Symptoms of performance degradation during multi-annual drought: a large-sample, multi-model study

Luca Trotter[1,1], Margarita Saft[1,1], Murray Cameron Peel[1,1], and Keirnan James Andrew Fowler[1,1]

[1]University of Melbourne

January 20, 2023

Abstract

Hydrologic models are essential tools to understand and plan for the effect of changing climates; however, they are known to underperform in transitory climate conditions. Research to date identifies the inadequacy of models to perform during prolonged drought, but falls short on pinpointing how and which specific aspects of model performance are affected. Here, we study five conceptual rainfall-runoff models and their performance in 155 Australian catchments which recently experienced a 13-year long dry period, with a focus on a wide range of performance metrics. We show that model performance degrades extensively during the drought across most metrics, with overestimation of flow volumes driving the decline and representation of shape and variability of the hydrograph and the flow-duration curve being more resilient to the prolonged dry climate. This indicates that the overestimation is not linked to specific flow regimes, but is the result of proportional flow decline throughout the hydrograph, suggesting engagement of multiple catchment processes in determining the changes in flow during the drought across high and low flow periods as well as through faster and slower flow routes. Additionally, we show that in most cases model performance does not recover after the end of the drought and that the multi-annual nature of the drought is the likely reason for exacerbated performance decline due to accumulation and aggravation of errors over subsequent dry years. By promoting detailed investigation of models' shortcomings, we hope to foster the development of more resilient model structures to improve applicability within climate change scenarios.

Symptoms of performance degradation during multi-annual drought: a large-sample, multi-model study

Luca Trotter[1], **Margarita Saft**[1], **Murray C. Peel**[1], **Keirnan J. A. Fowler**[1]

[1]Department of Infrastructure Engineering, University of Melbourne, Melbourne, Victoria, Australia

Key Points:

- We compare aspects of model performance during and after multi-annual drought against pre-drought performance
- Performance degradation is driven by bias in water balance estimates rather than errors in hydrograph shape
- Accumulation and aggravation of errors over multiple dry years exacerbates performance degradation

Corresponding author: L. Trotter, `l.trotter@unimelb.edu.au`

Abstract

Hydrologic models are essential tools to understand and plan for the effect of changing climates; however, they are known to underperform in transitory climate conditions. Research to date identifies the inadequacy of models to perform during prolonged drought, but falls short on pinpointing how and which specific aspects of model performance are affected. Here, we study five conceptual rainfall-runoff models and their performance in 155 Australian catchments which recently experienced a 13-year long dry period, with a focus on a wide range of performance metrics. We show that model performance degrades extensively during the drought across most metrics, with overestimation of flow volumes driving the decline and representation of shape and variability of the hydrograph and the flow-duration curve being more resilient to the prolonged dry climate. This indicates that the overestimation is not linked to specific flow regimes, but is the result of proportional flow decline throughout the hydrograph, suggesting engagement of multiple catchment processes in determining the changes in flow during the drought across high and low flow periods as well as through faster and slower flow routes. Additionally, we show that in most cases model performance does not recover after the end of the drought and that the multi-annual nature of the drought is the likely reason for exacerbated performance decline due to accumulation and aggravation of errors over subsequent dry years. By promoting detailed investigation of models' shortcomings, we hope to foster the development of more resilient model structures to improve applicability within climate change scenarios.

1 Introduction

Hydrological modelling is crucial for climate change assessment and adaptation studies. Atmospheric and climatic changes modify rainfall and temperature patterns, affecting water availability for humans and natural ecosystems, as well as the frequency and intensity of extreme hydroclimatic events (Milly et al., 2008). Future climate conditions are expected to deviate from observed historical records in many regions of the world (Hewitson et al., 2014) and hydrological models are a useful tool to assess risks associated with such changing climates as well as strategies and opportunities for adaptation and mitigation (Xu, 1999). Nevertheless, it is known that hydrological models underperform in changing climate conditions (Seibert, 2003; Peel & Blöschl, 2011). These limitations of contemporary hydrologic modelling are particularly evident under drying climate conditions, especially during multiyear drought (Coron et al., 2012; Deb & Kiem, 2020; Li et al., 2012; Vaze et al., 2010).

Drought is the most impactful and widespread natural disaster, threatening half of the earth's land surface (Mishra & Singh, 2010). In recent decades severe drought conditions have been reported in the Amazon (2005, 2010), Australia (1997–2009), California (2011–2014), Chile (2010–2018), China (2009–2011), Europe (2003, 2005), and the Horn of Africa (2011), amongst others (Feyen & Dankers, 2009; Sun & Yang, 2012; van Dijk et al., 2013; Mann & Gleick, 2015; Rowell et al., 2015; Marengo & Espinoza, 2016; Garreaud et al., 2020). Despite high levels of uncertainty in determining trends from changes in historical patterns of drought and attributing them to anthropogenic climate change (Dai & Zhao, 2017; Cook et al., 2018), the IPCC's sixth assessment report projects exacerbated risks of agricultural, ecological and hydrological drought in several regions of the world under future climate scenarios, driven by changed precipitation patterns, reduced soil moisture and increased potential evapotranspiration (Douville et al., 2021; Seneviratne et al., 2021). Because of this, the study of historical droughts as large-scale natural experiments can provide a unique insight into future climates of many drought-prone regions worldwide, which can inform scientific advancement and political action towards more farsighted climate adaptation strategies.

In particular, authors have studied the relationships between rainfall and streamflow anomalies during south-eastern Australia's Millennium drought, ca. 1997–2009, and discovered that during persistent drought, annual rainfall-runoff relationships shifted significantly in many of the catchments studied; causing reductions in streamflow disproportionate to the meteorological anomaly (Potter et al., 2010; Chiew et al., 2014; Saft et al., 2015, in preparation). In this context, the annual rainfall-runoff relationship is used to characterise a catchment's response to precipitation and any change in relationship over time can be symptomatic of a modification of a catchment's underlying hydrological behaviour through changes in its underlying processes or their relative prominence, affecting rainfall partitioning (Saft et al., 2015, in preparation). Very similar shifts in rainfall-runoff relationships during prolonged drought were more recently observed also in China (Gao et al., 2016; Tian et al., 2018; Zhang et al., 2018), California (Avanzi et al., 2019) and Chile (Alvarez-Garreton et al., 2021). Furthermore, the latest research out of south-eastern Australia suggests that the end of the dry spell is not always sufficient for catchments to recover and many catchments can persist in a low-flow state for several years after the drought, despite a return to pre-drought precipitation (Saft et al., in preparation; Peterson et al., 2021).

Such changes in hydrological response at the catchment level affect the reliability of hydrologic models' projections of streamflow and water availability. The aforementioned Millennium drought (MD), which affected an area of south-eastern continental Australia in excess of $1 \times 10^6\,\mathrm{km}^2$ between 1997–2009 (Verdon-Kidd & Kiem, 2009; van Dijk et al., 2013), exhibited these limitations of hydrologic modelling and calibration frameworks. As mentioned, the MD impacted on the hydrological behaviour of many catchments in the region, causing a shift in the long-term rainfall-runoff relationships of 50 % to 70 % of catchments in the southern Australian state of Victoria, many of which are still yet to recover (Saft et al., in preparation; Peterson et al., 2021). For these reasons, it has served as a case study for a number of studies aimed either at demonstrating the shortcomings of model structure and/or calibration methods in changing conditions (e.g. Vaze et al., 2010; Coron et al., 2012; Saft et al., 2016; Fowler et al., 2020) or suggesting methods to diagnose and improve modelling and calibration methods in nonstationary conditions (e.g. Fowler et al., 2016; Fowler, Coxon, et al., 2018). The results of these studies show a consistent degradation of hydrologic model performance when models calibrated on pre-MD data are forced with MD data (Coron et al., 2012), concentrating in catchments where a change in rainfall-runoff relationship had been observed (Saft et al., 2016). Such underperformance was shown to be mostly due to bias rather than variability, underlining that in conditions of systematic behavioural change, model ensembles are not an effective method to reduce uncertainty, and precision in simulated series isn't an indicator of low uncertainty (Saft et al., 2016).

In some cases, models can achieve more satisfactory calibration efficiency if they are shown both pre-MD and MD conditions by using a multi-objective approach to the calibration optimisation (Fowler et al., 2016). This seems to indicate that models are not structurally incapable of reproducing conditions before and during the drought and that better calibration strategies with different objective functions could help produce more reliable simulations in such changing climate conditions (Fowler, Peel, et al., 2018). However, the identification of a set of parameters able to perform over a range of climates, does not necessarily imply *adequacy* of the model to properly represent the underlying processes, but merely its ability to reproduce the observed hydrograph *well enough* (Fowler, Peel, et al., 2018; Fowler et al., 2020). Fowler et al. (2020) demonstrated this, by showing that none of the models tested were able to plausibly reproduce observed slow drying conditions observed in groundwater heads during the MD, either because they utilised the entire available storage variability in the pre-drought period, or because they failed to show any downward trend in their storage altogether (Fowler et al., 2020).

Previous research identified the inadequacy of hydrological models to perform during prolonged drought. However, due to their focus on only a couple of performance metrics (typically one overall goodness-of-fit measure and the volumetric bias), these studies largely fail to identify modes and reasons of such underperformance. This research aims at complementing existing research and providing a better understanding of how the Millennium drought affected the performance and behaviour of hydrological models. In order to address this goal, we look at a number of performance metrics useful to distinguish the ability of five hydrologic models to reproduce different portions of the hydrograph of 155 catchments in the southern Australian state of Victoria before, during and after the Millennium drought. We specifically aim to:

1. identify aspects of the flow regime that are more or less problematic for models to reproduce during and after the MD (when calibrated on pre-MD data); and
2. estimate how the performance of models during the years of the MD (and after) compares to their performance in individual years of similar dryness in the period before the drought.

Together with the focus on a more comprehensive set of performance metrics and addressing the issue of post-drought recovery by analysing model performance in the post-MD period, this study differentiates itself from previous ones by providing fairer and less biased estimates of model performance degradation by comparing MD and post-MD performance to a pre-MD evaluation benchmark, instead of the calibration performance.

2 Methods

The crux of the methods used to achieve the two objectives specified above is contained in section 2.5. Before that, we describe spatial and temporal extents of the analysis (§2.1) and the sources of data used (§2.2) and specify the settings used for calibration of hydrological modelling and their rationale (§2.3). In section 2.4, we describe the performance metrics used for this analysis, including reasoning for their use in this context.

2.1 Study extent

The spatial extent of this study is the state of Victoria. Victoria covers an area of approximately 230 000 km^2 in south-east Australia and is where some of the strongest impacts of the Millennium drought were felt (van Dijk et al., 2013). The catchments included in the research are the 155 catchments already used by Saft et al. (in preparation). Those catchments had been selected as mostly unimpaired by human influences on their flow regimes including regulation, known diversions, and land use changes (Saft et al., in preparation). The vast majority of catchments also have little to no groundwater extraction. The catchments included cover the width of Victoria from west to east on both sides of the Great Dividing Range. Climatically almost all catchments fall in the *Cfb* type according to the Köpper-Geiger classification, having a temperate climate, with no dry season and warm summers (Peel et al., 2007). Topographically they can broadly be divided between the eastern mountainous catchments, with headwaters on the Australian Alps, higher elevations and steeper slopes; and the western catchments, laying on flatter and lower ground. As seen in Figure 1c the former have generally higher annual precipitation than the latter. In the years of interest for this research, this set of catchments experienced a range of climatic and hydrological anomalies with several alternating periods of low and high rainfall and flow (Fig. 1a,b). All catchments experienced unusually persistent negative rainfall and streamflow anomalies during the Millennium drought; in many cases the streamflow deficits persisted after the end of the drought, despite a return to approximately average climatic conditions including a few wet years. Figure 1 also shows that western catchments experienced the highest reductions in streamflow during the drought, despite the rainfall anomalies being comparable between all catchments, this is consistent with findings from previous studies (Saft et al., 2015; Fowler et al., 2020).

The temporal extent of the analysis encompasses the period of available streamflow data in each catchment, typically starting in the 1960's (33.5 % of catchments) or 1970's (27.1 %). In the 29 catchments where streamflow data is available prior to 1950, 1950 is picked as the starting time for the analysis in order to ensure a more concurrent period of observation across the catchments. All but fifteen catchments have streamflow data running up to the end of the 2019 water year. Due to March and April typically being the driest months, hydrological or water years in this region conventionally start

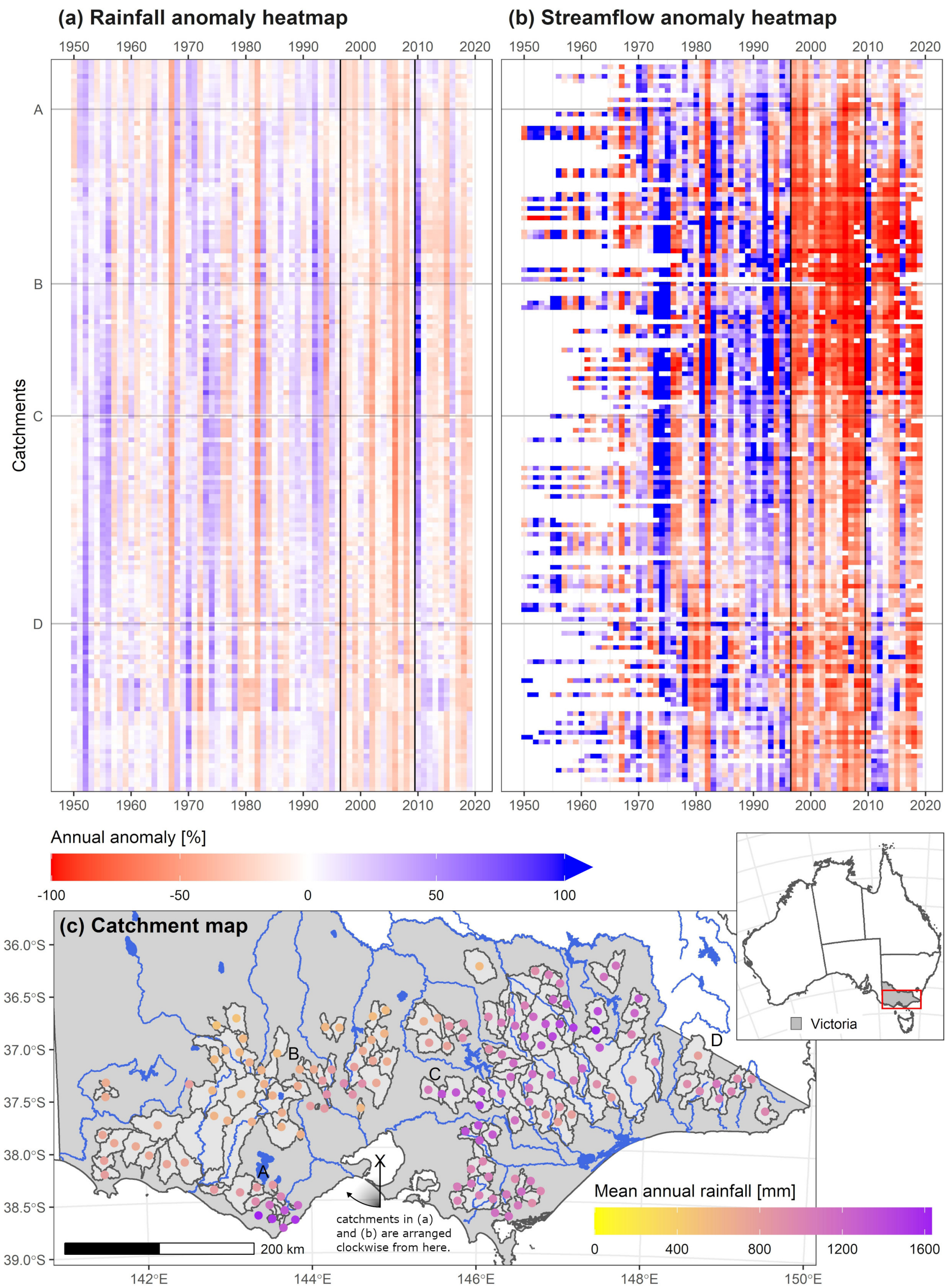

Figure 1. **(a, b)** Annual rainfall (a) and streamflow (b) anomalies for each of the catchments in this study. Each line represents a catchment. Catchments are arranged by the clockwise angle from the south axis created by connecting their centroid to the centre of Port Phillip Bay (point X in (c)). Catchments A, B, C and D are marked in (c) for reference. The vertical black lines indicate the extent of the Millennium drought. **(c)** Map of the catchments in this study with their mean annual rainfall. Each dot represent one catchment.

at the beginning of Autumn on 01 March and end on the last day of February of the subsequent year (Peterson et al., 2021).

The research period for each catchment is divided into three periods of interest: the pre-MD period, up to 1996; the MD, between 1997 and 2009; and the post-MD period, between 2010 and 2019 (or end of record). While there is some contention about the starting year of the drought (e.g. Kiem & Verdon-Kidd, 2010), these are generally the most accepted dates (CSIRO, 2012). Note that, in contrast to previous studies (e.g. Saft et al., 2015), the temporal extent of the MD in this study is not determined on a per-catchment basis.

2.2 Data sources

Gridded daily rainfall data are from the Australian Gridded Climate Data (AGCD) collection, formerly known as Australia Water Availability Project (AWAP). This dataset contains daily rainfall records interpolated from point measurements at a resolution of $0.05^{\circ}\times0.05^{\circ}$ (Jones et al., 2009). Gridded temperature (maximum and minimum) records, also interpolated from point measurements, as well as Morton's wet-environment potential evapotranspiration (Morton, 1983) data, both at the same resolution as the rainfall data, are from the SILO database (Jeffrey et al., 2001). Catchment average daily data were extracted for each of the catchments in this study. All the gridded climate data are complete at a daily timestep for the extent of this research.

The dataset of daily streamflow used for this research was collated, quality checked, infilled and used by Saft et al. (in preparation), from the WMIS portal of the Victorian Department of Environment, Land, Water and Planning (Saft et al., in preparation). As the dataset compiled by Saft et al. (in preparation) ended in 2016, it was updated for this study to extend to the end of the 2019 water year (i.e. 29 February 2020) with daily streamflow data gathered from the same source and following the same quality checks and procedures described by Saft et al. (in preparation) for consistency.

2.3 Hydrological modelling

Five conceptual, spatially lumped hydrological models are used in this study, namely IHACRES (Jakeman et al., 1990; Croke & Jakeman, 2004), GR4J (Perrin et al., 2003), SimHyd (Chiew et al., 2002), Sacramento (Burnash, 1995) and HBV (Lindström et al.,

Table 1. Characteristics of the hydrological models used in this study (Knoben, Freer, Fowler, et al., 2019)

Model name	Parameters	Stores		Routing functions
IHACRES	7	1	Soil moisture (deficit)	2
GR4J	4	2	Soil moisture	2
			Routing store	
SimHyd	7	3	Interception	0
			Soil moisture	
			Groundwater	
Sacramento	11	5	Soil moisture (5)	0
HBV	15	5	Snow store (2)	1
			Soil moisture (3)	

1997). These models were chosen to cover a range of complexities (see Table 1) and because of their widespread application in hydrological studies in and outside Australia, including in the same area and period of this study (Saft et al., 2016; Fowler et al., 2016, 2020). All models used were implemented within the MARRMoT modelling framework (Knoben, Freer, Fowler, et al., 2019; Trotter et al., in preparation).

Models were calibrated using the Covariance Matrix Adaptation Evolution Strategy, or CMA-ES (Hansen & Ostermeier, 1996; Hansen et al., 2003). CMA-ES is a widely used optimisation algorithm that performs favourably in hydrological model calibration in comparison to other algorithms (Arsenault et al., 2014). Additionally, it has been used successfully to calibrate models within the same geographical and temporal scope of this analysis (Fowler et al., 2016; Fowler, Coxon, et al., 2018) and it has also been applied in tandem with the MARRMoT modelling framework (Knoben et al., 2020).

The objective function used for the calibration is designed to ensure that models are able to reproduce both aspects of the high-flow and the low-flow portions of the hydrograph as well as ensure minimal volumetric bias (eq. 1).

$$E = \frac{1}{2}\left(KGE_Q + KGE_{Q^{0.2}}\right) - 5 \cdot |\ln(B+1)|^{2.5} \tag{1}$$

The model efficiency (E) in equation 1 is the combination of two additive parts. The first is the mean of two Kling-Gupta efficiencies, KGE (Gupta et al., 2009), one calculated using direct flows and one using their fifth root. The use of the fifth root of flows provides stronger weighing to small flows (Chiew et al., 1993) and is better suited to zero-flow conditions than the more common inverse or log transformations. The second addend of the model efficiency contains a bias penalisation, reducing the value of the efficiency as the volumetric bias (B) between simulated and observed streamflow deviates from 0 (Viney et al., 2009; Vaze et al., 2010). The use of a bias penalisation factor is motivated by the observation from previous studies that models applied to Millennium drought data showed a strongly biased response (Saft et al., 2016) and therefore it is desirable to minimise bias over the calibration period so that any bias in independent evaluation cannot be traced back to a similar error during calibration (Vaze et al., 2010). Models that did not achieve a calibration efficiency of at least 0.80 in a given catchment were calibrated a second time.

In order to reach the research goals set out in the introduction, models are calibrated on the even year of the available record in the pre-MD period. Model performance on pre-MD odd years is then used as a benchmark for MD and post-MD performance. The use of interlocking calibration and benchmarking periods is designed to expose models to the entire range of climate variability of the pre-MD period while striving to maintain climate conditions as similar as possible between calibration and benchmark. Kolmogorov-Smirnov tests were conducted to assess whether distributions of annual rainfall and potential ET in the two periods are significantly different. The p-values of the tests on rainfall (potential ET) data, adjusted using the false discovery rate method to account for the multiplicity of tests, are above 0.85 (0.5) for all catchments indicating that no significant difference in the distribution of rainfall (potential ET) exists between odd and even years in the pre-MD period. The model performance during the pre-MD odd years effectively represents how models would be expected to perform in evaluation had the climate remained stable.

2.4 Performance metrics

The metrics used to evaluate model performance are summarised in Table 2. This set of metrics is designed to assess the ability of models to reproduce different aspects of the hydrograph and they are grouped accordingly. Many of the metrics use biases to

Table 2. Model performance indicators used in this study. Equations S1 to S12 are given in the supporting information text S1.

Group	Metric	Description	eq.
Fit	*OF*	Objective function used for calibration.	1
Fit	*KGE*	Kling-Gupta efficiency (Gupta et al., 2009).	S1
Fit	*KGElo*	Kling-Gupta efficiency (Gupta et al., 2009) of fifth root of streamflows.	S2
Volumes	*Q**	Volumetric bias.	S3
Volumes	*Qbase**	Bias in baseflow volumes (Tallaksen & Van Lanen, 2004).	S4
Volumes	*Qlo**	Bias in low-flow portion of the FDC (Yilmaz et al., 2008).	S5
Volumes	*Qhi**	Bias in high-flow portion of the FDC (Yilmaz et al., 2008).	S6
Shape	*BFI**	Bias in the annual baseflow index (Tallaksen & Van Lanen, 2004).	S7
Shape	*FDCslp**	Bias in the slope of the mid-section of the annual FDC (Yilmaz et al., 2008).	S8
Shape	*sd**	Bias in the annual standard deviation.	S9
Shape	*r*	Pearson's correlation coefficient.	S10
Zeros	*pc0**	Bias in the percentage of zero-flows.	S11
Zeros	*TPR0*	True positive rate of zero flows.	S12

assess differences in statistical or hydrological properties of the observed and simulated timeseries. Note that the term *bias* here and throughout the text indicates a percentage difference between any observed and simulated quantity and is not limited to volumetric streamflow bias.

Performance metrics in the *fit* group are common performance metrics in hydrological modelling and represent summary goodness-of-fit measures to assess overall model performance. The form of the objective function (OF) and the use of the fifth-root transformation in *KGElo* have already been discussed. The volumetric bias ($Q*$) is also a standard hydrological performance index and it is useful to assess the ability of a model to reproduce the water balance (Yilmaz et al., 2008). Whereas $Q*$ indicates differences in the mean or central tendency between observed and simulated timeseries, $sd*$ indicates differences in their variability. Note that $Q*$ and $sd*$, albeit in their slightly different form of ratios instead of biases, are, together with r, components of KGE (Gupta et al., 2009).

Biases in the baseflow volume ($Qbase*$) and in the baseflow index ($BFI*$) tell how well a model simulates the delayed routing of flow and the speed of the hydrological response of a catchment respectively. Baseflow is the delayed portion of the hydrograph, associated with groundwater and other lagged sources of flow (Tallaksen & Van Lanen, 2004). Daily baseflow was obtained from the simulated and the observed hydrographs through the algorithm described by Tallaksen and Van Lanen (2004), using minimal flows of non-overlapping periods of 7 days. The baseflow index is the ratio of baseflow to flow and is an indicator of the hydrological response of the catchment: the smaller the index, the flashier the catchment (Tallaksen & Van Lanen, 2004).

The three metrics calculated from the flow-duration curve (i.e. $Qhi*$, $Qlo*$ and $FDCslp*$) are suggested by Yilmaz et al. (2008). The flow-duration curve (FDC) is also an indicator of the hydrological regime of a catchment (Westra et al., 2014). It has strong diagnostic power associated with dynamics of water storage and release within a catchment (Westra et al., 2014; McMillan, 2020). Here, we use the volumetric biases in the high-flow (exceedance $<$ 0.02) and low-flow (exceedance $>$ 0.7) portions of the FDC to assess the ability of models to reproduce the height of the peaks in the hydrograph and the volume in the low-flow periods respectively. The bias in the slope of the midsection ($0.2 <$ exceedance < 0.7) is a measure of the way a model reproduces the variability of the midrange flows and hence the speed of the transition from low- to high-flow conditions.

Finally, performance metrics in the *zero* group are included to evaluate the ability of models to reproduce cease-to-flow conditions. Low-flows, ephemerality and cease-to-flow conditions are intrinsic to Australia's hydrology (McMahon & Finlayson, 2003); nevertheless, models are especially deficient in their ability to reproduce such conditions (e.g. Ye et al., 1997). Metrics in this group are only calculated in 56 out of the original 155 catchments where the percentage of observed zero-flows in each of the three evaluation periods is at least 1 %. With regards to model simulations, daily flows below 5×10^{-4} mm/day are treated as zeros to match the precision of the observed streamflow data. *pc0** is an indicator of how models simulate the overall number of zero-flows in a given period; whereas *TPR0* represent the percentage of observed zeros actually modelled as such.

2.5 Data analysis

With 155 catchments, 5 models, 3 evaluation periods and 13 performance metrics, we find ourselves with upwards of 30 000 performance values to interpret. The following two sections describe the statistical methods used to analyse these data and achieve the two objectives stated in the introduction. In the next section, we describe the use of matched-pairs rank-biserial correlation coefficients to estimate changes in model performance in a consistent and comparable way, allowing us to identify which aspects of the flow regime are harder for models to reproduce during and after the drought (i.e. which metrics degrade most from their pre-MD values). In section 2.5.2, we describe the use of linear regressions to identify changes in the relationship between annual model performance and annual rainfall anomaly. We use an indicator variable to allow the linear models to shift their intercept at the onset and the end of the drought and use $t-$tests to evaluate whether the shift is significant.

2.5.1 Comparison of model performance across metrics and periods

Matched-pairs rank-biserial correlation is used to compare how model performance during and after the Millennium drought changes from the pre-MD evaluation period across the set of performance metrics. Matched-pairs rank-biserial correlation is a measurement of effect size for Wilcoxon's signed-ranks test (Wilcoxon, 1945) of statistical differences between two dependent samples (King & Minium, 2003). In the context of this research, the dependent samples in question are the levels of model performance in each catchment during each of the three evaluation periods: before, during and after the drought.

For each model, period of interest $\tau \in$ {MD, post-MD}, and performance metric E, the matched-pairs rank-biserial correlation coefficient r_c across all catchments was calculated following the four-step procedure below (King & Minium, 2003; Kerby, 2014). Except for the last step, this is identical to the calculation of Wilcoxon's test statistics.

1. For each catchment i, obtain the difference in performance between τ and pre-MD as $E_{\tau,i} - E_{\text{pre-MD},i}$.
2. Rank the absolute values of the differences from smallest to largest, and compute signed ranks by multiplying the signs of the differences to the ranks. Catchments where the difference in performance is zero are removed and the ranks of ties are averaged.
3. Sum the absolute values of the positive and negative ranks.
4. Calculate r_c as

$$r_c = \frac{R_+}{S} - \frac{R_-}{S}, \quad S = \frac{1}{2}n(n+1) \tag{2}$$

where R_+ and R_- are the sums of the ranks of the positive and negative differences respectively, calculated in step 3; and S is the total sum of ranks, which is computed from n, the number of catchments in the sample reduced by the number of catchments where the change in performance was zero.

Confidence intervals around r_c were calculated using the quantile method on 999 bootstraps. r_c is considered significantly different from zero, indicating that model performance did significantly shift from the pre-MD benchmark, if its two-sided 95 % confidence interval did not cross the zero.

Like other correlation metrics, the range of r_c is between -1 and 1. Interpretation of r_c is also similar to that of other correlation coefficients. A value of $r_c = 1$ (-1) indicates that all the differences $E_{\tau,i} - E_{\text{pre-MD},i}$ are positive (negative) and hence that for the given model the value of E is higher (lower) during τ than during the benchmarking period in all catchments. A value of $r_c = 0$ indicates that the ranks of the positive and negative differences in model performance between τ and pre-MD balance out over all the catchments.

The use of ranked differences allows comparison of changes in model performance across different performance metrics regardless of their range or sensitivity. This is a necessary requirement for this study, given that the set of metrics laid out in Table 2 have a variety of ranges and even the ones that share the same endpoints and optimal values are not 1-to-1 comparable. However, in order for the comparison to be meaningful it requires that the sign of the differences of all metrics have the same meaning (i.e. a positive difference is an improvement and a negative difference is a deterioration of performance). To comply with this requirement, all of the performance metrics based on bias are transformed by taking the opposite of their absolute values.

The use of ranked differences, while removing the need of distributional assumptions and allowing for comparison between metrics on different scales, carries the assumption that differences of metric values can be meaningfully ranked (King & Minium, 2003). Whether this assumption is fulfilled or not is somewhat subjective and dependent on the scale of the metric (e.g. Knoben, Freer, & Woods, 2019): is a drop in KGE from 1 to 0.5 comparable to a drop from -100 to -100.5? Should they be ranked in the same way, as the procedure to calculate r_c would? Most people familiar with the use of KGE to evaluate model performance would probably say that the former is a worse drop in performance than the latter, but they would also likely fail to quantify by how much: what is a drop in KGE from 1 to 0.5 comparable to when the starting point is as low as -100? For the purpose of this study, we have tested the influence of this assumption and concluded that it is unlikely to have significant impact on the results. Details are given in the supporting information text S2.

2.5.2 Comparison of annual model performance

The second aim of this study, as stated in the introduction, is to estimate how annual performance of models during the drought compares to their performance in pre-MD years of comparable wetness. Linear regressions of (transformed) annual performance metric as a function of annual rainfall anomaly are used, similarly to the procedure used by Saft et al. (2015) to identify significant changes in rainfall-runoff relationships on the same set of catchments.

For each catchment, (hydrologic) model, performance metric E and period $\tau \in \{\text{MD, post-MD}\}$, the model used for the regression is

$$BC(\tilde{E}) = \beta_1 \cdot P_a + \beta_2 \cdot I + \beta_0 + \varepsilon. \tag{3}$$

Where $BC(\tilde{E})$ is a Box-Cox transformation (Box & Cox, 1964) of the annual values of the performance metric; P_a is the annual rainfall anomaly, relative to the average pre-MD annual rainfall; and I is an indicator variable set to 0 for the years in pre-MD and 1 for the years in τ. Since the Box-Cox transformation requires strictly positive data, the annual performance was further transformed as $\tilde{E} = |E^* - E|$, where E^* represents the perfect score for each metric (i.e. 0 for the biases and 1 for all other metrics).

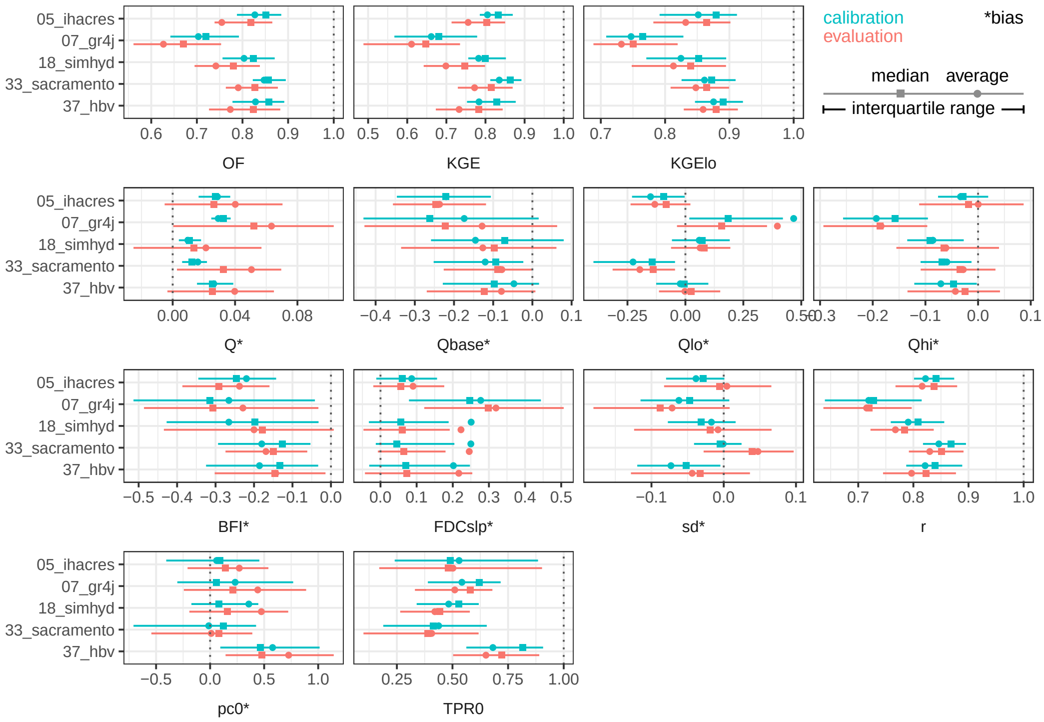

Figure 2. Model performance in the pre-MD period: comparison of all metrics between calibration (odd years) and evaluation (even years). Showing interquartile range, median and mean of performance across catchments. See Table 2 for the meaning of the performance metrics.

$\tilde{E}$ is therefore the distance from the perfect score and an increase in $\tilde{E}$ (and equally in $BC(\tilde{E})$) represents a decrease in performance.

Parameter β_2, associated with the indicator variable marking the period of interest from the benchmark, represents a shift in the intercept. We tested for the significance of this shift using a t-test ($\alpha = 0.05$) against the null-hypothesis that $\beta_2 = 0$. The outcome of the t-test was corrected with the false discovery rate approach (Benjamini & Hochberg, 1995) to control for the multiplicity of tests performed. Appropriate tests to check for normality and (lack of) autocorrelation were conducted on the residuals of the linear regressions (Haan, 2002).

3 Results

3.1 Model performance before the MD

All models perform very similarly during calibration, except for GR4J which has a lower calibration performance across most metrics. Models' average (median) calibration efficiency range from SimHyd's 0.80 (0.82) to Sacramento's 0.85 (0.86, same as HBV), with the exception of GR4J, which on average only reaches a value of the objective function of 0.70 (median = 0.72). As shown in Figure 2, the same pattern can be seen across the range of performance metrics, with the exception of the ones in the *zero* group, where GR4J's performance in the calibration period is in line with the other models. The difference in calibration performance between GR4J and the other models is most marked in the peak flow bias ($Qhi*$), the FDC slope bias ($FDCslp*$) and the correlation coefficient (r), this seems to indicate that GR4J performs worse than the other models in its ability to reproduce high flows. The same can be concluded by noticing that the difference between GR4J and the other models is larger for the standard KGE than for its transformed version. Reasons for the differences between GR4J and the other models are discussed in section 4.2.

Performance degradation from pre-MD calibration to pre-MD evaluation is limited and mostly occurs in volumetric bias and summary metrics, prioritising high flow metrics (i.e. OF and KGE). Although increments of change are not directly comparable across metrics, it is true that the changes were minor relative to the spread observed across all catchments, for most metrics under consideration. Figure 2 displays this in terms of aggregate (across catchments) model performance and is a confirmation that models are able to reproduce a range of aspects of the flow regime of an unseen hydrograph, given no significant changes in the underlying climate. The biggest changes to model performance from calibration to evaluation, relative to the spread of the data, occur in the summary performance metrics (the ones in the *fit* group, which are in the top row of Fig. 2) and in the volumetric bias ($Q*$). By this indicator, median KGE values decreased between 3.65 % (IHACRES) and 7.33 % (SimHyd), slightly less than the decrease in objective function median values (5.26 % to 9.59 %, Sacramento and SimHyd, respectively). Comparatively, median values of the transformed KGE decreased the least: only by between 1.38 % (GR4J) and 3.67 % (IHACRES) of the spread of $KGElo$ values in calibration. In terms of volumetric bias, median values did not actually increase extensively (up to 2.0 % for Sacramento and GR4J, and nearly zero for all other models), but the size of the interquartile range increased by at least 2.9 (HBV) and up to 8.3 (GR4J) times. This is due to the bias penalisation in the objective function. Inasmuch as the distributions shown in Figure 2 come from dependent samples, the same method and metric described in section 2.5.1 can be used to assess changes in performance from calibration to evaluation while taking in consideration changes in individual catchments. Values of r_c for this comparison are shown in Figures S1 and S2. They indicate that while there is some diversity between models (see for example the changes in the bias of the standard deviation, *sd**), the dataset-level conclusions above stand.

3.2 Effects of MD on performance

The matched-pairs rank-biserial correlation coefficients for each model and performance metric are shown in Figure 3. For each hydrologic model, the performance metrics are ordered from lowest to highest r_c during the MD period (round markers). Note, the bars here relate to the uncertainty in the chosen metric of rank-biserial correlation, which is different to the previous plot where the bars related to the range of values across the set of catchments. Performance metrics with the lowest (highest) r_c are the ones that degraded (improved) from the benchmark in the highest number of catchments. r_c values calculated across all models are shown in Figure S3. When looking at the order and extent of degradation from the benchmark of these metrics from Figures 3 and S3, it should be kept in mind that a lot of these metrics are not independent, especially the ones in the *fit* group as well as *Q**, *sd** and *r*, which make up the KGE and hence objective function, can be highly correlated. Correlation matrices for all metrics across all evaluation periods and models are shown in Figure S4.

For all the five models, overall model performance, as quantified by the summary performance metrics in the *fit* group, degrades during the drought in almost all catchments. r_c values for this group of metrics are always lower than -0.856 (IHACRES, *KGE*) for the comparison of MD performance to pre-MD evaluation performance. In terms of number of catchments, this results from models performing worse than the benchmark in between 129 and 151 catchments (or 83.2 % to 97.4 % of 155) depending on the metric and the model. On average models performed worse than they did in the benchmark period in 146 (94.3 %), 148 (95.5 %), and 137 (88.5 %) catchments for *OF*, *KGE* and *KGElo* respectively. With the exception of GR4J, model performance as measured by the trans-

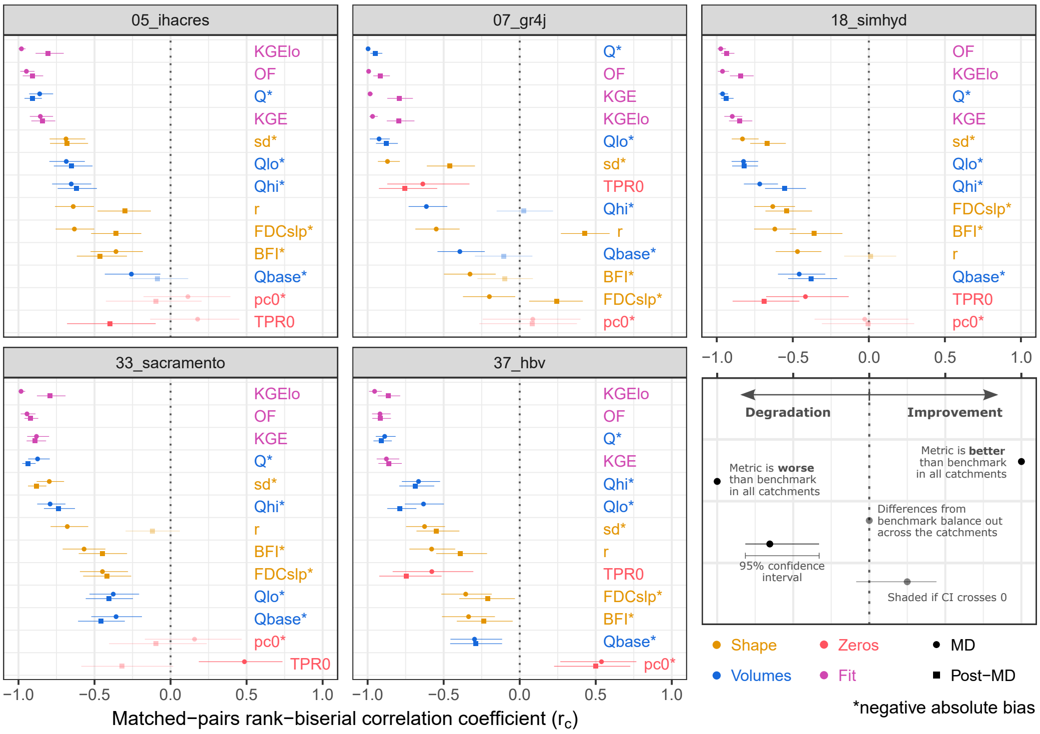

Figure 3. Changes in individual models performance from pre-MD evaluation (benchmark) to each of MD and post-MD. $r_c = -1$ ($+1$) indicate that the model performance according to that metric degrades (improves) from the benchmark in all catchments. Ranges indicate 95 % confidence intervals, points are faded when the CI crosses the zero. For each model, metrics are ordered from lowest to highest r_c for the MD period (round markers).

formed KGE (which gives greater weight to low flows) always degrades in more catchments than the performance measured in terms of untranformed KGE, resulting in lower r_c values.

Degradation of overall performance (as described above) is driven in large part by overestimation of the water balance. Amongst the other performance metrics, the only one whose r_c is consistently as low as the r_c of the *fit* metrics discussed above is the volumetric bias ($Q*$). Values of r_c for $Q*$ are always below -0.861 (IHACRES, MD) for all models and both periods of interest. This number is based on the negative absolute value of the bias and therefore only takes into consideration its distance from 0, in either direction. In reality the degradation of model performance in terms of water balance estimation is overwhelmingly driven by overestimation of streamflow: the average volumetric bias across all models and catchments during the benchmark period was 4.30 %, and it was positive (i.e. streamflow overestimated) in 110 catchments on average; during the drought, the average bias jumps to 67.8 % and the average number of catchments with overestimated streamflow become 130; even after the end of the drought, the average bias remains at 42.1 % (with 138 catchments with bias > 0, on average).

Compared to the volumetric bias, metrics representing the ability of models to reproduce hydrograph shape are less affected by the drought. The other two components of the KGE other than the bias are said to be indicators of the ability of a model to reproduce the shape of the hydrograph in terms of spread of flows ($sd*$) and hydrograph timing (r) (Gupta et al., 2009; Yilmaz et al., 2008). The r_c values for these two metrics are always higher than those of $Q*$, indicating that their performance degrades less consistently. Nevertheless, overall the bias in the standard deviation degrades in 115 to 134 catchments (or 74.2 % to 86.5 %) in the MD compared to the benchmark. After the drought, the number of catchments with $sd*$ worse than before the drought remains 98 to 133, depending on the model. In the pre-MD benchmark, the average value of $sd*$ was -1.42 % (i.e. slight underestimation), during (after) the drought the average becomes 48.7 % (55.9 %), with overestimation of the standard deviation of the flow occurring on average in 115 (121) catchments. The extent of degradation of the linear correlation coefficient between observed and simulated flows is smaller, with 98 to 116 catchments having worse r during the drought than in the benchmark period. Additionally, r is the only metric to recover after the drought based on its value of r_c. On individual models, r after the drought is found to be equivalent or better than during the benchmark period in 3 out of 5 models, and significantly less degraded than during the MD (non-overlapping 95 % confidence intervals) in 4 out of 5 models.

Overestimation of the water balance during and after the drought affects both high and low flows, driving down model performance. With respect to biases in the high- and low-flow portions of the flow duration curve (*Qhi** and *Qlo**), model performance degradation both during and after the drought is, just like in the case of the overall bias, driven by overestimation of flow amounts. This is most evident when looking at the volumes of the peak flow: in most catchments, models mildly underestimate it in the pre-MD benchmark period (-6.1 % to 0.0 %, on average for most models, -18.4 % for GR4J), but overestimate it during and after the drought (17.1 % to 89.8 % and 10.9 % to 26.6 %, on average respectively). In terms of absolute values (i.e. distance from the objective, 0), this overestimation causes a degradation in performance in at least two third of the catchments during the MD for all models (102 to 124), resulting in values of r_c between -0.614 (GR4J) and -0.792 (Sacramento). After the drought, r_c and extent of performance degradation in terms of *Qhi** are very similar for each model to their values during the MD; with the exception of GR4J. GR4J underestimates peak volumes before the drought in the majority of catchments (135 or 87.1 %). Therefore, the increase in the volumes estimated after the drought results in improved performance in most catchments, bringing GR4J's r_c for this metric in the post-MD period to be slightly positive and not statistically different from zero. The performance degradation in terms of volume estimates of the low-flow portion of the FDC is driven by the same mechanisms. Here the initial values of pre-MD bias are more varied from model to model (-19.7 % to 39.7 %) and the increase in percentage overestimation are much higher: on average higher than 130 % for each model and period with the exception of IHACRES, MD. However, the resulting values of r_c are similar to those for the peak flows.

The models' ability to reproduce the FDC shape is more resilient to the drought than their ability to reproduce volumes. The bias in the slope of the FDC's mid-section (*FDCslp**) degrades from pre-MD to MD (post-MD) in 90 to 114 (64 to 108) catchments, depending on the model. This results in values of r_c higher and closer to the zero than for *Qhi** and *Qlo**, indicating that this indicator tends to degrade less during and after the drought. Additionally, while with *Qhi** and *Qlo** there exists a clear increase in overestimation during the drought, the signal for *FDCslp** is less strong and while on average most models do overestimate the slope of the FDC during each of the three evaluation periods (simulating catchments with a flashier behaviour than in reality), the change in bias of FDC slope from pre-MD to MD is an increase in overestimation in only 48.6 % of catchment-model pairs; this value reduces to 33.5 % after the drought. Similarly to the bias in the slope of the FDC, the bias in the volume of baseflow and in the baseflow index (*Qbase** and *BFI**), indicators of a model's ability to reproduce catchments' flow regimes, were always amongst the least affected metrics during and after the drought in terms of r_c.

Finally, the two metrics of the *zeros* groups are consistently the least degraded during the drought, especially with regards to the estimation of the number of cease-to-flow days (*pc0**). This value is on average overestimated before the drought in all models, with average pre-MD values of *pc0** ranging from Sacramento's 1.0 % to HBV's 72.6 %. *pc0** is on average underestimated both during and after the drought (−1.5 % to −50.4 %, IHACRES, MD and SimHyd post-MD, respectively), as the number of zero-flow days increases. This results in an improvement in the estimation of the number of zero-flow days from pre-MD to MD (post-MD) in 21 to 39 (21 to 37) of the 56 catchments across which these metrics are calculated which causes r_c for this metric to never be significantly below the zero. With respect to *TPR0*, the percentage of zero-flow days actually modelled as such, r_c is significantly negative for three out of five models in the MD and for all models in the post-MD and it is the only metric consistently showing higher degradation after the drought compared to during the drought. Nevertheless, models' performance and performance changes according to this metric vary quite extensively and it is hard to establish generalisable patterns.

3.3 Annual model performance

Here we investigate model performance on interannual scale to separate the impact of multi-annual dry periods from impacts due to isolated dry years. For this, we fit the linear model in equation 3 to each combination of catchment, model, performance metric and period of interest: resulting in a total of 20 150 regressions. We use the fit to evaluate whether the relationship between model performance and annual rainfall anomaly changed significantly during each period of interest from the pre-drought evaluation benchmark. Figure 4 shows the percentage of catchments in each class of statistical significance for this change. In Figure 4 and in the next paragraphs, we present results only for some selected metrics from Table 2, namely the KGE, the volumetric bias and the biases of

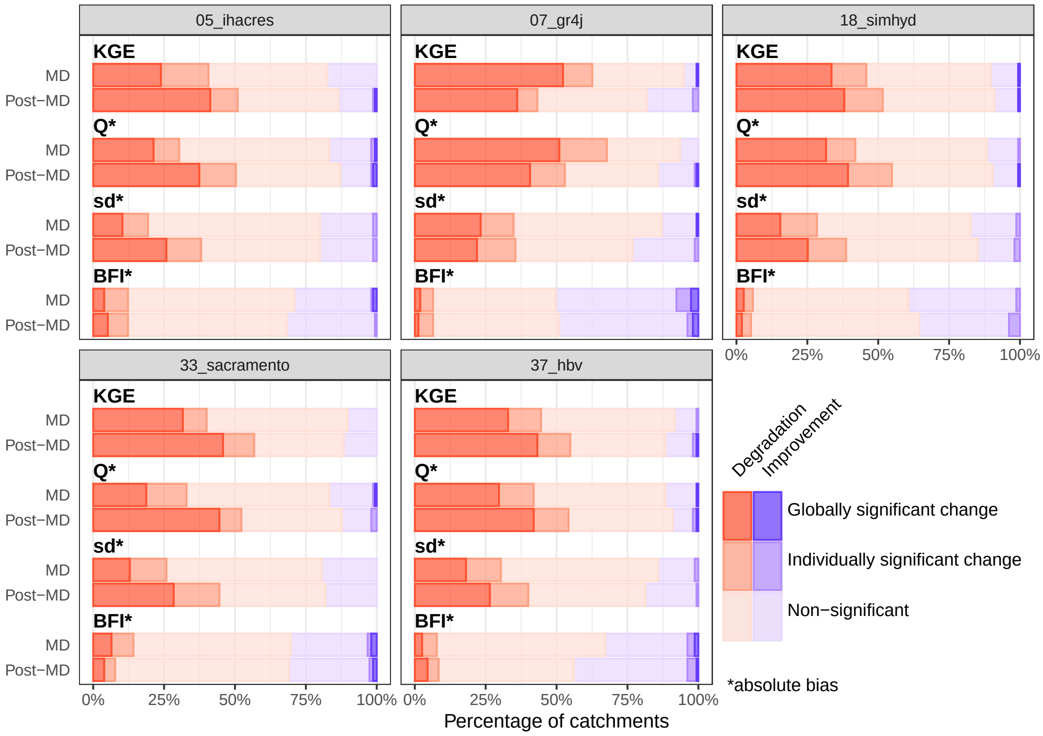

Figure 4. Changes in the relationship between annual model performance and annual rainfall anomaly, showing percentages of catchments in each class of statistical significance. Statistical significance is assessed with a t-test on the least-squares fitting of the period-specific intercept β_2 of the linear regression model in eq. 3.

standard deviations and the baseflow index. Results from the remainder of the metrics can be seen in figure S5. In the catchments represented in the red bars, the least-squares fitting on the linear model in eq. 3 resulted in $\hat{\beta}_2 > 0$, indicating that the model performance in the years of the drought or post drought ($I = 1$) is worse (i.e. further from the objective, in absolute value) than in the pre-MD evaluation years with a comparable rainfall anomaly. Conversely, regression models in the blue bars are where the fitting resulted in $\hat{\beta}_2 < 0$. Finally, the shading indicates the level of statistical significance of the value of $\hat{\beta}_2$ against the null-hypothesis that $\beta_2 = 0$.

During the drought, the change of KGE-to-anomaly relationship is individually significant and negative in between 40.0 % (Sacramento) and 45.8 % (SimHyd) of catchments for most model, with GR4J alone surpassing this and reaching 62.2 %. After the drought, these percentages increase, with nearly all models surpassing the threshold of half of the catchments with significantly degraded annual performance for a given P anomaly. GR4J is again exceptional, whereas it is the only model whose performance significantly degrades after the drought (again, for a given P anomaly) in less catchments than during the drought, bringing its behaviour in line to that of the other models in the post-MD period. Conversely, the number of catchments where the relationship changes significantly for the better (i.e. annual KGE is higher during or after the drought to expected from pre-MD years of similar P anomaly) is never above 3 (1.9 %). This results in the fact that, even if during the drought the results of this analysis actually show a non-significant change in the majority of catchments for most models, amongst the catchments where the shift is significant, it is overwhelmingly towards a degradation: at least 98.6 % of catchments with significant shifts during the drought and at least 95.7 % after the drought.

Similarly to what has been observed regarding overall performance degradation, degradation in the relationship between model performance and rainfall anomaly is driven in large part by errors in water balance estimation rather than hydrograph shape. The points made in the previous paragraph refer to model performance in terms of KGE, but the percentages and patterns described apply almost identically to the bias ($Q*$) as well. The relationship between bias and rainfall anomaly shifts significantly and negatively in 30.3 % to 41.8 % of catchments for most model, with again GR4J being the outlier with 67.7 %. Similarly as with the KGE, these percentages increase to at least 50.0 % after the drought for all models and decrease for GR4J. Amongst the catchments where the change in Q-to-anomaly relationship is significant, again the change is overwhelmingly towards a degradation: always at least 94.0 % of these catchments.

Whereas some of the patterns described above for KGE and $Q*$ (namely the relationships between GR4J and the other models, and relationship between MD and post-MD) are also similar for the bias of the standard deviations ($sd*$), the actual number of catchments where the change in perfomance-to-anomaly relationship is significant is lower (roughly halved in terms of global significance) than when performance is calculated in terms of KGE. Finally, with respect to the ability of models to estimate the baseflow index, the results show that in the greatest majority of catchments this was the same during and after the drought than it was in pre-MD years of similar rainfall anomaly. Here the results of the change analysis are non-significant in at least 128 (82.6 %, Sacramento, MD) and up to 144 (92.9 %, SimHyd, MD) catchments. Nevertheless, comparing the number of catchments within the same level of significance, we see again that the catchments where a degradation in performance occurs usually outnumber, albeit sometimes marginally, those where the performance is improved.

4 Discussion

In the introduction to this research, we set out to identify aspects of the flow regime and the hydrograph which are more or less problematic for models to reproduce when parameters calibrated on long-term average conditions are used to force a model using data from a period of drought. Additionally, we were interested in isolating the effects of the multi-annual drought from that of the drier conditions in individual years. Our results show extensive performance degradation during the years of the drought across catchments and models driven by overestimation of flow volumes. Replication of the shapes of the hydrograph and the flow duration curve is much more resilient to the drier climate. The analysis of performance in individual years and its relationship with annual rainfall anomaly shows that performance degradation cannot alone be attributed to drier conditions in individual years. In the metrics where most of the performance degradation occurred (i.e. summary performance metrics and volumetric biases), this is exacerbated by accumulation and aggravation of errors over the several subsequent dry years.

4.1 Relationship with existing literature

We show that degradation of model performance during the Millennium drought is largely driven by overestimation of flow volumes. This finding is in line with findings from previous studies on model performance during the Millennium drought (e.g. Saft et al., 2016) but the analysis here is considerably more in depth. Many of the catchments in this dataset experienced significant changes in their annual rainfall-runoff relationship (Saft et al., 2015, in preparation), these are essentially changes in water-balance and water partitioning and therefore intrinsically linked to streamflow volume. The overestimation of flow volumes and degradation of model performance shown here seems to be more widespread than the 50 % to 70 % of catchments shifted according to Saft et al. (2015) and Peterson et al. (2021). However, the numbers in those studies refer only to catchments where the shift in hydrologic response was found to be statistically significant, whereas here statistical significance is evaluated across all catchments. Systematic overestimation of streamflow indicates that models' mechanisms to delay flow and remove water from the system before it reaches the stream are not able to reproduce the decrease in streamflow observed during the drought. Previous research also showed that many models (including 4 of the 5 tested here) fail to realistically reproduce multi-annual declines in water stored during the drought (Fowler et al., 2020).

Failure to reproduce the long, slow dynamics described by Fowler et al. (2020) is also evident in the results of the annual performance analysis. The results presented point to the multi-year nature of the drought as a driver of the degradation of model performance and especially of the overestimation of flow volumes, caused by accumulation and aggravation of model errors as the dry spell persists over multiple years. This is supported by studies indicating the length and persistence of the Millennium drought as one cause of its disproportionate effects on hydrological systems (e.g. Murphy & Timbal, 2008; Potter et al., 2010) and by the observation that models are unsuited to reproduce multiyear drying conditions as they often deplete their entire storage variability within a single 1-year cycle (Fowler et al., 2020). However, the ability of models to reproduce the baseflow index during drought years is almost never different to their ability to estimate it during pre-drought years with a similar rainfall anomaly. This signals that flows generated via fast and slow mechanisms are similarly affected by drought, and models struggle to reproduce them both in a similar way. Their ratio, i.e. the baseflow index, is therefore less altered by drought and not affected by the same carry-over effect from year to year, which allows model to reproduce it better even after several subsequent dry years.

4.2 Exceptionalism of GR4J

There are some relevant differences in the ways models in this study behave. GR4J, in particular, was often flagged as an outlier. Contrary to previous studies (e.g. Saft et al., 2016; Fowler et al., 2016), GR4J's calibration performance (and performance before the drought, in general) is here lower than the performance of all the other models. Note that such studies used NSE (Saft et al., 2016) and KGE (Fowler et al., 2016) for calibration; the use of a different objective function here makes it impossible to compare performance across studies and only allows comparison across models within individual studies. Fowler et al. (2016) showed that GR4J calibrated similarly to other models within a single objective; however, it struggled more than the other models in finding good parameter sets to compromise between conflicting objectives (Fowler et al., 2016). This may play a role in reducing GR4J's calibration performance here, given that the objective function for this study requires models to consider high and low flows as well as bias. Additionally, it is possible that the difference in calibration performance seen here between GR4J and the other models has more to do with the latter performing better than they would *normally* do, rather than GR4J underperforming. This may be due to the use in this study of the MARRMoT implementation of each of these models. Amongst its design considerations, MARRMoT uses logistic smoothing of storage thresholds and a numerically stable timestepping scheme to reduce discontinuities in the response function and improve the calibration performance (Knoben, Freer, Fowler, et al., 2019). Compared to the other models, GR4J is less likely to benefit from such implementation, given that smoothing mechanisms are built into its constituting equations (Perrin et al., 2003).

The smaller flexibility of GR4J seen by Fowler et al. (2016) is also shown in the way it degrades more than the other models at the onset of the drought. Differently from the other models, GR4J contains a mechanism to regulate fluxes of water leaving (or entering) the system via a *groundwater exchange*. Albeit unrealistic within the Australian context, such a mechanism improves the performance of GR4J (Hughes et al., 2015) by *de facto* compensating for actual ET fluxes, which are dominant in these catchments (Fowler et al., 2021). However, GR4J's groundwater exchange is regulated by its parameter x_2, fixed throughout the simulation from its pre-MD calibration value, giving GR4J little flexibility to adapt this important water balance mechanism to a shifted hydrologic regime mid-simulation. This also makes GR4J more susceptible to errors due to accumulation of moisture deficits over multiple annual periods (Fig. 4). After the end of the drought, however, this mechanism might be what makes it easier for GR4J to recover some of the performance lost during the drought, compared to other models.

4.3 Post-drought recovery

According to most performance metrics, model performance does not recover after the end of the drought. Peterson et al. (2021) showed that a lot of the catchments where a hydrological shift occurred during the drought have not recovered to their pre-drought behaviour even years after the end of the dry spell. If the drop in performance is attributable at least in part to this changed hydrological behaviour, it is expected for the performance not to recover as long as rainfall-runoff relationships remain altered. Additionally, given that rainfall anomalies are by definition closer to their long-term average in this period, this also results in less of the models' performance degradation after the drought that is explainable alone by the climate anomaly, and hence the negative effect on the relationship between performance and anomaly in more catchments than during the drought (Fig. 4).

The fact that the correlation coefficient between observed and simulated streamflow is the only metric that consistently returned to pre-MD values after the drought is likely an indication that the dependency of streamflow on precipitation (and hence the ease with which models simulate streamflow timing from rainfall inputs) degrades during the drought and restores after the drought is finished, possibly thanks to restored near-surface soil moisture patterns. Additionally, it must be noted that amongst the many low (and zero) flows of the drought period, the correlation coefficient can be severely affected by the ability of models to simulate the timing of spells of above-average flow. After the end of the drought, with a more regular flow regime in many catchment, the correlation is likely to be less affected by individual high-flow outliers (Kim et al., 2015).

4.4 Limitations and further studies

Values of the matched-pairs rank-biserial correlation coefficients presented in the result section come from averaging model performance changes across the diversity of the catchments in the study. This makes non-extreme values of r_c hard to interpret, but it is the necessary cost of prioritising comparability of performance degradation across metrics. For example, consider the apparent resilience of the models to the drought according to the *zeros* metrics. Given the high diversity of performance for all models in this respect during calibration and the benchmark (Fig. 2), the fact that r_c often returns non-significant values does not actually entail that all models perform equally to the benchmark, but it's more likely a reflection of the volatility of model performance with respect to cease-to-flow conditions and may be the result of averaging model behaviour across catchments where they perform (and where their performance changes) very differently.

Another important limitation of such a large-sample approach is that it complicates general interpretation of the results in terms of model diagnostic and remedial actions. Whereas large-sample studies have immense value in the development of hydrological theories and models (Addor et al., 2019), model performance can be very catchment-specific and within a large set of catchments, it's rare for a single model to outperform all others across the landscape (e.g. Knoben et al., 2020). In this context, it is likely that the focus on aggregate results of this study obscures opportunities for remedial action and model improvement within specific (sets of) catchments. Nonetheless, our results uphold the call for model architectures to include longer memory components to keep track of moisture deficits across multiple annual cycles (Fowler et al., 2020) as well as more realistic representations of moisture removal mechanisms able to adapt to changing catchment conditions.

In this analysis, we present an easily generalisable methodology to assess and evaluate changes in model performance across periods and landscapes. We hope with this study to inspire further research in this space to expand our findings to additional models and regions affected by changing climates. Additionally, application to an even wider set of metrics, including metrics derived from hydrological signatures with specific links to catchment processes (see McMillan, 2020), would prove beneficial to estimate and diagnose models' realism in the face of changing hydrological behaviour. Within the scope of this study, we have already identified a shortcoming in the assessment of model performance in the face of cease-to-flow conditions. Given that there exists a relationship between ephemerality and drought-induced changes in catchment behaviour (see Saft et al., in preparation), we believe that ability of models to reproduce timing and extent of zero-flows during the drought should be further and better investigated with more appropriate and specifically designed metrics and indices.

5 Conclusions

In this study, we evaluated the effect of prolonged drought on hydrologic model performance. For this, we used 13 metrics of performance for five conceptual rainfall-runoff models, calibrated and run using data from 155 catchments in the Australian state of Victoria that experienced prolonged drought conditions. By using matched-pairs rank-biserial correlation to explore model performance changes across the performance metrics in a unified and comparable way, we observed extensive model degradation induced by the drought affecting all models tested. Particularly, we demonstrated that performance drops because of overestimation of flow volumes, whereas the ability of models to reproduce the shapes of the hydrograph and the flow duration curve is more resilient to the drought.

Additionally, we studied the relationship between annual model performance and rainfall anomaly and demonstrate that in many catchments, annual changes in catchment wetness during the drought cannot alone explain the degradation in model performance. This suggests that performance degradation is exacerbated by accumulation and aggravation of model errors as the rainfall anomaly persists over multiple years. In this context, we amplify calls from other researchers on the need to improve realism of model structures as a tool to improve applicability within climate change scenarios, especially with regards to multi-annual memory components.

Overall, the study presented testifies to the complexity of the challenges faced by hydrologists as they engage in simulation and analysis in nonstationary climate conditions. The extent of model performance degradation caused by ill-estimated volumes of streamflow is particularly concerning in the context of water availability studies for allocation and planning purposes. This is especially disquieting considering that models overestimate flow volumes, hence producing overly optimistic estimates of water availability during drought. In their current form and with common calibration methods, conceptual rainfall-runoff model simulations are not reliable for these objectives during extended drought.

Open Research

Model input data is described by Saft et al. (in preparation) and currently stored at `https://cloudstor.aarnet.edu.au/plus/s/A2M7Vqp6CU52SzU`. Model outputs and the rest of the data described in the supporting information text S3 is currently stored at `https://cloudstor.aarnet.edu.au/plus/s/GzcJ8ROItX9okdO`. The version of MARRMoT used for this study is described by Trotter et al. (in preparation) and currently stored at `https://github.com/ltrotter/MARRMoT`. These are temporary locations for the purpose of peer review, both datasets and the software package will be uploaded to an appropriate repository and shared via a DOI before acceptance and publication of this article.

Acknowledgments

This study received support from the Australian Research Council via Linkage Project LP180100796 *Observed streamflow generation changes: better understanding and modelling*, the Victorian Government Department of Environment, Land, Water and Planning, and Melbourne Water. KF also acknowledges support from LP170100598.

References

Addor, N., Do, H. X., Alvarez-Garreton, C., Coxon, G., Fowler, K. J. A., & Mendoza, P. A. (2019). Large-sample hydrology: recent progress, guidelines for new datasets and grand challenges. *Hydrological Sciences Journal*, 1–14. Retrieved from https://doi.org/10.1080/02626667.2019.1683182 doi: 10.1080/02626667.2019.1683182

Alvarez-Garreton, C., Pablo Boisier, J., Garreaud, R., Seibert, J., & Vis, M. (2021). Progressive water deficits during multiyear droughts in basins with long hydrological memory in Chile. *Hydrology and Earth System Sciences*, *25*(1), 429–446. doi: 10.5194/hess-25-429-2021

Arsenault, R., Poulin, A., Côté, P., & Brissette, F. (2014). Comparison of Stochastic Optimization Algorithms in Hydrological Model Calibration. *Journal of Hydrologic Engineering*, *19*(7), 1374–1384. Retrieved from http://ascelibrary.org/doi/10.1061/%28ASCE%29HE.1943-5584.0000938 doi: 10.1061/(ASCE)HE.1943-5584.0000938

Avanzi, F., Rungee, J., Maurer, T., Bales, R., Ma, Q., Glaser, S., & Conklin, M. (2019). Evapotranspiration feedbacks shift annual precipitation-runoff relationships during multi-year droughts in a Mediterranean mixed rain-snow climate. *Hydrology and Earth System Sciences Discussions*(August), 1–35. doi: 10.5194/hess-2019-377

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*. doi: 10.1111/j.2517-6161.1964.tb00553.x

Burnash, R. J. C. (1995). The NWS River Forecast System-catchment modeling. *Computer models of watershed hydrology*, 311–366.

Chiew, F. H. S., Peel, M. C., & Western, A. W. (2002). Application and testing of the simple rainfall runoff model Simhyd. *Mathematical Models of Small Watershed Hydrology and Applications*.

Chiew, F. H. S., Potter, N. J., Vaze, J., Petheram, C., Zhang, L., Teng, J., & Post,

D. A. (2014). Observed hydrologic non-stationarity in far south-eastern Australia: Implications for modelling and prediction. *Stochastic Environmental Research and Risk Assessment*, *28*(1), 3–15. doi: 10.1007/s00477-013-0755-5

Chiew, F. H. S., Stewardson, M. J., & McMahon, T. A. (1993). Comparison of six rainfall-runoff modelling approaches. *Journal of Hydrology*. doi: 10.1016/0022-1694(93)90073-I

Cook, B. I., Mankin, J. S., & Anchukaitis, K. J. (2018). Climate Change and Drought: From Past to Future. *Current Climate Change Reports*, *4*(2), 164–179. doi: 10.1007/s40641-018-0093-2

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

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

CSIRO. (2012). *Climate and water availability in south-eastern Australia: A synthesis of findings from Phase 2 of the South Eastern Australian Climate Initiative (SEACI)* (Tech. Rep.). Author.

Dai, A., & Zhao, T. (2017). Uncertainties in historical changes and future projections of drought. Part I: estimates of historical drought changes. *Climatic Change*, *144*(3), 519–533. Retrieved from `http://dx.doi.org/10.1007/s10584-016-1705-2` doi: 10.1007/s10584-016-1705-2

Deb, P., & Kiem, A. S. (2020). Evaluation of rainfall–runoff model performance under non-stationary hydroclimatic conditions. *Hydrological Sciences Journal*, *65*(10), 1667–1684. Retrieved from `https://doi.org/10.1080/02626667.2020.1754420` doi: 10.1080/02626667.2020.1754420

Douville, H., Raghavan, K., Renwick, J., Allan, R. P., Arias, P. A., Barlow, M., ... Zolina, O. (2021). Water Cycle Changes. In V. Masson-Delmotte et al. (Eds.), *Climate change 2021: The physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change* (p. 239). Cambridge University Press.

Feyen, L., & Dankers, R. (2009). Impact of global warming on streamflow drought in

Europe. *Journal of Geophysical Research Atmospheres*, *114*(17), 1–17. doi: 10.1029/2008JD011438

Fowler, K. J. A., Coxon, G., Freer, J., Peel, M. C., Wagener, T., Western, A. W., ... Zhang, L. (2018). Simulating Runoff Under Changing Climatic Conditions: A Framework for Model Improvement. *Water Resources Research*, *54*(12), 9812–9832. doi: 10.1029/2018WR023989

Fowler, K. J. A., Coxon, G., Freer, J. E., M. Knoben, W. J., Peel, M. C., Wagener, T., ... Zhang, L. (2021). Towards more realistic runoff projections by removing limits on simulated soil moisture deficit. *Journal of Hydrology*, 126505. doi: 10.1016/j.jhydrol.2021.126505

Fowler, K. J. A., Knoben, W. J. M., Peel, M. C., Peterson, T., Ryu, D., Saft, M., ... Western, A. W. (2020). Many commonly used rainfall-runoff models lack long, slow dynamics: implications for runoff projections. *Water Resources Research*, *56*(5). doi: 10.1029/2019wr025286

Fowler, K. J. A., Peel, M. C., Western, A., & Zhang, L. (2018). Improved Rainfall-Runoff Calibration for Drying Climate: Choice of Objective Function. *Water Resources Research*, *54*(5), 3392–3408. doi: 10.1029/2017WR022466

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. Retrieved from `http://doi.wiley.com/10.1002/2015WR018068` doi: 10.1002/2015WR018068

Gao, Z., Zhang, L., Zhang, X., Cheng, L., Potter, N., Cowan, T., & Cai, W. (2016). Long-term streamflow trends in the middle reaches of the Yellow River Basin: Detecting drivers of change. *Hydrological Processes*, *30*(9), 1315–1329. doi: 10.1002/hyp.10704

Garreaud, R. D., Boisier, J. P., Rondanelli, R., Montecinos, A., Sepúlveda, H. H., & Veloso-Aguila, D. (2020). The Central Chile Mega Drought (2010–2018): A climate dynamics perspective. *International Journal of Climatology*, *40*(1), 421–439. doi: 10.1002/joc.6219

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*(1-2), 80–91. doi:

10.1016/j.jhydrol.2009.08.003

Haan, C. T. (2002). *Statistical methods in hydrology* (2nd ed.). Ames, Iowa.: Iowa State Press. doi: 10.1201/9780429423116-36

Hansen, N., Müller, S. D., & Koumoutsakos, P. (2003). Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). *Evolutionary Computation*. doi: 10.1162/106365603321828970

Hansen, N., & Ostermeier, A. (1996). Adapting arbitrary normal mutation distributions in evolution strategies: the covariance matrix adaptation. In *Proceedings of the ieee conference on evolutionary computation.* doi: 10.1109/icec.1996.542381

Hewitson, B., Janetos, A. C., Carter, T. R., Giorgi, F., Jones, R. G., Kwon, W. T., … Van Aalst, M. K. (2014). Regional context. In *Climate change 2014: Impacts, adaptation and vulnerability: Part b: Regional aspects: Working group ii contribution to the fifth assessment report of the intergovernmental panel on climate change.* Cambridge University Press. doi: 10.1017/CBO9781107415386.001

Hughes, J. D., Potter, N. J., & Zhang, L. (2015). Is inter-basin groundwater exchange required in rainfall-runoff models: The Australian context. *Proceedings - 21st International Congress on Modelling and Simulation, MODSIM 2015*(December), 2423–2429. doi: 10.36334/modsim.2015.l14.hughes

Jakeman, A. J., Littlewood, I. G., & Whitehead, P. G. (1990). Computation of the instantaneous unit hydrograph and identifiable component flows with application to two small upland catchments. *Journal of Hydrology*. doi: 10.1016/0022-1694(90)90097-H

Jeffrey, S. J., Carter, J. O., Moodie, K. B., & Beswick, A. R. (2001). Using spatial interpolation to construct a comprehensive archive of Australian climate data. *Environmental Modelling and Software*, *16*(4), 309–330. doi: 10.1016/S1364-8152(01)00008-1

Jones, D. A., Wang, W., & Fawcett, R. (2009). High-quality spatial climate datasets for Australia. *Australian Meteorological and Oceanographic Journal*, *58*(4), 233–248. doi: 10.22499/2.5804.003

Kerby, D. S. (2014). The Simple Difference Formula: An Approach to Teaching Nonparametric Correlation. *Innovative Teaching*, *3*(1). doi: 10.2466/11.it.3.1

Kiem, A. S., & Verdon-Kidd, D. C. (2010). Hydrology and Earth System Sciences Towards understanding hydroclimatic change in Victoria, Australia-preliminary insights into the "Big Dry". *Hydrology and Earth System Sciences*, *14*, 433–445. Retrieved from `www.hydrol-earth-syst-sci.net/14/433/2010/`

Kim, Y., Kim, T. H., & Ergün, T. (2015). The instability of the Pearson correlation coefficient in the presence of coincidental outliers. *Finance Research Letters*, *13*, 243–257. Retrieved from `http://dx.doi.org/10.1016/j.frl.2014.12.005` doi: 10.1016/j.frl.2014.12.005

King, B. M., & Minium, E. W. (2003). *Statistical reasoning in psychology and education.* (4th ed. ed.). J. Wiley and Sons. Retrieved from `https://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=cat00006a&AN=melb.b2813620&site=eds-live&scope=site&custid=s2775460`

Knoben, W. J. M., Freer, J. E., Fowler, K. J. A., Peel, M. C., & Woods, R. A. (2019). Modular Assessment of Rainfall-Runoff Models Toolbox (MARRMoT) v1.2: An open-source, extendable framework providing implementations of 46 conceptual hydrologic models as continuous state-space formulations. *Geoscientific Model Development*, *12*(6), 2463–2480. doi: 10.5194/gmd-12-2463-2019

Knoben, W. J. M., Freer, J. E., Peel, M. C., Fowler, K. J. A., & Woods, R. A. (2020). A Brief Analysis of Conceptual Model Structure Uncertainty Using 36 Models and 559 Catchments. *Water Resources Research*, *56*(9), 1–23. doi: 10.1029/2019WR025975

Knoben, W. J. M., Freer, J. E., & Woods, R. A. (2019). Technical note: Inherent benchmark or not? Comparing Nash-Sutcliffe and Kling-Gupta efficiency scores. *Hydrology and Earth System Sciences*, *23*(10), 4323–4331. doi: 10.5194/hess-23-4323-2019

Li, C. Z., Zhang, L., Wang, H., Zhang, Y. Q., Yu, F. L., & Yan, D. H. (2012). The transferability of hydrological models under nonstationary climatic conditions. *Hydrology and Earth System Sciences*, *16*(4), 1239–1254. doi: 10.5194/hess-16-1239-2012

Lindström, G., Johansson, B., Persson, M., Gardelin, M., & Bergström, S. (1997).

Development and test of the distributed HBV-96 hydrological model. *Journal of Hydrology*. doi: 10.1016/S0022-1694(97)00041-3

Mann, M. E., & Gleick, P. H. (2015). Climate change and California drought in the 21st century. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(13), 3858–3859. doi: 10.1073/pnas.1503667112

Marengo, J. A., & Espinoza, J. C. (2016). Extreme seasonal droughts and floods in Amazonia: Causes, trends and impacts. *International Journal of Climatology*, *36*(3), 1033–1050. doi: 10.1002/joc.4420

McMahon, T. A., & Finlayson, B. L. (2003). Droughts and anti-droughts: The low flow hydrology of Australian rivers. *Freshwater Biology*, *48*(7), 1147–1160. doi: 10.1046/j.1365-2427.2003.01098.x

McMillan, H. (2020). Linking hydrologic signatures to hydrologic processes: A review. *Hydrological Processes*, *34*(6), 1393–1409. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1002/hyp.13632` doi: 10.1002/hyp.13632

Milly, P. C. D., Betancourt, J., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W., Lettenmaier, D. P., & Stouffer, R. J. (2008). Stationarity Is Dead: Whither Water Management? *New Series*, *319*(5863), 573–574. doi: 10.1126/science.ll51915

Mishra, A. K., & Singh, V. P. (2010). A review of drought concepts. *Journal of Hydrology*, *391*(1-2), 202–216. Retrieved from `http://dx.doi.org/10.1016/j.jhydrol.2010.07.012` doi: 10.1016/j.jhydrol.2010.07.012

Morton, F. I. (1983). Operational estimates of areal evapotranspiration and their significance to the science and practice of hydrology. *Journal of Hydrology*. doi: 10.1016/0022-1694(83)90177-4

Murphy, B. F., & Timbal, B. (2008). A review of recent climate variability and climate change in southeastern Australia. *International Journal of Climatology*, *28*(7), 859–879. Retrieved from `http://doi.wiley.com/10.1002/joc.1627` doi: 10.1002/joc.1627

Peel, M. C., & Blöschl, G. (2011). Hydrological modelling in a changing world. *Progress in Physical Geography*, *35*(2), 249–261. doi: 10.1177/0309133311402550

Peel, M. C., Finlayson, B. L., & McMahon, T. A. (2007). Updated world map of the

Köppen-Geiger climate classification. *Hydrology and Earth System Sciences*. doi: 10.5194/hess-11-1633-2007

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

Peterson, T. J., Saft, M., Peel, M. C., & John, A. (2021). Watersheds may not recover from drought. *Science*, *372*(6543), 745–749. Retrieved from https://www.sciencemag.org/lookup/doi/10.1126/science.abd5085 doi: 10.1126/science.abd5085

Potter, N. J., Chiew, F. H. S., & Frost, A. J. (2010). An assessment of the severity of recent reductions in rainfall and runoff in the Murray-Darling Basin. *Journal of Hydrology*, *381*(1-2), 52–64. doi: 10.1016/j.jhydrol.2009.11.025

Rowell, D. P., Booth, B. B., Nicholson, S. E., & Good, P. (2015). Reconciling past and future rainfall trends over East Africa. *Journal of Climate*, *28*(24), 9768–9788. doi: 10.1175/JCLI-D-15-0140.1

Saft, M., Peel, M. C., & Peterson, T. J. (in preparation). *Explaining hydrological shift and non-recovery after prolonged drought.*

Saft, M., Peel, M. C., Western, A. W., Perraud, J. M., & Zhang, L. (2016). Bias in streamflow projections due to climate-induced shifts in catchment response. *Geophysical Research Letters*, *43*(4), 1574–1581. doi: 10.1002/2015GL067326

Saft, M., Western, A. W., Zhang, L., Peel, M. C., & Potter, N. J. (2015). The influence of multiyear drought on the annual rainfall-runoff relationship: An Australian perspective. *Water Resources Research*, *51*(4), 2444–2463. doi: 10.1002/2014WR015348

Seibert, J. (2003). Reliability of Model Predictions Outside Calibration Conditions. *Nordic Hydrology*, *34*(5), 1–13. doi: 10.2166/nh.2003.0019

Seneviratne, S. I., Zhang, X., Adnan, M., Badi, W., Dereczynski, C., Luca, A. D., ... Zhou, B. (2021). Weather and Climate Extreme Events in a Changing Climate. In V. Masson-Delmotte et al. (Eds.), *Climate change 2021: The physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change* (p. 366). Cambridge University Press. Retrieved from https://www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC_AR6_WGI_Chapter_11.pdf

Sun, C., & Yang, S. (2012). Persistent severe drought in southern China during winter-spring 2011: Large-scale circulation patterns and possible impacting factors. *Journal of Geophysical Research Atmospheres*, *117*(10). doi: 10.1029/2012JD017500

Tallaksen, L. M., & Van Lanen, H. A. J. (2004). *Hydrological Drought: Processes and Estimation Methods for Streamflow and Groundwater* (Vol. 48). Amsterdam, London: Elsevier B.V.

Tian, W., Liu, X., Liu, C., & Bai, P. (2018). Investigation and simulations of changes in the relationship of precipitation-runoff in drought years. *Journal of Hydrology*, *565*(June), 95–105. Retrieved from `https://doi.org/10.1016/j.jhydrol.2018.08.015` doi: 10.1016/j.jhydrol.2018.08.015

Trotter, L., Knoben, W. J. M., Fowler, K. J. A., Saft, M., & Peel, M. C. (in preparation). *Modular Assessment of Rainfall–Runoff Models Toolbox (MARRMoT) v2.0: an object-oriented implementation of all your favourite hydrologic models for improved speed and readability.*

van Dijk, A. I. J. M., Beck, H. E., Crosbie, R. S., de Jeu, R. A. M., Liu, Y. Y., Podger, G. M., ... Viney, N. R. (2013). The Millennium Drought in southeast Australia (2001-2009): Natural and human causes and implications for water resources, ecosystems, economy, and society. *Water Resources Research*, *49*(2), 1040–1057. Retrieved from `http://doi.wiley.com/10.1002/wrcr.20123` doi: 10.1002/wrcr.20123

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. Retrieved from `http://dx.doi.org/10.1016/j.jhydrol.2010.09.018` doi: 10.1016/j.jhydrol.2010.09.018

Verdon-Kidd, D. C., & Kiem, A. S. (2009). Nature and causes of protracted droughts in southeast Australia: Comparison between the Federation, WWII, and Big Dry droughts. *Geophysical Research Letters*, *36*(22). doi: 10.1029/2009GL041067

Viney, N. R., Perraud, J., Vaze, J., Chiew, F. H. S., Post, D. A., & Yang, A. (2009). The usefulness of bias constraints in model calibration for regionalisation to ungauged catchments. In *Proceedings of the 18 th world imacs / mod-*

sim congress. Cairns, Australia. Retrieved from http://mssanz.org.au/modsim09

Westra, S., Thyer, M., Leonard, M., Kavetski, D., & Lambert, M. (2014). A strategy for diagnosing and interpreting hydrological model nonstationarity. *Water Resources Research*. doi: 10.1002/2013WR014719

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, *1*(6), 80–83.

Xu, C. Y. (1999). Climate change and hydrologic models: A review of existing gaps and recent research developments. *Water Resources Management*, *13*(5), 369–382. Retrieved from https://link.springer.com/article/10.1023/A:1008190900459 doi: 10.1023/A:1008190900459

Ye, W., Bates, B. C., Viney, N. R., & Sivapalan, M. (1997). Performance of conceptual rainfall-runoff models in low-yielding ephemeral catchments. *Water Resources Research*, *33*(1), 153–166.

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), 1–18. doi: 10.1029/2007WR006716

Zhang, Y., Feng, X., Wang, X., & Fu, B. (2018). Characterizing drought in terms of changes in the precipitation-runoff relationship: A case study of the Loess Plateau, China. *Hydrology and Earth System Sciences*, *22*(3), 1749–1766. doi: 10.5194/hess-22-1749-2018

Supporting Information for "Symptoms of performance degradation during multi-annual drought: a large-sample, multi-model study"

Luca Trotter[1], Margarita Saft[1], Murray C. Peel[1], Keirnan J. A. Fowler[1]

[1]Department of Infrastructure Engineering, University of Melbourne, Melbourne, Victoria, Australia

Contents of this file

Introduction

Text S1. In this section, we provide the formulas used to calculate the performance metrics in Table 1.

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \tag{S1}$$

$$KGElo = 1 - \sqrt{(r_{r5}-1)^2 + (\alpha_{r5}-1)^2 + (\beta_{r5}-1)^2} \tag{S2}$$

$$Q^* = \frac{\sum_{t=1}^{N} Qs_t}{\sum_{t=1}^{N} Qo_t} - 1 \tag{S3}$$

$$Qbase^* = \frac{\sum_{t=1}^{N} b(Qs)_t}{\sum_{t=1}^{N} b(Qo)_t} - 1 \tag{S4}$$

$$Qlo^* = \frac{\sum_{t \in L_s} Qs_t}{\sum_{t \in L_o} Qo_t} - 1 \tag{S5}$$

$$Qhi^* = \frac{\sum_{t \in H_s} Qs_t}{\sum_{t \in H_o} Qo_t} - 1 \tag{S6}$$

$$BFI^* = \frac{\sum_{y \in Y} \left(\frac{\sum_{t \in y} b(Qs)_t}{\sum_{t \in y} Qs_t} \right)}{\sum_{y \in Y} \left(\frac{\sum_{t \in y} b(Qo)_t}{\sum_{t \in y} Qo_t} \right)} - 1 \tag{S7}$$

$$FDCslp^* = \frac{\sum_{y \in Y} \left(\log(\{Qs_{t \in y}\}_{80}) - \log(\{Qs_{t \in y}\}_{30}) \right)}{\sum_{y \in Y} \left(\log(\{Qo_{t \in y}\}_{80}) - \log(\{Qo_{t \in y}\}_{30}) \right)} - 1 \tag{S8}$$

$$sd^* = \frac{\sum_{y \in Y} \sigma(Qs_{t \in y})}{\sum_{y \in Y} \sigma(Qo_{t \in y})} - 1 \tag{S9}$$

$$r = \frac{1}{N-1} \sum_{t=1}^{N} \left(\frac{Qs_t - \mu(Qs)}{\sigma(Qs)} \right) \left(\frac{Qo_t - \mu(Qo)}{\sigma(Qo)} \right) \tag{S10}$$

$$pc0^* = \frac{n(\{t | Qs_t < 5 \times 10^{-4}\})}{n(\{t | Qo_t = 0\})} - 1 \tag{S11}$$

$$TPR0 = \frac{n(\{t | Qs_t < 5 \times 10^{-4} \wedge Qo_t = 0\})}{n(\{t | Qo_t = 0\})} \tag{S12}$$

Where:

- Qo and Qs are observed and simulated streamflow, respectively;
- $\mu(\cdot)$ and $\sigma(\cdot)$ are mean and standard deviation of the quantity in parentheses;
- in eq. S1, r comes from eq. S10, $\alpha = \frac{\mu(Qs)}{\mu(Qo)}$, and $\beta = \frac{\sigma(Qs)}{\sigma(Qo)}$;
- in eq. S2, r_{r5}, α_{r5}, and β_{r5} retain the same definitions, with flows transformed to their fifth root;
- t indicates a timestep and N is the total number of timesteps with valid observations;
- $b(\cdot)$ indicates the algorithm described by Tallaksen and Van Lanen (2004) to calculate baseflow;
- $H_s = \{t|Qs_t > \{Qs\}_{98}\}$ and $H_o = \{t|Qo_t > \{Qo\}_{98}\}$ are the sets of timesteps where Qs and Qo have exceedance probability < 0.02, respectively;
- $L_s = \{t|Qs_t < \{Qs\}_{30}\}$ and $L_o = \{t|Qo_t < \{Qo\}_{30}\}$ are the sets of timesteps where Qs and Qo have exceedance probability > 0.7, respectively;
- $\{\cdot\}_p$ indicates the $p-$th percentile of the quantity in the curly brackets;
- y is a (water) year, and Y is the set of years with less than 15 missing observations;
- $n(\cdot)$ denotes the cardinality of the set in parentheses.

Text S2. It is impossible to definitively determine whether the assumption of meaningful rankability of differences is fulfilled for the performance metrics in this study. Therefore, we assess the applicability of the matched-pairs rank biserial correlation coefficient (r_c), which requires this assumption, by evaluating how its value changes when a monotonous transformation is applied to the performance metrics. Specifically, we take all the unbounded metrics (E) and bound them to the interval $[-1, 1]$, using the following transformation:

$$E_{bnd} = \frac{E}{2 \pm E} \tag{S13}$$

where the sign at the denominator is $-$ for the metrics whose original range was $(-\infty, 1]$ (i.e. the two KGEs and the objective function), and $+$ for all the biases, whose original range was $[-1, \infty)$. The bounding performed by eq. S13 was proposed by Mathevet, Michel, Andréassian, and Perrin (2006) to bound the Nash-Sutcliffe efficiency metric and is extended here to the biases by changing the sign at the denominator. Using the example of the KGE, the effect of this transformation on the performance differences is to give more weight (i.e. higher rank) to changes in KGE closer to 1 compared to those of the same magnitude in the negative realm. This is arguably a better encoding for the differences in KGE performance. However, our aim here is not to discuss or prove this, but to assess what impact this transformation has on r_c values and orders for this specific dataset and set of performance metrics and hence evaluate the importance of the assumption of meaningful rankability.

The result of this comparison are shown as scatter plots in figure S6. These plots show the value of r_c for each metric in its unbound (x-axis) and bound (y-axis) versions. While there are a few changes in the order of the metrics, the only metric whose values of r_c calculated with the two methods are not compatible within their 95 % confidence intervals is `pc_0` and only in GR4J and HBV and only in the *Post-MD* period. These differences suggest that the level of degradation of the metrics in the *zeros* groups may be underestimated, especially after the drought. However, the results and findings of our study are not affected by the transformation, hence supporting the use of r_c to quantify metric degradation regardless of the assumption of meaningful rankability of metrics' differences.

Text S3. The dataset provided with this publication contain the following:

1. calibrated parameter sets for each model and catchment in the study;

2. timeseries of model simulated streamflow for each catchment and model;

3. values of each performance metric in Table 1 for each catchment and model combination during calibration, each evaluation period and each individual year in the evaluation period;

4. values of matched-pairs rank-biserial correlation coefficient (r_c) for each performance metric and model (see §2.5.1); and

5. results of the annual linear regression for each metric and model (see §2.5.2).

References

Mathevet, T., Michel, C., Andréassian, V., & Perrin, C. (2006). A bounded version of the Nash-Sutcliffe criterion for better model assessment on large sets of basins. *IAHS-AISH Publication*(307), 211–219.

Tallaksen, L. M., & Van Lanen, H. A. J. (2004). *Hydrological Drought: Processes and Estimation Methods for Streamflow and Groundwater* (Vol. 48). Amsterdam, London: Elsevier B.V.

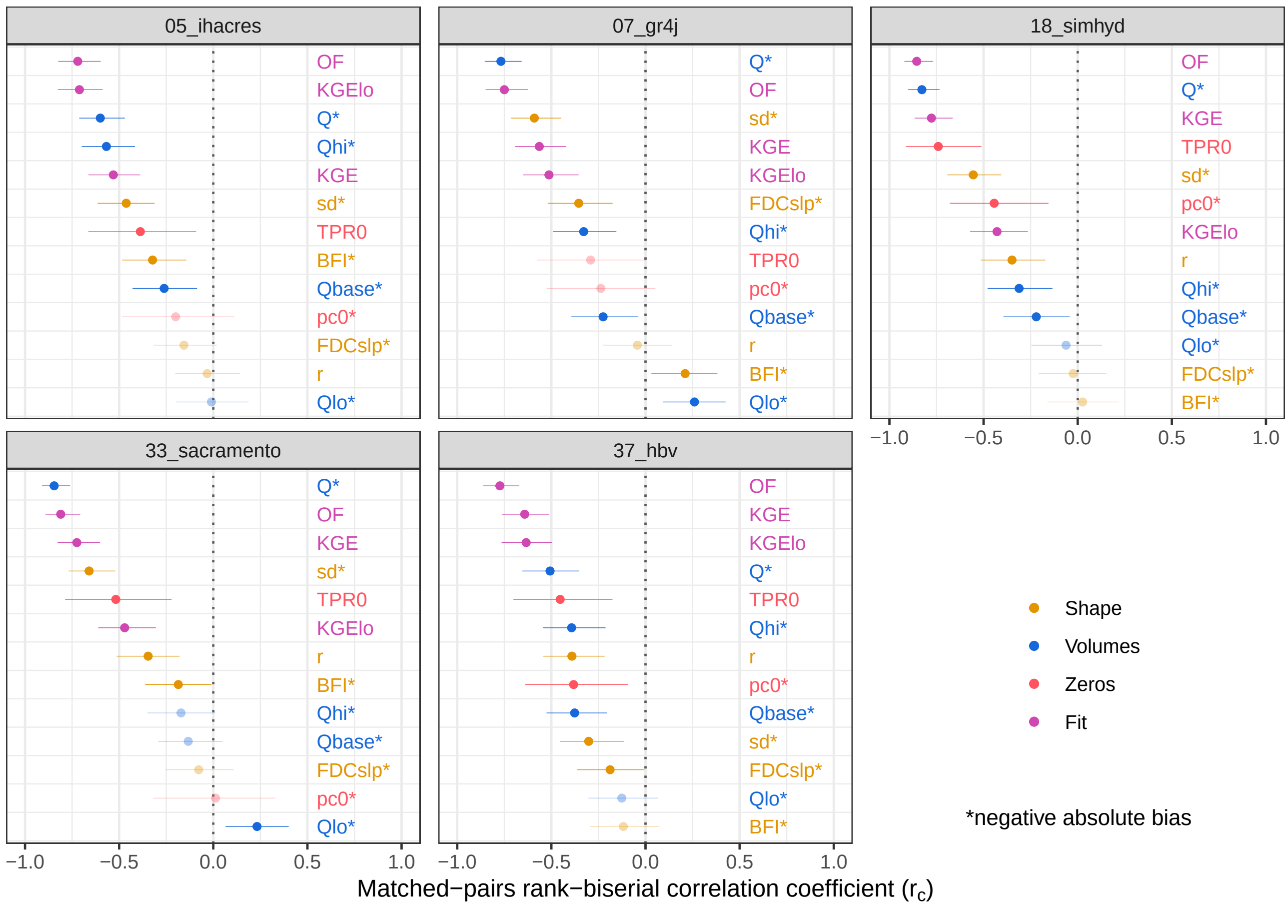

Figure S1. Changes in individual model performance from calibration to *Pre-MD* evaluation. See Fig. 3 for details.

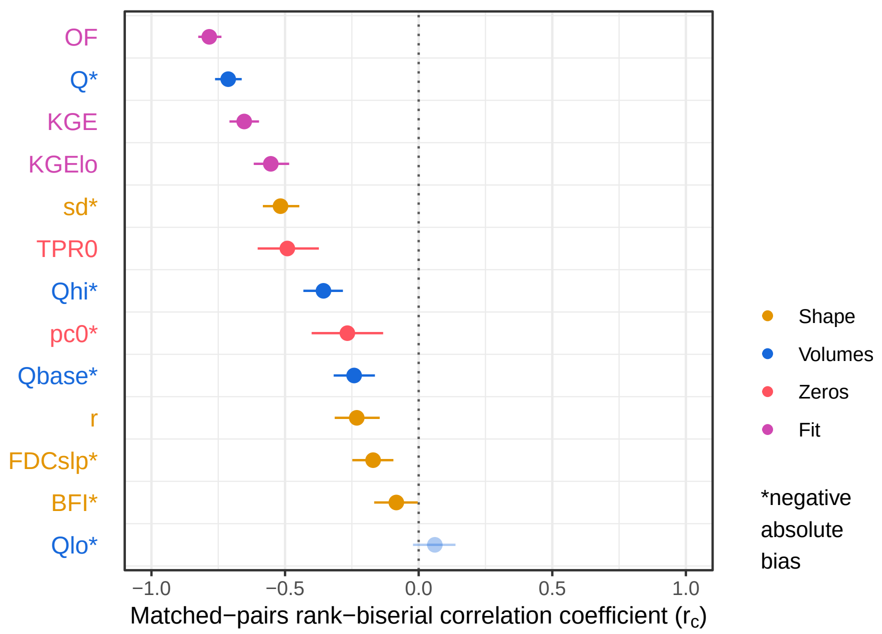

Figure S2. Changes in all models performance from calibration to *Pre-MD* evaluation. See Fig. 3 for details.

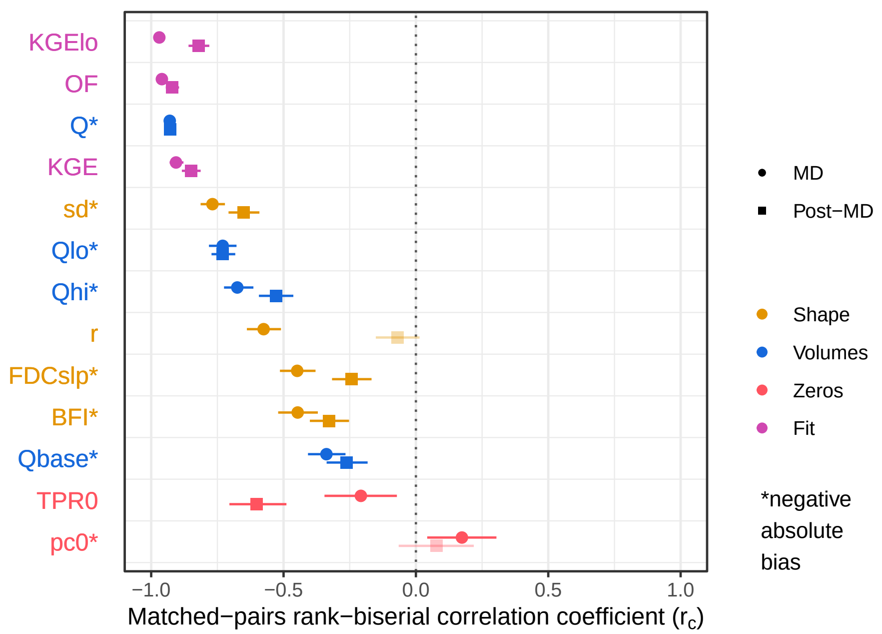

Figure S3. Changes in performance from pre-MD evaluation (benchmark) to each of MD and post-MD across all models. See Fig. 3 for details.

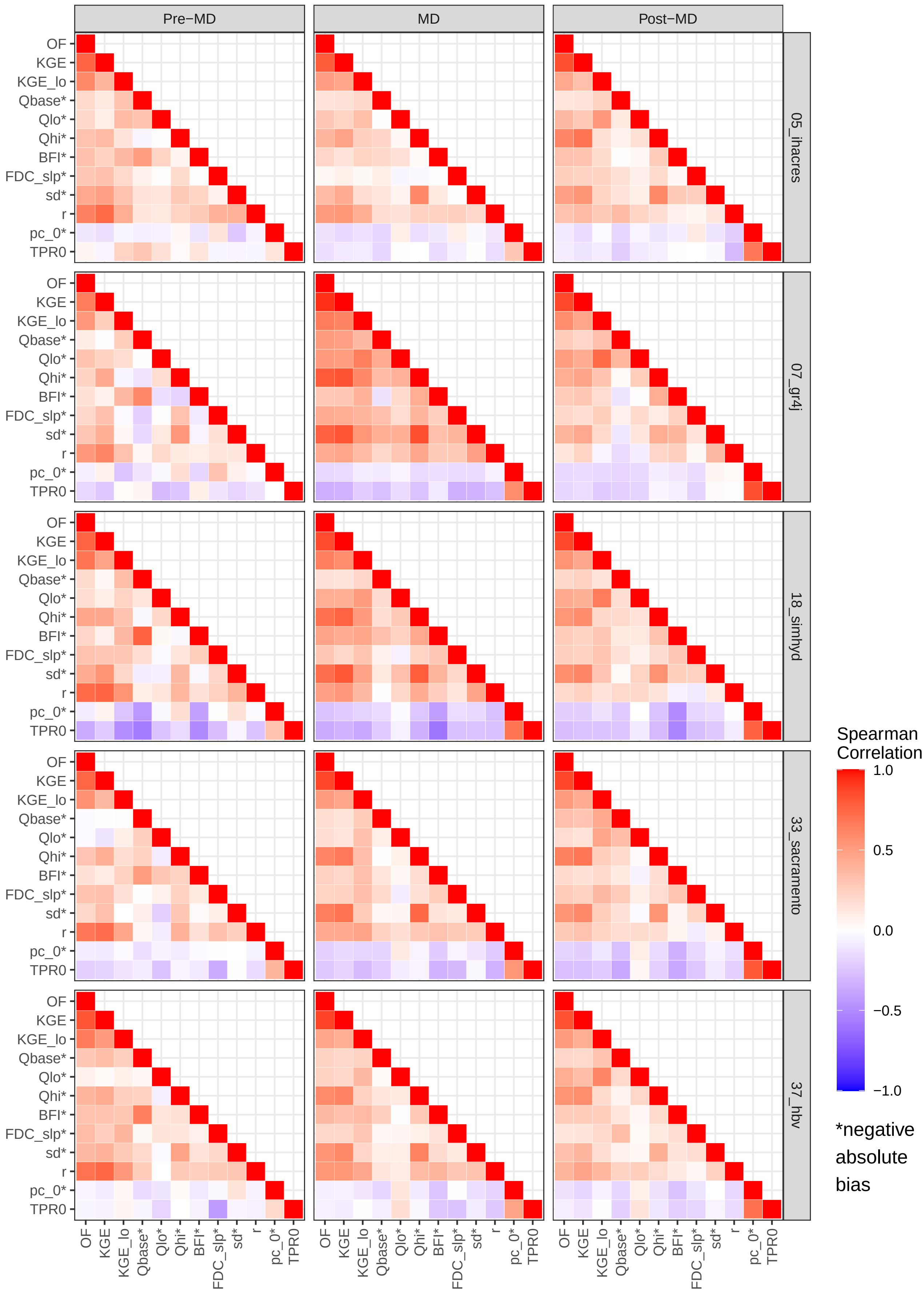

Figure S4. Spearman correlation matrices for each performance indicator model and evaluation period.

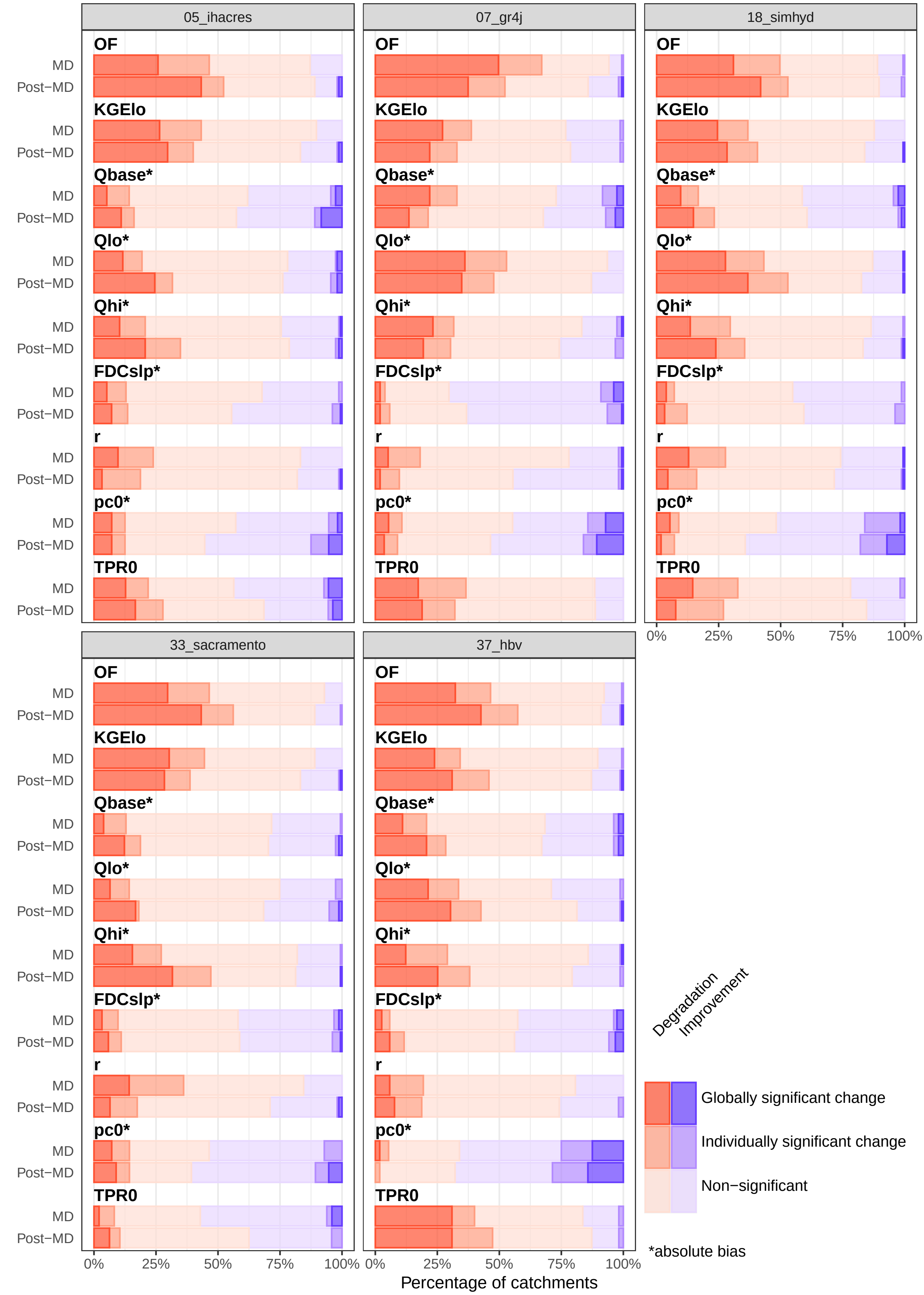

Figure S5. Percentages of catchments in each class of statistical significance of the change in the relationship between annual model performance and annual rainfall anomaly.

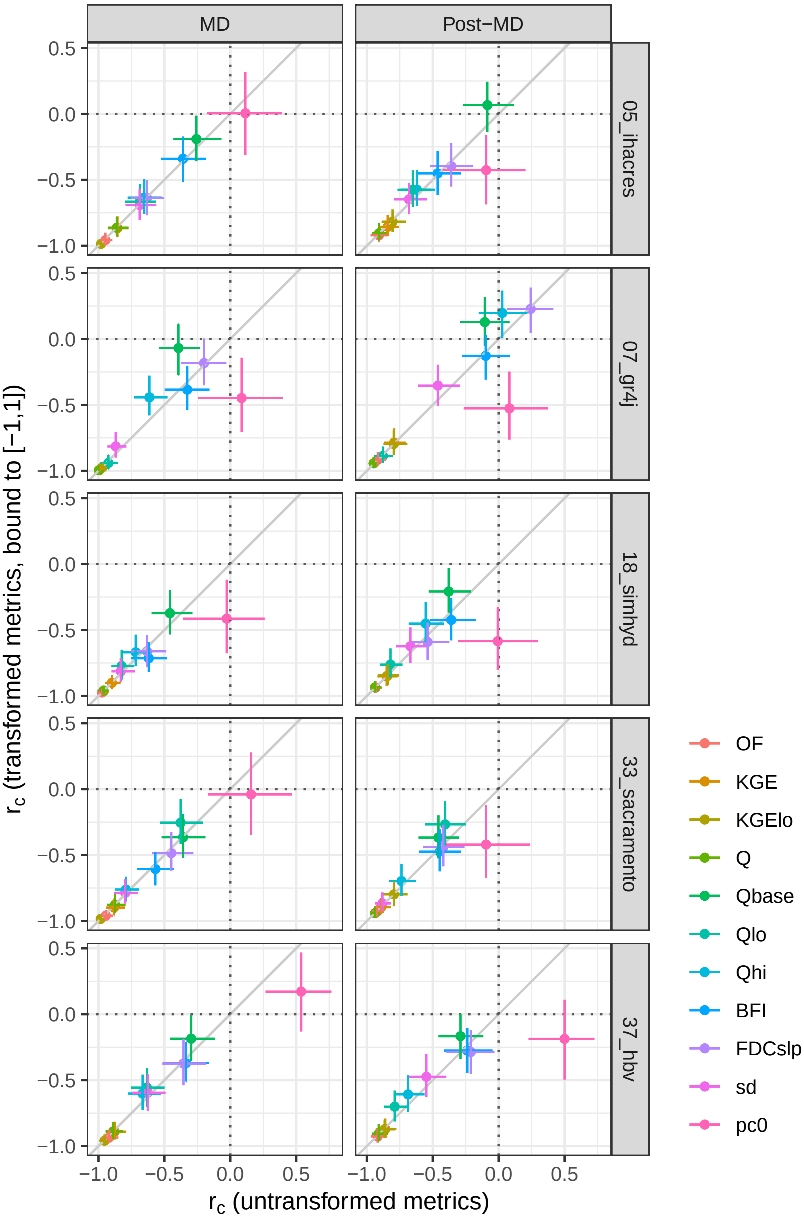

Figure S6. Comparison of the values of r_c using bounded or unbounded versions of the performance metrics.