	2.9	16.6	0.4	0.6	0.4	0	3.3	1	3.9	2.2	0.4	15.4	1
MRC1	4.9	3.8	1.9	14	0.6	1	0.8	0.3	4	0.7	7.9	6	2.6	47.8	0.9
MRC2	4.8	4.2	1.8	12.9	0.5	1.2	0.9	0.2	10	1.1	6.9	5.4	1.8	23	0.7
MRC3	5.8	4.7	2	15.8	0.5	1.1	0.7	0.2	6.6	0.9	8.2	5.6	1.6	32.8	0.8
MRC4	5.5	4.5	1.9	13.4	0.5	1.3	0.9	0.1	10.4	1.1	7.7	6.3	0.6	48	2.9
MRC5	5.2	4	1.8	14	0.6	1	0.6	0.2	5.3	1	6.3	5.3	1.2	20	0.6
MRC6	5.1	4.1	1.8	13.5	0.5	1	0.7	0.1	5.5	0.9	7.1	5.5	0.6	31.8	0.8
MRC7	5.4	4.5	2.1	11.8	0.5	1.3	1	0.2	5.3	0.7	7.4	6	1.8	23.2	0.7
MRC8	5.1	4.4	1.8	13.5	0.5	1.1	0.9	0	5.5	0.8	7.7	5.6	0.4	45.2	0.9
HJA1	7.6	5.8	0.3	22.7	0.7	0.6	0.6	0.1	1.6	0.5	5.4	4.9	1.1	10.8	0.4
HJA2	8.8	6.3	0.9	28.4	0.8	0.6	0.6	0.1	2.3	0.6	5.2	4.8	0.8	10.8	0.4
HJA3	8	5.6	0.8	25.6	0.8	0.7	0.6	0.1	1.6	0.5	5.2	4.8	1.1	10.8	0.4
HJA4	7.8	5.5	1.3	25.4	0.7	0.9	0.8	0.3	3.3	0.5	5.8	5.1	2.3	13.7	0.4
HJA5	8.4	5.6	1.5	29.7	0.8	0.9	0.7	0.2	3.3	0.6	5.4	4.8	2	13.7	0.4
HJA6	9.9	6.4	1.4	33.3	0.8	0.8	0.6	0.2	3.3	0.7	5	4.6	1.3	13.7	0.5
HJA7	10.2	6.9	0.9	37.1	0.9	0.7	0.6	0	1.9	0.6	4.9	4.2	1	10.8	0.5
HJA8	10.8	7.5	1.5	38.6	0.8	0.8	0.7	0.1	3.3	0.6	5.4	4.8	1.8	13.7	0.4

Mean RR (computed across all observed events) was 0.02 at the PMRW site, 0.31 at the HRM site, 0.21 at the UP1 site, 0.30 at the CCW site, 0.23-0.39 at the MRC sites, and 0.14-0.34 at the HJA sites. Other summary statistics of event response metrics are presented through boxplots in Figure 2-4. The RR metric was the least variable across sites, with medians between 0.09 and 0.38 – except for the PMRW which had a smaller median RR. T_c medians ranged between 41 and 69 hours at most sites but were generally lower at the PMRW, HRM, and CCW sites. Values of T_c were quasi-constant at the HRM, TRC, and CCW, while T_c varied more at other sites. Pronounced inter-site differences were observed with other metrics. For Q_{TOT}, medians and interquartile range were largest for the HJA sites (median > 5.6 mm) and smallest for the PMRW (median = 0.04 mm). Similarly, medians of I_{abs} were the greatest for the HJA sites (> 7.6 mm) and the smallest at the CCW (1.0 mm). The interquartile range for I_{abs} was largest for the HJA sites and smallest for the MRC sites. The medians and inter-quartile range for Q_{MAX} were the largest for the MRC sites and noticeably smaller for the PMRW, UP1, and CCW sites. When comparing response timing metrics (other than T_c) across sites, the greatest median values were observed for the HJA sites, notably for T_{LR} (> 23.0 hours), T_r (> 19.0 hours), T_{LP} (> 57.5 hours), and T_{LC} (> 31.8 hours). The inter-quartile range for those response timing metrics was largest at the HJA, UP1, and TRC, while medians and inter-quartile ranges were typically smallest at the MRC, CCW, HRM, and PMRW sites.

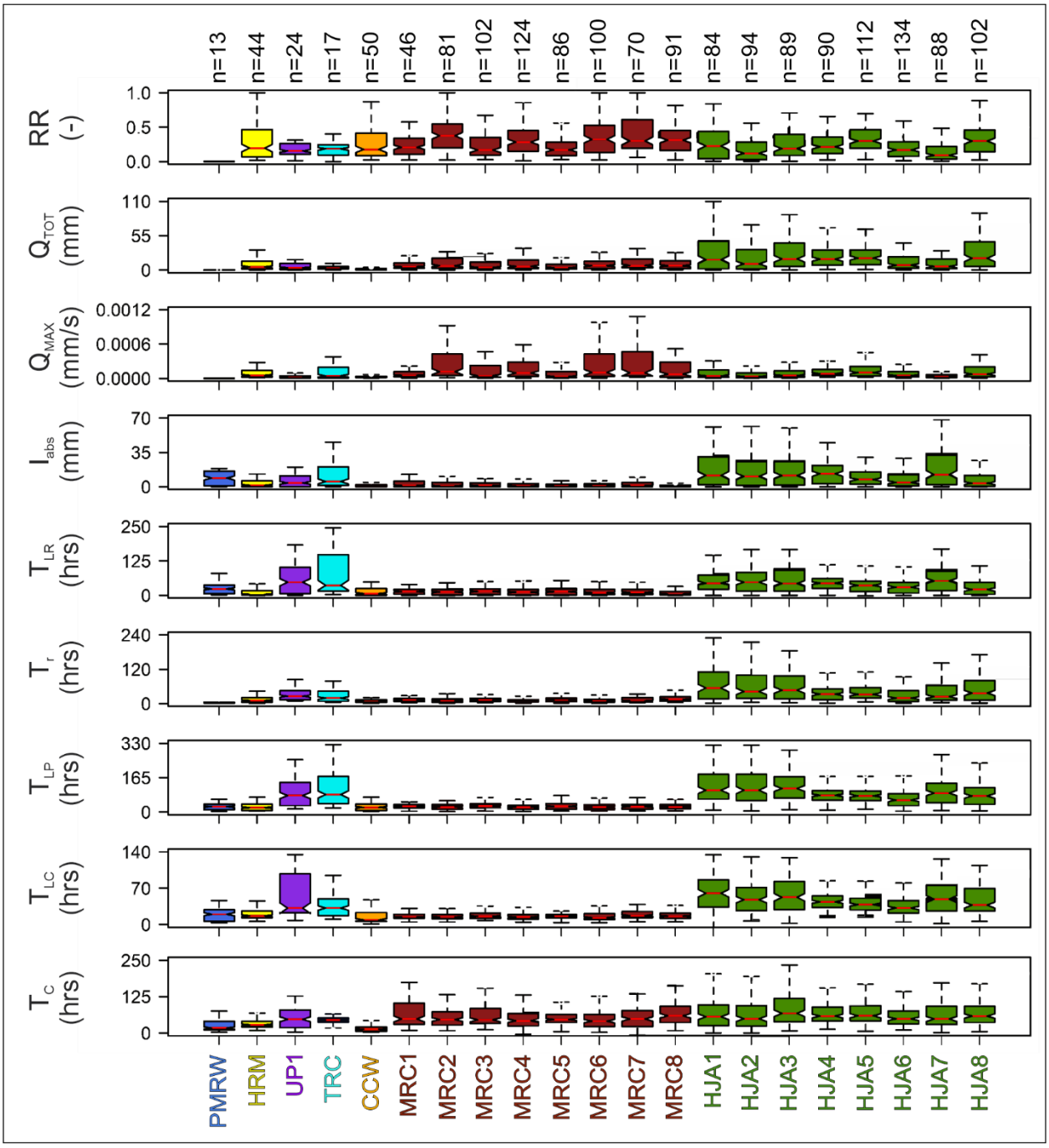


Figure 2-4. Boxplots showing temporal variability in response metrics at each site. The number of events for each site is shown by *n*. Horizontal red lines in boxplots are the median values computed across all events at each site and each box spans from the 25th to 75th percentiles.

Scatter plots comparing coefficients of variation of response metrics and meteorological factors are displayed in Figure 2-5. Event RR typically varied more than meteorological factors for the PMRW and TRC sites and was equally variable as meteorological factors for the HRM, UP1, MRC, and HJA sites. The Q_{TOT} and Q_{MAX} metrics were generally more variable than meteorological factors, except for the UP1 site. The T_{LR} metric varied more than meteorological factors at the PMRW, HRM, UP1, TRC, and CCW sites, and was as variable as most meteorological factors for the MRC and HJA sites. Similarly, T_r was more variable than meteorological factors at the PMRW, HRM, TRC, and HJA sites, equally variable at the CCW and MRC sites, and less variable at the UP1 site. The T_{LP} metric was more variable than storage-driven meteorological factors for the PMRW, HRM, and UP1 sites, equally variable as storage-driven meteorological factors at the TRC and HJA sites, and less variable than storage-driven meteorological factors at the CCW and MRC. Lastly, T_c was more variable than most meteorological factors for the PMRW, UP1, and MRC4, while for remaining sites, T_c was as variable as most storage-driven meteorological factors.

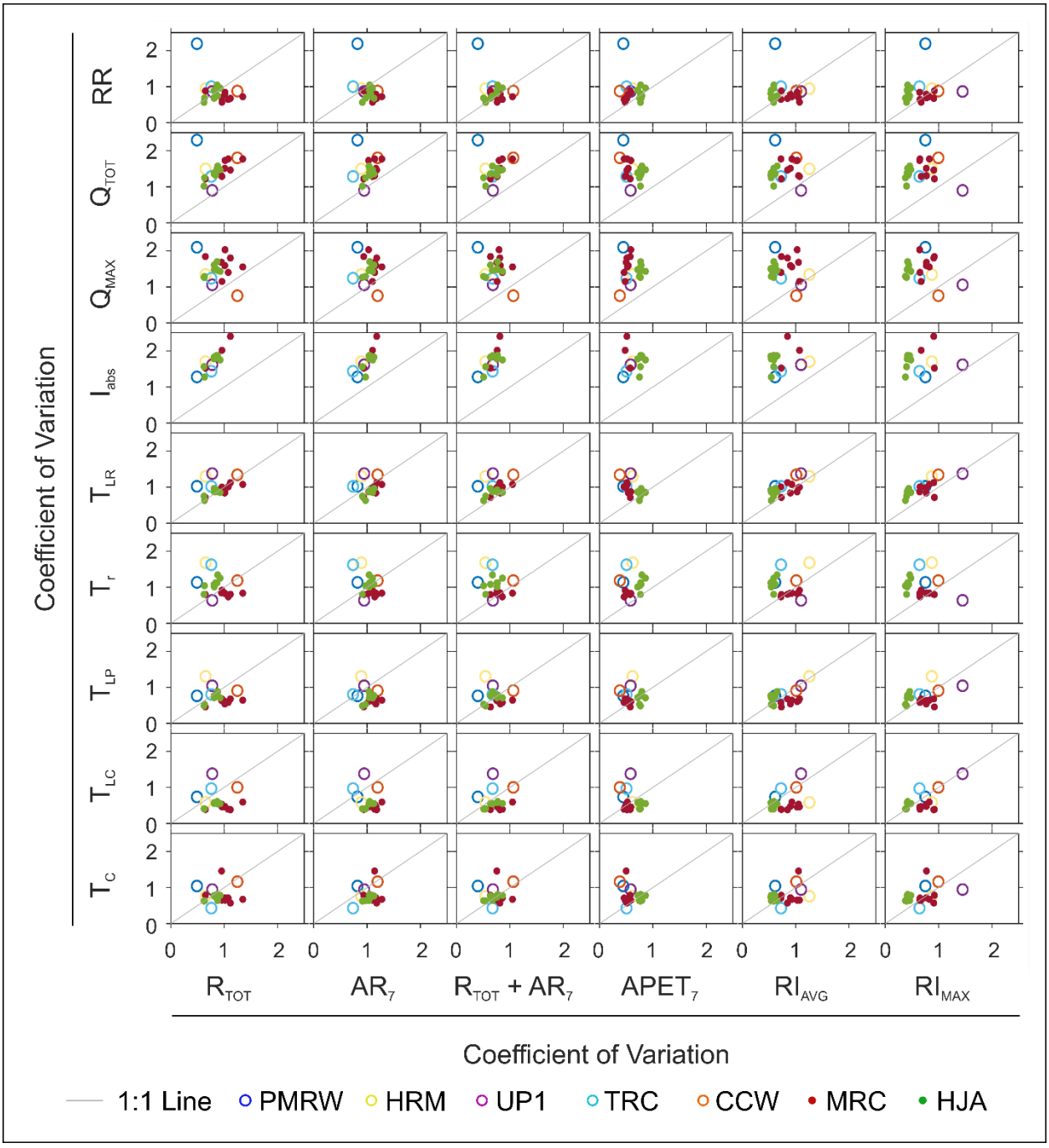


Figure 2-5. Scatter plots showing the relationship between the variability of response magnitude and timing metrics (y-axis, shown via the coefficient of variation) and the variability of storage- and intensity-driven meteorological factors (x-axis, shown as the coefficient of variation). Points/circles are for individual sites and are color-coded by study area.

2.3.2 Influence of meteorological factors on event response variability

Variation partitioning showed that across sites, 50% or less of temporal variability in all response metrics was explained by the storage- and intensity-driven meteorological factors considered in the present study (Figure 2-6A). This means that a large proportion of the variability in all response metrics was attributable to factors other than those considered in the present study. The variability that could be explained was not controlled by the same factors at all sites. For instance, at the CCW and MRC8 sites, the temporal variability in all response metrics was primarily attributable to pure storage effects, followed by combined storage-intensity effects (Figure 2-6A). At other sites such as UP1, MRC1, HJA4, HJA5, and HJA8, however, combined storage-intensity effects only explained a small proportion (< 3%) of the variability in all response metrics.

When only response magnitude metrics were considered, the total proportion of variability explained was higher than when all response metrics were considered, generally 50% or more (Figure 2-6B). At the UP1, MRC2, MRC4, MRC6, MRC7, and all HJA sites, variability in response magnitude was primarily attributable to pure storage effects, followed by combined storage-intensity effects and then pure intensity effects. At the MRC1, MRC3, MRC5, MRC8, and CCW sites, however, combined storage-intensity effects explained most of the variability in response magnitude metrics (Figure 2-6C). The temporal variability in response timing metrics was the most difficult to explain, i.e., the analyses led to the lowest proportions of explained variation (Figure 2-6C). Response timing variability at 17 of 21 sites was primarily attributable to pure storage effects, followed by pure intensity effects and combined storage-intensity effects. The only exceptions were the MRC1, MRC3, and MRC5 sites where response timing variability

was mostly due to pure intensity effects, and the CCW where combined storage-intensity effects were the main source of response timing variability.

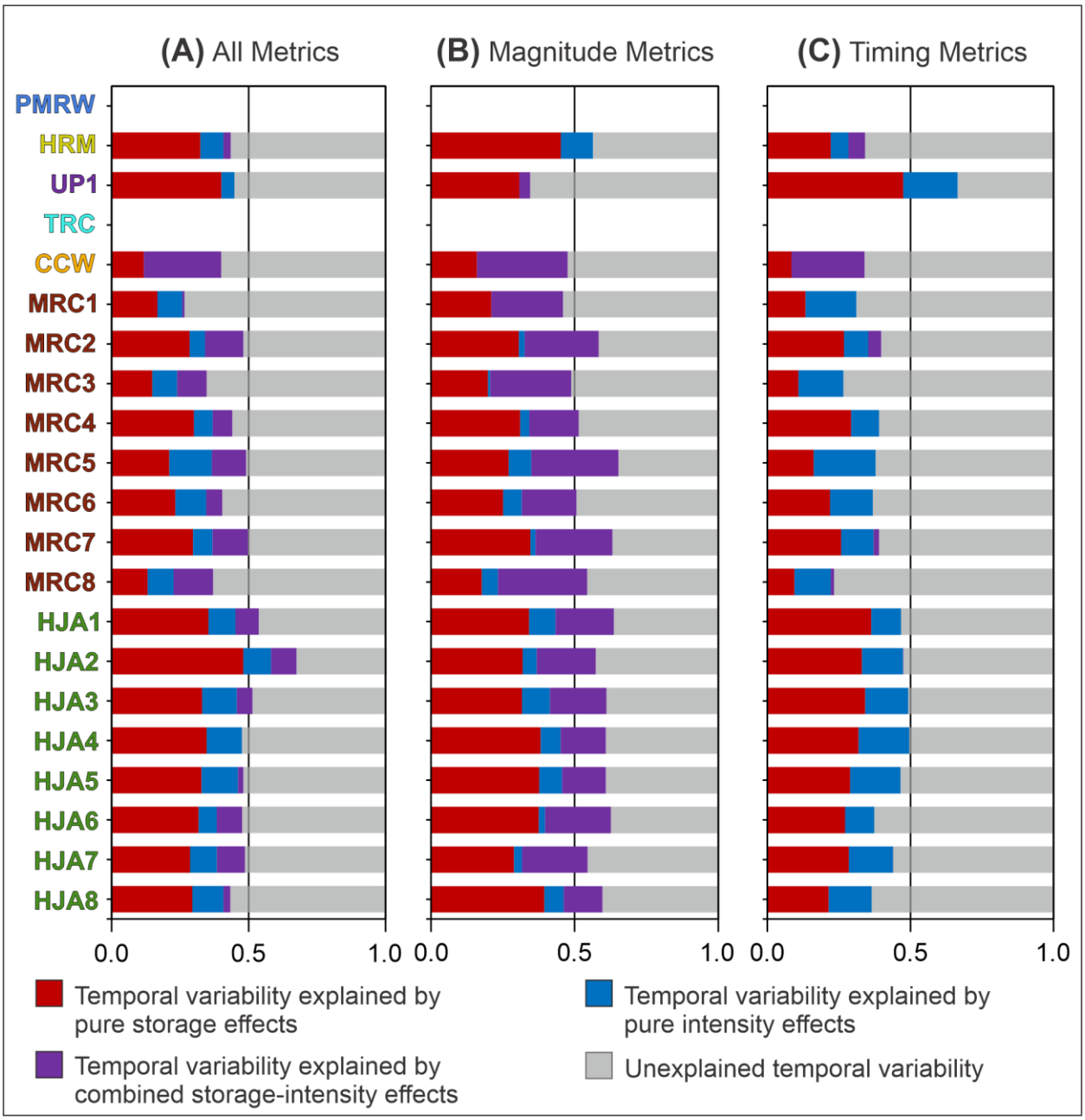


Figure 2-6. Stacked bars indicating proportions of temporal variability in all response metrics (A), response magnitude metrics only (B), and response timing metrics only (C) that are explained (and unexplained) by pure storage effects, pure intensity effects, and combined storage-intensity effects. When no bars are shown, collinearity between meteorological factors was detected and variation partitioning was not performed.

2.3.3 Influence of site characteristics on event response

Partial correlation coefficients (rho) between summary statistics of response metrics and site characteristics are presented in Table 2-5. All correlations that were statistically significant at the 95% level are reported, but only the strongest (here, |rho| > 0.65 are considered representative of moderate to strong correlations) are discussed and interpreted below. Negative correlations between summary statistics of response timing metrics and site characteristics were most common: mean annual temperature and mean annual PET were strongly correlated with mean T_r, $T_{LP,}$ and T_{LC} (rho <= -0.69); median T_{LR}, T_r, $T_{LC,}$ and T_{LP} (rho <= -0.65); and maximum T_{LR}, T_{LC} and T_{LP} (rho <= -0.68). Summary statistics of response magnitude metrics were generally positively correlated with mean annual precipitation: those correlations were strongest for maximum Q_{TOT} (rho = 0.69).

Table 2-5. Partial correlation coefficients (rho) between summary statistics of response metrics and select site characteristics. Only statistically significant rho values at the 95% level ($p < 0.05$) are displayed. Med.: median, Min.: minimum, Max.: maximum, CV: coefficient of variation, DA: site drainage area, T: mean annual temperature, P: mean annual precipitation, PET: mean annual potential evapotranspiration.

			DA	T	P	PET			DA	T	P	PET
Mean	Response magnitude metrics	RR					Response timing metrics	T_{LR}	-0.53	-0.64		-0.60
		Q_{TOT}			0.65	-0.48		T_r		-0.69		-0.76
		Q_{MAX}		0.56	0.50			T_{LC}	-0.56	-0.79		-0.76
		I_{abs}		-0.53		-0.54		T_{LP}		-0.71		-0.72
								T_c			0.59	
Med.		RR						T_{LR}	-0.57	-0.65		-0.62
		Q_{TOT}			0.56			T_r		-0.68		-0.75
		Q_{MAX}						T_{LC}		-0.71		-0.74
		I_{abs}		-0.58		-0.54		T_{LP}		-0.67		-0.69
								T_c			0.57	
Min.		RR			0.49			T_{LR}			-0.54	
		Q_{TOT}			0.54			T_r		-0.48		-0.70
		Q_{MAX}						T_{LC}				
		I_{abs}						T_{LP}		-0.48		-0.48
								T_c				
Max.		RR						T_{LR}	-0.58	-0.73		-0.68
		Q_{TOT}			0.69			T_r	-0.46	-0.63		
		Q_{MAX}		0.57				T_{LC}	-0.63	-0.75		-0.68
		I_{abs}			0.61			T_{LP}	-0.55	-0.80		-0.81
								T_c				
CV		RR			-0.50			T_{LR}				
		Q_{TOT}		0.50		0.55		T_r				
		Q_{MAX}		0.58		0.60		T_{LC}		-0.56	-0.59	
		I_{abs}						T_{LP}		-0.63		-0.57
								T_c				

There were also statistically significant correlations between proportions of explained variability in response metrics – obtained from variation partitioning – and all site characteristics except relief and mean slope (Table 2-6). The strongest relationships were between the importance of combined storage-intensity effects on response magnitude and mean annual values of temperature (rho = 0.81) and PET (rho = 0.79). There were also strong, negative correlations between the importance of pure storage effects on response timing and drainage area and mean annual values of temperature and PET (rho <= -0.75). The proportion of variability in all response metrics explained uniquely by storage-driven meteorological factors was strongly correlated with drainage area (rho = -0.68), mean annual values of temperature (rho = -0.77), and PET (rho = -0.78). There was also a strong correlation between the importance of combined storage-intensity effects on response timing metrics and drainage area (rho = 0.77).

Table 2-6. Partial correlation coefficients (rho) between variation partitioning fractions and select site characteristics. Only statistically significant rho values at the 95% level ($p < 0.05$) are displayed. DA: site drainage area, SD Slope: standard deviation of site slope, T: mean annual temperature, P: mean annual precipitation, PET: mean annual potential evapotranspiration.

	Meteorological effects (or lack thereof) on event response variability (as estimated through variation partitioning)	DA	SD Slope	T	P	PET
All response metrics	Pure storage effects	-0.68		-0.77		-0.78
	Pure intensity effects					
	Combined storage-intensity effects	0.63		0.53		0.52
	Unexplained					
Response magnitude metrics	Pure storage effects			-0.55		-0.57
	Pure intensity effects		-0.53			
	Combined storage-intensity effects	0.59		0.81		0.79
	Unexplained					
Response timing metrics	Pure storage effects	-0.75		-0.78		-0.78
	Pure intensity effects	-0.51				
	Combined storage-intensity effects	0.77				
	Unexplained					

2.3.4 Metrics capturing the maximum variability in site-specific hydrologic response

The site-specific PCA of all response metrics revealed that between 73% and 96% of the variance was captured by the first three PCs (Figure 2-7). Depending on the site considered, the number of metrics with loadings of |0.45| or more on the first three PCs ranged between two (PMRW) and seven (MRC1 and HJA5). For 14 out of 21 sites, response timing metrics were the most important contributors to the first three PCs. However, for the MRC7, HJA7, and HJA8 sites, the same numbers of response magnitude metrics and response timing metrics had high loadings on the first three PCs. For the TRC, MRC4, and HJA5 sites, it was mostly response

magnitude metrics that had high loadings on the first three PCs. As a result, there was a large range in the number of sites for which individual metrics were identified as the biggest contributors to the first three PCs and hence deemed most important for capturing intra-site (temporal) response variability. For response timing metrics, T_{LR}, $T_{LP,}$ and T_c were deemed most important for 16, 16, and 15 sites out of 21, respectively. As for response magnitude metrics, I_{abs} and Q_{TOT} were deemed important for capturing the temporal variability in hydrologic response for 15 and 10 sites out of 21, respectively. Metrics RR and Q_{MAX} were deemed most important for capturing response variability for less than 10 sites out of 21.

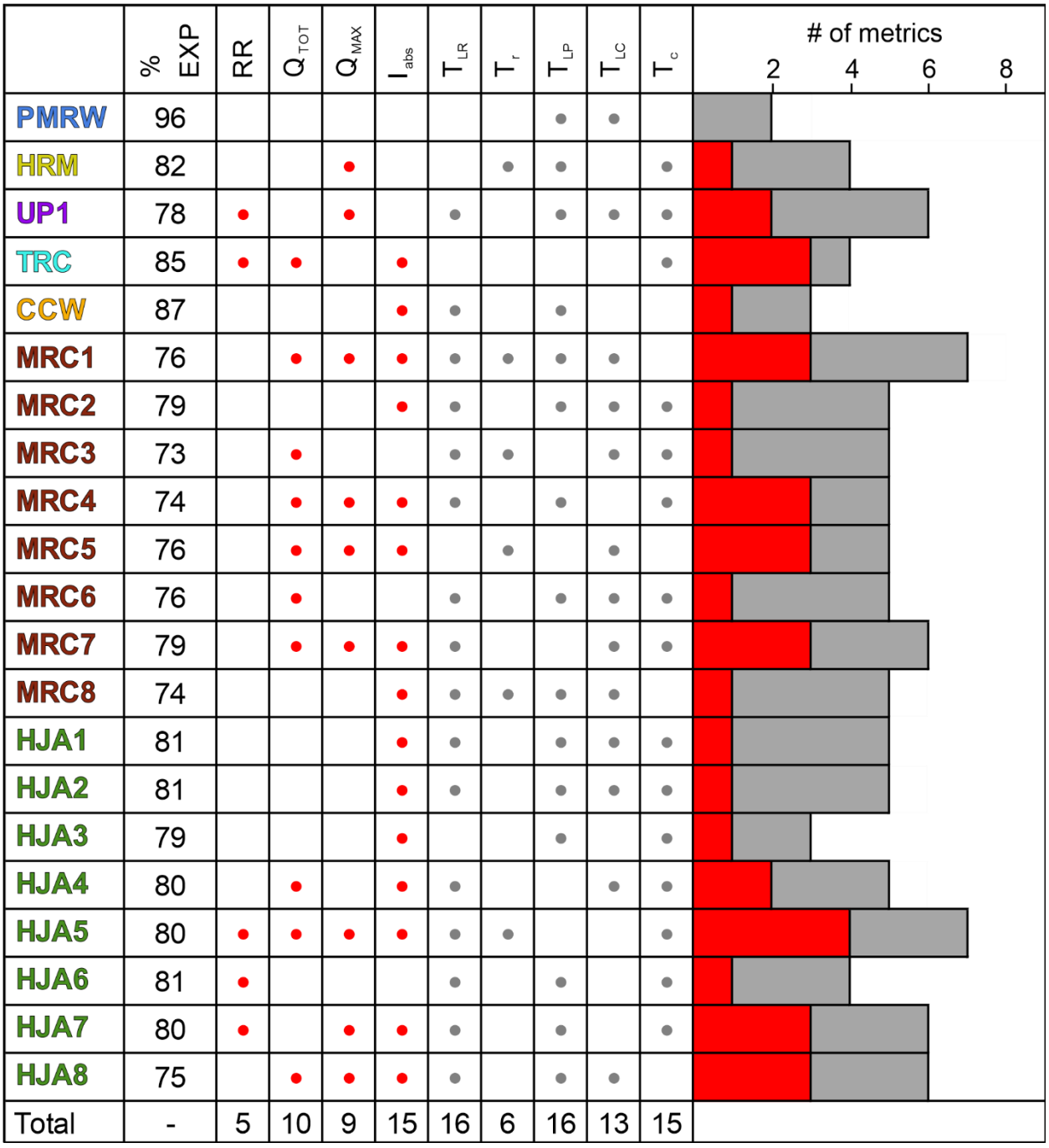


Figure 2-7. Summary of principal component analysis results. Red and grey dots show response magnitude and timing characteristics that have significant loadings (>|0.45|) on the first three principal components (PCs). Red and grey bars show the relative importance of response versus timing metrics as major contributors to the first three PCs. The total percentage of intrasite

(temporal) variability in hydrograph response captured by the first three PCs is reported in the "% EXP" column. The number of sites for which each metric was deemed most important for explaining intrasite (temporal) response variability is reported in the "Total" (bottom) row. Refer to the text for the meaning of abbreviations.

To take the results of PCA further, magnitude and timing metrics that were deemed most important for capturing temporal variability in response across most sites were used to classify rainfall-runoff events. Site-specific scatter plots were built, where response magnitude and timing metrics were plotted on the x- and y-axis, respectively. Plot axes ranged from 0 to the 75^{th} percentile plus 1.5 times the inter-quartile range of each site-specific metric to capture most events while excluding outliers. Each axis was then split by the median of the corresponding metric, resulting in a four-quadrant empirical classification space indicating whether an event response could be labelled as a) low magnitude and fast timing; b) low magnitude and slow timing; c) high magnitude and fast timing, or d) high magnitude and slow timing. The classification space was used to identify the dominant event type at each site and not to assess any causal relationship between response magnitude and timing. Event classification using I_{abs} and T_{LR} (Figure 2-8) showed that across most sites, low magnitude and fast timing events (i.e., events plotting in the lower left quadrant) were dominant. This was particularly visible at the HRM, TRC, CCW, and MRC8 sites where median values of I_{abs} and T_{LR} were small – compared to the range of observed values. The second most common event type was high magnitude and slow timing (i.e., events plotting in the upper right quadrant), and for many sites, multiple events plotted directly on quadrant boundaries. When event classification was done using Q_{TOT} and T_{LP} (Figure 2-9), low magnitude and fast timing events were dominant at all sites except the PMRW

where low magnitude and slow timing events were the most numerous. Generally, high magnitude and slow timing events were also common – except for at the PMRW, TRC, CCW, MRC1, MRC3, and MRC4 sites where low magnitude and slow timing events were more numerous. This qualitative analysis, therefore, revealed some commonalities between sites – in terms of dominant event types –, although results for the PMRW, UP1, and TRC sites need to be interpreted with caution due to their relatively smaller number of events considered in the current paper, compared to other sites.

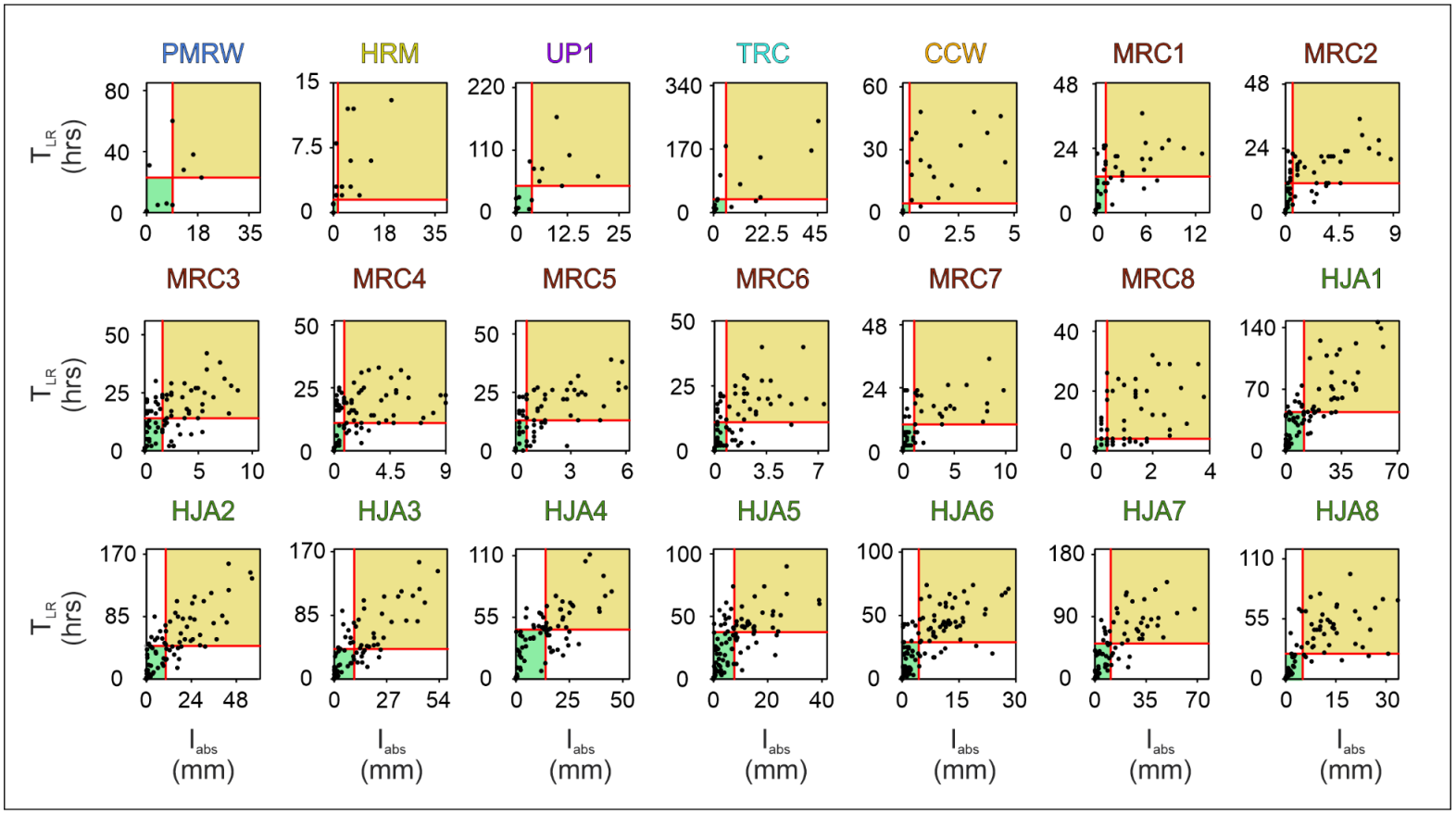


Figure 2-8. Site-specific scatter plots showing I_{abs} (x-axis) and T_{LR} (y-axis) response metrics. Black dots are for individual rainfall-runoff events, while red lines are the median values of magnitude and timing metrics across all events. Axes maxima are limited to the 75th percentile + 1.5 interquartile range. Low- and high-magnitude events are located to the left and right of the vertical red lines, respectively. Fast- and slow-timing events are located below and above the horizontal red lines, respectively. Green and yellow boxes indicate quadrants containing the largest and second-largest number of events, respectively. Refer to Table 2-2 for metric definitions.

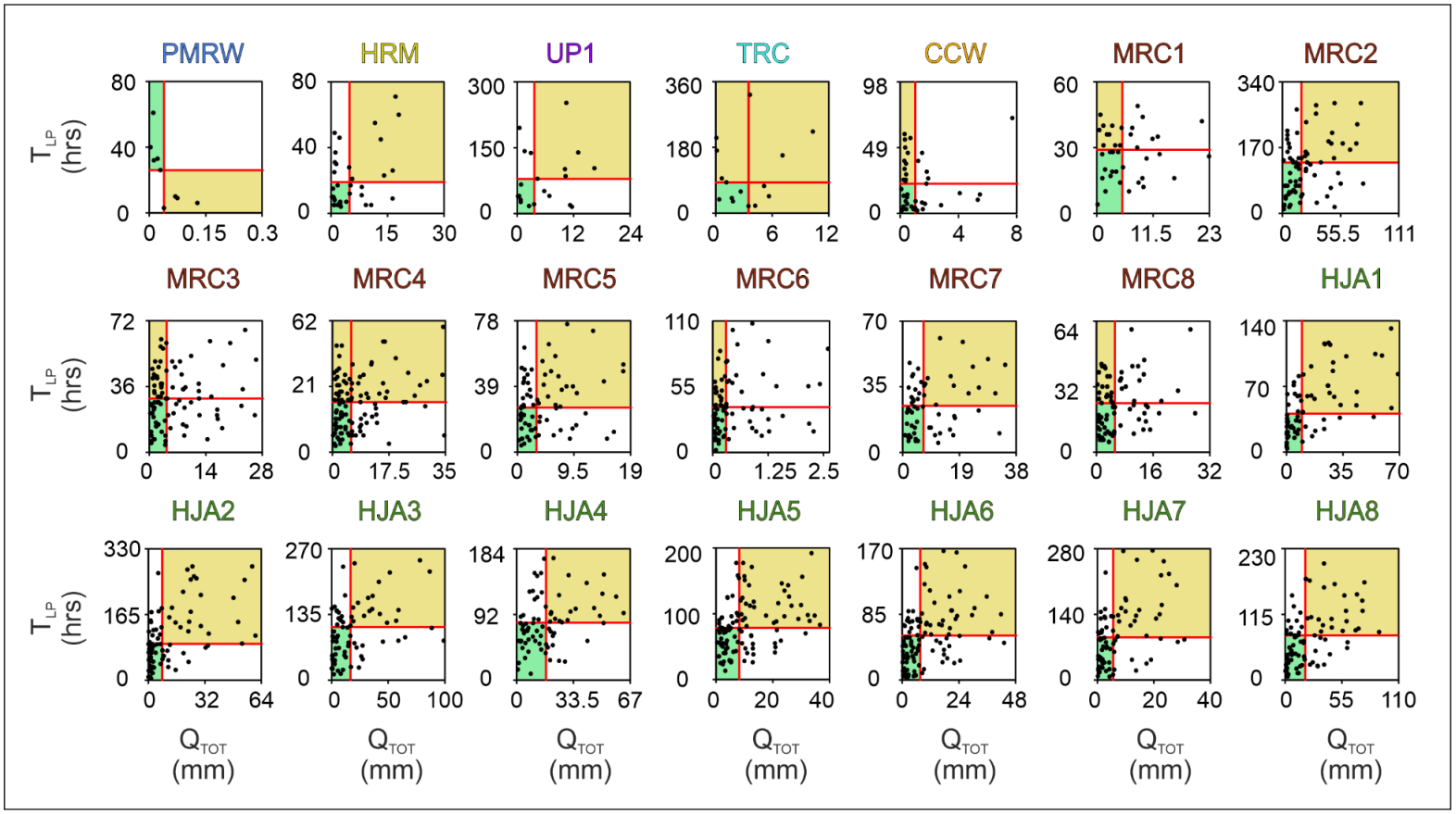


Figure 2-9. Site-specific scatter plots showing Q_{TOT} (x-axis) and T_{LP} (y-axis) response metrics. Black dots are for individual rainfall-runoff events, while red lines are the median values of magnitude and timing metrics across all events. Axes maxima are limited to the 75th percentile + 1.5 interquartile range. Low- and high-magnitude events are located to the left and right of the vertical red lines, respectively. Fast- and slow-timing events are located below and above the horizontal red lines, respectively. Green and yellow boxes indicate quadrants containing the largest and second-largest number of events, respectively. Refer to Table 2-2 for metric definitions.

2.4 Discussion

2.4.1 Controls of spatial variability in event response

Evaluating the predictability of event response metrics from site characteristics led to interesting findings about controls of spatial variability in event response, especially given the comprehensive character of the datasets relied upon in the current study. Indeed, despite differences in record length for the data examined here, the events captured for each site were very diverse and a reflection of wide-ranging conditions. Event-specific response metrics and meteorological factors varied significantly for most sites, as shown by site-specific boxplots (Figure 2-4) and coefficients of variation (Table 2-3 and Table 2-4). Research previously carried out at the 21 sites included in the current study reported comparable results for one key response metric, namely the runoff ratio or RR. The literature reports that mean event RR ranged from 0.0004 to 0.098 for the PMRW site (Tromp-van Meerveld & McDonnell, 2006b). Mean event RR was 0.30 at the HRM site (Ali et al., 2010) and 0.22 at the UP1 site (Oswald et al., 2011), and it was typically below 0.3 at the TRC site (Woods et al., 2013) and below 0.46 at the MRC sites (McMillan et al., 2014). Event RR ranged from 0 to 0.57 for HJA sites (Graham & McDonnell, 2010). Research works published for those study sites, however, typically did not report other response magnitude or timing metrics, preventing further comparisons between our findings and previously published work.

In relation to controls of spatial variability on event response, it was mostly response timing metrics that were strongly correlated with site characteristics. The strongest of those correlations were negative and involved mean annual temperature and mean annual PET. Other

studies have reported negative correlations between mean RR and mean annual PET (e.g., Merz & Blöschl, 2009). Negative correlations between mean annual temperature or PET and summary statistics of timing metrics suggest that sites with typically slower timing events have relatively high mean annual temperature and PET estimates, and vice versa. This result highlights differences in storage deficits needing to be met before flow is generated.

The results from the nested sites (i.e., HJA and MRC sites sorted in order of increasing drainage area) did not suggest any strong control exerted by watershed size on event response (Figure 2-4). While other studies reported positive correlations between drainage area and RR, Q_{MAX}, $T_{LC,}$ and T_{LP} values (Acreman & Sinclair, 1986; Dingman, 2015; Holtan & Overton, 1963), such correlations were either not statistically significant or not strong in the present study (Table 2-5). The lack of statistically significant correlations between drainage area and metrics such as RR and Q_{MAX} might simply be because those metrics were area-normalized. As for positive correlations between drainage area and response metrics such as T_{LC}, $T_{LP,}$ and T_{LR}, while they make sense because the time taken for rainfall to be routed to a channel and subsequently a watershed outlet is often greater for sites with larger drainage areas (Dingman, 2015), they were not found in the present study. It has long been acknowledged that the relationship between drainage area and rainfall-runoff processes is difficult to interpret due to a lack of uniformity within the landscape and because drainage area is related to other physiographic characteristics (Gregory & Walling, 1973). Thus, some have suggested that relationships between drainage area and response metrics are influenced by other factors like catchment morphology and network complexity (Beven et al., 1988). The methodological approach adopted in the current study did not allow us to confirm this, despite the reliance on partial correlation analysis and the consideration of physiographic variables such as elevation

range and standard deviation of the slope as indicators of topographic heterogeneity. It is also worth mentioning that partial correlation analysis was done both using area-normalized metrics (data shown in this paper), and metrics that were not area-normalized (data not shown in this paper). The results of those two sets of analyses were not significantly different and hence led to similar interpretations. The lack of correlations with drainage area or physiographic variables in the present study suggests that landscape characteristics alone cannot be used to predict either typical event response characteristics (i.e., mean and median metric values), or extreme event characteristics (i.e., minimum and maximum metric values), or the degree of event response temporal variability (i.e., coefficient of variation of metrics across all events).

There are several possible explanations for the lack of interpretable correlations between response metrics and physiographic variables. First, some have argued that controls on hydrological responses are interacting (Merz & Blöschl, 2009; Yadav et al., 2007) or hierarchal, with climatic controls taking precedence over geology, soil characteristics, and topography (Devito et al., 2005). Given the wide range of climatic conditions present across study sites, it is perhaps not surprising that physiographic controls were not dominant predictors of typical event response. Second, it has been proposed that there are three first-order controls on streamflow characteristics within a given climatic regime: the T^3 template of typology, topography, and topology (Buttle, 2006). However, the methodology adopted in the present study did not allow a full application of the T^3 template: the physiographic variables considered here (i.e., drainage area, relief, slope) encompass topography, but they do not explicitly account for differences in drainage network connectivity (topology) or partitioning between vertical and lateral flowpaths (typology), which are known to influence event response characteristics. Besides, the consideration of the spatial arrangement (i.e., areal extent and adjacency) of different types of

parent material, vegetation, soil texture, or percentages of soil organic matter content may have led to different correlation results. Such physiographic variables were not considered in the present study due to the difficulty in quantifying them with consistent data and methods across all sites.

2.4.2 Controls of temporal variability in event response

One of the research questions listed at the beginning of the current paper targeted the evaluation of the relative influence of storage- and intensity-driven meteorological factors on event response metrics, which had not been pursued before via variation partitioning. While variation partitioning has a relatively long history (Borcard et al., 1992), having been applied in thousands of ecological studies to model species-environment relationships (Peres-Neto et al., 2006), hydrological applications remain rare (e.g., Ali et al., 2010). Similarities can be drawn across applications – for example, variation partitioning exercises in natural systems often highlight a large portion (~50 %) of unexplained variability, regardless of the explanatory variables considered (Ali et al., 2010; Borcard et al., 1992; Borcard & Legendre, 1994). Hence, the results from the variation partition analyses performed in the present study are not unusual, with proportions of unexplained variability near or sometimes above 50% (Figure 2-6). Regardless of whether variation partitioning was performed considering all response metrics, magnitude metrics only, or timing metrics only, response variability was largely explained by pure storage effects for most sites. This finding is certainly aligned with countless other studies which have identified the importance of storage effects in explaining response variability in a wide range of environments (e.g., Bishop et al., 2011; McNamara et al., 2011; Oswald et al.,

2011; Seibert et al., 2011; Shaw et al., 2012; Spence, 2007; Tromp-van Meerveld & McDonnell, 2006b, 2006a). However, in the present study, pure-storage effects were not always the most important for explaining response variability: combined storage-intensity effects or pure intensity effects were deemed important at some sites to explain variability in response magnitude and timing metrics. It is also clear that combined storage-intensity effects often explained more variability in response magnitude metrics than pure intensity effects, and the reverse was true for response timing metrics. This suggests that intensity-driven meteorological factors, individually and in combination, are important for explaining variability in response metrics for many environments, including some with highly contrasted climatic and physiographic characteristics. At the event scale, intensity-driven meteorological factors are most commonly explored in circumstances where infiltration-excess overland flow is known or thought to dominate (Horton, 1933), which is usually the case in semi-arid to arid (e.g., Cammeraat, 2002) or in infiltration-limited (e.g., Granger et al., 1984, 1984) environments. The current study, however, shows that even at sites where saturation-excess flow processes have been observed, (e.g., sites such as the HRM, MRC, and UP1; Ali et al., 2011, McMillan et al., 2014, Oswald et al., 2011), pure intensity effects or combined storage-intensity effects were quantifiable (Figure 2-6).

The notion of combined storage-intensity effects is interesting, from a process standpoint, as it may either reflect additive (cumulative) influences of different types of meteorological factors or effects that are synergistic or interactive. However, from a statistical standpoint, interpreting combined storage-intensity effects is not straightforward as variation partitioning does not permit combined effects to be tested for statistical significance (Borcard et al., 1992). The statistical literature also insists that combined effects in variation partitioning analyses do not necessarily imply the interaction of terms, as is the case in ANOVA-like analyses, for

instance (Peres-Neto et al., 2006). Drawing from the ecological literature, it may also be the case that combined-effects reflect indirect links, i.e., cases where factors X and Y might be uncorrelated to one another but may both be influenced by the same third factor Z, thus creating apparent joint or combined effects (Borcard & Legendre, 1994). For the present study, based on previous statistical literature (Borcard et al., 1992; Borcard & Legendre, 1994; Gilbert & Bennett, 2010), it is hypothesized that combined storage-intensity effects represent redundancy in response variation that is explained by both pure storage effects and pure intensity effects. The present study results not only seem to indicate that intensity-driven meteorological factors are important for explaining response variability, but also that the relative importance of storage-driven and intensity-driven meteorological effects on response variability can be predicted from select site characteristics. For example, drainage area and mean annual values of temperature and PET were strong predictors of the effects of meteorological factors on response variability. More specifically, drainage area was positively correlated with combined storage-intensity effects on response timing metrics (Table 2-6), hinting that the larger the watershed, the greater the combined influence of storage-driven and intensity-driven meteorological factors on the timing of hydrologic response. Mean annual values of temperature and PET were positively correlated with combined storage-intensity effects on response magnitude metrics (Table 2-6), suggesting that the drier the climate, the more likely it is for joint storage and intensity effects to determine the magnitude of hydrologic response.

2.4.3 Effectively capturing the spatio-temporal variability in event response

Through PCA it was possible to assess which event metrics are most important for capturing variability in hydrologic response. The present study underlined that for every single site considered, multiple metrics were needed to capture dominant patterns of variability in event response. While site-specific temporal response variability was overwhelmingly captured by the first three PCs (Figure 2-7), the number of response metrics needed to do so changed drastically across sites. For the PMRW site, for instance, temporal variability in event response was best captured by timing metrics only, suggesting similar response magnitude metrics across events. For all other sites, however, both magnitude and timing metrics were identified as necessary to capture the maximum amount of variability in event response: those metrics illustrate not only runoff generation but also the distribution of runoff over time, which are key elements of the rainfall-runoff transformation process (Beven, 2011). In the literature, response timing has been described as a reflection of catchment or hillslope wetness conditions (Ali et al., 2010; Carey & Woo, 2001; Dingman, 2015), and a large body of work has demonstrated the significant influence that those conditions exert on event response (Ali & Roy, 2010a; Anderson & Burt, 1978; Biron et al., 1999; Bonell, 1993; Grayson et al., 1997; James & Roulet, 2009). PCA results from the current study are consistent with the published literature, in that they show response timing metrics to be critical in the characterization of response variability. For all sites, multiple response timing metrics were needed to capture most of response variability, thus hinting at different aspects of the rainfall-runoff transformation process being worth consideration. Across sites, T_{LR} and T_{LP} were often highlighted as important for capturing response variability, which is evidence of the role of antecedent moisture conditions and hydrologic abstractions that may

affect the initiation of runoff (T_{LR}) and peak runoff. At some sites, T_c was important for capturing variability in hydrological responses, suggesting that drainage and recession dynamics are also important to consider when evaluating rainfall-runoff dynamics. Very few response magnitude metrics had loadings greater than |0.45| for the first three PCs, but one notable exception was I_{abs}, which implies that the storage deficit that needed to be satisfied before runoff initiation was important for explaining response variability at many sites. Key response magnitude and timing metrics also proved useful for event classification, as they helped identify that event response is predominantly low magnitude and fast timing, or high magnitude and slow timing (Figure 2-8 and Figure 2-9). Worth mentioning is the fact that many of the key response timing and magnitude metrics (i.e., T_{LR}, T_{LP}, I_{abs} and Q_{TOT}) identified in the present study via PCA are not the same metrics that are most commonly used in the literature. Indeed, only select studies have used multiple response metrics to characterize event response (e.g., Ali et al., 2010; Carey & Woo, 2001; Post & Jakeman, 1996) and the vast majority have used individual response magnitude metrics, particularly RR and Q_{MAX}. In the present study, RR and Q_{MAX} were only important for capturing response variability at five and nine sites out of 21, respectively. While RR and Q_{MAX} have proven useful in other efforts, they appear to be less effective than other response magnitude metrics – particularly I_{abs} and Q_{TOT} – for representing response variability from one event to another.

2.5 Conclusion

This study used a data-driven approach to understand spatial and temporal variability in hydrologic response across 21 sites with contrasting drainage area, topography, climate, and

geology to investigate how controls on these responses vary across sites. Key findings from this study include:

(1) Differences in the variability of event response metrics across sites were more pronounced for response timing metrics, compared to response magnitude metrics.

(2) For most sites, combined storage-intensity effects explained a large portion of the variability in both response magnitude and response timing metrics, highlighting the importance of explicitly considering intensity effects in hydrologic studies.

(3) Climatic variables (mean annual temperature and potential evapotranspiration) were strong controls in predicting typical values of events response metrics and their degree of temporal variability. Physiographic characteristics were less frequently correlated to summary statistics of response metrics, suggesting that physiographic characteristics alone do not effectively account for the variability in hydrologic response.

(4) Response timing metrics, especially those related to runoff or hydrograph response initiation, were critical for capturing variability in site-specific hydrologic response. This hints at the importance of data-driven approaches for selecting appropriate response metrics for hydrological process conceptualization and inter-site comparisons, as opposed to relying on the most commonly used metrics in the literature.

(5) The predominant event response type across sites could be qualitatively described as low magnitude and fast timing, or as high magnitude and slow timing, despite significant differences in climate and physiography across sites.

Overall, the present study contributes to the growing knowledge of event-specific hydrologic responses and their controls. It is recommended that similar studies be done across a greater number of geographic areas using even larger datasets comprising sub-daily records over

multiple years: such studies could further the evaluation of magnitude versus timing metrics, confirm the dual importance of storage-driven and intensity-driven meteorological factors, and better quantify landscape-hydrology interactions.

2.6 References

Acreman, M. C., & Sinclair, C. D. (1986). Classification of drainage basins according to their physical characteristics; an application for flood frequency analysis in Scotland. *Journal of Hydrology*, *84*(3), 365–380. https://doi.org/10.1016/0022-1694(86)90134-4

Ali, G., L'Heureux, C., Roy, A., Turmel, M.-C., & Courchesne, F. (2011). Linking spatial patterns of perched groundwater storage and stormflow generation processes in a headwater forested catchment. *Hydrological Processes*, *25*(25), 3843–3857. https://doi.org/10.1002/hyp.8238

Ali, G., & Roy, A. (2010a). A case study on the use of appropriate surrogates for antecedent moisture conditions (AMCs). *Hydrology and Earth System Sciences*, *14*(10), 1843–1861.

Ali, G., & Roy, A. (2010b). Shopping for hydrologically representative connectivity metrics in a humid temperate forested catchment. *Water Resources Research*, *46*(12). http://onlinelibrary.wiley.com/doi/10.1029/2010WR009442/full

Ali, G., Roy, A., Turmel, M.-C., & Courchesne, F. (2010). Multivariate analysis as a tool to infer hydrologic response types and controlling variables in a humid temperate catchment. *Hydrological Processes*, *24*(20), 2912–2923. https://doi.org/10.1002/hyp.7705

Ali, G., Tetzlaff, D., McDonnell, J., Soulsby, C., Carey, S., Laudon, H., McGuire, K., Buttle, J., Seibert, J., & Shanley, J. (2015). Comparison of threshold hydrologic response across

northern catchments. *Hydrological Processes*, *29*(16), 3575–3591. https://doi.org/10.1002/hyp.10527

Ambroise, B. (2004). Variable 'active' versus 'contributing' areas or periods: A necessary distinction. *Hydrological Processes*, *18*(6), 1149–1155. https://doi.org/10.1002/hyp.5536

Anderson, M. G., & Burt, T. P. (1978). The role of topography in controlling throughflow generation. *Earth Surface Processes*, *3*(4), 331–344.

Bates, Bc., & Pilgrim, D. H. (1983). Investigation of storage-discharge relations for river reaches and runoff routing models. TRANS. INST. ENG. AUSTRAL. CIVIL ENG., (3), 153-161.

Betson, R. (1964). What is watershed runoff? *Journal of Geophysical Research*, *69*(8), 1541–1552.

Beven, K. (2001). On fire and rain (or predicting the effects of change). *Hydrological Processes*, *15*(7), 1397–1399.

Beven, K. (2011). *Rainfall-runoff modelling: The primer*. John Wiley & Sons.

Beven, K., Wood, E., & Sivapalan, M. (1988). On hydrological heterogeneity—Catchment morphology and catchment response. *Journal of Hydrology*, *100*(1), 353–375. https://doi.org/10.1016/0022-1694(88)90192-8

Biron, P. M., Roy, A. G., Courschesne, F., Hendershot, W. H., Côté, B., & Fyles, J. (1999). The effects of antecedent moisture conditions on the relationship of hydrology to hydrochemistry in a small forested watershed. *Hydrological Processes*, *13*(11), 1541–1555. https://doi.org/10.1002/(SICI)1099-1085(19990815)13:11<1541::AID-HYP832>3.0.CO;2-J

Bishop, K., Seibert, J., Nyberg, L., & Rodhe, A. (2011). Water storage in a till catchment. II: Implications of transmissivity feedback for flow paths and turnover times. *Hydrological Processes*, *25*(25), 3950–3959. https://doi.org/10.1002/hyp.8355

Blume, T., Zehe, E., & Bronstert, A. (2007). Rainfall—Runoff response, event-based runoff coefficients and hydrograph separation. *Hydrological Sciences Journal*, *52*(5), 843–862. https://doi.org/10.1623/hysj.52.5.843

Bonell, M. (1993). Progress in the understanding of runoff generation dynamics in forests. *Journal of Hydrology*, *150*(2), 217–275. https://doi.org/10.1016/0022-1694(93)90112-M

Borcard, D., Gillet, F., & Legendre, P. (2011). *Numerical ecology with R*. Springer Science & Business Media.

Borcard, D., & Legendre, P. (1994). Environmental control and spatial structure in ecological communities: An example using oribatid mites (Acari, Oribatei). *Environmental and Ecological Statistics*, *1*(1), 37–61. https://doi.org/10.1007/BF00714196

Borcard, D., Legendre, P., & Drapeau, P. (1992). Partialling out the Spatial Component of Ecological Variation. *Ecology*, *73*(3), 1045–1055. https://doi.org/10.2307/1940179

Bosch, J. M., & Hewlett, J. D. (1982). A review of catchment experiments to determine the effect of vegetation changes on water yield and evapotranspiration. *Journal of Hydrology*, *55*(1–4), 3–23.

Buttle, J. (2006). Mapping first-order controls on streamflow from drainage basins: The T3 template. Hydrological Processes: An International Journal, 20(15), 3415–3422. https://doi.org/10.1002/hyp.6519

Cammeraat, L. H. (2002). A review of two strongly contrasting geomorphological systems within the context of scale. *Earth Surface Processes and Landforms*, *27*(11), 1201–1222. https://doi.org/10.1002/esp.421

Carey, S. K., & Woo, M. (2001). Slope runoff processes and flow generation in a subarctic, subalpine catchment. Journal of Hydrology, 253(1–4), 110–129. https://doi.org/10.1016/S0022-1694(01)00478-4

Detty, J. M., & McGuire, K. J. (2010). Threshold changes in storm runoff generation at a till-mantled headwater catchment. *Water Resources Research*, *46*(7), W07525. https://doi.org/10.1029/2009WR008102

Devito, K., Creed, I., Gan, T., Mendoza, C., Petrone, R., Silins, U., & Smerdon, B. (2005). A framework for broad-scale classification of hydrologic response units on the Boreal Plain: Is topography the last thing to consider? *Hydrological Processes, 19*(8), 1705–1714. https://doi.org/10.1002/hyp.5881

Dingman, S. L. (2015). *Physical hydrology*. Waveland press.

Dixon, P., & Palmer, M. W. (2003). VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science*, *14*(6), 927–930. https://doi.org/10.1658/1100-9233(2003)014[0927:VAPORF]2.0.CO;2

Dunne, T. (1978). Field studies of hillslope flow processes. *Hillslope Hydrology*, 227–293.

Dunne, T., & Black, R. D. (1970). Partial area contributions to storm runoff in a small New England watershed. *Water Resources Research*, *6*(5), 1296–1311.

Fedora, M. A., & Beschta, R. L. (1989). Storm runoff simulation using an antecedent precipitation index (API) model. *Journal of Hydrology*, *112*(1–2), 121–133.

Gilbert, B., & Bennett, J. R. (2010). Partitioning variation in ecological communities: Do the numbers add up? *Journal of Applied Ecology*, *47*(5), 1071–1082. https://doi.org/10.1111/j.1365-2664.2010.01861.x

Graham, C. B., & McDonnell, J. J. (2010). Hillslope threshold response to rainfall: (2) Development and use of a macroscale model. *Journal of Hydrology*, *393*(1–2), 77–93. https://doi.org/10.1016/j.jhydrol.2010.03.008

Graham, C. B., Woods, R. A., & McDonnell, J. J. (2010). Hillslope threshold response to rainfall: (1) A field based forensic approach. *Journal of Hydrology*, *393*(1–2), 65–76. https://doi.org/10.1016/j.jhydrol.2009.12.015

Granger, R. J., Gray, D. M., & Dyck, G. E. (1984). Snowmelt infiltration to frozen prairie soils. *Canadian Journal of Earth Sciences*, 21(6), 669–677. https://doi.org/10.1139/e84-073

Grayson, R. B., Western, A. W., Chiew, F. H., & Blöschl, G. (1997). Preferred states in spatial soil moisture patterns: Local and nonlocal controls. *Water Resources Research*, 33(12), 2897–2908. https://doi.org/10.1029/97WR02174

Gregory, K. J., & Walling, D. E. (1973). *Drainage basin form and process; a geomorphological approach*. Edward Arnold.

Hannah, D. M., Smith, B. P., Gurnell, A. M., & McGregor, G. R. (2000). An approach to hydrograph classification. Hydrological Processes, 14(2), 317–338. https://doi.org/10.1002/(SICI)1099-1085(20000215)14:2<317::AID-HYP929>3.0.CO;2-T

Hargreaves, G. H., & Samani, Z. A. (1982). Estimating potential evapotranspiration. *Journal of the Irrigation and Drainage Division*, *108*(3), 225–230.

Hewlett, J. D., & Hibbert, A. R. (1967). Factors affecting the response of small watersheds to precipitation in humid areas. *Forest Hydrology*, 275–290.

Holtan, H. N., & Overton, D. E. (1963). Analyses and application of simple hydrographs. *Journal of Hydrology*, *1*(3), 250–264.

Holtan, H. N., & Overton, D. E. (1965). Storage-flow hysteresis in hydrograph synthesis. *Journal of Hydrology*, *2*(4), 309–323.

Horton, R. E. (1933). The Rôle of infiltration in the hydrologic cycle. *Transactions, American Geophysical Union*, *14*(1), 446. https://doi.org/10.1029/TR014i001p00446

James, A., & Roulet, N. (2007). Investigating hydrologic connectivity and its association with threshold change in runoff response in a temperate forested watershed. *Hydrological Processes*, *21*(25), 3391–3408. https://doi.org/10.1002/hyp.6554

James, A., & Roulet, N. (2009). Antecedent moisture conditions and catchment morphology as controls on spatial patterns of runoff generation in small forest catchments. *Journal of Hydrology*, *377*(3–4), 351–366. https://doi.org/10.1016/j.jhydrol.2009.08.039

Jones, J. (2006). Intersite comparisons of rainfall-runoff processes. In *Encyclopedia of Hydrological Sciences*. John Wiley & Sons, Ltd.

Jones, J. A. (2000). Hydrologic processes and peak discharge response to forest removal, regrowth, and roads in 10 small experimental basins, western Cascades, Oregon. Water Resources Research, 36(9), 2621–2642. https://doi.org/10.1029/2000WR900105

Jones, J. A., & Swanson, F. J. (2001). Hydrologic inferences from comparisons among small basin experiments. *Hydrological Processes*, *15*(12), 2363–2366. https://doi.org/10.1002/hyp.474

Laudon, H., Sjöblom, V., Buffam, I., Seibert, J., & Mörth, M. (2007). The role of catchment scale and landscape characteristics for runoff generation of boreal streams. *Journal of Hydrology*, *344*(3), 198–209. https://doi.org/10.1016/j.jhydrol.2007.07.010

Legendre, P., & Legendre, L. F. (2012). *Numerical ecology* (3rd English ed.). Elsevier.

Lehmann, P., Hinz, C., McGrath, G., Tromp-van Meerveld, H. J., & McDonnell, J. J. (2007). Rainfall threshold for hillslope outflow: An emergent property of flow pathway connectivity. *Hydrol. Earth Syst. Sci.*, *11*(2), 1047–1063. https://doi.org/10.5194/hess-11-1047-2007

McDonnell, J. (2013). Are all runoff processes the same? *Hydrological Processes*, *27*(26), 4103–4111. https://doi.org/10.1002/hyp.10076

McDonnell, J., Sivapalan, M., Vaché, K., Dunn, S., Grant, G., Haggerty, R., Hinz, C., Hooper, R., Kirchner, J., Roderick, M. L., Selker, J., & Weiler, M. (2007). Moving beyond heterogeneity and process complexity: A new vision for watershed hydrology. *Water Resources Research*, *43*(7), W07301. https://doi.org/10.1029/2006WR005467

McGlynn, B. L., & McDonnell, J. (2003). Quantifying the relative contributions of riparian and hillslope zones to catchment runoff. *Water Resources Research*, *39*(11), 1310. https://doi.org/10.1029/2003WR002091

McKee, A., & Druliner, P. (1998). *HJ Andrews Experimental Forest.* http://andrewsforest.oregonstate.edu/pubs/pdf/pub2415.pdf

McMillan, H., Gueguen, M., Grimon, E., Woods, R., Clark, M., & Rupp, D. E. (2014). Spatial variability of hydrological processes and model structure diagnostics in a 50 km2 catchment. *Hydrological Processes*, *28*(18), 4896–4913. https://doi.org/10.1002/hyp.9988

McNamara, J. P., Tetzlaff, D., Bishop, K., Soulsby, C., Seyfried, M., Peters, N. E., Aulenbach, B. T., & Hooper, R. (2011). Storage as a Metric of Catchment Comparison. *Hydrological Processes*, *25*(21), 3364–3371. https://doi.org/10.1002/hyp.8113

Merz, R., & Blöschl, G. (2009). A regional analysis of event runoff coefficients with respect to climate and catchment characteristics in Austria. *Water Resources Research*, *45*(1). https://doi.org/10.1029/2008WR007163

Mosley, M. P. (1979). Streamflow generation in a forested watershed, New Zealand. *Water Resources Research*, 15(4), 795–806. https://doi.org/10.1029/WR015i004p00795

Mulvaney, T. J. (1851). On the use of self-registering rain and flood gauges in making observations of the relations of rainfall and flood discharges in a given catchment. *Proceedings of the Institution of Civil Engineers of Ireland*, *4*, 19–31.

Oksanen, J., Kindt, R., Legendre, P., O'Hara, B., Stevens, M. H. H., Oksanen, M. J., & Suggests, M. (2007). The vegan package. *Community Ecology Package*, *10*, 631–637.

Oswald, C., Richardson, M. C., & Branfireun, B. A. (2011). Water storage dynamics and runoff response of a boreal Shield headwater catchment. *Hydrological Processes*, *25*(19), 3042–3060. https://doi.org/10.1002/hyp.8036

Peres-Neto, P. R., Legendre, P., Dray, S., & Borcard, D. (2006). Variation Partitioning of Species Data Matrices: Estimation and Comparison of Fractions. *Ecology*, *87*(10), 2614–2625. https://doi.org/10.1890/0012-9658(2006)87[2614:VPOSDM]2.0.CO;2

Post, D. A., & Jakeman, A. J. (1996). Relationships between catchment attributes and hydrological response characteristics in small Australian mountain ash catchments. Hydrological Processes, 10(6), 877–892. https://doi.org/10.1002/(SICI)1099-1085(199606)10:6<877::AID-HYP377>3.0.CO;2-T

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Reaney, S. M., Bracken, L. J., & Kirkby, M. J. (2007). Use of the Connectivity of Runoff Model (CRUM) to investigate the influence of storm characteristics on runoff generation and connectivity in semi-arid areas. *Hydrological Processes*, *21*(7), 894–906. https://doi.org/10.1002/hyp.6281

Redding, T. E., & Devito, K. J. (2008). Lateral flow thresholds for aspen forested hillslopes on the Western Boreal Plain, Alberta, Canada. *Hydrological Processes*, *22*(21), 4287–4300. https://doi.org/10.1002/hyp.7038

Ross, C., Ali, G., Bansah, S., & Laing, J. R. (2017). Evaluating the Relative Importance of Shallow Subsurface Flow in a Prairie Landscape. *Vadose Zone Journal*, *16*(5). https://doi.org/10.2136/vzj2016.10.0096

Seibert, J., Bishop, K., Nyberg, L., & Rodhe, A. (2011). Water storage in a till catchment. I: Distributed modelling and relationship to runoff. *Hydrological Processes*, *25*(25), 3937–3949. https://doi.org/10.1002/hyp.8309

Shaw, D. A., Vanderkamp, G., Conly, F. M., Pietroniro, A., & Martz, L. (2012). The Fill–Spill Hydrology of Prairie Wetland Complexes during Drought and Deluge. *Hydrological Processes*, *26*(20), 3147–3156. https://doi.org/10.1002/hyp.8390

Sidle, R. C., Tsuboyama, Y., Noguchi, S., Hosoda, I., Fujieda, M., & Shimizu, T. (2000). Stormflow generation in steep forested headwaters: A linked hydrogeomorphic paradigm. *Hydrological Processes*, *14*(3), 369–385. https://doi.org/10.1002/(SICI)1099-1085(20000228)14:3<369::AID-HYP943>3.0.CO;2-P

Sokal, R. R., & Rohlf, F. J. (1995). Biometry (3rd edn). *WH Freman and Company: New York.*

Spence, C. (2007). On the relation between dynamic storage and runoff: A discussion on thresholds, efficiency, and function. *Water Resources Research*, *43*(12), W12416. https://doi.org/10.1029/2006WR005645

Tallaksen, L. M. (1995). A review of baseflow recession analysis. *Journal of Hydrology*, *165*(1), 349–370. https://doi.org/10.1016/0022-1694(94)02540-R

Tang, W., & Carey, S. K. (2017). HydRun: A MATLAB toolbox for rainfall–runoff analysis. *Hydrological Processes*, *31*(15), 2670–2682. https://doi.org/10.1002/hyp.11185

Tani, M. (1997). Runoff generation processes estimated from hydrological observations on a steep forested hillslope with a thin soil layer. *Journal of Hydrology*, *200*(1), 84–109. https://doi.org/10.1016/S0022-1694(97)00018-8

Te Chow, V. (1959). *Open-channel hydraulics* (Vol. 1). McGraw-Hill.

Te Chow, V., Maidment, D. R., & Mays, L. W. (1988). *Applied hydrology*. McGraaw Hill.

Thornthwaite, C. W. (1948). An approach toward a rational classification of climate. *Geographical Review*, 38(1), 55–94. https://doi.org/10.2307/210739

Tromp-van Meerveld, H. J., J, H., James, A. L., McDonnell, J. J., & Peters, N. E. (2008). A reference data set of hillslope rainfall-runoff response, Panola Mountain Research Watershed, United States. *Water Resources Research*, *44*(6). https://doi.org/10.1029/2007WR006299

Tromp-van Meerveld, H. J., & McDonnell, J. J. (2006a). Threshold relations in subsurface stormflow: 1. A 147-storm analysis of the Panola hillslope. *Water Resources Research*, *42*(2). https://doi.org/10.1029/2004WR003778

Tromp-van Meerveld, H. J., & McDonnell, J. J. (2006b). Threshold relations in subsurface stormflow: 2. The fill and spill hypothesis. *Water Resources Research*, *42*(2). https://doi.org/10.1029/2004WR003800

Uchida, T., McDonnell, J. J., & Asano, Y. (2006). Functional intercomparison of hillslopes and small catchments by examining water source, flowpath and mean residence time. Journal of Hydrology, 327(3–4), 627–642. https://doi.org/10.1016/j.jhydrol.2006.02.037

Western, A. W., & Grayson, R. B. (1998). The Tarrawarra Data Set: Soil moisture patterns, soil characteristics, and hydrological flux measurements. *Water Resources Research*, *34*(10), 2765–2768. https://doi.org/10.1029/98WR01833

Whipkey, R. Z. (1965). Subsurface Stormflow from Forested Slopes. *International Association of Scientific Hydrology. Bulletin*, *10*(2), 74–85. https://doi.org/10.1080/02626666509493392

Woods, R., Grayson, R., Western, A., Duncan, M., Wilson, D., Young, R., Ibbitt, R., Henderson, R., & McMahon, T. (2013). Experimental Design and Initial Results from the Mahurangi River Variability Experiment: MARVEX. In *Land Surface Hydrology, Meteorology, and Climate: Observations and Modeling* (pp. 201–213). American Geophysical Union. http://onlinelibrary-wiley-com.uml.idm.oclc.org/doi/10.1029/WS003p0201/summary

Yadav, M., Wagener, T., & Gupta, H. (2007). Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins. *Advances in Water Resources*, *30*(8), 1756–1774. https://doi.org/10.1016/j.advwatres.2007.01.005

CHAPTER 3. EVALUATING THE UBIQUITY OF THRESHOLDS IN RAINFALL-RUNOFF RESPONSE ACROSS CONTRASTING ENVIRONMENTS

3.1 Introduction

Studies examining hydrologic response to climatic inputs at hillslope and small catchment scales have shown highly nonlinear runoff behaviour (Beven et al., 1988; McDonnell et al., 2007; Sivapalan et al., 2002; Sivapalan, 2006). While these studies have greatly advanced our understanding of runoff generation processes, the "uniqueness of place" (Beven, 2000) inherent to isolated studies has resulted in limited transferability of some findings, making generalization across sites difficult (McDonnell et al., 2007; Scaife & Band, 2017; Sivapalan, 2006). The difficulty in generalizing some process conceptualizations has motivated a shift in focus towards emergent properties, i.e., properties that cannot be predicted from individual landscape components but reflect landscape heterogeneity and process complexity (Lehmann et al., 2007; McDonnell et al., 2007). Thresholds in runoff response are one of these properties and are generally defined as critical moments in time or points in space at which runoff behaviour rapidly changes (Ali et al., 2013; Phillips, 2006). For critical moments in time, thresholds are typically defined as values of one or multiple meteorological factors that trigger a nonlinear change in hydrologic response characteristics. Thresholds are assessed through the evaluation of scatter plots that compare hydrologic response metrics (y-axis) to meteorological factors (x-axis). To date, threshold-related research has mostly taken place on hillslopes and catchments in temperate or humid environments (e.g., Ali et al., 2011; Detty & McGuire, 2010; Graham et al., 2010; James & Roulet, 2007; Lehmann et al., 2007; Mosley, 1979; Sidle et al., 2000; Tani, 1997; Tromp-van Meerveld & McDonnell, 2006; Whipkey, 1965). Fewer studies focused on semi-arid, snow- and permafrost-dominated, or prairie pothole landscapes (e.g., Cammeraat, 2002; Kim et al., 2004; Mielko & Woo, 2006; Reaney et al., 2007; Shaw et al., 2012; Stichling & Blackwell,

1957). Very few inter-site comparisons have been done (e.g., Ali et al., 2015), making the ubiquity (or lack thereof) of thresholds difficult to evaluate (Weiler et al., 2006). Relying on existing literature alone is insufficient, given that studies that have observed threshold behaviour have used different pairs of meteorological factors and response metrics, with different degrees of consideration of antecedent moisture conditions.

One major source of variability among threshold studies is the use of different hydrologic response metrics. Indeed, the response metrics which are typically represented on scatter plots are total discharge, peak discharge, and the runoff ratio, which all refer to the magnitude of the hydrological response. Hyetograph-hydrograph analysis performed for individual hydrologic events, however, yields a much broader list of response metrics that could be considered when evaluating the presence of threshold behaviour. With respect to response magnitude, for instance, the initial abstraction – defined as the storage deficit that needs to be satisfied before the initial event hydrograph rise – could also be used. Furthermore, response timing metrics – such as event duration, response lag, time of rise, centroid lag, lag-to-peak, centroid lag-to-peak, and time of concentration (Dingman, 2015) – might also be subject to threshold effects, as they reflect different measures of water travel time throughout an event. Response timing metrics have not previously been evaluated in the context of hydrologic thresholds.

Another major source of variability among threshold studies is the use of different meteorological factors. Numerous studies have shown that a storage threshold must be exceeded to trigger hillslope contributions to runoff (Detty & McGuire, 2010; Kim et al., 2004; Oswald et al., 2011; Spence & Woo, 2003; Tromp-van Meerveld & McDonnell, 2006a, 2006b). The initiation of runoff generation contingent on rainfall depth thresholds has been of particular interest and reported values of such thresholds range between 20 and 150 mm (Mosley, 1979;

Redding & Devito, 2008; Sidle et al., 2000; Tani, 1997; Tromp-van Meerveld & McDonnell, 2006a; Weiler et al., 2006; Whipkey, 1965). Storage thresholds based on variables other than rainfall depth have also been documented. For example, thresholds of soil moisture and the combination of soil moisture and total event rainfall have been linked to runoff response (Detty & McGuire, 2010; James & Roulet, 2007). Similarly, threshold behaviour has been reported between runoff response and water table levels (Ali et al., 2011; Detty & McGuire, 2010; Kim et al., 2004). Others have identified storage thresholds related to water accumulation in surface depressions or subsurface depressions located at interfaces between subsurface materials with distinct texture, density, or degree of cementation: once filled, these depressions can transmit water downslope, allowing for connectivity of adjacent depressions and/or coupling between hillslopes and stream networks (Mielko & Woo, 2006; Oswald et al., 2011; Spence & Woo, 2003; Tromp-van Meerveld & McDonnell, 2006a, 2006b). Changes in dynamic storage caused by evaporation have also been shown to affect threshold behaviour – in one case, the time before a storage threshold was met was prolonged by evaporation (Mielko & Woo, 2006). Thresholds related to rainfall intensity, however, have received much less attention in the literature than storage thresholds (Ali et al., 2015): these thresholds are rates of water addition to a site (i.e., rainfall intensity) that, once exceeded, result in a notable change in response. This is distinct from the storage thresholds that involve specific volumes or depths of water (e.g., total event rainfall) to be exceeded. Rainfall intensity thresholds have been used to characterize rainfall-runoff responses (Ali et al., 2015) and to discriminate between runoff generation processes (Scaife & Band, 2017), particularly for environments where infiltration-excess overland flow is a dominant runoff generation mechanism (Cammeraat, 2002; Reaney et al., 2007). However, studies in humid, temperate forested catchments with high soil infiltration capacity and forest

interception capacity have shown that rainfall intensity exerts little influence on runoff processes (Graham et al., 2010; Tromp-van Meerveld & McDonnell, 2006a). While the importance of rainfall intensity thresholds where infiltration is limited has been clearly demonstrated (e.g., Cammeraat, 2002; Laudon et al., 2007; Mielko & Woo, 2006; Reaney et al., 2007; Shanley & Chalmers, 1999), the extent to which rainfall intensity thresholds govern or influence nonlinear runoff dynamics in other environments remains unclear.

Given all the possible response metrics and meteorological factors that can be estimated for individual hydrological events, numerous pairs can theoretically be used to evaluate whether thresholds are present. Different pairs of meteorological factors and response metrics may lead to very different conclusions or hypotheses about the presence or absence of threshold behaviour. The choice could be made between meteorological factors that reflect precipitation only (e.g., total rainfall, average and maximum rainfall intensity, and antecedent rainfall) or factors that consider hydrologic abstractions (e.g., antecedent potential evapotranspiration and effective rainfall, estimated as the difference between gross rainfall and evapotranspiration). Little to no guidance exists regarding which meteorological factors and response metrics to use in what situation. Another factor that may have a large influence on the evaluation of threshold behaviour is antecedent conditions. While antecedent moisture conditions have long been confirmed as a first-order control on storage thresholds (e.g., Ali et al., 2011; Detty & McGuire, 2010; James & Roulet, 2007; Kim et al., 2004), it has also been demonstrated that the effectiveness of proxies of antecedent moisture conditions depends on the duration over which they are computed (Ali & Roy, 2010; Ross et al., 2017). Whether or not thresholds are sensitive to antecedent conditions over specific durations of time preceding a rainfall-runoff event has not yet been examined.

In light of the aforementioned research gaps, the goal of the current study was to examine the prevalence of threshold behaviour for a large number of sites. The focus of this study was less on estimating specific threshold values (i.e., the values of meteorological factors leading to changes in hydrologic response) and their controls, but rather on widely testing for the presence of threshold behaviour. This focus on the presence or absence of thresholds across a wide range of relationships provides an opportunity to determine the ubiquity (or lack thereof) of hydrologic thresholds, regardless of the processes that led to their emergence or the precise values of the meteorological factors that control them. To ensure that a wide variety of landscape types and hydrologic events were included in this evaluation, we tested for the presence of thresholds using existing datasets from one experimental hillslope and twenty catchments located in seven geographic areas with contrasting climate, topography, geology, soil properties, and land cover. While focusing on these sites, three specific research questions were addressed:

(1) To what extent are total event rainfall thresholds associated with nonlinear changes in the response magnitude metrics that are most commonly used in the literature (e.g., runoff ratio and peak discharge)?

(2) Are thresholds also associated with relationships that involve other meteorological factors related to rainfall and evapotranspiration, and/or response timing metrics?

(3) Is threshold behaviour dependent on antecedent conditions over short-, medium- and longer-term durations?

3.2 Methods

3.2.1 Study sites

Existing data for twenty-one sites distributed across seven study areas were used in this study (Figure 3-1). Three of the sites are located in Canada: the Catfish Creek Watershed (CCW), the IISD Experimental Lakes Area Lake 658 UP1 catchment (UP1), and the Hermine catchment (HRM). Nine sites are in the United States: the Panola Mountain Research Watershed experimental hillslope (PMRW) and eight catchments of the HJ Andrews Experimental Forest (HJA1 – HJA8). One site is in Australia – the Tarrawarra catchment (TRC) – and eight sites are nested catchments of the larger Mahurangi River catchment (MRC1 – MRC8) in New Zealand. Physiographic features of individual sites and long-term regional climate variables are summarized in Table 3-1. Each study area has been described in detail as part of previous hydrologic research (Ali et al., 2011; McKee & Druliner, 1998; Oswald et al., 2011; Ross et al., 2017; Ross et al., 2019; Tromp-van Meerveld et al., 2008; Western & Grayson, 1998; Woods et al., 2013).

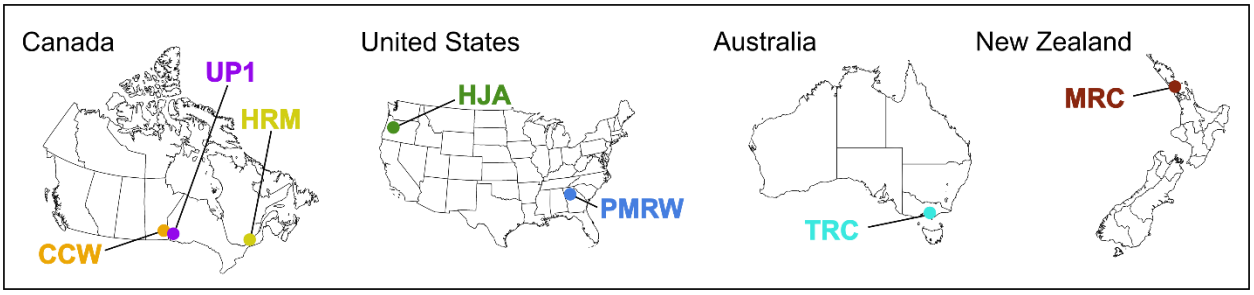


Figure 3-1. Location of the twenty-one sites spanning seven study areas selected for this study. Eight nested catchments were selected for the HJA and MRC. For full, non-abbreviated names, refer to Section 3.2.1.

Table 3-1. Site-specific drainage area (DA), relief and regional mean annual values of temperature (T), potential evapotranspiration (PET), precipitation (P), and proportion of P that falls as rain (P_R).

	DA (km^2)	Relief (m)	T (°C)	PET (mm)	P (mm) / P_R
PMRW	1.0×10^{-3}	12.5	17.4	835	1240 / 1.0
HRM	0.05	27.0	6.9	508	1150 / 0.7
UP1	0.08	63.1	3.2	451	708 / 0.75
TRC	0.11	34.3	15.6	724	820 / 1.0
CCW	145.32	77.0	2.2	444	530 / 0.8
MRC1	0.51	104.8	15.6	716	1600 / 1.0
MRC2	0.71	116.4	15.6	716	1600 / 1.0
MRC3	2.30	113.6	15.6	716	1600 / 1.0
MRC4	2.63	210.8	15.6	716	1600 / 1.0
MRC5	2.65	155.1	15.6	716	1600 / 1.0
MRC6	2.96	73.1	15.6	716	1600 / 1.0
MRC7	4.61	30.1	15.6	716	1600 / 1.0
MRC8	24.80	286.6	15.6	716	1600 / 1.0
HJA1	0.13	129.3	7.5	489	2300 / 0.80
HJA2	0.15	155.1	7.1	475	2300 / 0.80
HJA3	0.21	126.6	7.8	492	2200 / 0.80
HJA4	0.60	432.0	8.7	525	2500 / 0.75
HJA5	0.96	408.3	8.9	528	2600 / 0.75
HJA6	1.01	276.6	8.7	523	2400 / 0.75
HJA7	14.36	860.7	8.8	527	2600 / 0.60
HJA8	62.42	1206.0	7.4	485	2400 / 0.75

3.2.2 Rainfall-runoff event delineation, meteorological factors, and response metrics

The datasets associated with the study sites had observation frequencies and record lengths ranging from 1 minute to 1 hour, and from 6 months to 5 years, respectively. Each site, including nested catchments of the HJA and MRC, had its own, independent hydrological (discharge) and climatic (rainfall, air temperature) datasets, with the slight exception that the

MRC shared temperature data across catchments. To maintain consistency across sites, all data were aggregated to a 1-hour frequency: rainfall data were aggregated using the sum of observations over hourly periods, while temperature and discharge data were aggregated using the mean of observations. Rainfall-runoff events were identified from site-specific rainfall and discharge data using the MATLAB toolbox HydRun (Tang & Carey, 2017). The HydRun toolbox matches rainfall events to runoff events. First, the toolbox separates base flow from streamflow using a recursive digital filter, then extracts runoff events based on the shape of the hydrograph. Runoff event start and endpoints are determined using a local-minimum approach. Rainfall events are defined as adjacent rainfall observations separated by a rainless period. A total of 1,641 rainfall-runoff events were delineated and retained for further analysis, with the number of events per site ranging from 13 to 134. A comprehensive set of event-based meteorological factors and response metrics were derived. Meteorological factors that quantify rainfall depths include the total event rainfall (R_{TOT}), as well as amounts of rainfall over a given temporal window preceding a rainfall event (antecedent rainfall). A range of antecedent window durations, x, were considered, where x is equal to 1, 3, 5, 7, 10, 14 or 30 day(s). Cumulative antecedent rainfall (AR_X) and the sum of event and antecedent rainfall ($R_{TOT} + AR_X$) were computed over each antecedent window. Average and maximum event rainfall intensity values (RI_{AVG} and RI_{MAX}) were calculated to quantify the rate of water addition to a site during an event. To account for some hydrologic abstractions and their effect on hydrologic response, antecedent evapotranspiration ($APET_X$) and effective rainfall (estimated by subtracting antecedent potential evapotranspiration from antecedent rainfall: AR_X-$APET_X$) were also estimated for each antecedent window. To characterize event hydrologic responses, response magnitude and response timing metrics were computed. Response magnitude metrics (magnitude

metrics) included the initial abstraction (I_{abs}), peak event discharge (Q_{MAX}), total event runoff (Q_{TOT}), and the runoff ratio (RR). Response timing metrics (timing metrics) included the response lag (T_{LR}), the lag-to-peak (T_{LP}), and the time of concentration (T_c) (Dingman, 2015). Table 3-2 shows response metrics that were used in this study, along with their abbreviations and definitions. Additional methodological details pertaining to rainfall-runoff event delineation, including examples of typical event rainfall-runoff responses for each site, have been previously reported, along with a detailed account of response metric and meteorological factor computations (refer to Ross et al., 2019).

Table 3-2. Names, abbreviations, and definitions of response metrics used in this study.

Response metric	Definition
Response magnitude metrics	
Runoff ratio (RR)	The fraction of event rainfall that becomes runoff
Peak discharge (Q_{MAX})	Area-normalized maximum event discharge
Total runoff (Q_{TOT})	Area-normalized total event runoff
Initial abstraction (I_{abs})	Storage deficit satisfied before hydrograph response, estimated as the amount of event rainfall occurring before the initial hydrograph rise
Response timing metrics	
Response lag (T_{LR})	Time elapsed between the beginning of rainfall and the initial hydrograph rise
Lag-to-peak (T_{LP})	Time elapsed between the beginning of rainfall and peak discharge
Time of concentration (T_c)	Time elapsed between the end of rainfall and the end of hydrograph response

3.2.3 Testing for the presence of hydrologic thresholds

To date, studies have assessed scatter plots for the presence of thresholds either visually (e.g., Ali et al., 2011; Detty & McGuire, 2010; Graham et al., 2010; James & Roulet, 2007; Lehmann et al., 2007; McGlynn & McDonnell, 2003; Mosley, 1979; Sidle et al., 2000; Tani,

1997; Tromp-van Meerveld & McDonnell, 2006a; Whipkey, 1965), or by using piecewise linear regression analysis (PRA) (e.g., Oswald et al., 2011; Scaife & Band, 2017). In the current study, we tested for the presence of thresholds using PRA via the segmented package in R (Muggeo, 2008). This approach was selected as it can be automated, and it is less subject to user bias than visual methods. When possible, PRA identifies a breakpoint in a relationship. Scatter plots for which no breakpoint could be identified were not analyzed further, as a threshold was deemed not to be present. When a breakpoint was identified, however, the goodness-of-fit of the piecewise linear model (R^2) was estimated, along with the Akaike Information Criterion (AIC) of the piecewise linear model (Sakamoto et al., 1986; Yafune et al., 2005), the breakpoint value, the standard error of the estimated breakpoint, and the slope for each linear segment (m_1 and m_2 below and above the breakpoint, respectively). Four criteria needed to be met before an identified breakpoint was considered a threshold, namely:

- The goodness-of-fit of the piecewise linear regression model must be moderate to strong ($R^2 > 0.45$).
- The AIC of the piecewise linear regression model must be more than two units below the AIC of the simple linear regression model.
- The percent difference between the slope of the first segment (m_1) and the slope of the second segment (m_2) must be greater than 10%.
- The slope of the second segment (m_2) must be greater than 0.

The AIC criterion was used to compare competing models (i.e., the linear model versus the piecewise linear model) and to identify the most parsimonious model with respect to model fit and model complexity (Sakamoto et al., 1986). A piecewise linear model with an AIC two or more units below the AIC of the linear model indicates the data provide more support for the

piecewise linear model than the simpler linear model (Sakamoto et al., 1986; Yafune et al., 2005). These criteria are consistent with the literature, which defines a threshold as a critical moment in time at which runoff behaviour rapidly changes (Ali et al., 2013; Detty & McGuire, 2010; Phillips, 2006). It should be noted that while multiple diagnostic shapes of nonlinear rainfall-runoff relationships have been described in the literature (Ali et al., 2013), most studies identify thresholds for "hockey-stick" relationships (Ali et al., 2015; Detty & McGuire, 2010; Graham & McDonnell, 2010; Tani, 1997; Weiler, 2005). Previous studies that have identified thresholds from "hockey stick" relationships have typically forced m_1 to equal zero (Scaife & Band, 2017). In the current study, however, a m_1 value of zero was not assumed since doing so may exclude possible slower stormflow generation (Scaife & Band, 2017).

To address research question 1, 84 site-specific scatter plots showing magnitude metrics (I_{abs}, Q_{MAX}, $Q_{TOT,}$ and RR) against R_{TOT} were constructed and evaluated, using the aforementioned procedure, to determine if thresholds were present. Similarly, to evaluate the presence of thresholds in relationships based on a wider set of response metrics and meteorological factors (research question 2), the same procedure was applied to an additional 4,473 scatter plots (i.e., 213 plots per site) showing magnitude metrics (I_{abs}, Q_{MAX}, $Q_{TOT,}$ and RR) and timing metrics (T_{LR}, T_{LC}, and T_c) against meteorological factors representing different rainfall depths (R_{TOT}, $AR_{X,}$ and AR_X+R_{TOT}), rainfall intensity (RI_{AVG} and RI_{MAX}) and hydrologic abstractions related to evapotranspiration (AR_X-$APET_X$ and $1/APET_X$ (hereafter referred to as $APET_X$)). For each relationship evaluated for a given site, the absence of a threshold could mean one of two things: either that a threshold does not exist and therefore does not affect hydrological behaviour, or that a threshold exists but does not affect all aspects of hydrological behaviour and

therefore only arises when specific meteorological factors and/or response metrics are considered.

Across research questions 1 and 2, 217 different pairs of meteorological factors and response metrics were analyzed for each site. These pairs can be divided into six distinct groups, according to the specific pair examined, namely: (i) response magnitude as a function of rainfall depth (60 pairs) ; (ii) response magnitude as a function of rainfall intensity (8 pairs); (iii) response magnitude as a function of hydrologic abstractions related to evapotranspiration (56 pairs); (iv) response timing as a function of rainfall depth (45 pairs); (v) response timing as a function of rainfall intensity (6 pairs); and (vi) response timing as a function of hydrologic abstractions related to evapotranspiration (42 pairs). The extent to which thresholds were observed for these six groups was quantified using the threshold detection frequency (F):

$$F = \frac{\text{\# of pairs for which a threshold was observed}}{\text{total \# of pairs in the group}} \quad . \qquad \text{Equation 3-1}$$

To examine whether the presence of threshold behaviour varied depending on antecedent moisture conditions (research question 3), the number of AR_X, AR_X+R_{TOT}, $APET_X$, and AR_X-APET_X thresholds that were observed for magnitude and timing metrics were compared across antecedent window durations of 1, 3, 5, 7, 10, 14 and 30 day(s). The number of thresholds observed for each antecedent window duration was also evaluated for individual sites, to examine inter-site variability in the role of antecedent window duration on threshold dynamics.

3.3 Results

3.3.1 Total event rainfall thresholds for response magnitude

Scatter plots showing magnitude metrics against R_{TOT} suggested the presence of R_{TOT} thresholds at 14 (out of 21) sites (Table 3-3). Three out of four potential thresholds were observed at the PMRW site, while two out of four potential thresholds were observed at three sites, and a single threshold was observed at ten sites. The prevalence of R_{TOT} thresholds varied among magnitude metrics: I_{abs}, Q_{MAX}, Q_{TOT}, and RR were controlled by a R_{TOT} threshold at six, five, seven, and one site(s), respectively. These R_{TOT} thresholds were variable, ranging from 37 mm (MRC3) to 335 mm (HJA7). In contrast, R_{TOT} threshold values at the PMRW site were similar regardless of the magnitude metric considered (49 – 51 mm). The goodness-of-fit (R^2) of the piecewise linear model varied across sites and magnitude metrics (Table 3-3). For instance, three sites (PMRW, MRC5, and MRC6) had R^2 values larger than 0.90. For magnitude metrics other than RR, piecewise linear models involving Q_{TOT} had the highest median R^2 value ($R^2 = 0.83$), followed by I_{abs} ($R^2 = 0.72$) and Q_{MAX} ($R^2 = 0.67$). Across sites with R_{TOT} thresholds for response magnitude metrics, the AIC of piecewise linear model ranged from 3.75 (R_{TOT} threshold for I_{abs} at TRC) to 210.19 (R_{TOT} threshold for I_{abs} at MRC5) units below the AIC of the comparable simple linear model. The standard errors (SE) associated with the estimated R_{TOT} threshold values ranged from 0 to 20 mm depending on the site and response metric considered. For example, the 277 mm R_{TOT} threshold for I_{abs} observed at the MRC6 site had an SE of 0 mm, while the 161 mm R_{TOT} threshold for Q_{MAX} observed at the HJA4 site had an SE of 20 mm.

Table 3-3. Site-specific results from PRA for relationships involving response magnitude metrics and R_{TOT}. For sites where a threshold was observed, the R^2 column indicates the goodness-of-fit measure for the piecewise linear model, the threshold column shows the R_{TOT} threshold value, and the SE column indicates the standard error associated with the threshold value.

	I_{abs}			Q_{MAX}			Q_{TOT}			RR		
Site	R^2	Threshold (mm)	SE (mm)	R^2	Threshold (mm)	SE (mm)	R^2	Threshold (mm)	SE (mm)	R^2	Threshold (mm)	SE (mm)
PMRW	-	-	-	0.95	51	2	0.85	49	3	0.80	49	3
HRM	-	-	-	-	-	-	-	-	-	-	-	-
UP1	0.62	65	5	-	-	-	-	-	-	-	-	-
TRC	0.75	66	4	-	-	-	-	-	-	-	-	-
CCW	-	-	-	-	-	-	0.83	37	2	-	-	-
MRC1	-	-	-	0.86	73	3	-	-	-	-	-	-
MRC2	-	-	-	-	-	-	-	-	-	-	-	-
MRC3	-	-	-	-	-	-	-	-	-	-	-	-
MRC4	-	-	-	-	-	-	-	-	-	-	-	-
MRC5	0.93	141	0	-	-	-	-	-	-	-	-	-
MRC6	0.91	277	0	-	-	-	-	-	-	-	-	-
MRC7	-	-	-	-	-	-	0.85	161	11	-	-	-
MRC8	-	-	-	-	-	-	-	-	-	-	-	-
HJA1	0.68	331	13	-	-	-	-	-	-	-	-	-
HJA2	-	-	-	-	-	-	-	-	-	-	-	-
HJA3	-	-	-	-	-	-	-	-	-	-	-	-
HJA4	-	-	-	0.66	161	20	0.79	173	19	-	-	-
HJA5	-	-	-	0.59	169	16	0.69	170	17	-	-	-
HJA6	-	-	-	0.67	149	20	0.84	157	11	-	-	-
HJA7	0.58	335	14	-	-	-	-	-	-	-	-	-
HJA8	-	-	-	-	-	-	0.79	119	17	-	-	-

Examples of Q_{TOT} versus R_{TOT} scatter plots for all sites, regardless of whether a threshold was present or not, are shown in Figure 3-2. When a R_{TOT} threshold was present and led to a change in response magnitude, the slope of the relationship after the threshold was greater than that before the threshold (i.e., $m_2 > m_1$). Generally, m_1 was greater than 0; a nonlinear relationship with a typical hockey-stick shape (i.e., $m_1 \approx 0$) was only observed at the PMRW.

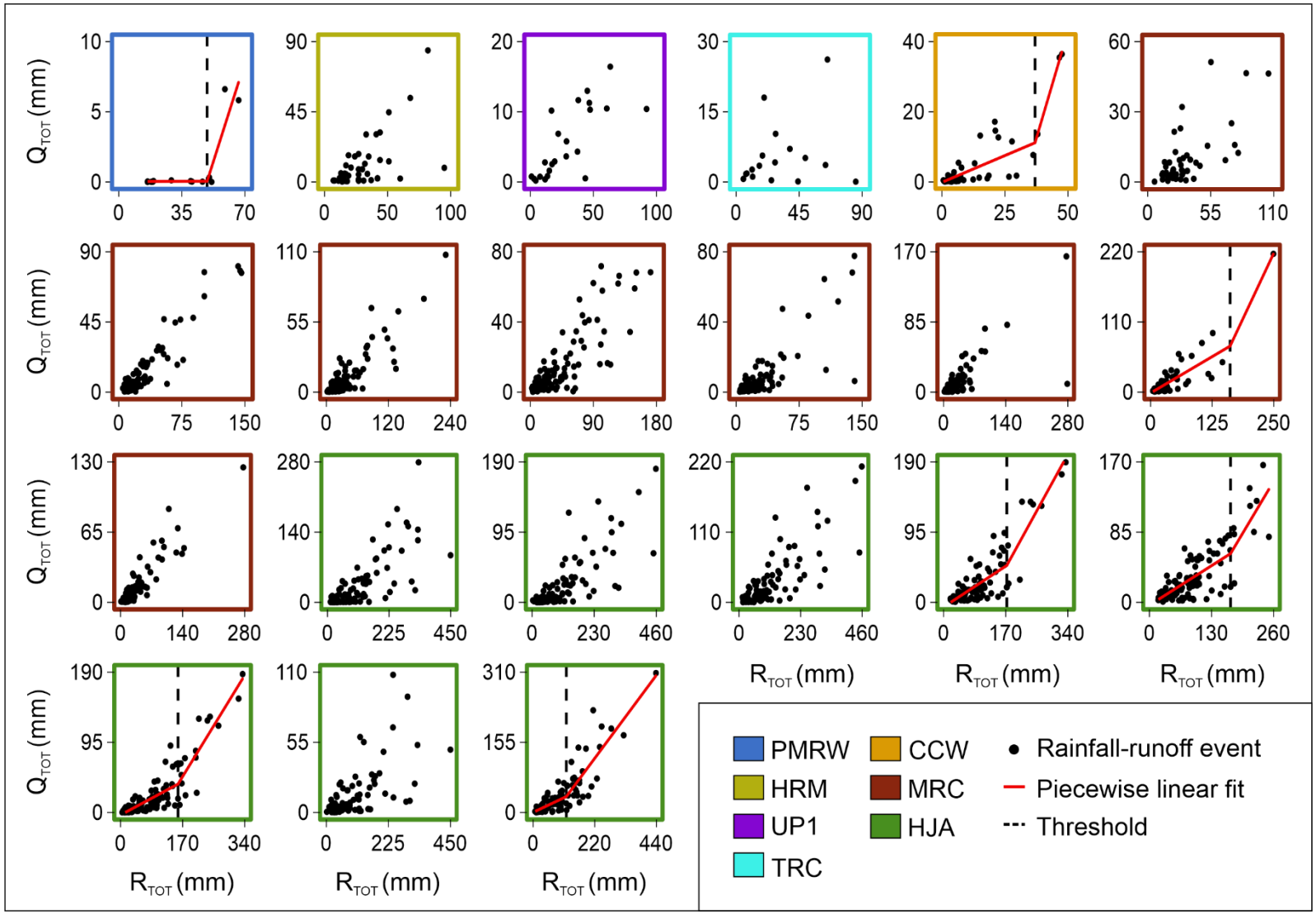


Figure 3-2. Site-specific scatter plots of Q_{TOT} against R_{TOT}. For sites with thresholds, the best fit lines for the piecewise linear model and the threshold value are shown. Note that each site has unique x-axis and y-axis ranges, for better readability.

3.3.2 Thresholds involving a wider set of meteorological factors and response metrics

Scatter plots showing a wider set of magnitude and timing metrics against meteorological factors related to rainfall, that consider evapotranspiration or rainfall intensity, suggested threshold behaviour at numerous sites (Figure 3-3). At 20 sites, thresholds of AR_X+R_{TOT} for response magnitude were observed. Thresholds of R_{TOT}, AR_X, $APET_X$, and AR_X-$APET_X$ for response magnitude were observed at fourteen, fifteen, one, and sixteen sites, respectively. Rainfall intensity thresholds (RI_{AVG} and/or RI_{MAX}) for response magnitude were observed at the PMRW, MRC5, and MRC7 sites, but were absent at other sites. Thresholds for response timing were less common than thresholds for response magnitude, except for the MRC4 site. When present, thresholds for response timing were mostly attributable to AR_X+R_{TOT} and AR_X-$APET_X$.

The extent to which thresholds were observed for different relationships is summarized in Table 3-4 using threshold detection frequencies (i.e., F values, refer to section 2.3). In terms of thresholds for response magnitude, rainfall depth thresholds had the largest detection frequencies (F mean: 14%, F range: 0-22%) at all 20 sites. In contrast, thresholds in rainfall intensity for response magnitude had comparatively small detection frequencies (F mean: 2%, range: 0-12%), as did thresholds in evapotranspiration-related factors (F mean: 4%, F range: 0-7%). In general, thresholds for response timing metrics were less common and were observed at only eight sites. Thresholds for response timing, however, had relatively large detection frequencies at the PMRW (F: 0-24%) and UP1 (F: 0-14%) sites.

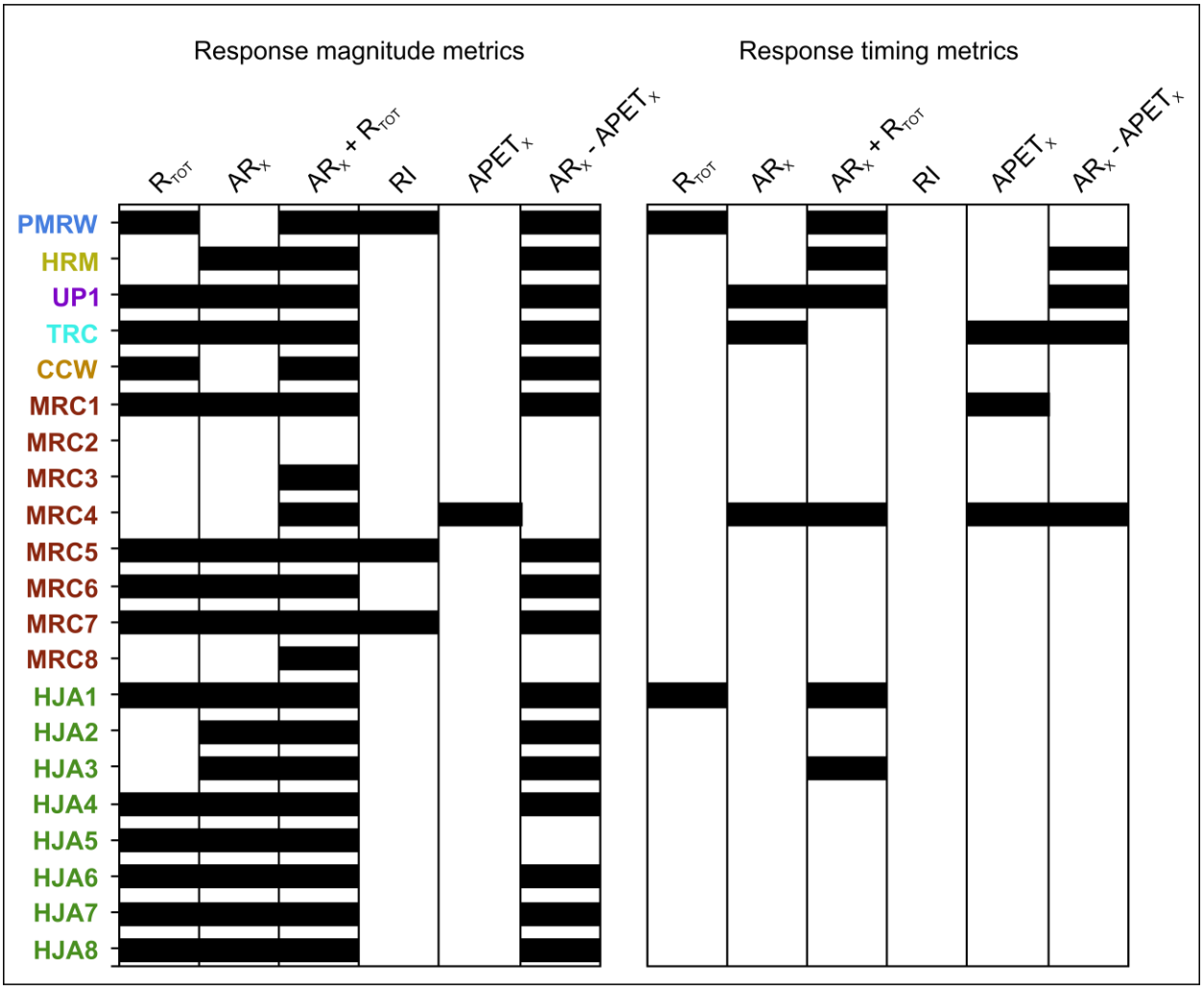


Figure 3-3. Summary of PRA results (i.e., presence/absence of thresholds for at least one metric) across sites, for different pairs of response metrics and meteorological factors. Results are separated by response metric type (magnitude – left, and timing – right). Each column is associated with a different type of meteorological factor. RI includes both RI_{AVG} and RI_{MAX}.

Table 3-4. Threshold detection frequencies (F values) for a wide set of response metrics and meteorological factors. Frequencies are categorized by response metric and meteorological factor type. The total number of possible thresholds is shown in the 'Total' row. The average threshold frequency for each category is shown in the 'Average' row.

	Magnitude vs Rainfall Depth	Magnitude vs Rainfall Intensity	Magnitude vs Abstractions	Timing vs Rainfall Depth	Timing vs Rainfall Intensity	Timing vs Abstractions
Total	60	8	56	45	6	42
Site	Frequency (%)	Frequency (%)	Frequency (%)	Frequency (%)	Frequency (%)	Frequency (%)
PMRW	22	12	2	24	0	0
HRM	18	0	7	2	0	2
UP1	22	0	7	11	0	14
TRC	12	0	7	4	0	7
CCW	13	0	2	0	0	0
MRC1	22	0	9	0	0	2
MRC2	0	0	0	0	0	0
MRC3	7	0	0	0	0	0
MRC4	2	0	2	4	0	5
MRC5	13	12	2	0	0	0
MRC6	20	0	7	0	0	0
MRC7	18	12	4	0	0	0
MRC8	12	0	0	0	0	0
HJA1	17	0	5	9	0	0
HJA2	12	0	5	0	0	0
HJA3	10	0	4	2	0	0
HJA4	12	0	2	0	0	0
HJA5	15	0	0	0	0	0
HJA6	15	0	2	0	0	0
HJA7	15	0	4	0	0	0
HJA8	12	0	2	0	0	0
Average	14	2	3	3	0	1

Figure 3-4 includes scatter plots showing T_{LR} against AR_1 – a relationship devoid of threshold behaviour for all twenty-one sites. While thresholds, as defined in this study, were not observed for these scatter plots, some notable features suggest behavioural changes in response timing at specific antecedent rainfall amounts. For example, many scatter plots for the MRC and HJA sites show variance collapse, where T_{LR} converges for events with AR_1 larger than a specific value. At MRC1 and MRC2, this value is approximately 10 mm, while for most catchments of the HJA, this value ranges between 15 and 20 mm. Figure 3-5 shows scatter plots of Q_{TOT} against AR_1+R_{TOT}, a relationship for which threshold behaviour was observed at eight sites. These thresholds are visually similar to thresholds that were observed in relation to research question 1 (i.e., in Figure 3-2): most thresholds had a non-zero m_1 slope and a larger m_2 slope.

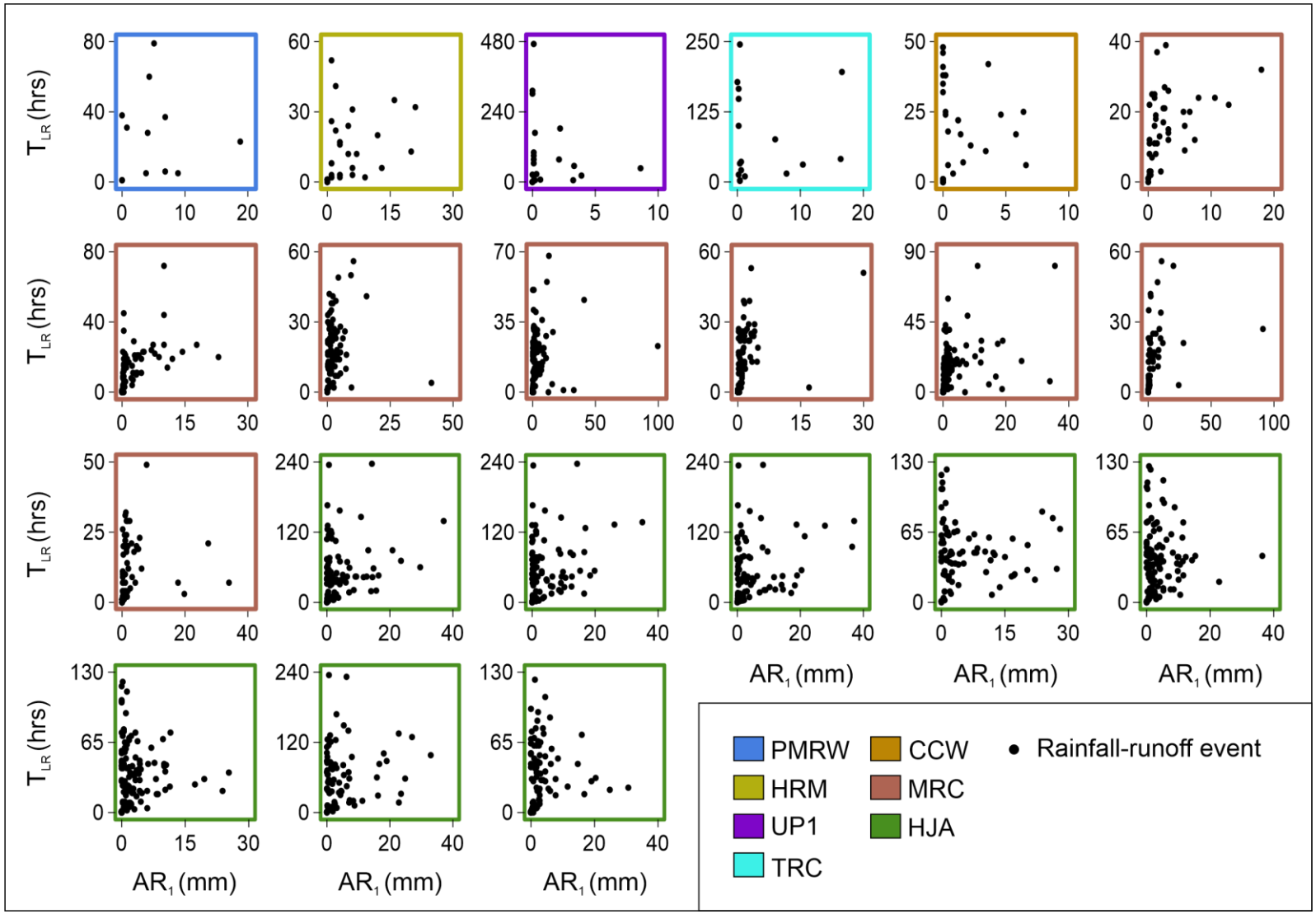


Figure 3-4. Site-specific scatter plots of T_{LR} against AR_1. Note that each site has unique x-axis and y-axis ranges.

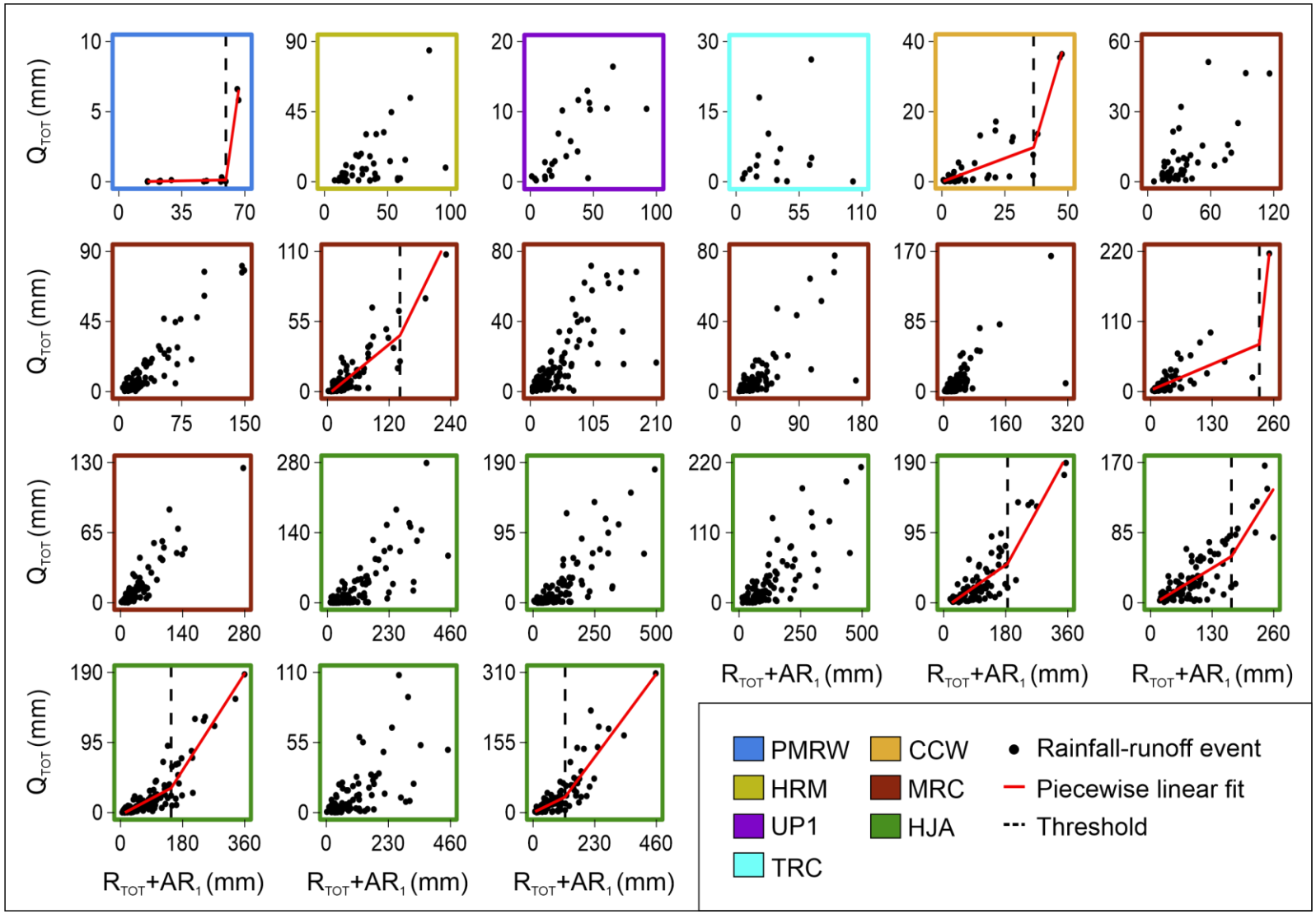


Figure 3-5. Site-specific scatter plots of Q_{TOT} against $R_{TOT}+AR_1$. For sites with thresholds, the best fit line for the piecewise linear model and the threshold value are shown. Note that each site has unique x-axis and y-axis ranges.

One auxiliary question investigated in this study was whether the same threshold value of a given meteorological factor could trigger a change in different response metrics. This could only be done for sites with sufficiently large numbers of thresholds (Figure 3-6). For the PMRW, and select sites of the HJA, R_{TOT}+AR_1 thresholds for Q_{TOT} were mostly similar to R_{TOT} thresholds for Q_{MAX} (i.e., points plotting near the 1:1 line in Figure 3-6A). Similarly, for the PMRW and HJA1 sites, R_{TOT}+AR_{10} thresholds for I_{abs} were similar to R_{TOT}+AR_{10} thresholds for T_{LR} (Figure 3-6B).

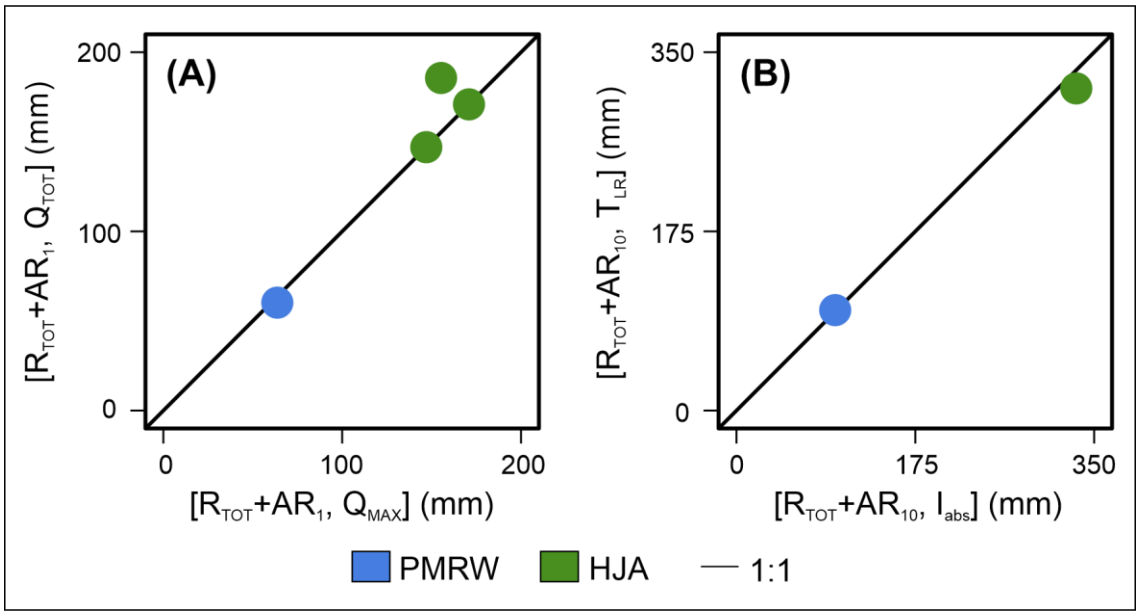


Figure 3-6. Scatter plots comparing threshold values of a given meteorological factor triggering a change in response metrics. Plot (A) includes R_{TOT}+AR_1 thresholds for Q_{MAX} and Q_{TOT}, while plot (B) includes R_{TOT}+AR_{10} thresholds for I_{abs} and T_{LR}. Sites plotting on or near the 1:1 line indicates similar threshold values for the x-axis and y-axis relationships. Note that each panel has unique x-axis and y-axis ranges.

3.3.3 Thresholds of factors calculated over different antecedent window durations

The number of precipitation and potential evaporation thresholds that was observed at short- (5 days or less), medium- (between 5 and 14 days) and longer-term (14 days or more) antecedent durations is shown in Figure 3-7. This figure shows that AR_X (Figure 3-7A), AR_X+R_{TOT} (Figure 3-7B), $APET_X$ (Figure 3-7C), and AR_X-$APET_X$ (Figure 3-7D) thresholds for response timing were uncommon, regardless of antecedent duration. Except for $APET_X$, threshold behaviour involving these factors was, however, more frequently observed for response magnitude. AR_X (Figure 3-7A) and AR_X-$APET_X$ (Figure 3-7D) thresholds for response magnitude were most commonly observed for 3-day to 10-day antecedent durations (Figure 3-7A). Interestingly, the number of AR_X+R_{TOT} thresholds observed for response magnitude decreased, albeit not monotonically, with increasing antecedent duration. $APET_X$ thresholds for response magnitude were uncommon and were only observed for the 30-day antecedent duration (Figure 3-7C).

When thresholds for each antecedent window duration were separated by site, a few patterns emerged. Figure 3-8 reiterates that across all sites, $APET_X$ thresholds were uncommon, regardless of antecedent duration. For the HRM site, AR_X and AR_X-$APET_X$ thresholds were observed for most antecedent durations, except for the 1-day antecedent duration, while AR_X+R_{TOT} thresholds were more frequently observed at medium antecedent durations. At the UP1 and TRC sites, threshold behaviour was more commonly observed for factors calculated over medium antecedent durations. Relationships at the CCW rarely showed threshold behaviour; however, a small number of AR_X+R_{TOT} thresholds were observed at each antecedent duration. For the MRC and HJA sites, threshold behaviour was mostly observed for short and

medium antecedent window durations. In contrast, at the PMRW site, AR_X+R_{TOT} thresholds were common regardless of antecedent duration, and AR_X-$APET_X$ thresholds were only observed at shorter antecedent durations.

Lastly, the nested configuration of the MRC and HJA catchments was used to evaluate whether threshold behaviour was influenced by drainage area. Since these study areas include only eight nested catchments each, bar charts rather than statistical techniques were used to qualitatively assess possible correlations between drainage area and the number of relationships for which threshold behaviour was observed. Slightly more AR_X, AR_X+R_{TOT}, and AR_X-$APET_X$ thresholds were observed at the smallest HJA sites than larger HJA sites. However, bar charts in Figure 3-9A and Figure 3-9B do not suggest other relationships between drainage area and the number of thresholds observed.

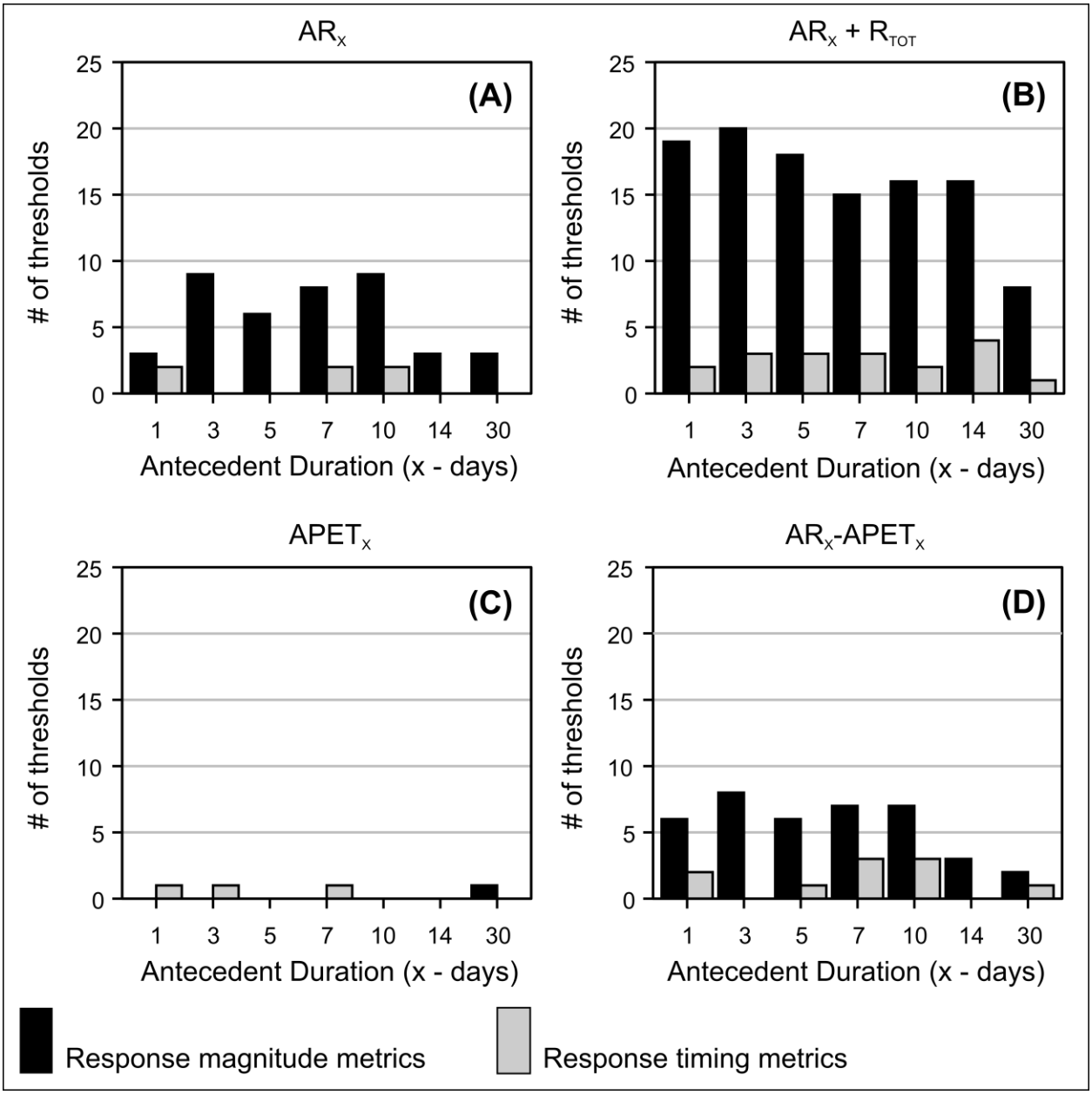


Figure 3-7. Bar charts indicating the number of thresholds observed for all sites at 1, 3, 5, 7, 10, 14, and 30-day antecedent window durations. Individual plots (A), (B), (C) and (D) consider the number of AR_X, AR_X+R_{TOT}, $APET_X$, and AR_X-$APET_X$ thresholds, respectively.

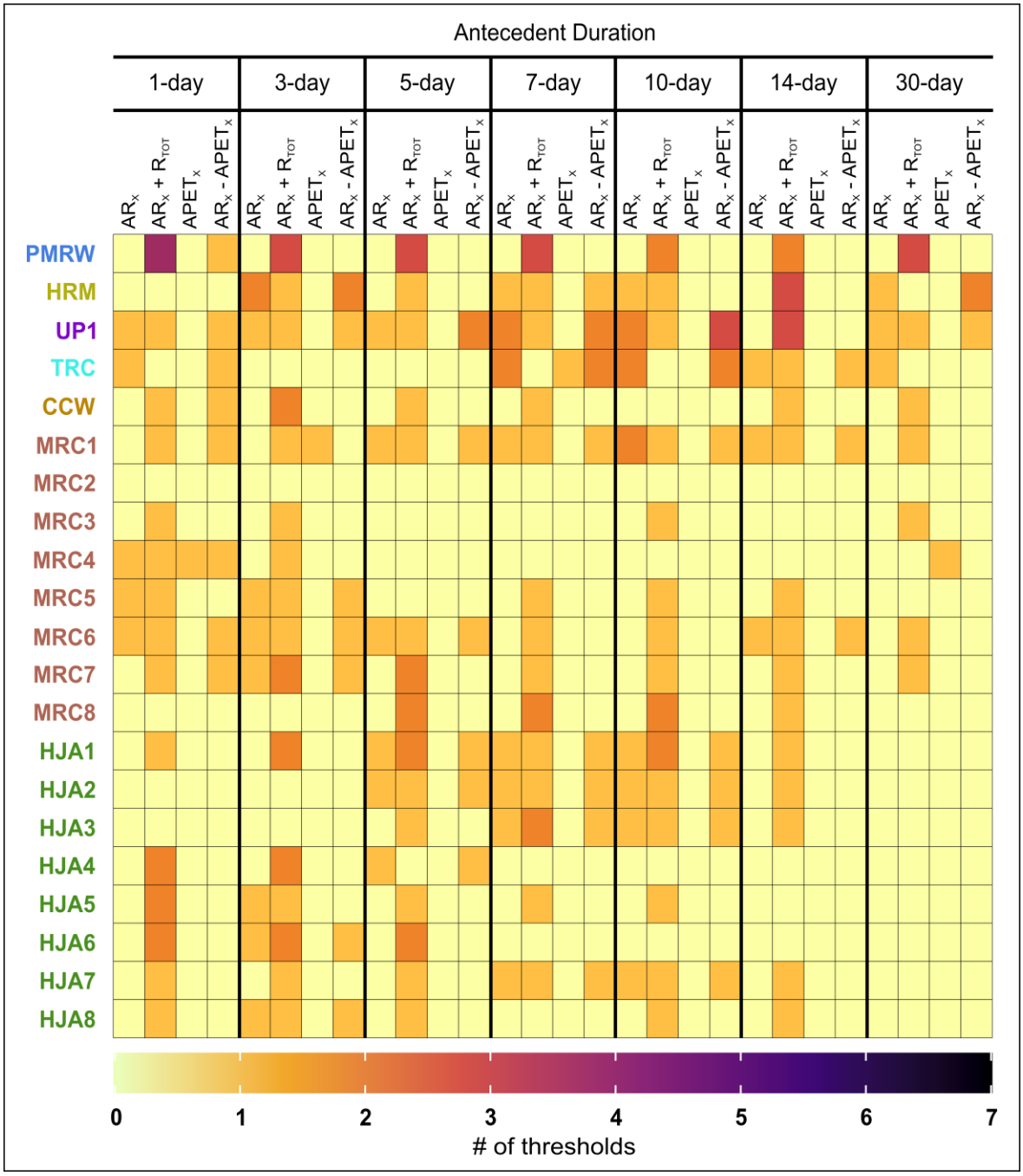


Figure 3-8. Heatmap summarizing the number of thresholds observed at each antecedent window duration. Each cell indicates the number of thresholds that were observed at each site for a specific type of meteorological factor and a specific antecedent window duration.

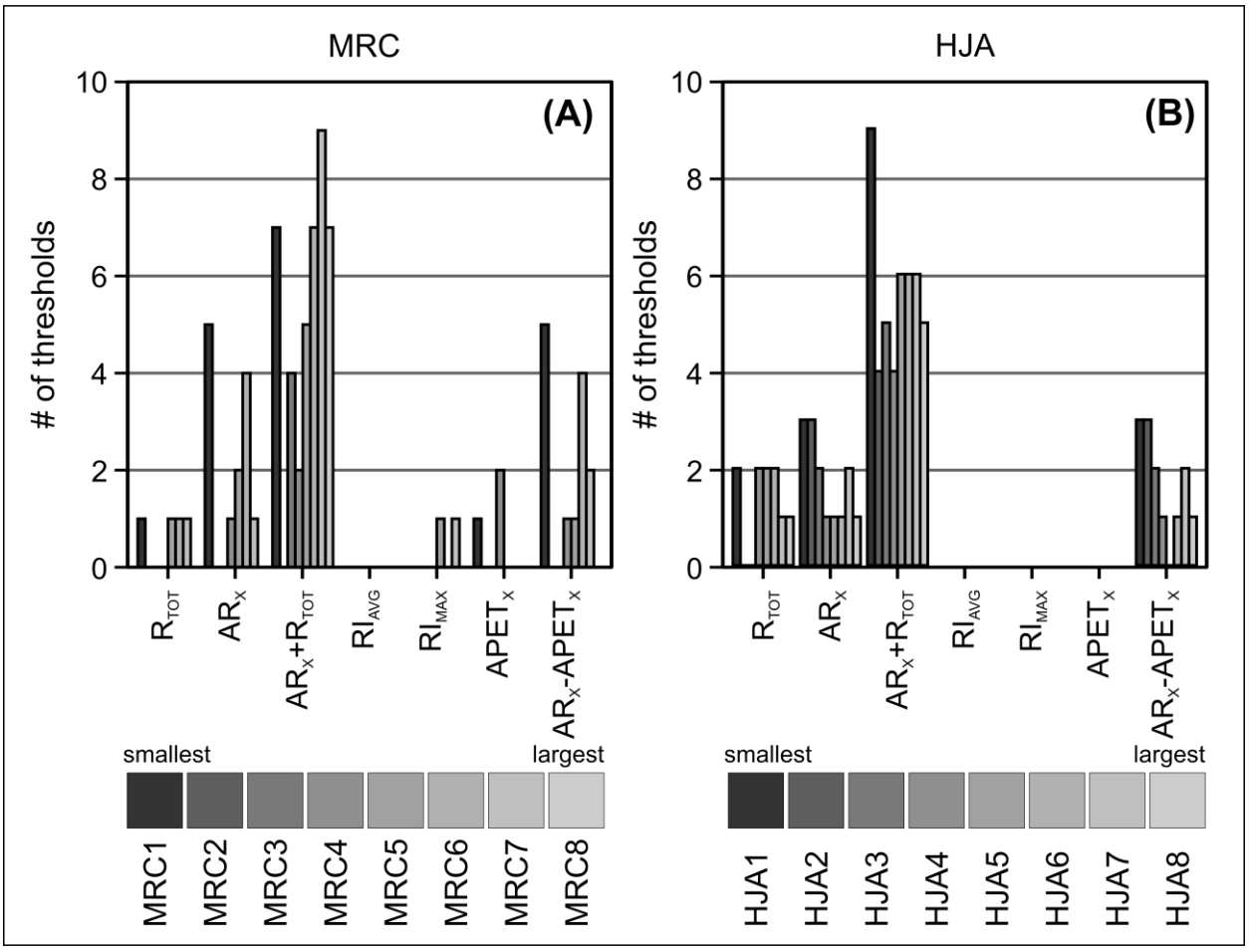


Figure 3-9. Bar charts showing the number of thresholds observed for seven different meteorological factor types across nested catchments of the MRC (A) and HJA (B).

3.4 Discussion

3.4.1 Threshold behaviour for different response metrics and meteorological factors

Total event rainfall (i.e., R_{TOT}) thresholds for response magnitude were observed for most (14 out of 21) study sites, but not for every magnitude metric at every site. R_{TOT} thresholds for $I_{abs,}$ Q_{MAX}, and Q_{TOT} were more common than R_{TOT} thresholds for RR. This suggests variability

in the presence of nonlinear response at critical R_{TOT} values for different aspects of response magnitude. Previous work using the same input data also showed that $I_{abs,}$ Q_{MAX}, and Q_{TOT} are more effective at capturing the variability of hydrologic response among events, compared to RR (Ross et al., 2019). Past research at the PMRW site observed a rainfall threshold for trench flow (Tromp-van Meerveld & McDonnell, 2006a), which is analogous to the R_{TOT} threshold for Q_{TOT} observed in the current study. Other studies have observed storage thresholds for sites of this study, albeit using different meteorological factors and response metrics. For the UP1 catchment, thresholds were observed for relationships between event runoff and S_{TOTAL} (Oswald et al., 2011), where S_{TOTAL} is an estimate of storage derived from soil depth, soil properties, topography, water table data, and precipitation data. Similarly, at HJA sites, thresholds were observed for scatter plots involving streamflow and the sum of gross precipitation and the antecedent soil moisture index (ASI), where the ASI was estimated from soil moisture data collected at four soil profiles (Detty & McGuire, 2010). Since these studies used response metrics and meteorological factors derived from data other than precipitation, temperature, and streamflow data, their results are not directly comparable to the results of the current study. That said, although not directly comparable, some threshold values reported here (e.g., R_{TOT} threshold of 335 mm at HJA7) are larger than those reported in previous studies (e.g., Detty & McGuire, 2010). This may be explained by the fact that these previous studies reported critical meteorological factor values required to initiate response (i.e., runoff-initiation thresholds), while the thresholds observed in the current study were not necessarily runoff-initiation thresholds. Most runoff initiation thresholds have been attributed to response from matrix flow, macropore flow, or pipe flow (McGuire & McDonnell, 2010; Tromp-van Meerveld & McDonnell, 2006b; Uchida et al., 2005; Wei et al., 2020). However, others have identified relationships with

multiple thresholds: in these multi-threshold cases, the runoff initiation threshold is followed by a secondary threshold (referred to as a "rise threshold") when additional rainfall triggers a transition to different runoff generation processes (Wei et al., 2020). The thresholds are therefore sequential and manifest as multiple, distinct breakpoints at two or more different moments in time (Wei et al., 2020). In the case of the HJA7 site, the larger threshold value that was observed in the present study could be associated with a "rise threshold": it may be caused by the addition of rainfall onto wetted soils that results in the activation of matrix or preferential flow in hillslope soils, thereby increasing the hillslope-riparian-stream connectivity (McGuire & McDonnell, 2010), and leading to a nonlinear response change (Sidle et al., 2000; Wei et al., 2020). If such a hypothesis were true, however, it would raise the question of why the PRA did not identify the runoff initiation threshold but rather identified a subsequent "rise threshold". Multi-breakpoint PRA approaches or alternate threshold identification methods are warranted to better evaluate the presence of sequential initiation and rise thresholds.

In the recent literature, threshold-mediated runoff response has typically been documented using response magnitude metrics like Q_{TOT}, Q_{MAX}, and RR. In the present study, the consideration of relationships involving I_{abs}, which is a less commonly used response magnitude metric, showed thresholds at all sites, except for three of the MRC sites. Since by definition, I_{abs} is the storage deficit that must be exceeded before the initial hydrograph rise (Dingman, 2015), an observed threshold for I_{abs} may also be interpreted as the presence of more than one threshold in a single relationship. Thresholds for timing metrics were uncommon, compared to thresholds for response magnitude. It should be noted that this study only considered T_{LR}, T_{LP}, and T_c: while many timing metrics can be calculated, previous work at the

sites examined in the present study identified T_{LR}, T_{LP}, and T_c as the most important timing metrics for capturing temporal variability in hydrologic response (Ross et al., 2019).

In terms of meteorological factors, previous studies mostly considered water volumes or depths, but less emphasis was put on hydrologic abstractions (e.g., evapotranspiration and effective rainfall) and rainfall intensity. In the present study, rainfall intensity, $APET_X$, and AR_X-$APET_X$ thresholds were observed at three, three, and seventeen sites, respectively. This suggests that rainfall intensity, potential evapotranspiration, and effective rainfall can be important controls of nonlinear hydrologic response. Even for some humid areas with high infiltration capacity like the MRC, where rainfall intensity is not expected to strongly influence runoff processes, rainfall intensity thresholds for response magnitude were observed. One reason for the presence of rainfall intensity thresholds at sites where infiltration capacity is not a limiting factor could be a strong correlation between rainfall depth and rainfall intensity. For the sites included in the present study, however, the Spearman's rank correlation (rho) coefficient between those two variables varied significantly ($0 \leq |rho| \leq 0.75$). For study areas where rainfall intensity thresholds were observed, correlations between R_{TOT} and RI_{MAX} were strong ($rho > 0.6$), but this was also true for study areas where rainfall intensity thresholds were not observed (e.g., HJA sites: $0.48 \leq |rho| \leq 0.75$). Interestingly, in dryer Prairie environments like the CCW, where infiltration-excess overland flow is assumed to be a dominant runoff generation mechanism (Fang et al., 2007), rainfall intensity thresholds were not observed. This suggests that thresholds of different meteorological factor types (i.e., rainfall depth, hydrologic abstractions related to evapotranspiration, or rainfall intensity) may not always be predicted based on *a priori* hypotheses about dominant runoff generation mechanisms.

3.4.2 Controls on threshold behaviour

We explored the potential influence of antecedent conditions and drainage area on threshold behaviour. AR_X+R_{TOT} thresholds for response magnitude were observed at 20 out of 21 sites. For most sites, AR_X+R_{TOT} and AR_X-$APET_X$ thresholds were more common than individual AR_X thresholds, R_{TOT} thresholds, and $APET_X$ thresholds. These findings indicate that response magnitude is often contingent on both event and antecedent conditions, which confirms conclusions from other studies. Notably, previous research at the HJA observed a combined ASI + gross precipitation threshold, while thresholds controlled exclusively by ASI or gross precipitation were absent (Detty & McGuire, 2010). In the current study, the presence of threshold behaviour was variable among meteorological factors calculated over different antecedent durations. For example, threshold behaviour was more commonly observed for AR_X+R_{TOT} and AR_X factors calculated over 3-, 5-, 7-, and 10-day antecedent durations and for AR_X-$APET_X$ factors calculated over 7- and 10-day antecedent durations. This result indicates that overall, catchment memory over three to ten days preceding an event is an important control on nonlinear response for sites considered in this study. This result also suggests that nonlinear response behaviour is dependent on antecedent moisture conditions over specific periods preceding a rainfall-runoff event.

The nested configuration of the MRC and HJA sites allowed for the qualitative assessment of correlations between the number of thresholds observed and drainage area. Since larger catchments often require more spatially extensive rainfall to initiate runoff response compared to smaller catchments, and large rainfall events are less common than smaller, spatially isolated rainfall events, a negative correlation between drainage area and the number of

thresholds observed was anticipated. Only the number of AR_X, AR_X+R_{TOT}, and AR_X-APET_X thresholds that were observed at HJA sites appeared qualitatively related to catchment drainage area. Also, this was not the case for other meteorological factor types or the MRC sites. The lack of consistent correlations for different factor types may indicate an intricate relationship between drainage area and threshold behaviour. Others have suggested that different controls on hydrologic response may interact (Merz & Blöschl, 2009; Yadav et al., 2007), which is one potential reason for these inconsistencies. Alternatively, the influence of drainage area on hydrologic response may be spatially and/or temporally variable due to the relative influence of other controls. For example, Devito et al. (2005) suggested that the importance of different controls on response is hierarchical with climatic controls being the most important. One last possibility is that drainage area inadequately represents a scale-based control on threshold behaviour. Indeed, there is a clear distinction between catchment drainage area, which is topographically defined, and the spatial extent of temporally variable contributing areas that directly affect hydrologic response (Ambroise, 2004; Betson, 1964; Hewlett & Hibbert, 1967). There remain opportunities to assess how the scale of drainage network connectivity or the partitioning between vertical and lateral flowpaths might influence threshold behaviour. This could be explored in future research by assessing threshold behaviour within existing frameworks like the T^3 template of typology, topography and topology (Buttle, 2006). It should be noted that while others have done correlation analysis relating physiographic variables to actual threshold values (Ali et al., 2015), here the focus was rather on evaluating potential controls on the presence or absence of threshold behaviour. Only the potential effect of drainage area on the presence of thresholds was considered since drainage area is the only physiographic variable that could be isolated effectively.

3.4.3 Study limitations related to rainfall-runoff events

In the present study, we evaluated threshold behaviour using a large variety of response metric and meteorological factor pairs for twenty-one sites from seven study areas using a standardized approach. One limitation of our study, however, is the exclusion of snowmelt and rain-on-snow events. While snow is an important component of the hydrologic cycle for some of the selected sites, only rainfall-triggered events were considered. This methodological choice was driven by the fact that high-frequency snowfall and snowmelt data were not available, and HydRun is only designed to delineate rainfall-runoff events (Tang & Carey, 2017). Furthermore, while some past studies did identify seasonal differences in threshold behaviour (Scaife & Band, 2017; Wei et al., 2020), such differences were not explored here. Although many rainfall-runoff events were considered in the present study, the number of events per site varied significantly due to shorter measurement periods for some sites and the fact that drier sites tend to have fewer rainfall-runoff events. These rainfall-runoff events were therefore not separated based on season or year, even though some sites are subject to seasonal changes (e.g., leaf-on versus leaf-off periods) or annual differences in climatic conditions (e.g., wet versus dry versus average years) that may change how meteorological factors influence hydrologic response.

3.4.4 Confronting conceptual and operational threshold definitions

In hydrology, while conceptual definitions of thresholds are well documented in the literature, varying approaches are being used to “operationalize” those definitions, i.e., to identify criteria and statistical models thanks to which thresholds can be identified from data.

Thresholds in runoff response are generally conceptually defined as critical moments in time or points in space at which runoff behaviour rapidly changes (Ali et al., 2013; Phillips, 2006). Multiple operational threshold definitions adhere to this conceptual definition, including definitions associated with nonlinear input-output relationships that are typically represented by diagnostic shapes (e.g., hockey-stick, sigmoidal, Dirac, step-wise) (Ali et al., 2013; Lehmann et al., 2007; Saffarpour et al., 2016). Hockey-stick shaped thresholds have received the most attention in the literature (e.g., Ali et al., 2015; Detty & McGuire, 2010; Graham & McDonnell, 2010; Tani, 1997; Weiler, 2005) and this type of threshold focuses on a change in slope in an input-output relationship. As mentioned in Section 3.2.3, most threshold studies have assessed potential threshold behaviour associated with changes in slope visually (e.g., Ali et al., 2011; Detty & McGuire, 2010; Graham et al., 2010; James & Roulet, 2007; Lehmann et al., 2007; McGlynn & McDonnell, 2003; Mosley, 1979; Sidle et al., 2000; Tani, 1997; Tromp-van Meerveld & McDonnell, 2006a; Whipkey, 1965), and few of these studies describe explicit criteria for visual threshold identification. The PRA that was used in the current study and other studies (e.g., Oswald et al., 2011; Scaife & Band, 2017) also focuses on thresholds associated with a change in slope. In this context, threshold identification using PRA typically involves a statistical assessment of whether the data offers more support for the piecewise linear model than a competing model (usually the simple linear model). In the current study, criterion 1 (related to the R^2 of the piecewise linear model) and criterion 2 (related to the AIC of the piecewise linear model and simple linear model) were used to verify the appropriateness of the piecewise linear model for representing the observed response behaviour. Criterion 2 is a statistically robust approach for comparing the piecewise linear model and simple linear model (Sakamoto et al., 1986) and also accounts for the influence of sample size (i.e., number of rainfall-runoff events).

For all relationships that were deemed threshold mediated in this paper, the AIC of the piecewise linear model was between 2.01 and 221.31 units below the AIC of the simple linear model. In the current study, additional criteria related to relationship slope were also imposed on the PRA to ensure that relationships deemed threshold mediated were associated with a noticeable change in response. Specifically, criterion 3 stipulated that the percent difference between the below-threshold slope (m_1) and the above-threshold slope (m_2) must be greater than 10%. It should be noted that the existing literature offers no guidance about the minimum slope difference required to assess the "rapidly changing runoff behaviour" associated with the conceptual definition of hydrological thresholds. Besides, it is likely that the required minimum slope difference is site-specific and depends on climate, drainage area, soil characteristics, and/or other factors. To assess the potential effects of the minimum slope difference criterion on threshold detection results, a sensitivity analysis was conducted: PRA were performed with the minimum percent difference between the below-threshold and above-threshold slopes set at 10, 20, 30, 40, and 50%. This sensitivity analysis showed that threshold detection for the sites considered in this study did not change considerably across the 10-50% minimum slope difference range that was tested (Appendix A-1). It is also notable that in the present study and other recent ones (e.g., Scaife & Band, 2017; Wei et al., 2020), the exclusion of possible slower stormflow generation was avoided by not assuming a below-threshold slope (i.e., m_1) of 0. This is fundamentally different from operational definitions of threshold behaviour found in other studies that only assess a below-threshold regime that is characterized by negligible runoff response (i.e., $m_1 = 0$).

In general, the use of PRA to assess the presence or absence of threshold behaviour – in the present study and others – can be considered a narrow interpretation of conceptual threshold definitions, since it allows only threshold behaviour associated with a change in slope to be

considered. Other possible operational threshold definitions are associated with a change in response variance. The present study showed that some notable changes in hydrological response that adhere to conceptual threshold definitions were not identified as breakpoints by PRA. One example of this is presented in Figure 3-4, where some scatter plots suggest variance collapse in T_{LR} at critical AR_1 values. One possible explanation for this response change is that multiple runoff processes with unique flow pathways initially contribute to hydrological response but once a critical AR_1 value is exceeded, one of those pathways becomes dominant or the signals of multiple processes with unique pathways become integrated. One hypothetical example of this might be for sites that exhibit hydrologic response as a result of saturation-excess flow after critical volumes of rainfall are exceeded. In such a case, multiple subsurface runoff generation mechanisms with distinct flow pathways (e.g., matrix flow, macropore flow) may be active, but once response from saturation-excess flow is triggered, the signal of these other mechanisms may be overshadowed. As variance collapse and other patterns indicative of response change are not typically labeled as thresholds (despite fitting the textual definition), hydrologists may want to consider tools other than PRA to identify any notable changes in hydrological behaviour, including those that cannot be represented by monotonic mathematical functions.

3.4.5 Typology of threshold behaviour

In hydrology, classification schemes and typologies have facilitated the conceptualization of complex systems. There are useful classifications of catchments (e.g., McDonnell & Woods, 2004; Wagener et al., 2007), runoff generation mechanisms (e.g., McDonnell, 2013), hysteretic processes (e.g., Evans & Davies, 1998) and even hydrologic models (e.g., Beven, 2011).

Numerous studies have observed nonlinearities in rainfall-runoff responses and have demonstrated diverse and complex threshold behaviours. However, a typology of threshold dynamics does not yet exist. The large number of threshold behaviours observed in the current study provides an opportunity to group thresholds in at least three different ways.

First, thresholds can be grouped based on the number of response metrics that simultaneously change behaviour at a single critical meteorological factor value. In most cases in the present study, an individual threshold influenced a single response metric. However, in some instances, a threshold influenced multiple response metrics (Figure 3-6). Therefore, two types of threshold dynamics could be distinguished based on the number of response metrics influenced by a single critical value of a meteorological factor. **Restricted threshold dynamics** could refer to a critical value of a meteorological factor that triggers a change in a single response metric, while **pervasive threshold dynamics** could refer to a critical value of a meteorological factor that triggers a change in two or more response metrics. Both restricted and pervasive threshold dynamics were observed in this study: R_{TOT} thresholds for response magnitude metrics at the MRC6 and PMRW site hinted at restricted and pervasive threshold dynamics, respectively.

Second, the present study considered the effects of antecedent conditions, computed over several durations, on threshold behaviour. Some sites were shown to be more or less sensitive to short-, medium- and longer-term antecedent conditions, which is indicative of different, and highly site-specific memory effects. Figure 3-8 notably shows that some thresholds may be present over a specific antecedent duration, while others are present across very different antecedent durations and therefore seem unaffected by antecedent conditions. Three types of threshold dynamics associated with different memory effects could, therefore, be distinguished. Threshold behaviour that is present solely when considering a short antecedent duration (i.e., 5

days or less) would indicate **short-memory threshold dynamics**, while threshold behaviour that is present solely when considering a long antecedent duration (i.e., 14 days or more) would indicate **long-memory threshold dynamics**. As for threshold behaviour that is present regardless of antecedent duration, it could be put in a category akin to **memory-independent threshold dynamics.** In the current study, AR_X thresholds observed at the HJA6 site suggest short-memory threshold dynamics, while AR_X thresholds observed at the TRC site were more aligned with long-memory threshold dynamics. The AR_X+R_{TOT} thresholds observed at the PMRW site appear to be memory-independent.

Third, the present study evaluated scatter plots involving meteorological factors representing rainfall depths, rainfall intensity and hydrologic abstractions related to evapotranspiration. Considering such a large number of factors was important, since their different effects on watershed dynamic storage and runoff generation mechanisms (e.g., infiltration-excess versus saturation-excess) are well documented (e.g., Kirchner, 2009; McDonnell, 2013). Threshold dynamics could therefore by categorized based on the dominant hydrological processes they are assumed to represent or depend on. A **single-process threshold dynamics** category could be used to label a response metric that changes because of a rainfall depth threshold, or a rainfall intensity threshold, or a threshold associated with a hydrologic abstraction. A **multi-process threshold dynamics** category could rather be used to indicate that a response metric is governed by multiple thresholds across more than one meteorological factor type. Examples of such process-threshold dynamics could be observed in the present study: nonlinearities in response at the HRM site seemed to fit the single-process threshold dynamics conceptualization, while changes in response at five out of eight MRC sites seemed to be controlled by multi-process threshold dynamics.

Overall, the proposed typology distinguishes seven different types of threshold dynamics based on three criteria: (1) the number of response metrics that simultaneously change behaviour at a single critical meteorological factor value; (2) the memory effects on thresholds; and (3) the underlying hydrological processes leading to threshold behaviour (Figure 3-10). The types of threshold dynamics derived under the three criteria are not mutually exclusive, as highlighted above using site-specific examples from the present study. Distinguishing threshold dynamics based on criteria 1 may prove particularly useful for modelling exercises since pervasive threshold dynamics would be a strong rationalization for prescribing unique model structure or parameterization for meteorological factor values smaller or larger than a threshold value. Similarly, discriminating threshold dynamics based on criteria 2 might help characterize the resilience of catchments to antecedent conditions that may trigger drastic increases in response. Lastly, distinguishing threshold dynamics based on criteria 3 may facilitate the understanding of how different processes uniquely (or non-uniquely) control nonlinear rainfall-runoff behaviour for individual sites. More broadly, a typology of threshold dynamics could promote standardized descriptions of nonlinear rainfall-runoff behaviour and enable easier inter-site comparisons, provided that the typology can be confirmed across a larger range of sites and conditions where thresholds are present. Beyond validation efforts for multiple sites, further analysis is needed to assess the appropriateness of the proposed typology, especially in scenarios where threshold dynamics may change over seasonal and multi-year timescales, or before and after ecohydrological disturbances.

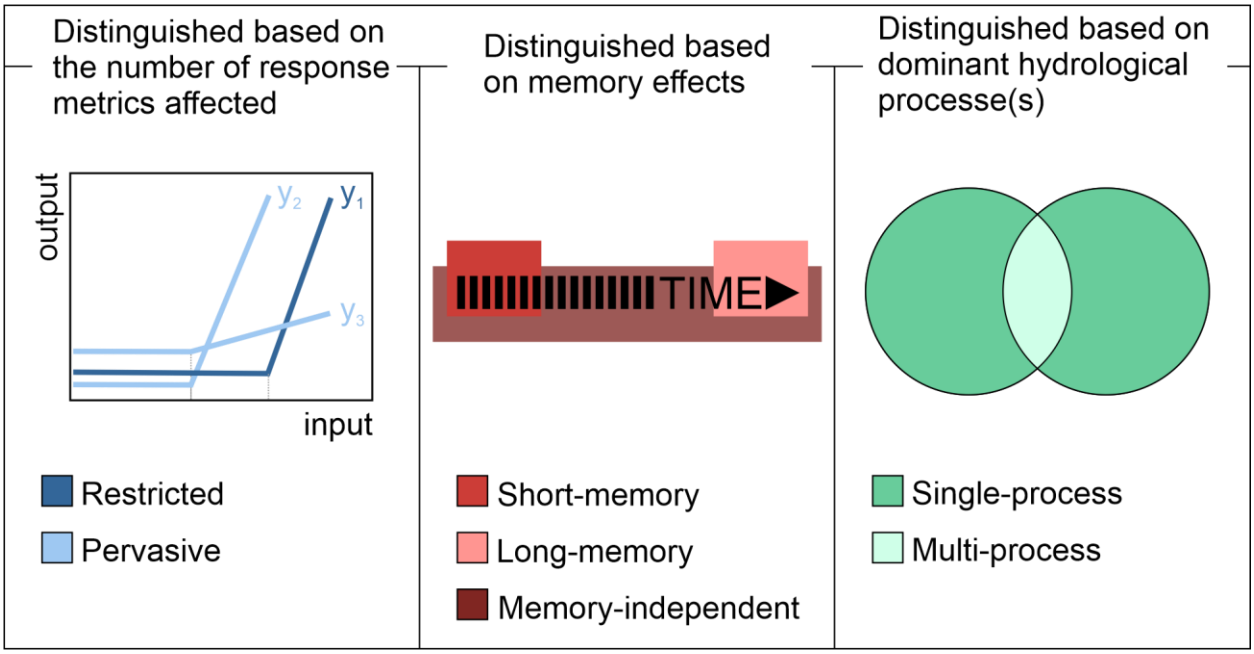


Figure 3-10. Visual representation of the proposed typology distinguishing seven different types of threshold dynamics based on three criteria. Since the three criteria are not mutually exclusive, threshold dynamics can belong to three different types.

3.5 Conclusion

This study contributes to the growing knowledge surrounding hydrologic thresholds in three novel ways. First, an unprecedented suite of meteorological factor and response metric pairs was used to evaluate the presence or absence of threshold behaviour for 21 study sites in seven different geographic areas. Secondly, this study shifted focus from specific threshold values, that are difficult to compare between sites, to the presence or absence of threshold behaviour. This distinction was important, as it allowed the ubiquitous nature of thresholds to be highlighted using commonly available data. Lastly, the present study comprehensively assessed

antecedent conditions as a control of threshold behaviour. Key findings from this study include the following:

(1) Total event rainfall thresholds are one form of rainfall depth threshold affecting response magnitude metrics at 14 out of 21 sites.

(2) Some meteorological factors other than total event rainfall, and which reflect other measures of rainfall depth, rainfall intensity, or hydrologic abstractions related to evapotranspiration, also appear as strong determinants of nonlinear response.

(3) Threshold behaviour is sensitive to both event conditions and antecedent conditions.

(4) Threshold behaviour is sensitive to antecedent conditions over specific durations preceding a rainfall-runoff event.

(5) A typology of threshold dynamics emerges from the data synthesis.

A direct extension of this work would be to further assess the effectiveness of the proposed typology of threshold dynamics. Additional studies evaluating the influence of physiographic, soil, and climatic controls on threshold behaviour, and studies examining changes in threshold behaviour over time, would also be valuable. Lastly, robust characterization of other possible nonlinear relationship shapes (e.g., sigmoidal, dirac, variance collapse) would be beneficial and may lead to automated change identification techniques that capture nuanced, threshold-mediated hydrologic responses.

3.6 References

Ali, G., L'Heureux, C., Roy, A., Turmel, M.-C., & Courchesne, F. (2011). Linking spatial patterns of perched groundwater storage and stormflow generation processes in a

headwater forested catchment. *Hydrological Processes*, *25*(25), 3843–3857. https://doi.org/10.1002/hyp.8238

Ali, G., Oswald, C., Spence, C., Cammeraat, E., McGuire, K., Meixner, T., & Reaney, S. M. (2013). Towards a unified threshold-based hydrological theory: Necessary components and recurring challenges. *Hydrological Processes*, *27*(2), 313–318. https://doi.org/10.1002/hyp.9560

Ali, G., & Roy, A. (2010). Shopping for hydrologically representative connectivity metrics in a humid temperate forested catchment. *Water Resources Research*, *46*(12). http://onlinelibrary.wiley.com/doi/10.1029/2010WR009442/full

Ali, G., Tetzlaff, D., McDonnell, J., Soulsby, C., Carey, S., Laudon, H., McGuire, K., Buttle, J., Seibert, J., & Shanley, J. (2015). Comparison of threshold hydrologic response across northern catchments. *Hydrological Processes*, *29*(16), 3575–3591. https://doi.org/10.1002/hyp.10527

Ambroise, B. (2004). Variable 'active' versus 'contributing' areas or periods: A necessary distinction. *Hydrological Processes*, *18*(6), 1149–1155. https://doi.org/10.1002/hyp.5536

Betson, R. (1964). What is watershed runoff? *Journal of Geophysical Research*, *69*(8), 1541–1552.

Beven, K. (2011). *Rainfall-runoff modelling: The primer*. John Wiley & Sons.

Beven, K. J. (2000). Uniqueness of place and process representations in hydrological modelling. *Hydrology and Earth System Sciences*, *4*(2), 203–213. https://doi.org/10.5194/hess-4-203-2000

Beven, K., Wood, E., & Sivapalan, M. (1988). On hydrological heterogeneity—Catchment morphology and catchment response. *Journal of Hydrology, 100*(1), 353–375. https://doi.org/10.1016/0022-1694(88)90192-8

Buttle, J. (2006). Mapping first-order controls on streamflow from drainage basins: The T3 template. *Hydrological Processes, 20*(15), 3415–3422. https://doi.org/10.1002/hyp.6519

Cammeraat, L. H. (2002). A review of two strongly contrasting geomorphological systems within the context of scale. *Earth Surface Processes and Landforms*, *27*(11), 1201–1222. https://doi.org/10.1002/esp.421

Detty, J. M., & McGuire, K. J. (2010). Threshold changes in storm runoff generation at a till-mantled headwater catchment. *Water Resources Research; Washington*, *46*(7). http://dx.doi.org.uml.idm.oclc.org/10.1029/2009WR008102

Devito, K., Creed, I., Gan, T., Mendoza, C., Petrone, R., Silins, U., & Smerdon, B. (2005). A framework for broad-scale classification of hydrologic response units on the Boreal Plain: Is topography the last thing to consider? *Hydrological Processes, 19*(8), 1705–1714. https://doi.org/10.1002/hyp.5881

Dingman, S. L. (2015). *Physical hydrology*. Waveland press.

Evans, C., & Davies, T. D. (1998). Causes of concentration/discharge hysteresis and its potential as a tool for analysis of episode hydrochemistry. *Water Resources Research*, *34*(1), 129–137. https://doi.org/10.1029/97WR01881

Fang, X., Minke, A., Pomeroy, J., Brown, T., Westbrook, C., Guo, X., & Guangul, S. (2007). A review of Canadian Prairie hydrology: Principles, modelling and response to land use and drainage change. *University of Saskatchewan Centre for Hydrology Report*, *2*. http://www.usask.ca/hydrology/papers/Fang_et_al_2007.pdf

Graham, C. B., & McDonnell, J. J. (2010). Hillslope threshold response to rainfall: (2) Development and use of a macroscale model. *Journal of Hydrology*, *393*(1–2), 77–93. https://doi.org/10.1016/j.jhydrol.2010.03.008

Graham, C. B., Woods, R. A., & McDonnell, J. J. (2010). Hillslope threshold response to rainfall: (1) A field based forensic approach. *Journal of Hydrology*, *393*(1–2), 65–76. https://doi.org/10.1016/j.jhydrol.2009.12.015

Hewlett, J. D., & Hibbert, A. R. (1967). Factors affecting the response of small watersheds to precipitation in humid areas. *Forest Hydrology*, 275–290.

James, A., & Roulet, N. (2007). Investigating hydrologic connectivity and its association with threshold change in runoff response in a temperate forested watershed. *Hydrological Processes*, *21*(25), 3391–3408. https://doi.org/10.1002/hyp.6554

Johnson, S. L., Wondzell, S. M., Jones, J. A., Swanson, F. J., Henshaw, D., & Downing, G. (2019). Long-term research and new findings from experimental catchments at H.J. Andrews Experimental Forest, Oregon. *AGU Fall Meeting Abstracts*, *13*. http://adsabs.harvard.edu/abs/2019AGUFMPA13B1006J

Kim, H. J., Sidle, R. C., Moore, R. D., & Hudson, R. (2004). Throughflow variability during snowmelt in a forested mountain catchment, coastal British Columbia, Canada. *Hydrological Processes*, *18*(7), 1219–1236. https://doi.org/10.1002/hyp.1396

Kirchner, J. W. (2009). Catchments as simple dynamical systems: Catchment characterization, rainfall-runoff modeling, and doing hydrology backward. *Water Resources Research*, *45*(2). https://doi.org/10.1029/2008WR006912

Laudon, H., Sjöblom, V., Buffam, I., Seibert, J., & Mörth, M. (2007). The role of catchment scale and landscape characteristics for runoff generation of boreal streams. *Journal of Hydrology*, *344*(3), 198–209. https://doi.org/10.1016/j.jhydrol.2007.07.010

Lehmann, P., Hinz, C., McGrath, G., Tromp-van Meerveld, H. J., & McDonnell, J. J. (2007). Rainfall threshold for hillslope outflow: An emergent property of flow pathway connectivity. *Hydrol. Earth Syst. Sci.*, *11*(2), 1047–1063. https://doi.org/10.5194/hess-11-1047-2007

McDonnell, J. (2013). Are all runoff processes the same? *Hydrological Processes*, *27*(26), 4103–4111. https://doi.org/10.1002/hyp.10076

McDonnell, J., Sivapalan, M., Vaché, K., Dunn, S., Grant, G., Haggerty, R., Hinz, C., Hooper, R., Kirchner, J., Roderick, M. L., Selker, J., & Weiler, M. (2007). Moving beyond heterogeneity and process complexity: A new vision for watershed hydrology. *Water Resources Research*, *43*(7), W07301. https://doi.org/10.1029/2006WR005467

McDonnell, J., & Woods, R. (2004). On the need for catchment classification. *Journal of Hydrology*, *299*(1), 2–3. https://doi.org/10.1016/j.jhydrol.2004.09.003

McGlynn, B. L., & McDonnell, J. (2003). Role of discrete landscape units in controlling catchment dissolved organic carbon dynamics. *Water Resources Research*, *39*(4), 1090. https://doi.org/10.1029/2002WR001525

McGuire, K. J., & McDonnell, J. J. (2010). Hydrological connectivity of hillslopes and streams: Characteristic time scales and nonlinearities. *Water Resources Research*, *46*(10). https://doi.org/10.1029/2010WR009341

McKee, A., & Druliner, P. (1998). *HJ Andrews Experimental Forest.* http://andrewsforest.oregonstate.edu/pubs/pdf/pub2415.pdf

Merz, R., & Blöschl, G. (2009). A regional analysis of event runoff coefficients with respect to climate and catchment characteristics in Austria. *Water Resources Research*, *45*(1). https://doi.org/10.1029/2008WR007163

Mielko, C., & Woo, M. (2006). Snowmelt runoff processes in a headwater lake and its catchment, subarctic Canadian Shield. *Hydrological Processes*, *20*(4), 987–1000. https://doi.org/10.1002/hyp.6117

Mosley, M. P. (1979). Streamflow generation in a forested watershed, New Zealand. *Water Resources Research*, 15(4), 795–806. https://doi.org/10.1029/WR015i004p00795

Muggeo, V. M. (2008). Segmented: An R package to fit regression models with broken-line relationships. *R News*, *8*(1), 20–25.

Oswald, C., Richardson, M. C., & Branfireun, B. A. (2011). Water storage dynamics and runoff response of a boreal Shield headwater catchment. *Hydrological Processes*, *25*(19), 3042–3060. https://doi.org/10.1002/hyp.8036

Phillips, J. D. (2006). Evolutionary geomorphology: Thresholds and nonlinearity in landform response to environmental change. *Hydrol. Earth Syst. Sci.*, *10*(5), 731–742. https://doi.org/10.5194/hess-10-731-2006

Reaney, S. M., Bracken, L. J., & Kirkby, M. J. (2007). Use of the Connectivity of Runoff Model (CRUM) to investigate the influence of storm characteristics on runoff generation and connectivity in semi-arid areas. *Hydrological Processes*, *21*(7), 894–906. https://doi.org/10.1002/hyp.6281

Redding, T. E., & Devito, K. J. (2008). Lateral flow thresholds for aspen forested hillslopes on the Western Boreal Plain, Alberta, Canada. *Hydrological Processes*, *22*(21), 4287–4300. https://doi.org/10.1002/hyp.7038

Ross, C., Ali, G., Bansah, S., & Laing, J. R. (2017). Evaluating the Relative Importance of Shallow Subsurface Flow in a Prairie Landscape. *Vadose Zone Journal*, *16*(5). https://doi.org/10.2136/vzj2016.10.0096

Ross, C. A., Ali, G., Spence, C., Oswald, C., & Casson, N. (2019). Comparison of event-specific rainfall–runoff responses and their controls in contrasting geographic areas. *Hydrological Processes*, *33*(14), 1961–1979. https://doi.org/10.1002/hyp.13460

Saffarpour, S., Western, A. W., Adams, R., & McDonnell, J. J. (2016). Multiple runoff processes and multiple thresholds control agricultural runoff generation. *Hydrology and Earth System Sciences*, 20(11), 4525–4545. https://doi.org/10.5194/hess-20-4525-2016

Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1986). Akaike information criterion statistics. *Dordrecht, The Netherlands: D. Reidel*, *81*.

Scaife, C. I., & Band, L. E. (2017). Nonstationarity in threshold response of stormflow in southern Appalachian headwater catchments. *Water Resources Research*, *53*(8), 6579–6596. https://doi.org/10.1002/2017WR020376

Shanley, J. B., & Chalmers, A. (1999). The effect of frozen soil on snowmelt runoff at Sleepers River, Vermont. *Hydrological Processes*, *13*(12-13), 1843–1857. https://doi.org/10.1002/(SICI)1099-1085(199909)13:12/13<1843::AID-HYP879>3.0.CO;2-G

Shaw, D. A., Vanderkamp, G., Conly, F. M., Pietroniro, A., & Martz, L. (2012). The Fill–Spill Hydrology of Prairie Wetland Complexes during Drought and Deluge. *Hydrological Processes*, *26*(20), 3147–3156. https://doi.org/10.1002/hyp.8390

Sidle, R. C., Tsuboyama, Y., Noguchi, S., Hosoda, I., Fujieda, M., & Shimizu, T. (2000). Stormflow generation in steep forested headwaters: A linked hydrogeomorphic paradigm.

Hydrological Processes, *14*(3), 369–385. https://doi.org/10.1002/(SICI)1099-1085(20000228)14:3<369::AID-HYP943>3.0.CO;2-P

Sivapalan, M. (2006). Pattern, Process and Function: Elements of a Unified Theory of Hydrology at the Catchment Scale. In *Encyclopedia of Hydrological Sciences*. John Wiley & Sons, Ltd. https://doi.org/10.1002/0470848944.hsa012

Sivapalan, M., Jothityangkoon, C., & Menabde, M. (2002). Linearity and nonlinearity of basin response as a function of scale: Discussion of alternative definitions. *Water Resources Research*, *38*(2). https://doi.org/10.1029/2001WR000482

Spence, C., & Woo, M. (2003). Hydrology of subarctic Canadian shield: Soil-filled valleys. *Journal of Hydrology*, *279*(1), 151–166. https://doi.org/10.1016/S0022-1694(03)00175-6

Stichling, W., & Blackwell, S. R. (1957). Drainage area as a hydrologic factor on the glaciated Canadian prairies. *International Association for Scientific Hydrology Publication*, *45*.

Tang, W., & Carey, S. K. (2017). HydRun: A MATLAB toolbox for rainfall–runoff analysis. *Hydrological Processes*, *31*(15), 2670–2682. https://doi.org/10.1002/hyp.11185

Tani, M. (1997). Runoff generation processes estimated from hydrological observations on a steep forested hillslope with a thin soil layer. *Journal of Hydrology*, *200*(1), 84–109. https://doi.org/10.1016/S0022-1694(97)00018-8

Tromp-van Meerveld, H. J., J, H., James, A. L., McDonnell, J. J., & Peters, N. E. (2008). A reference data set of hillslope rainfall-runoff response, Panola Mountain Research Watershed, United States. *Water Resources Research*, *44*(6). https://doi.org/10.1029/2007WR006299

Tromp-van Meerveld, H. J., & McDonnell, J. J. (2006a). Threshold relations in subsurface stormflow: 1. A 147-storm analysis of the Panola hillslope. *Water Resources Research*, *42*(2). https://doi.org/10.1029/2004WR003778

Tromp-van Meerveld, H. J., & McDonnell, J. J. (2006b). Threshold relations in subsurface stormflow: 2. The fill and spill hypothesis. *Water Resources Research*, *42*(2). https://doi.org/10.1029/2004WR003800

Uchida, T., Tromp-van Meerveld, I., & McDonnell, J. J. (2005). The role of lateral pipe flow in hillslope runoff response: An intercomparison of non-linear hillslope response. *Journal of Hydrology*, *311*(1), 117–133. https://doi.org/10.1016/j.jhydrol.2005.01.012

Wagener, T., Sivapalan, M., Troch, P., & Woods, R. (2007). Catchment Classification and Hydrologic Similarity. *Geography Compass*, *1*(4), 901–931. https://doi.org/10.1111/j.1749-8198.2007.00039.x

Wei, L., Qiu, Z., Zhou, G., Kinouchi, T., & Liu, Y. (2020). Stormflow threshold behaviour in a subtropical mountainous headwater catchment during forest recovery period. *Hydrological Processes*, *34*(8), 1728–1740. https://doi.org/10.1002/hyp.13658

Weiler, M. (2005). An infiltration model based on flow variability in macropores: Development, sensitivity analysis and applications. *Journal of Hydrology*, *310*(1–4), 294–315. https://doi.org/10.1016/j.jhydrol.2005.01.010

Weiler, M., McDonnell, J. J., Meerveld, I. T., & Uchida, T. (2006). Subsurface Stormflow. In *Encyclopedia of Hydrological Sciences*. John Wiley & Sons, Ltd. https://doi.org/10.1002/0470848944.hsa119

Western, A. W., & Grayson, R. B. (1998). The Tarrawarra Data Set: Soil moisture patterns, soil characteristics, and hydrological flux measurements. *Water Resources Research*, *34*(10), 2765–2768. https://doi.org/10.1029/98WR01833

Whipkey, R. Z. (1965). Subsurface Stormflow from Forested Slopes. *International Association of Scientific Hydrology. Bulletin*, *10*(2), 74–85. https://doi.org/10.1080/02626666509493392

Woods, R., Grayson, R., Western, A., Duncan, M., Wilson, D., Young, R., Ibbitt, R., Henderson, R., & McMahon, T. (2013). Experimental Design and Initial Results from the Mahurangi River Variability Experiment: MARVEX. In *Land Surface Hydrology, Meteorology, and Climate: Observations and Modeling* (pp. 201–213). American Geophysical Union. http://onlinelibrary-wiley-com.uml.idm.oclc.org/doi/10.1029/WS003p0201/summary

Yadav, M., Wagener, T., & Gupta, H. (2007). Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins. *Advances in Water Resources*, *30*(8), 1756–1774. https://doi.org/10.1016/j.advwatres.2007.01.005

Yafune, A., Narukawa, M., & Ishiguro, M. (2005). A Note on Sample Size Determination for Akaike Information Criterion (AIC) Approach to Clinical Data Analysis. *Communications in Statistics - Theory and Methods*, *34*(12), 2331–2343. https://doi.org/10.1080/03610920500257295

CHAPTER 4. RAINFALL-RUNOFF MODEL EVALUATION USING MULTIPLE HYDROLOGIC DESCRIPTORS

4.1 Introduction

Both detailed field measurements and rainfall-runoff models contribute greatly to better understanding and predicting hydrologic response at the catchment scale (Beven & Kirkby, 1979). Rainfall-runoff models are routinely calibrated to identify parameter sets that offer the best fit between simulated and observed hydrologic response according to one or more performance measures (e.g., Nash-Sutcliffe efficiency, Kling-Gupta efficiency) (Beven, 2011; Clark et al., 2011; Son & Sivapalan, 2007). The typical calibration approach uses performance measures to label model simulations as behavioural or nonbehavioural: behavioural simulations meet a specific performance measure criterion, while nonbehavioural simulations do not (Beven, 2011; Beven & Binley, 1992, 2014; Spear & Hornberger, 1980). Labelling simulations in this way is aligned with the concept of equifinality: multiple unique parameter sets may lead to simulations with a comparable model fit (Beven, 2006, 2011). However, the exact performance measure criterion used to distinguish between behavioural and nonbehavioural simulations is arbitrary (Clark et al., 2011; Vrugt et al., 2009), and behavioural simulations do not always have high fidelity (with fidelity defined as the degree to which a model simulation reproduces hydrologic processes observed in nature). It is common for simulations with good performance measure scores to poorly reproduce low flows (Staudinger et al., 2011), mischaracterize high flows, or lead to inaccurate flood frequency estimates (Mizukami et al., 2019).

While calibration may allow hydrologists to maximize model fit and identify behavioural simulations, distinguishing between behavioural simulations based on their fidelity necessitates post-calibration model evaluation (Beven, 2011; Kelleher et al., 2017). Model evaluation assesses model outputs relative to observations of the runoff response (Beven, 2011), and can include either the consideration of auxiliary data other than streamflow (e.g., Ala-aho et al.,

2017; Son & Sivapalan, 2007; Stadnyk et al., 2013), or the consideration of enhanced information extracted from streamflow data (e.g., Boyle et al., 2000; Yilmaz et al., 2008). On one hand, auxiliary data, including soil moisture, water chemistry, stable isotope composition, and groundwater levels, have been used to evaluate and distinguish behavioural simulations (Ala-aho et al., 2017; Franks et al., 1998; Grayson et al., 2002; Kelleher et al., 2017; Koch et al., 2016; Kuraś et al., 2011; Lamb et al., 1998; Seibert & McDonnell, 2002; Son & Sivapalan, 2007; Stadnyk et al., 2013; Wealands et al., 2005). Model evaluations involving auxiliary data are typically more computationally demanding, but they can help assess model fidelity by facilitating the falsification of behavioural simulations (Clark et al., 2015; Hill et al., 2016). On the other hand, modelling exercises that rely solely on streamflow data have been criticized as oversimplistic (Kirchner, 2006). However, it is important to distinguish models that use raw streamflow data from models that rely on information extracted from processed streamflow data. Indeed, continuous streamflow data may be subjected to transformation, trend analysis, event-based rainfall-runoff analysis, or frequency filtering, and the outcomes of these analyses can be used to develop hypotheses about different aspects of catchment response (Clark et al., 2011). Past studies have, notably, performed model evaluation on streamflow data that had been transformed through baseflow separation (e.g., Dunn, 1999; Gallart et al., 2007; Kroll et al., 2004), or separated into periods of recession or periods driven by precipitation (e.g., Boyle et al., 2000). Enhancing the use of streamflow data in model evaluation is, therefore, appealing since it leverages data that is readily available while going beyond a simple statistical assessment of fit between continuous observed and simulated data. The use of streamflow records can be enhanced by calculating a range of measures related to the flow duration curve, rainfall-runoff event response metrics, and hydrologic thresholds.

Firstly, the flow duration curve (FDC) describes the probability of flow exceeding a specific value and is an indicator of catchment function (Blöschl et al., 2013; Dingman, 2015; Vogel & Fennessey, 1994, 1995; Yilmaz et al., 2008). As such, the FDC has previously been used in a variety of ways for model evaluation (Blazkova & Beven, 2009; Herbst et al., 2009; Ley et al., 2016; Westerberg et al., 2011; Yilmaz et al., 2008; Yu & Yang, 2000). Like continuous flow timeseries, complete FDCs derived from observed and simulated data can be compared using a performance measure. Alternatively, model fidelity can be assessed by comparing specific points of interest on the observed and simulated FDCs, such as points associated with high flows from storm events (e.g., Westerberg et al., 2011). With that approach, measures of bias are used to compare segments of the observed and simulated FDC that are related to different, presumed dominant flow processes (Gronz, 2013; Yilmaz et al., 2008). For example, the FDC middle-segment can be related to the vertical redistribution of water within the soil profile: a low middle-segment slope indicates a response from slower and more sustained groundwater flow, while a steeper middle-segment slope indicates a flashy response delivered via overland flow due to a small soil storage capacity (Clark et al., 2011; Ley et al., 2016; Yilmaz et al., 2008). In addition to continuous streamflow records, the FDC presents process specific information that can be used to distinguish simulations based on how well they reproduce water volumes that are associated with different flow frequency classes and dominant flow generation mechanisms (Yilmaz et al., 2008).

Secondly, rainfall-runoff event response metrics relate event rainfall to the timing and magnitude of hydrologic response (Dingman, 2015; Tang & Carey, 2017). While these metrics are commonly used in process hydrology studies, as they provide first-order information on catchment function (Carey & Woo, 2001; Post & Jakeman, 1996; Yair & Raz-Yassif, 2004),

their use in modelling studies have been variable. For instance, some event-based rainfall-runoff models (i.e., models that simulate individual rainfall-runoff events) (Hossain et al., 2019; Singh, 1995) have been calibrated and evaluated by comparing observed and simulated event response metrics (e.g., peak discharge, total flow, and lag-to-peak) (Loague & Freeze, 1985), sometimes by relying on multi-objective functions based on these event response metrics (e.g., Yang et al., 2004). When it comes to the evaluation of continuous simulation rainfall-runoff models (i.e., models that simulate longer periods that include multiple events and inter-event periods) (Hossain et al., 2019; Singh, 1995), some have used small sets of representative rainfall-runoff events for model calibration (Brath et al., 2004; Tan et al., 2008). Individual storm events have also been used in Bayesian analysis of input uncertainty for continuous simulation rainfall-runoff models (Kavetski et al., 2006a, 2006b; Vrugt et al., 2009). There are multiple advantages to using event response metrics to assess the fidelity of model simulations, as these metrics can capture different aspects of hydrologic response. In particular, response metrics related to the travel time of precipitation inputs to the catchment outlet (e.g., lag-to-peak and centroid lag-to-peak) depend on rainfall timing as well as fixed (e.g., drainage area and slope) and dynamic (e.g., vegetation and catchment wetness) catchment characteristics (Dingman, 2015). The simulation of these timing metrics is often affected by model temporal resolution and assumptions about precipitation rate (e.g., constant or variable rainfall intensity), which has implications for catchment responses with variable time-to-peak or rainfall intensity that is not linearly associated with peak discharge (Wooding, 1965; Woods & Sivapalan, 1999). Using response timing metrics towards evaluating continuous rainfall-runoff models may, therefore, help identify simulations that adequately capture dominant runoff generation mechanisms and flow pathways at short timescales.

Thirdly, most hydrologic responses have a nonlinear dependence on rainfall inputs (Sivapalan et al., 2002) and numerous studies have shown threshold behaviour in rainfall-runoff relationships (Detty & McGuire, 2010; Mosley, 1979; Redding & Devito, 2008; Sidle et al., 2000; Tani, 1997; Tromp-van Meerveld & McDonnell, 2006b, 2006a; Weiler et al., 2006; Whipkey, 1965). Thresholds in rainfall-runoff relationships are defined as critical moments in time when runoff behaviour rapidly changes (Ali et al., 2013; Phillips, 2006) and as such, they are thought to be emergent catchment properties that reflect landscape heterogeneity and process complexity (Lehmann et al., 2007; McDonnell et al., 2007; Spence, 2010). It is also worth noting that in the literature, thresholds have been related to a variety of processes. In most cases, threshold behaviour in runoff response has been associated with the exceedance of a catchment storage deficit (Ali et al., 2015; Detty & McGuire, 2010; Lehmann et al., 2007; Oswald et al., 2011; Tromp-van Meerveld & McDonnell, 2006a, 2006b). However, threshold behaviour has also been associated with precipitation rate overwhelming the maximum rate of infiltration, the spatial connectedness of runoff generating areas, and the activation of different runoff generation mechanisms (Cammeraat, 2002; James & Roulet, 2009; Reaney et al., 2007; Scaife et al., 2020; Scaife & Band, 2017; Wei et al., 2020). Increasingly, thresholds in rainfall-runoff relationships are being used to characterize and compare catchment responses (Ali et al., 2015), but reproducing threshold-driven response dynamics using rainfall-runoff models is challenging (Thyer et al., 2009). While thresholds in hydrologic response related to storage capacity are often represented in rainfall-runoff models using storage buckets (Staudinger et al., 2011), other threshold types are not commonly considered, and models that can be used to examine thresholds in rainfall-runoff relationships for a range of environmental conditions are scarce (Mirus & Loague, 2013). Therefore, model evaluations that incorporate hydrologic thresholds may

facilitate the falsification of behavioural simulations based on their ability to effectively predict a variety of important catchment functions.

In light of the above-cited literature, there remain opportunities to enhance the use of streamflow data in hydrologic modelling to better characterize both model fidelity and model uncertainty. Indeed, parameter uncertainty is a primary source of uncertainty in hydrologic models that complicates hydrologic process descriptions and the assessment of relationships between parameters, processes, and catchment characteristics (Li et al., 2010; Zhang et al., 2016). While the shape of parameter distributions can affect model uncertainty (Benke et al., 2008) and lead to model fidelity tradeoffs (Hallouin et al., 2020), some studies have shown that model evaluation using auxiliary data or enhanced streamflow data can reduce the uncertainty in poorly defined parameters (Kelleher et al., 2017; Kuczera & Mroczkowski, 1998). Flow duration curves have been widely used to evaluate model simulations of continuous streamflow, while event response timing metrics have mostly been used to evaluate event-based models, and thresholds of rainfall-runoff relationships have not yet been used in model evaluation. This chapter, therefore, aims to assess how model evaluation – using measures of bias related to the flow duration curve, event response timing, and hydrologic thresholds – may help identify high-fidelity model simulations by distinguishing between behavioural parameter sets based on their ability to reproduce specific aspects of catchment response. This goal is pursued through two research questions:

(1) To what extent can behavioural simulations be labelled as “high-fidelity” based on their ability to reproduce (i) the flow duration curve, (ii) rainfall-runoff event response timing metrics, and (iii) hydrologic thresholds?

(2) Do the parameter distributions of behavioural simulations that can adequately reproduce specific aspects of catchment response differ from those of behavioural simulations that cannot?

4.2 Methods

4.2.1 Study site and data

The ~50 km^2 Mahurangi River catchment is located approximately 70 km north of Auckland, New Zealand, and comprises steep to gently rolling lowlands with land use/land cover of pasture, plantation forest (primarily *Pinus radiate*), and native forests (Woods et al., 2013). A 24.8 km^2 sub-catchment of the Mahurangi River catchment was the focus of this study (Figure 4-1. The sub-catchment of the Mahurangi River Catchment that was used in this study, including its location within New Zealand, digital elevation model, and channel network. m.a.s.l: meters above sea level.). This sub-catchment has ~290 m of relief and soils are clay loam with a maximum depth of 1 m developed on Waitemata sandstones (Woods et al., 2013).

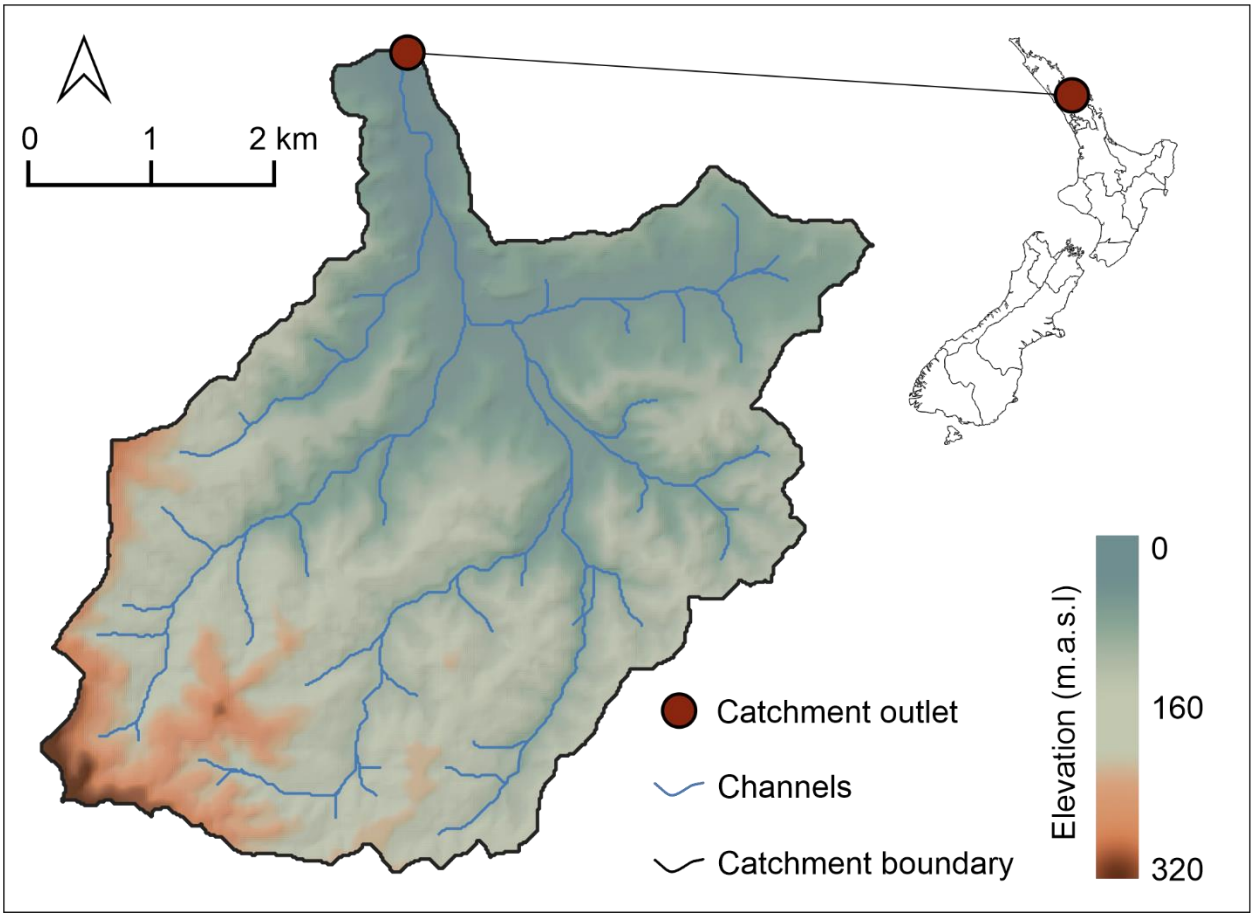


Figure 4-1. The sub-catchment of the Mahurangi River Catchment that was used in this study, including its location within New Zealand, digital elevation model, and channel network. m.a.s.l: meters above sea level.

For this study, we used high-frequency rainfall (2-minute frequency), flow (2-minute frequency), and temperature (1-hour frequency) data that were collected from July 1997 to September 2001 for the Mahurangi River Variability Experiment (Woods et al., 2013). Rainfall depths were measured using standard 200 mm collectors and 0.2 mm tipping buckets; streamflow was estimated using a site-specific weir equation from water levels that were measured using a float and counterweight with compound v-notch weir; and weather data was from a nearby (~ 10 km) meteorological station (Woods et al., 2013). Rainfall and flow data

were aggregated to a 1-hour frequency (Appendix B-1): rainfall data were aggregated using the sum, while flow data were aggregated using the arithmetic mean. In terms of regional long-term climate, the mean annual temperature is ~15.6 ºC and the mean annual precipitation and potential evapotranspiration are ~1,600 mm and ~716 mm, respectively (Ross et al., 2019; Woods et al., 2013). The mean annual potential evapotranspiration (PET) was determined by summing monthly PET estimates obtained using the Thornthwaite equation (Thornthwaite, 1948).

4.2.2 Hydrologic descriptors

This study considered different aspects of catchment response captured by three hydrologic descriptors: (1) the flow duration curve of the study period, (2) rainfall-runoff event response timing metrics, and (3) hydrologic thresholds. The FDC showing the probability, expressed as % of the time, of flow equalling or exceeding a specific value (Dingman, 2015; Vogel & Fennessey, 1994) was computed for the observed continuous flow timeseries using the hydroTSM package in R (Zambrano-Bigiarini, 2012). The FDC was separated into four distinct segments (Appendix B-2) that represent critical catchment functions (Ley et al., 2011; Ley et al., 2016; Yilmaz et al., 2008). These segments include low flows (exceedance probability > 70 %), the middle-segment (exceedance probability between 20 and 70 %), high and medium flows (exceedance probability between 2 and 20 %), and very high flows (exceedance probability < 2 %) (Gronz, 2013; Yilmaz et al., 2008).

Observed rainfall-runoff events were characterized using input meteorological factors and output hydrologic response metrics. Rainfall-runoff analysis, including the delineation of rainfall-runoff events using the MATLAB toolbox HydRun (Tang & Carey, 2017), was

previously performed on the catchment featured in this study (Ross et al., 2019). Meteorological factors and response metrics from Chapter 2 were used in the current study. Event-specific meteorological factors include the total rainfall (R_{TOT}), the average rainfall intensity (RI_{AVG}), the maximum rainfall intensity (RI_{MAX}), and the sum of event rainfall (i.e., R_{TOT}) and antecedent rainfall that occurred over a duration (X) preceding the rainfall-runoff event ($R_{TOT} + AR_x$). $R_{TOT}+AR_X$ was computed for antecedent durations (X) of 3, 7, and 14 days. Event response metrics include the total flow (Q_{TOT}), the peak flow (Q_{MAX}), the lag-to-peak (T_{LP} – the time between the beginning of rainfall and the peak event flow), and the centroid lag-to-peak (T_{LPC} – the time between the rainfall centroid and peak event flow) (Dingman, 2015). More details regarding meteorological factor and response metric calculations, statistical summaries, and examples of typical rainfall-runoff events are featured in Ross et al. (2019), where the catchment featured in the current study is referred to as MRC8.

Event response timing metrics, specifically T_{LP} and T_{LPC}, were used in the current study to represent hydrologic response at the event scale. These metrics characterize the timing of peak event flow relative to key parts of the event hyetograph (i.e., the beginning of rainfall and the hyetograph centroid). Summary statistics of T_{LP} and T_{LPC} for the seventy-four rainfall-runoff events are shown in Appendix B-3. Select event meteorological factors (R_{TOT}, RI_{AVG}, RI_{MAX}, $AR_3+ R_{TOT}$, $AR_7+ R_{TOT,}$ and $AR_{14}+ R_{TOT}$) and response metrics (Q_{TOT} and Q_{MAX}) were used to assess rainfall-runoff relationships for threshold behaviour. Twelve scatterplots showing response magnitude metrics against meteorological factors were constructed and evaluated for thresholds using piecewise linear regression analysis (PRA) (Oswald et al., 2011; Scaife & Band, 2017) via the segmented package in R (Muggeo, 2008). These twelve input-output scatter plots were intended to represent three types of potential threshold behaviour: thresholds related to

rainfall intensity, thresholds related to event rainfall depth, and thresholds related to event plus antecedent rainfall depth. When possible, PRA identifies a breakpoint in an input-output pair and fits two linear segments to the event data: these two segments correspond to data on each side of the breakpoint. Pairs for which no breakpoint was identified were assumed to not have a threshold. However, if a breakpoint was identified, the goodness-of-fit (R^2) of the piecewise linear model was estimated, along with the Akaike Information Criterion (AIC) of the piecewise linear model (Sakamoto et al., 1986; Yafune et al., 2005) and the breakpoint value. The slopes of the linear segments preceding (m_1) and following (m_2) the breakpoint were also calculated. Breakpoints identified during PRA were only considered thresholds if (1) the R^2 of the piecewise linear model was moderate to strong ($R^2 > 0.45$), (2) the AIC of the piecewise linear regression model was more than two units below the AIC of the simple linear regression model, (3) the percent difference between m_1 and m_2 exceeded 10%, and (4) m_2 was greater than zero. These criteria were imposed to ensure that only input-output pairs characterized by rapid changes in runoff behaviour at critical input values were considered threshold mediated (Ali et al., 2013; Detty & McGuire, 2010; Phillips, 2006). Five of the twelve input-output pairs that were tested met these criteria, indicating threshold behaviour. Thresholds identified for observed input-output pairs are shown in Appendix B-4.

4.2.3 Model selection and calibration

The GR5H model, which is a lumped, hourly, bucket-type, continuous rainfall-runoff model (Ficchì, 2017; Ficchì et al., 2019; Perrin et al., 2003) was selected for this study. There are two major components of the GR5H model: (1) a production module that models the water

balance, comprising an interception function, a soil moisture accounting store, and a groundwater exchange function; and (2) a flow routing module comprising unit hydrographs and a nonlinear store (Ficchì, 2017; Ficchì et al., 2019; Le Moine, 2008; Perrin et al., 2003). GR5H is a parsimonious model with only five parameters: the maximum capacity of the production store (X_1 – mm), the groundwater exchange coefficient (X_2 – mm/hr), the reference capacity of the routing store one time-step ahead (X_3 – mm), the time base of the unit hydrograph (X_4 – hours), and the level of the routing store at which the flux exchange changes sign (X_5 – [-]). The GR5H model was developed from the GR4J model, using an empirical comparative approach carried out on over a thousand catchments in France and Australia (Ficchì et al., 2019; Le Moine, 2008; Perrin et al., 2003). The main differences between GR5H and its predecessor, the four-parameter GR4J model, are that GR5H operates at the hourly time-step rather than the daily time-step, and GR5H includes a more complex groundwater exchange function associated with the additional fifth free parameter (Ficchì et al., 2019; Le Moine, 2008). The groundwater exchange function is an explicit, albeit conceptual, representation of groundwater/surface water interactions and contributes to improved low flow simulations (Ficchì et al., 2019; Le Moine, 2008). GR5H does not assume that deeply infiltrating water that bypasses the topographic catchment is negligible, which is an assumption of many rainfall-runoff models (Le Moine, 2008). Perrin et al. (2003) and Ficchì et al. (2019) offer detailed descriptions of the GR model structure and the associated model equations. While the GR models were initially developed based on catchments in France, they have been applied in a wide variety of climate and flow conditions, and changes in the GR5H version are associated with improved flux coherence and simulation realism for most catchments (Coron et al., 2017; Ficchì, 2017; Ficchì et al., 2019; Le Moine, 2008).

In the current study, the GR5H model was implemented in R using the airGR package (Coron et al., 2017; Coron et al., 2020) and the aforementioned temperature, rainfall, and flow data. Data were separated into two subsets: six months, starting in July 1997, were used for model warm-up while the remaining data were used for model calibration. In total, 500,000 simulations were performed using parameter values that were randomly sampled from the feasible parameter space using Monte Carlo methods. The feasible parameter space was derived from established GR5H calibration routines in the airGR package and previous GR model implementations (Coron et al., 2020; Ficchì, 2017; Ficchì et al., 2019; Le Moine, 2008; Perrin et al., 2003). For model calibration, to evaluate the agreement between the observed and predicted continuous streamflow timeseries of each Monte Carlo simulation (n = 500,000), the Kling-Gupta Efficiency (KGE) was calculated:

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \qquad \text{Equation 4-1}$$

where α is the relative variability ($\sigma_{sim}/\sigma_{obs}$), β is the bias (μ_{sim}/μ_{obs}), r is the linear correlation coefficient ($Cov/\sigma_{sim}\sigma_{obs}$), with Cov being the covariance between simulated and observed values. The KGE values were used to label model simulations as behavioural or nonbehavioural (Beven, 2011; Beven & Binley, 2014; Freer et al., 1996; Spear & Hornberger, 1980). In this study, simulations with KGE $\geq$ 0.3 were labelled behavioural and simulations with KGE < 0.3 were labelled nonbehavioural. After calibration, behavioural model simulations were evaluated to assess how different hydrologic response descriptors (listed in Section 4.2.2) may help distinguish behavioural, low-fidelity simulations from behavioural, high-fidelity simulations.

4.2.4 Post-calibration model evaluation

Behavioural simulations were considered to exhibit different degrees of fidelity based on their ability to reproduce different aspects of the observed hydrologic response captured by three descriptors: the FDC, rainfall-runoff event response timing metrics, and thresholds in rainfall-runoff relationships. The FDC of the simulated streamflow timeseries associated with each behavioural simulation was computed in the same way as the observed FDC. To evaluate model fidelity for the FDC, the percent bias (Pbias) between four segments of the observed and simulated FDCs were calculated (Gronz, 2013; Ley et al., 2011; Ley et al., 2016; Yilmaz et al., 2008). The FLV is the Pbias between observed and simulated flow volumes of the low flow exceedance probability segment:

$$\mathrm{FLV} = \frac{\int_{0.7}^{1}\left(\log(\mathrm{QS}_{\mathrm{p}}^{\mathrm{sim}})-\log(\mathrm{QS}^{\mathrm{sim}})\right)\mathrm{dp}-\int_{0.7}^{1}\left(\log(\mathrm{QS}_{\mathrm{p}}^{\mathrm{obs}})-\log(\mathrm{QS}^{\mathrm{min}})\right)\mathrm{dp}}{\int_{0.7}^{1}\left(\log(\mathrm{QS}_{\mathrm{p}}^{\mathrm{obs}})-\log(\mathrm{QS}^{\mathrm{sim}})\right)\mathrm{dp}} * 100 \qquad \text{Equation 4-2}$$

where $\mathrm{QS}_{\mathrm{p}}^{\mathrm{sim}}$ (mm/hr) is the predicted flow at exceedance probability P, $\mathrm{QS}_{\mathrm{p}}^{\mathrm{obs}}$ (mm/hr) is the observed flow at exceedance probability P and $\mathrm{QS}^{\mathrm{min}}$ is the minimum of $\mathrm{QS}^{\mathrm{sim}}$ and $\mathrm{QS}^{\mathrm{obs}}$ (Gronz, 2013; Ley et al., 2016; Yilmaz et al., 2008). The FMS is calculated as follows:

$$\mathrm{FMS} = \frac{\left(\log(\mathrm{QS}_{0.2}^{\mathrm{sim}})-\log(\mathrm{QS}_{0.7}^{\mathrm{sim}})\right)-\left(\log(\mathrm{QS}_{0.2}^{\mathrm{obs}})-\log(\mathrm{QS}_{0.7}^{\mathrm{min}})\right)}{\left(\log(\mathrm{QS}_{0.2}^{\mathrm{obs}})-\log(\mathrm{QS}_{0.7}^{\mathrm{sim}})\right)} * 100 \qquad \text{Equation 4-3}$$

and is the Pbias between the slopes of the observed and simulated FDC middle segment. The FMV and FHV are the Pbias between observed and simulated flow volumes of the medium and high flow exceedance probability segment, and the very high flow exceedance probability segment of the FDC:

$$\mathrm{FMV} = \frac{\int_{0.02}^{0.2} \mathrm{QS_p^{sim}}\,\mathrm{dp} - \int_{0.02}^{0.2} \mathrm{QS_p^{obs}}\,\mathrm{dp}}{\int_{0.02}^{0.2} \mathrm{QS_p^{obs}}\,\mathrm{dp}} * 100 \qquad \text{Equation 4-4}$$

and

$$\mathrm{FHV} = \frac{\int_{0}^{0.02} \mathrm{QS_p^{sim}}\,\mathrm{dp} - \int_{0}^{0.02} \mathrm{QS_p^{obs}}\,\mathrm{dp}}{\int_{0}^{0.02} \mathrm{QS_p^{obs}}\,\mathrm{dp}} * 100. \qquad \text{Equation 4-5}$$

These measures of bias have been described and used elsewhere (Casper et al., 2012; Gronz, 2013; Herbst et al., 2009; Ley et al., 2011; Ley et al., 2016; Yilmaz et al., 2008).

To evaluate model fidelity for event response timing, the T_{LP} and T_{LPC} of each simulated rainfall-runoff event were estimated for each behavioural simulation. The start of rainfall and end of event hydrologic response for rainfall-runoff events delineated from observed data were used to isolate event periods (n = 74) in the simulated streamflow timeseries. The Pbias between event T_{LP} and T_{LPC} of the observed and simulated data were calculated to evaluate how well simulations reproduced rainfall-runoff event response timing.

To evaluate model fidelity with respect to thresholds in rainfall-runoff relationships, the Q_{TOT} and Q_{MAX} of each rainfall-runoff event were computed for each behavioural simulation. The simulated event Q_{TOT} and Q_{MAX} and meteorological factors from rainfall-runoff events were used to construct twelve input-output scatterplots for each behavioural simulation. For each behavioural simulation, input-output pairs were evaluated for threshold behaviour using the procedure described in Section 4.2.2. If a threshold was confirmed in a simulated input-output pair, a compound Pbias value was computed between the observed and simulated threshold value, observed and simulated segment slopes (m_1 and m_2), and observed and simulated R^2 and AIC (associated with the piecewise linear regression model) according to Equation 4-6:

$$\text{Compound threshold bias} = \left[\frac{\begin{array}{c}(\text{sim}_{\text{th}} - \text{obs}_{\text{th}}) + (\text{sim}_{\text{m}_1} - \text{obs}_{\text{m}_1}) + \\ (\text{sim}_{\text{m}_2} - \text{obs}_{\text{m}_2}) + (\text{sim}_{\text{R}^2} - \text{obs}_{\text{R}^2}) + \\ (\text{sim}_{\text{AIC}} - \text{obs}_{\text{AIC}})\end{array}}{(\text{obs}_{\text{th}} + \text{obs}_{\text{m}_1} + \text{obs}_{\text{m}_2} + \text{obs}_{\text{R}^2} + \text{obs}_{\text{AIC}})} \right] * 100 \quad \text{Equation 4-6}$$

where sim_{th} and obs_{th} are the simulated and observed threshold values, sim_{m_1} and obs_{m_1} are the simulated and observed slopes preceding the threshold, sim_{m_2} and obs_{m_2} are the simulated and observed slopes following the threshold, sim_{R^2} and obs_{R^2} are the simulated and observed segmented linear model fit, and sim_{AIC} and obs_{AIC} are the simulated and observed segmented linear model AIC. Some of the observed input-output pairs (five of twelve) selected for this study had threshold behaviour, while others did not (seven of twelve) (Appendix B-4). This allowed behavioural simulations to be evaluated based on their association with true-positive, false-positive, true-negative, or false-negative threshold identification.

Overall, model fidelity was assessed using eleven measures of bias, including FDC biases (4 measures), rainfall-runoff event response timing biases (2 measures), and threshold biases (5 measures). The optimal bias values are 0, and values closest to 0 imply a better reproduction of the related process. Positive and negative values for biases indicate model overestimation and model underestimation, respectively. Throughout the rest of this paper, biases are expressed as absolute values, and negative and positive biases are described as underestimation and overestimation, respectively. Assessments of fidelity for behavioural simulations were made for single descriptors and multiple descriptors. Single-descriptor approaches were used to independently assess the ability of a simulation to minimize FDC biases, timing biases, or threshold biases. Multi-descriptor approaches rather assessed the ability of a simulation to minimize biases across all descriptors. In the multi-descriptor assessment, a simulation could potentially minimize no biases, biases of a single descriptor, biases of two descriptors, or biases

of all three descriptors. Behavioural simulations with Pbias ≤ 15% for the aforementioned biases were deemed to have low descriptor-specific or low multi-descriptor bias, and therefore to exhibit descriptor-specific or multi-descriptor fidelity.

4.2.5 Assessing parameter distributions

For each model parameter, a two-sample Kolmogorov-Smirnov test was performed to compare the parameter distributions of behavioural simulations that met the 15% Pbias criterion for different measures of bias against the parameter distributions of the remaining behavioural simulations. Using the FLV measure of bias and the X_1 parameter as an example: in this test, one sample includes the X_1 parameter values of all behavioural simulations that met the 15% criterion for the FLV, while the second sample includes the X_1 parameter values of all behavioural simulations that did not meet the 15% criterion for the FLV. The null hypothesis was that the X_1 parameter values of these two samples were from the same distribution, while the alternative hypothesis was that the X_1 parameter values of these two samples were not from the same distribution (Conover, 1998). In this study, the 5% significance level was used to reject the null hypothesis. Similar two-sample Kolmogorov-Smirnov tests were performed to compare the parameter distributions of behavioural simulations that met the 15% Pbias criterion for different groups of bias measures related to a single descriptor (e.g., all bias measures related to the FDC) and the remaining behavioural simulations.

4.3 Results

4.3.1 Behavioural simulations

Of the 500,000 simulations that were performed in this study, 8339 simulations were labelled behavioural (i.e., KGE $\geq$ 0.3). The KGE scores of behavioural simulations ranged from 0.30 to 0.92: over 6000 simulations had KGE scores between 0.30 and 0.50, while only five simulations had a KGE score greater than 0.90 (Figure 4-2). It should be noted that simulations shown in Figure 4-2 include flow predictions for isolated periods of missing flow observations, as these gaps were not necessarily mirrored in the input rainfall data. However, predictions for missing flow observations were not included in KGE score calculations. The visual agreement between observed and simulated streamflow for behavioural simulations improved with increasing KGE scores (Figure 4-2). However, regardless of KGE scores, several simulations significantly overestimated low flows (Figure 4-2). Also, some behavioural simulations erroneously predicted the location and/or magnitude of flow peaks.

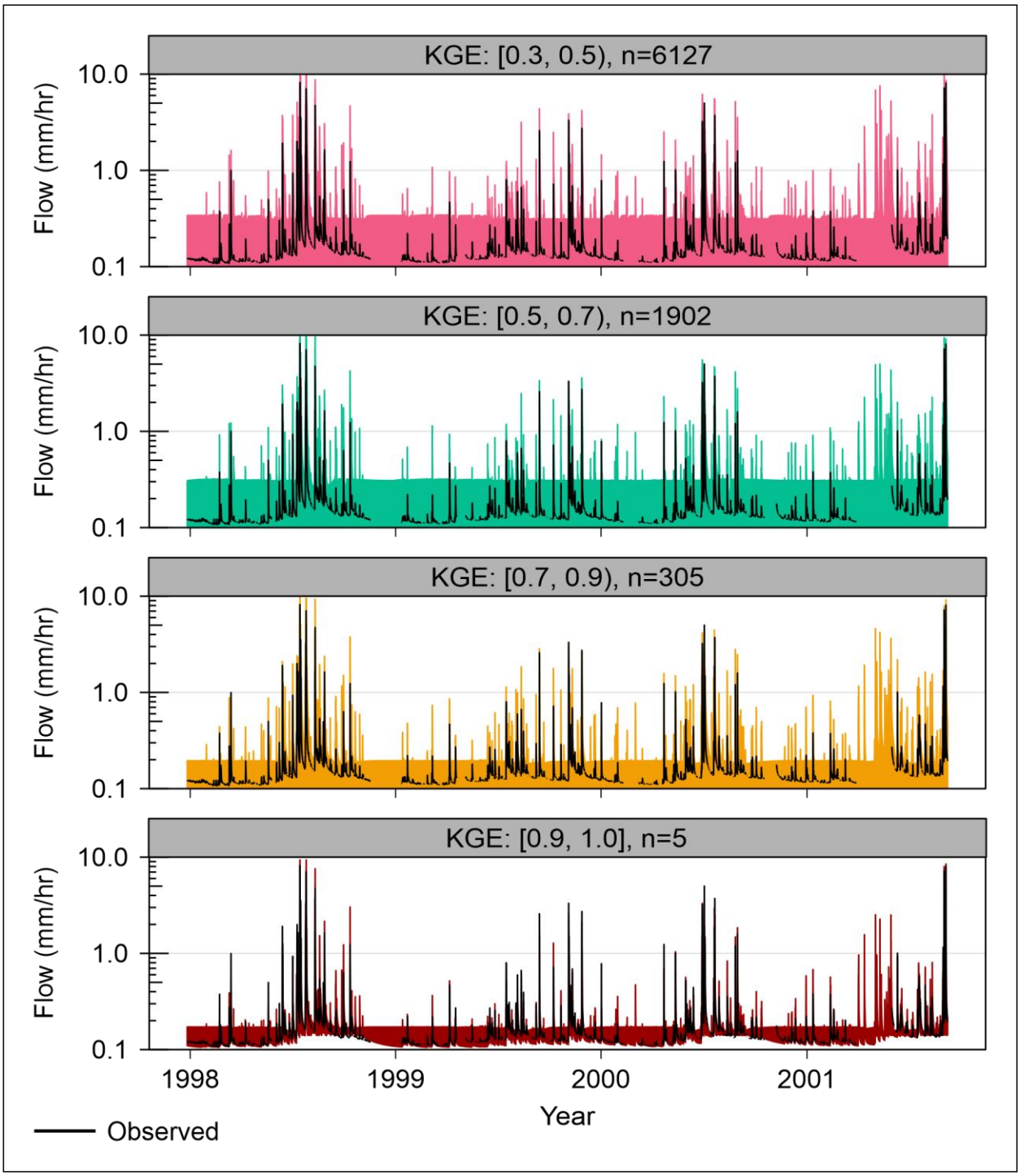


Figure 4-2. Observed flow timeseries and flow ranges of simulated flow timeseries associated with behavioural parameter sets. Simulation flow ranges are colour-coded by KGE score range.

4.3.2 Single-descriptor model evaluation

First, model evaluation was performed based on FDC biases only. Among the four FDC biases, the 15% Pbias criterion was met for the FLV by the fewest behavioural simulations, followed by the FMS, FMV, and FHV (Table 4-1 and Figure 4-3). More than three times as many behavioural simulations adequately reproduced FDC segments represented by the FHV and FMV than the FMS and FLV. In detail, 3054, 2137, 978, and 678 behavioural simulations had FHV, FMV, FMS, and FLV ≤ 15%, respectively. Only 94 behavioural simulations had all four FDC biases ≤ 15%. More than half (51.5%) of the behavioural simulations that met the 15% criterion for the FLV mostly underestimated the flow volume. Similarly, most of the behavioural simulations that met the 15% criterion for the FMS (52.5%) and FHV (53.5%) underestimated the middle-segment slope and the flow volume of the high flow exceedance probability segment, respectively. In contrast, 51.6% of the behavioural simulations with FMV ≤ 15% overestimated the associated flow volume.

Table 4-1. The number (percentage shown in brackets) of behavioural simulations that met the 15% Pbias criterion for different measures of bias.

	Pbias ≤ 15%
	Simulations (%)
Flow duration curve	
FLV	678 (8.1)
FMS	978 (11.7)
FMV	2137 (25.6)
FHV	3054 (36.6)
All FDC biases	94 (1.1)
Event-specific response timing metrics	
T_{LP}	3714 (44.5)
T_{LPC}	968 (11.6)
All response timing biases	968 (11.6)
Event rainfall threshold relationship	
R_{TOT}, Q_{TOT}	5032 (60.3)
Antecedent plus event rainfall threshold relationships	
AR_3+R_{TOT}, Q_{TOT}	6307 (75.6)
AR_7+R_{TOT}, Q_{TOT}	7633 (91.5)
AR_7+R_{TOT}, Q_{MAX}	5619 (67.4)
AR_{14}+R_{TOT}, Q_{TOT}	8077 (96.9)
All AR_X+R_{TOT} threshold biases	4081 (48.9)
All threshold biases	2272 (27.3)

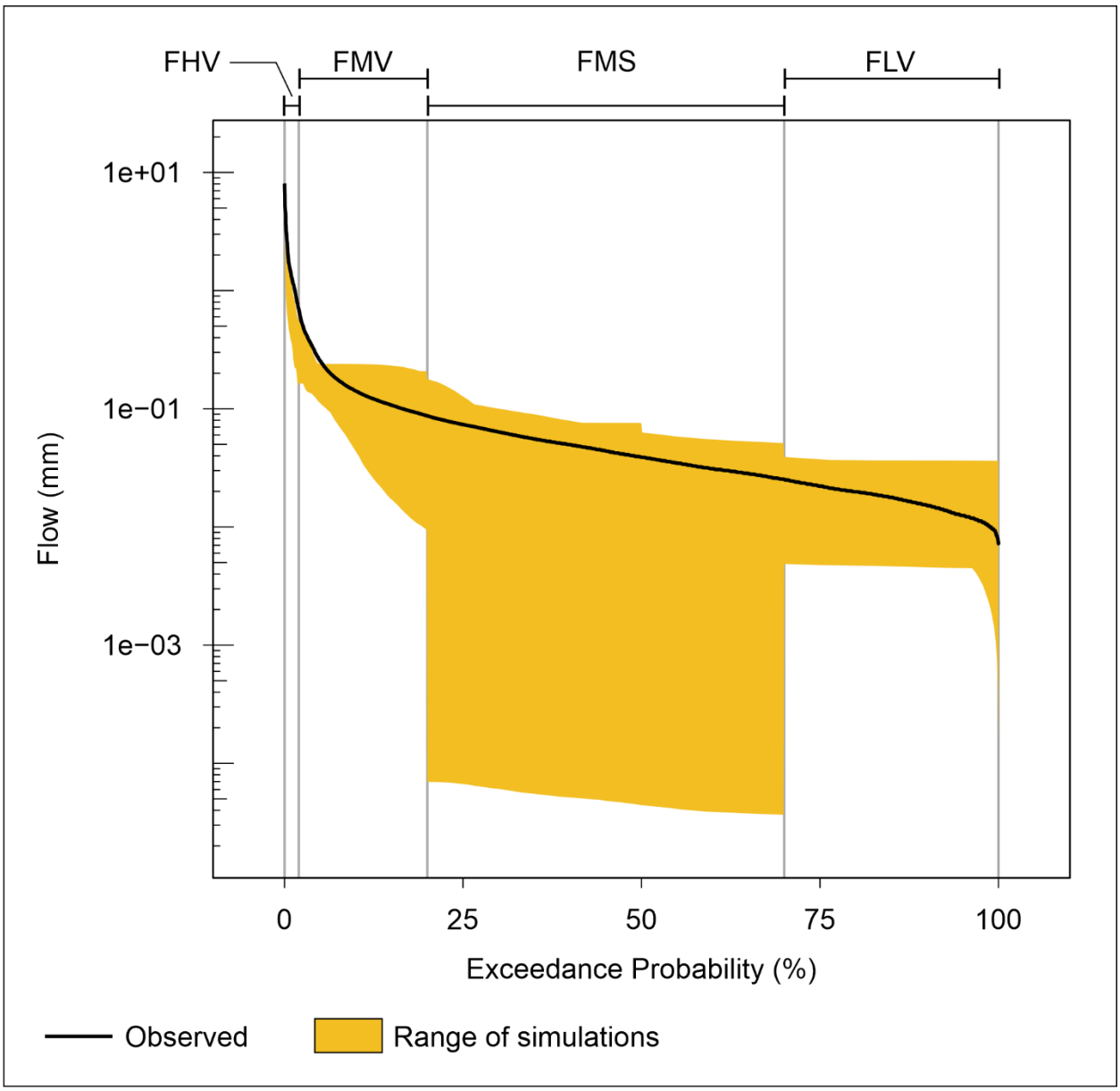


Figure 4-3. Observed flow duration curve and range of simulated flow duration curves associated with behavioural simulations with FLV, FMS, FMV, or FHV ≤ 15% (shown in yellow).

Second, model evaluation was performed based on response timing biases only. There were 3714 and 968 behavioural simulations that reproduced event T_{LP} (Figure 4-4A) and T_{LPC} (Figure 4-4B) with Pbias ≤ 15%, respectively (Table 4-1 and Figure 4-4). All the behavioural

simulations that reproduced T_{LPC} with Pbias ≤ 15% also reproduced T_{LP} with Pbias ≤ 15%. The event T_{LP} and T_{LPC} were underestimated, overestimated, and predicted exactly by 48%, 44%, and 8% of behavioural simulations that reproduced both T_{LPC} and T_{LP} with Pbias ≤ 15%.

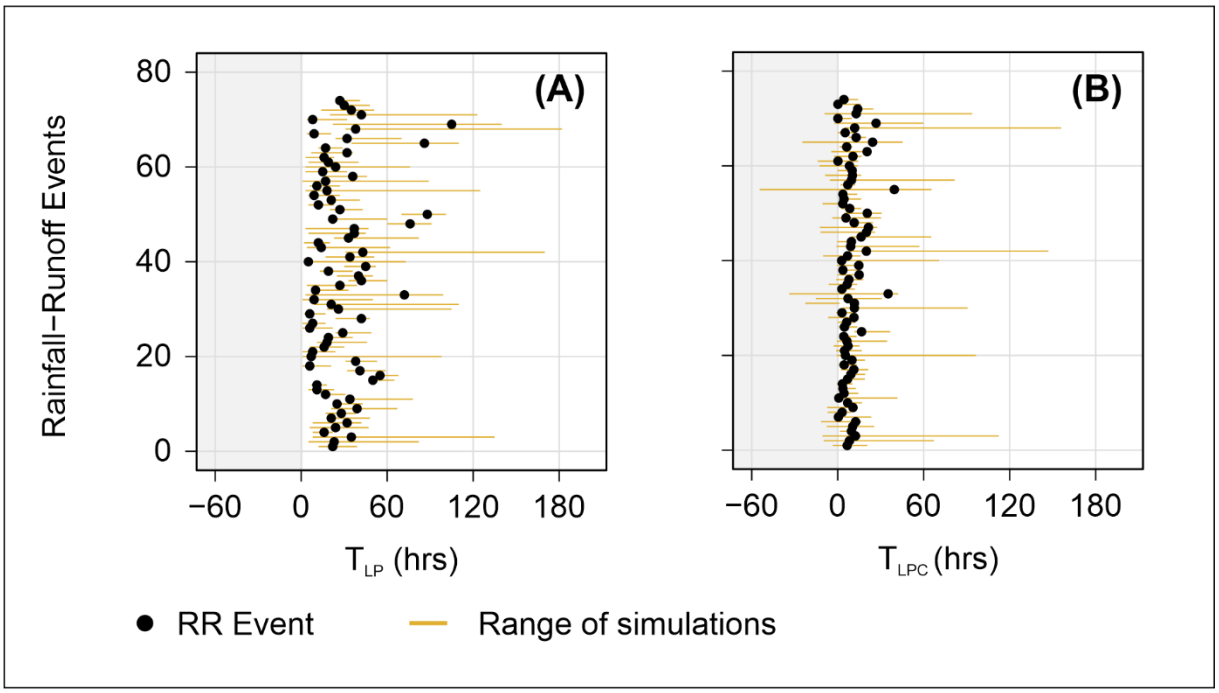


Figure 4-4. T_{LP} and T_{LPC} of rainfall-runoff (RR) events (n = 74) in the observation data, and bars showing the T_{LP} (A) and T_{LPC} (B) ranges across all events for behavioural simulations with Pbias ≤ 15%. Plot areas associated with negative response timing values are shaded grey.

Third, model evaluation was performed based on threshold biases only. There was variability in threshold identification between behavioural simulations. Behavioural simulations could be associated with true-positives (i.e., a threshold detected from observed and simulated data), false-positives (i.e., a threshold not detected from observed data but detected from simulated data), false-negatives (i.e., a threshold detected from observed data but not detected from simulated data) and true-negatives (i.e., a threshold not detected from observed and

simulated data) (Table 4-2). Most behavioural simulations were associated with true-positives or true-negatives. However, there were some false-negatives: the observed thresholds of R_{TOT} - Q_{TOT}, $AR_7+R_{TOT} - Q_{TOT}$, and $AR_7+R_{TOT} - Q_{MAX}$ pairs were not detected for 240, 490, and 1228 behavioural simulations, respectively. Similarly, there were some false-positives: thresholds of $R_{TOT} - Q_{MAX}$, $RI_{AVG} - Q_{MAX}$, $RI_{MAX} - Q_{MAX}$, $AR_3+R_{TOT} - Q_{MAX}$, and $AR_{14}+R_{TOT} - Q_{MAX}$ pairs were erroneously detected for 6631, 64, 2275, 8252, and 7817 behavioural simulations, respectively.

Table 4-2. Number (percentages, shown in brackets) of behavioural simulations for which thresholds were identified (or not) for the twelve input-output pairs evaluated in this study. "NA": Options that are inapplicable given the presence or absence of threshold behaviour in the observed data.

	True-positive	False-negative	False-positive	True-negative
R_{TOT}, Q_{TOT}	8099 (97.1)	240 (2.9)	NA	NA
R_{TOT}, Q_{MAX}	NA	NA	6631 (79.5)	1708 (20.5)
RI_{AVG}, Q_{TOT}	NA	NA	0 (0.0)	8339 (100.0)
RI_{AVG}, Q_{MAX}	NA	NA	64 (0.8)	8275 (99.2)
RI_{MAX}, Q_{TOT}	NA	NA	0 (0.0)	8339 (100.0)
RI_{MAX}, Q_{MAX}	NA	NA	2275 (27.3)	6064 (72.7)
AR_3+R_{TOT}, Q_{TOT}	8338 (100.0)	1 (0.0)	NA	NA
AR_3+R_{TOT}, Q_{MAX}	NA	NA	8252 (99.0)	87 (1.0)
AR_7+R_{TOT}, Q_{TOT}	7849 (94.1)	490 (5.9)	NA	NA
AR_7+R_{TOT}, Q_{MAX}	7111 (85.3)	1228 (14.7)	NA	NA
$AR_{14}+R_{TOT}$, Q_{TOT}	8333 (99.9)	6 (0.1)	NA	NA
$AR_{14}+R_{TOT}$, Q_{MAX}	NA	NA	7817 (93.7)	522 (6.3)

The number of behavioural simulations that adequately reproduced thresholds of the observed data (Appendix B-4) varied depending on the input-output pair considered (Table 4-1 and Figure 4-5), and 2272 behavioural simulations had all five threshold biases with Pbias ≤ 15%. A slightly larger number of behavioural simulations had Pbias ≤ 15% for thresholds that considered antecedent rainfall (i.e., pairs involving AR_X+R_{TOT}). Of the five observed threshold relationships, the $AR_{14}+R_{TOT}$ - Q_{TOT} pair was adequately reproduced by the largest number of behavioural simulations. In total, 4081 behavioural simulations were able to adequately reproduce the thresholds associated with all four pairs involving AR_X+R_{TOT}. The range of variability of the segmented linear fit for most simulated input-output pairs was narrower before the threshold had been exceeded, compared to after it had been exceeded (Figure 4-5).

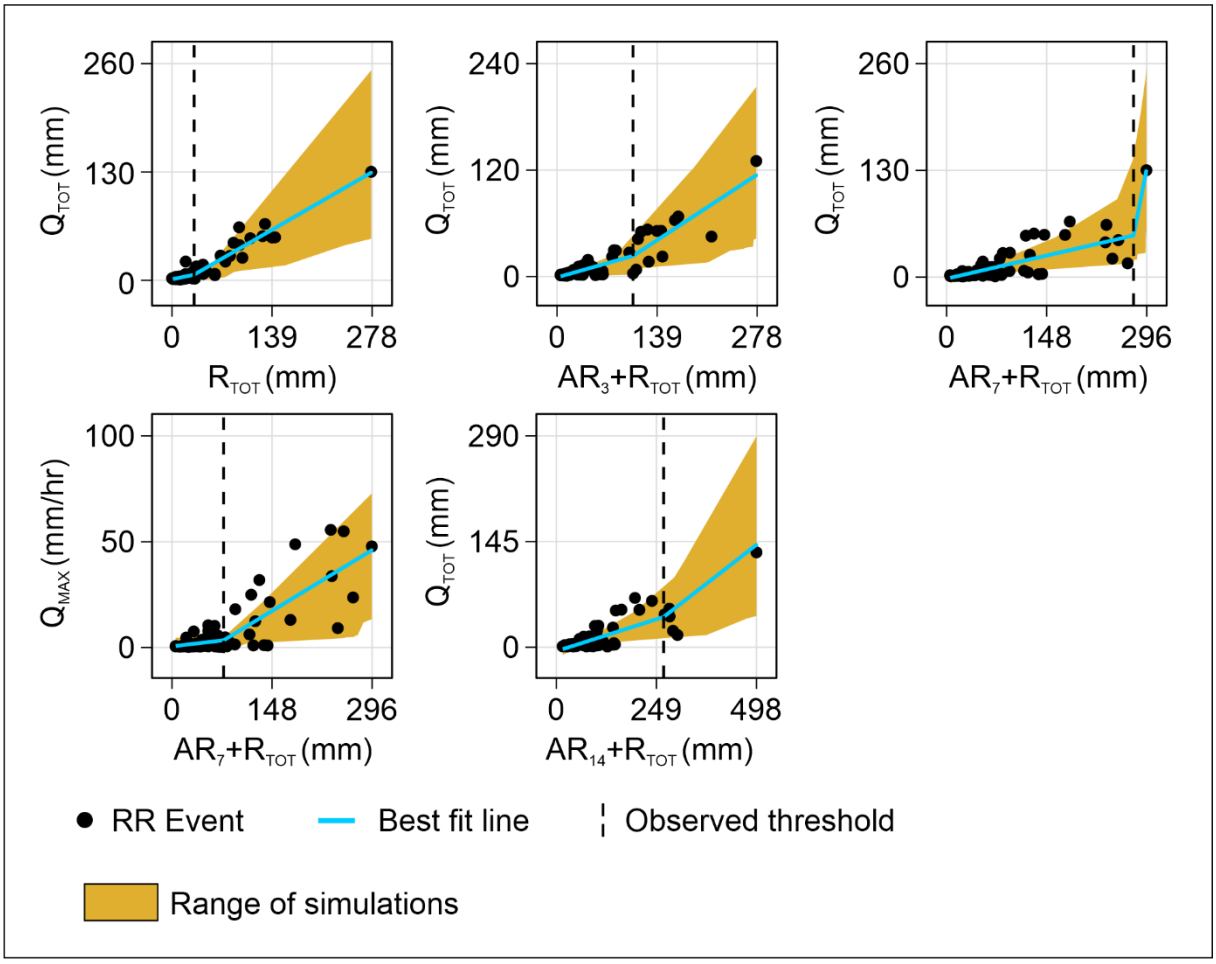


Figure 4-5. Thresholds associated with input-output pairs. In each panel, black dots represent individual observed rainfall-runoff (RR) events. The teal line indicates the piecewise linear model derived from the observed data, and the dashed black line indicates the observed threshold value. The range of piecewise linear models for behavioural simulations with compound Pbias ≤ 15% for each input-output pair is shown in yellow.

The KGE scores of behavioural simulations that met the 15% Pbias criterion for different measures of bias were similar (Figure 4-6). Compared to the median KGE score of 0.41 for all behavioural simulations, behavioural simulations that had low FDC biases (i.e., Pbias ≤ 15%) had a median KGE score of ~0.42, regardless of which FDC segment was considered. The

median KGE score of behavioural simulations that had low timing metric biases was slightly higher: 0.44 for T_{LP} and 0.45 for T_{LPC}. For behavioural simulations that had low threshold biases, the median KGE score ranged from 0.39 (AR_7+R_{TOT} - Q_{MAX} pair) to 0.42 (R_{TOT} - Q_{TOT} pair). The 75th percentile and maximum KGE score of low-bias simulations were variable. For behavioural simulations that had low FDC biases, the 75th percentile KGE score ranged from 0.51 (FMS) to 0.54 (FHV), and the maximum ranged from 0.81 (FLV) to 0.92 (FMV). For behavioural simulations with low timing biases, the 75th percentile KGE score ranged from 0.54 (T_{LPC}) to 0.55 (T_{LP}), and the maximum from 0.85 (T_{LPC}) to 0.92 (T_{LP}). Lastly, for threshold biases, the 75th percentile KGE score ranged from 0.47 (AR_7+R_{TOT} - Q_{MAX} pair) to 0.52 (R_{TOT} - Q_{TOT} pair), and the maximum ranged from 0.91 (AR_7+R_{TOT} - Q_{MAX} pair) to 0.92 (other observed thresholds). These compared to the 75th percentile and maximum KGE score of 0.51 and 0.92 for all behavioural simulations.

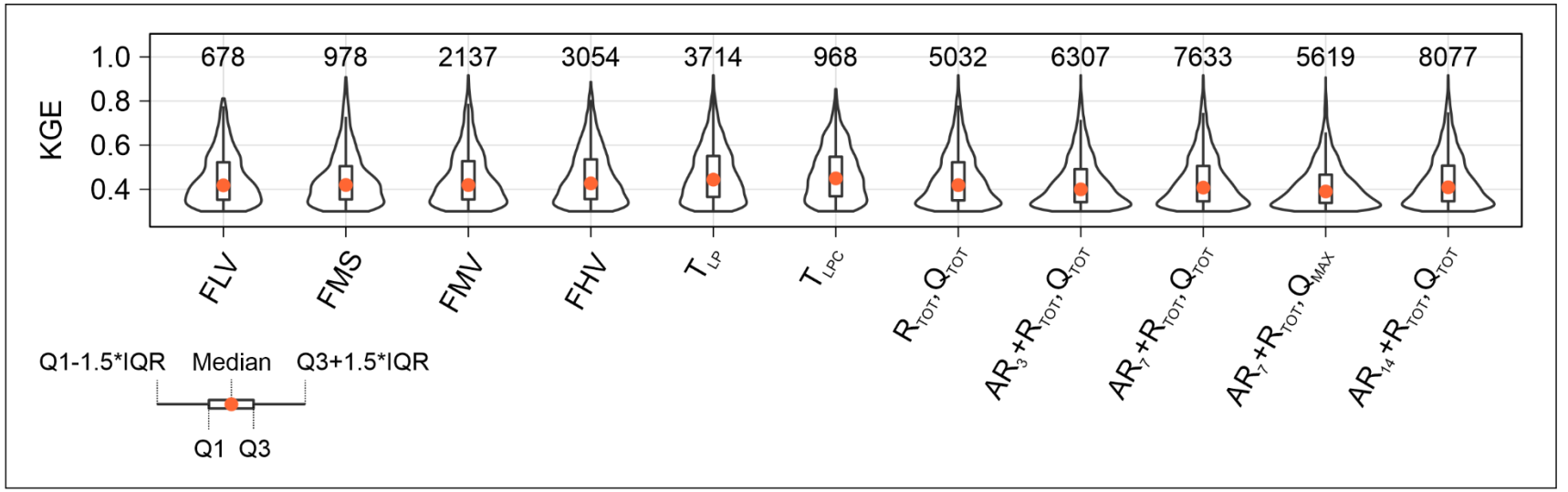


Figure 4-6. Violin plots showing the distribution of KGE scores for behavioural simulations that met the 15% Pbias criterion for different measures of bias, with the x-axis showing each measure of bias. The number of simulations is indicated above each violin plot.

4.3.3 Multi-descriptor model evaluation

Subsets of behavioural simulations were formed comprising simulations with [I] low FDC biases only, [II] low timing biases only, [III] low threshold biases, or low biases related to combinations of hydrologic descriptors (i.e., I and II, I and III, II and III, or I, II, and III). These subsets were then compared by KGE score (Figure 4-7). Of the behavioural simulations with a KGE score between 0.3 and 0.5, 31.0% had low FDC and threshold biases, 30.2% had low threshold biases only, 20.3% had low timing and threshold biases, and 18.6% had low FDC, timing, and threshold biases. The proportion of simulations with KGE scores between 0.3 and 0.5 that were a part of other behavioural simulation subsets (i.e., I, II, and I and II) was negligible (i.e., < 1%). Of the behavioural simulations with a KGE score between 0.5 and 0.7, a large proportion (37.0%) had low FDC, timing, and threshold biases, while 23.1% of these behavioural simulations had low FDC and threshold biases, 22.7% had low timing and threshold biases, and 17.3% had low threshold biases only. Again, the proportion of behavioural simulations with KGE scores between 0.5 and 0.7 that were part of other subsets (i.e., I, II, and I and II) was negligible. For behavioural simulations with a KGE score between 0.7 and 0.9, none had low FDC biases only, or low timing biases only. 47.5% of these behavioural simulations had low FDC, timing, and threshold biases, 20.3% had low FDC and threshold biases, 16.1% had low timing and threshold biases, and 16.1% had low threshold biases only. Finally, 60.0% of the behavioural simulations with a KGE score of 0.9 or greater had low timing and threshold biases, while the remaining 40.0% had low FDC, timing, and threshold biases.

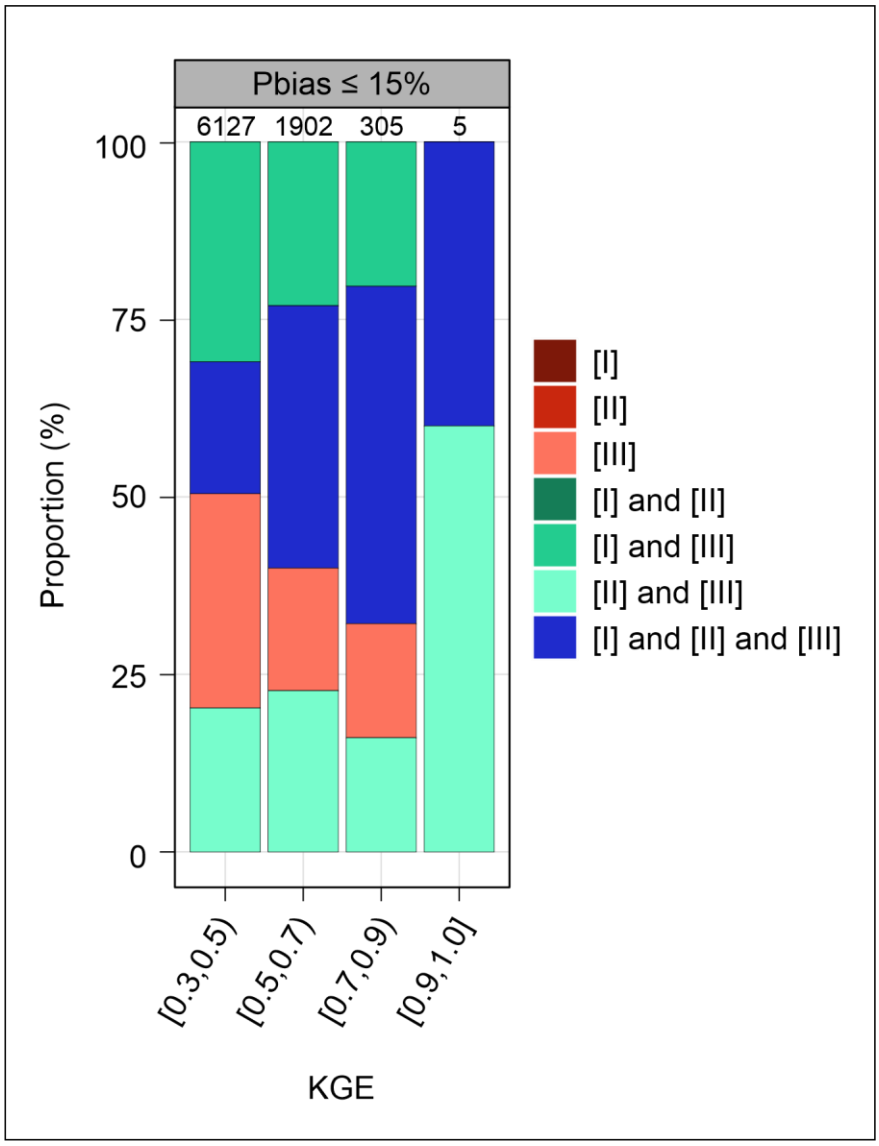


Figure 4-7. Bar charts showing the proportion of simulations at different KGE score ranges with [I] low FDC biases only, [II] low timing biases only, [III] low threshold biases only, or low biases related to combinations of these descriptors. The number of behavioural simulations for each KGE score range is shown above each bar.

The minimum, median, and maximum KGE scores of behavioural simulation subsets (see Section 4.4.3) were also identified (Table 4-3). The behavioural simulations with the largest minimum KGE scores had low FDC, timing, and threshold biases (minimum KGE = 0.39). The

median KGE scores of behavioural simulation subsets ranged from 0.37 to 0.51, and simulations with the largest median KGE scores had low FDC, timing, and threshold biases (median KGE = 0.51). The maximum KGE scores of behavioural simulation subsets ranged from 0.45 to 0.92, and simulations with the largest maximum KGE scores had low threshold biases only (maximum KGE = 0.87), low FDC and threshold biases (maximum KGE = 0.89), low timing and threshold biases (maximum KGE = 0.91), or low FDC, timing, and threshold biases (maximum KGE = 0.92).

The observed flow timeseries and the modelled flow timeseries associated with the behavioural simulations that had the maximum KGE scores from each behavioural simulation subset (see Table 4-3) are shown in Figure 4-8. Of these behavioural simulations, those with low threshold biases only (Figure 4-8A – KGE = 0.87), or low FDC, timing, and threshold biases (Figure 4-8C – KGE = 0.92) closely resembled the observed flow timeseries. In contrast, the simulations with low biases for other descriptor combinations reproduced low-flow conditions poorly – overestimating flow volumes and inaccurately predicting small variations in low-flow response (Figure 4-8A, B, and C).

Table 4-3. Minimum, median, and maximum KGE scores of behavioural simulation subsets that had [I] low FDC biases only, [II] low timing biases only, [III] low threshold biases only, or low biases related to combinations of these descriptors. Columns are grouped by the number of low biases. Simulations are separated based on behavioural simulation subsets. Simulations with the maximum KGE score for each group are bolded and their flow timeseries are shown in Figure 4-8. "NA": cases where the group range exceeds the possible number of bias measures. "-": cases where no simulation of a given subset reproduced the associated number of biases.

# of Biases	1-3			4-6			7-9			10-11		
KGE												
	Minimum	Median	Maximum	Minimum	Median	Maximum	Minimum	Median	Maximum	Minimum	Median	Maximum
[I]	-	-	-	-	-	-	NA	NA	NA	NA	NA	NA
[II]	-	-	-	NA	NA	NA	NA	NA	NA	NA	NA	NA
[III]	0.30	0.38	**0.87**	0.30	0.37	0.85	NA	NA	NA	NA	NA	NA
[I] and [II]	-	-	-	-	-	-	NA	NA	NA	NA	NA	NA
[I] and [III]	0.30	0.47	0.80	0.30	0.40	**0.89**	0.30	0.38	0.71	NA	NA	NA
[II] and [III]	0.30	0.48	0.83	0.30	0.42	**0.91**	0.30	0.38	0.76	NA	NA	NA
[I] and [II] and [III]	0.36	0.46	0.62	0.30	0.51	**0.92**	0.30	0.44	0.85	0.39	0.42	0.45

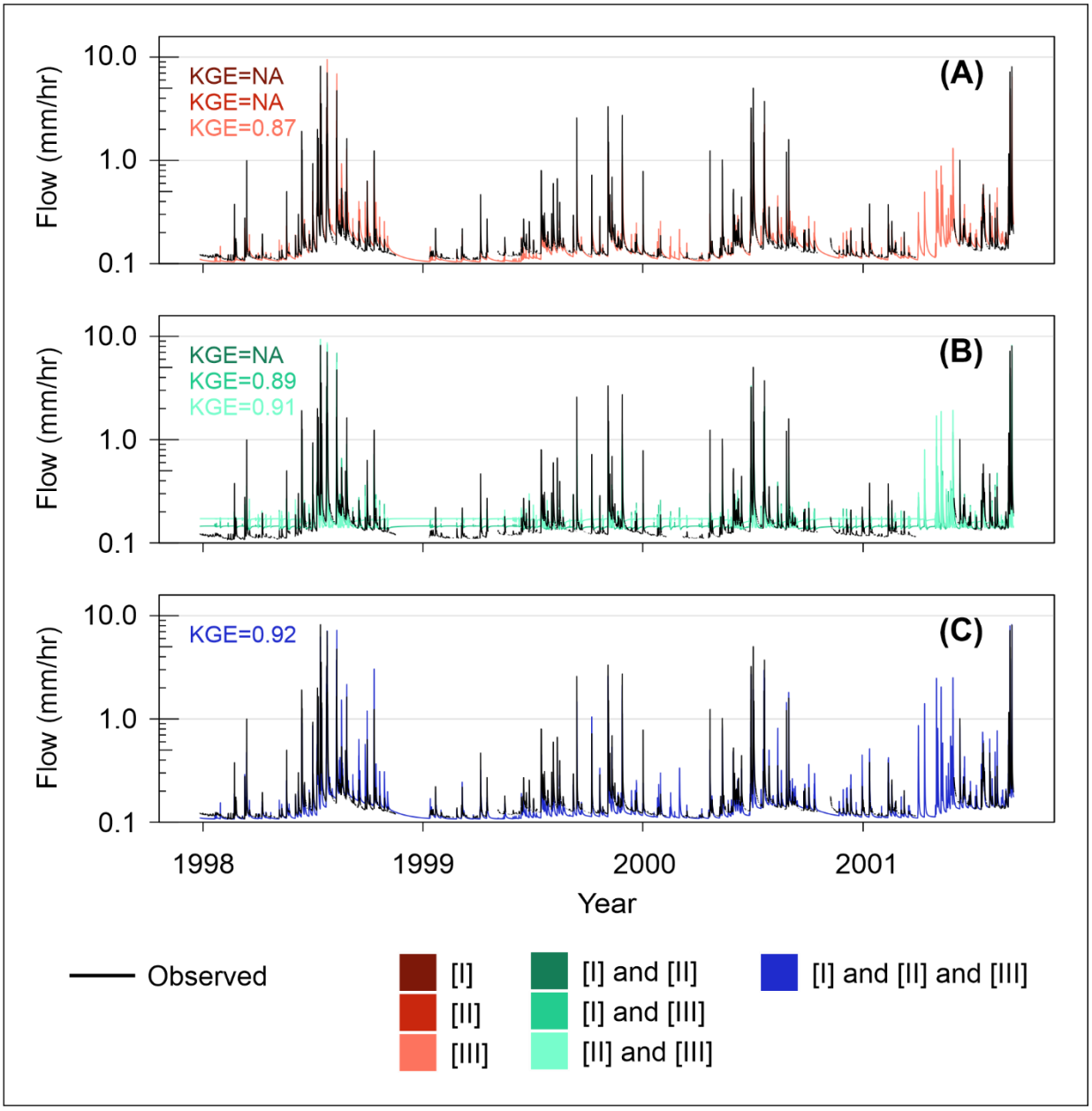


Figure 4-8. Observed and modelled flow timeseries, the latter colour-coded according to [I] low FDC biases only, [II] low timing biases only, [III] low threshold biases only (panel A); or low biases related to combinations of two of these descriptors (panel B); or low biases related to all three descriptors (panel C). Modelled flow timeseries are of behavioural simulations with the maximum KGE scores that had low biases related to different descriptors (shown in bold in Table 4-3). "NA": cases where no simulation no simulations were part of a given subset.

4.3.4 Model parameterization and parameter distributions

Variability in parameter values across behavioural simulations differed between parameters (Table 4-4). As indicated by the coefficient of variation (CV), X_5 was the least variable (CV = 0.55), while X_2 was the most variable (CV = -1.30), followed by X_3 (CV = 0.92) , X_4 (CV = 0.78), and X_1 (CV = 0.71). The KGE score was relatively insensitive to X_1, X_2, and X_5 (Figure 4-9): values of X_1 and X_2 span much of their feasible ranges, regardless of simulation KGE. In contrast, simulations with higher KGE scores were associated with X_3 and X_4 values within a relatively narrow range.

Table 4-4. Summary statistics of parameter values and KGE scores for behavioural simulations. SD: standard deviation and CV: coefficient of variation. See Section 4.2.3 for parameter abbreviations. Table columns are presented independently and do not imply row-wise relationships between parameters and the KGE descriptive statistics.

	X_1 (mm)	X_2 (mm/hr)	X_3 (mm)	X_4 (hours)	X_5 (-)	KGE
Mean	620.034	-1.497	140.253	6.856	0.290	0.440
Median	493.303	-0.672	97.619	5.347	0.266	0.409
Minimum	100.279	-9.974	0.017	0.500	0.001	0.300
Maximum	1999.212	-3.607e-05	499.742	24.952	0.999	0.916
SD	440.210	1.952	128.871	5.348	0.160	0.117
CV	0.710	-1.304	0.919	0.780	0.550	0.265

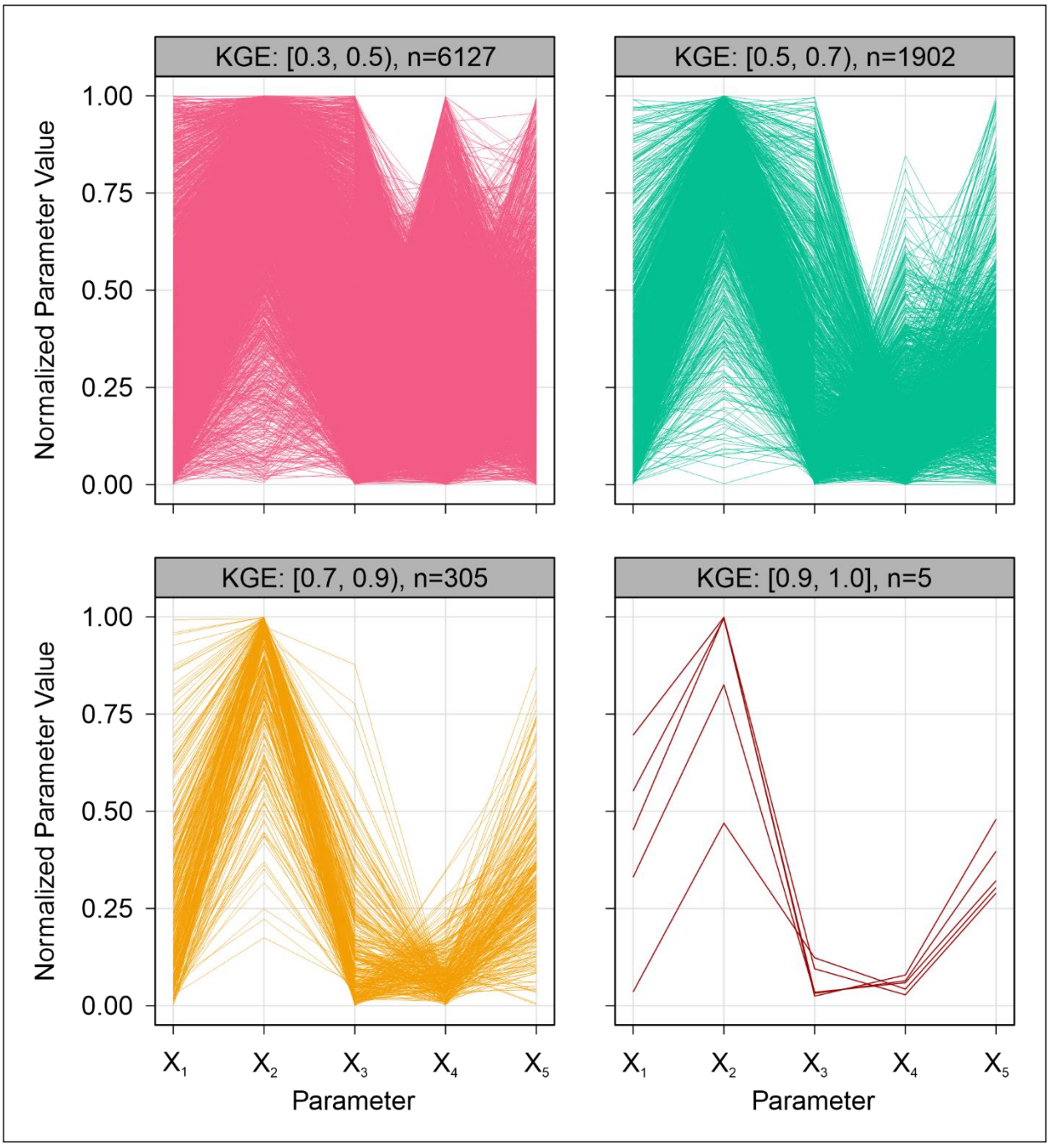


Figure 4-9. Coordinate plots showing normalized parameter values for behavioural simulations. Parameters were normalized by subtracting the minimum parameter value and dividing by the range of parameter values across all behavioural simulations. Each line is associated with one behavioural simulation, and simulations are colour-coded by KGE score range.

According to two-sample Kolmogorov-Smirnov test results, behavioural simulations that met the 15% Pbias criterion for different measures of bias mostly had parameter distributions that were different from those of the remaining behavioural simulations in terms of median, variability, or distribution shape (Table 4-5). The few instances where the differences between the parameter distributions of behavioural simulations that met one or several 15% Pbias criteria and the parameter distributions of remaining behavioural simulations were not statistically significant involved the X_5 parameter. Behavioural simulations that met the 15% Pbias criterion for groups of biases related to a single descriptor also mostly had parameter distributions that differed from that of the remaining behavioural simulations. For example, behavioural simulations that met the 15% Pbias criteria for all FDC biases or all timing metrics did not have parameter values from the same distribution as the remaining behavioural simulations. However, behavioural simulations that met the 15% Pbias criteria for all AR_X+R_{TOT} thresholds did not have parameter distributions that were different from that of the remaining behavioural simulations at a statistically significant level.

Table 4-5. p-values of two-sample Kolmogorov-Smirnov tests that were performed to compare the parameter distributions of behavioural simulations that met the 15% Pbias criterion for different measures of bias to the parameter distributions of the remaining behavioural simulations. "-": cases with too few simulations to perform statistical testing.

	X_1	X_2	X_3	X_4	X_5
Flow duration curve					
FLV	0.01	0.00	0.00	0.00	0.00
FMS	0.00	0.00	0.00	0.00	0.00
FMV	0.00	0.00	0.00	0.00	0.00
FHV	0.00	0.00	0.00	0.00	0.00
All FDC biases	0.00	0.00	0.00	0.00	0.00
Event-specific response timing metric					
T_{LP}	0.00	0.00	0.00	0.00	0.00
T_{LPC}	0.00	0.00	0.00	0.00	0.25
All response timing biases	0.00	0.00	0.00	0.00	0.00
Event rainfall threshold relationship					
R_{TOT}, Q_{TOT}	0.00	0.00	0.00	0.00	0.00
Event plus antecedent rainfall threshold relationships					
AR_3+R_{TOT}, Q_{TOT}	0.00	0.00	0.00	0.00	0.00
AR_7+R_{TOT}, Q_{TOT}	0.00	0.00	0.00	0.00	0.00
AR_7+R_{TOT}, Q_{MAX}	0.00	0.00	0.00	0.00	0.00
AR_{14}+R_{TOT}, Q_{TOT}	0.00	0.00	0.00	0.00	0.00
All AR_X+R_{TOT} threshold biases	0.64	0.79	0.18	0.42	0.33

4.4 Discussion

The overall goal of this study was to assess if hydrologic descriptors derived from readily available flow timeseries could be used in post-calibration model evaluation to facilitate the identification of high-fidelity simulations. Behavioural simulations (n = 8339) were identified

based on a performance measure criterion (KGE ≥ 0.3) from 500,000 Monte Carlo model simulations. Visual inspection of the observed and simulated flow timeseries showed that many of these behavioural simulations reproduced response dynamics at higher flow conditions reasonably well (Figure 4-2). However, the flow volumes and response dynamics for lower flows were often poorly estimated. This was unsurprising, as behavioural model simulations identified based on total streamflow are often biased toward the interpretation of rapid runoff (Dunn, 1999).

4.4.1 From descriptor-specific biases to process interpretations

The ability of behavioural simulations to minimize measures of bias related to the FDC, event response timing, or hydrologic thresholds may offer some insights into process representation within the GR5H model. Relatively few behavioural simulations had low FDC biases related to low flow volumes and the long-term sustainability of baseflow (i.e., FLV) or the vertical redistribution of soil moisture (i.e., FMS) (Vogel & Fennessey, 1994; Yilmaz et al., 2008). These findings are consistent with literature that corroborates the difficulty of predicting baseflow volumes and low flow response dynamics (Dunn, 1999; Gallart et al., 2007; Kroll et al., 2004). A comparatively large number of behavioural simulations had low FDC biases related to the hydrologic response to larger rainfall events (i.e., FMV and FHV). This also coincides with findings from other studies that suggest that many rainfall-runoff models involve trade-offs in terms of adequately reproducing base-flow dominated conditions and responses associated with larger rainfall events (Ley et al., 2016).

The timing of peak discharge in response to rainfall was adequately predicted by many of the behavioural simulations identified in this study. This indicates that the model adequately reproduced the location of flow peaks, relative to the start and centroid of event rainfall, and that travel times of runoff generation mechanisms active over short durations are well represented (Dingman, 2015). However, ~15% of the behavioural simulations that reproduced T_{LPC} with Pbias ≤ 15% erroneously predicted negative T_{LPC} values for some events, signalling peak event discharge preceding the event rainfall centroid. This is possible since behavioural simulations that reproduced T_{LPC} with Pbias ≤ 15% need not reproduce the response timing of all events equally well. As an example, behavioural simulations that accurately reproduced the timing of events with average T_{LPC} values (see Appendix B-3) may have inaccurately predicted the timing of longer or more flashy events. It should be noted that event T_{LP} and T_{LPC} only partially characterize event-scale response timing dynamics, and other timing metrics are informative for other aspects of response (e.g., lag to rise, time of concentration). We limited the number of response timing metrics considered in this study to avoid performing rainfall-runoff event delineation from simulated flow timeseries for each event and behavioural simulation, which most other response timing metrics would have required. Other studies have made similar compromises to avoid rainfall-runoff delineation from simulated timeseries (e.g., Yilmaz et al., 2008), and have used alternate timing-related indices, like the bias in peak flow timing (Yang et al., 2004; Yilmaz et al., 2005).

Simulations with low threshold biases mostly involved AR_X+R_{TOT}. In the literature, similar threshold behaviour related to rainfall and catchment antecedent conditions (quantified by antecedent rainfall, soil moisture, or water table levels) has been associated with processes that vary in space and time, like increased hillslope-riparian-stream connectivity (James &

Roulet, 2009; Oswald et al., 2011; Spence & Woo, 2003; Tromp-van Meerveld & McDonnell, 2006a, 2006b) or the activation of faster flowpaths (Detty & McGuire, 2010; Scaife et al., 2020). As such, some may find it surprising that threshold behaviour involving AR_X+R_{TOT} could be reproduced by the lumped GR5H model that ignores spatial heterogeneity. This, however, is likely attributable to the fact that AR_X+R_{TOT} thresholds were not derived from spatially distributed data. AR_X+R_{TOT} thresholds are spatially aggregated proxies for spatially heterogeneous processes that are expressed in terms of the ensuing hydrologic response observed at the catchment outlet. Low bias for AR_X+R_{TOT} thresholds in this study may, therefore, suggest that these processes are sufficiently represented by the model, albeit imperfectly since the process spatial heterogeneity is not fully characterized. Slightly fewer behavioural simulations had a low bias for the observed threshold behaviour of the $R_{TOT} – Q_{TOT}$ pair (Table 4-1 and Figure 4-6), which may indicate that significant changes in catchment storage over shorter event-scale periods are more difficult to predict. This inference is further supported by a large number of false positives for the observed threshold behaviour of the $R_{TOT} – Q_{MAX}$ pair (Table 4-2), which is also related to catchment storage at the rainfall-runoff event scale. Besides the $R_{TOT} – Q_{MAX}$ pair, behavioural simulations that were associated with false-positives or false-negatives for threshold behaviour mostly involved Q_{MAX} (Table 4-2). Apart from the threshold biases calculated in this study, assessing behavioural simulations for threshold-related false-positives and false-negatives was an additional tool for assessing model fidelity that was made possible by performing model evaluation using hydrologic thresholds. It is important to mention that limited process inferences could be made from threshold biases given the way they were computed. Indeed, threshold biases were compound and simultaneously considered the m_1, m_2, threshold value, R^2, and AIC of the segmented linear model: it is, therefore, possible that a simulation with

a low threshold bias underestimated one characteristic of threshold behaviour (e.g., m_1) and overestimated another characteristic of threshold behaviour (e.g., m_2), making the compound bias value difficult to interpret.

4.4.2 No-, low-, moderate-, and high-fidelity model simulations

While most behavioural simulations had a KGE score between 0.3 and 0.5, many behavioural simulations rather had a KGE score exceeding 0.5 and 0.7 (Figure 4-2 and Figure 4-7). These findings articulate the need for model evaluations that consider additional constraints to identify simulations that best reproduce the observed hydrologic response. Here, we discuss the use of different measures of bias, post-calibration, to evaluate the ability of behavioural simulations to reproduce different aspects of hydrologic response captured by the flow duration curve, rainfall-runoff event response timing metrics, and hydrologic thresholds. For discussion purposes, behavioural simulations that did not achieve any low biases will be referred to as no-fidelity; behavioural simulations that achieved low biases related to one descriptor only will be referred to as low-fidelity; and simulations that achieved low biases related to two or three descriptors will be referred to as moderate-fidelity and high-fidelity, respectively.

No behavioural simulations had the lowest possible fidelity in the context of this study (i.e., no-fidelity). Approximately 26% of behavioural simulations had low biases for aspects of response related to only one descriptor (i.e., low-fidelity - Figure 4-7). For example, a behavioural simulation with low threshold biases only was unable to adequately predict response timing on short timescales (i.e., T_{LP} and/or T_{LPC}), volumes associated with baseflow dominated conditions (i.e., FLV), volumes associated with large rainfall events (i.e., FMV and FHV), or the

vertical redistribution of soil moisture in the catchment (i.e., FMS). All low-fidelity behavioural simulations had low threshold biases and relatively low KGE scores, between 0.3 and 0.5 (Figure 4-7). However, some low-fidelity simulations that had low threshold biases only did achieve higher performance (maximum KGE = 0.87) and closely resembled the observed continuous flow timeseries (Figure 4-8A), indicating that some low-fidelity simulations could have provided "the right answers for the wrong reasons" (Kirchner, 2006). This conclusion is similar to that reached by other studies, which acknowledged that simulations deemed high-performance based on a single performance measure applied to the continuous streamflow timeseries often poorly replicate catchment functions represented by specific hydrologic descriptors or auxiliary data (Casper et al., 2012; Kelleher et al., 2017; Ley et al., 2011; Ley et al., 2016).

Many behavioural simulations had low biases related to two or three hydrologic descriptors (Figure 4-8). These simulations were moderate-fidelity or high-fidelity, as they adequately estimated a broader range of processes. However, the maximum possible fidelity level in the context of this study, i.e., the fidelity associated with minimizing eleven bias measures, was achieved by two behavioural simulations with relatively low KGE scores of 0.39 and 0.45. Simulations with low biases related to two descriptors had either low FDC and threshold biases or low timing and threshold biases. Approximately 24% of behavioural simulations had at least one low bias measure associated with each of the three descriptors (Figure 4-7). Overall, the median KGE scores of simulations with low biases related to two or three descriptors were slightly higher than that of simulations with low biases related to a single descriptor, but no such pattern was observed in the maximum KGE scores (Figure 4-7 and Table 4-3). This indicates that the fidelity of behavioural simulations is not strongly related to simulation performance according to the KGE. For example, behavioural simulations with at

least one low bias measure associated with each of the three descriptors could be associated with KGE scores as high as 0.92 (Table 4-3 and Figure 4-8). Conversely, behavioural simulations with at least one low bias measure associated with each of the three descriptors could also be associated with KGE scores as low as 0.30 (Table 4-3). Also, there was no indication that the minimum, median, or maximum KGE score increases with the number of low biases for high-fidelity simulations (Table 4-3): their median KGE scores ranged from 0.42 to 0.51, and by most accounts would not be considered high-performance. Furthermore, behavioural simulations with four to six low biases related to all three descriptors had KGE scores as high as 0.92, while behavioural simulations with seven to nine low biases related to all three descriptors had KGE scores as low as 0.30. This indicates that the KGE score of behavioural simulations did not increase monotonically with the number of biases ≤ 15% considered. These results are similar to other studies that have also shown that model simulations ranging in fidelity can achieve similar performance measure scores, suggesting that performance measures and model fidelity are not strongly related (Kelleher et al., 2017; Ley et al., 2016).

4.4.3 Parameter distributions of low-, moderate-, and high-fidelity behavioural simulations

A large variety of parameter sets found throughout the feasible parameter space achieved the behavioural criterion defined in this study (Figure 4-2, Figure 4-7, and Table 4-4). The X_1 and X_2 parameter values covered a wide range, regardless of KGE score, indicating that the KGE score calculated for the continuous streamflow timeseries was inadequate for constraining these parameters. Others have shown that a performance measure based on the continuous flow timeseries is inadequate for distinguishing between the parameter sets of behavioural simulations

(Dunn, 1999; Kelleher et al., 2017; Kroll et al., 2004; Ley et al., 2016), and more recently this point has been used to advocate for more robust, purpose-dependent methods for identifying parameter sets of high-fidelity model simulations (Knoben et al., 2019; Schwemmle et al., 2020). The parameter values of behavioural simulations that met the 15% Pbias criterion for different measures of bias were generally shown to be from different distributions than that of the remaining behavioural simulations (Table 4-5). This was, however, not always the case. The X_5 parameter values of behavioural simulations that reproduced T_{LPC} with Pbias ≤ 15% were not from a different distribution than that of the remaining behavioural simulations. This was also true for parameter values of behavioural simulations that had all AR_X+R_{TOT} threshold biases with Pbias ≤ 15%. Since model evaluation was conducted post-calibration, biases had no bearing on the model parameterization process. However, differences between the parameter distributions of behavioural simulations with one or more biases < 15% and the remaining behavioural simulations suggest that these biases can be used to distinguish behavioural simulations and constrain parameter ranges. For example, behavioural simulations with FLV ≤ 15% had parameter values for all five parameters that were not from the same distribution as the parameters associated with the remaining behavioural simulations. Therefore, the parameter distributions of these behavioural simulations with FLV ≤ 15% could be used to narrow the range of acceptable parameter values. This information could be used to falsify other behavioural simulations, or it could be used to confine the parameter space during automated parameter optimization. Furthermore, since each measure of bias is related to a different aspect of catchment response, the fact that the parameter distributions of behavioural simulations that minimize specific biases differs from the remaining behavioural simulations may be an indication that specific parameter value ranges could be associated with different functions of the

sub-catchment featured in this study. As an example, parameter values associated with the level of the routing store at which the flux exchange changes sign (i.e., X_5) differed between behavioural simulations with FLV or FHV ≤ 15%. The X_5 values of behavioural simulations with FLV ≤ 15% ranged from 0.17 to 0.99 and the X_5 values of behavioural simulations with FHV ≤ 15% ranged from 0.00 to 0.99. This demonstrates that different X_5 parameter value ranges, although overlapping, were associated with high-fidelity in terms of low-flow volumes and the long-term sustainability of baseflow (i.e., FLV) and high-fidelity in terms of catchment response to large rainfall events (i.e., FHV).

4.4.4 Results sensitivity to the Pbias criterion

Most studies that have assessed models using auxiliary data or enhanced streamflow data have set a single performance measure criterion to indicate whether a simulation adequately reproduced the data being evaluated. For example, Kelleher et al. (2017) assessed behavioural simulations on a range of constraints using one performance measure criterion for each constraint. Alternatively, others have compared simulations using measures of bias for segments of the FDC to identify the best-performing simulation rather than setting a specific acceptability criterion (e.g., Ley et al., 2016). In this paper, model simulations were assessed on their ability to minimize FDC biases, response timing biases, and threshold biases using a 15% Pbias criterion. We do not compare the relative merits of performance measures other than the Pbias, as this is discussed elsewhere (Krause et al., 2005; Ley et al., 2016; Moriasi et al., 2015). However, we were interested in how sensitive our results were to the specific criterion chosen, so the same analyses were also performed using 5% and 25% Pbias criteria. Results for these criteria are

shown in the Appendix B-5 – Appendix B-13. For most measures of bias, 30 to 65% fewer behavioural simulations met the 5% Pbias criterion than the 15% Pbias criterion (Appendix B-5). Similarly, for most measures of bias, 2 to 40% more behavioural simulations met the 25% Pbias criterion than the 15% Pbias criterion. The number of behavioural simulations that adequately reproduced threshold behaviour for pairs involving AR_7+R_{TOT} or AR_{14}+R_{TOT} and Q_{TOT} was similar at the 15% and 25% Pbias criteria. Results and conclusions related to no-, low-, moderate-, and high-fidelity assessments were mostly similar, regardless of the criterion used, with some exceptions. For example, 47% of behavioural simulations with timing bias within 5% overestimated T_{LP} and T_{LPC}, while the opposite was true for simulations that met the 15% or 25% criterion. Also, the median KGE scores of behavioural simulations were similar, regardless of the criterion satisfied, but the corresponding maximum KGE scores of these simulations decreased with stricter Pbias criteria (Appendix B-9). The percentage of simulations with low biases varied only slightly depending on the Pbias criterion considered. For example, fewer behavioural simulations had low threshold bias only at 15% and 25% criteria compared to the 5% criterion. Overall, model evaluation using three Pbias criteria showed that for the present study, interpretations of model fidelity are not highly sensitive to the exact Pbias criterion considered.

4.5 Conclusion

Rainfall-runoff models are valuable tools that facilitate predictions of hydrologic response and offer a means of process-based hypothesis testing. However, even for relatively simple rainfall-runoff models like the GR5H model, multiple unique parameter sets can achieve the same high level of performance according to common performance measures applied across

the entire continuous streamflow timeseries. These high-performance parameter sets often misrepresent some aspects of hydrologic response. In this study, we performed a post-calibration model evaluation on behavioural simulations (KGE $\geq$ 0.3) using eleven measures of bias related to specific functions of a sub-catchment of the Mahurangi River catchment in New Zealand. This was done to assess the influence of such an evaluation on parameter uncertainty and the identification of high-fidelity model simulations that accurately represent a range of real-world hydrologic processes. One novel contribution of this study was that, in addition to model evaluation using common hydrologic descriptors like the flow duration curve, we also used characteristics of individual rainfall-runoff events and thresholds of rainfall-runoff relationships. All three descriptors that biases were calculated for were derived from the continuous streamflow timeseries, thereby making enhanced use of readily available data. Key findings of this study are listed below.

(1) Some low-fidelity behavioural simulations that failed to minimize many process-related measures of bias attained high KGE scores, indicating a lack of connection between the KGE of a behavioural simulation and how well a behavioural simulation predicts catchment functions.

(2) The parameter distributions of behavioural simulations that adequately reproduced different aspects of hydrologic response differed from those associated with other behavioural simulations, indicating that process-based measures of bias can be used to constrain parameter space.

(3) Like many other studies, behavioural simulations were found to exhibit trade-offs between adequately predicting baseflow-dominated conditions and predicting the response timing and volumes associated with large rainfall events.

(4) Most behavioural simulations accurately predicted threshold behaviour involving antecedent rainfall plus event rainfall. Emergent catchment properties, like thresholds, therefore, appear to be effective and underutilized tools to support model evaluation by helping to distinguish behavioural simulations with similar performance measure scores.

Overall, findings from this study indicate that readily available streamflow data can be effectively leveraged to identify high-fidelity simulations and constrain parameter space. It would be beneficial for future studies to assess the usefulness of such an evaluation of rainfall-runoff models of varying complexity. We especially anticipate the further use of hydrologic thresholds in modelling exercises as they are increasingly being used to characterize and compare catchment responses.

4.6 References

Ala-aho, P., Tetzlaff, D., McNamara, J. P., Laudon, H., & Soulsby, C. (2017). Using isotopes to constrain water flux and age estimates in snow-influenced catchments using the STARR (Spatially distributed Tracer-Aided Rainfall–Runoff) model. *Hydrology and Earth System Sciences*, *21*(10), 5089–5110. https://doi.org/10.5194/hess-21-5089-2017

Ali, G., Oswald, C., Spence, C., Cammeraat, E., McGuire, K., Meixner, T., & Reaney, S. M. (2013). Towards a unified threshold-based hydrological theory: Necessary components and recurring challenges. *Hydrological Processes*, *27*(2), 313–318. https://doi.org/10.1002/hyp.9560

Ali, G., Tetzlaff, D., McDonnell, J., Soulsby, C., Carey, S., Laudon, H., McGuire, K., Buttle, J., Seibert, J., & Shanley, J. (2015). Comparison of threshold hydrologic response across

northern catchments. *Hydrological Processes*, *29*(16), 3575–3591. https://doi.org/10.1002/hyp.10527

Benke, K. K., Lowell, K. E., & Hamilton, A. J. (2008). Parameter uncertainty, sensitivity analysis and prediction error in a water-balance hydrological model. *Mathematical and Computer Modelling*, *47*(11), 1134–1149. https://doi.org/10.1016/j.mcm.2007.05.017

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology*, 320(1–2), 18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007

Beven, K. (2011). *Rainfall-runoff modelling: The primer*. John Wiley & Sons.

Beven, K., & Binley, A. (1992). The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes*, *6*(3), 279–298.

Beven, K., & Binley, A. (2014). GLUE: 20 years on. *Hydrological Processes*, *28*(24), 5897–5918. https://doi.org/10.1002/hyp.10082

Beven, K. J., & Kirkby, M. J. (1979). A physically based, variable contributing area model of basin hydrology/Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant. *Hydrological Sciences Journal*, *24*(1), 43–69.

Blazkova, S., & Beven, K. (2009). A limits of acceptability approach to model evaluation and uncertainty estimation in flood frequency estimation by continuous simulation: Skalka catchment, Czech Republic. *Water Resources Research*, *45*(12). https://doi.org/10.1029/2007WR006726

Blöschl, G., Sivapalan, M., Savenije, H., Wagener, T., & Viglione, A. (2013). *Runoff prediction in ungauged basins: Synthesis across processes, places and scales*. Cambridge University Press.

Boyle, D. P., Gupta, H. V., & Sorooshian, S. (2000). Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods. *Water Resources Research*, *36*(12), 3663–3674. https://doi.org/10.1029/2000WR900207

Brath, A., Montanari, A., & Toth, E. (2004). Analysis of the effects of different scenarios of historical data availability on the calibration of a spatially-distributed hydrological model. *Journal of Hydrology*, *291*(3), 232–253. https://doi.org/10.1016/j.jhydrol.2003.12.044

Cammeraat, L. H. (2002). A review of two strongly contrasting geomorphological systems within the context of scale. *Earth Surface Processes and Landforms*, *27*(11), 1201–1222. https://doi.org/10.1002/esp.421

Carey, S. K., & Woo, M. (2001). Slope runoff processes and flow generation in a subarctic, subalpine catchment. Journal of Hydrology, 253(1–4), 110–129. https://doi.org/10.1016/S0022-1694(01)00478-4

Casper, M. C., Grigoryan, G., Gronz, O., Gutjahr, O., Heinemann, G., Ley, R., & Rock, A. (2012). Analysis of projected hydrological behavior of catchments based on signature indices. *Hydrology and Earth System Sciences*, *16*(2), 409–421. https://doi.org/10.5194/hess-16-409-2012

Clark, M. P., Kavetski, D., & Fenicia, F. (2011). Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resources Research*, *47*(9). https://doi.org/10.1029/2010WR009827

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., Freer, J. E., Gutmann, E. D., Wood, A. W., Brekke, L. D., Arnold, J. R., Gochis, D. J., & Rasmussen, R. M. (2015). A unified approach for process-based hydrologic modeling: 1. Modeling

concept. *Water Resources Research*, *51*(4), 2498–2514. https://doi.org/10.1002/2015WR017198

Conover, W. J. (1998). *Practical nonparametric statistics* (Vol. 350). John Wiley & Sons.

Coron, L., Thirel, G., Delaigue, O., Perrin, C., & Andréassian, V. (2017). The suite of lumped GR hydrological models in an R package. *Environmental Modelling & Software*, *94*, 166–171. https://doi.org/10.1016/j.envsoft.2017.05.002

Coron, Laurent, Delaigue, O., Thirel, G., Perrin, C., & Michel, C. (2020). *airGR: Suite of GR Hydrological Models for Precipitation-Runoff Modelling. R package version 1.4.3.65*. [Data set]. Portail Data INRAE. https://doi.org/10.15454/EX11NA

Detty, J. M., & McGuire, K. J. (2010). Threshold changes in storm runoff generation at a till-mantled headwater catchment. *Water Resources Research; Washington*, *46*(7). http://dx.doi.org.uml.idm.oclc.org/10.1029/2009WR008102

Dingman, S. L. (2015). *Physical hydrology*. Waveland press.

Dunn, S. M. (1999). Imposing constraints on parameter values of a conceptual hydrological model using baseflow response. *Hydrology and Earth System Sciences Discussions*, *3*(2), 271–284.

Ficchì, A. (2017). *An adaptive hydrological model for multiple time-steps: Diagnostics and improvements based on fluxes consistency* [Phdthesis, Université Pierre et Marie Curie - Paris VI]. https://tel.archives-ouvertes.fr/tel-01619102

Ficchì, A., Perrin, C., & Andréassian, V. (2019). Hydrological modelling at multiple sub-daily time steps: Model improvement via flux-matching. *Journal of Hydrology*, *575*, 1308–1327. https://doi.org/10.1016/j.jhydrol.2019.05.084

Franks, S. W., Gineste, P., Beven, K. J., & Merot, P. (1998). On constraining the predictions of a distributed model: The incorporation of fuzzy estimates of saturated areas into the calibration process. *Water Resources Research*, *34*(4), 787–797. https://doi.org/10.1029/97WR03041

Freer, J., Beven, K., & Ambroise, B. (1996). Bayesian Estimation of Uncertainty in Runoff Prediction and the Value of Data: An Application of the GLUE Approach. *Water Resources Research*, *32*(7), 2161–2173. https://doi.org/10.1029/95WR03723

Gallart, F., Latron, J., Llorens, P., & Beven, K. (2007). Using internal catchment information to reduce the uncertainty of discharge and baseflow predictions. *Advances in Water Resources*, *30*(4), 808–823.

Grayson, R. B., Blöschl, G., Western, A. W., & McMahon, T. A. (2002). Advances in the use of observed spatial patterns of catchment hydrological response. *Advances in Water Resources*, *25*(8), 1313–1334. https://doi.org/10.1016/S0309-1708(02)00060-X

Gronz, O. (2013). *Nutzung von Abflussprozessinformation in LARSIM*. Retrieved July 15, 2020, from https://ubt.opus.hbz-nrw.de/opus45-ubtr/frontdoor/index/index/docId/599

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Hallouin, T., Bruen, M., & O'Loughlin, F. E. (2020). Calibration of hydrological models for ecologically relevant streamflow predictions: A trade-off between fitting well to data and estimating consistent parameter sets? *Hydrology and Earth System Sciences*, *24*(3), 1031–1054.

Herbst, M., Casper, M. C., Grundmann, J., & Buchholz, O. (2009). Comparative analysis of model behaviour for flood prediction purposes using Self-Organizing Maps. *Natural Hazards and Earth System Sciences*, *9*(2), 373–392. https://doi.org/10.5194/nhess-9-373-2009

Hill, M. C., Kavetski, D., Clark, M., Ye, M., Arabi, M., Lu, D., Foglia, L., & Mehl, S. (2016). Practical Use of Computationally Frugal Model Analysis Methods. *Groundwater*, *54*(2), 159–170. https://doi.org/10.1111/gwat.12330

Hossain, S., Hewa, G. A., & Wella-Hewage, S. (2019). A Comparison of Continuous and Event-Based Rainfall–Runoff (RR) Modelling Using EPA-SWMM. *Water*, *11*(3), 611. https://doi.org/10.3390/w11030611

James, A., & Roulet, N. (2009). Antecedent moisture conditions and catchment morphology as controls on spatial patterns of runoff generation in small forest catchments. *Journal of Hydrology*, *377*(3–4), 351–366. https://doi.org/10.1016/j.jhydrol.2009.08.039

Kavetski, D., Kuczera, G., & Franks, S. W. (2006a). Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. *Water Resources Research*, *42*(3). https://doi.org/10.1029/2005WR004368

Kavetski, D., Kuczera, G., & Franks, S. W. (2006b). Bayesian analysis of input uncertainty in hydrological modeling: 2. Application. *Water Resources Research*, *42*(3). https://doi.org/10.1029/2005WR004376

Kelleher, C., McGlynn, B., & Wagener, T. (2017). Characterizing and reducing equifinality by constraining a distributed catchment model with regional signatures, local observations, and process understanding. *Hydrology and Earth System Sciences*, *21*(7), 3325–3352. https://doi.org/10.5194/hess-21-3325-2017

Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research*, *42*(3), W03S04. https://doi.org/10.1029/2005WR004362

Knoben, W. J., Freer, J. E., & Woods, R. A. (2019). Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrology and Earth System Sciences*, *23*(10), 4323–4331.

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

Krause, P., Boyle, D. P., & Bäse, F. (2005). Comparison of different efficiency criteria for hydrological model assessment. *Advances in geosciences*, 5, 89-97.

Kroll, C., Luz, J., Allen, B., & Vogel, R. M. (2004). Developing a Watershed Characteristics Database to Improve Low Streamflow Prediction. *Journal of Hydrologic Engineering*, *9*(2), 116–125. https://doi.org/10.1061/(ASCE)1084-0699(2004)9:2(116)

Kuczera, G., & Mroczkowski, M. (1998). Assessment of hydrologic parameter uncertainty and the worth of multiresponse data. *Water Resources Research*, *34*(6), 1481–1489. https://doi.org/10.1029/98WR00496

Kuraś, P. K., Alila, Y., Weiler, M., Spittlehouse, D., & Winkler, R. (2011). Internal catchment process simulation in a snow-dominated basin: Performance evaluation with spatiotemporally variable runoff generation and groundwater dynamics. *Hydrological Processes*, *25*(20), 3187–3203. https://doi.org/10.1002/hyp.8037

Lamb, R., Beven, K., & Myrabø, S. (1998). Use of spatially distributed water table observations to constrain uncertainty in a rainfall–runoff model. *Advances in Water Resources*, *22*(4), 305–317. https://doi.org/10.1016/S0309-1708(98)00020-7

Le Moine, N. (2008). *Le bassin versant de surface vu par le souterrain: Une voie d'amélioration des performances et du réalisme des modèles pluie-débit ?* (p. 348) [Phdthesis, Doctorat Géosciences et Ressources Naturelles, Université Pierre et Marie Curie Paris VI]. https://hal.inrae.fr/tel-02591478

Lehmann, P., Hinz, C., McGrath, G., Tromp-van Meerveld, H. J., & McDonnell, J. J. (2007). Rainfall threshold for hillslope outflow: An emergent property of flow pathway connectivity. *Hydrol. Earth Syst. Sci.*, *11*(2), 1047–1063. https://doi.org/10.5194/hess-11-1047-2007

Ley, R., Casper, M. C., Hellebrand, H., & Merz, R. (2011). Catchment classification by runoff behaviour with self-organizing maps (SOM). *Hydrology and Earth System Sciences*, *15*(9), 2947–2962. https://doi.org/10.5194/hess-15-2947-2011

Ley, Rita, Hellebrand, H., Casper, M. C., & Fenicia, F. (2016). Comparing classical performance measures with signature indices derived from flow duration curves to assess model structures as tools for catchment classification. *Hydrology Research*, *47*(1), 1–14. https://doi.org/10.2166/nh.2015.221

Li, Z., Shao, Q., Xu, Z., & Cai, X. (2010). Analysis of parameter uncertainty in semi-distributed hydrological models using bootstrap method: A case study of SWAT model applied to Yingluoxia watershed in northwest China. *Journal of Hydrology*, *385*(1), 76–83. https://doi.org/10.1016/j.jhydrol.2010.01.025

Loague, K. M., & Freeze, R. A. (1985). A Comparison of Rainfall-Runoff Modeling Techniques on Small Upland Catchments. *Water Resources Research*, *21*(2), 229–248. https://doi.org/10.1029/WR021i002p00229

McDonnell, J., Sivapalan, M., Vaché, K., Dunn, S., Grant, G., Haggerty, R., Hinz, C., Hooper, R., Kirchner, J., Roderick, M. L., Selker, J., & Weiler, M. (2007). Moving beyond heterogeneity and process complexity: A new vision for watershed hydrology. *Water Resources Research*, *43*(7), W07301. https://doi.org/10.1029/2006WR005467

Mirus, B. B., & Loague, K. (2013). How runoff begins (and ends): Characterizing hydrologic response at the catchment scale. *Water Resources Research*, *49*(5), 2987–3006. https://doi.org/10.1002/wrcr.20218

Mizukami, N., Rakovec, O., Newman, A. J., Clark, M. P., Wood, A. W., Gupta, H. V., & Kumar, R. (2019). On the choice of calibration metrics for "high-flow" estimation using hydrologic models. *Hydrology and Earth System Sciences*, *23*(6), 2601–2614. https://doi.org/10.5194/hess-23-2601-2019

Moriasi, D. N., Gitau, M. W., Pai, N., & Daggupati, P. (2015). Hydrologic and water quality models: Performance measures and evaluation criteria. *Transactions of the ASABE*, *58*(6), 1763–1785.

Mosley, M. P. (1979). Streamflow generation in a forested watershed, New Zealand. *Water Resources Research*, 15(4), 795–806. https://doi.org/10.1029/WR015i004p00795

Muggeo, V. M. (2008). Segmented: An R package to fit regression models with broken-line relationships. *R News*, *8*(1), 20–25.

Oswald, C., Richardson, M. C., & Branfireun, B. A. (2011). Water storage dynamics and runoff response of a boreal Shield headwater catchment. *Hydrological Processes, 25*(19), 3042–3060. https://doi.org/10.1002/hyp.8036

Perrin, C., Michel, C., & Andréassian, V. (2003). Improvement of a parsimonious model for streamflow simulation. *Journal of Hydrology, 279*(1), 275–289. https://doi.org/10.1016/S0022-1694(03)00225-7

Phillips, J. D. (2006). Evolutionary geomorphology: Thresholds and nonlinearity in landform response to environmental change. *Hydrol. Earth Syst. Sci., 10*(5), 731–742. https://doi.org/10.5194/hess-10-731-2006

Post, D. A., & Jakeman, A. J. (1996). Relationships between catchment attributes and hydrological response characteristics in small Australian mountain ash catchments. *Hydrological Processes, 10*(6), 877–892.

Reaney, S. M., Bracken, L. J., & Kirkby, M. J. (2007). Use of the Connectivity of Runoff Model (CRUM) to investigate the influence of storm characteristics on runoff generation and connectivity in semi-arid areas. *Hydrological Processes, 21*(7), 894–906. https://doi.org/10.1002/hyp.6281

Redding, T. E., & Devito, K. J. (2008). Lateral flow thresholds for aspen forested hillslopes on the Western Boreal Plain, Alberta, Canada. *Hydrological Processes, 22*(21), 4287–4300. https://doi.org/10.1002/hyp.7038

Ross, C. A., Ali, G., Spence, C., & Courchesne, F. (2021). Evaluating the Ubiquity of Thresholds in Rainfall-Runoff Response Across Contrasting Environments. *Water Resources Research*, e2020WR027498. https://doi.org/10.1029/2020WR027498

Ross, C. A., Ali, G., Spence, C., Oswald, C., & Casson, N. (2019). Comparison of event-specific rainfall–runoff responses and their controls in contrasting geographic areas. *Hydrological Processes*, *33*(14), 1961–1979. https://doi.org/10.1002/hyp.13460

Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1986). Akaike information criterion statistics. *Dordrecht, The Netherlands: D. Reidel*, *81*.

Scaife, C. I., & Band, L. E. (2017). Nonstationarity in threshold response of stormflow in southern Appalachian headwater catchments. *Water Resources Research*, *53*(8), 6579–6596. https://doi.org/10.1002/2017WR020376

Scaife, C. I., Singh, N. K., Emanuel, R. E., Miniat, C. F., & Band, L. E. (2020). Non-linear quickflow response as indicators of runoff generation mechanisms. *Hydrological Processes*, 34(13), 2949-2964.

Schwemmle, R., Demand, D., & Weiler, M. (2020). Technical note: Diagnostic efficiency – specific evaluation of model performance. *Hydrology and Earth System Sciences Discussions*, 1–15. https://doi.org/10.5194/hess-2020-237

Seibert, J., & McDonnell, J. J. (2002). On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria model calibration. *Water Resources Research*, *38*(11), 23-1-23–14. https://doi.org/10.1029/2001WR000978

Sidle, R. C., Tsuboyama, Y., Noguchi, S., Hosoda, I., Fujieda, M., & Shimizu, T. (2000). Stormflow generation in steep forested headwaters: A linked hydrogeomorphic paradigm. *Hydrological Processes*, *14*(3), 369–385. https://doi.org/10.1002/(SICI)1099-1085(20000228)14:3<369::AID-HYP943>3.0.CO;2-P

Singh, V. P. (1995). *Computer models of watershed hydrology*. Water Resources Publications. https://catalog.hathitrust.org/Record/003186137

Sivapalan, M., Jothityangkoon, C., & Menabde, M. (2002). Linearity and nonlinearity of basin response as a function of scale: Discussion of alternative definitions. *Water Resources Research*, *38*(2). https://doi.org/10.1029/2001WR000482

Son, K., & Sivapalan, M. (2007). Improving model structure and reducing parameter uncertainty in conceptual water balance models through the use of auxiliary data. *Water Resources Research*, *43*(1). https://doi.org/10.1029/2006WR005032

Spear, R. C., & Hornberger, G. M. (1980). Eutrophication in peel inlet—II. Identification of critical uncertainties via generalized sensitivity analysis. *Water Research*, *14*(1), 43–49.

Spence, C. (2010). A Paradigm Shift in Hydrology: Storage Thresholds Across Scales Influence Catchment Runoff Generation. *Geography Compass*, *4*(7), 819–833. https://doi.org/10.1111/j.1749-8198.2010.00341.x

Spence, Christopher, & Woo, M. (2003). Hydrology of subarctic Canadian shield: Soil-filled valleys. *Journal of Hydrology*, *279*(1), 151–166. https://doi.org/10.1016/S0022-1694(03)00175-6

Stadnyk, T. A., Delavau, C., Kouwen, N., & Edwards, T. W. D. (2013). Towards hydrological model calibration and validation: Simulation of stable water isotopes using the isoWATFLOOD model. *Hydrological Processes*, *27*(25), 3791–3810. https://doi.org/10.1002/hyp.9695

Staudinger, M., Stahl, K., Seibert, J., Clark, M. P., & Tallaksen, L. M. (2011). Comparison of hydrological model structures based on recession and low flow simulations. *Hydrology and Earth System Sciences*, 15(11), 3447-3459.

Tan, S. B., Chua, L. H., Shuy, E. B., Lo, E. Y.-M., & Lim, L. W. (2008). Performances of Rainfall-Runoff Models Calibrated over Single and Continuous Storm Flow Events.

Journal of Hydrologic Engineering, *13*(7), 597–607. https://doi.org/10.1061/(ASCE)1084-0699(2008)13:7(597)

Tang, W., & Carey, S. K. (2017). HydRun: A MATLAB toolbox for rainfall–runoff analysis. *Hydrological Processes*, *31*(15), 2670–2682. https://doi.org/10.1002/hyp.11185

Tani, M. (1997). Runoff generation processes estimated from hydrological observations on a steep forested hillslope with a thin soil layer. *Journal of Hydrology*, *200*(1), 84–109. https://doi.org/10.1016/S0022-1694(97)00018-8

Thornthwaite, C. W. (1948). An approach toward a rational classification of climate. *Geographical Review*, *38*(1), 55–94. https://doi.org/10.2307/210739

Thyer, M., Renard, B., Kavetski, D., Kuczera, G., Franks, S. W., & Srikanthan, S. (2009). Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis. *Water Resources Research*, *45*(12). https://doi.org/10.1029/2008WR006825

Tromp-van Meerveld, H. J., & McDonnell, J. J. (2006a). Threshold relations in subsurface stormflow: 1. A 147-storm analysis of the Panola hillslope. *Water Resources Research*, *42*(2). https://doi.org/10.1029/2004WR003778

Tromp-van Meerveld, H. J., & McDonnell, J. J. (2006b). Threshold relations in subsurface stormflow: 2. The fill and spill hypothesis. *Water Resources Research*, *42*(2). https://doi.org/10.1029/2004WR003800

Vogel, R. M., & Fennessey, N. M. (1994). Flow-duration curves. I: New interpretation and confidence intervals. *Journal of Water Resources Planning and Management*, *120*(4), 485–504.

Vogel, R. M., & Fennessey, N. M. (1995). Flow duration curves II: A review of applications in water resources planning 1. *JAWRA Journal of the American Water Resources Association*, *31*(6), 1029–1039.

Vrugt, J. A., ter Braak, C. J. F., Gupta, H. V., & Robinson, B. A. (2009). Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling? *Stochastic Environmental Research and Risk Assessment*, *23*(7), 1011–1026. https://doi.org/10.1007/s00477-008-0274-y

Wealands, S. R., Grayson, R. B., & Walker, J. P. (2005). Quantitative comparison of spatial fields for hydrological model assessment—some promising approaches. *Advances in Water Resources*, *28*(1), 15–32. https://doi.org/10.1016/j.advwatres.2004.10.001

Wei, L., Qiu, Z., Zhou, G., Kinouchi, T., & Liu, Y. (2020). Stormflow threshold behaviour in a subtropical mountainous headwater catchment during forest recovery period. *Hydrological Processes*, *34*(8), 1728–1740. https://doi.org/10.1002/hyp.13658

Weiler, M., McDonnell, J. J., Meerveld, I. T., & Uchida, T. (2006). Subsurface Stormflow. In *Encyclopedia of Hydrological Sciences*. John Wiley & Sons, Ltd. https://doi.org/10.1002/0470848944.hsa119

Westerberg, I. K., Guerrero, J. L., Younger, P. M., Beven, K. J., Seibert, J., Halldin, S., Freer, J. E., & Xu, C. Y. (2011). Calibration of hydrological models using flow-duration curves. *Hydrology and Earth System Sciences*, *15*(7), 2205–2227. https://doi.org/10.5194/hess-15-2205-2011

Whipkey, R. Z. (1965). Subsurface Stormflow from Forested Slopes. *International Association of Scientific Hydrology. Bulletin*, *10*(2), 74–85. https://doi.org/10.1080/02626666509493392

Wooding, R. A. (1965). A hydraulic model for the catchment-stream problem: II. Numerical solutions. *Journal of Hydrology*, *3*(3–4), 268–282.

Woods, R., Grayson, R., Western, A., Duncan, M., Wilson, D., Young, R., Ibbitt, R., Henderson, R., & McMahon, T. (2013). Experimental Design and Initial Results from the Mahurangi River Variability Experiment: MARVEX. In *Land Surface Hydrology, Meteorology, and Climate: Observations and Modeling* (pp. 201–213). American Geophysical Union. http://onlinelibrary-wiley-com.uml.idm.oclc.org/doi/10.1029/WS003p0201/summary

Woods, R., & Sivapalan, M. (1999). A synthesis of space-time variability in storm response: Rainfall, runoff generation, and routing. *Water Resources Research*, *35*(8), 2469–2485. https://doi.org/10.1029/1999WR900014

Yafune, A., Narukawa, M., & Ishiguro, M. (2005). A Note on Sample Size Determination for Akaike Information Criterion (AIC) Approach to Clinical Data Analysis. *Communications in Statistics - Theory and Methods*, *34*(12), 2331–2343. https://doi.org/10.1080/03610920500257295

Yair, A., & Raz-Yassif, N. (2004). Hydrological processes in a small arid catchment: Scale effects of rainfall and slope length. *Geomorphology*, *61*(1–2), 155–169.

Yang, T.-C., Yu, P.-S., Kuo, C.-M., & Wang, Y.-C. (2004). Application of Fuzzy Multiobjective Function on Storm-Event Rainfall-Runoff Model Calibration. *Journal of Hydrologic Engineering*, *9*(5), 440–445. https://doi.org/10.1061/(ASCE)1084-0699(2004)9:5(440)

Yilmaz, K. K., Gupta, H. V., & Wagener, T. (2008). A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model. *Water Resources Research*, *44*(9). https://doi.org/10.1029/2007WR006716

Yilmaz, K. K., Hogue, T. S., Hsu, K.-L., Sorooshian, S., Gupta, H. V., & Wagener, T. (2005). Intercomparison of Rain Gauge, Radar, and Satellite-Based Precipitation Estimates with Emphasis on Hydrologic Forecasting. *Journal of Hydrometeorology*, *6*(4), 497–517. JSTOR.

Yu, P.-S., & Yang, T.-C. (2000). Using synthetic flow duration curves for rainfall–runoff model calibration at ungauged sites. *Hydrological Processes*, *14*(1), 117–133. https://doi.org/10.1002/(SICI)1099-1085(200001)14:1<117::AID-HYP914>3.0.CO;2-Q

Zambrano-Bigiarini, M. (2012). HydroTSM: Time series management, analysis and interpolation for hydrological modelling. *R Package Version 0.3*, *3*.

Zhang, J., Li, Y., Huang, G., Chen, X., & Bao, A. (2016). Assessment of parameter uncertainty in hydrological model using a Markov-Chain-Monte-Carlo-based multilevel-factorial-analysis method. *Journal of Hydrology*, *538*, 471–486.

CHAPTER 5. CHARACTERIZING THRESHOLDS IN RAINFALL-RUNOFF RESPONSE: CAN 3D REPRESENTATIONS HELP?

5.1 Introduction

Catchment responses to precipitation are regularly threshold mediated and can be difficult to predict from the simple aggregation of small-scale processes (McDonnell et al., 2007; Sivapalan et al., 2002; Sivapalan, 2006). As a result, there is growing interest in emergent properties that reflect landscape heterogeneity and process complexity (Lehmann et al., 2007; McDonnell et al., 2007). Thresholds in precipitation-runoff relationships are such emergent properties: defined as critical moments in time or points in space that coincide with a significant change in runoff behaviour (Ali et al., 2013; Phillips, 2006), they are increasingly used to characterize nonlinear catchment dynamics (e.g., Ali et al., 2015; Detty & McGuire, 2010; Scaife & Band, 2017; Spence, 2007). Two-dimensional (2D) scatter plots, i.e., hydrologic response metrics plotted against meteorological factors, are routinely used to detect thresholds in precipitation-runoff relationships (e.g., Ali et al., 2013; Scaife & Band, 2017). A wide variety of meteorological factors can be featured on these plots that quantify the volume or depth of water received by a catchment, the rate of water added to a catchment, or hydrologic abstraction caused by evapotranspiration (Ali et al., 2013). Critical values of these factors triggering significant changes in hydrologic response metrics are identified as thresholds. Researchers, to date, have mostly investigated thresholds in the form of critical values of precipitation (Mosley, 1979; Redding & Devito, 2008; Sidle et al., 2000; Tani, 1997; Whipkey, 1965), soil moisture (James & Roulet, 2007), water table levels (Ali et al., 2011; Kim et al., 2004), depression storage (Mielko & Woo, 2006; Oswald et al., 2011; Spence & Woo, 2003; Tromp-van Meerveld & McDonnell, 2006a) or some combination of those variables (Detty & McGuire, 2010; Scaife & Band, 2017; Wei et al., 2020). Fewer studies examined rainfall intensity thresholds, primarily in infiltration-

limited environments (Cammeraat, 2002; Reaney et al., 2007), and thresholds involving factors related to hydrologic abstraction caused by evapotranspiration have not been assessed.

The relative lack of research on thresholds involving rainfall intensity or hydrologic abstraction stems from the fact that most hydrological research is carried in temperate humid, forested catchments where rainfall intensity and evapotranspiration are believed to exert little influence on runoff processes (Graham et al., 2010; Tromp-van Meerveld & McDonnell, 2006a). However, recent studies conducted in a range of environments have challenged that assumption: variables reflecting rainfall depth, rainfall intensity, and hydrologic abstraction related to antecedent evapotranspiration have been shown to strongly influence hydrologic response variability (Cammeraat, 2002; Reaney et al., 2007; Ross et al., 2019). Moreover, the possibility of threshold-mediated responses affected by multiple meteorological factors simultaneously, or by factor interactions, has rarely been considered, which is at odds with a range of ecohydrological process conceptualizations. For example, at the hillslope or catchment scale, factors related to water storage (e.g., soil depth) and rainfall intensity have been used to establish conditions that favour different runoff generation mechanisms (Dingman, 2015; Dunne, 1978). At the vegetation stand scale, rainfall intensity, evaporation, and antecedent canopy storage determine if and when canopy interception capacity is exceeded, thereby controlling the volume and rate of net rainfall reaching the ground and triggering different runoff generation mechanisms (Dingman, 2015). Along the same lines, at the soil matrix scale, both the volume of water and its rate of delivery affect the development of soil saturation that can influence hillslope and catchment hydrologic response (Fetter, 2018).

While the literature on runoff generation, canopy interception, and soil physics discusses how catchment response may be governed by multiple meteorological factors, traditional

methods for detecting thresholds are not well-suited to assess nonlinearities dictated by multiple factors. Indeed, regardless of whether threshold identification is performed through visual identification from 2D scatter plots (e.g., Ali et al., 2015; Detty & McGuire, 2010; Tromp-van Meerveld & McDonnell, 2006a) or piecewise regression analysis (e.g., Oswald et al., 2011; Scaife & Band, 2017), only one response metric and one meteorological factor are considered, thereby implicitly perpetuating the notion that catchment response can be explained by a single factor. This approach differs from ecology, where ecosystem response is often assessed in terms of at least two explanatory factors and analyses carried in three-dimensional (3D) space have helped unravel multi-factor relationships and detect thresholds when they exist (Andersen et al., 2009; Kinzig et al., 2006; Limburg et al., 2002; Lintz et al., 2011). Such a multi-factor approach generally relies on the estimation of 3D response surfaces – as opposed to 2D response curves – and can include the computation of statistical parameters that quantify the abruptness of changes in response across an entire response curve or surface (Lintz et al., 2011). Of course, linear relationships that are devoid of thresholds do not have abrupt changes in response. For those nonlinear relationships with thresholds, however, the quantification of the abruptness of response changes across entire curves or surfaces addresses the diverse ways in which thresholds may manifest and is not limited to quantifying a change in slope at a single inflection point. Specifically, strong thresholds are those characterized by sharper (i.e., more abrupt) changes in response, while weaker thresholds have smoother (i.e., less abrupt) changes in response.

Making distinctions between sharp and smooth thresholds has been previously rationalized in ecology to better understand and predict the conditions that lead to changes in ecosystem response, and to assess the effectiveness of different threshold identification techniques (Ficetola & Denoël, 2009). In hydrology, distinguishing between smooth and sharp

changes in response may similarly contribute to hydrologic response predictability, and it may more generally improve understanding of how meteorological inputs are transformed into runoff. The introduction of a threshold strength parameter, originally developed to quantify the abruptness of ecological response change, allows for such distinctions to be made regardless of the number of explanatory factors considered or whether response changes are present at one or more inflection point(s) (e.g., Lintz et al., 2011). Multi-factor approaches have also been used to gain insight into the factors that underlie system response. In the case of dual-factor relationships, for instance, distinctions can be made between factors that are independent of one another and influence response simultaneously (i.e., additive effects), as opposed to factors that simultaneously influence response but are not independent of one another (i.e., interactive effects) (Antony, 2014; Bailey, 2008).

The extent to which such 3D characterization approaches can be borrowed from other disciplines and applied to hydrology to assess threshold-mediated catchment response has yet to be explored. The goal of the present study is, therefore, to model 3D surfaces to illustrate hydrologic responses as functions of multiple meteorological factor pairs. Specifically, while borrowing approaches used in ecology and other disciplines, 3D response surfaces will be analyzed to: (1) compare the strength of thresholds identified in three dimensions to the strength of thresholds considered individually in two dimensions, and (2) describe potential interactions between meteorological factors.

5.2 Methods

5.2.1 Study sites and rainfall-runoff event characterization

This study focused on sixteen catchments: eight sub-catchments of the Mahurangi River Catchment in New Zealand (MRC1-MRC8) and eight nested catchments located within the HJ Andrews Experimental Forest in the United States (HJA1-HJA8). Both regions have relatively humid climates and the catchments included in this study cover a range of drainage areas (MRC: 0.51 – 24.80 km^2 and HJA: 0.13 – 62.42 km^2). Catchments of the MRC are primarily rangeland and forested, while catchments of the HJA are forested. These sites have been the subject of previous hydrological research and have been described extensively by others (McKee & Druliner, 1998; Woods et al., 2013). Rainfall-runoff event delineation was performed on five-year records of rainfall, discharge, and temperature for each site using the HydRun toolbox in MATLAB® (Tang & Carey, 2017). Event response was characterized by total event runoff (Q_{TOT}). Event-specific meteorological factors were also derived. Factors that quantify rainfall depths include the total event rainfall (R_{TOT}) and the sum of R_{TOT} and 7-day antecedent rainfall ($R_{TOT} + AR_7$). The average event rainfall intensity (RI_{AVG}) was calculated to quantify the rate of water added to a catchment during an event. To account for some hydrologic abstractions and their effect on catchment response, antecedent evapotranspiration over a 7-day antecedent period was calculated. Additional details regarding the chosen response metric and meteorological factors, along with descriptions of response dynamics for the MRC and HJA sites, are available in Ross et al. (2019).

5.2.2 Threshold strength computations

To quantify and compare the abruptness of changes in hydrologic responses identified from 2D and 3D plots, one consistent method had to be selected to model event-specific point data as continuous curves in 2D, and as continuous surfaces in 3D. Locally weighted polynomial regression (LWPR) was chosen, as it approximates the underlying regression function of nonlinear relationships (Cleveland & Devlin, 1988). LWPR has previously been used in change point analyses for water quality data (e.g., Huang et al., 2017) and for estimating lake volumes (Lall et al., 2006). In total, 64 response curves (2D) illustrating Q_{TOT} as a function of a single meteorological factor were modelled using LWPR. Similarly, 64 response surfaces (3D) were modelled using LWPR to illustrate Q_{TOT} as a function of rainfall depth and rainfall intensity or rainfall depth and antecedent potential evapotranspiration. Both the response curves and response surfaces that were modelled using LWPR were controlled by parameters that establish the degree of the polynomial, the size of the neighbourhood of data points used for each local polynomial fit, and the weight of each data point based on their distance from the estimation point (Cleveland & Devlin, 1988; Rajagopalan & Lall, 1998). There are numerous methods to optimize parameter selection for LWPR, including ordinary and generalized cross-validation or the finite prediction error (Lall et al., 2006). Rather than performing parameter optimization for any single relationship or site, curves and surfaces were modelled for all sites via LWPR using a consistent parameter set to facilitate the comparison of response curve and response surface characteristics. A local linear polynomial regression model (i.e., degree of 1) was used, which is typical of LWPR applications (Huang et al., 2017). The neighbourhood size controls the degree of smoothing and a relatively modest neighbourhood size comprising 25% of the available data

points was selected. This value was selected to avoid under- or over-smoothing that is associated with smaller or larger neighbourhoods, respectively. Lastly, the default tri-cubic weighting function was adopted to establish the relative importance of data points based on their distance from the estimation point. Response curves and surfaces constructed using LWPR were not extrapolated beyond the available data. The fit of modelled curves and surfaces relative to original point data was assessed using the coefficient of determination (R^2), which is the square of the sample Pearson correlation coefficient between the observed data and the predicted values. Only curves and surfaces with moderate to strong R^2 values ($R^2 > 0.45$) were deemed adequate and retained for further analysis.

Quantitative and qualitative approaches were used to evaluate and compare response curves and response surfaces. To address the first research objective, the abruptness of response changes across entire response curves and response surfaces was quantified using a statistical parameter called threshold strength, which was computed using the procedure described in Lintz et al. (2011). The following procedure outlines the computation of threshold strength for 3D response surfaces. Threshold strength is the product of the bimodality of the frequency distribution of response captured by the standard deviation (σ_Z) and monotonicity (M):

$$\text{Threshold strength} = 2\sigma_Z M. \qquad \text{Equation 5-1}$$

The standard deviation of the response is as follows:

$$\sigma_Z = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(Z_i - \bar{Z})^2}, \qquad \text{Equation 5-2}$$

where N is the number of points on the gridded response surface. As in Lintz et al. (2011), the denominator in Equation 5-2 is N rather than N-1, as the standard deviation is used to describe shape rather than a population sample. Monotonicity is computed using a moving window of nine adjacent points on a 3D response surface discretized into an equally spaced grid (100 by 100

increments). At each grid position, the window comprises four pairs of opposing vectors that share a common center point: using cardinal directions, those vectors can be called NS, NESW, EW, and NWSE (Figure 5-1). Vectors depart from monotonicity if the vector endpoints are both above or below the center point, and departure from monotonicity for each vector is determined as follows:

$$NS_{i,j} = \min\left\{\left|Z_{i+1,j+2} - Z_{i+1,j+1}\right|, \left|Z_{i+1,j} - Z_{i+1,j+1}\right|\right\}, \quad \text{Equation 5-3}$$

$$NESW_{i,j} = \min\left\{\left|Z^{*}_{i+2,j+2} - Z_{i+1,j+1}\right|, \left|Z^{*}_{i,j} - Z_{i+1,j+1}\right|\right\}, \quad \text{Equation 5-4}$$

$$EW_{i,j} = \min\left\{\left|Z_{i,j+1} - Z_{i+1,j+1}\right|, \left|Z_{i+2,j+1}\right|\right\}, \quad \text{Equation 5-5}$$

and

$$NWSE_{i,j} = \min\left\{\left|Z^{*}_{i,j+2} - Z_{i+1,j+1}\right|, \left|Z^{*}_{i+2,j} - Z_{i+1,j+1}\right|\right\} \quad \text{Equation 5-6}$$

where Z is a point within a window, and i and j are indices of the point position on the uniform surface grid. For each window, diagonally oriented vectors are shortened via interpolation (represented as Z^{*}) to create a circular window. The departure from monotonicity for each window is determined by summing the vector departures from monotonicity. Similarly, the departure from monotonicity of an entire response surface (S) is determined by summing all of the window departures:

$$S = \sum_{i=1}^{n-2}\sum_{j=1}^{n-2}\left(NS_{i,j} + NESW_{i,j} + EW_{i,j} + NWSE_{i,j}\right) \quad \text{Equation 5-7}$$

where n is the number of points in each grid axis. The average departure from monotonicity for a response surface (K) is calculated by dividing S by the number of paired, opposing vectors for the surface. A negative exponential function of K is used to calculate surface monotonicity (M):

$$M = e^{-950K}. \quad \text{Equation 5-8}$$

As in Lintz et al. (2011), an exponential coefficient of 950 was selected to ensure that the lower range of M approaches zero.

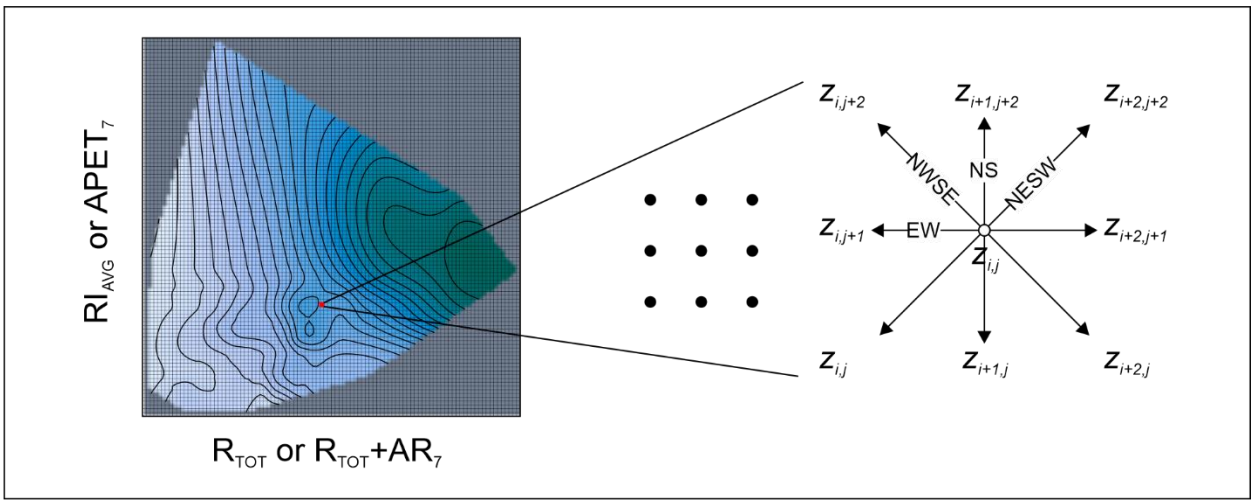


Figure 5-1. The departure from monotonicity is calculated from a gridded three-dimensional surface. For each grid cell, a moving window of nine adjacent points comprises four pairs of opposing vectors that share a common center point. Figure adapted from Lintz et al. (2011).

While the procedure described above outlines the computation of threshold strength for 3D response surfaces, a comparable procedure was used to determine the threshold strength of 2D response curves. Specific computational details for 2D threshold strength can be found in Lintz et al. (2011). Values of threshold strength range from 0 to 1: a value of zero suggests that no threshold is present, while non-zero values are indicative of threshold-mediated responses that increase according to the abruptness of response change (Lintz et al., 2011). The percent difference between the threshold strength of 2D response curves and that of 3D response surfaces featuring a common meteorological factor was computed, to compare threshold strength in two and three dimensions. Such comparisons are legitimized by previous work that demonstrated equivalence between the threshold strength of 2D response curves and 3D response surfaces generated from the same function (Lintz et al., 2011). A negative percent difference value indicates that the threshold strength of a 2D curve is smaller than the threshold strength of the

comparable 3D surface. Since there were a variety of catchment drainage areas, relationships between threshold strength values and drainage area were assessed using Spearman's rank correlation analysis. The Mann-Whitney test (McGrew & Monroe, 1993) was also applied to identify statistically significant differences in threshold strength between the two geographic regions considered in this study, i.e., MRC catchments versus HJA catchments.

5.2.3 Characterization of meteorological factor effects

While the threshold strength parameter quantifies the abruptness of changes in response curves and response surfaces, it does not quantify the extent to which factor interactions determine nonlinear response in three dimensions. Documenting the presence or absence of such interactions is, however, critical in multi-factor relationships. Hence, concerning the second research objective, multi-factor interactions were assessed using contour plots derived from 3D response surfaces. These contour plots were used as qualitative diagnostic tools to interpret meteorological factor effects on hydrologic response; however, they are not used to evaluate threshold strength. It should be noted that the contour maps considered in the present study may not be interpreted in the same manner as topographic maps. Indeed, with topographic maps, the x and y axes both illustrate independent geographic coordinates expressed in similar units while the contours depict changes in elevation. In the present study, however, the x-axes are associated with rainfall depths and the y-axes are associated with rainfall intensity or antecedent potential evapotranspiration, which do not necessarily share a common unit of measurement and are not necessarily independent from one another, from a physical (process) standpoint. The interpretation of contour plots was therefore done according to principles described by Antony

(2014) for the specific purpose of evaluating potential relationships between two factors that contribute to a modelled response surface. Contour lines that are straight and oriented perpendicularly to any axis are illustrative of a response being mostly influenced by a single factor (i.e., main effects), specifically the factor illustrated by the axis that contour lines are perpendicular to. Conversely, curved contour lines are indicative of a response being strongly determined by the interactions between the two factors underlying the 3D response surface. Contour plots were therefore classified based on their dominant contour line shape, i.e., straight (S), mostly straight (MS), mostly curved (MC) and curved (C), to allow for quick comparisons across sites.

5.3 Results

5.3.1 Threshold strength in 2D response curves and 3D response surfaces

Across all sites, response curves, and surfaces modelled using LWPR had R^2 values ranging from 0.09 to 0.91 (median: 0.51). Based on the chosen $R^2 > 0.45$ cut-off, 24 of 32 curves featuring a rainfall depth, 1 of 16 curves featuring $RI_{AVG,}$ and 60 of 64 surfaces were deemed adequate and retained for further analysis. Threshold strength values for those retained response curves and response surfaces are reported in Table 5-1. Across all sites, the range of threshold strength values for 2D curves was 0.19 – 0.71 (median: 0.57), while the range of threshold strength values for 3D surfaces was 0.35 – 0.68 (median: 0.48). Select examples of response curves and response surfaces along with their threshold strength values are featured in Figure 5-2: they highlight differences between weaker thresholds associated with smoother changes in

response (Figure 5-2A and Figure 5-2C) and stronger thresholds associated with sharper changes in response (Figure 5-2B and Figure 5-2D). Correlation coefficients, or *rho* values, between threshold strength and drainage area were low and not statistically significant, regardless of whether 3D surfaces (*rho* = -0.25, *p-value* = 0.06) or 2D curves featuring a volume factor (*rho* = 0.36, *p-value* = 0.08) were considered. The Mann-Whitney test also revealed that there was no statistically significant difference in threshold strength between the HJA catchments and the MRC catchments, regardless of whether 3D surfaces (*p-value* = 0.83) or 2D curves featuring a volume factor (*p-value* = 0.11) were considered. Too few 2D curves featuring RI_{AVG} or $APET_7$ were retained after LWPR, thereby preventing the evaluation of potential correlations between threshold strength and drainage area, or the assessment of regional differences in threshold strength.

Table 5-1. Threshold strength values for the 2D response curves and the 3D response surfaces with $R^2 > 0.45$ that were modelled in this study.

	2D Threshold Strength				3D Threshold Strength			
	R_{TOT}	$R_{TOT}+AR_7$	RI_{AVG}	$APET_7$	R_{TOT}/RI_{AVG}	$R_{TOT}/APET_7$	$R_{TOT}+AR_7/RI_{AVG}$	$R_{TOT}+AR_7/APET_7$
MRC1	0.34	-	-	-	0.43	0.42	-	0.46
MRC2	0.19	0.19	-	-	0.48	0.52	0.42	0.41
MRC3	0.51	0.58	0.19	-	0.38	0.44	0.57	0.51
MRC4	0.57	-	-	-	0.51	0.55	0.55	0.42
MRC5	0.62	-	-	-	0.59	0.44	0.62	0.48
MRC6	0.71	-	-	-	0.59	0.68	0.50	0.46
MRC7	0.45	-	-	-	0.35	0.36	0.45	0.53
MRC8	0.53	0.50	-	-	0.46	0.47	0.56	0.49
HJA1	0.69	0.61	-	-	0.66	0.43	0.59	0.44
HJA2	0.57	0.64	-	-	0.50	0.44	0.57	0.42
HJA3	0.56	0.45	-	-	0.47	0.47	0.48	0.45
HJA4	0.55	-	-	-	0.49	0.42	0.56	0.56
HJA5	0.58	-	-	-	0.47	0.42	0.51	-
HJA6	0.60	0.60	-	-	0.48	0.44	0.48	0.48
HJA7	0.58	-	-	-	0.50	0.53	-	-
HJA8	0.61	0.50	-	-	0.49	0.50	0.39	0.48

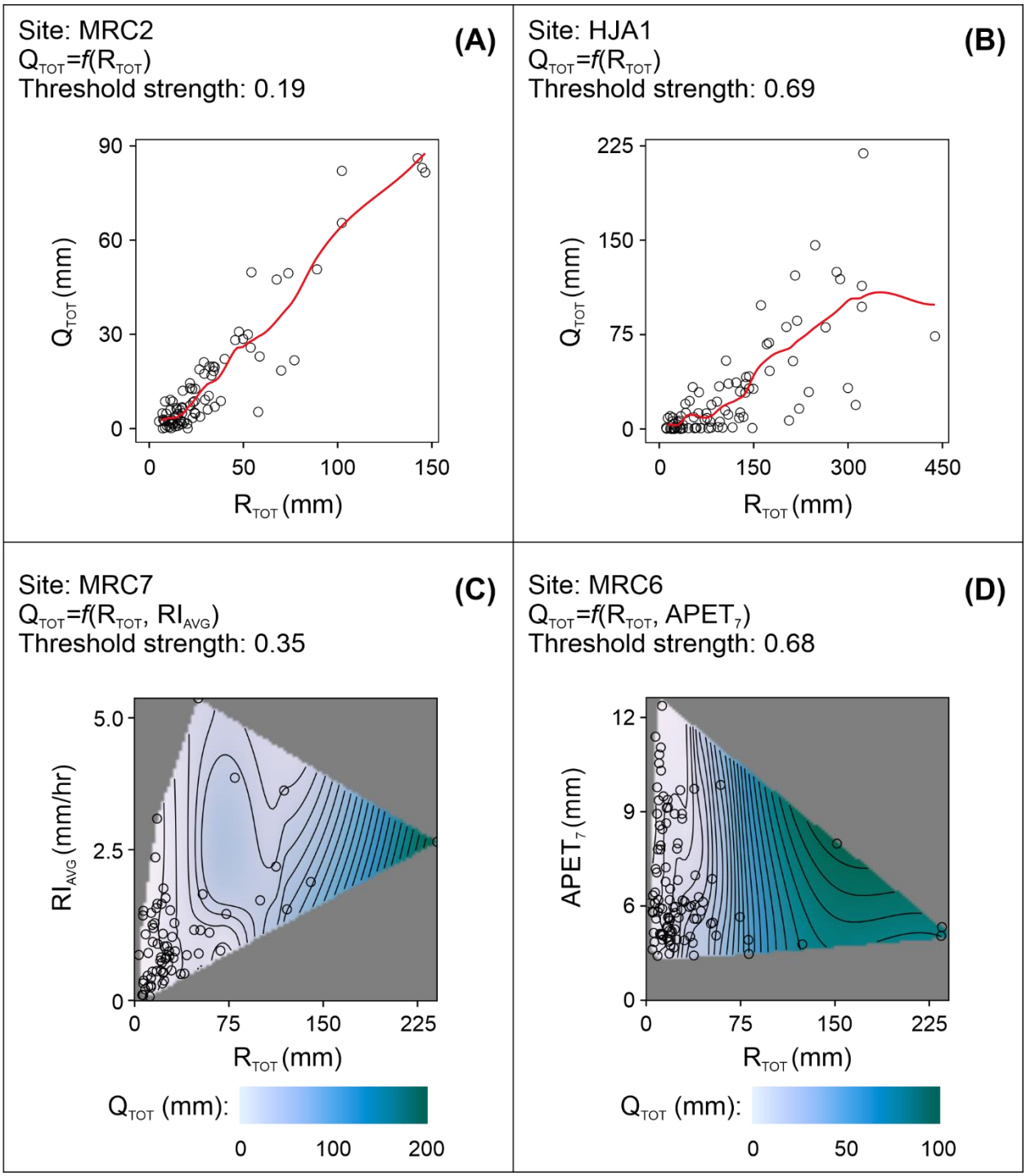


Figure 5-2. Select curves and surfaces modelled using LWPR that yielded low and high threshold strength values. Examples of curves with low and high threshold strength are featured in panels (A) and (B), respectively. Likewise, examples of surfaces with low and high threshold strength are featured in panels (C) and (D), respectively.

The percent differences in threshold strength between 2D curves and 3D surfaces sharing a common meteorological factor are reported in Table 5-2. In most cases, the threshold strength calculated for a 2D curve was greater than the threshold strength estimated for a 3D surface sharing a common meteorological factor, as indicated by positive percent differences. Of the 32 surfaces featuring R_{TOT}, 28 of those surfaces had weaker thresholds than response curves featuring R_{TOT}. However, this observation was not consistent across all sites and factors: in 10 cases, the percent difference in threshold strength between 2D curves and 3D surfaces that were constructed using the same meteorological factor was negative, which indicates a stronger threshold in a 3D surface compared to a 2D curve. The percent difference between threshold strength values for 2D curves featuring R_{TOT}+AR_7 and 3D surfaces including R_{TOT}+AR_7 and RI_{AVG} was negative for three of eight sites. Similarly, the percent difference between the response curve featuring RI_{AVG} and response surfaces featuring RI_{AVG} and R_{TOT} or R_{TOT}+AR_7 was negative. Additionally, there was one case where the threshold strength was nearly equal between a 2D curve and a 3D surface: at the HJA3 site, the threshold strength percent difference between the 2D curve featuring R_{TOT}+AR_7 and the 3D surface featuring R_{TOT}+AR_7 and $APET_7$ was only 1.0%.

Table 5-2. Percent differences between the threshold strength of 2D response curves and the threshold strength of 3D response surfaces that share a common meteorological factor.

	R_{TOT} / RI_{AVG}		R_{TOT} /$APET_7$		R_{TOT}+AR_7 / RI_{AVG}		R_{TOT}+AR_7 / $APET_7$	
	R_{TOT}	RI_{AVG}	R_{TOT}	$APET_7$	R_{TOT}+AR_7	RI_{AVG}	R_{TOT}+AR_7	$APET_7$
MRC1	-21.9	-	-19.0	-	-	-	-	-
MRC2	-60.2	-	-63.5	-	-54.4	-	-53.5	-
MRC3	35.2	-50.3	16.6	-	2.1	-67.0	12.4	-
MRC4	10.4	-	2.9	-	-	-	-	-
MRC5	5.2	-	39.0	-	-	-	-	-
MRC6	18.9	-	4.1	-	-	-	-	-
MRC7	26.2	-	22.6	-	-	-	-	-
MRC8	15.4	-	12.4	-	-10.3	-	2.4	-
HJA1	5.1	-	61.6	-	2.9	-	38.2	-
HJA2	13.8	-	29.6	-	12.9	-	51.0	-
HJA3	17.3	-	18.7	-	-5.6	-	1.0	-
HJA4	12.8	-	32.2	-	-	-	-	-
HJA5	21.4	-	38.4	-	-	-	-	-
HJA6	25.9	-	36.7	-	23.2	-	24.1	-
HJA7	15.1	-	8.2	-	-	-	-	-
HJA8	24.7	-	23.0	-	28.2	-	3.9	-

5.3.2 Interactions (and lack thereof) between meteorological factors

Concerning 3D response surfaces, classifying contour plots as S, MS, MC or C showed that only six surfaces could be adequately described by a single class. To perform a classification of contour line patterns while considering the heterogeneity present in 3D space, contour plots were divided into four quadrants representing low rainfall depth and low RI_{AVG} or $APET_7$ values (quadrant 1), high rainfall depth and low RI_{AVG} or $APET_7$ values (quadrant 2), high rainfall depth and high RI_{AVG} or $APET_7$ values (quadrant 3), and low rainfall depth and high RI_{AVG} or $APET_7$

values (quadrant 4) (Figure 5-3A). Examples of quadrant-specific classifications of contour patterns are shown in Figure 5-3B. A total of 240 quadrants (i.e., 16 sites × 4 response surfaces per site × 4 quadrants per response surface, excluding 4 surfaces that were not retained for analysis) were analyzed across all sites: 72 quadrants were classified as C, 66 were classified as MC, 63 were classified as MS, and 3 were classified as S. Additionally, 39 quadrants had insufficient data for classification, as the constructed surfaces did not include extrapolation beyond the available data. The distribution of classes across quadrants was not uniform: quadrants were rarely classified as S, quadrant 1 was frequently classified as C, and quadrants 2, 3, and 4 were most frequently classified as MS or MC (Figure 5-4). The relative importance of straight and curved contour lines across surfaces resulted in the creation of linear, angular, or radial gradients. Linear gradients illustrate responses that increase progressively from low to high along a single axis or factor (Figure 5-5A); angular gradients illustrate responses that increase with both factors (Figure 5-5B), and radial gradients show hotspots of elevated response (Figure 5-5C). Close examination of all 60 response surfaces evaluated in this study (all plots not shown) revealed that 15 surfaces had linear gradients, 16 surfaces had angular gradients, and 29 surfaces had radial gradients. Response surfaces also revealed that in 3D, significant changes in runoff behaviour did not correspond to individual points where a break in slope is observed, as is the case in 2D. Rather, significant changes in runoff behaviour corresponded to planes or lines along discontinuities in contour plot colour, which we refer to here as threshold fronts (see yellow lines in Figure 5-6). It should be noted that the location of threshold fronts in Figure 5-6 was identified visually and is, therefore, approximate.

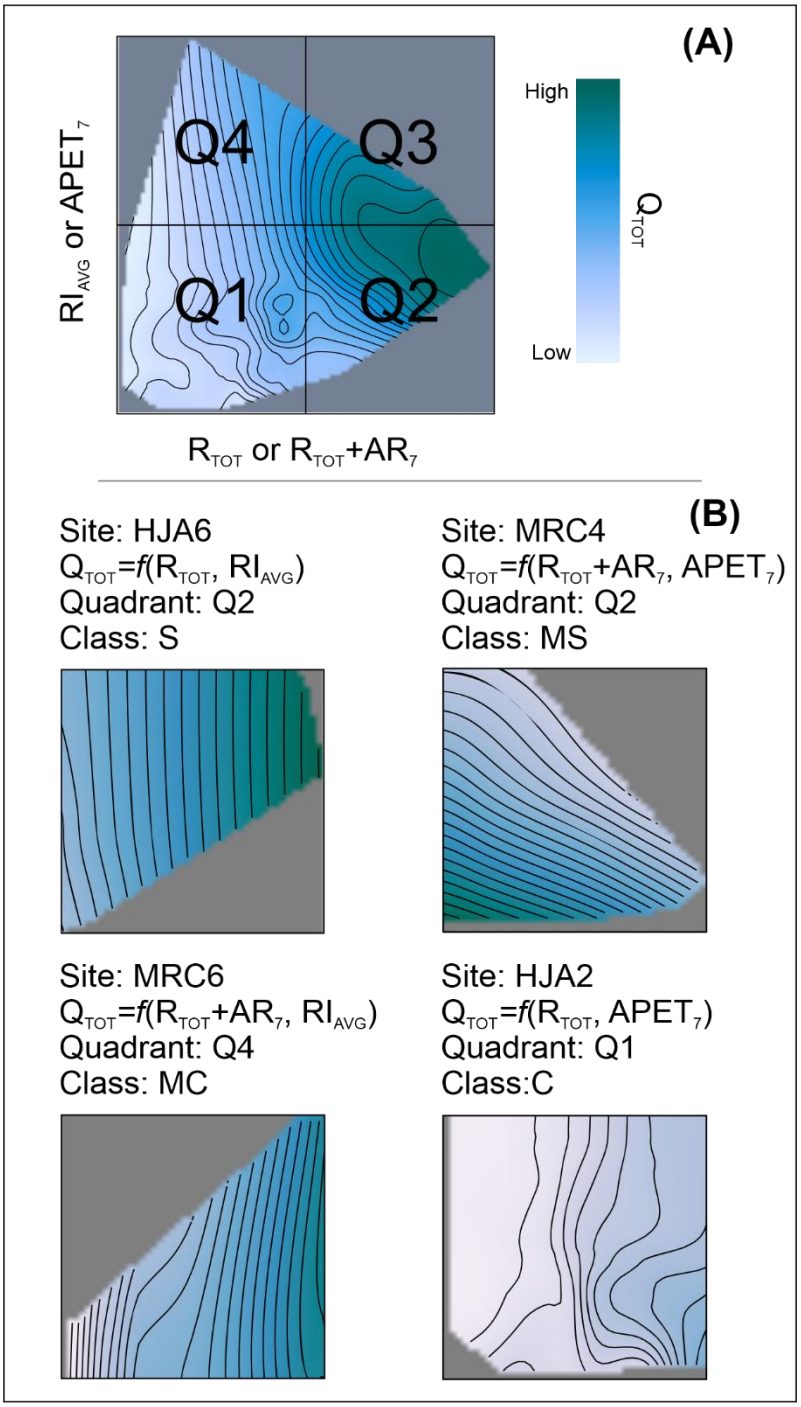


Figure 5-3. (A) Response surface separated into four quadrants. (B) Examples of contour pattern classifications (S: straight; MS: mostly straight; MC: mostly curved; C: curved). Axis ticks and labels are omitted for readability. Data points and axis ticks and labels are omitted for readability. See Appendix C-1 for a figure showing data points.

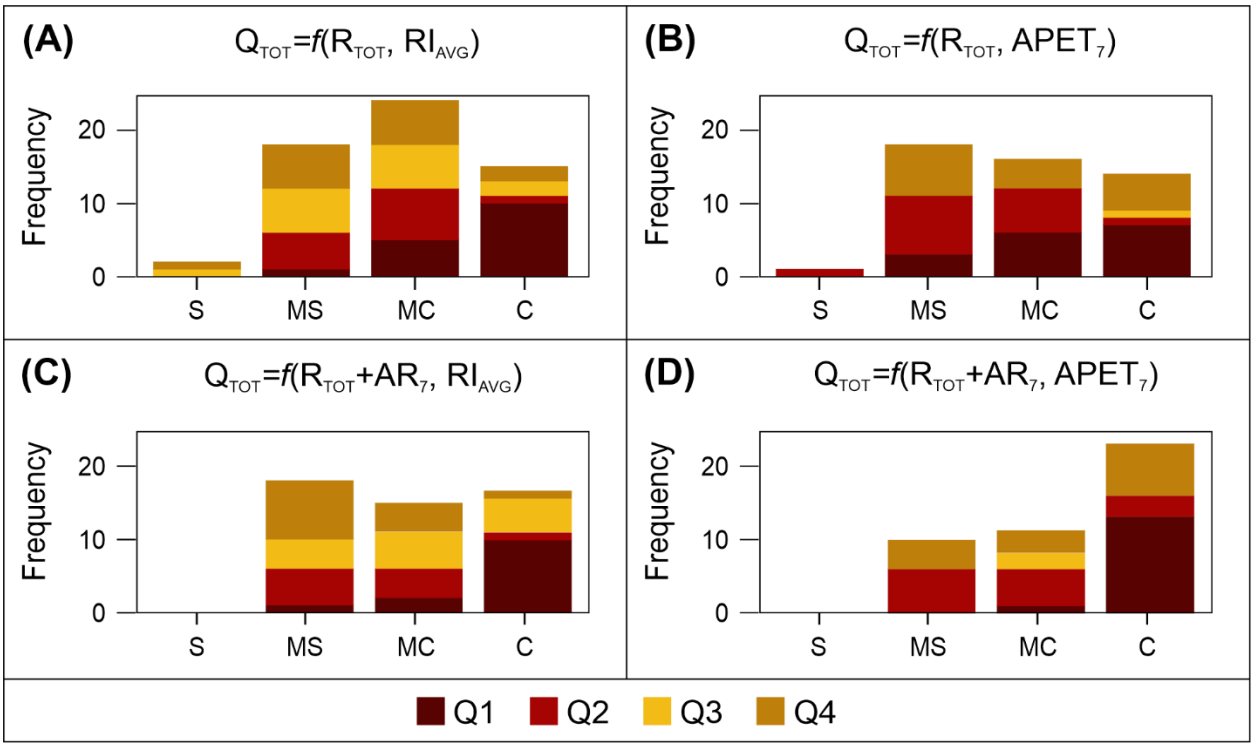


Figure 5-4. Frequency of contour pattern classes observed across all sites and summarized by quadrant for the 3D response surfaces evaluated in this study. Results are separated by relationship factors: (A) Q_{TOT} as a function of R_{TOT} and RI_{AVG}; (B) Q_{TOT} as a function of R_{TOT} and $APET_7$, (C) Q_{TOT} as a function of $R_{TOT}+AR_7$ and RI_{AVG}, and (D) Q_{TOT} as a function of $R_{TOT}+AR_7$ and $APET_7$.

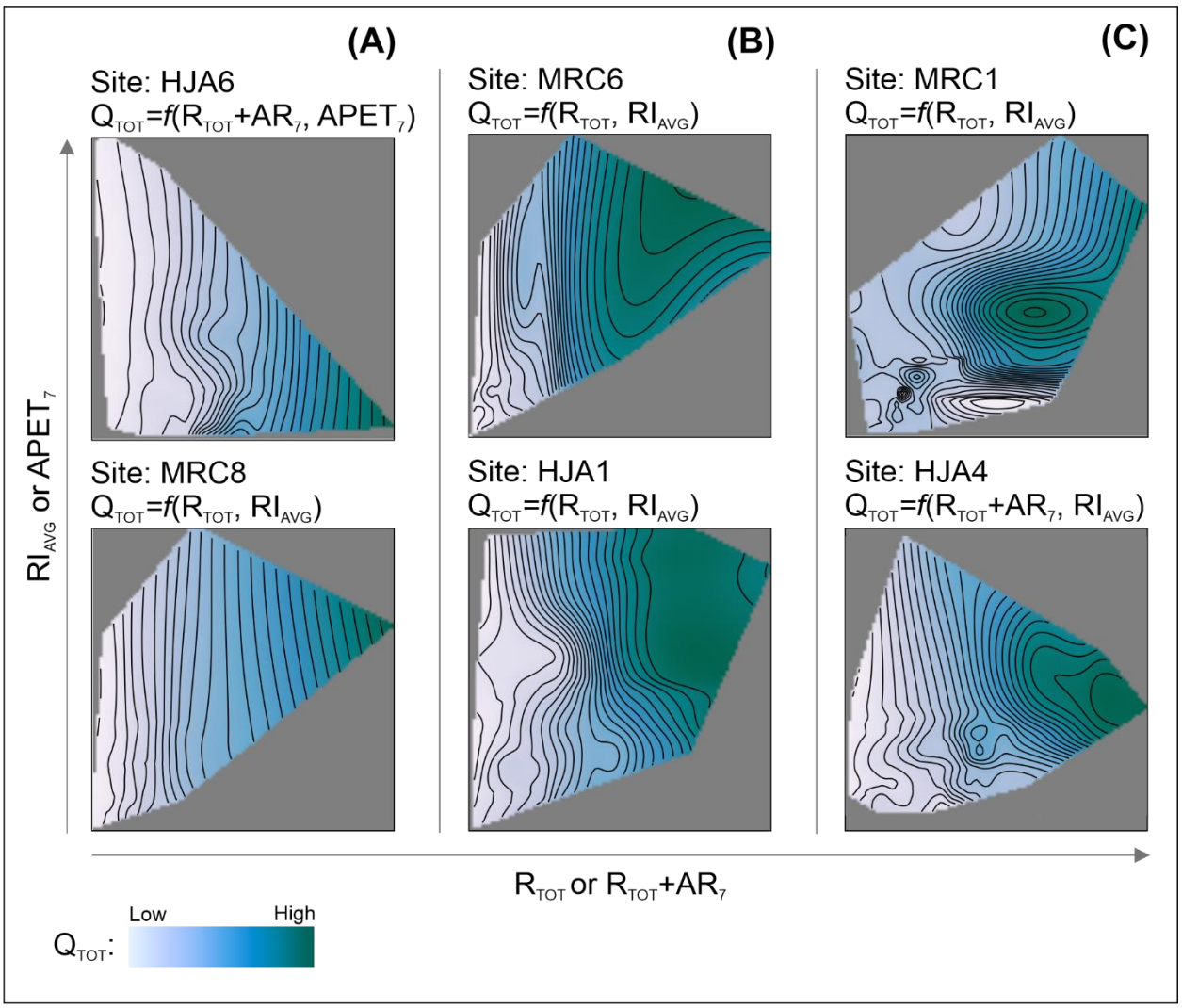


Figure 5-5. Examples of response surfaces modelled in this study: (A) linear gradient; (B) angular gradient; and (C) radial gradient. Axis ticks and labels are omitted for readability. Data points and axis ticks and labels are omitted for readability. See Appendix C-2 for a figure showing data points.

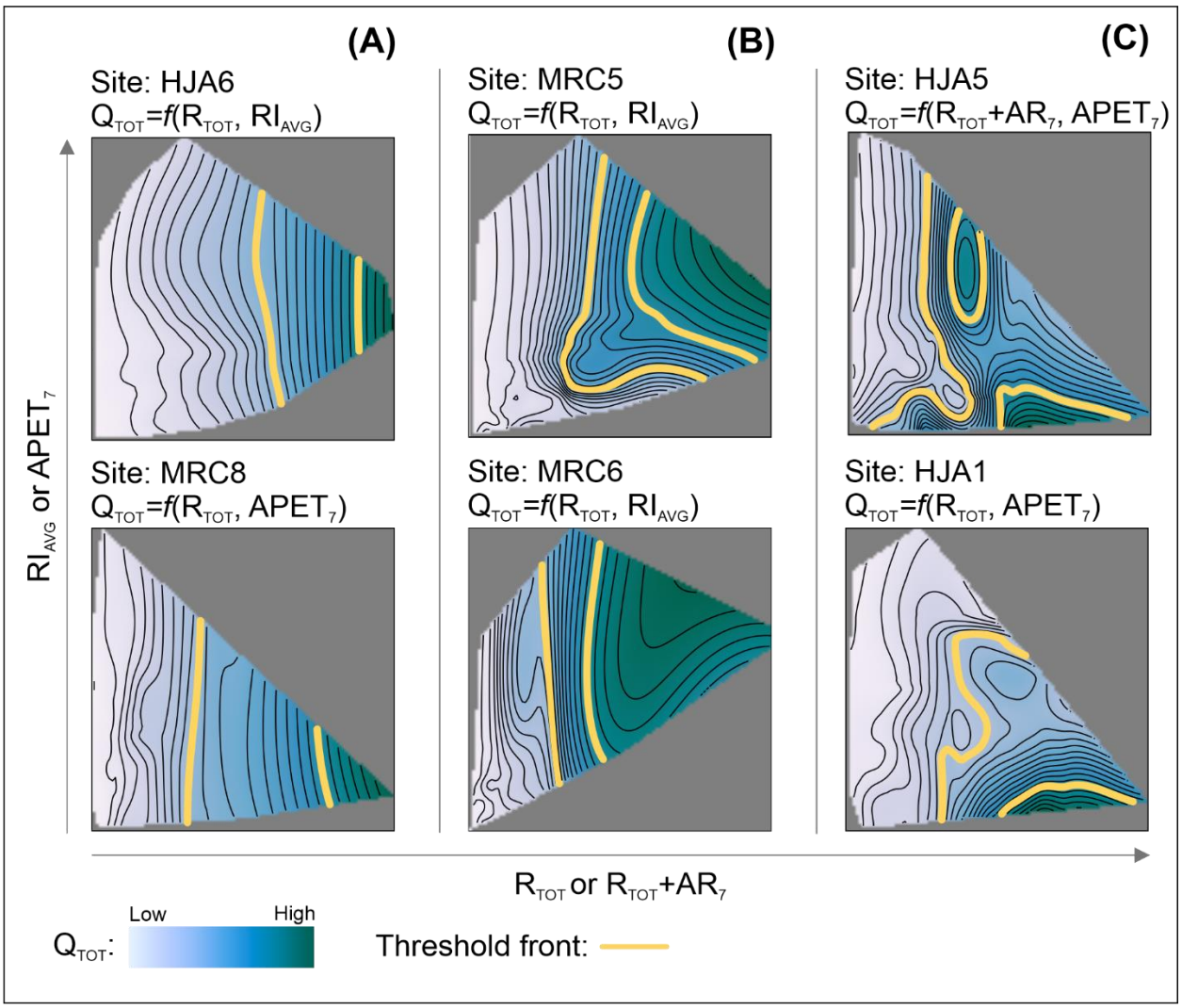


Figure 5-6. Examples of response surfaces modelled in this study with threshold fronts highlighted in yellow: (A) linear gradient; (B) angular gradient; and (C) radial gradient. Axis ticks and labels are omitted for readability. Data points and axis ticks and labels are omitted for readability. See Appendix C-3 for a figure showing data points.

5.4 Discussion

5.4.1 How does threshold strength vary among response curves and surfaces?

The non-zero values of the threshold strength parameter estimated in this study suggest that curves depicting Q_{TOT} as a function of rainfall depth, rainfall intensity, or antecedent evapotranspiration, and surfaces depicting Q_{TOT} as a function of both rainfall depth and rainfall intensity or rainfall depth and antecedent evapotranspiration, are threshold-mediated. Non-zero threshold strength values covered a wide range, though, from 0.19 to 0.71: as this parameter is sensitive to small changes in monotonicity and bimodality, it can be used to distinguish thresholds with smoother changes in response (Figure 5-2A and Figure 5-2C) from those with sharper changes in response (Figure 5-2B and Figure 5-2D), in a manner superior to what could be achieved visually. It is important to note that the response changes illustrated in Figure 5-2A and Figure 5-2B are more subtle than those which have typically been labelled as "thresholds" in previous hydrological studies. Indeed, characterizations of hydrologic thresholds are typically guided by diagnostic shapes (e.g., hockey-stick, step or Heaviside function, Dirac function, and sigmoid function) where pre- and post-threshold conditions are delimited by a single breakpoint, i.e., a point at which the slope of the input-output relationship changes (Ali et al., 2013). Conversely, the response curves illustrated in Figure 5-2A and Figure 5-2B – which are both labelled as threshold-mediated according to threshold strength computations – do not showcase a single, major breakpoint but rather multiple, smoother changes. This apparent discrepancy is not surprising, given inherent differences in threshold identification approaches. The recent hydrology literature (Oswald et al., 2011; Scaife & Band, 2017; Wei et al., 2020) almost

exclusively relies on breaks of slope to identify thresholds, which has led to a narrow range of possible nonlinear relationships (i.e., the aforementioned diagnostic shapes, see Ali et al., 2013) that consider thresholds in the form of steep, step-like transitions. The threshold strength parameter, however, does not merely evaluate break(s) in slope, but rather monotonicity and the presence of bimodal frequency distributions across an entire response curve or surface. The threshold strength parameter, therefore, allows for a much broader range of possible nonlinear relationships, from small undulations (yielding non-zero but low threshold strength values) to "staircases" and step-like features (yielding higher threshold strength values). The high sensitivity associated with the threshold strength parameter has been deemed useful in ecology, not only for when the goal is to explore a wide variety of nonlinear relationships, but also when the goal is to identify systems that are approaching step-like threshold behaviours and may exhibit them at a later time in response to new or cumulative stressors. Although similar goals could be pursued in hydrology, to date hydrologists have used identification methods that have been biased towards step-like thresholds. Lintz et al. (2011) tested 48 different potential shapes of nonlinear relationships and suggested that step-like threshold shapes were not distinguishable when threshold strength values were below 0.72. Future hydrological studies could, therefore, rely on the threshold strength parameter to address targeted questions related to how abrupt or "step-like" a change in response needs to be to warrant consideration, and whether nonlinear relationships that do not showcase step-like changes can also provide insights into runoff processes.

The results of this study also suggest that the ability to distinguish differences in threshold-mediated responses using threshold strength may be useful for evaluating the effects of antecedent conditions on response. In one previous study, gross precipitation-only thresholds

were not visually detected, but a combined gross precipitation plus antecedent soil moisture index threshold was visually detected at a humid continental site (Detty & McGuire, 2010). In the present study, while response was characterized as threshold-mediated in curves featuring R_{TOT} and in curves featuring R_{TOT}+AR_7, the changes in response for curves featuring R_{TOT} were comparatively smoother, based on the smaller threshold strength values. It is, therefore, likely that response thresholds that consider rainfall depth but not antecedent conditions, and that were associated with lower threshold strength values, would go undetected during a visual examination. It is worth noting that numerous response curves modelled using LWPR were not deemed adequate, based on low R^2 values, and were therefore not analyzed in terms of threshold strength. While only 8 of 32 curves featuring rainfall depth were discarded because of a poor LWPR fit (i.e., $R^2 < 0.45$), 15 out of 16 curves featuring RI_{AVG} and all 16 curves featuring $APET_7$ were discarded. This suggests that for the catchments featured in this study, rainfall depth considered in insolation may influence Q_{TOT}, but that is not the case for rainfall intensity or antecedent potential evapotranspiration considered in isolation. This observation is in agreement with other studies that indicated that rainfall intensity and potential evapotranspiration exert little to no influence on hydrologic response in humid environments (Graham et al., 2010; Tromp-van Meerveld & McDonnell, 2006a).

One objective of the current study was to compare the strength of dual-factor thresholds identified in 3D to that of individual rainfall depth, rainfall intensity, and antecedent potential evapotranspiration thresholds in 2D. The predominance of positive percent differences (Table 5-2) implies that thresholds identified in 2D response curves were stronger than thresholds in 3D response surfaces that shared a common rainfall depth factor. One potential explanation for this is that differences in the distribution of data points used to construct response curves and

response surfaces led to smoother transitions between data points in 3D than in 2D. Potential bias in the construction of response curves and response surfaces was mitigated by using the same LWPR parameterization in two- and three-dimensions. One other possibility is that hydrologic response may be mostly controlled by a single factor. Subsequently, the consideration of multiple factors in those cases may result in a positive (or near 0) percent difference between 2D and 3D threshold strength. That said, 2D threshold strength was not always greater than 3D threshold strength: in 10 cases, the percent difference between 2D and 3D thresholds was negative, suggesting that the threshold strength of some of the observed nonlinear hydrologic responses was increased by considering both rainfall depth and rainfall intensity or antecedent potential evapotranspiration. Cases with 3D thresholds that were stronger than 2D thresholds sharing a common factor were primarily observed at MRC sites (9 of 10 cases). Furthermore, most cases (9 of 10) exhibiting negative percent difference occurred at small headwater catchments (i.e., drainage area < 2.5 km^2). Those results could mean that the dual consideration of multiple meteorological factors for characterizing threshold dynamics is critical in headwater catchments but less important in larger catchments. While this hypothesis could not be tested in a statistically robust manner in the present study, due to the relatively small number of cases with negative percent difference, it is supported by existing literature on headwater catchment dynamics, with multiple studies that have addressed the effects of both water volumes and rainfall intensities on response. For example, the depth of rainfall and the rainfall intensity has been deemed important control factors on peak discharge and total runoff for a small Mediterranean headwater catchment in the Central Spanish Pyrenees (Seeger et al., 2004). Similarly, one study focusing on humid headwater catchments in India showed that the depth of rainfall and rainfall intensity can strongly influence the size of the contributing area during

runoff events (Putty & Prasad, 2000). Others have also shown that rainfall depth and rainfall intensity are linked to the response characteristics of flash floods and debris flow in headwater catchments (Borga et al., 2014).

5.4.2 To what extent do meteorological factors interact?

This study also relied on the shape of contour lines associated with 3D response surfaces to describe interactions between rainfall depths and RI_{AVG} or $APET_7$. In general, quadrant-specific classifications showed that Q_{TOT} was affected by factor interactions, as evidenced by curved or mostly curved patterns observed in at least one quadrant (Figure 5-4). Quadrant-specific classifications were useful in revealing that the presence/absence of factor interactions are dependent on the range of meteorological factor values considered. Quadrant 1 was mostly classified as C, which may imply that factor interactions are strongly present when both factor values are low. Conversely, quadrants 2 and 4 were commonly classified as MS: those two quadrants have one low factor value and one high factor value, which is intuitively consistent with a response pattern being mostly affected by a single factor. Lastly, in quadrant 3, where both factor values are high, no dominant class (or dominant contour line shape) was identifiable, suggesting a low degree of predictability of control factor dynamics in those conditions.

A comparison of Figure 5-3B with Figure 5-5A, Figure 5-5B, and Figure 5-5C provides information for interpreting different 3D response surface features (i.e., contour line shape and gradient type) in terms of factor effects. Indeed, linear gradients had mostly straight contour lines oriented perpendicular to an axis, which is indicative of main effects being exerted by a single factor (Antony, 2014). Notably, all the quadrants classified as S and MS had contour lines

oriented vertically or sub-vertically, which indicates that these responses are mainly affected by rainfall depth. Not a single quadrant classified as S or MS exhibited response mainly affected by RI_{AVG} or $APET_7$ (i.e., contour lines oriented horizontally or sub-horizontally). Conversely, angular and radial gradients had mostly curved and curved contour lines, which indicates factor interactions (Antony, 2014).

Response gradient type and contour line shape also have implications for predicting hydrologic response as a function of two meteorological factors. Indeed, hydrologic response predictability would undoubtedly be the highest when linear gradients and mostly straight or straight contour lines are present. Such graphical features illustrate the fact that hydrologic response is predominantly influenced by a single factor. In such a scenario, one may question the need for a three-dimensional approach. One advantage of the three-dimensional approach is that it does not require *a priori* information about which of the two considered factors is dominant. For angular response gradients, the predictability of hydrologic response is complicated by the diversity of possible surface features. In the present study, surfaces with angular response gradients were associated with quadrants classified as MS, MC, and C. Additionally, the relative influence of individual factor effects on response is unknown for angular gradients. In this study, the orientation of angular response gradients relative to the x and y axes were variable, demonstrating that the effects of individual factors on response was unequal. The difficulty in determining the exact gradient orientation increases the difficulty of predicting hydrologic response. Finally, hydrologic responses with radial response gradients would likely be most difficult to predict. As observed in this study (Figure 5-5C and Figure 5-6C), surfaces with radial response gradients have hotspots of elevated response. The difficulty in predicting the location of hotspots in a response surface could be compared to the difficulties associated with identifying

global maxima in parameter distributions in the context of hydrologic modelling (Kavetski et al., 2006). The identification of global maxima in parameter space has been pursued using a range of optimization techniques, including particle swarm and differential evolution (Gill et al., 2006; Zhang et al., 2009). It is unclear whether similar optimization techniques could be used to further characterize hydrological responses showcasing radial gradients. Additional research is warranted, as the underlying physical processes that drive factor interactions and the formation of response hotspots are not well understood, and simple yet robust methods for predicting hotspot geometry (e.g., circular or oblong) have not been identified.

The yellow lines – or threshold fronts – displayed in Figure 5-6 suggest that when 3D response surfaces are examined, several combinations of critical rainfall depths and RI_{AVG} or $APET_7$ can lead to significant changes in runoff behaviour when exceeded. Historically, research efforts targeting the characterization of nonlinear hydrologic response relied on 2D response curves as well as the premise of a single critical threshold value (e.g., Ali et al., 2015; Detty & McGuire, 2010; Tromp-van Meerveld & McDonnell, 2006a). However, recent studies have identified multiple thresholds in a single rainfall-runoff relationship and linked them to process-based interpretations (e.g., Scaife & Band, 2017; Wei et al., 2020). The latter studies are more aligned with the threshold fronts showcased in the present study. Conceptual similarities can be drawn between threshold fronts and the notion of equifinality in hydrologic modelling, i.e., the possibility that multiple unique parameter sets may lead to equally plausible simulations of reality (Beven, 2006, 2011). While the principle of equifinality is used to illustrate the complexity of evaluating unique parameter sets associated with behavioural simulations, the notion of threshold fronts could similarly be used to demonstrate the range of process dynamics that trigger changes in hydrologic response. Consequently, two plausible hypotheses related to

threshold fronts are suggested here and could be tested in the future within a modelling framework, namely that: 1) the length of a threshold front is positively correlated with the number of thresholds driving the transformation of precipitation into runoff, and 2) the complexity of the threshold front shape is inversely related to how easily individual threshold values can be identified or predicted.

One step towards testing those hypotheses could be the construction of 3D response surfaces from synthetic data that is continuous across the entire surface domain. This data would allow for 2D cross-sections to be taken from the 3D surfaces at specific factor values to evaluate the relationship shape and the presence of multiple thresholds. This exercise was not possible in this study, since surfaces constructed using LWPR were modelled from point data that varied in availability and density across the surface domain. Such an exercise would also allow the influence of LWPR parameterization and data features like quantity, distribution, accuracy, and noise on response surface features, including contour line geometry, to be further scrutinized. This may help resolve some of the uncertainty associated with qualitative interpretations of surface features, including the reproducibility of features within domain subsets, and interactions among multiple meteorological factors. While this study employed a relatively conservative approach by assessing generalized response surface patterns, like dominant contour line shape, stronger consideration of these methodological limitations is needed to advance future attempts to characterize hydrologic thresholds in three dimensions.

5.4.3 Do underlying factor interactions determine response threshold strength?

While answering whether underlying factor interactions determine response threshold strength was not a primary goal of this study, it is a question worth examining, as it may contribute insights towards the predictability of hydrologic response. Many of the response surfaces analyzed in this study are characterized by curved contour lines that are a product of factor interactions, which raises the question of whether factor interactions lead to stronger thresholds in hydrologic response. Figure 5-7 shows that there is no clear difference in the distribution of threshold strength between surfaces comprising quadrants mostly classified as MC or C, and surfaces that have one or no quadrant classified as MC or C. Those results are consistent with literature surrounding the development of the threshold strength parameter. Potential linkages between threshold strength and factor interactions can be further evaluated by quadrant (Figure 5-8). Surfaces with a classification of C or MC in quadrant 1 tended to result in lower median threshold strength values, compared to surfaces with quadrant 1 classified as S or MS (Figure 5-8C). Similarly, median threshold strength values were lower for surfaces with quadrant 4 classified as C or MC, as opposed to S or MS (Figure 5-8A). However, in contrast, no significant differences in threshold strength were discernable between different classifications for quadrant 2 and quadrant 3 (Figure 5-8B and Figure 5-8D). This might suggest that the threshold strength of an entire surface may be more sensitive to the presence (or absence) of factor interactions at low rainfall depths, than at high rainfall depths. While the number of surfaces included in this study is too small to conduct robust statistical analyses to determine potential differences in threshold strength controlled by the dominant contour line shape in specific quadrants, Figure 5-8C suggests that factor interactions at pre-threshold, low rainfall depth

conditions may influence hydrologic thresholds. This is consistent with other studies that have evaluated the influence of initial conditions on hydrologic response dynamics. For example, in a study conducted in Germany, initial soil moisture microstates were deemed important for determining hydrologic response and response predictability (Zehe & Blöschl, 2004). Specifically, that study found that the predictability of rainfall-runoff event response increased when elevated initial soil moisture conditions were combined with more intense event rainfall (Zehe & Blöschl, 2004). The importance of factor interactions and response dynamics at low factor values for determining hydrologic thresholds – as alluded to in the current study – is also consistent with other work that identified initial conditions as influential controls on both the timing and spatial pattern of runoff (Mirus & Loague, 2013).

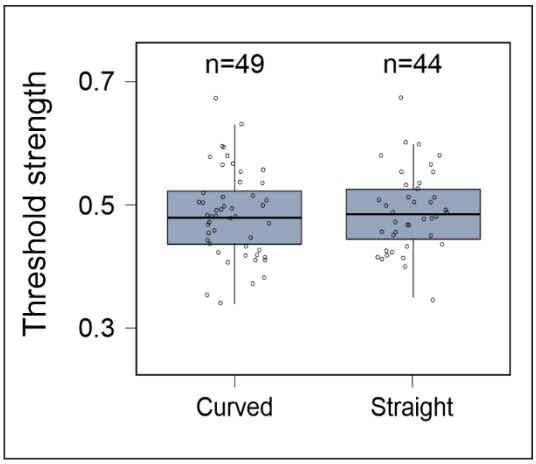


Figure 5-7. Boxplots showing the distribution of threshold strength for surfaces with two or more quadrants classified as MC or C (curved), and surfaces with one or no quadrants classified as MC or C (straight). The number of surfaces designated as curved or straight is denoted by n. Horizontal black lines in boxplots are representative of median values, whereas each box spans from the 25^{th} to 75^{th} percentiles. Lower and upper whiskers extend from the median value to 1.5 times the inter-quartile range. Outliers are not shown.

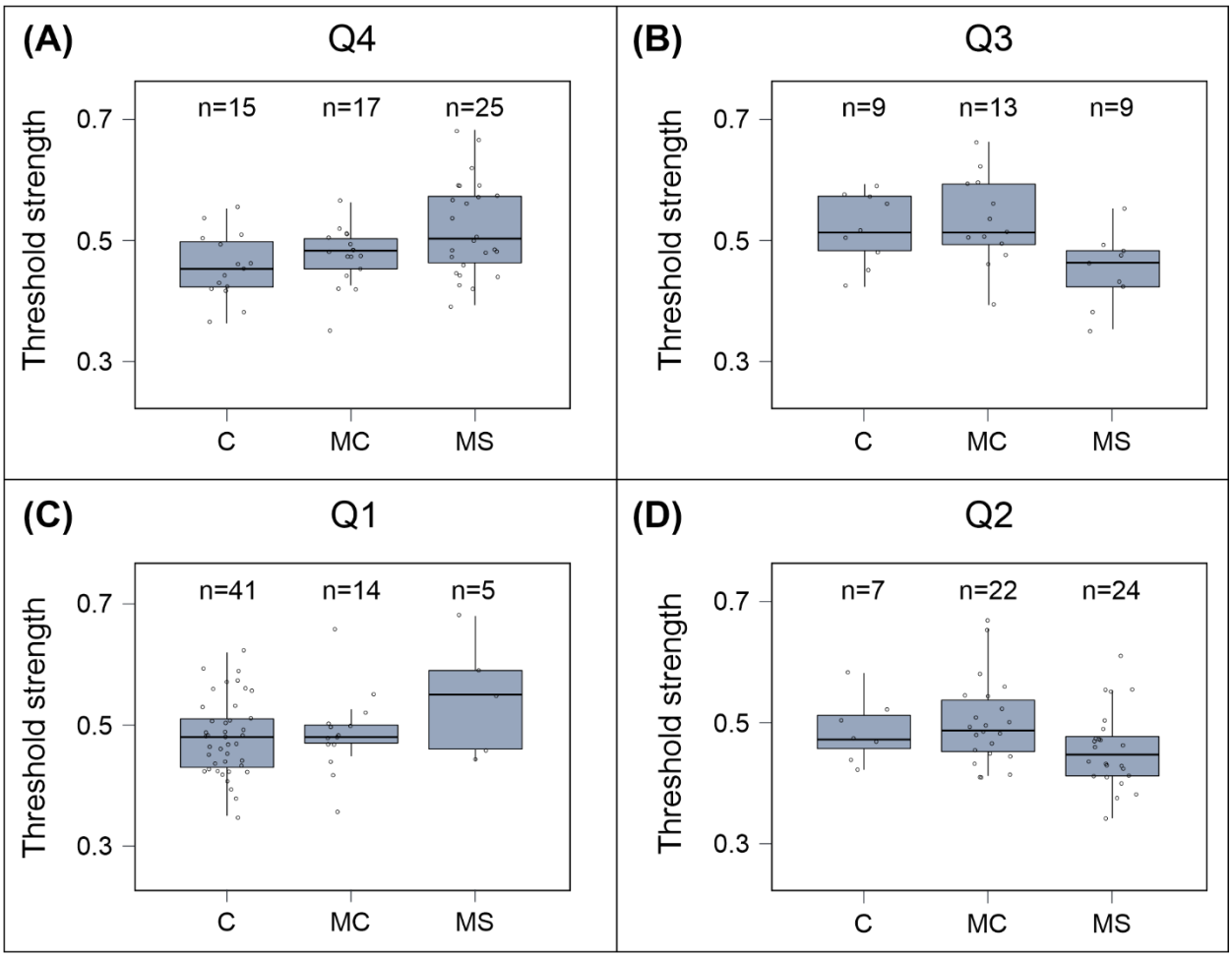


Figure 5-8. Boxplots showing the distribution of threshold strength for each quadrant and each dominant contour line classification. Classifications of S were omitted, as there were too few data points to build boxplots. The number of surfaces that had a specific classification for the corresponding quadrant is denoted by n. Horizontal black lines in boxplots are representative of median values, whereas each box spans from the 25th to 75th percentiles. Lower and upper whiskers extend from the median value to 1.5 times the inter-quartile range. Outliers are not shown.

5.4.4 Do three-dimensional approaches show promise for characterizing hydrologic thresholds?

While attempting to characterize hydrologic thresholds from 2D response curves and 3D response surfaces, the present study identified the advantages and limitations of some quantitative and qualitative tools, but it also highlighted important, open-ended questions. For instance, the aforementioned results suggest that the threshold strength parameter may prove useful for robust comparisons across threshold studies. Indeed, in hydrology, thresholds have typically been characterized by their absolute value and/or the shape of the nonlinear 2D response curve. Concerning specific threshold values, inter-site comparisons have been quite rare (e.g., Ali et al., 2015) and are complicated by the fact that not all studies use the same response metrics and meteorological factors for detecting thresholds. Furthermore, visual threshold detection methods applied to 2D plots can be complicated by scattering near the inflection point and user bias (Oswald et al., 2011). With respect to nonlinear responses dictated by thresholds, while several diagnostic shapes (e.g., hockey-stick, step or Heaviside function, Dirac function, and sigmoid function) have been identified (Ali et al., 2013), those diagnostic shapes are limited to 2D response curves. Robust comparisons based on shape are also inhibited by the fact that the processes underlying each diagnostic shape are unknown. Since the threshold strength parameter can be determined independently of uncertainty and user bias related to visual identification methods, without making assumptions about relationship shape, and regardless of dimensionality, it has the potential to help comparisons across studies. It should, however, be noted that for other quantitative relationship descriptors, like R^2 for correlative relationships, guidelines exist in the literature to translate descriptor values into relationship qualifiers, i.e., weak, moderate, or strong (Rousseau et al., 2018). A similar operational guideline for translating

and communicating threshold strength within the [0-1] range does not exist. As shown in Figure 5-2, response curves and surfaces associated with low threshold strength values can showcase very smooth changes in response that most hydrologists would not necessarily identify as thresholds after a visual assessment. An operational guideline might help describe changes in response that may be too subtle to detect visually, especially if they are deemed important for process understanding.

Another element worth discussing, about 3D response surfaces, is whether linkages could be made between surface features and the frequency with which threshold-crossing conditions are met. For instance, when evaluating threshold fronts associated with a linear response gradient such as those shown in Figure 5-6A, individual threshold fronts can be crossed a single time along the x-axis, but they cannot be crossed along the y-axis. When dealing with an angular response gradient such as those displayed in Figure 5-6B, however, the same threshold front can be crossed twice along the y-axis in some cases. For threshold fronts associated with radial gradients (i.e., hotspots) such as the one in Figure 5-6C, threshold fronts can be crossed multiple times regardless of direction. Future research could investigate scenarios with multiple threshold-crossing conditions and suggest process-based interpretations to explain them. One other context where multiple threshold-crossing conditions have been examined is in the literature surrounding memory persistence in soil moisture (e.g., Ghannam et al., 2016), which could be used to guide analyses of threshold-crossing conditions in rainfall-runoff relationships.

Lastly, in the present study, most response surfaces (60 of 64) constructed using LWPR had R^2 values greater than 0.45, when compared against input data. However, many curves (particularly those featuring an intensity factor) and some surfaces had lower R^2 values. Curves and surfaces with R^2 values below 0.45 were omitted from subsequent analyses to limit biases in

the estimation of threshold strength and the identification of key features of response surfaces. The R^2 criterion used to omit some curves and surfaces is somewhat arbitrary, and in fact, adjusting the smoothing parameter for LWPR has a strong influence on the resulting R^2 value. Therefore, in no way is the R^2 criterion implemented here absolute. Further work is needed to ascertain the relative influence of data distribution, data accuracy, and the degree of smoothing on the quality of fit for modelled curves and surfaces. Future studies could also be launched to evaluate the suitability of other techniques for modelling curves and surfaces (e.g., classification and regression trees, non-parametric multiplicative regression), with the specific goal of using those modelled curves and surfaces for characterizing and comparing threshold-mediated hydrologic responses. Such studies would also help address important questions about the threshold strength parameter. For instance, questions surrounding the numeric stability of the threshold strength parameter for quantifying the abruptness of change between different constant and/or dynamic states.

5.5 Conclusion

This study aimed to (1) compare the strength of thresholds identified in three dimensions to that of thresholds identified in two dimensions and (2) describe potential interactions between meteorological factors. The most novel contributions of this study include the extension of traditional threshold analysis to include two meteorological factors, the application of tools to quantify the strength of hydrologic thresholds, and the introduction of a qualitative classification approach for evaluating the presence of meteorological factor interactions driving hydrologic response. Key findings from this study include the identification of:

(1) **Smooth and sharp hydrologic thresholds.** Thresholds in rainfall-runoff relationships vary considerably in strength. However, the consideration of two factors (i.e., 3D assessments) does not necessarily lead to stronger thresholds than the consideration of one factor (i.e., 2D assessments).

(2) **Non-negligible factor interactions.** Not only do such interactions exist, but those that occur at low factor values appear to influence the strength of hydrologic thresholds.

(3) **Complex threshold fronts.** Hydrologic thresholds may exist at several combinations of critical factor values on a single three-dimensional response surface.

(4) **Wide-ranging degrees of hydrologic predictability.** Hydrologic responses characterized by gradients affected by two factors and factor interactions (i.e., angular or radial gradients) are likely more difficult to predict than response with linear gradients.

Findings from this study highlight some advantages of using three-dimensional approaches to characterize thresholds in rainfall-runoff relationships. The computation of the threshold strength parameter allows for an objective comparison of thresholds that differ based on the abruptness of response change. Further, the evaluation of response gradients and dominant contour line shapes on 3D surfaces depicting hydrologic response as a function of two meteorological factors allows factor interactions to be identified, when present. There are also areas of opportunity that may increase the usefulness of multi-factor approaches related to hydrologic thresholds. Importantly, findings from this study may encourage hydrologists to re-assess how thresholds are defined and identified for a broad range of possible nonlinear response changes. Future studies should further investigate the usefulness of three-dimensional approaches for characterizing hydrologic thresholds by considering a larger variety of sites and events across a range of hydroclimatic conditions and spatial scales.

5.6 References

Ali, G., L'Heureux, C., Roy, A., Turmel, M.-C., & Courchesne, F. (2011). Linking spatial patterns of perched groundwater storage and stormflow generation processes in a headwater forested catchment. *Hydrological Processes*, *25*(25), 3843–3857. https://doi.org/10.1002/hyp.8238

Ali, G., Oswald, C., Spence, C., Cammeraat, E., McGuire, K., Meixner, T., & Reaney, S. M. (2013). Towards a unified threshold-based hydrological theory: Necessary components and recurring challenges. *Hydrological Processes*, *27*(2), 313–318. https://doi.org/10.1002/hyp.9560

Ali, G., Tetzlaff, D., McDonnell, J., Soulsby, C., Carey, S., Laudon, H., McGuire, K., Buttle, J., Seibert, J., & Shanley, J. (2015). Comparison of threshold hydrologic response across northern catchments. *Hydrological Processes*, *29*(16), 3575–3591. https://doi.org/10.1002/hyp.10527

Andersen, T., Carstensen, J., Hernandez-Garcia, E., & Duarte, C. M. (2009). Ecological thresholds and regime shifts: Approaches to identification. *Trends in Ecology & Evolution*, *24*(1), 49–57.

Antony, J. (2014). *Design of experiments for engineers and scientists*. Elsevier.

Bailey, R. A. (2008). *Design of comparative experiments* (Vol. 25). Cambridge University Press.

Beven, K. (2006). A manifesto for the equifinality thesis. *Journal of Hydrology*, *320*(1–2), 18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007

Beven, K. (2011). *Rainfall-runoff modelling: The primer*. John Wiley & Sons.

Borga, M., Stoffel, M., Marchi, L., Marra, F., & Jakob, M. (2014). Hydrogeomorphic response to extreme rainfall in headwater systems: Flash floods and debris flows. *Journal of Hydrology*, *518*, 194–205. https://doi.org/10.1016/j.jhydrol.2014.05.022

Cammeraat, L. H. (2002). A review of two strongly contrasting geomorphological systems within the context of scale. *Earth Surface Processes and Landforms*, *27*(11), 1201–1222. https://doi.org/10.1002/esp.421

Cleveland, W. S., & Devlin, S. J. (1988). Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*, *83*(403), 596–610. JSTOR. https://doi.org/10.2307/2289282

Detty, J. M., & McGuire, K. J. (2010). Threshold changes in storm runoff generation at a till-mantled headwater catchment. *Water Resources Research; Washington*, *46*(7). http://dx.doi.org.uml.idm.oclc.org/10.1029/2009WR008102

Dingman, S. L. (2015). *Physical hydrology*. Waveland press.

Dunne, T. (1978). Field studies of hillslope flow processes. *Hillslope Hydrology*, 227–293.

Francesco Ficetola, G., & Denoël, M. (2009). Ecological thresholds: An assessment of methods to identify abrupt changes in species–habitat relationships. *Ecography*, *32*(6), 1075–1084. https://doi.org/10.1111/j.1600-0587.2009.05571.x

Ghannam, K., Nakai, T., Paschalis, A., Oishi, C. A., Kotani, A., Igarashi, Y., Kumagai, T., & Katul, G. G. (2016). Persistence and memory timescales in root-zone soil moisture dynamics. *Water Resources Research*, *52*(2), 1427–1445. https://doi.org/10.1002/2015WR017983

Gill, M. K., Kaheil, Y. H., Khalil, A., McKee, M., & Bastidas, L. (2006). Multiobjective particle swarm optimization for parameter estimation in hydrology. *Water Resources Research*, *42*(7). https://doi.org/10.1029/2005WR004528

Graham, C. B., Woods, R. A., & McDonnell, J. J. (2010). Hillslope threshold response to rainfall: (1) A field based forensic approach. *Journal of Hydrology*, *393*(1–2), 65–76. https://doi.org/10.1016/j.jhydrol.2009.12.015

Huang, H., Wang, Z., Xia, F., Shang, X., Liu, Y., Zhang, M., Dahlgren, R. A., & Mei, K. (2017). Water quality trend and change-point analyses using integration of locally weighted polynomial regression and segmented regression. *Environmental Science and Pollution Research International; Heidelberg*, *24*(18), 15827–15837. http://dx.doi.org.uml.idm.oclc.org/10.1007/s11356-017-9188-x

James, A., & Roulet, N. (2007). Investigating hydrologic connectivity and its association with threshold change in runoff response in a temperate forested watershed. *Hydrological Processes*, *21*(25), 3391–3408. https://doi.org/10.1002/hyp.6554

Kavetski, D., Kuczera, G., & Franks, S. W. (2006). Calibration of conceptual hydrological models revisited: 2. Improving optimisation and analysis. *Journal of Hydrology*, *320*(1), 187–201. https://doi.org/10.1016/j.jhydrol.2005.07.013

Kim, H. J., Sidle, R. C., Moore, R. D., & Hudson, R. (2004). Throughflow variability during snowmelt in a forested mountain catchment, coastal British Columbia, Canada. *Hydrological Processes*, *18*(7), 1219–1236. https://doi.org/10.1002/hyp.1396

Kinzig, A., Ryan, P., Etienne, M., Allison, H., Elmqvist, T., & Walker, B. (2006). Resilience and Regime Shifts: Assessing Cascading Effects. *Ecology and Society*, *11*(1). https://doi.org/10.5751/ES-01678-110120

Lall, U., Moon, Y.-I., Kwon, H.-H., & Bosworth, K. (2006). Locally weighted polynomial regression: Parameter choice and application to forecasts of the Great Salt Lake. *Water Resources Research*, *42*(5).

Lehmann, P., Hinz, C., McGrath, G., Tromp-van Meerveld, H. J., & McDonnell, J. J. (2007). Rainfall threshold for hillslope outflow: An emergent property of flow pathway connectivity. *Hydrol. Earth Syst. Sci.*, *11*(2), 1047–1063. https://doi.org/10.5194/hess-11-1047-2007

Limburg, K. E., O'Neill, R. V., Costanza, R., & Farber, S. (2002). Complex systems and valuation. *Ecological Economics*, *41*(3), 409–420. https://doi.org/10.1016/S0921-8009(02)00090-3

Lintz, H. E., McCune, B., Gray, A. N., & McCulloh, K. A. (2011). Quantifying ecological thresholds from response surfaces. *Ecological Modelling*, *222*(3), 427–436. https://doi.org/10.1016/j.ecolmodel.2010.10.017

McDonnell, J., Sivapalan, M., Vaché, K., Dunn, S., Grant, G., Haggerty, R., Hinz, C., Hooper, R., Kirchner, J., Roderick, M. L., Selker, J., & Weiler, M. (2007). Moving beyond heterogeneity and process complexity: A new vision for watershed hydrology. *Water Resources Research*, *43*(7), W07301. https://doi.org/10.1029/2006WR005467

McGrew, J., Lembo, A., & Monroe, C. (1993). *An introduction to statistical problem solving in geography*. Waveland Press, Inc.

McKee, A., & Druliner, P. (1998). *HJ Andrews Experimental Forest.* http://andrewsforest.oregonstate.edu/pubs/pdf/pub2415.pdf

Mielko, C., & Woo, M. (2006). Snowmelt runoff processes in a headwater lake and its catchment, subarctic Canadian Shield. *Hydrological Processes*, *20*(4), 987–1000. https://doi.org/10.1002/hyp.6117

Mirus, B. B., & Loague, K. (2013). How runoff begins (and ends): Characterizing hydrologic response at the catchment scale. *Water Resources Research*, *49*(5), 2987–3006. https://doi.org/10.1002/wrcr.20218

Mosley, M. P. (1979). Streamflow generation in a forested watershed, New Zealand. *Water Resources Research*, 15(4), 795–806. https://doi.org/10.1029/WR015i004p00795

Oswald, C., Richardson, M. C., & Branfireun, B. A. (2011). Water storage dynamics and runoff response of a boreal Shield headwater catchment. *Hydrological Processes*, *25*(19), 3042–3060. https://doi.org/10.1002/hyp.8036

Phillips, J. D. (2006). Evolutionary geomorphology: Thresholds and nonlinearity in landform response to environmental change. *Hydrol. Earth Syst. Sci.*, *10*(5), 731–742. https://doi.org/10.5194/hess-10-731-2006

Putty, M. R. Y., & Prasad, R. (2000). Runoff processes in headwater catchments—An experimental study in Western Ghats, South India. *Journal of Hydrology*, *235*(1), 63–71. https://doi.org/10.1016/S0022-1694(00)00262-6

Rajagopalan, B., & Lall, U. (1998). Locally weighted polynomial estimation of spatial precipitation. *Journal of Geographic Information and Decision Analysis*, *2*(2), 44–51.

Reaney, S. M., Bracken, L. J., & Kirkby, M. J. (2007). Use of the Connectivity of Runoff Model (CRUM) to investigate the influence of storm characteristics on runoff generation and connectivity in semi-arid areas. *Hydrological Processes*, *21*(7), 894–906. https://doi.org/10.1002/hyp.6281

Redding, T. E., & Devito, K. J. (2008). Lateral flow thresholds for aspen forested hillslopes on the Western Boreal Plain, Alberta, Canada. *Hydrological Processes*, *22*(21), 4287–4300. https://doi.org/10.1002/hyp.7038

Ross, C., Ali, G., Spence, C., Oswald, C., & Casson, N. (2019). Comparison of event-specific rainfall–runoff responses and their controls in contrasting geographic areas. *Hydrological Processes*, *33*(14), 1961–1979. https://doi.org/10.1002/hyp.13460.

Rousseau, R., Egghe, L., & Guns, R. (2018). *Becoming metric-wise: A bibliometric guide for researchers*. Chandos Publishing.

Scaife, C. I., & Band, L. E. (2017). Nonstationarity in threshold response of stormflow in southern Appalachian headwater catchments. *Water Resources Research*, *53*(8), 6579–6596. https://doi.org/10.1002/2017WR020376

Seeger, M., Errea, M.-P., Beguería, S., Arnáez, J., Martí, C., & García-Ruiz, J. M. (2004). Catchment soil moisture and rainfall characteristics as determinant factors for discharge/suspended sediment hysteretic loops in a small headwater catchment in the Spanish pyrenees. *Journal of Hydrology*, *288*(3), 299–311. https://doi.org/10.1016/j.jhydrol.2003.10.012

Sidle, R. C., Tsuboyama, Y., Noguchi, S., Hosoda, I., Fujieda, M., & Shimizu, T. (2000). Stormflow generation in steep forested headwaters: A linked hydrogeomorphic paradigm. *Hydrological Processes*, *14*(3), 369–385. https://doi.org/10.1002/(SICI)1099-1085(20000228)14:3<369::AID-HYP943>3.0.CO;2-P

Sivapalan, M. (2006). Pattern, Process and Function: Elements of a Unified Theory of Hydrology at the Catchment Scale. In *Encyclopedia of Hydrological Sciences*. John Wiley & Sons, Ltd. https://doi.org/10.1002/0470848944.hsa012

Sivapalan, M., Jothityangkoon, C., & Menabde, M. (2002). Linearity and nonlinearity of basin response as a function of scale: Discussion of alternative definitions. *Water Resources Research*, *38*(2). https://doi.org/10.1029/2001WR000482

Spence, C. (2007). On the relation between dynamic storage and runoff: A discussion on thresholds, efficiency, and function. *Water Resources Research*, *43*(12), W12416. https://doi.org/10.1029/2006WR005645

Spence, Christopher, & Woo, M. (2003). Hydrology of subarctic Canadian shield: Soil-filled valleys. *Journal of Hydrology*, *279*(1), 151–166. https://doi.org/10.1016/S0022-1694(03)00175-6

Tang, W., & Carey, S. K. (2017). HydRun: A MATLAB toolbox for rainfall–runoff analysis. *Hydrological Processes*, *31*(15), 2670–2682. https://doi.org/10.1002/hyp.11185

Tani, M. (1997). Runoff generation processes estimated from hydrological observations on a steep forested hillslope with a thin soil layer. *Journal of Hydrology*, *200*(1), 84–109. https://doi.org/10.1016/S0022-1694(97)00018-8

Tromp-van Meerveld, H. J., & McDonnell, J. J. (2006a). Threshold relations in subsurface stormflow: 1. A 147-storm analysis of the Panola hillslope. *Water Resources Research*, *42*(2). https://doi.org/10.1029/2004WR003778

Tromp-van Meerveld, H. J., & McDonnell, J. J. (2006b). Threshold relations in subsurface stormflow: 2. The fill and spill hypothesis. *Water Resources Research*, *42*(2). https://doi.org/10.1029/2004WR003800

Wei, L., Qiu, Z., Zhou, G., Kinouchi, T., & Liu, Y. (2020). Stormflow threshold behaviour in a subtropical mountainous headwater catchment during forest recovery period. *Hydrological Processes*, *34*(8), 1728–1740. https://doi.org/10.1002/hyp.13658

Whipkey, R. Z. (1965). Subsurface Stormflow from Forested Slopes. *International Association of Scientific Hydrology. Bulletin*, *10*(2), 74–85. https://doi.org/10.1080/02626666509493392

Woods, R., Grayson, R., Western, A., Duncan, M., Wilson, D., Young, R., Ibbitt, R., Henderson, R., & McMahon, T. (2013). Experimental Design and Initial Results from the Mahurangi River Variability Experiment: MARVEX. In *Land Surface Hydrology, Meteorology, and Climate: Observations and Modeling* (pp. 201–213). American Geophysical Union. http://onlinelibrary-wiley-com.uml.idm.oclc.org/doi/10.1029/WS003p0201/summary

Zehe, E., & Blöschl, G. (2004). Predictability of hydrologic response at the plot and catchment scales: Role of initial conditions. *Water Resources Research*, *40*(10), W10202. https://doi.org/10.1029/2003WR002869

Zhang, X., Srinivasan, R., Zhao, K., & Liew, M. V. (2009). Evaluation of global optimization algorithms for parameter calibration of a computationally intensive hydrologic model. *Hydrological Processes*, *23*(3), 430–441. https://doi.org/10.1002/hyp.7152

CHAPTER 6. SYNTHESIS AND CONCLUSION

6.1 Summary and major findings

In hydrology, a unified theory is needed to increase the inter-site and cross-scale transferability of findings and process interpretations from individual studies conducted at unique places (Beven, 2000; Kirchner, 2006, 2009; McDonnell et al., 2007). Hydrologic thresholds have been identified as a potential foundation for such a theory (Ali et al., 2013; Spence, 2010), as they are emergent properties that reflect process and landscape heterogeneity (Lehmann et al., 2007). The primary goal of this thesis was to contribute to the development of a unified threshold-based hydrological theory by resolving key limitations surrounding our theoretical and operational understanding of hydrologic thresholds. Indeed, most threshold research has been limited to hillslopes and small catchments with humid to temperate climates, while fewer studies have focused on other environments (Ali et al., 2013, 2015). To fill this knowledge gap, research questions were pursued using existing data of twenty-one sites from seven study areas. These sites vary in scale, and study areas have contrasting climate, topography, soils, geology, and land-use/land-cover. This thesis was guided by four objectives:

(1) Assess the spatial and temporal variability in rainfall-runoff event dynamics and the influence of fixed (e.g., topography) and dynamic (e.g., climate) controls on hydrologic response across a range of scales and environments.

(2) Evaluate the ubiquity of threshold behaviour in rainfall-runoff event response, with a special focus on rainfall depth thresholds, rainfall intensity thresholds, and thresholds in hydrologic abstraction from evapotranspiration.

(3) Appraise the potential benefits of constraining rainfall-runoff model outputs using multiple hydrologic descriptors, including thresholds.

(4) Characterize the simultaneous influence of multiple, potentially interacting, meteorological factors on threshold mediated hydrologic response.

In Chapter 2, analyses were carried out to assess the spatial and temporal variability in rainfall-runoff event dynamics and the influence of fixed (e.g., topography) and dynamic (e.g., climate) controls on hydrologic response across twenty-one sites (research objective 1). This was key in determining the best response metrics for capturing temporal variability in response and characterizing the relative influence of different meteorological factors on response – knowledge that was essential to address the research questions of this thesis. Findings from Chapter 2 showed that most rainfall-runoff event responses could be qualitatively described as low magnitude and fast timing, or as high magnitude and slow timing, even though there were significant differences in the climates and physiographic features of the study areas considered. These analyses also confirmed significant spatial and temporal variability in the magnitude and timing of hydrologic response at the rainfall-runoff event scale. Some timing metrics (e.g., lag-to-peak) were as important or more important than more commonly considered response magnitude metrics (e.g., runoff ratio) for describing site-specific response variability overall. Furthermore, a large suite of meteorological factors was analyzed to determine their effects on hydrologic response variability. These different meteorological factors are presumed to be associated with storage-driven processes, intensity-driven processes, or combinations of both storage-driven and intensity-driven processes. The observed temporal variability in site-specific hydrologic response was often attributed to the influence of meteorological factors associated with intensity-driven processes, articulating the importance of considering the effects of these processes on hydrologic response. It should be noted that in Chapter 2, the response metrics and meteorological factors used to characterize rainfall-runoff events were grouped based on the

processes that they were presumed to be associated with. Alternatively, meteorological factors can be grouped by their unit of measure or based on how they most directly affect the hydrologic response. Each of these grouping approaches has specific consequences. Grouping meteorological factors based on the processes that they are presumed to be associated with may help ascertain the relative importance of different processes in determining hydrologic response. However, in some cases, deriving straightforward process-based meteorological factor groups can be difficult. For example, antecedent potential evapotranspiration can be considered as an intensity-driven factor or as a storage-driven factor: antecedent potential evapotranspiration is related to the intensity of solar radiation (Dingman, 2015); however, antecedent potential evapotranspiration influences changes in dynamic storage. Grouping meteorological factors based on their unit of measure is intuitive and straightforward. However, this type of grouping has no process-based rationale and most factors can be represented as rates, making unit-based groupings relatively meaningless. Grouping meteorological factors based on how they most directly affect hydrologic response provides a compromise since this type of grouping is meaningful in terms of how different factors are related to hydrologic response and since it is intuitive. As such, the latter form of grouping was adopted in Chapters 3, 4, and 5. Meteorological factor groups include factors that quantify rainfall depth, factors that quantify rainfall intensity, and factors that quantify hydrologic abstraction caused by evapotranspiration.

Hydrologic response metrics and meteorological factors that were derived in Chapter 2 were used in Chapter 3 to evaluate the ubiquity of threshold behaviour across the twenty-one sites considered in this thesis (research objective 2). Importantly, the potential influence of rainfall depth, rainfall intensity, and hydrologic abstraction caused by evapotranspiration on nonlinearities in response magnitude and timing was assessed. In doing so, Chapter 3 filled a

critical gap in our current knowledge of threshold behaviour, as the available literature has mostly focused on rainfall depth thresholds for hillslope or catchment response magnitude (e.g., Ali et al., 2011; Detty & McGuire, 2010; James & Roulet, 2007; Kim et al., 2004; Mosley, 1979; Oswald et al., 2011; Redding & Devito, 2008; Sidle et al., 2000; Tani, 1997; Tromp-van Meerveld & McDonnell, 2006a; Weiler et al., 2006; Whipkey, 1965), while fewer studies have assessed rainfall intensity thresholds (e.g., Cammeraat, 2002; Reaney et al., 2007) or hydrologic abstraction caused by evapotranspiration. Additionally, this chapter assessed antecedent conditions as a control on a broad range of threshold behaviours. This assessment went beyond other threshold studies that quantified antecedent condition effects (e.g., Detty & McGuire, 2010; James & Roulet, 2007; Tromp-van Meerveld & McDonnell, 2006a) by considering: 1) the influence of both antecedent rainfall and hydrologic abstractions calculated over a wide range of durations on threshold behaviour, and 2) the influence of these factors on potential nonlinearities in response. Results from Chapter 3 demonstrated the pervasiveness of total event rainfall thresholds that affect response magnitude, as these thresholds were observed for fourteen of twenty-one sites. Results indicate that measures of antecedent rainfall and event rainfall depth, as well as rainfall intensity and hydrologic abstractions from evapotranspiration, can also be important controls on threshold behaviour. This finding was applicable even for more humid environments with high infiltration capacity, where rainfall intensity is assumed to exert little or no influence on the hydrologic response (Graham & McDonnell, 2010; Tromp-van Meerveld & McDonnell, 2006a). While not all sites had thresholds for the same input-output pairs, threshold behaviour was observed at twenty out of twenty-one of the sites evaluated in this thesis. Chapter 3 also presented evidence of threshold behaviour being sensitive to both event conditions and antecedent conditions over a wide range of antecedent durations. Combined, results of Chapter 3

offered an opportunity for synthesis, leading to the development of a typology that distinguishes threshold dynamics based on the extent to which thresholds affect different aspects of event-specific response magnitude and response timing, memory effects on thresholds, and the processes leading to threshold behaviour. The proposed typology of threshold dynamics may provide a useful tool for future threshold-based studies.

Chapter 4 focused on a rainfall-runoff model evaluation approach implemented using data from the MRC8 catchment. Rainfall-runoff model simulations were assessed on their ability to reproduce a range of catchment functions that are captured by the flow duration curve, rainfall-runoff event response timing metrics, and hydrologic thresholds (research objective 3). Since these hydrologic descriptors capture catchment functions, they are considered to be process-based information. While previous model evaluations have been performed using the flow duration curve (Blazkova & Beven, 2009; Herbst et al., 2009; Ley et al., 2016; Westerberg et al., 2011; Yilmaz et al., 2008; Yu & Yang, 2000) and event response metrics (Loague & Freeze, 1985; Yang et al., 2004), Chapter 4 was the first model evaluation that incorporated emergent threshold behaviour. Findings from Chapter 4 showed that behavioural simulations identified based on a single performance measure (i.e., KGE) calculated over the entire continuous streamflow timeseries were inadequate for identifying simulations that can reproduce real-world processes, necessitating post-calibration model evaluation using process-based information. Behavioural simulations ranged broadly in their ability to minimize measures of bias associated with specific hydrologic descriptors. Like other studies (e.g., Dunn, 1999), model simulations reproduced flow volumes and response timing associated with large rainfall events reasonably well, while fewer simulations adequately reproduced baseflow-dominated conditions. In terms of threshold behaviour, many behavioural simulations adequately reproduced observed

threshold behaviour for relationships involving antecedent plus event rainfall and response magnitude. Key findings of Chapter 4 were that model evaluation on process-based measures of bias could be used to identify simulations that reproduced observed hydrologic behaviour and to constrain model parameter space. Overall, including thresholds in the model evaluation process showed that emergent catchment properties can be valuable tools for falsifying model outputs.

In Chapter 5, a three-dimensional (3D) approach was taken to characterize the simultaneous influence of two, potentially interacting, meteorological factors on threshold behaviour in hydrologic response for the MRC and HJA catchments (research objective 4). While numerous studies have demonstrated that the variability in hydrologic response is determined by multiple processes, threshold studies have considered hydrologic response as a function of a single meteorological factor. In this chapter, thresholds detected from 3D surfaces showing response as a function of two meteorological factors were compared against thresholds involving the same meteorological factors detected from two-dimensional (2D) curves, this using a parameter of threshold strength. Results indicated that 2D thresholds are often stronger than 3D thresholds and that there is significant variability in the abruptness of threshold behaviour. Furthermore, 3D surfaces showed characteristics associated with interactions between meteorological factors that can lead to complex surface geometries that may present challenges for hydrologic response predictability. While 2D thresholds were often stronger, they are inadequate for characterizing meteorological factor interactions that were shown to affect threshold behaviour in some cases. Importantly, results showed that hydrologic thresholds may exist at several combinations of critical meteorological factor values on a single 3D response surface. Combined, findings from Chapter 5 indicate that multiple meteorological factors may

influence the emergence of nonlinear response and that the form of hydrologic thresholds may differ from common notions of what a threshold is.

6.2 Study limitations

This thesis was made possible using existing data from multiple study areas. There were several limitations associated with the data and data-processing tools that were used. Undoubtedly, data accuracy influenced the meteorological factors and response metrics calculated during rainfall-runoff event analysis and all subsequent results and interpretations presented throughout this thesis. While standard measurement procedures were used to obtain input data at all twenty-one sites (Chapter 1), methodology varied by jurisdiction and the data have indeed undergone varying degrees of quality control and processing by the original investigators. Unfortunately, no formal uncertainty analysis could be performed for the data used in this thesis due to a lack of site-specific information about measurement error. Relying on the literature, error in point measurements of rainfall range from 5 to 15% for long-term data and as high as 75% for storm data (Winter, 1981). It is also unclear how well meteorological measurements from a single station captured hillslope- or catchment-wide conditions, which was assumed in this thesis. Discharge estimates are also likely to be associated with significant error. While the level of uncertainty is highly dependent on the technique used to establish site-specific rating curves (e.g., USGS velocity-area measurements), error rates can be as high as 50% (McMillan et al., 2018). It is unclear how measurement error affected the findings presented in this thesis. More specifically, it is unclear how much of the unexplained variance in response (Chapter 2) and threshold mediated relationships (Chapter 3) was associated with limitations in

measurement or missing explanatory variables. Furthermore, measurement errors would have propagated throughout the modelling exercise presented in Chapter 4 and may have influenced the features of modelled 3D response surfaces (Chapter 5). Therefore, one limitation of this thesis is that the influence of measurement error in input data is not fully understood and the absence of replicates or nested data for the sites considered in this thesis prevented a more exhaustive assessment of input data measurement error (McMillan et al., 2018).

In addition to measurement error, the record length of hydrometric and meteorological data ranged from one to five years, depending on the study site. Also, data of more northern sites (i.e., CCW, UP1, and HRM) have seasonal gaps because below freezing conditions preclude runoff. These two data features had ramifications throughout the thesis. First, shorter record length for some sites prevented the characterization of longer-term temporal variability in rainfall-runoff event response and the factors that control that response (Chapter 2). Second, since some sites had shorter data records or were significantly drier than others (e.g., TRC), there was considerable variability in the number of rainfall-runoff events that were delineated for each site. In Chapter 3, this prevented questions from being answered, like how many rainfall-runoff event data points are needed to ensure consistency in threshold detection? Also, in Chapter 3, for some sites, there were too few rainfall-runoff events to conduct a uniform analysis of seasonal variability in threshold behaviour across sites. In Chapter 4, record length limited the number of sites that were suitable for the described modelling exercise. In Chapter 5, the number of rainfall-runoff events per site affected site selection, as many events are needed to decrease uncertainty in modelled 3D response surfaces. Along similar lines, this thesis relied heavily on meteorological factors and response metrics of rainfall-runoff events that were delineated using HydRun (Tang & Carey, 2017). While HydRun facilitated the quasi-automated delineation of

hundreds of rainfall-runoff events, this tool is not designed to delineate snowmelt or rain-on-snow events. As such, only rainfall-initiated events were considered, even though some of the study sites are influenced by snowmelt. Other studies that have involved event delineation have also focused on rainfall-initiated events only or assessed rainfall-initiated and snowmelt-initiated events independently as they require different event delineation tools (Blume et al., 2007; Penna et al., 2016).

Many process-based inferences were made throughout this thesis related to observed (Chapters 3 and 5) and predicted (Chapter 4) threshold behaviour. In some cases, these inferences were related to processes that vary in both space and time. One limitation of this study is that spatial data was not used to verify these process-based hypotheses. Other studies have used spatial data of topography, bedrock, soils, vegetation, soil moisture, and water table depth to explain threshold behaviour (e.g., Detty & McGuire, 2010; Freer et al., 2002; Graham et al., 2010; Graham & McDonnell, 2010; James & Roulet, 2007; Kim et al., 2004; Laudon et al., 2007; Lehmann et al., 2007; Mielko & Woo, 2006; Oswald et al., 2011; Redding & Devito, 2008; Scaife & Band, 2017; Tromp-van Meerveld & McDonnell, 2006a, 2006b). Integrating spatial data throughout this thesis would have been a large undertaking. While the inclusion of spatial data would have increased the number of potential research questions, fewer sites would have been considered as spatial data was not available for every site.

6.3 Synthesis across thesis chapters and recommendations for future work

6.3.1 From response variability to hydrologic thresholds

Amalgamating findings from Chapters 2 and 3 may help ascertain whether the degree of temporal variability for a given response metric coincides with that metric's involvement in threshold mediated relationships. Most event-specific response magnitude and response timing metrics for sites of this thesis are characterized by considerable temporal variability (Chapter 2). For response magnitude metrics, all sites had relatively low temporal variability in RR, except for the PMRW (Table 6-1). In contrast, Q_{TOT} and Q_{MAX} were slightly more variable than RR at most sites and the temporal variability of I_{abs} varied among sites (Table 6-1). Overall, response magnitude metrics with more temporal variability did not necessarily contribute to a larger number of threshold mediated input-output relationships. For example, I_{abs} was highly variable across events at the MRC6 site (CV = 4.34), but only 52% of input-output relationships involving I_{abs} were threshold mediated. Response timing metrics varied less across events at each site than response magnitude metrics. Threshold mediated relationships involving response timing metrics were also rare: there were fifty-two cases across all sites and metrics where timing metrics do not exhibit threshold mediated relationships. Also, response metrics that were deemed most important for explaining site-specific response variability overall according to principal component loadings calculated in Chapter 2 did not necessarily contribute to more threshold mediated relationships than other response metrics (Chapter 3). Combined, these results indicate that response metric temporal variability alone is not a meaningful surrogate measure of threshold mediated hillslope or catchment response for the sites considered in this

thesis. Such a surrogate measure would have allowed future threshold-focused studies to consider fewer input-output pairs because the response metrics that are most likely to be associated with threshold behaviour could have potentially been identified based on their temporal variability.

Table 6-1. Site- and metric-specific temporal variability is shown by the coefficient of variation (CV). Asterisks (*) show response magnitude and timing metrics that were important for explaining site-specific response temporal variability (i.e., principal component loadings >|0.45|). The percentage of input-output pairs that were threshold mediated for each site and metric is shown (%).

	RR		Q_{TOT}		Q_{MAX}		I_{abs}		T_{LR}		T_{LP}		T_c	
	CV	%	CV	%	CV	%	CV	%	CV	%	CV	%	CV	%
PMRW	2.19	10	2.30	10	2.10	13	1.28	16	1.02	13	0.77*	0	0.74	10
HRM	0.95	0	1.50	13	1.34*	23	1.70	13	1.30	7	1.30*	0	0.59*	0
UP1	0.87*	13	0.90	0	1.06*	0	1.62	42	1.38*	10	1.04*	0	1.38*	7
TRC	1.00*	0	1.29*	0	1.24	0	1.44*	36	1.02	13	0.81	0	0.97*	3
CCW	0.88	0	1.81	20	0.76	7	2.51*	3	1.34*	0	0.91*	0	1.00	0
MRC1	0.89	3	1.22*	0	1.84*	52	1.52*	3	0.72*	0	0.46*	0	0.39	3
MRC2	0.57	0	1.29	0	1.15	0	2.02*	0	1.00*	0	0.68*	0	0.46*	0
MRC3	0.79	0	1.51*	13	1.59	0	2.70	0	0.84*	0	0.54	0	0.47*	0
MRC4	0.70	0	1.31*	0	1.68*	0	2.85*	7	0.99*	0	0.63*	0	0.54*	13
MRC5	0.85	0	1.73*	0	2.03*	3	4.34*	29	0.87	0	0.61	0	0.46	0
MRC6	0.72	0	1.77*	0	1.56	0	4.24	52	1.08*	0	0.64*	0	0.60*	0
MRC7	0.65	0	1.77*	23	1.40*	3	2.56*	19	1.01*	0	0.58	0	0.40*	0
MRC8	0.68	0	1.47	13	1.80	10	2.40*	0	1.13*	0	0.68*	0	0.38	0
HJA1	0.90	0	1.39	0	1.48	0	1.78*	42	0.87*	13	0.72*	0	0.58*	0
HJA2	1.05	0	1.51	0	1.70	0	1.88*	32	0.89*	0	0.79*	0	0.61*	0
HJA3	0.81	0	1.32	0	1.51	0	1.86*	26	0.93	3	0.78*	0	0.58*	0
HJA4	0.69	0	1.25*	10	1.29	10	1.28*	7	0.63*	0	0.53	0	0.41*	0
HJA5	0.57*	0	1.02*	7	1.27*	7	1.56*	16	0.75*	0	0.49	0	0.41*	0
HJA6	0.75*	0	1.58	13	1.46	13	1.87	7	0.87*	0	0.88*	0	0.55*	0
HJA7	0.96*	0	1.48	0	1.43*	0	1.75*	36	0.85*	0	0.71*	0	0.55*	0
HJA8	0.68	0	1.36*	19	1.61*	0	1.83*	7	0.96*	0	0.77*	0	0.56	0

6.3.2 False-positive and false-negative threshold detection in GR5H behavioural simulations

Some of the threshold behaviours that were observed in Chapter 3 were erroneously predicted in Chapter 4 (i.e., false-positive and false-negative threshold detection). Through synthesis across Chapters 2, 3, and 4, the potential influence of response metric and meteorological factor temporal variability on the erroneous prediction of thresholds can be assessed. Answering this question may help determine why some threshold behaviours could not be predicted by the GR5H model. For data of the MRC8 site that was evaluated across Chapters 2, 3, and 4, the input meteorological factor with the greatest temporal variability was R_{TOT} (Table 6-2). Rainfall intensity factors had slightly lower temporal variability. Also, rainfall depth factors that quantify antecedent and event rainfall decreased in temporal variability as the antecedent duration increased. Similarly, the temporal variability of Q_{TOT} (CV = 1.47) was slightly lower than Q_{MAX} (CV = 1.80). The R_{TOT}-Q_{MAX} input-output pair includes the meteorological factor and response metric with the greatest temporal variability and this input-output pair was associated with false-positive threshold identification for 79.5% of behavioural simulations in Chapter 4. However, the temporal variability of the R_{TOT}-Q_{TOT} pair was similar and threshold behaviour for this input-output pair was appropriately detected by 97.1% of behavioural simulations. Similarly, AR_{14}+R_{TOT} exhibited relatively low temporal variability, yet thresholds for the AR_{14}+R_{TOT}-Q_{MAX} pair were associated with false-positive threshold detection for 93.7% of behavioural simulations. These results indicate that meteorological factor or response metric temporal variability is not informative for determining why the GR5H model was able to predict some thresholds better than others.

Table 6-2. The coefficient of variation (CV) of meteorological factors (input) and response metrics (output) for the MRC8 catchment and the percentage of behavioural simulations associated with true-positive, false-negative, false-positive, and true-positive threshold detection for each input-output pair. "NA": Options that are inapplicable given the presence or absence of threshold behaviour in the observed data.

	Input CV	Output CV	True-positive	False-negative	False-positive	True-negative
R_{TOT}, Q_{TOT}	1.12	1.47	97.1	2.9	NA	NA
R_{TOT}, Q_{MAX}	1.12	1.80	NA	NA	79.5	20.5
RI_{AVG}, Q_{TOT}	0.85	1.47	NA	NA	0.0	100.0
RI_{AVG}, Q_{MAX}	0.85	1.80	NA	NA	0.8	99.2
RI_{MAX}, Q_{TOT}	0.91	1.47	NA	NA	0.0	100.0
RI_{MAX}, Q_{MAX}	0.91	1.80	NA	NA	27.3	72.7
AR_3+R_{TOT}, Q_{TOT}	1.01	1.47	100.0	0.0	NA	NA
AR_3+R_{TOT}, Q_{MAX}	1.01	1.80	NA	NA	99.0	1.0
AR_7+R_{TOT}, Q_{TOT}	0.89	1.47	94.1	5.9	NA	NA
AR_7+R_{TOT}, Q_{MAX}	0.89	1.80	85.3	14.7	NA	NA
AR_{14}+R_{TOT}, Q_{TOT}	0.78	1.47	99.9	0.1	NA	NA
AR_{14}+R_{TOT}, Q_{MAX}	0.78	1.80	NA	NA	93.7	6.3

6.3.3 Observed threshold behaviour and the emergent nature of thresholds in runoff response

Rainfall thresholds observed at the hillslope or catchment scale have been referred to as emergent properties (Ali et al., 2011, 2015; Lehmann et al., 2007; Sidle et al., 2001; Spence, 2010; Weiler, 2005) because they are associated with the integration of responses at the next smaller spatial scale (Lehmann et al., 2007). The broad range of threshold mediated behaviours that were observed and quantified in Chapters 3 and 4 may provide insights on the emergent nature of thresholds in runoff response. Emergence is a hallmark of complex systems (Fromm, 2004; Ottino, 2004) and through self-organization, many complex natural systems become

structured into a higher organized state by their internal processes (Fromm, 2004; Yates, 2012). In the philosophical and scientific literature, conceptions of emergence are hotly contested (Bedau, 1997; Chalmers, 2006; Gillett, 2016; O'Connor, 2020). Common to these different conceptions of emergence are wholes that both depend on their parts and are autonomous from their parts (Humphreys & Imbert, 2013).

Beyond the general notion of emergence that is commonly attributed to hydrologic thresholds, multiple proposed typologies can be used to distinguish between emergent phenomena based on specific criteria (Fromm, 2005; Wilson, 2016). Some typologies characterize two types of emergence – weak emergence and strong emergence (e.g., Chalmers, 2006; Wilson, 2016). Others have proposed typologies that include nominal, weak, and strong emergence (Bedau, 1997) or weak, ontological, and strong emergence (Gillett, 2016). There are even more complex typologies of emergence that include taxa of simple, weak, multiple, and strong emergence as well as multiple sub-taxa (Fromm, 2005). Furthermore, some taxonomies that share the same taxa names do not share the same taxa definitions. Indeed, emergent phenomena can be typified in a variety of ways - however, such an enterprise requires an in-depth understanding of system dynamics, including the identification and characterization of component interactions (Fromm, 2004).

Hydrologic thresholds observed at the hillslope or catchment scale have not been characterized using one of these typologies and at this time, such an effort may be fruitless – except for intensely studied hillslopes and small catchments, our knowledge of component interactions at the next smaller spatial scale is incomplete. Most conceptualizations of hydrologic thresholds feature response magnitude as a function of a single meteorological factor. While one approach for considering the simultaneous effects of two, potentially interacting, factors on

threshold behaviour was explored in Chapter 5, the dimensionality of that approach could be expanded in future work to consider the influence of multiple, potentially interacting, factors on hydrologic response (Lintz et al., 2011). This would allow for a more complete understanding of how hydrologic thresholds might fit into one of the available typologies of emergence, which could offer a valuable framework for a unified threshold-based hydrological theory.

6.3.4 Perceptions of threshold behaviour versus the conceptual threshold definition

In hydrology, thresholds are conceptually defined as a critical moment in time or point in space at which runoff behaviour rapidly changes (Ali et al., 2013; Phillips, 2006). In Chapter 3, only relationships with hockey-stick shape threshold behaviour were deemed threshold mediated. These thresholds were identified via linear piecewise regression analysis using criteria that were stricter than any other criteria presented in the hydrologic threshold literature. These hockey-stick thresholds are one operational threshold definition that is associated with a change in the relationship slope. This type of threshold has been the focus of most threshold-based research in hydrology, with specific attention being drawn to relationships that exhibit very abrupt changes in slope (Ali et al., 2013, 2015; Detty & McGuire, 2010; Tromp-van Meerveld & McDonnell, 2006b, 2006a). While hockey-stick thresholds have certainly received the most attention in the literature, other operational threshold definitions exist, including those that are described using diagnostic shapes (Ali et al., 2013) and those that are associated with changes in response variance. In Chapter 3, multiple relationships exhibited variance collapse, where variability in response sharply declined once a critical input value was reached. This behaviour was not considered threshold mediated, even though it satisfies the conceptual definition of a hydrologic

threshold. This shows that multiple operational threshold definitions exist, and they tend to be narrower than the conceptual threshold definition, which does not include notions of abruptness, shape, or slope. Future studies should, therefore, explore more holistic methodologies and tools that might help identify the broad range of threshold behaviours that adhere to the conceptual threshold definition. Furthermore, in Chapter 5, examples of threshold behaviour that varied in abruptness across entire 2D response curves and 3D response surfaces (i.e., threshold strength) were observed. This indicates that in pursuit of a unified threshold-based hydrologic theory, work is needed to assess when a threshold is abrupt enough to be considered an important determinant for understanding hillslope- and catchment-scale hydrologic response.

6.4 Novel contributions and remaining challenges

The overarching goal of this thesis was to contribute to a unified threshold-based hydrological theory. It built upon previous studies that have demonstrated that characterizing threshold behaviour is key to our current understanding of hydrological processes at the hillslope and catchment scales (Ali et al., 2013; Detty & McGuire, 2010; Saffarpour et al., 2016; Spence, 2010; Weiler et al., 2006). However, as demonstrated throughout this thesis, our current understanding of threshold behaviour has limitations. Novel contributions of this thesis that resolve some of these knowledge gaps are highlighted below:

- In Chapter 2, univariate and multivariate statistical techniques were used to characterize variability in hydrologic response and a range of response controls for twenty-one sites across seven diverse study areas (research objective 1). These analyses showed

considerable temporal variability in event response and illustrated the importance of considering the influence of intensity-driven processes on hydrologic response.

- In Chapter 3, a robust and consistent approach was taken to assess the ubiquity of threshold behaviour across a range of study areas (research objective 2). This work was unprecedented in virtue of the breadth of meteorological factors and response metrics considered as well as the assessment of antecedent conditions as a control on threshold behaviour. Threshold behaviour was observed across twenty out of the twenty-one sites considered and in addition to rainfall depth, rainfall intensity and hydrologic abstractions caused by evapotranspiration proved to be important controls on threshold behaviour for some sites. The variety of sites and threshold behaviours assessed allowed for the first typology of threshold dynamics to be proposed.
- In Chapter 4, a model evaluation was performed that assessed behavioural model simulations on their ability to predict key catchment functions. This was the first study to explicitly incorporate emergent threshold information into the model evaluation process (research objective 3). This work showed that emergent properties can be powerful in identifying behavioural simulations that adequately predict real-world processes and that the enhanced use of readily available data is an effective mode of model evaluation.
- In Chapter 5, nonlinearities in hydrologic response were assessed as a function of two meteorological factors (research objective 4). This was the first study to consider the influence of potential meteorological factor interactions on threshold mediated hydrologic response. Results of this analysis showed that while 2D thresholds are typically more abrupt, they are unable to characterize factor interactions that can lead to complex and highly nonlinear response surfaces. These findings challenge hydrologists to

consider a wider range of threshold behaviours and to evaluate response nonlinearities as a function of multiple, potentially interacting meteorological factors.

This thesis led to many interesting results that encouragingly bring us closer to a unified threshold-based hydrologic theory. There are, however, some remaining knowledge gaps that were not explored in this thesis. For one, very few studies have considered seasonal or long-term variability in threshold behaviour (e.g., Scaife & Band, 2017) and as suggested in Chapter 3 of this thesis, seasonally variable factors may influence nonlinear response and may affect key features that we currently use to characterize 2D thresholds, like the relationship slope following threshold exceedance. Additionally, there is a need to better understand how data amount and data distribution affect the identification and interpretation of threshold behaviour. For example, for a site with a specific event size distribution, it is unknown which portion of that distribution needs to be known for thresholds to be effectively characterized. It is also hoped that some of the questions that were explored in Section 6.3 – and others – can galvanize the hydrologic community into drafting a list of the additional investigations that are needed before the goal of a unified threshold-based hydrological theory can be achieved.

6.5 References

Ali, G., L'Heureux, C., Roy, A., Turmel, M.-C., & Courchesne, F. (2011). Linking spatial patterns of perched groundwater storage and stormflow generation processes in a headwater forested catchment. *Hydrological Processes*, *25*(25), 3843–3857. https://doi.org/10.1002/hyp.8238

Ali, G., Oswald, C., Spence, C., Cammeraat, E., McGuire, K., Meixner, T., & Reaney, S. M. (2013). Towards a unified threshold-based hydrological theory: Necessary components and recurring challenges. *Hydrological Processes*, *27*(2), 313–318. https://doi.org/10.1002/hyp.9560

Ali, G., Tetzlaff, D., McDonnell, J., Soulsby, C., Carey, S., Laudon, H., McGuire, K., Buttle, J., Seibert, J., & Shanley, J. (2015). Comparison of threshold hydrologic response across northern catchments. *Hydrological Processes*, *29*(16), 3575–3591. https://doi.org/10.1002/hyp.10527

Bedau, M. A. (1997). Weak emergence. *Philosophical Perspectives*, *11*, 375–399.

Beven, K. J. (2000). Uniqueness of place and process representations in hydrological modelling. *Hydrology and Earth System Sciences*, *4*(2), 203–213. https://doi.org/10.5194/hess-4-203-2000

Blazkova, S., & Beven, K. (2009). A limits of acceptability approach to model evaluation and uncertainty estimation in flood frequency estimation by continuous simulation: Skalka catchment, Czech Republic. *Water Resources Research*, *45*(12). https://doi.org/10.1029/2007WR006726

Blume, T., Zehe, E., & Bronstert, A. (2007). Rainfall—Runoff response, event-based runoff coefficients and hydrograph separation. *Hydrological Sciences Journal*, *52*(5), 843–862. https://doi.org/10.1623/hysj.52.5.843

Cammeraat, L. H. (2002). A review of two strongly contrasting geomorphological systems within the context of scale. *Earth Surface Processes and Landforms*, *27*(11), 1201–1222. https://doi.org/10.1002/esp.421

Chalmers, D. J. (2006). Strong and weak emergence. *The Re-Emergence of Emergence*, 244–256.

Detty, J. M., & McGuire, K. J. (2010). Threshold changes in storm runoff generation at a till-mantled headwater catchment. *Water Resources Research; Washington*, *46*(7). http://dx.doi.org.uml.idm.oclc.org/10.1029/2009WR008102

Dingman, S. L. (2015). *Physical hydrology*. Waveland press.

Dunn, S. M. (1999). Imposing constraints on parameter values of a conceptual hydrological model using baseflow response. *Hydrology and Earth System Sciences Discussions*, *3*(2), 271–284.

Freer, J., McDonnell, J. J., Beven, K. J., Peters, N. E., Burns, D. A., Hooper, R. P., Aulenbach, B., & Kendall, C. (2002). The role of bedrock topography on subsurface storm flow. *Water Resources Research*, *38*(12). https://doi.org/10.1029/2001WR000872

Fromm, J. (2004). *The emergence of complexity* (p. 23). Kassel: Kassel university press.

Fromm, J. (2005). Types and forms of emergence. *ArXiv Preprint Nlin/0506028*.

Gillett, C. (2016). *Reduction and emergence in science and philosophy*. Cambridge University Press.

Graham, C. B., & McDonnell, J. J. (2010). Hillslope threshold response to rainfall: (2) Development and use of a macroscale model. *Journal of Hydrology*, *393*(1–2), 77–93. https://doi.org/10.1016/j.jhydrol.2010.03.008

Graham, C. B., Woods, R. A., & McDonnell, J. J. (2010). Hillslope threshold response to rainfall: (1) A field based forensic approach. *Journal of Hydrology*, *393*(1–2), 65–76. https://doi.org/10.1016/j.jhydrol.2009.12.015

Herbst, M., Casper, M. C., Grundmann, J., & Buchholz, O. (2009). Comparative analysis of model behaviour for flood prediction purposes using Self-Organizing Maps. *Natural Hazards and Earth System Sciences*, *9*(2), 373–392. https://doi.org/10.5194/nhess-9-373-2009

Humphreys, P., & Imbert, C. (2013). *Models, Simulations, and Representations* (Vol. 9). Routledge.

James, A., & Roulet, N. (2007). Investigating hydrologic connectivity and its association with threshold change in runoff response in a temperate forested watershed. *Hydrological Processes*, *21*(25), 3391–3408. https://doi.org/10.1002/hyp.6554

Kim, H. J., Sidle, R. C., Moore, R. D., & Hudson, R. (2004). Throughflow variability during snowmelt in a forested mountain catchment, coastal British Columbia, Canada. *Hydrological Processes*, *18*(7), 1219–1236. https://doi.org/10.1002/hyp.1396

Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research*, *42*(3), W03S04. https://doi.org/10.1029/2005WR004362

Kirchner, J. W. (2009). Catchments as simple dynamical systems: Catchment characterization, rainfall-runoff modeling, and doing hydrology backward. *Water Resources Research*, *45*(2). https://doi.org/10.1029/2008WR006912

Laudon, H., Sjöblom, V., Buffam, I., Seibert, J., & Mörth, M. (2007). The role of catchment scale and landscape characteristics for runoff generation of boreal streams. *Journal of Hydrology*, *344*(3), 198–209. https://doi.org/10.1016/j.jhydrol.2007.07.010

Lehmann, P., Hinz, C., McGrath, G., Tromp-van Meerveld, H. J., & McDonnell, J. J. (2007). Rainfall threshold for hillslope outflow: An emergent property of flow pathway

connectivity. *Hydrol. Earth Syst. Sci.*, *11*(2), 1047–1063. https://doi.org/10.5194/hess-11-1047-2007

Ley, R., Hellebrand, H., Casper, M. C., & Fenicia, F. (2016). Comparing classical performance measures with signature indices derived from flow duration curves to assess model structures as tools for catchment classification. *Hydrology Research*, *47*(1), 1–14. https://doi.org/10.2166/nh.2015.221

Lintz, H. E., McCune, B., Gray, A. N., & McCulloh, K. A. (2011). Quantifying ecological thresholds from response surfaces. *Ecological Modelling*, *222*(3), 427–436. https://doi.org/10.1016/j.ecolmodel.2010.10.017

Loague, K. M., & Freeze, R. A. (1985). A Comparison of Rainfall-Runoff Modeling Techniques on Small Upland Catchments. *Water Resources Research*, *21*(2), 229–248. https://doi.org/10.1029/WR021i002p00229

McDonnell, J., Sivapalan, M., Vaché, K., Dunn, S., Grant, G., Haggerty, R., Hinz, C., Hooper, R., Kirchner, J., Roderick, M. L., Selker, J., & Weiler, M. (2007). Moving beyond heterogeneity and process complexity: A new vision for watershed hydrology. *Water Resources Research*, *43*(7), W07301. https://doi.org/10.1029/2006WR005467

McMillan, H. K., Westerberg, I. K., & Krueger, T. (2018). Hydrological data uncertainty and its implications. *WIREs Water*, 5(6), e1319. https://doi.org/10.1002/wat2.1319

Mielko, C., & Woo, M. (2006). Snowmelt runoff processes in a headwater lake and its catchment, subarctic Canadian Shield. *Hydrological Processes*, *20*(4), 987–1000. https://doi.org/10.1002/hyp.6117

Mosley, M. P. (1979). Streamflow generation in a forested watershed, New Zealand. *Water Resources Research*, *15*(4), 795–806. https://doi.org/10.1029/WR015i004p00795

O'Connor, T. (2020). *Emergent Properties*. https://plato.stanford.edu/archives/fall2020/entries/properties-emergent/

Oswald, C., Richardson, M. C., & Branfireun, B. A. (2011). Water storage dynamics and runoff response of a boreal Shield headwater catchment. *Hydrological Processes*, *25*(19), 3042–3060. https://doi.org/10.1002/hyp.8036

Ottino, J. M. (2004). Engineering complex systems. *Nature*, *427*(6973), 399–399.

Penna, D., van Meerveld, H. J., Zuecco, G., Dalla Fontana, G., & Borga, M. (2016). Hydrological response of an Alpine catchment to rainfall and snowmelt events. *Journal of Hydrology*, *537*, 382–397. https://doi.org/10.1016/j.jhydrol.2016.03.040

Phillips, J. D. (2006). Evolutionary geomorphology: Thresholds and nonlinearity in landform response to environmental change. *Hydrol. Earth Syst. Sci.*, *10*(5), 731–742. https://doi.org/10.5194/hess-10-731-2006

Reaney, S. M., Bracken, L. J., & Kirkby, M. J. (2007). Use of the Connectivity of Runoff Model (CRUM) to investigate the influence of storm characteristics on runoff generation and connectivity in semi-arid areas. *Hydrological Processes*, *21*(7), 894–906. https://doi.org/10.1002/hyp.6281

Redding, T. E., & Devito, K. J. (2008). Lateral flow thresholds for aspen forested hillslopes on the Western Boreal Plain, Alberta, Canada. *Hydrological Processes*, *22*(21), 4287–4300. https://doi.org/10.1002/hyp.7038

Saffarpour, S., Western, A. W., Adams, R., & McDonnell, J. J. (2016). Multiple runoff processes and multiple thresholds control agricultural runoff generation. *Hydrology and Earth System Sciences*, 20(11), 4525–4545. https://doi.org/10.5194/hess-20-4525-2016

Scaife, C. I., & Band, L. E. (2017). Nonstationarity in threshold response of stormflow in southern Appalachian headwater catchments. *Water Resources Research*, *53*(8), 6579–6596. https://doi.org/10.1002/2017WR020376

Sidle, R. C., Noguchi, S., Tsuboyama, Y., & Laursen, K. (2001). A conceptual model of preferential flow systems in forested hillslopes: Evidence of self-organization. *Hydrological Processes*, *15*(10), 1675–1692.

Sidle, R. C., Tsuboyama, Y., Noguchi, S., Hosoda, I., Fujieda, M., & Shimizu, T. (2000). Stormflow generation in steep forested headwaters: A linked hydrogeomorphic paradigm. *Hydrological Processes*, *14*(3), 369–385. https://doi.org/10.1002/(SICI)1099-1085(20000228)14:3<369::AID-HYP943>3.0.CO;2-P

Spence, C. (2010). A Paradigm Shift in Hydrology: Storage Thresholds Across Scales Influence Catchment Runoff Generation. *Geography Compass*, *4*(7), 819–833. https://doi.org/10.1111/j.1749-8198.2010.00341.x

Tang, W., & Carey, S. K. (2017). HydRun: A MATLAB toolbox for rainfall–runoff analysis. *Hydrological Processes*, *31*(15), 2670–2682. https://doi.org/10.1002/hyp.11185

Tani, M. (1997). Runoff generation processes estimated from hydrological observations on a steep forested hillslope with a thin soil layer. *Journal of Hydrology*, *200*(1), 84–109. https://doi.org/10.1016/S0022-1694(97)00018-8

Tromp-van Meerveld, H. J., & McDonnell, J. J. (2006a). Threshold relations in subsurface stormflow: 1. A 147-storm analysis of the Panola hillslope. *Water Resources Research*, *42*(2). https://doi.org/10.1029/2004WR003778

Tromp-van Meerveld, H. J., & McDonnell, J. J. (2006b). Threshold relations in subsurface stormflow: 2. The fill and spill hypothesis. *Water Resources Research*, *42*(2). https://doi.org/10.1029/2004WR003800

Weiler, M. (2005). An infiltration model based on flow variability in macropores: Development, sensitivity analysis and applications. *Journal of Hydrology*, *310*(1–4), 294–315. https://doi.org/10.1016/j.jhydrol.2005.01.010

Weiler, M., McDonnell, J. J., Meerveld, I. T., & Uchida, T. (2006). Subsurface Stormflow. In *Encyclopedia of Hydrological Sciences*. John Wiley & Sons, Ltd. https://doi.org/10.1002/0470848944.hsa119

Westerberg, I. K., Guerrero, J. L., Younger, P. M., Beven, K. J., Seibert, J., Halldin, S., Freer, J. E., & Xu, C. Y. (2011). Calibration of hydrological models using flow-duration curves. *Hydrology and Earth System Sciences*, *15*(7), 2205–2227. https://doi.org/10.5194/hess-15-2205-2011

Whipkey, R. Z. (1965). Subsurface Stormflow from Forested Slopes. *International Association of Scientific Hydrology. Bulletin*, *10*(2), 74–85. https://doi.org/10.1080/02626666509493392

Wilson, J. (2016). Metaphysical emergence: Weak and strong. In *Metaphysics in contemporary physics* (pp. 345–402). Brill Rodopi.

Winter, T. C. (1981). Uncertainties in Estimating the Water Balance of Lakes1. *JAWRA Journal of the American Water Resources Association*, 17(1), 82–115. https://doi.org/10.1111/j.1752-1688.1981.tb02593.x

Yang, T.-C., Yu, P.-S., Kuo, C.-M., & Wang, Y.-C. (2004). Application of Fuzzy Multiobjective Function on Storm-Event Rainfall-Runoff Model Calibration. *Journal of Hydrologic Engineering*, *9*(5), 440–445. https://doi.org/10.1061/(ASCE)1084-0699(2004)9:5(440)

Yates, F. E. (2012). *Self-organizing systems: The emergence of order*. Springer Science & Business Media.

Yilmaz, K. K., Gupta, H. V., & Wagener, T. (2008). A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model. *Water Resources Research*, *44*(9). https://doi.org/10.1029/2007WR006716

Yu, P.-S., & Yang, T.-C. (2000). Using synthetic flow duration curves for rainfall–runoff model calibration at ungauged sites. *Hydrological Processes*, *14*(1), 117–133. https://doi.org/10.1002/(SICI)1099-1085(200001)14:1<117::AID-HYP914>3.0.CO;2-Q

APPENDICES

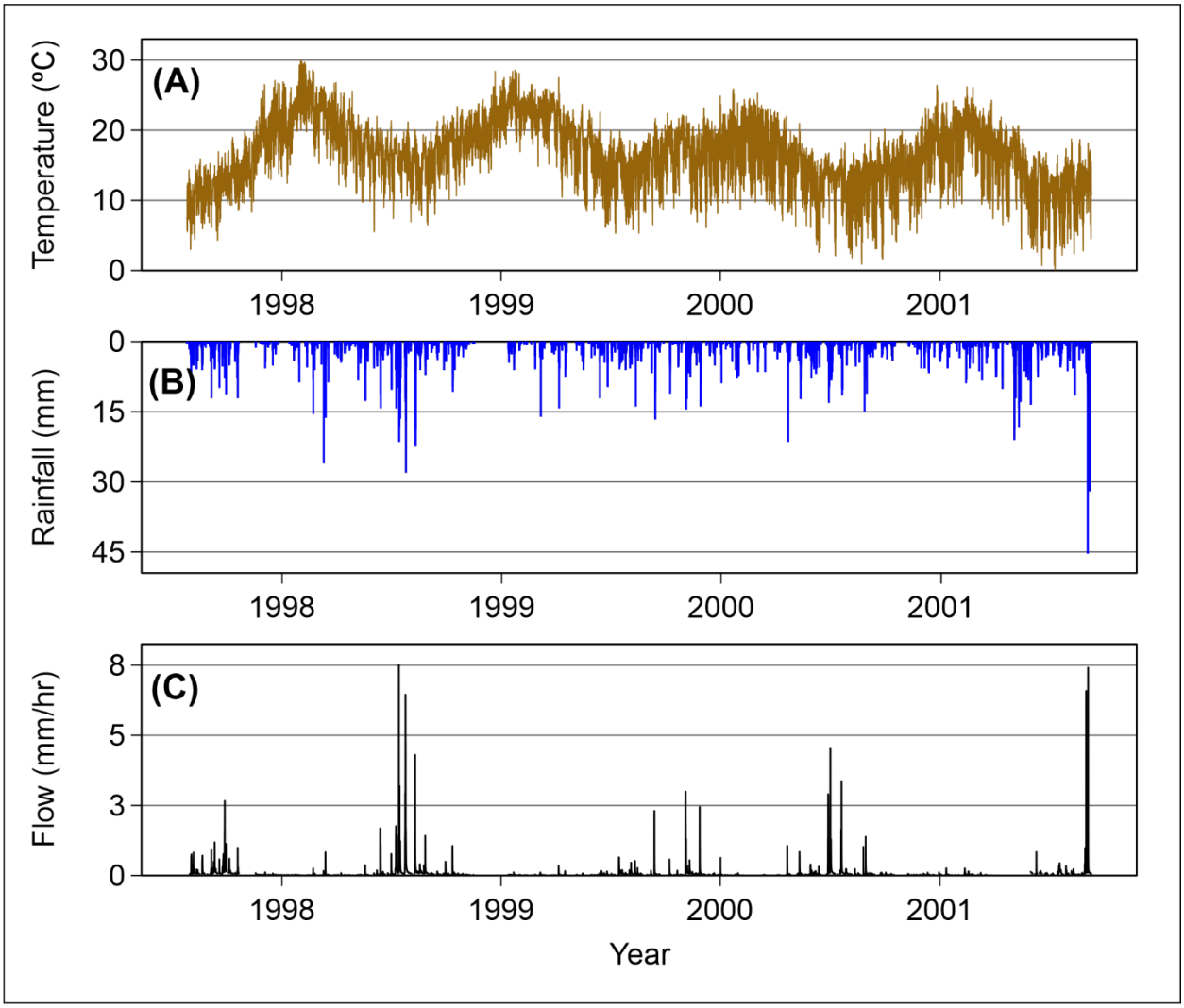
APPENDIX A. Supplemental Materials Related to Chapter 3

Appendix A-1. Site-specific threshold detection frequencies for all 217 meteorological factor – response metric pairs when using five different values for the below- and above-threshold slope percent difference criterion. The total number of possible thresholds is shown in the 'Total' row. The average threshold frequency for each category is shown in the 'Average' row.

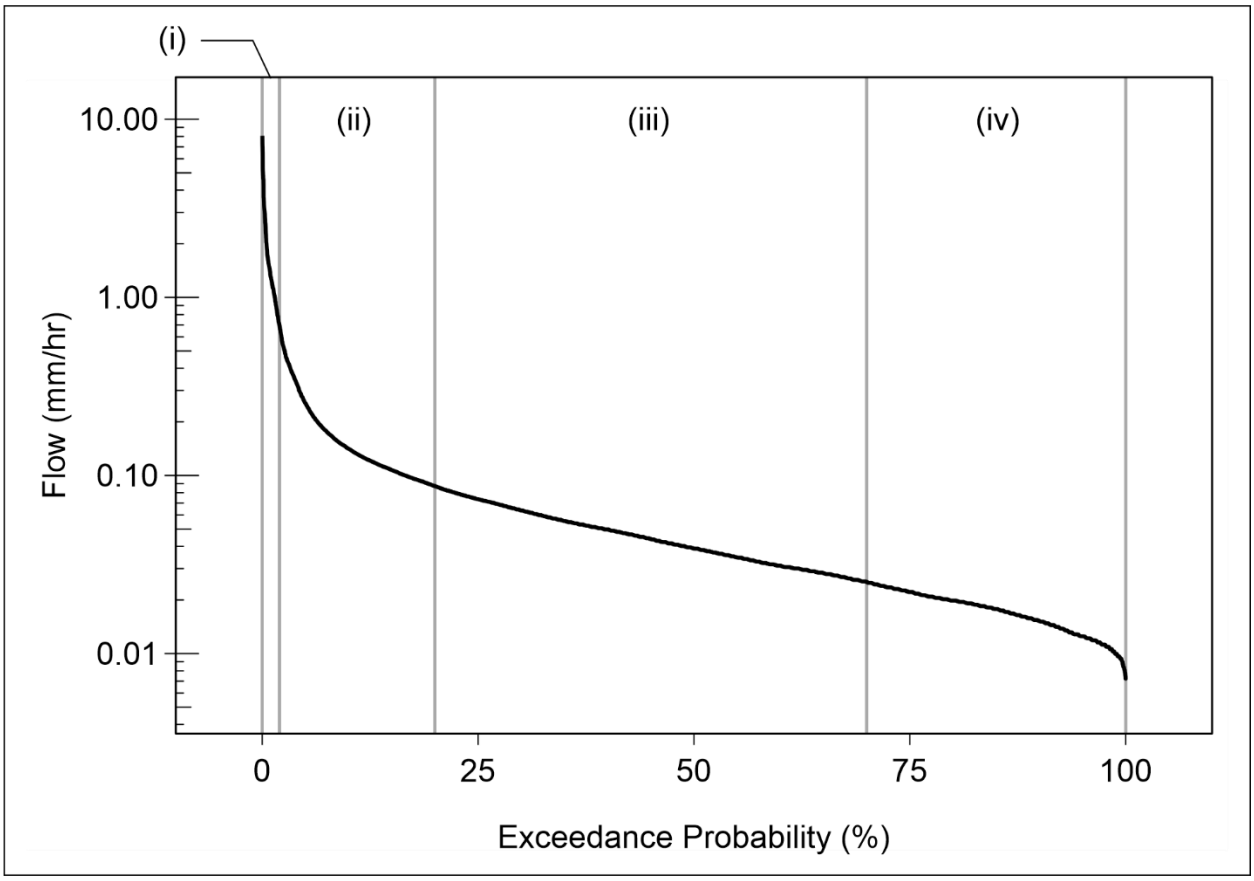
% difference	10%	20%	30%	40%	50%
Total	217	217	217	217	217
Site	Frequency (%)	Frequency (%)	Frequency (%)	Frequency (%)	Frequency (%)
PMRW	12	12	12	12	12
HRM	8	8	8	8	8
UP1	13	13	13	13	13
TRC	7	7	7	7	7
CCW	4	4	4	4	4
MRC1	9	9	9	9	9
MRC2	0	0	0	0	0
MRC3	2	2	2	1	1
MRC4	3	3	3	3	3
MRC5	5	5	5	5	5
MRC6	7	7	7	7	7
MRC7	7	7	6	6	6
MRC8	3	3	3	3	3
HJA1	8	8	7	7	7
HJA2	5	5	5	5	5
HJA3	4	4	4	4	4
HJA4	4	4	4	4	3
HJA5	4	4	4	4	3
HJA6	5	5	4	4	3
HJA7	5	5	5	5	5
HJA8	4	4	4	4	4
Average	6	6	6	5	5

APPENDIX B. Supplemental Materials Related to Chapter 4

Appendix B-1. Timeseries (July 1997 – September 2001) of the hourly temperature (A), rainfall (B), and flow (C) data that were used in this study.



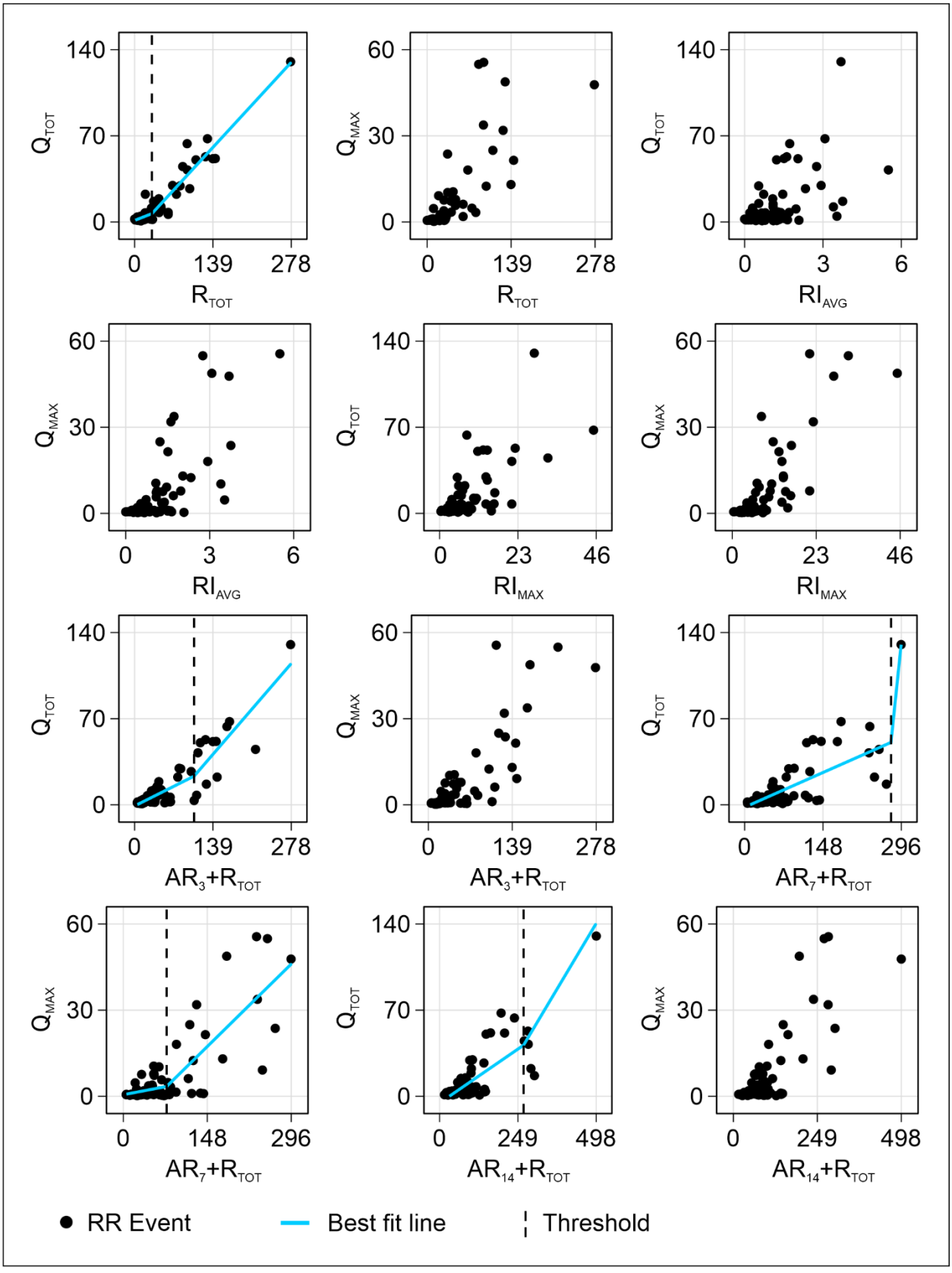
Appendix B-2. Flow duration curve of the observed hydrograph with key segments delineated by vertical lines: very high flows (i), high and medium flows (ii), the middle-slope, and (iii) low flows (iv).



Appendix B-3. Summary statistics of event (n = 74) lag-to-peak (T_{LP}) and centroid lag-to-peak (T_{LPC}). SD: standard deviation and CV: coefficient of variation.

	T_{LP}	T_{LPC}
Mean	29.34	9.58
Median	26.50	8.31
Minimum	5.00	0.17
Maximum	105.00	35.20
SD	20.27	6.66
CV	0.69	0.70

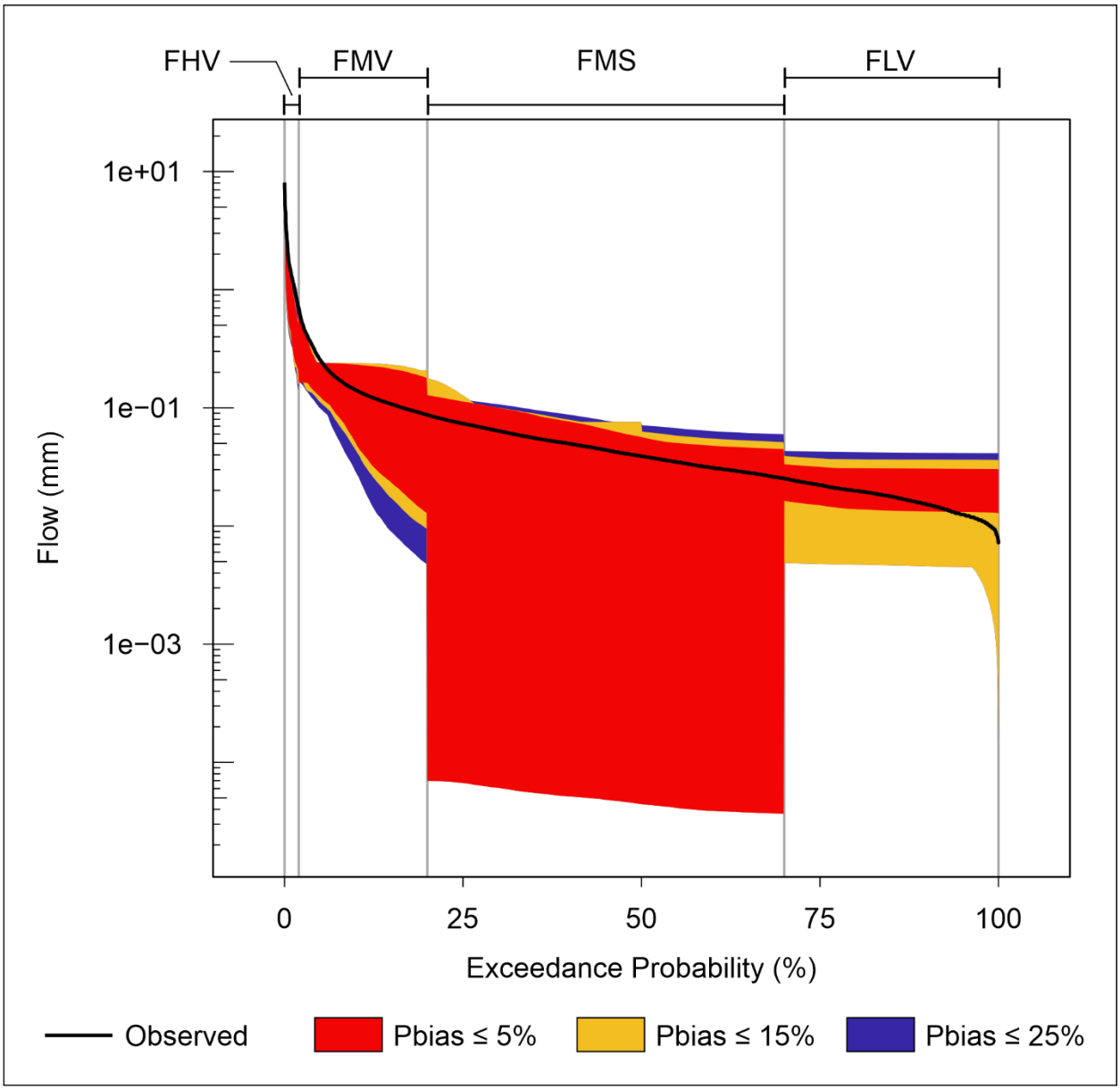
Appendix B-4. Observed, rainfall-runoff relationships involving input-output pairs. Black dots represent individual rainfall-runoff (RR) events from the observed data. For cases where threshold behaviour was observed, teal lines indicate the piecewise linear model of the observed data and dashed black lines indicate the threshold value.


Q_{TOT}
R_{TOT}
Q_{MAX}
RI_{AVG}
RI_{MAX}
AR_3+R_{TOT}
AR_7+R_{TOT}
$AR_{14}+R_{TOT}$
RR Event
Best fit line
Threshold

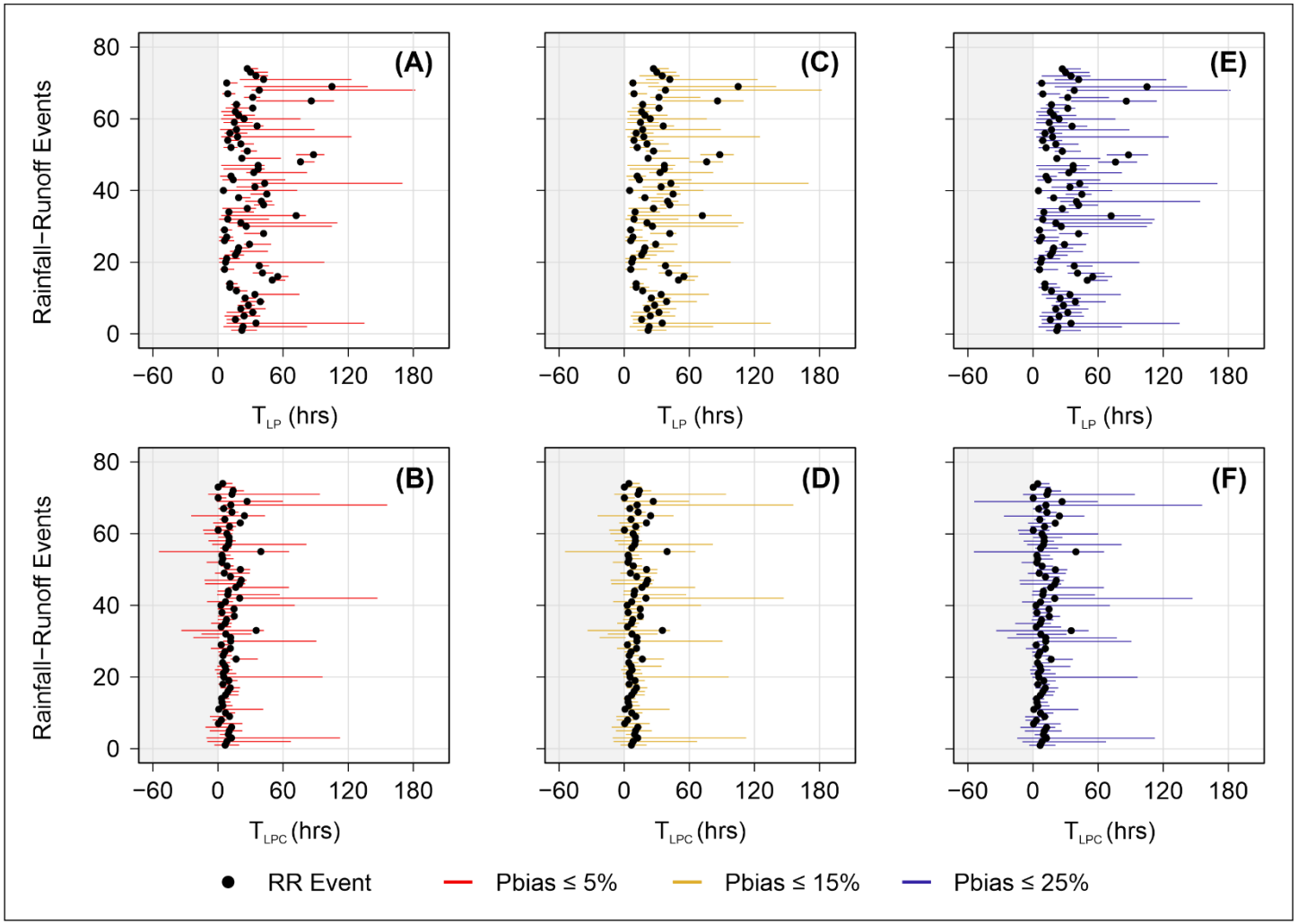
Appendix B-5. The number (percentage shown in brackets) of behavioural simulations that met the 5%, 15%, and 25% Pbias criteria for different measures of bias.

	Pbias ≤ 5%	Pbias ≤ 15%	Pbias ≤ 25%
	Simulations (%)	Simulations (%)	Simulations (%)
Flow duration curves			
FLV	253 (3.0)	678 (8.1)	1127 (13.5)
FMS	325 (3.9)	978 (11.7)	1625 (19.5)
FMV	704 (8.4)	2137 (25.6)	3500 (42.0)
FHV	1033 (12.4)	3054 (36.6)	4811 (57.7)
All FDC biases	3 (0.0)	94 (1.1)	418 (5.0)
Event-specific response timing metrics			
T_{LP}	1159 (13.9)	3714 (44.5)	5976 (71.7)
T_{LPC}	336 (4.0)	968 (11.6)	1640 (19.7)
All response timing biases	336 (4.0)	968 (11.6)	1640 (19.7)
Event rainfall threshold relationship			
R_{TOT}, Q_{TOT}	1602 (19.2)	5032 (60.3)	7449 (89.3)
Antecedent plus event rainfall threshold relationships			
AR_3+R_{TOT}, Q_{TOT}	2188 (26.2)	6307 (75.6)	8101 (97.2)
AR_7+R_{TOT}, Q_{TOT}	4915 (58.9)	7633 (91.5)	7844 (94.1)
AR_7+R_{TOT}, Q_{MAX}	2487 (29.8)	5619 (67.4)	6878 (82.5)
AR_{14}+R_{TOT}, Q_{TOT}	4511 (54.1)	8077 (96.9)	8329 (99.9)
All AR_X+R_{TOT} threshold biases	148 (1.8)	4081 (48.9)	6482 (77.7)
All threshold biases	3 (0.0)	2272 (27.3)	5674 (68.0)

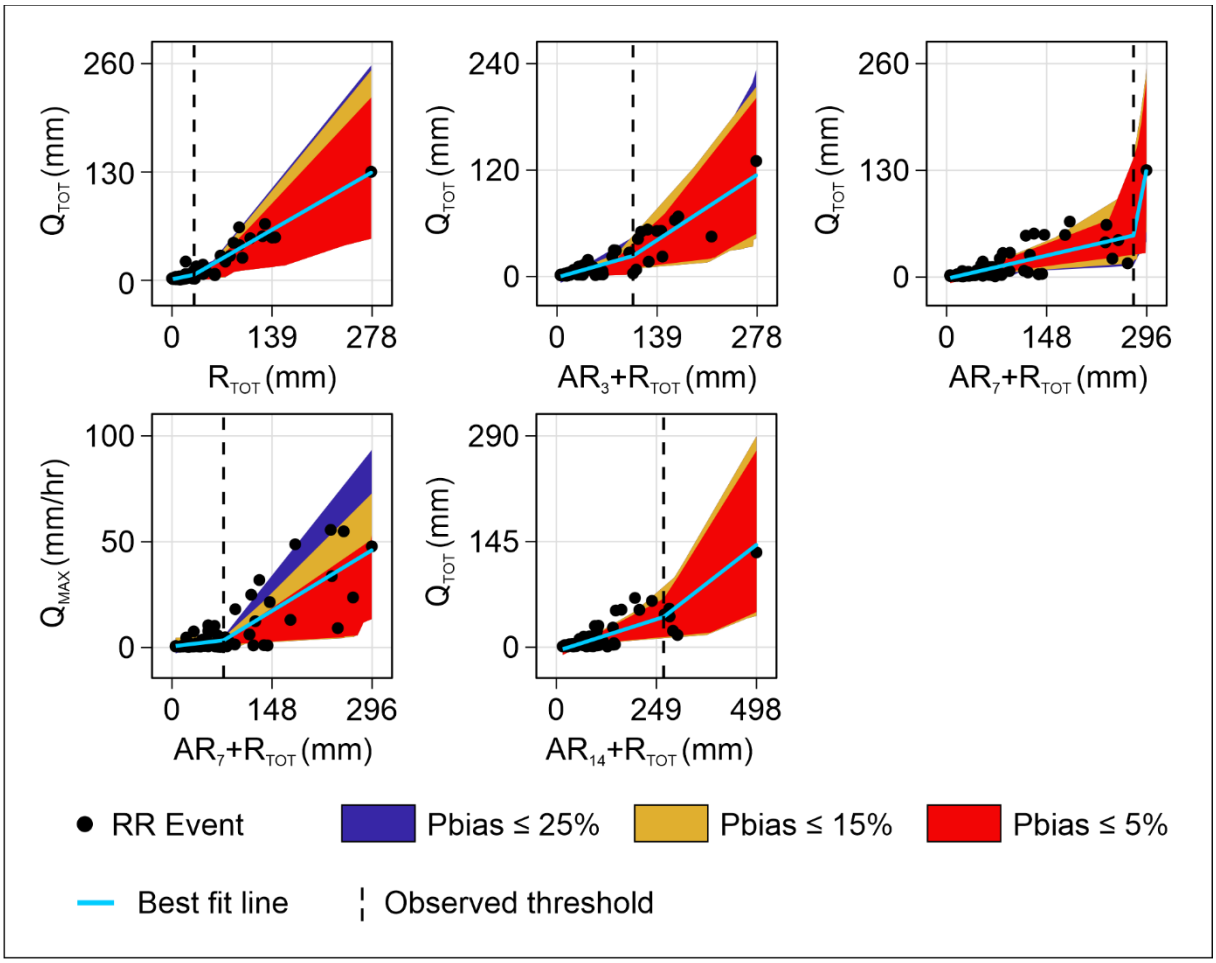
Appendix B-6. Observed flow duration curve and range of simulated flow duration curves associated with behavioural simulations with FLV, FMS, FMV, or FHV ≤ 5%, ≤ 15%, and ≤ 25% (shown in red, yellow, and blue, respectively).



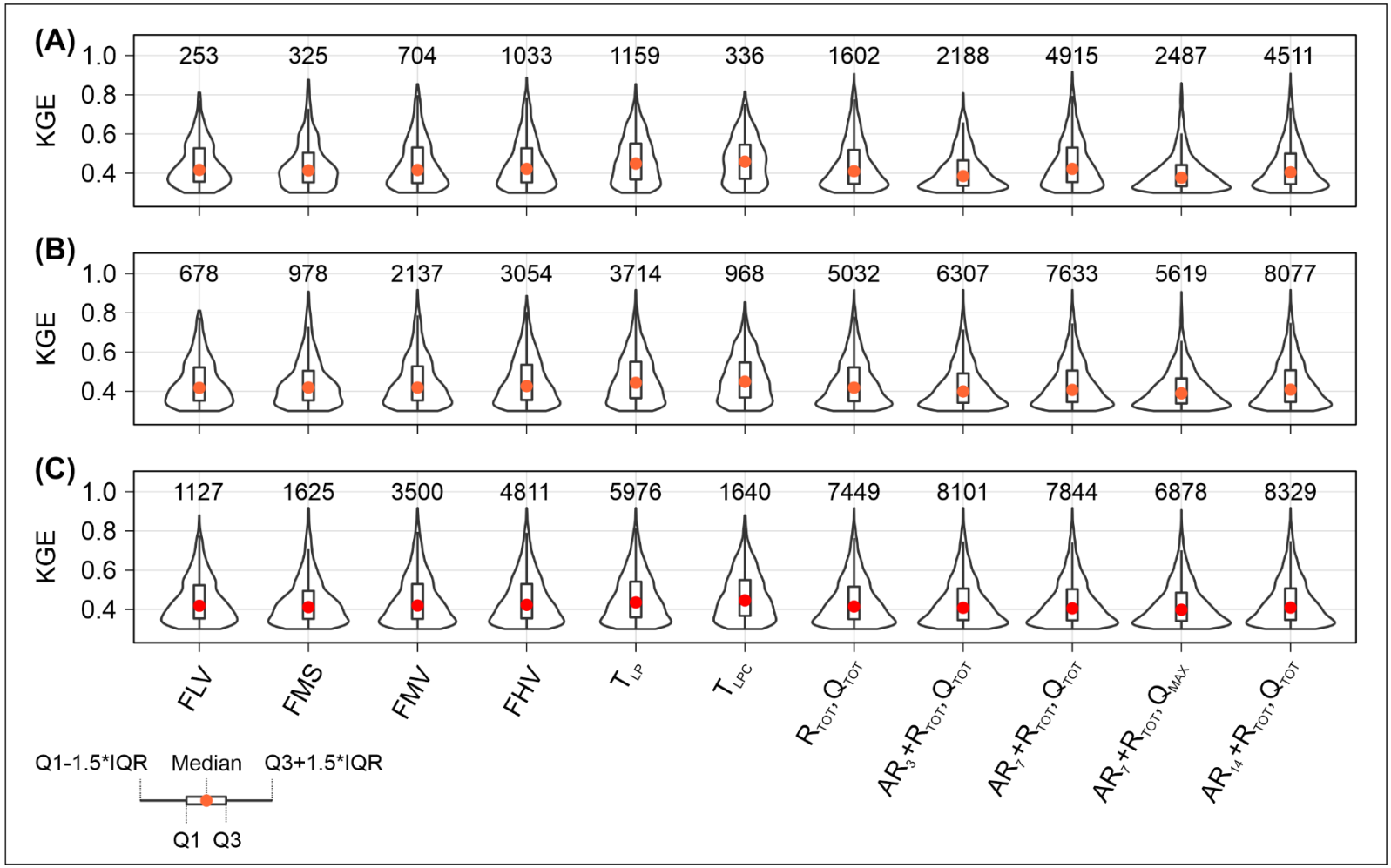
Appendix B-7. T_{LP} and T_{LPC} of rainfall-runoff (RR) events (n = 74) in the observation data and colour-coded bars showing the T_{LP} (A, B, and C) and T_{LPC} (D, E, and F) ranges across all events for behavioural simulations with Pbias ≤ 5% (A and D), ≤ 15% (B and E), and ≤ 25% (C and F). Plot areas associated with negative response timing metric values are shaded grey.



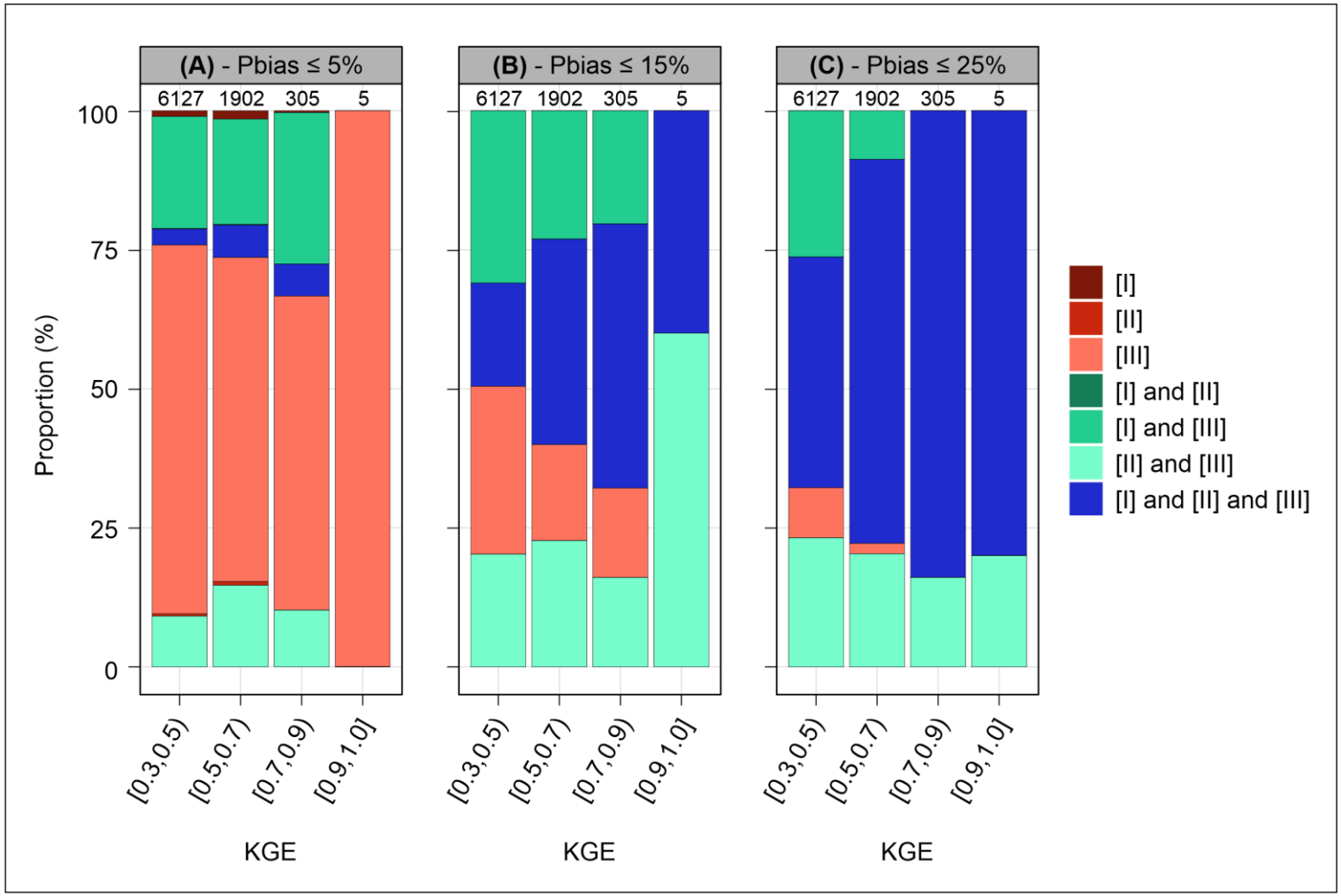
Appendix B-8. Thresholds associated with input-output pairs. In each panel, black dots represent individual rainfall-runoff (RR) events. The teal line indicates the piecewise linear model of the observed data and the dashed black line indicates the observed threshold value. Colour-coded envelopes show the ranges of the piecewise linear models for behavioural simulations with compound Pbias ≤ 5%, ≤ 15%, and ≤ 25%.



Appendix B-9. Violin plots showing the distribution of KGE scores for behavioural simulations that met the 5% (A), 15% (B), and 25% (C) Pbias criterion for different measures of bias, with the x-axis showing each measure of bias. The number of simulations is indicated above each violin plot.



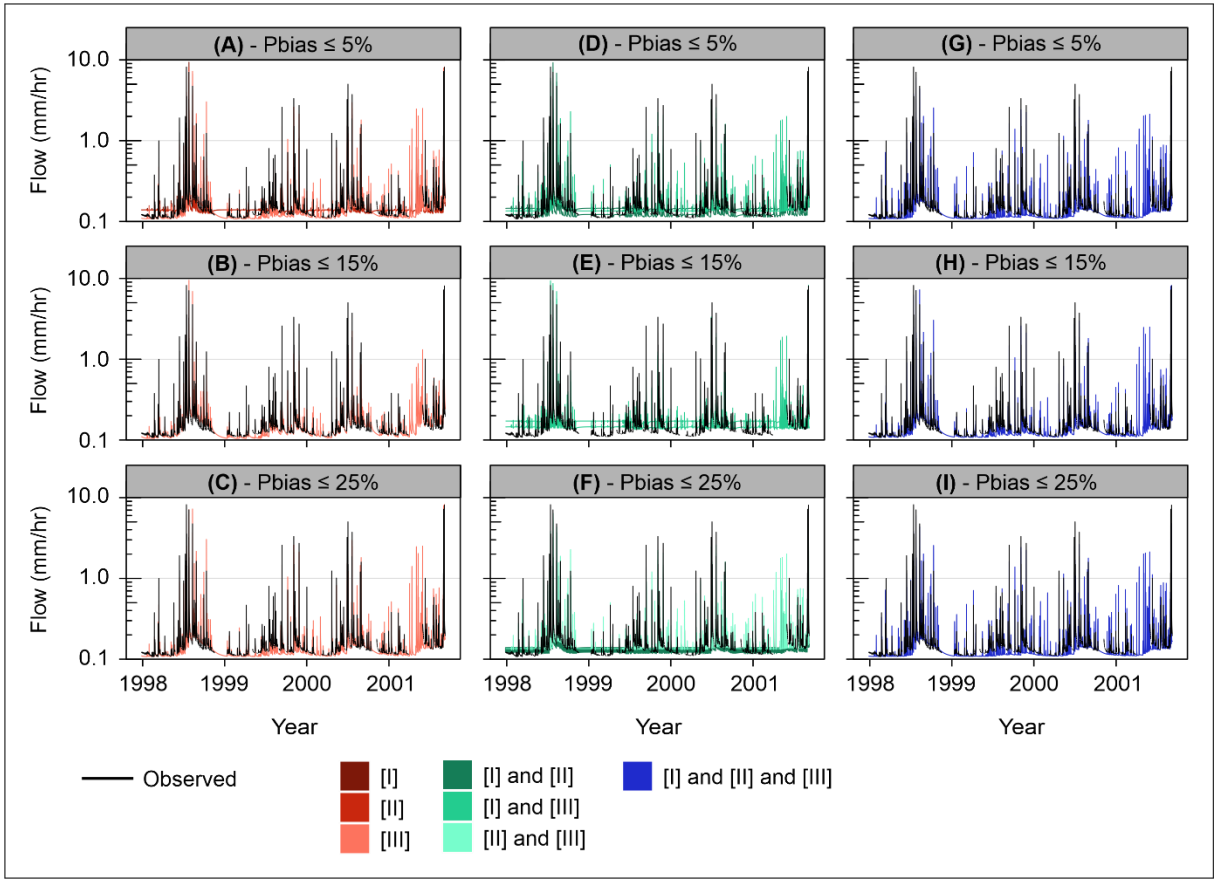
Appendix B-10. Bar charts showing the proportion of behavioural simulations at different KGE score ranges with Pbias ≤ 5% (A), ≤ 15% (B), and ≤ 25% (C) for [I] FDC biases only, [II] timing biases only, [III] threshold biases only, or biases related to combinations of these descriptors. The number of behavioural simulations for each KGE score range is shown above each bar.



Appendix B-11. Minimum, median, and maximum KGE scores of behavioural simulation subsets that had Pbias ≤ 5%, ≤ 15%, and ≤ 25% for [I] FDC biases only, [II] timing biases only, [III] thresholds biases only, or biases related to combinations of these descriptors. Columns are grouped by the number of biases. Simulations are separated based on behavioural simulation subsets. Simulations with the maximum KGE score for each group are bolded and their flow timeseries are shown in Appendix B-12. "NA": cases where the group range exceeds the possible number of bias measures. "-": cases where no simulation of a given subset reproduced the associated number of biases.

# of Biases	1-3			4-6			7-9			10-11		
	KGE											
	Minimum	Median	Maximum	Minimum	Median	Maximum	Minimum	Median	Maximum	Minimum	Median	Maximum
	Pbias ≤ 5%											
[I]	0.31	0.44	**0.76**	-	-	-	NA	NA	NA	NA	NA	NA
[II]	0.30	0.42	**0.69**	NA	NA	NA	NA	NA	NA	NA	NA	NA
[III]	0.30	0.40	**0.92**	0.30	0.38	0.83	NA	NA	NA	NA	NA	NA
[I] and [II]	0.31	0.36	**0.66**	-	-	-	NA	NA	NA	NA	NA	NA
[I] and [III]	0.30	0.43	**0.89**	0.30	0.39	0.82	0.40	0.40	0.40	NA	NA	NA
[II] and [III]	0.30	0.45	**0.85**	0.30	0.44	0.82	-	-	-	NA	NA	NA
[I] and [II] and [III]	0.30	0.60	0.82	0.30	0.46	**0.85**	0.31	0.52	0.67	-	-	-
Pbias ≤ 15%												
[I]	-	-	-	-	-	-	NA	NA	NA	NA	NA	NA
[II]	-	-	-	NA	NA	NA	NA	NA	NA	NA	NA	NA
[III]	0.30	0.38	**0.87**	0.30	0.37	0.85	NA	NA	NA	NA	NA	NA
[I] and [II]	-	-	-	-	-	-	NA	NA	NA	NA	NA	NA
[I] and [III]	0.30	0.47	0.80	0.30	0.40	**0.89**	0.30	0.38	0.71	NA	NA	NA
[II] and [III]	0.30	0.48	0.83	0.30	0.42	**0.91**	0.30	0.38	0.76	NA	NA	NA
[I] and [II] and [III]	0.36	0.46	0.62	0.30	0.51	**0.92**	0.30	0.44	0.85	0.39	0.42	0.45

Pbias ≤ 25%												
[I]	-	-	-	-	-	-	NA	NA	NA	NA	NA	NA
[II]	-	-	-	NA	NA	NA	NA	NA	NA	NA	NA	NA
[III]	0.30	0.35	0.59	0.30	0.35	**0.69**	NA	NA	NA	NA	NA	NA
[I] and [II]	-	-	-	-	-	-	NA	NA	NA	NA	NA	NA
[I] and [III]	0.51	0.54	0.58	0.30	0.38	**0.66**	0.30	0.36	0.66	NA	NA	NA
[II] and [III]	-	-	-	0.30	0.40	**0.91**	0.30	0.40	0.76	NA	NA	NA
[I] and [II] and [III]	0.39	0.42	0.45	0.30	0.49	0.90	0.30	0.45	**0.92**	0.30	0.46	0.77



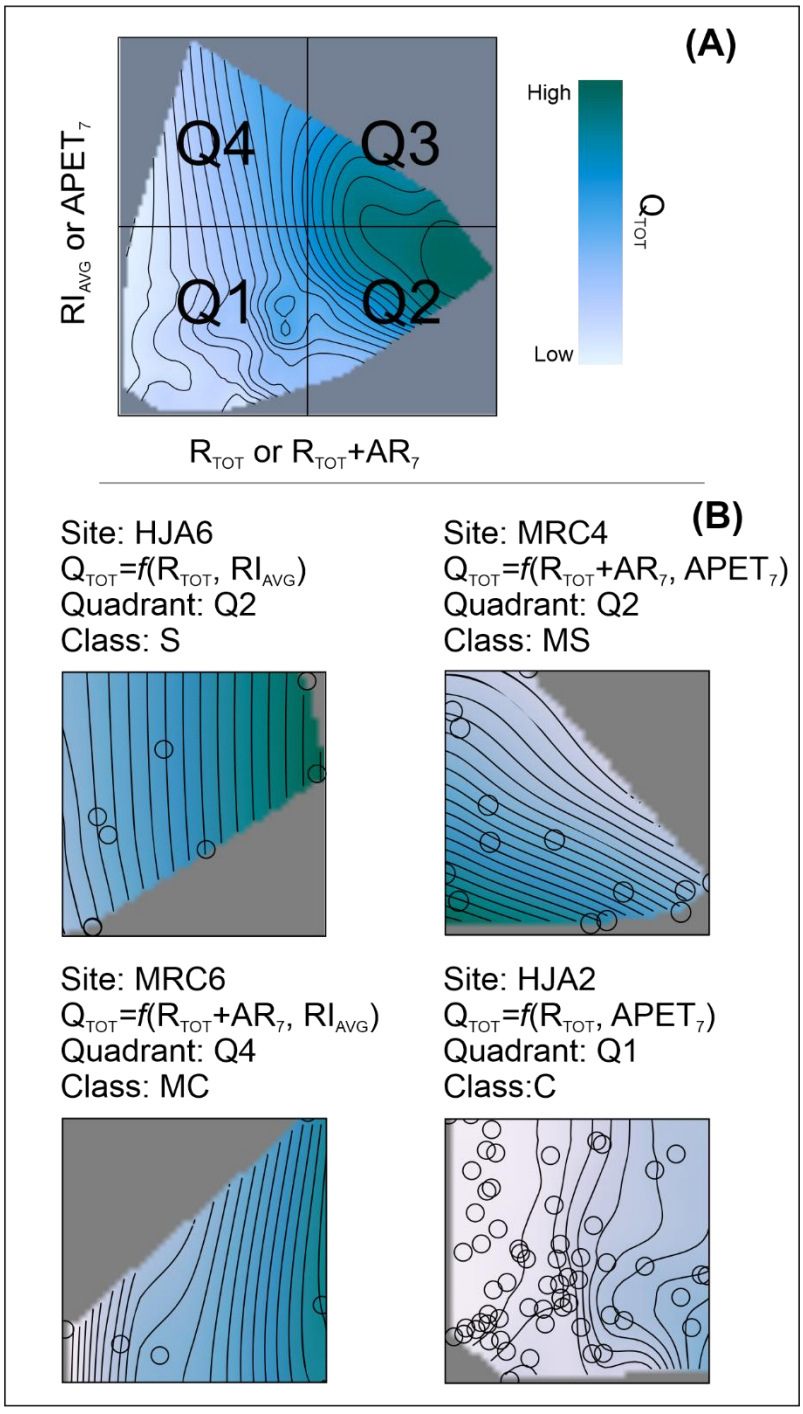
Appendix B-12. Observed and modelled flow timeseries of behavioural simulations that met the 5% (A, D, and G), 15% (B, E, and H), and 25% (C, F, and I) Pbias criteria for [I] FDC biases only, [II] timing biases only, [III] threshold biases only (A, B, and C), or low biases related to combinations of two of these descriptors (D, E, and F), or low biases related to all three descriptors (G, H, and I). Modelled flow timeseries are of behavioural simulations with the maximum KGE scores that had low biases related to different descriptors (shown in bold in Appendix B-11).

Appendix B-13. p-values of two-sample Kolmogorov-Smirnov tests that were performed to compare the parameter value distributions of behavioural simulations that met the 5%, 15%, and 25% Pbias criteria for different measures of bias against that of the remaining behavioural simulations. "-": cases with too few simulations to perform statistical testing.

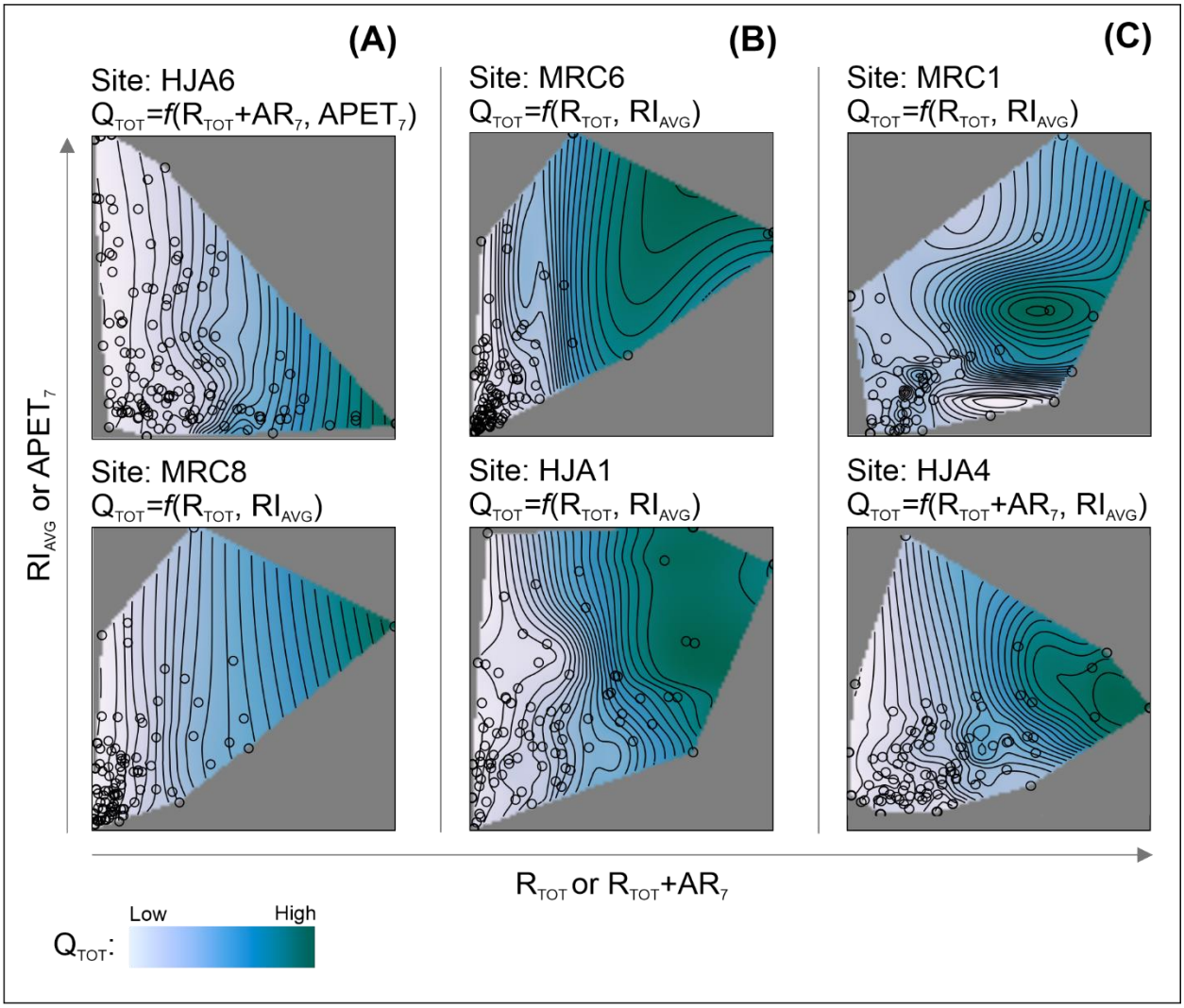
	Pbias ≤ 5%					Pbias ≤ 15%					Pbias ≤ 25%				
Parameter	X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5
Flow duration curve															
FLV	0.09	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FMS	0.00	0.00	0.00	0.29	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FMV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FHV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
All FDC biases	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Event-specific response timing metric															
T_{LP}	0.00	0.00	0.00	0.00	0.16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
T_{LPC}	0.21	0.03	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.16
All response timing biases	0.00	0.00	0.00	0.00	0.16	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00
Event rainfall threshold relationship															
R_{TOT}, Q_{TOT}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.00
Event plus antecedent rainfall threshold relationships															
AR_3+R_{TOT}, Q_{TOT}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AR_7+R_{TOT}, Q_{TOT}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
AR_7+R_{TOT}, Q_{MAX}	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AR_{14}+R_{TOT}, Q_{TOT}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.48	0.49
All AR_X+R_{TOT} biases	0.00	0.00	0.00	0.00	0.00	0.64	0.79	0.18	0.42	0.33	-	-	-	-	-

APPENDIX C. Supplemental Materials Related to Chapter 5

Appendix C-1. (A) Response surface separated into four quadrants. (B) Examples of contour pattern classifications (S: straight; MS: mostly straight; MC: mostly curved; C: curved). Axis ticks and labels are omitted for readability.



Appendix C-2. Examples of response surfaces modelled in this study: (A) linear gradient; (B) angular gradient; and (C) radial gradient. Axis ticks and labels are omitted for readability.



Appendix C-3. Examples of response surfaces modelled in this study with threshold fronts highlighted in yellow: (A) linear gradient; (B) angular gradient; and (C) radial gradient. Axis ticks and labels are omitted for readability.