Check for updates

Guo Danlu (Orcid ID: 0000-0003-1083-1214)
Johnson Fiona (Orcid ID: 0000-0001-5708-1807)
Marshall Lucy, Amanda (Orcid ID: 0000-0003-0450-4292)

Assessing the potential robustness of conceptual rainfall-runoff models under a changing climate

Danlu Guo[1,2], Fiona Johnson[1], and Lucy Marshall[1]

[1] Water Research Centre, School of Civil and Environmental Engineering, University of New South Wales, Sydney, NSW 2052, Australia.

[2] Department of Infrastructure Engineering, The University of Melbourne, Parkville, VIC 3010, Australia.

Corresponding author: Danlu Guo (danlu.guo@unimelb.edu.au)

Key Points:

1. We studied potential robustness of conceptual rainfall-runoff models, as variability in predictions from different calibration periods.
2. Use of stochastic weather generator allows us to assess potential robustness for climate conditions that extrapolates beyond observations.
3. The approach is transferable across case studies to inform model selection and calibration strategies for climate change assessments.

Author Manuscript

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1029/2018WR022636

This article is protected by copyright. All rights reserved.

Author Manuscript

Abstract

Conceptual rainfall runoff (CRR) models are commonly used to assess the potential impact of climate change on water resources systems. However, they are often characterized by poorer performance when used to simulate a different climate compared to that of the calibration period. This is generally referred to as low model robustness and these issues have been thoroughly explored using historical data. However, the implications of robustness are unknown for a changing climate where models may have to operate under conditions that lie beyond existing observations. This study extends these ideas to evaluate the 'potential robustness' of different CRR models in the context of a changing climate. To achieve this aim, we combine a generalized split-sample test framework with a stochastic weather generator. This allows us to assess the variabilities in runoff predictions obtained from using different calibration periods within each CRR model. We tested the potential robustness on three catchments with contrasting hydro-climatic conditions. We observed a consistent higher potential robustness in all models under drier conditions at all catchments. The three catchments illustrate contrasting patterns in the relative potential robustness of the three CRR models, which are related to both the structures of the CRR models and the unique catchment characteristics, highlighting the need of case-specific assessment. This study illustrates a transferable empirical testing strategy to understanding variabilities in CRR model predictions. This approach can improve our knowledge of model behavior, and thus informs the suitability of alternative models to simulate catchments hydrology under a changing climate.

Plain language summary

Conceptual rainfall runoff (CRR) models are commonly used to understand how climate change may affect river flows, floods and droughts. These models need to be calibrated to historical data. However, they tend to not perform well when used to simulate wetter or drier conditions. This property is referred to as 'low model robustness'. This study extends this idea to focus on the uncertainty in model predictions under a changing climate, which is defined as 'potential robustness'. By generating synthetic rainfall data, we can represent a range of possible future changes in rainfall, including scenarios that are very different from the historical climate. We tested the potential robustness of three CRR models on three river catchments. There are contrasting patterns in the relative potential robustness among models, which are related to both the structure of the CRR models and the unique characteristics of individual catchments. This means that potential robustness cannot be estimated in advance of a modelling study and a result catchment-specific testing is essential. This study also illustrates a transferable approach to perform such testing. The potential robustness approach can improve our knowledge of model behaviors, and help to choose suitable CRR models to assess climate change impacts.

This article is protected by copyright. All rights reserved.

Author Manuscript

1 Introduction

Climate change is expected to have significant impacts on water resources availability (CSIRO and Bureau of Meteorology, 2015; Hauser et al., 2009; IPCC, 2014; Turral et al., 2011). These impacts are assessed by downscaling projections of large-scale changes in atmospheric variables from general circulation models. Conceptual rainfall runoff (CRR) models are then used to translate these projected changes in atmospheric variables to local runoff (e.g. see Akhtar et al., 2008; Chiew et al., 2009). Projected changes in runoff can inform expected changes in different runoff regimes including high and low flows (Prudhomme et al., 2003; Wilby & Harris, 2006), catchment yield (Haque et al., 2015), water quality (Crossman et al., 2013; Wilby et al., 2006), water supply security (Christensen et al., 2004; Paton et al., 2013, 2014) and flood risk (Kay & Jones, 2012). During this modelling process, a 'cascade of uncertainty' is created through the choices of climate change scenarios, climate models, downscaling strategies and CRR models (Bastola et al., 2011b; Clark et al., 2016; Kay & Davies, 2008; Kay et al., 2009)

The uncertainties in CRR models arise from data errors, model structure and non-uniqueness in model parameters (Broderick et al., 2016). In addition, a more complex type of CRR model uncertainty is involved in climate impact assessments, for which CRR models generally lack robustness across different climate conditions. A lack of robustness is defined as the drop in model performance when a simulation period exhibits different climate conditions to those of the calibration period (Bastola et al., 2011a; Broderick et al., 2016; Coron et al., 2014; Coron et al., 2012; Merz et al., 2011; Vaze et al., 2010; Zheng et al., 2018), and can thus be combined effects of uncertainties in model input, structure and parameters. Two findings are common to robustness studies despite the variability in locations and models used. Firstly, robustness can be used to identify climate conditions that have greater impacts on model performance. For example, model performance is mainly related to changes in rainfall from the calibration period to the validation period (Bastola et al., 2011a; Coron et al., 2012). Greater reductions in model performance are observed when simulations conditions are drier than the calibration period (Vaze et al., 2010). Secondly, investigating robustness can recommend better ways to use CRR models so that model performance is less affected by robustness issues. For two Irish catchments, model performance was less affected when calibration period is sufficiently long to capture both wet and dry conditions (i.e. containing both a dry and a wet decades), and the difference in rainfall between calibration and simulation periods is within 10% (Bastola et al., 2011a). Slightly larger ranges of changes (-15% to 20%) in rainfall difference led to acceptable transferability of models for 61 Australian catchments (Vaze et al., 2010).

The implications of CRR robustness under a changing climate remains unknown, since potential changes in climate may be beyond observed variability (Johnson & Sharma, 2011; Rajah et al., 2014; Wasko & Sharma, 2015; Westra et al., 2013; Westra et al., 2014; Zheng et al., 2015). This gap motivates us to assess the 'potential robustness' for CRR models. To predict for a plausible future climate condition with a specific CRR model, we define the potential robustness by the variability of predictions obtained from different versions of that CRR model

This article is protected by copyright. All rights reserved.

Author Manuscript

calibrated to different historical periods. This differs from previous studies where robustness has been assessed by the variability in historical model performance. In contrast, using our definition, high potential robustness describes a model that has low variability in runoff predictions for each of the future climate scenarios considered. Assessing potential robustness allows us to compare the ranges of prediction variations across multiple CRR models, from which we can then relate any inter-model differences to the differences in model structures. Quantifying this variability in future CRR model predictions is critical to understanding the uncertainties associated with alternative CRR models and various calibration options for runoff predictions. Ultimately it may identify the best CRR models and calibration strategies for climate change impact assessments.

We develop the concept of 'potential robustness' by addressing the following three research questions:

1. How variable are historical runoff simulations and performances obtained from different calibration periods?
2. How variable are runoff predictions obtained from different calibration periods, when used to model future climate change?
3. How is the variability in runoff predictions related to way physical processes are represented in each CRR model?

We test the potential robustness of three conceptual rainfall-runoff models for three catchments in Australia with very different hydroclimatic conditions. The CRR models have contrasting structures for their soil moisture accounting (SMA) routines, which enables us to identify relationships between the potential robustness and the structure of CRR models. To assess the potential robustness, the CRR models require inputs that represent a range of possible future changes in rainfall. We therefore generate multiple sets of synthetic time-series of rainfall data from historical observations with a stochastic weather generator, following a recently developed inverse approach for climate impact assessments (Guo et al., 2018). The differences in potential robustness across three CRR models are investigated by considering the behavior of the simulated effective rainfall and store levels.

The paper is organized as follows. Section 2 introduces the three CRR models and the case study catchments. Section 3 details the analytical approach, including the calibration of each of the three CRR models at each catchment to multiple historical periods (section 3.2), and the simulations and analyses we used to assess the historical and potential robustness (section 3.3). The robustness results are presented in section 4, followed by discussions in section 5 on the specific implications of the results on runoff prediction under a changing climate. The study is then summarized and concluded in section 6.

This article is protected by copyright. All rights reserved.

Author Manuscript

2 Models and Case Studies

2.1 Models

To assess the potential robustness of runoff predictions, we focused on three CRR models with different methods for runoff production. All models are lumped conceptual models that simulate runoff at a daily time step. Their SMA routines consist of contrasting structures, which affect the conversion of rainfall to effective rainfall (i.e. for runoff production), as summarized in Figure 1. These structural differences lead to different runoff responses to rainfall and thus potentially contrasting runoff predictions under changes in climate conditions. All models were implemented with R package *hydromad* (http://hydromad.catchment.org/) (Andrews & Guillaume, 2013; Andrews et al., 2011).

Figure 1 around here

The first model, GR4J (Perrin et al., 2003) has a single production store. The model requires two input variables, PET and rainfall (P). Interception is treated as a store with zero capacity, so that each day can be either 'wet' (P > PET), which produces a net rainfall P_n = P – PET, or 'dry' (P < PET) with no net rainfall. Runoff is predominately generated by net rainfall on wet days. A small amount of store percolation also contributes to runoff which maintains non-zero runoff for dry days. For any wet day, the model estimates the effective rainfall U (i.e. the rainfall used for runoff production) as the sum of percolation from the store (Perc), and a portion of P_n (kP_n) which depends on the store level relative to the store capacity, S/S_{max}, as:

$$\mathrm{U_{GR4J}} = \mathrm{Perc} + kP_n$$

$$kP_n = P_n - \frac{S_{max}\left[1-\left(\frac{S}{S_{max}}\right)^2\right]\times tanh\left(\frac{P_n}{S_{max}}\right)}{1+\frac{S}{S_{max}}\times tanh\left(\frac{P_n}{S_{max}}\right)} \quad \textbf{(1)}$$

As such, the only free parameter to be calibrated in the model is S_{max}, which represents the capacity of production store. The model controls the rate of replenishing and extraction of water into/from the store according to the actual store levels, to ensure that store level is always between 0 and S_{max}.

In contrast to GR4J, the production store in the second model, AWBM (Boughton, 2004) is represented by three individual stores with increasing capacities (S_{1_max}, S_{2_max} and S_{3_max}) and actual store levels within each store (S_1, S_2 and S_3). Each store occupies a fixed fraction of the total catchment area, denoted as A_1, A_2 and A_3 (where $A_1 + A_2 + A_3 = 1$), respectively. The original version of AWBM requires calibration of all the three store capacities and two of the store fraction areas i.e. a total of five free parameters (Boughton, 1993; Boughton & Carroll, 1993). A self-calibrated version of AWBM was later developed (AWBM2002, see Boughton, 2004),

This article is protected by copyright. All rights reserved.

which uses constant fraction areas of $A_2 = 0.433$, $A_3 = 0.433$ and thus $A_1 = 1 - A_2 - A_3 = 0.134$. Besides, a single area-averaged store capacity, S_{max}, is used to define the three store capacities, as:

$$S_{1\_max} = 0.01 \times S_{max}/A_1$$
$$S_{2\_max} = 0.33 \times S_{max}/A_2$$
$$S_{3\_max} = 0.66 \times S_{max}/A_3 \quad \textbf{(2)}$$

On each day P is added to each of the three stores, and the area-weighted sum of any excess from any store ΔP_i for i = 1,2 and 3 becomes effective rainfall U. Therefore, effective rainfall is calculated based on a step function of rainfall excess, which depends on the status of the three individual stores:

$$\mathrm{U_{AWBM=}}\begin{cases} 0, \text{when } \dfrac{S}{S_{max}} < 0.01 \\ A_1 \times \Delta P_1, \text{when } 0.01 \le \dfrac{S}{S_{max}} < 0.33 \\ (A_1 + A_2) \times \Delta P_2, \text{when } 0.33 \le \dfrac{S}{S_{max}} < 0.66 \\ \Delta P_3, \text{when } \dfrac{S}{S_{max}} \ge 0.66 \end{cases}$$

$$\text{where } \Delta P_i = S_i + \mathrm{P} - S_{i\_max}, S = \sum S_i \times A_i \text{ for } i = 1,2,3 \quad \textbf{(3)}$$

In the third model, IHACRES_CMD (Croke & Jakeman, 2004), the level of the store is represented by the catchment moisture deficit (CMD), which is the difference between the actual and the saturation levels of the store. Therefore, CMD has a lower bound of 0 which represents saturation. For each day, all rainfall P directly fills the store without explicit considering interception. The only free parameter in the model, d, represents the threshold of CMD below which runoff production starts. When CMD is below d, the proportion of rainfall P that becomes effective rainfall U is a function of the instantaneous CMD, with a simplest linear form of:

$$\frac{d\mathrm{U}}{d\mathrm{P}} = 1 - \min(1, \frac{\mathrm{CMD}}{d}) \quad \textbf{(4)}$$

The U produced at each time step can then be estimated by integrating equation 4, which becomes equation 5 once estimated evapotranspiration from the store is incorporated (Croke & Jakeman, 2004):

$$\mathrm{U_{CMD}} = \begin{cases} \mathrm{P} - \mathrm{CMD} \times [1 - \exp\left(-\frac{\mathrm{P}}{d}\right)], \text{when } \mathrm{CMD} < d \\ \mathrm{P} - \mathrm{CMD} + d \exp\left(-\left[\frac{\mathrm{P} - (\mathrm{CMD} - d)}{d}\right]\right), \text{when } d \le \mathrm{CMD} < d + \mathrm{P} \\ 0, \text{when } \mathrm{CMD} \ge d + \mathrm{P} \end{cases} \quad \textbf{(5)}$$

Author Manuscript

This article is protected by copyright. All rights reserved.

Author Manuscript

We use the three-parameter GR4J routing model (Perrin et al., 2003) for all three CRR models. As such, any differences between model simulations can be attributed specifically to the structural differences in SMA routines. The effective rainfall simulated from the SMA models is partitioned by two unit hydrographs UH1 and UH2, as: 1) 90% of the effective rainfall is routed by UH1, with the time base of x_4, which then feeds into a non-linear routing store of capacity x_3 and a groundwater exchange rate x_2; and 2) the remaining 10% of the effective rainfall is routed by another single unit hydrograph UH2 with the time base of $2x_4$.

2.2 Case Studies

We illustrate our approach on three Australian catchments, Scott Creek (South Australia), Black River (Tasmania) and Coen River (Queensland). Figure 2 shows the locations of the three case study catchments, which are within climatologically different regions, as defined in the Australian Köppen climate classifications of Stern et al. (2000). The CRR models (section 2.1) were calibrated using the catchment average rainfall, PET and runoff for a 45-year period (1970–2014). Daily catchment average rainfall and PET were extracted from the Australian Water Availability Project (AWAP) gridded dataset (Raupach et al., 2009, 2012), and daily runoff was obtained from the Australian Bureau of Meteorology. Table 1 summarizes key characteristics of the case study catchments.

Figure 2 around here

Table 1 around here

The potential robustness of the three CRR models was assessed, by simulating four key attributes that represent different runoff regimes, namely:

1. the daily average runoff for all days (*Qavg*);
2. the daily average base flow runoff (*Qbase*) estimated using the Lyne-Hollick filter in R package *hydrostats*, with a default alpha value of 0.925 (Bond, 2016);
3. a measure of peak flow using the 95^{th} percentile of non-zero daily runoff (*Q5*); and
4. the average daily runoff in the wettest season, which is Australian winter for Scott Creek and Black River (*QJJA*), and Australian summer for Coen River (*QDJF*).

Table 2 shows the corresponding baseline values of the four runoff attributes for each catchment over the study period (1970-2014).

Table 2 around here

This article is protected by copyright. All rights reserved.

Author Manuscript

3 Methodology

3.1 Overview

We first assess the model robustness based on the historical performance of each CRR model across multiple calibration periods for each catchment. We then compare the potential robustness using runoff predictions obtained from the three models. Building on these, we further investigate how potential robustness relates to model structure. These are achieved via two main steps:

1. Calibration of each CRR model with multiple historical periods (section 3.2), yielding:
   a) one 'full-period (FP)' model, calibrated to the entire period of historical data;
   b) multiple 'partial-period (PP)' models calibrated to different subsets of the historical data, following the generalized split-sample test approach in Coron et al. (2012).
2. Simulation with the calibrated FP and PP versions of each CRR model from Step 1, to evaluate their historical and potential robustness (section 3.3), by using:
   a) the entire period of observed rainfall data to assess the robustness of the model performance for historical climate (section 3.3.1);
   b) possible future rainfall scenarios to assess the potential robustness of runoff prediction under climate change, as detailed in section 3.3.2. We then further look at the simulated changes of effective rainfall and store levels to investigate the relationship between potential robustness and model structure.

A schematic of the study approach and how it links to the results at each catchment is shown in Figure 3, with each individual step detailed in the subsequent sections.

Figure 3 around here

3.2 Calibration CRR models

At each catchment, a FP model was first established for each CRR model with the full 45 years of data. The full period was then divided by a 10-year moving window to construct 35 PP versions of each CRR model. The choice of 10 years as calibration period follows the recommendation for generalized split-sample test in Coron et al. (2012), which considers a trade-off between: 1) maximizing the number of subsets which can represent contrasting climate conditions; and 2) maximizing the length of each calibration period to ensure the estimated parameters converge to an acceptable level. Figure 4 shows the variability in rainfall and runoff at both an annual and a 10-year time scales. When comparing across catchments, Black River has much less inter-annual fluctuation in both rainfall and runoff, which thus identifies more

This article is protected by copyright. All rights reserved.

'average years' (Figure 4(b)). This also leads to more stable rainfall and runoff conditions at the 10-year time scale at Black River, despite a slight decreasing trend in runoff.

Figure 4 around here

All the FP and PP models were calibrated with the Shuffled Complex Evolution (SCE) algorithm (Duan et al., 1993), with the objective function of Nash-Sutcliff Efficiency (NSE), which has been widely used in rainfall-runoff modelling (Gupta et al., 2009). Due to parameter equifinality the best-fit parameters may vary when the calibration is repeated with different initial parameter values (Shin et al., 2015). Therefore, for each PP model we repeated the SCE calibration five times with different random starting seed, which allows us to take into account parameter variability in assessing both historical and potential robustness.

3.3 CRR Model Simulations for Robustness Assessments

3.3.1 Robustness in historical performance

Our assessment of CRR model robustness starts with the historical performance of each CRR, following the approach in existing studies (Bastola et al., 2011a; Coron et al., 2014; Coron et al., 2012; Merz et al., 2011; Vaze et al., 2010). For each case study, we simulate historical runoff from the FP model and the 35 PP versions for each CRR model (section 3.2), using the historical rainfall and PET data. The variations of PP model performance within each of the three CRR models are then compared to establish the relative historical robustness of these models. The effect of parameter variability on this robustness is also assessed by simulation using the alternative parameter sets identified from the repeated calibration. Note that one difference this has with the previous approach is that, instead of the model performance on different partial periods, we focus on the model performance on the entire period of historical observations. This ensures that the durations of data used to assess the historical and potential robustness (as detailed in section 3.3.2) are consistent.

3.3.2 Potential robustness in runoff prediction

To represent possible changes in rainfall under a changing climate at each case study, we considered possible changes in a) average rainfall between -30% and +10%, with steps of 10%; and b) extreme rainfall (99^{th} percentile) between +10% and +30%, with steps of 5%. These possible changes are based on a common range of rainfall projections for the three case study locations by the year 2090 (CSIRO and Bureau of Meteorology, 2015; IPCC, 2014). For each case study, the future rainfall time series were generated using the inverse approach in Guo et al. (2018). This approach uses a stochastic weather generator combined with optimization to generate climate time-series from historical data, and thus can be a useful tool to represent specific plausible changes for assessing potential impacts of climate change (Culley et al., 2016; Guo et al., 2017b).

Author Manuscript

This article is protected by copyright. All rights reserved.

Author Manuscript

We applied a widely used stochastic weather generator WGEN (Richardson, 1981), which models the daily rainfall for each month with a first order Markov chain for occurrence, and a gamma distribution for intensity on wet days (as detailed in Guo et al., 2018). For each of the rainfall scenarios (i.e. -30%, -20%, -10%, 0% and +10% in average rainfall, and +10%, +15%, +20%, +25% and +30% in 99th percentile rainfall), we determined the best-fit WGEN parameter set that would produce rainfall time-series to represent such change in average rainfall. This is achieved with a genetic algorithm (GA) due to its proven efficiency particularly for solving high-dimensional optimization problems in hydrological studies (Cheng et al., 2002; Gibbs et al., 2012; Shafii & De Smedt, 2009)

We used a one-at-a-time perturbation for the average and extreme rainfall, respectively. For possible changes in average rainfall, the objective function to be minimized was:

$$F_{obj} = \sum \sqrt{\left[\frac{(PD_{targ} - PD_{his})}{PD_{his}} - \frac{(PD_{sim} - PD_{his})}{PD_{his}}\right] * 100 + \left[1 - \frac{Pex99_{sim}}{Pex99_{his}}\right] * 100} \quad \textbf{(6)}$$

Equation 6 measures how well a synthetic rainfall time-series from WGEN represents the target levels of average change in daily rainfall, *PD*, as well as maintaining the historical levels of the 99th percentile of daily rainfall, *Pex99*. This is evaluated based on the Euclidean distance between the target levels (subscript *targ*) and simulated levels (subscript *sim*) for *PD* and *Pex99*. As mentioned previously we considered five possible target levels of *PD* (-30%, -20%, -10%, 0 and +10%) to be simulated with WGEN, which therefore defined the five levels of PDtarg. We focus changes in the average rainfall first, so that the target level of the 99th percentile of rainfall is fixed at the historical value. To ensure that the two attributes have equal influence on the objective function, distances for *PD* and *Pex99* are represented as percentage changes relative to their historical values (subscript *his*) (see Table 1).

In a similar way, to determine the WGEN parameters to represent possible changes in 99th percentile rainfall, the objective function to be minimized was:

$$F_{obj} = \sum \sqrt{\left[\frac{(Pex99_{targ} - Pex99_{his})}{Pex99_{his}} - \frac{(Pex99_{sim} - Pex99_{his})}{Pex99_{his}}\right] * 100 + \left[1 - \frac{PD_{sim}}{PD_{his}}\right] * 100} \quad \textbf{(7)}$$

The Markov transition probabilities for rainfall occurrence were assumed to be the same as observed in historical data. To simulate the rainfall intensity for each month, two Gamma parameters (α and β) are required, leading to a total of 24 parameters to be determined. The search range for the Gamma parameters were within ±20% of their values from the historical rainfall. This limited search range ensures that: 1) the seasonality of rainfall intensity in the synthetic time-series are generally consistent with the historical condition (i.e. avoid generation of 'unrealistic' rainfall seasonality); and 2) the search space for the Gamma parameters is constrained so that the optimization algorithm can converge in an efficient manner.

The convergence criterion was set to a value of 0.1 for both objective functions (equations 6 and 7). In addition, as suggested in Guo et al. (2018), during these optimization

This article is protected by copyright. All rights reserved.

processes, the random seed of the WGEN was held constant. This is because the stochastic generator can introduce random behavior to the generated hydro-meteorological time series, which can mislead the optimization algorithm and thereby slow down the optimization process. To represent natural variability at each case study, for each rainfall change condition we ran the optimization five times which each producing independent WGEN parameter sets, and thus resulting in five different synthetic time-series of rainfall.

We have checked the plausibility of the synthetic rainfall time-series corresponding to the modelled rainfall change at each case study (see Figure S1 in the Supporting Information). As an expected result of constraining the monthly WGEN parameters (as discussed in the parameter optimization process above), the synthetic rainfall time-series generally preserve the monthly patterns within the historical data. In addition, the variability that is represented by WGEN within each scenario is smaller than the variability across the five different climate change scenarios.

Synthetic time-series for PET were generated together with each synthetic rainfall time-series, with the assumption that the observed correlations between rainfall and PET at the case study are maintained under a changing climate. To achieve this, the historical rainfall and PET data was first used to construct a daily regression model for PET. The model relates PET on dry days with PET on the previous day, while for wet days both rainfall on the current day and PET on the previous day are used as independent variables (see details in Srikanthan & Zhou, 2003). The fitted regression was then used to generate a PET time-series corresponding to each synthetic rainfall time-series.

Having generated the synthetic rainfall and PET time-series corresponding to a range of possible changes in rainfall at each case study, we used these as input data to each of GR4J, AWBM and CMD to obtain runoff predictions. Considering the different calibration strategies used in section 3.2, for each set of rainfall and PET time-series, runoff predictions were obtained from both the FP model and the 20 PP models for each CRR model. The variations of predictions across the PP versions within each of the three CRR models are then compared to establish their relative potential robustness. In addition, the effect of parameter variability on this potential robustness is also assessed by simulation using the alternative parameter sets identified from the repeated calibration (section 3.2).

Following the comparison of potential robustness, we also aim to relate differences in potential robustness to the structure of individual CRR models. To achieve this, we first looked at the variation in the calibrated parameters from the 35 PP models, for all three models and at all case studies. We then focused on the simulated levels of (a) production store; (b) effective rainfall and (c) routing store, obtained from the 35 PP versions of each CRR model, in response to possible changes in rainfall.

Author Manuscript

This article is protected by copyright. All rights reserved.

Author Manuscript

4 Results

4.1 Historical performance and robustness of runoff simulation for three CRR models

Figure 5 shows the calibration performance of GR4J, AWBM and CMD at the three catchments. The NSE values are generally satisfactory (greater than 0.65) for all catchments and FP and PP models, with the effects of parameter uncertainty shown in smaller dots. Parameter uncertainty has little impact on the performances of the PP models. The three FP models have comparable performance at both Scott Creek (Figure 5(a)) and Coen River (Figure 5(c)). However, for Black River (Figure 5(b)), the FP AWBM model performs slightly worse than GR4J and CMD. The relative performance of the PP versions of the three models is consistent with that of the three FP models.

Figure 5 around here

Figure 5 also illustrates the variability of PP model performance across calibration periods for all three catchments. Black River has the most stable PP model performances across different calibration periods, with NSEs only vary by approximately 0.1 for each of the three CRR models. For Scott Creek, the NSE differences are between 0.13 and 0.16, whilst Coen River has the largest variability of PP model performances with NSE differences around 0.20. The variability in PP model performances is related to the variability of rainfall and runoff conditions across the calibration periods at the three catchments (Figure 4) as Black River has the most stationary rainfall and runoff conditions over time. For Scott Creek, lower performances occur in low-flow periods, such as for calibration periods starting from 1970 to 1976. In contrast, models perform better in Coen River in drier periods (e.g. calibration periods starting between mid 1970s and mid 1990s).

Figure 6 illustrates the historical robustness of each CRR in simulating the full study period for each catchment (1970-2014). This is summarized by five different performance metrics, namely NSE, as well as percentage biases in the four runoff attributes (Table 1). The performance variation across the PP models is shown with boxes and whiskers. Red and orange represent simulations with and without the effect of parameter variability, respectively (i.e. repeated calibration and single calibration as detailed in section 3.2). As a reference, we show the performance of the FP models with large black dots.

Figure 6 around here

The FP versions of all three models (black dots) have comparably high performance illustrated by the NSEs at Scott Creek and Coen River (Figures 6(a) and 6(k)), with AWBM being slightly weaker at Black River (Figure 6(f)). AWBM consistently underestimates the mean flows for all catchments (Figures 6(b), (g) and (l)).

We first assess historical robustness using the single calibration results (orange boxes) first. Unsurprisingly the PP models of all the CRR models have lower overall performance than their corresponding FP models (lower NSE and higher bias). This suggests that lower

This article is protected by copyright. All rights reserved.

Author Manuscript

performance is to be expected when using shorter calibration periods, consistent with the findings of Bastola et al. (2011a). GR4J, AWBM and CMD have comparable robustness when evaluated with NSE (Figures 6(a), (f) and (k)), with similar ranges of degradation of NSE of the PP models when compared to their corresponding FP versions at each catchment.

Differences in model robustness are more evident when looking at different flow regimes although the results are catchment dependent. For Scott Creek, the three models have similar historical robustness in representing different flow characteristics. The robustness of AWBM in modelling Black River is lower than the other two models when evaluated on the mean flow and base flow (Figures 6(g) and (h)). Coen River has consistently higher variability (i.e. lower robustness) across all runoff attributes than other two catchments. The performance of GR4J for Coen River is particularly interesting as it has clearly lower robustness in simulating the average flow, base flow and summer flow (Figures 6(l), (m) and (o)).

In addition, the effects of parameter variability obtained from the repeated calibration (red boxes) on model robustness are generally smaller compared to those from different PP models. Therefore, in the subsequent results sections (4.2 and 4.3) we focus on the results from the single calibration only.

4.2 Potential robustness of runoff prediction for three CRR models

The future potential robustness of the models has been assessed by considering the variability in the simulations due to changing future rainfall. Figure 7 shows the potential robustness in predicting the four runoff attributes when future average rainfall is varied between -30% to +10%. For each catchment, predictions were made with the FP version of each of three models (larger dots), as well as their PP versions obtained from single calibration (boxes and whiskers). To provide consistency across the three CRR models, all runoff predictions are presented as percentage changes relative to the historical simulation from the FP version of each model. Results simulated with the other four replicates for each possible rainfall condition are shown in Figures S2 to S6 in Supplementary Information, which are generally consistent with those presented in Figure 7.

Figure 7 around here

Before assessing the potential robustness of the three models, we first compare the runoff predictions from the FP versions of these models (large dots). For Black River, predictions from all three FP models are consistent with differences less than 7%. Scott Creek and Coen River shows more distinct inter-model differences, with similar estimates from GR4J and CMD, while those from AWBM are consistently lower (i.e. leading to greater decreases and smaller increases in predicted runoff, compared with GR4J and CMD). Nevertheless, it is worth noting that inter-model consistency (e.g. predictions from FP GR4J and CMD) does not necessarily suggest high accuracy, as models that derive similar simulations may be biased in the same way. A good example of this is that GR4J and CMD have similar degree of overestimation for both the high

This article is protected by copyright. All rights reserved.

Author Manuscript

flow and winter flow at Scott Creek (Figures 6(d) and (e)). This is a common limitation of robustness assessment and will be further discussed in section 5.

The predictions from the PP versions of the three models (boxes and whiskers) scatter around the corresponding FP models, which have consistent patterns with the historical results. To better summarize the potential robustness of the three CRR models, we present the absolute range of variation in the PP predictions from each model in Figure 8.

Figure 8 around here

At each catchment, the variations across the PP predictions are always smaller under drier climate conditions compared to wetter ones for all three models. For example, in Scott Creek, average runoff varies across the PP predictions by 12%, 11% and 20% for GR4J, AWBM and CMD for the driest future simulations (-30% in average rainfall). For the wettest condition (+10% in average rainfall), these ranges increase to 32%, 31% and 40%, respectively. The possible causes for this are discussed in more detail in section 4.3.

The three catchments have different relative rankings of potential robustness across the three models, particularly for the mean flow and base flow predictions which has greater inter-model differences in potential robustness. For Scott Creek, CMD has the lowest potential robustness in predicting mean flow. CMD and AWBM both have low potential robustness in predicting base flow. Contrasting results are obtained for Black River and Coen River, where AWBM and GR4J have the lowest potential robustness, respectively. These relative ranking of models based on their potential robustness is consistent with those found for historical robustness in Figure 6. Historical robustness thus extrapolates well to potential robustness for any particular catchment. However, compared with historical robustness, potential robustness results can offer additional information, which illustrates the values of undertaking these additional simulations. For example, for all catchments and across all runoff attributes, we observed that predictions for dryer conditions have smaller uncertainties than those within simulations for current conditions. The converse is also true with wetter conditions leading to larger uncertainties in runoff predictions. This information is useful to understand how a particular model may perform in the future for the expected changes in precipitation, for any particular catchment.

The range of the variabilities is also dependent on the runoff attributes to predict and the catchment of interest. These suggest that potential robustness of CRR models are specific to case studies. Therefore, in the following section we investigate the interactions between case-specific robustness patterns and CRR model structures for individual catchments.

4.3 Relationship between potential robustness and CRR model structures

The previous section demonstrated that the relative potential robustness of each model varies for the different catchments. To investigate the possible drivers, in this section we first look at the ranges of the parameters of the 35 PP versions of the three CRR models in each catchment, and how these propagate to impact runoff predictions. Figure 9(a) shows the

This article is protected by copyright. All rights reserved.

calibrated parameter values for the store capacities in the individual SMA models (S_{max} for GR4J and AWBM, and d for CMD, see section 2.1), and Figures 9(b)-(d) show the calibrated routing parameters (x_2, x_3 and x_4) for the common GR4J-routing model.

Figure 9 around here

For all three CRR models, the parameter values are more sensitive to the calibration periods for Coen River. The large variability in the PP parameter values suggest a common difficulty for all the three CRR models to identify a unique parameter set across all calibration periods at Coen River. These results can be linked back with the highly variable PP model performances at Coen river shown in Figure 5, which together highlight a common structural limitation of all the three CRR models in simulating a catchment that has highly variable rainfall and runoff conditions. Furthermore, at Coen River, GR4J generally has the highest variability in the calibrated parameter values of the store capacity Figure 9(a), as well as the routing parameters of x_2 and x_3 in Figures 9(b) and (c).

How does the variation in calibrated parameters propagate through to the runoff predictions from different PP models? To answer this, we compare the simulated effective rainfall and store levels from the FP and PP versions of GR4J, AWBM and CMD. Three examples were found to be useful to explain the potential robustness results.

We first explain the low potential robustness of AWBM for Black River, focusing on the -30% change in average rainfall. We illustrate this simulations from the year 1983 (Figure 10), which represents an average year for Black River (as in Figure 4). The simulated production store levels from both GR4J and CMD are relatively consistent across different PP models (i.e. calibration periods). However, simulations from AWBM are highly variable, which can be a result of the relatively large variability of the calibrated store capacity in AWBM (Figure 9). Consequently, simulations of effective rainfall from different AWBM PP models are also highly variable. Variability is particularly evident in the middle of the year when the simulated store levels rise above and then drop below the calibrated store threshold levels. This is related to the structure of AWBM, which uses calibrated store threshold levels to define rainfall-runoff relationship as step functions (as detailed in section 2.1). For PP models that are calibrated to different periods, the calibrated store thresholds change to reflect changes in the climate conditions. Consequently, the rainfall-runoff relationships simulated by AWBM vary across different PP models, especially at times when the simulated store levels from some PP models exceed specific threshold, while some other PP models suggest store levels below that threshold. Apart from the contrasting ranges of simulated production store levels, the ranges of variation in the simulated routing stores are similar across three CRR models. Therefore, different model potential robustness at Black River are mainly related to the uncertainties in the calibrated production store capacity, which can cause high variability in the simulated production store levels especially in AWBM.

Figure 10 around here

Author Manuscript

This article is protected by copyright. All rights reserved.

Author Manuscript

Figure 11 around here

Figure 12 around here

The second illustrative example investigates causes for the low potential robustness of GR4J for Coen River, with two extreme future rainfall conditions considered in this study, namely, -30% and +10% changes in average rainfall (Figure 11 and 12, respectively). Again, we focus on simulations for an average year for Coen River (1977, as in Figure 4). Like the simulation for Black River, the store levels from AWBM are still highly variables across calibration periods. However, in contrast to Black River, GR4J and CMD also both have large variability in simulated production stores. These variabilities are consistent with the wide range of calibrated values for the production store capacities in Figure 9. Despite the commonly large variability in the simulated production store, the routing store obtained from GR4J have clearly higher variability, which is a result of the high variability in the calibrated values of the routing parameters x_2 and x_3 (Figure 9).

The results of these two simulations for Coen River (Figures 11 and 12) explain why potential robustness is generally lower under wetter conditions (as observed in Figure 8). For any wet day, effective rainfall simulated from different PP simulations are highly variable. However, for a dry day, the effective rainfall from all PP simulations are consistently equal to 0. Therefore, under wetter conditions, more wet days will lead to higher variability in the simulated effective rainfall, and thus causing higher variability in the predicted runoff i.e. lower potential robustness.

The above results illustrate how uncertainties in parameters obtained from different calibration periods can propagate via simulations of the production store and routing store, which both lead to uncertainties on simulated runoff and thus affecting potential robustness. The variation in parameters across different calibration periods is closely related to the variability of rainfall runoff conditions over time at each catchment. This explains why individual case studies have different patterns of potential robustness for different CRR models, as observed in Figure 8.

5 Discussion

In this study, we estimated the potential robustness of runoff prediction for three CRR models, considering possible future changes in rainfall. This study confirms previous findings that longer calibration periods lead to more consistent runoff simulations, because they are more likely to sample a wider range of dry and wet conditions (Bastola et al., 2011a; Vaze et al., 2010). In addition, for each catchment, the relative uncertainty ranges in predictions amongst the three CRR models are consistent with the historical simulations. Therefore, CRR models that have high robustness in historical performance at a catchment are also likely to produce more consistent runoff predictions for the future. However, the novelty of this study is to assess robustness with climate data that represent future climate conditions, which lie outside the range of observed variability. The use of a stochastic weather generator allows us to stress testing CRR models beyond conditions within existing records, and thus explore a larger continuum of model

This article is protected by copyright. All rights reserved.

behavior. By observing the behavior GR4J, AWBM and CMD under possible climate change conditions, one of the key findings of this study is that the relative uncertainty and sensitivity amongst models extrapolate reasonably well from the historical robustness.

Previous studies have provided theoretical understanding of model responses as a function of model structure. For example, Kavetski and Kuczera (2007) concluded that models with strong threshold can introduce excessive sensitivity to small changes in parameters. The impacts of climate variability on hydrologic simulations have also been studied, which can greatly affect the predicted runoff under a changing climate (e.g. Bastola et al., 2011a; Coron et al., 2012). Relating to these previous findings, the potential robustness approach presented in this study integrates the effects of model variability (e.g. structural differences in models) and climatic variability (e.g. variability in the rainfall and runoff observations that the models are calibrated to). Due to this complexity of effects we consider, we acknowledge that there is currently no clear theoretical interpretation of the patterns observed. This is well illustrated with the highly variable potential robustness results from this study (section 4.2) across different catchments, especially when comparing the relative potential robustness between GR4J and AWBM for Black River (in which AWBM has the lowest potential robustness) and Coen River (in which AWBM has higher potential robustness than GR4J). Contrasting potential robustness results between the two catchments may suggest that a threshold-type model with a larger number of parameters in defining the production store parameters (e.g. AWBM) can be more flexible in adapting to catchments with high inter-annual variabilities in hydrological conditions (e.g. Coen River), and thus can lead to more robust runoff predictions at these catchments.

These results suggest that it is difficult to use our current understanding of CRR models to predict the uncertainty and sensitivity of these models a-priori under a changing climate across catchments. Therefore, we recommend individual assessments that are specific to catchments and runoff attributes of interest, to improve the understanding of how uncertainty and sensitivity will propagate in future unknown climate conditions. In addition to these recommendations, we have also illustrated an empirical testing strategy for such case-specific assessment, which is transferable to any other catchments to quantify the robustness of runoff predictions from CRR models, and can thus provide recommendations on model selection and calibration strategies for predicting runoff under a changing climate.

It is important to note that the potential robustness we presented in this study is based on the agreement of runoff predictions, which only indicates the relative consistency of CRR models under a changing climate. However, due to the lack of ‘true observations’, we cannot assess the modelling bias (i.e. accuracy) of future predictions. A model with high potential robustness (i.e. with high consistency in runoff predictions) could still be biased. Assessing potential robustness can provide additional insights on model behaviors besides their historical performance. This is illustrated with our finding that for all catchments, runoff attributes and CRR models, runoff predictions have lower uncertainty for drier future conditions. This offers a different perspective to previous studies which have suggested that CRR model performance is

Author Manuscript

This article is protected by copyright. All rights reserved.

poorer when simulating historical dry periods (Fowler et al., 2016; Vaze et al., 2010). Potential robustness is a useful framework to understand the full picture of model performance under a changing climate, which requires information on both model accuracy and reliability (Dessai et al., 2009).

As the first study to illustrate the assessment of potential robustness of CRR models, we focus on possible changes in average rainfall conditions. The analyses presented here were also repeated for changes in extreme rainfall (99^{th} percentile daily values) and the results were consistent with those presented here (see Figures S7 to S11 in Supporting Information). Further insights into CRR model robustness could be gained by exploring a wider range of possible hydroclimatic changes expected to affect streamflow under a changing climate. These include changes in other rainfall characteristics, such as increasing intensity of extreme sub-daily and short-duration rainfall (Wasko & Sharma, 2015; Westra et al., 2013; Westra et al., 2014; Zheng et al., 2015), temporal distributions (Rajah et al., 2014) and low-frequency variability (Johnson & Sharma, 2011; Kwon et al., 2009). Furthermore, potential future variations in temperature, solar radiation and thus evapotranspiration, may also have a substantial impact on catchment runoff (Chiew & McMahon, 1991; Guo et al., 2017a; Prudhomme & Williamson, 2013). The potential robustness framework presented here could assess the impact of any of these changes.

Natural climate variability is likely underrepresented in our approach, with five replicates of synthetic rainfall time-series to sample variability in each specific change of rainfall. It would be possible to increase the number of replicates to sample a wider range of variability at daily or seasonal scales. However, representing natural variability is likely a fundamental limitation when using stochastic weather generators to generate synthetic climate data, due to the lack of long-term memory of these weather generators (Chen et al., 2010; Wilks, 1999). The capacity of the potential robustness framework would be greatly enhanced with the development of modelling tools to better generate future hydro-climate sequences and changes (e.g. Bennett et al., 2018; Guo et al., 2016).

6 Summary and Conclusions

Recently, robustness has been used to assess conceptual rainfall-runoff model performance (e.g. Bastola et al., 2011; Coron et al., 2012; Vaze et al., 2010). Compared to traditional performance assessments based on a single period of historical observation, asessing robustness improves the understanding of whether a model structure is sufficient to simulate runoff under a range of different observed climate conditions. This information has been used to infer a model's capacity to simulate future climate change conditions (e.g. Coron et al., 2014; Fowler et al., 2016). However, observed climate can only represent existing natural variability, whereas future changes may be much larger. Therefore, this study extends the existing robustness method to extrapolate beyond observed climate conditions, and thereby enhance the relevance of robustness assessments for a climate change context.

Author Manuscript

This article is protected by copyright. All rights reserved.

We used a stochastic weather generator to represent a range of plausible rainfall conditions for three catchments with contrasting hydro-climate conditions. This allowed us to explore the model behavior for conditions different from historical observations for these catchments. We calibrated GR4J, AWBM and CMD to multiple historical periods at three catchments. using the generalized split-sample method. We then used these calibrated models to predict runoff for the plausible rainfall changes considered. Potential robustness was defined as the variability in runoff predictions within each CRR model from using different calibration periods.

A longer calibration period provides more consistent runoff predictions under a changing climate. The relative uncertainty and sensitivity amongst models extrapolate reasonably well from the historical robustness. However, the new insight obtained from this study is a consistently higher potential robustness in all models under drier conditions for all catchments, which implies lower uncertainties in predicting runoff in drier-than-current climates. The three catchments had different relative potential robustness for the three models. This is a result of both the model structures and the unique characteristics of each catchment, highlighting the need of case-specific assessment for potential robustness.

Potential robustness enables us to understand which CRR models are likely to produce more robust runoff predictions for changing climate conditions, with less dependency on the historical conditions used to calibrate the model. This understanding can inform modelling decisions for climate impact assessments for the specific catchment. This study also formulates a flexible approach which can be transferred to assess CRR model behaviors and sensitivities at other case studies, under a changing climate.

Acknowledgments, Samples, and Data

The historical streamflow data for the three catchments are available from the Australian Bureau of Meteorology Water Data Online, available from: http://www.bom.gov.au/waterdata/. The historical rainfall and PET data were from the Australian Water Availability Project (AWAP), at http://www.csiro.au/awap/. The authors would like to help Jie Jian for her help with identifying data sources. The authors also wish to thank the editor, the associate editor and the three anonymous reviewers for their thoughtful comments for the revision of the manuscript.

Author Manuscript

This article is protected by copyright. All rights reserved.

Author Manuscript

References

Akhtar, M., Ahmad, N., & Booij, M. J. (2008). The impact of climate change on the water resources of Hindukush–Karakorum–Himalaya region under different glacier coverage scenarios. *Journal of Hydrology, 355*(1–4), 148-163. doi:http://dx.doi.org/10.1016/j.jhydrol.2008.03.015

Andrews, F., & Guillaume, J. (2013). hydromad: Hydrological Model Assessment and Development. R package version 0.9-18. Retrieved 15/07/2014 http://hydromad.catchment.org/

Andrews, F. T., Croke, B. F. W., & Jakeman, A. J. (2011). An open software environment for hydrological model assessment and development. *Environmental Modelling & Software, 26*(10), 1171-1185.

Bastola, S., Murphy, C., & Sweeney, J. (2011a). Evaluation of the transferability of hydrological model parameters for simulations under changed climatic conditions. *Hydrol. Earth Syst. Sci. Discuss., 2011*, 5891-5915. doi:10.5194/hessd-8-5891-2011

Bastola, S., Murphy, C., & Sweeney, J. (2011b). The role of hydrological modelling uncertainties in climate change impact assessments of Irish river catchments. *Advances in Water Resources, 34*(5), 562-576. doi:http://dx.doi.org/10.1016/j.advwatres.2011.01.008

Bennett, B., Culley, S., & Westra, S. (2018). *foreSIGHT: Systems Insights from Generation of Hydroclimatic Timeseries. R package version 0.9.2.* . Retrieved from: https://CRAN.R-project.org/package=foreSIGHT

Bond, N. (2016). hydrostats: Hydrologic Indices for Daily Time Series Data. R package version 0.2.5. Retrieved from https://CRAN.R-project.org/package=hydrostats

Boughton, W. (2004). The Australian water balance model. *Environmental Modelling & Software, 19*(10), 943-956.

Boughton, W. C. (1993). *A hydrograph-based model for estimating the water yield of ungauged catchments*. Paper presented at the Hydrology and Water Resources Conference, Institution of Engineers, Australia.

Boughton, W. C., & Carroll, D. G. (1993). *A simple combined water balance/flood hydrograph model*. Paper presented at the Hydrology and Water Resources Conference, Institution of Engineers, Australia.

Broderick, C., Matthews, T., Wilby, R. L., Bastola, S., & Murphy, C. (2016). Transferability of hydrological models and ensemble averaging methods between contrasting climatic periods. *Water Resources Research, 52*(10), 8343-8373. doi:doi:10.1002/2016WR018850

Chen, J., Brissette, F. P., & Leconte, R. (2010). A daily stochastic weather generator for preserving low-frequency of climate variability. *Journal of Hydrology, 388*(3), 480-490. doi:https://doi.org/10.1016/j.jhydrol.2010.05.032

Cheng, C. T., Ou, C. P., & Chau, K. W. (2002). Combining a fuzzy optimal model with a genetic algorithm to solve multi-objective rainfall–runoff model calibration. *Journal of Hydrology, 268*(1), 72-86. doi:https://doi.org/10.1016/S0022-1694(02)00122-1

Chiew, F. H. S., & McMahon, T. A. (1991). The applicability of Morton's and Penman's evapotranspiration estimates in rainfall-runoff modelling. *JAWRA Journal of the American Water Resources Association, 27*(4), 611-620.

Chiew, F. H. S., Teng, J., Vaze, J., Post, D. A., Perraud, J. M., Kirono, D. G. C., et al. (2009). Estimating climate change impact on runoff across southeast Australia: Method, results, and implications of the modeling method. *Water Resources Research, 45*(10), W10414. doi:10.1029/2008WR007338

This article is protected by copyright. All rights reserved.

Author Manuscript

Christensen, N. S., Wood, A. W., Voisin, N., Lettenmaier, D. P., & Palmer, R. N. (2004). The Effects of Climate Change on the Hydrology and Water Resources of the Colorado River Basin. *Climatic Change*, *62*(1), 337-363. doi:10.1023/B:CLIM.0000013684.13621.1f

Clark, M. P., Wilby, R. L., Gutmann, E. D., Vano, J. A., Gangopadhyay, S., Wood, A. W., et al. (2016). Characterizing Uncertainty of the Hydrologic Impacts of Climate Change. *Current Climate Change Reports*, *2*(2), 55-64. doi:10.1007/s40641-016-0034-x

Coron, L., Andréassian, V., Perrin, C., Bourqui, M., & Hendrickx, F. (2014). On the lack of robustness of hydrologic models regarding water balance simulation: a diagnostic approach applied to three models of increasing complexity on 20 mountainous catchments. *Hydrol. Earth Syst. Sci.*, *18*(2), 727-746. doi:10.5194/hess-18-727-2014

Coron, L., Andréassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M., et al. (2012). Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments. *Water Resources Research*, *48*(5), n/a-n/a. doi:10.1029/2011WR011721

Croke, B. F. W., & Jakeman, A. J. (2004). A catchment moisture deficit module for the IHACRES rainfall-runoff model. *Environmental Modelling and Software with Environment Data News*, *19*(1), 1-5. doi:10.1016/j.envsoft.2003.09.001

Crossman, J., Futter, M. N., Oni, S. K., Whitehead, P. G., Jin, L., Butterfield, D., et al. (2013). Impacts of climate change on hydrology and water quality: Future proofing management strategies in the Lake Simcoe watershed, Canada. *Journal of Great Lakes Research*, *39*(1), 19-32. doi:http://dx.doi.org/10.1016/j.jglr.2012.11.003

CSIRO and Bureau of Meteorology. (2015). *Climate Change in Australia Information for Australia's Natural Resource Management Regions: Technical Report*. Retrieved from Australia:

Culley, S., Noble, S., Yates, A., Timbs, M., Westra, S., Maier, H. R., et al. (2016). A bottom-up approach to identifying the maximum operational adaptive capacity of water resource systems to a changing climate. *Water Resources Research*, *52*(9), 6751-6768. doi:10.1002/2015WR018253

Duan, Q., Gupta, V. K., & Sorooshian, S. (1993). Shuffled complex evolution approach for effective and efficient global minimization. *Journal of optimization theory and applications*, *76*(3), 501-521.

Fowler, K. J. A., Peel, M. C., Western, A. W., Zhang, L., & Peterson, T. J. (2016). Simulating runoff under changing climatic conditions: Revisiting an apparent deficiency of conceptual rainfall-runoff models. *Water Resources Research*, *52*(3), 1820-1846. doi:10.1002/2015WR018068

Gibbs, M. S., Maier, H. R., & Dandy, G. C. (2012). A generic framework for regression regionalization in ungauged catchments. *Environmental Modelling & Software*, *27-28*, 1-14. doi:https://doi.org/10.1016/j.envsoft.2011.10.006

Guo, D., Westra, S., & Maier, H. R. (2016). An inverse approach to perturb historical rainfall data for scenario-neutral climate impact studies. *Journal of Hydrology, Advance online publication*. doi:http://dx.doi.org/10.1016/j.jhydrol.2016.03.025

Guo, D., Westra, S., & Maier, H. R. (2017a). Impact of evapotranspiration process representation on runoff projections from conceptual rainfall-runoff models. *Water Resources Research*, *53*. doi:10.1002/2016WR019627

Guo, D., Westra, S., & Maier, H. R. (2017b). Use of a scenario-neutral approach to identify the key hydro-meteorological attributes that impact runoff from a natural catchment. *Journal of Hydrology*, *554*, 317-330. doi:https://doi.org/10.1016/j.jhydrol.2017.09.021

This article is protected by copyright. All rights reserved.

Author Manuscript

Guo, D., Westra, S., & Maier, H. R. (2018). An inverse approach to perturb historical rainfall data for scenario-neutral climate impact studies. *Journal of Hydrology, 556*, 877-890. doi:https://doi.org/10.1016/j.jhydrol.2016.03.025

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. doi:https://doi.org/10.1016/j.jhydrol.2009.08.003

Haque, M. M., Rahman, A., Hagare, D., Kibria, G., & Karim, F. (2015). Estimation of catchment yield and associated uncertainties due to climate change in a mountainous catchment in Australia. *Hydrological Processes, 29*(19), 4339-4349. doi:10.1002/hyp.10492

Hauser, R., Archer, S., Backlund, P., Hatfield, J., Janetos, A., Lettenmaier, D., et al. (2009). The effects of climate change on US Ecosystems. *US Global Change Research Program*.

IPCC. (2014). *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Retrieved from Cambridge, United Kingdom and New York, NY, USA:

Johnson, F., & Sharma, A. (2011). Accounting for interannual variability: A comparison of options for water resources climate change impact assessments. *Water Resources Research, 47*(4), n/a-n/a. doi:10.1029/2010WR009272

Kavetski, D., & Kuczera, G. (2007). Model smoothing strategies to remove microscale discontinuities and spurious secondary optima in objective functions in hydrological calibration. *Water Resources Research, 43*(3). doi:doi:10.1029/2006WR005195

Kay, A. L., & Davies, H. N. (2008). Calculating potential evaporation from climate model data: A source of uncertainty for hydrological climate change impacts. *Journal of Hydrology, 358*(3–4), 221-239. doi:http://dx.doi.org/10.1016/j.jhydrol.2008.06.005

Kay, A. L., Davies, H. N., Bell, V. A., & Jones, R. G. (2009). Comparison of uncertainty sources for climate change impacts: flood frequency in England. *Climatic Change, 92*(1-2), 41-63. doi:10.1007/s10584-008-9471-4

Kay, A. L., & Jones, R. G. (2012). Comparison of the use of alternative UKCP09 products for modelling the impacts of climate change on flood frequency. *Climatic Change, 114*(2), 211-230. doi:10.1007/s10584-011-0395-z

Kwon, H.-H., Lall, U., & Obeysekera, J. (2009). Simulation of daily rainfall scenarios with interannual and multidecadal climate cycles for South Florida. *Stochastic Environmental Research and Risk Assessment, 23*(7), 879-896.

Merz, R., Parajka, J., & Blöschl, G. (2011). Time stability of catchment model parameters: Implications for climate impact analyses. *Water Resources Research, 47*(2), n/a-n/a. doi:10.1029/2010WR009505

Paton, F. L., Maier, H. R., & Dandy, G. C. (2013). Relative magnitudes of sources of uncertainty in assessing climate change impacts on water supply security for the southern Adelaide water supply system. *Water Resources Research, 49*(3), 1643-1667. doi:10.1002/wrcr.20153

Paton, F. L., Maier, H. R., & Dandy, G. C. (2014). Including adaptation and mitigation responses to climate change in a multiobjective evolutionary algorithm framework for urban water supply systems incorporating GHG emissions. *Water Resources Research, 50*(8), 6285-6304. doi:10.1002/2013WR015195

Perrin, C., Michel, C., & Andréassian, V. (2003). Improvement of a parsimonious model for streamflow simulation. *Journal of Hydrology, 279*(1), 275-289.

This article is protected by copyright. All rights reserved.

Author Manuscript

Prudhomme, C., Jakob, D., & Svensson, C. (2003). Uncertainty and climate change impact on the flood regime of small UK catchments. *Journal of Hydrology, 277*(1), 1-23. doi:https://doi.org/10.1016/S0022-1694(03)00065-9

Prudhomme, C., & Williamson, J. (2013). Derivation of RCM-driven potential evapotranspiration for hydrological climate change impact analysis in Great Britain: a comparison of methods and associated uncertainty in future projections. *Hydrology and Earth System Sciences, 17*(4), 1365-1377.

Rajah, K., O'Leary, T., Turner, A., Petrakis, G., Leonard, M., & Westra, S. (2014). Changes to the temporal distribution of daily precipitation. *Geophysical Research Letters, 41*(24), 8887-8894.

Raupach, M., Briggs, P., Haverd, V., King, E., Paget, M., & Trudinger, C. (2009). *Australian water availability project (AWAP): CSIRO marine and atmospheric research component: final report for phase 3*. Retrieved from

Raupach, M., Briggs, P., Haverd, V., King, E., Paget, M., & Trudinger, C. (2012). *Australian Water Availability Project. CSIRO Marine and Atmospheric Research, Canberra, Australia.* Retrieved from: http://www.csiro.au/awap

Shafii, M., & De Smedt, F. (2009). Multi-objective calibration of a distributed hydrological model (WetSpa) using a genetic algorithm. *Hydrol. Earth Syst. Sci., 13*(11), 2137-2149. doi:10.5194/hess-13-2137-2009

Shin, M.-J., Guillaume, J. H. A., Croke, B. F. W., & Jakeman, A. J. (2015). A review of foundational methods for checking the structural identifiability of models: Results for rainfall-runoff. *Journal of Hydrology, 520*, 1-16. doi:https://doi.org/10.1016/j.jhydrol.2014.11.040

Srikanthan, R., & Zhou, S. L. (2003). Stochastic Generation of Climate Data. Working Document 03/12.

Stern, H., De Hoedt, G., & Ernst, J. (2000). Objective classification of Australian climates. *Australian Meteorological Magazine, 49*(2), 87-96.

Turral, H., Burke, J., & Faurès, J.-M. (2011). *Climate change, water and food security*: FAO.

Vaze, J., Post, D. A., Chiew, F. H. S., Perraud, J. M., Viney, N. R., & Teng, J. (2010). Climate non-stationarity – Validity of calibrated rainfall–runoff models for use in climate change studies. *Journal of Hydrology, 394*(3–4), 447-457. doi:http://dx.doi.org/10.1016/j.jhydrol.2010.09.018

Wasko, C., & Sharma, A. (2015). Steeper temporal distribution of rain intensity at higher temperatures within Australian storms. *Nature Geoscience, 8*, 527. doi:10.1038/ngeo2456

https://www.nature.com/articles/ngeo2456#supplementary-information

Westra, S., Alexander, L. V., & Zwiers, F. W. (2013). Global increasing trends in annual maximum daily precipitation. *Journal of Climate, 26*(11), 3904-3918.

Westra, S., Fowler, H. J., Evans, J. P., Alexander, L. V., Berg, P., Johnson, F., et al. (2014). Future changes to the intensity and frequency of short-duration extreme rainfall. *Reviews of Geophysics, 52*(3), 522-555. doi:10.1002/2014RG000464

Wilby, R. L., & Harris, I. (2006). A framework for assessing uncertainties in climate change impacts: Low -flow scenarios for the River Thames, UK. *Water Resources Research, 42*(2). doi:doi:10.1029/2005WR004065

Wilby, R. L., Whitehead, P. G., Wade, A. J., Butterfield, D., Davis, R. J., & Watts, G. (2006). Integrated modelling of climate change impacts on water resources and quality in a lowland catchment: River Kennet, UK. *Journal of Hydrology, 330*(1), 204-220. doi:https://doi.org/10.1016/j.jhydrol.2006.04.033

This article is protected by copyright. All rights reserved.

Wilks, D. S. (1999). Interannual variability and extreme-value characteristics of several stochastic daily precipitation models. *Agricultural and Forest Meteorology, 93*(3), 153-169. doi:http://dx.doi.org/10.1016/S0168-1923(98)00125-7

Zheng, F., Maier, H. R., Wu, W., Dandy, G. C., Gupta, H. V., & Zhang, T. (2018). On Lack of Robustness in Hydrological Model Development Due to Absence of Guidelines for Selecting Calibration and Evaluation Data: Demonstration for Data *Water Resources Research, 54*(2), 1013-1030. doi:doi:10.1002/2017WR021470

Zheng, F., Westra, S., & Leonard, M. (2015). Opposing local precipitation extremes. *Nature Climate Change, 5*, 389. doi:10.1038/nclimate2579

https://www.nature.com/articles/nclimate2579#supplementary-information

Author Manuscript

This article is protected by copyright. All rights reserved.

Author Manuscript

Tables

Table 1. Key properties and hydrologic features of the three case study catchments.

	Scott Creek	**Black River**	**Coen River**
Area (km^2)	29	318.5	170
Köppen climate classification	Distinctly dry (and warm) summer	No dry season (mild summer)	Savanna
Daily average rainfall (mm/d)	2.37	3.54	3.88
Daily average runoff (mm/d)	0.34	1.65	1.88
Daily average PET (mm/d)	3.26	2.52	5.07
Runoff ratio, Q/P	0.14	0.47	0.49
Aridity, PET/P	1.38	0.71	1.31
Annual average wet days (P > 0.1mm)	173	249	131

Table 2. Definitions and baseline values of the four runoff attributes which are used to assess the potential robustness of three CRR models. Note the difference in the choice of dominant flow season for Coen River.

Runoff attribute (mm/d)	**Definition**	**Scott Creek**	**Black River**	**Coen River**
Qavg	Daily average flow over all days	0.34	1.65	1.88
Qbase	Daily average base flow	0.13	0.81	0.68
Q5	95^{th} percentile of daily flow over all days when flow occurs (Q>0)	1.38	6.41	9.71
QJJA	Daily flow averaged over winter (June, July and August)	0.80	3.30	-
QDJF	Daily flow averaged over summer (December, January and February)	-	-	3.34

This article is protected by copyright. All rights reserved.

Author Manuscript

Figure Captions

Figure 1. Different ways to relate the conversion from rainfall to effective rainfall depending on store levels in GR4J, AWBM and CMD (note the use of a reversed x-axis to represent moisture deficit). Contribution from percolation in GR4J is neglected. See text below for the definitions of symbols, and detailed description of model structures.

Figure 2. Map of three case study catchments. Coloring relates to Köppen climate classifications from Stern et al. (2000).

Figure 3. Schematic of the study approach and linkage to results for each case study catchment.

Figure 4. Rainfall and runoff conditions for the 35 sub-periods used for calibrating the PP models at each case study (columns). Panels (a)-(c) show the variability in annual rainfall and runoff, panels (d)-(f) illustrate conditions for each 10-year calibration period (a-axes indicate starting year of each 10-year calibration period). All rainfall and runoff conditions are summarized as percentage anomalies to that of the entire calibration period (45 years).

Figure 5. Calibration performance for the FP version of each CRR model (dashed line), and those for the 35 corresponding PP versions of each CRR model (dots), summarized by NSE. The x-axes show the starting year of the 10-year period that each PP model is calibrated to.

Figure 6. Robustness of three CRRs in the performance of simulating historical runoff over the full study period (1970-2014) at Scott Creek (panels a-e), Black River (panels f-j) and Coen River (panels k-o), summarized as NSE and the percentage bias in four runoff attributes including: daily mean flow, daily mean base flow, high flow and daily mean flow in the dominant flow season (winter for Scott and Black and summer for Coen).

Figure 7. Potential robustness of three CRRs at Scott Creek (panels (a)-(d)), Black River (panels (e)-(h)) and Coen River (panels (i)-(l)), for the predictions of changes in daily mean flow, daily mean base flow, high flow (95 percentiles of all non-zero daily flows) and daily mean flow in the dominant runoff season (winter for Scott Creek and Black River, and summer for Coen River). Runoff predictions are in response to different plausible changes in average rainfall (x-axes), and are presented as relative percentage change to the historical simulation from the FP version of each CRR model (y-axes).

Figure 8. Absolute range of predicted changes across PP versions of GR4J, AWBM and CMD, for four runoff attributes as in Figure 7. Panels (a) to (d) are used for each runoff attribute, and x-axes indicate different case studies.

Figure 9. Ranges of the parameters of the 35 PP versions of GR4J, AWBM and CMD, at each of the three catchments.

Figure 10. Simulated effective rainfall as a proportion of rainfall on wet days (first row), and simulated levels of: daily production store, as ratios to the production store capacity (first row); effective rainfall (second row); and daily routing store, as ratios to the routing store capacity

This article is protected by copyright. All rights reserved.

(third row) at Black River, for a 30% decrease in average rainfall. Results are presented in three columns for GR4J (left), AWBM (middle) and CMD (right), respectively, and for the year 1983 which represents average rainfall and runoff conditions. Red dashed lines for AWBM shows capacities of the three individual production stores.

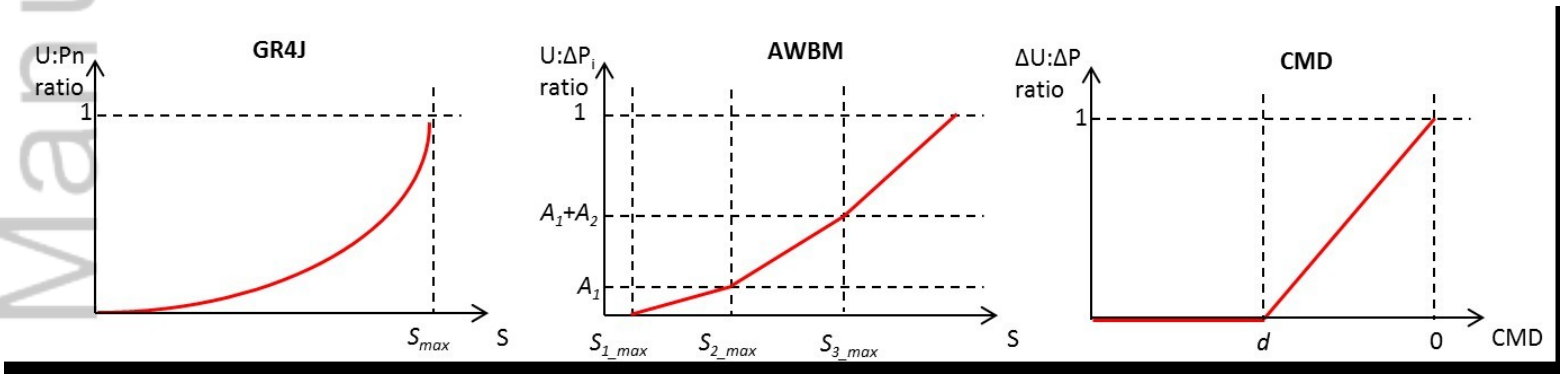
Figure 11. Simulated effective rainfall as a proportion of rainfall on wet days (first row), and simulated levels of: daily production store, as ratios to the production store capacity (first row); effective rainfall (second row); and daily routing store, as ratios to the routing store capacity (third row) at Coen River, for a 30% decrease in average rainfall. Results are presented in three columns for GR4J (left), AWBM (middle) and CMD (right), respectively, and for the year 1977 which represents average rainfall and runoff conditions. Red dashed lines for AWBM shows capacities of the three individual production stores.

Figure 12. Simulated effective rainfall as a proportion of rainfall on wet days (first row), and simulated levels of: daily production store, as ratios to the production store capacity (first row); effective rainfall (second row); and daily routing store, as ratios to the routing store capacity (third row) at Coen River, for a 10% increase in average rainfall. Results are presented in three columns for GR4J (left), AWBM (middle) and CMD (right), respectively, and for the year 1977 which represents average rainfall and runoff conditions. Red dashed lines for AWBM shows capacities of the three individual production stores.

Author Manuscript

This article is protected by copyright. All rights reserved.

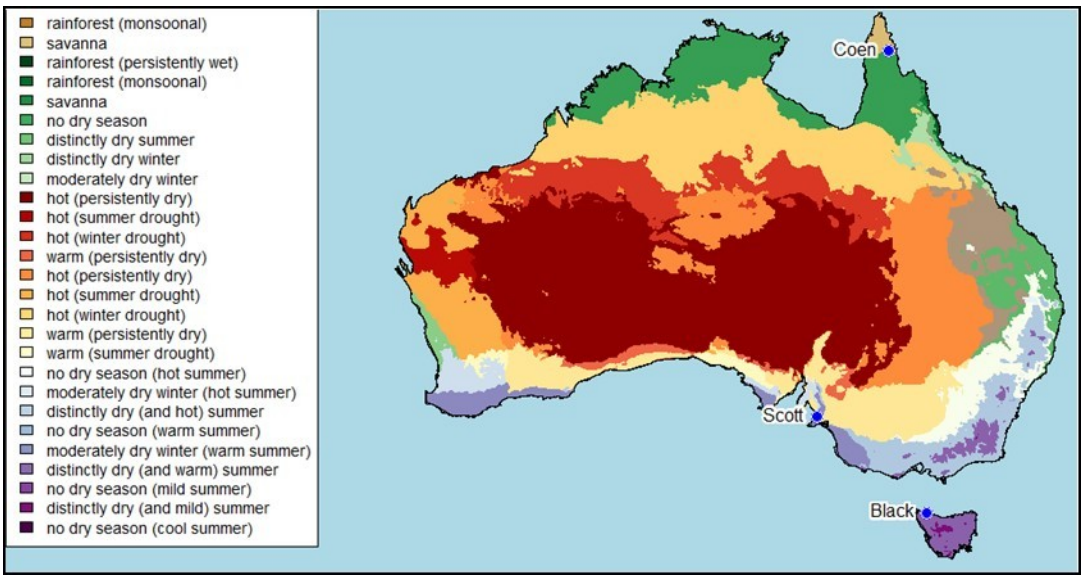
Author Manuscript



2018WR022636-f01-z-.jpg

This article is protected by copyright. All rights reserved.

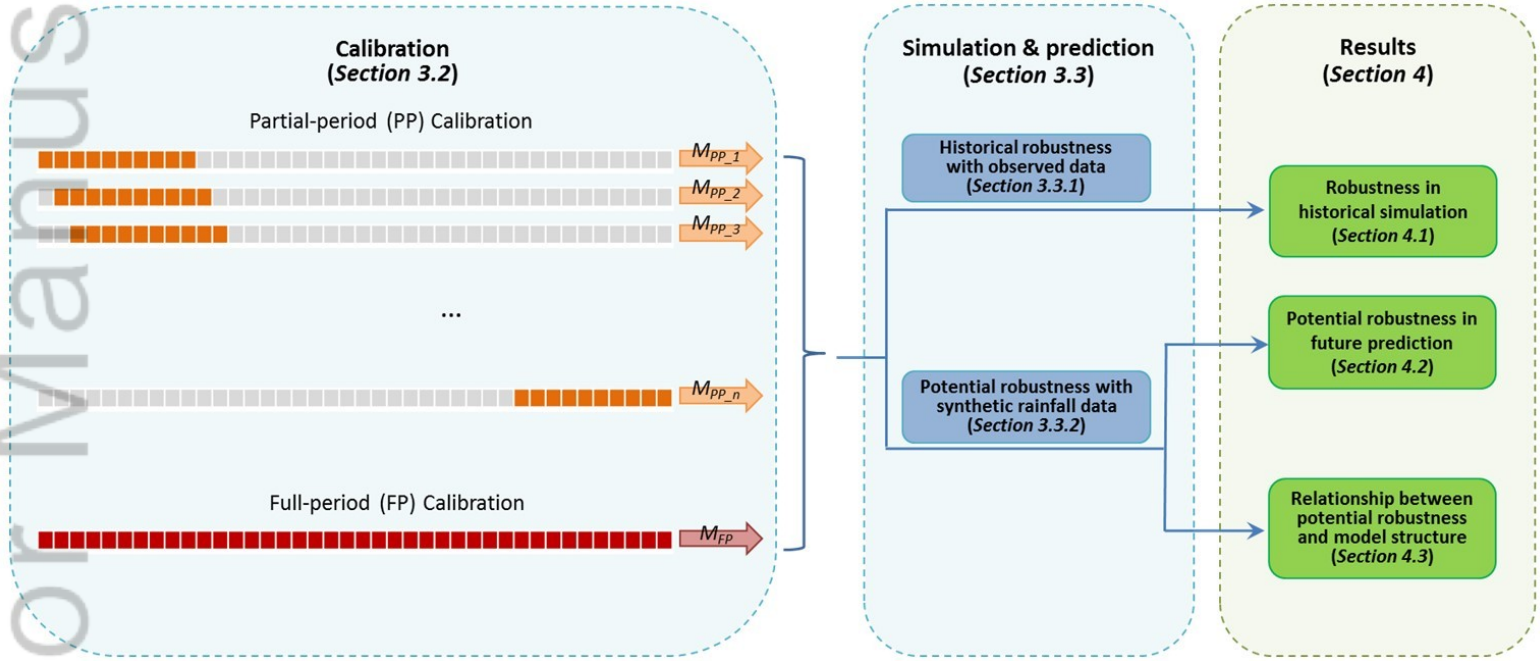
Author Manuscript



2018WR022636-f02-z-.jpg

This article is protected by copyright. All rights reserved.

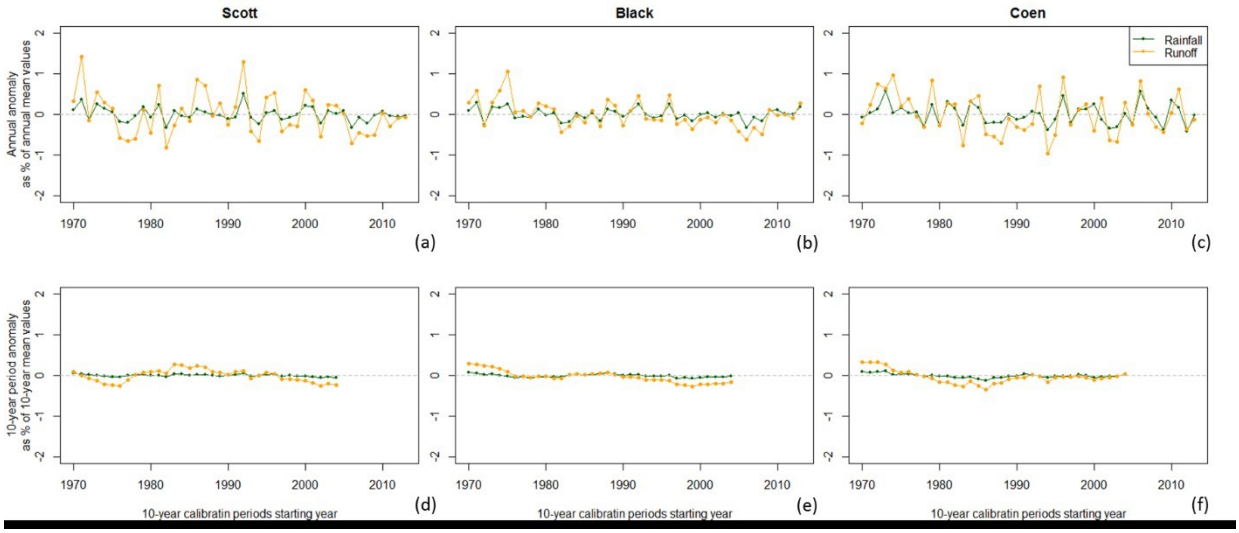
Author Manuscript



2018WR022636-f03-z-.jpg

This article is protected by copyright. All rights reserved.

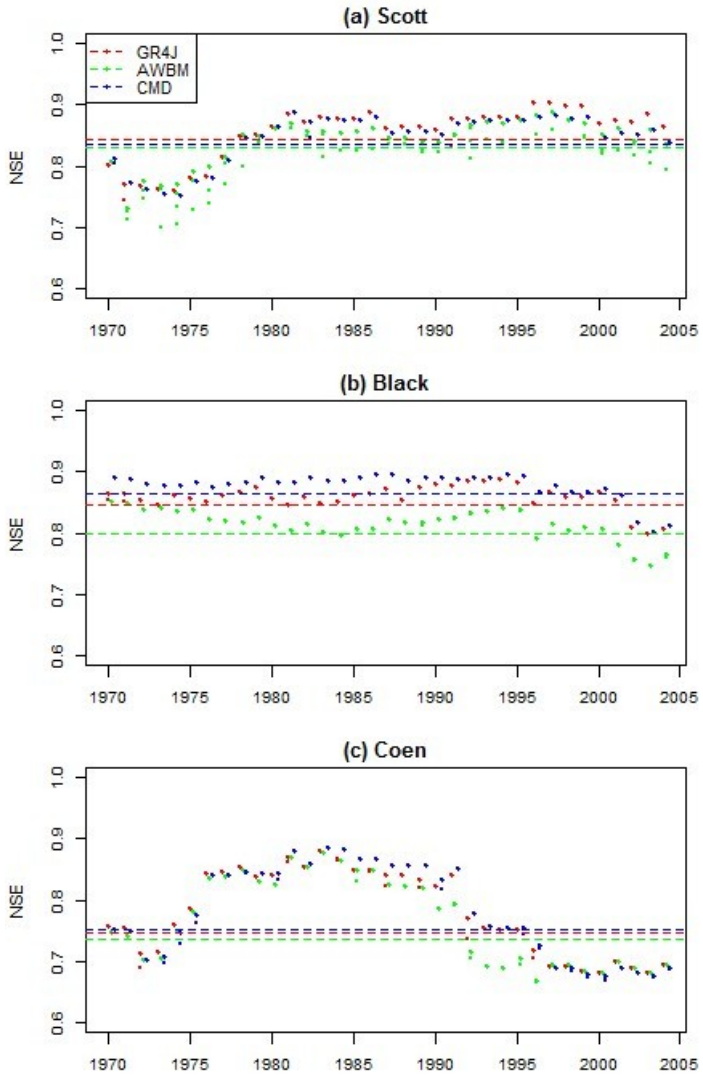
Author Manuscript



2018WR022636-f04-z-.jpg

This article is protected by copyright. All rights reserved.

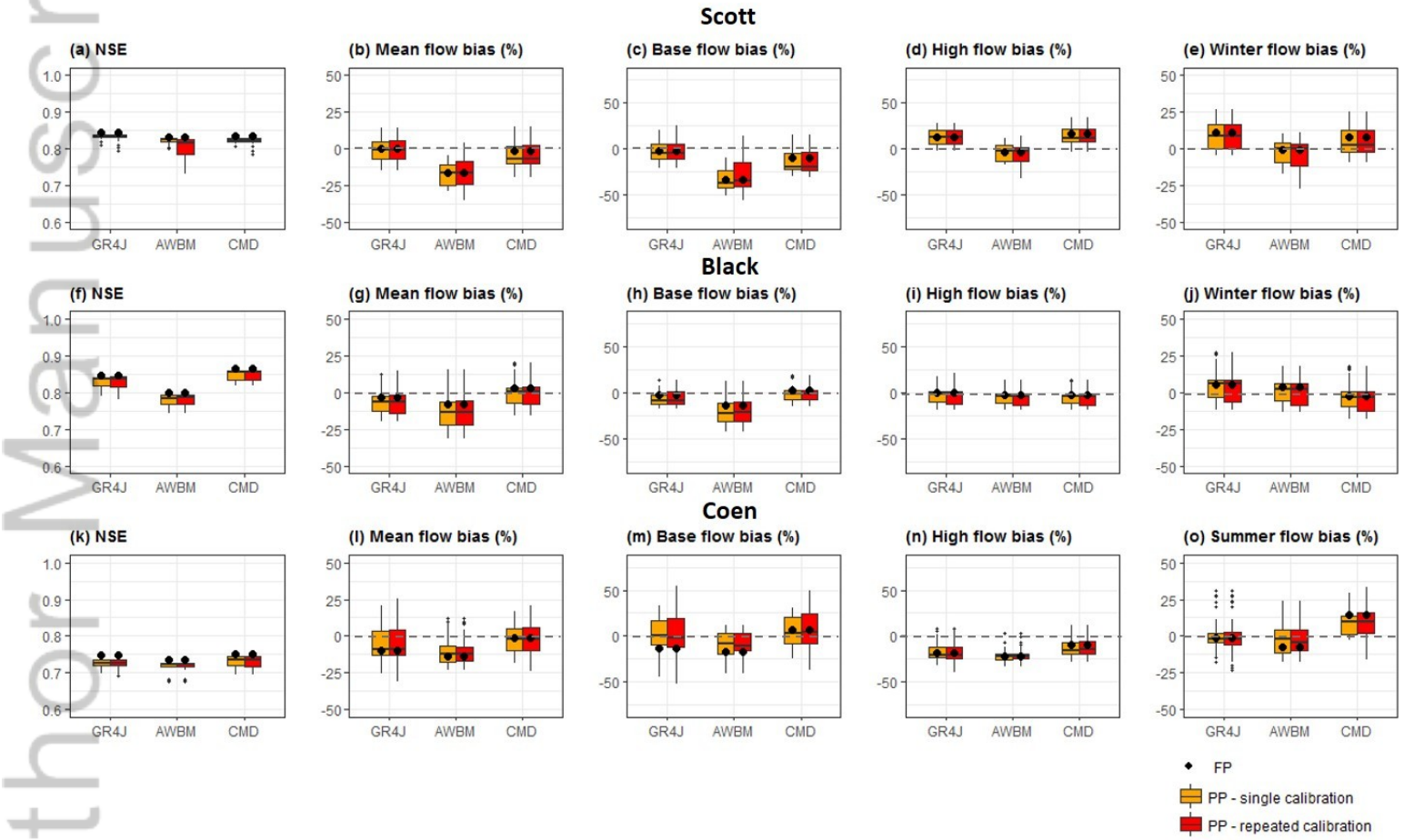
Author Manuscript



2018WR022636-f05-z-.jpg

This article is protected by copyright. All rights reserved.

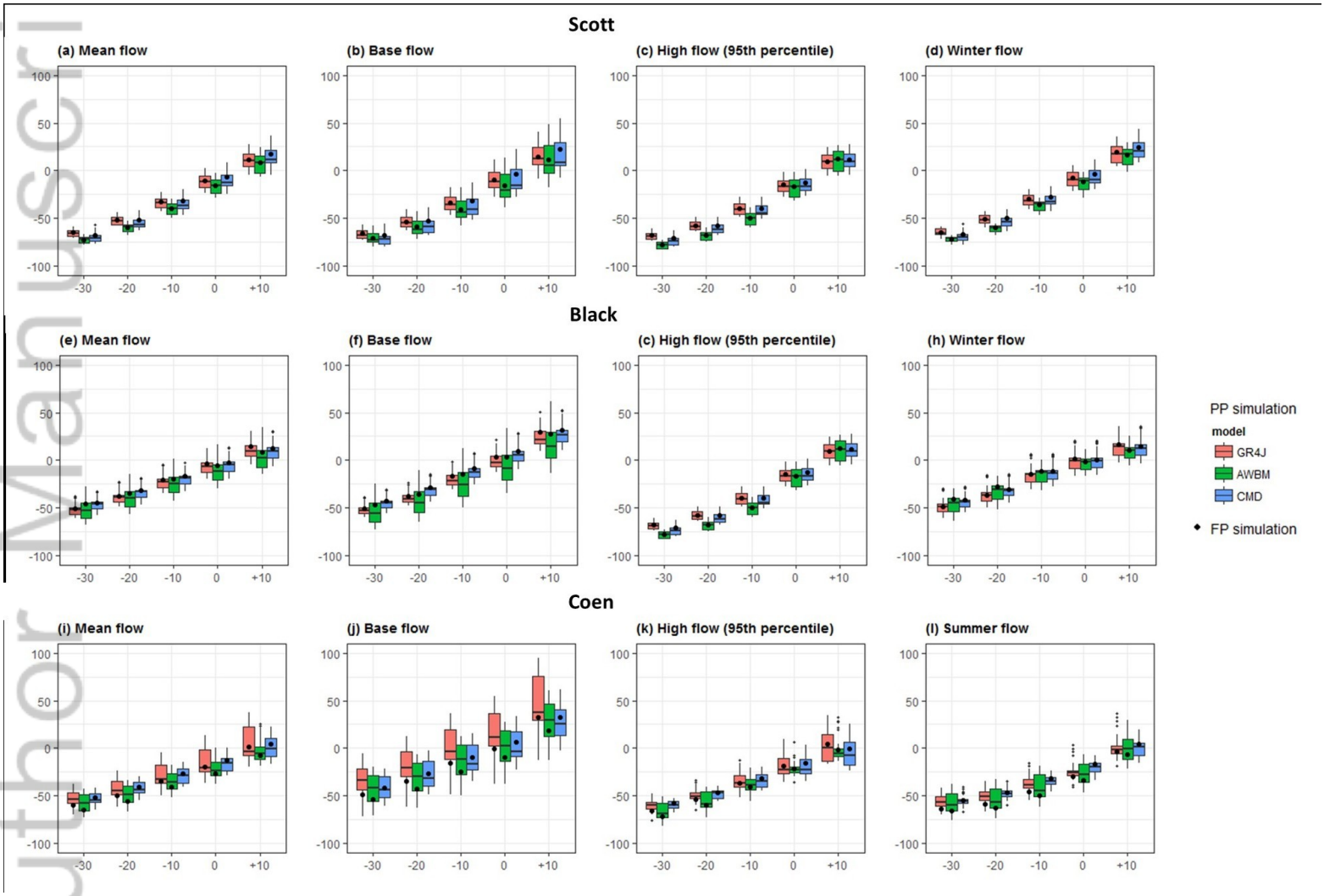
Author Manuscript



2018WR022636-f06-z-.jpg

This article is protected by copyright. All rights reserved.

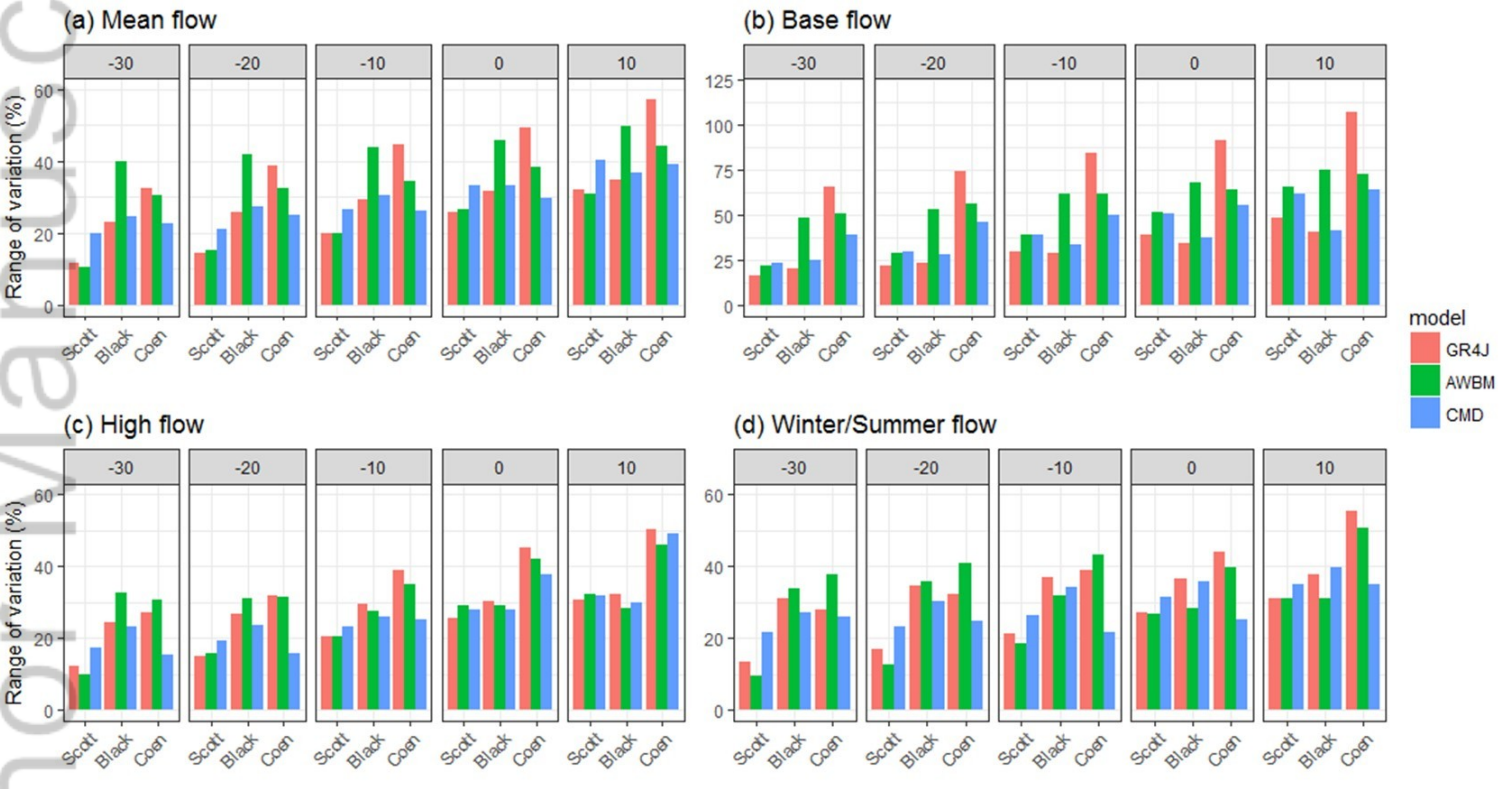
Author Manuscript



2018WR022636-f07-z-.jpg

This article is protected by copyright. All rights reserved.

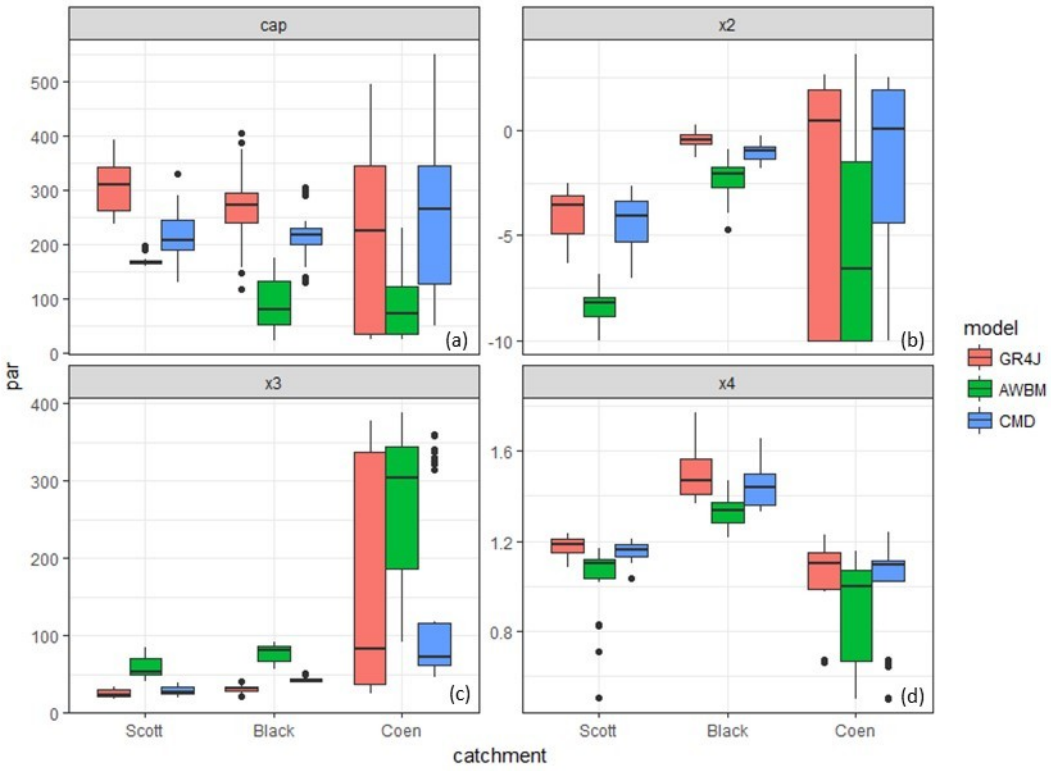
Author Manuscript



2018WR022636-f08-z-.jpg

This article is protected by copyright. All rights reserved.

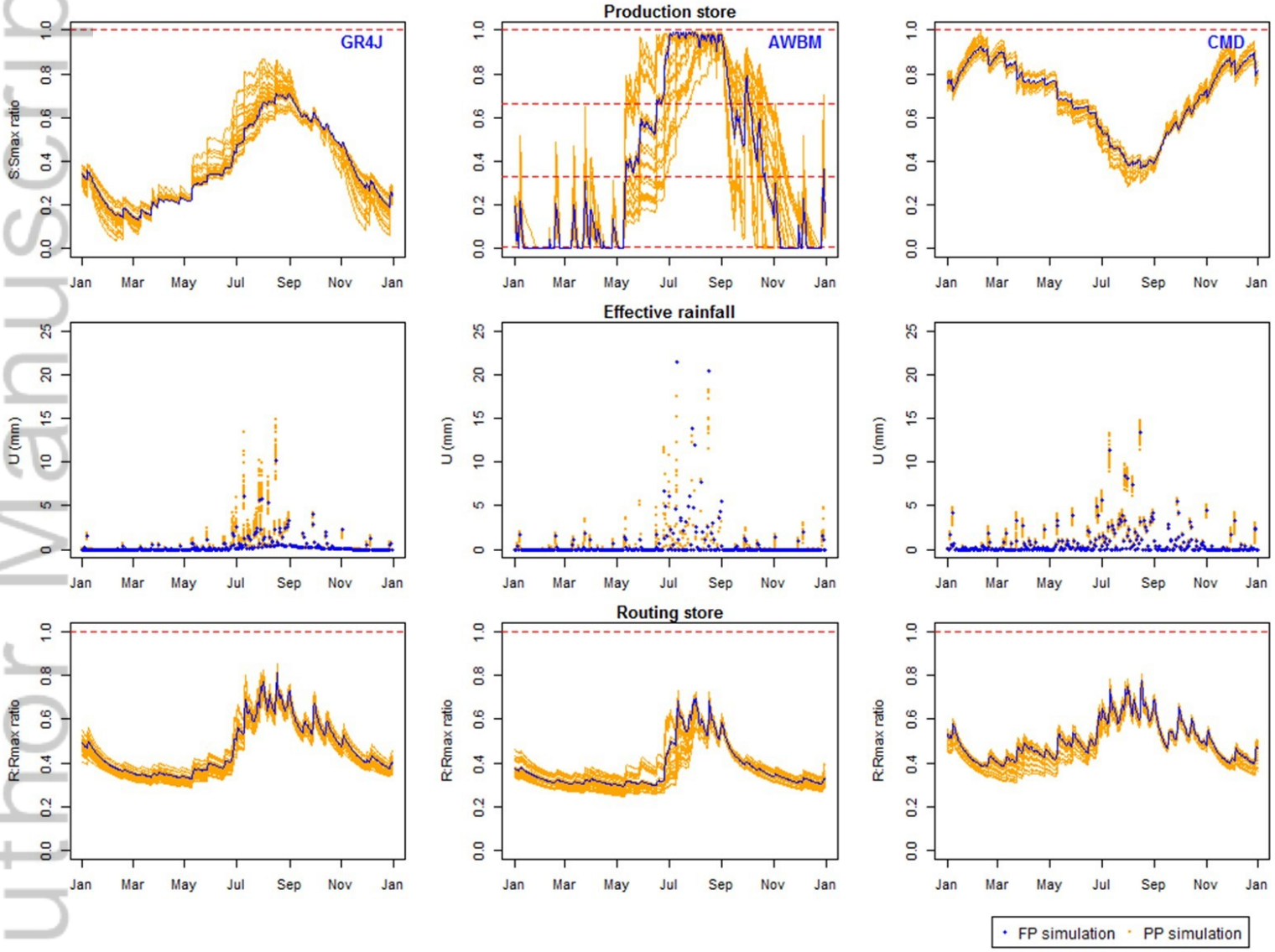
Author Manuscript



2018WR022636-f09-z-.jpg

This article is protected by copyright. All rights reserved.

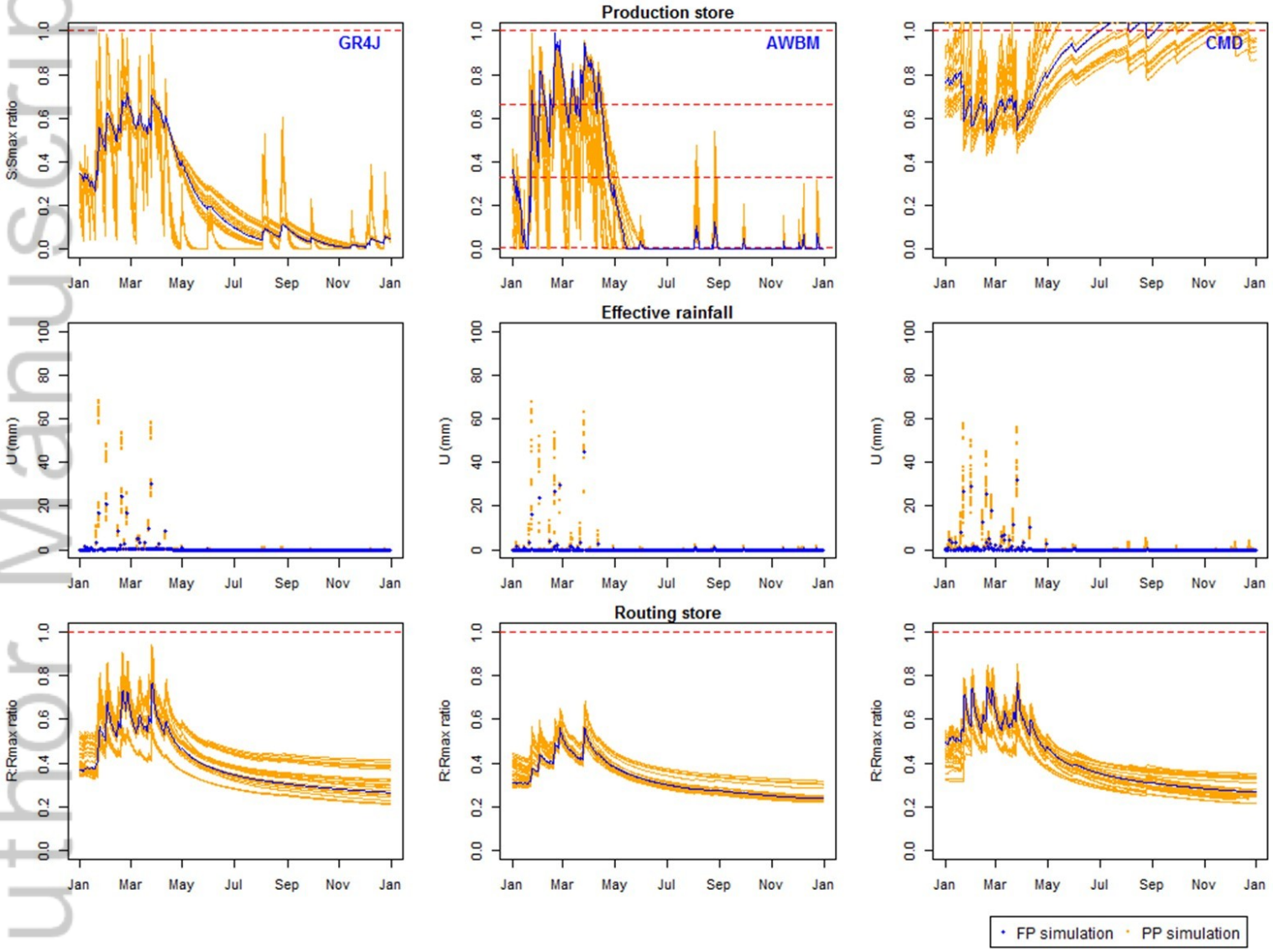
Author Manuscript



2018WR022636-f10-z-.jpg

This article is protected by copyright. All rights reserved.

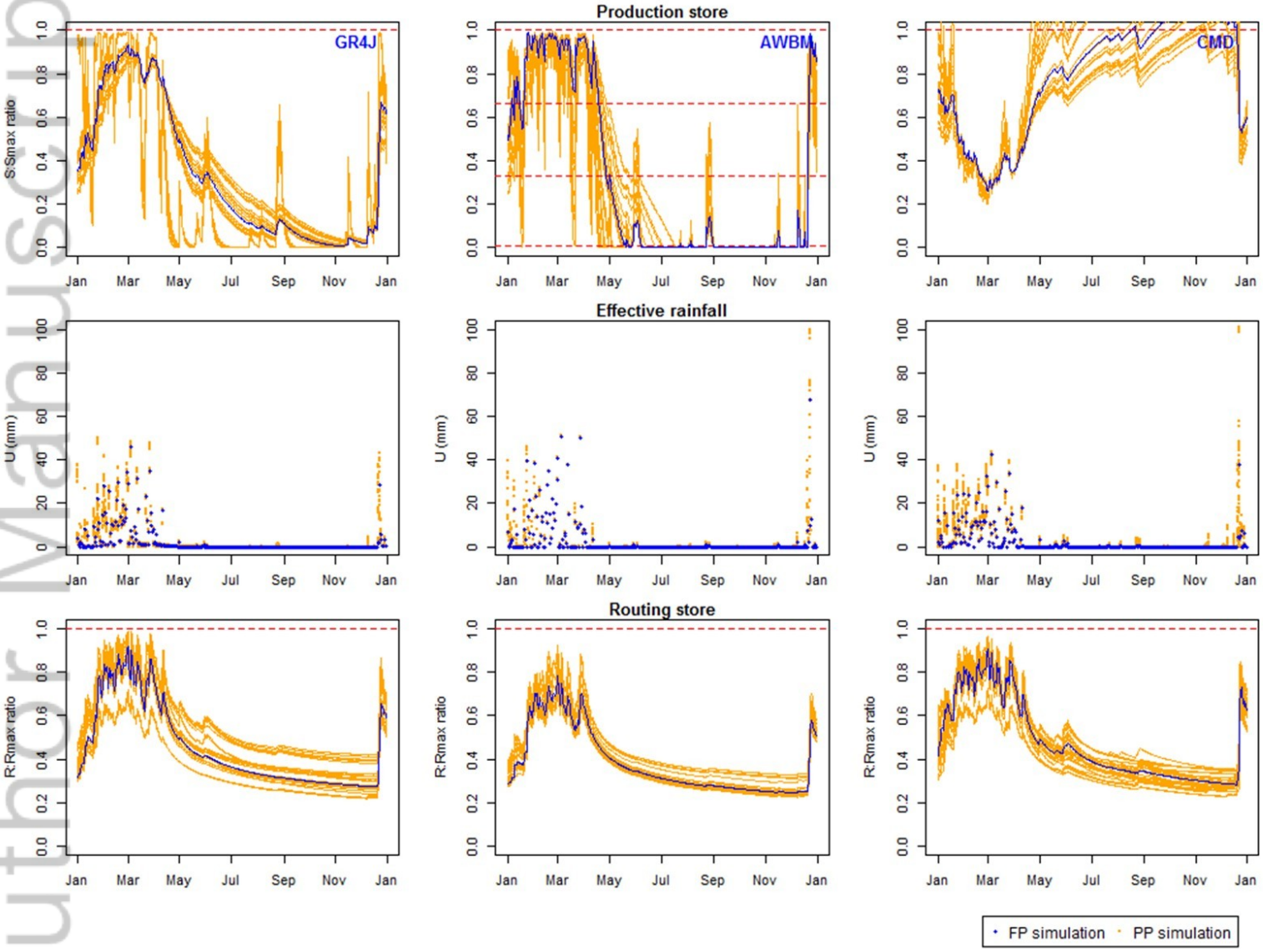
Author Manuscript



2018WR022636-f11-z-.jpg

This article is protected by copyright. All rights reserved.

Author Manuscript



2018WR022636-f12-z-.jpg

This article is protected by copyright. All rights reserved.