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 underperform in transitory climate conditions. Existing research identifies models' inadequacy to perform during prolonged drought, but falls short on pinpointing which specific aspects of model performance are affected. We study five conceptual rainfall-runoff models and their performance in 155 Australian catchments which recently experienced a 13-year long drought. We use a wide range of performance metrics and a methodology based on ranked differences to a benchmark to fairly compare levels of degradation across metrics and periods. We show model performance degrading extensively during and after the drought, largely driven by overestimation of flow. Representation of shape and variability of hydrograph and flow-duration curve are more resilient to the prolonged dry climate and rarely more degraded during the multi-annual drought that on isolated dry years in the pre-drought record. Conversely, volumetric error suffers from significant exacerbation over the multiple subsequent dry years. This indicates that catchment retention times and rates of storage depletion storage are significantly less affected by the drought than amounts of streamflow produced, pointing to a mismatch between reduction of influxes and out-fluxes during the drought. We also identify a deficiency of models to delay and remove flow before it reaches the stream and keep track of moisture deficits over multiple dry seasons. By promoting rigorous investigation of models' shortcomings, we hope to foster the development of more robust model structures and/or calibration frameworks 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, 20052015), 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; Hanel et al., 2018; 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 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; Fowler et al., 2022). 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), Chile (Alvarez-Garreton et al., 2021) and Europe (Massari et al., 2022). 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 (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 (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).

Models' inability to extrapolate satisfactorily to previously unseen climate conditions is intrinsically linked to the issue of parameter instability. Parameter instability refers to the dependence of model parameters to the characteristics of the specific period that they have been optimised for (Brigode et al., 2013). In the context of a transient climate, parameter instability is a considerable source of uncertainty as parameters that are optimised for past conditions are unlikely to be adequate for projection under a future climate (Seiller et al., 2012; Stephens et al., 2020), even when calibration sequences contains periods that are similar to the changed conditions (Trotter et al., 2021). In order to evaluate the adequacy of models for use in these scenarios, many more or less complex validation strategies have been proposed (e.g. Merz et al., 2011; Seiller et al., 2012; Thirel et al., 2015; Motavita et al., 2019; Royer-Gaspard et al., 2021), all stemming from the idea of differential split-sample testing – DSST (Kleme, 1986). DSST consists of utilising periods of constrasting conditions in the historical record for calibration and evaluation in order to force the model to extrapolate into previously unseen conditions. DSST studies have consistently observed strongly biased responses in the evaluation periods, especially for bucket-type conceptual rainfall-runoff models (e.g. Vaze et al., 2010; Seiller et al., 2012; Brigode et al., 2013; Broderick et al., 2016), but also in semidistributed hydrological models (e.g. Merz et al., 2011; Duethmann et al., 2020), distributed ecohydrological models (e.g. Stephens et al., 2020), and groundwater recharge models of various complexity (e.g. Moeck et al., 2018).

In some cases, models can achieve more satisfactory performance if they are shown both humid and dry conditions by using a multi-objective approach to the calibration optimisation (e.g. Fowler et al., 2016; Smith et al., 2019). In the context of the Millennium drought, 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.

Gaining a deeper understanding of how well (or poorly) models reproduce specific aspects of the hydrograph (e.g. volumes of high and low flows, timing of peak flow, transition from high- to low-flow regime) is of great interest for water management and allocation purposes, additionally it provides us with important hints on how hydrological processes are affected by these long drought events and possible model remediation strategies. This is particularly true when comparing performance during persistent drought to that during "regular" dry years. 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 traditional DSST studies 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 155 of the catchments already used by Peterson et al. (2021). Those catchments had been selected as mostly unimpaired by human influences on their flow regimes including regulation, known diversions, and land use changes (Peterson et al., 2021). 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 unexpected streamflow deficits persisted years after the end of the meteorological drought (Peterson et al., 2021), which broke in 2010 due to heavy rainfall and extensive flooding throughout the region (CSIRO, 2012), despite a return to approximately average climatic conditions including a few wet years (Peterson et al., 2021). 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 at the beginning of Australian Autumn on 01 March and end on the last day of February of the subsequent year (Peterson et al., 2021).

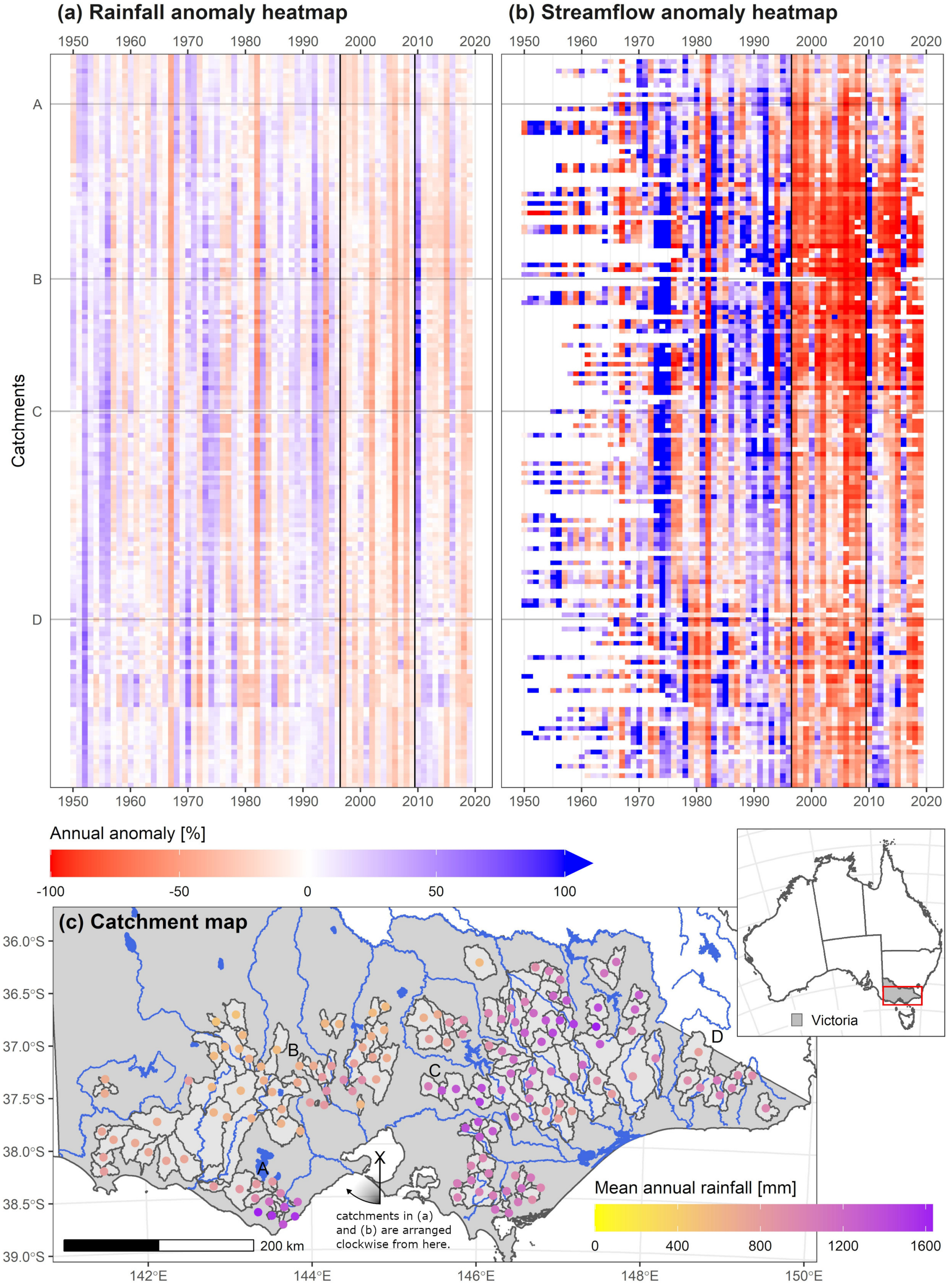

Figure 1. **(a, b)** Annual rainfall (a) and streamflow (b) anomalies for each of the catchments in this study. Each line represents a catchment. Catchments are arranged by the clockwise angle from the south axis created by connecting their centroid to the centre of Port Phillip Bay (point X in (c)). 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.

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°×0.05° (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 calculated for each of the catchments in this study by averaging the values of each grid pixel falling inside each catchment, weighted by the catchment coverage. 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 and used by Peterson et al. (2021), from the WMIS portal of the Victorian Department of Environment, Land, Water and Planning (Peterson et al., 2021). As the dataset compiled by Peterson et al. (2021) 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 Peterson et al. (2021) 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., 1997). These models were chosen to cover a range of complexities (see Table 1) and be-

Table 1. Characteristics of the hydrological models used in this study (Knoben 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)	

cause 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 et al., 2019; Trotter et al., 2022).

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).

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

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 form and constants of the bias penalisation addend to Eq. 1 were proposed by Viney et al. (2009). 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, careful consideration is given to what portion of the data to use for model calibration and what portion(s) to reserve for independent evaluation. Specifically, we require four independent samples: namely one calibration sequence, which must be in the Pre-MD period; and three evaluation sequences, one in the pre-MD period, one in the MD and one in the post-MD period. The pre-MD evaluation sequence is necessary to guarantee that model performance degradation is assessed by comparing model performance on non-training data only. In practice, model performance during the MD and the post-MD (evaluation) periods will be compared to their performance during this pre-MD evaluation sequence, to assess changes in performance caused by the need for the models to extrapolate to previously unseen climate conditions. The pre-MD evaluation period, therefore, must be designed as to minimise extrapolation, as model performance during this period should represent how models would be expected to perform in evaluation had the climate remained stable. To achieve this objective, we use a systematic blocking strategy (Roberts et al., 2017) of alternate years for the pre-MD period, using only the even years of the record for calibration. Model performance on pre-MD odd years is then used as an evaluation benchmark for MD and post-MD performance. Kolmogorov-Smirnov tests were conducted to ensure that distributions of annual rainfall and potential ET in the two pre-MD periods are not significantly different (i.e. that minimal extrapolation is being requested of the models). 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.

In practice, for calibration, models run from 5 years before the first day of the first water year with observed streamflow measured and up to the end of the last PreMD water year (i.e. 28 February 1997). The objective function optimised for calibration is calculated on the streamflow from these runs, using data from even water years starting from the first year with observed streamflow. E.g. if streamflow data for a given catchment starts in 1973, the model is ran from 1968 to 1996 (incl.) and the objective function is calculated using observed and simulated flows from all the even years between 1974 and 1996. This gives the models 5 years to warm up and stabilise their state variables. After calibration, parametrised models are ran again from the same starting point (i.e. using the same 5-year warm up time) and until the end of the observed record. The single record of simulated streamflow is then used to calculate model performance during odd PreMD year, all MD years and all available PostMD years.

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 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, $cv*$ indicates differences in their variability. Note that $Q*$ and $cv*$, albeit after some algebraic manipulation, are, together with r, components of KGE (Gupta et al., 2009). The use of

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	*cv**	Bias in the annual coefficient of variation.	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

the coefficient of variation (ratio of standard deviation to mean) to evaluate flow variability removes the dependency to flow volumes and hence to volumetric bias.

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), i.e. by connecting the minima of non-overlapping periods of 7 days of flow whose value is deemed a turning point by comparison with adjacent minima (see Tallaksen & Van Lanen, 2004, §5.3, for details). The R package *lfstat* (Koffler et al., 2016) was used to compute baseflow. 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 biases in the volume of the peak-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

To compare how model performance during and after the Millennium drought changes from the pre-MD evaluation period across the set of performance metrics, we need to take into consideration that the metrics in Table 2 are on different scales and have different sensitivities and even the ones that share the same endpoints and optimal values are not 1-to-1 comparable. To ensure comparability between the levels of degradation for each metric, we make use of signed ranked differences rather than absolute values of the metrics. Whereas this reduces the information content in the data, it is a necessary compromise to achieve the objectives of this study.

Specifically, we use matched-pairs rank-biserial correlation coefficients to quantify changes in performance across the set of catchments. 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). *Rank-biserial* correlation measures the correlation between a ranked variable (here the values of performance according to each metric) and a dichotomy (i.e. a set of 0s and 1s indicating if the performance in question comes from the pre-MD or the MD/post-MD). It effectively measures if and to what extent the numerical values associated with the value of 1 in the dichotomy are larger or smaller than those having a value of 0 (Cureton, 1956). The use of ranks removes the need for distributional assumptions and, as mentioned, in our case it allows us to compare results from metrics on different scales. The *matched-pairs* version of rank-biserial correlation is its extension to dependent samples, i.e. the case when each numeric value with dichotomy= 1 has a corresponding value with dichotomy=

0. This is the case for this research where the same sets of catchments are used to evaluate model performance in each of the three periods.

For each model, period of interest $\tau \in \{\text{MD, post-MD}\}$, 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. The R package *effectsize* (Ben-Shachar et al., 2020) was used to calculate r_c and its confidence intervals

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.

Application of this metric to our dataset requires two assumptions to be fulfilled: (1) 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); and (2) that differences of metric values can be meaningfully ranked (King & Minium, 2003). To comply with the first requirement, all of the performance metrics based on bias are transformed by taking the opposite of their absolute values. Regarding the second assumption, it is impossible to evaluate objectively its validity. For the purpose of this study, we have tested its influence and concluded that it is unlikely to have significant impact on the results. Details and further discussion 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. The relevance of this analysis stems from the observation, already highlighted in the Introduction, that models generally perform worse during drier periods. Within this study, we verified this by testing for correlation between annual values of each performance metric and annual rainfall anomaly. We observed a general negative correlation between the metrics' distance to their perfect score and rainfall anomaly in most catchments and for all metrics except those in the *zeros* group (see Fig. S4). In particular, we notice that this correlation is more significant for those metrics where higher levels of degradation is observed (see §3.2). For this part of the analysis, we account for this correlation to assess whether it is sufficient to explain the changes in performance observed during (and after) the MD.

To account for this correlation, we used dual-intercept linear regressions of (transformed) annual performance metric as a function of annual rainfall anomaly, 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, used here to linearise the relationship, 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). $\tilde{E}$ 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. We applied the regression model in equation 3 to all combinations of model, period τ and performance metric, excluding the metrics in the *zeros* group. Many of the annual values of these metrics are undetermined due to zeros in their denominator. Additionally, as mentioned, the negative correlation between the remaining annual values of these metrics and the rainfall anomaly indicates a relationship that is often opposite that of the other metrics. Amongst the other metrics, there are also instances where annual values are undetermined: this occurs in ephemeral catchments with many zero flows in a single year. For example, when more than 30 % of observed flows in a year is zero, the denominator of *Qlo** for that year is zero. In a few cases, catchments cease to produce flow for entire years during the MD causing all metrics to be undetermined in those years. To ensure stable enough regression results despite these missing data points, we excluded any regression where the number of points for each period is less than 3, this excluded 35 possible regressions from our analysis. Additionally, we imposed a minimum amount of overlap between the domains of the independent variable P_a associated with $I = 0$ and $I = 1$ for each regression. In particular, for each regression, we looked at the range of P_a values that is common between the two periods studies (i.e. $[min(P_{a,I=0}), max(P_{a,I=0})] \cap [min(P_{a,I=1}), max(P_{a,I=1})]$) and eliminated possible regressions where for either of the two periods considered the percentage of P_a value within this range is smaller than 25 %. The overlap criterion was introduced to limit the risk of a significant change in intercept being identified as an artifact of a false slope. An additional 130 possible regressions were removed from the analysis due to this criterion. Finally, appropriate tests to check for normality, homoscendasticity and (lack of) autocorrelation were conducted on the residuals of the remaining 16 885 linear regressions (Haan, 2002). Only 70 (0.41 %) did not pass at least one of these tests, these were also removed from the analysis (see supporting information text S3).

3 Results

3.1 Model performance before the MD

Median and average values and interquartile ranges of all performance metrics in the pre-MD period, for both calibration (even years) and evaluation (odd years), are shown in Figure 2. As the figure shows, all models calibrated very well. Models' average calibration efficiency range from SimHyd's 0.81 to Sacramento's 0.85. Median values range from 0.82 (also SimHyd) to 0.88 (HBV). Because of the bias penalisation in the equation for the objective function, volumetric bias in calibration (Q^*) is almost absent (average and median below 3 % for all models) and performance in the calibration period of *KGE* and *KGElo* have very similar values to those of *OF*. SimHyd consistently has the lowest values for these metrics while Sacramento and HBV alternate for the highest values. Looking at the other metrics, however, reveals that there is no consistency in the ranking of models across all the metrics and all of the models underperform compared to all the others according to at least one metric.

With regards to the other metrics, all models tend to underestimate peak-flow and baseflow volumes (negative *Qhi** and *Qbase**), whereas they are inconsistent with low-flow volumes: GR4J and SimHyd having positive average bias (as high as 18.9 % for GR4J) and the other three models having negative average bias, with Sacramento underestimating low-flow volumes the most at −22.2 %, on average. The BFI is also underestimated by all models, as expected by the observation that they slightly overestimate flow and underestimate baseflow. The slope of the FDC is overestimated on average by all models. Additionally, we note that absolute values in calibration of *Qlo** and especially of *Qbase** and *BFI** are generally higher than those of other bias metrics such as *Qhi**, Q^* and *cv**. As discussed in the Methodology section (see §2.5.1), however, these performance metrics are not one-to-one comparable and there is no objective way of determining whether

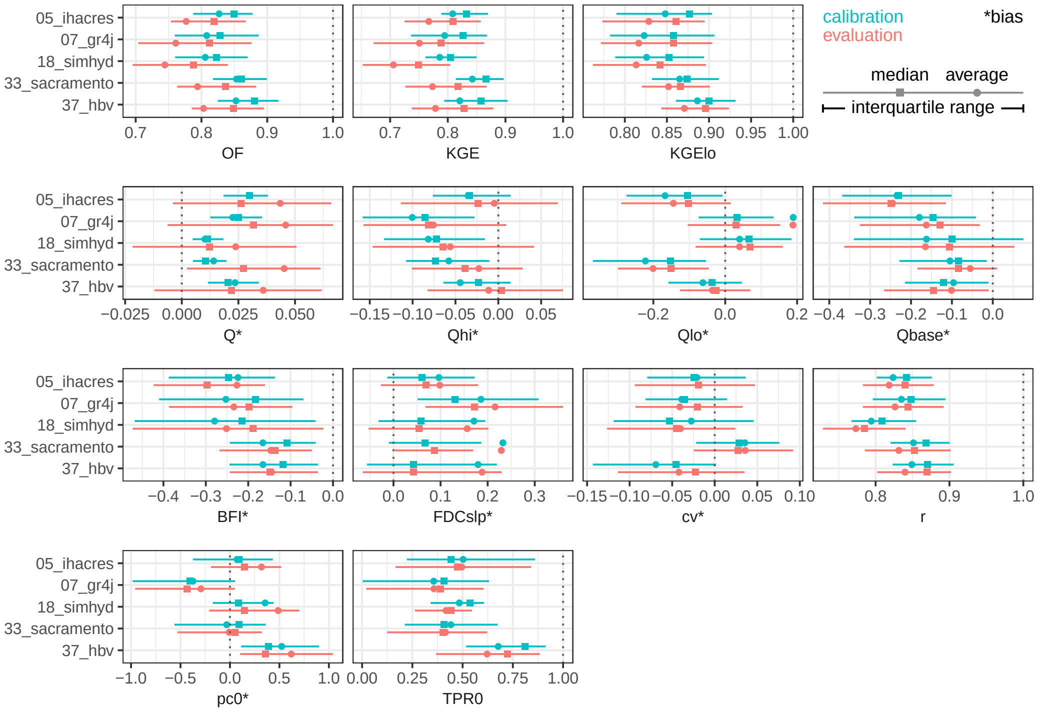

Figure 2. Model performance in the pre-MD period: comparison of all metrics between calibration (odd years) and evaluation (even years). Showing interquartile range, median and mean of performance across catchments. See Table 2 for the meaning of the performance metrics.

a value of a metric corresponds to a higher or lower overall model performance than the value of another metric.

Performance during pre-MD evaluation (red markers in Figure 2) is very similar to the performance during calibration for most metrics and models, with the biggest differences in volumetric bias and summary metrics, prioritising high flow metrics (i.e. *OF* and *KGE*). According to a Wilcoxon test, *OF* and *KGE* are the only metrics where the difference between the calibration and evaluation performance before the MD is significant (all models except for GR4J). For the models Sacramento and HBV, the Wilcoxon test also detected significant differences between calibration and Pre-MD evaluation values of the metrics *Q** (Sacramento) and *Qhi** (HBV). This 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.

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 S1, while Figure S2 shows mean, median and interquartile range of performance for each metric and model in each of the three evaluation periods. When looking at the order and extent of degradation from the benchmark of these metrics from Figures 3 and S1, 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 $Q*$, $cv*$ and r which make up the KGE and hence the objective function, can be highly correlated. Correlation matrices for all metrics across all evaluation periods and models are shown in Figure S3.

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 -0.868 (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 130 and 149 catchments (or 83.9 % to 96.1 % of 155) depending on the metric and the model. On average models performed worse than they did in the benchmark period in 146 (94.2 %), 134 (86.5 %), and 146 (94.2 %) catchments for *OF*, *KGE* and *KGElo* respectively. With the exception of GR4J, model performance as measured by the transformed 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 streamflow volumes. 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 -0.883 (IHACRES, MD) for

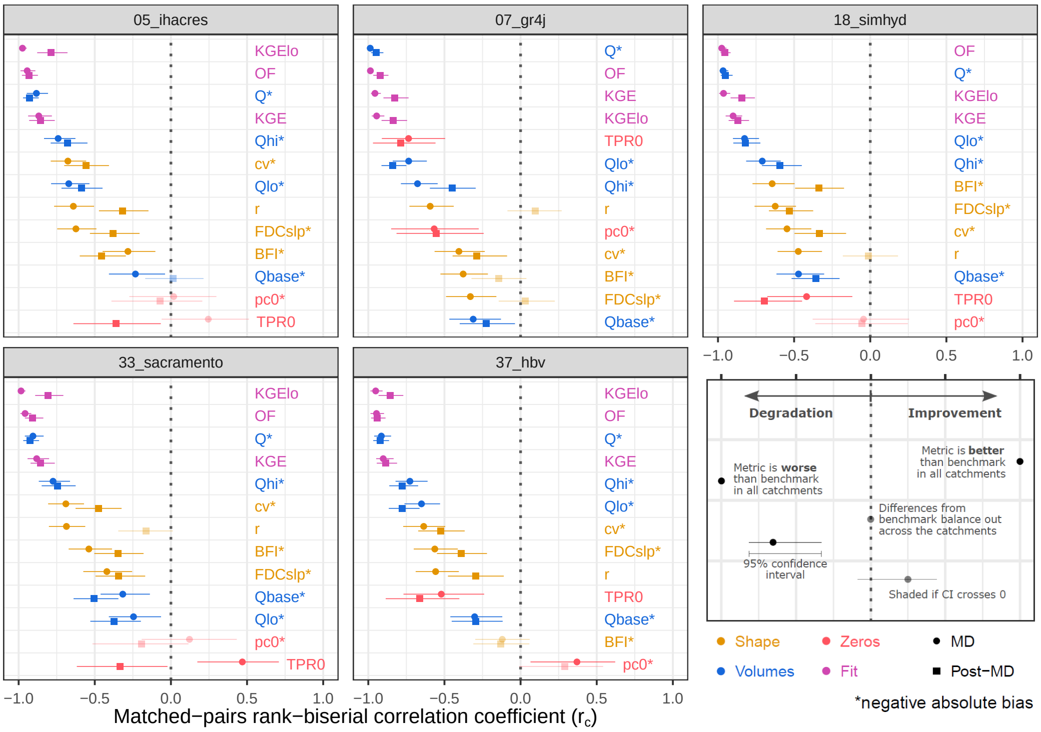

Figure 3. Changes in individual models performance from pre-MD evaluation (benchmark) to each of MD and post-MD. $r_c = -1$ $(+1)$ indicate that the model performance according to that metric degrades (improves) from the benchmark in all catchments. Ranges indicate 95 % confidence intervals, points are 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).

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 3.88 %, and it was positive (i.e. streamflow overestimated) in 109 catchments on average; during the drought, the average bias jumps to 54.3 % and the average number of catchments with overestimated streamflow become 126; even after the end of the drought, the average bias remains at 40.0 % (with 136 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 ($cv*$) 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 coefficient of variation degrades in 100 to 116 catchments (or 64.5 % to 74.8 %) in the MD compared to the benchmark. After the drought, the number of catchments with $cv*$ worse than before the drought remains 90 to 111, depending on the model. In the pre-MD benchmark, one model (Sacramento) slightly overestimated the variability of flows (average pre-MD $cv*$=3.57 %), during (after) the drought, the overestimation increased to on average 12.3 % (5.60 %). Of the other four models which instead tended to underestimate variability in the pre-MD period (−3.58 % on average), three increased the extent of the underestimation (up to 2.8 times, HBV) during and after the drought. This is clearly visible in Figure S2 and overall resulted in values of r_c ranging from −0.69 to −0.41 for the MD and −0.56 to −0.29 for the post-MD. The extent of degradation of the linear correlation coefficient between observed and simulated flows is similar, with 102 to 116 catchments having worse r during the drought than in the benchmark period and r_c values between −0.69 to −0.47. However, 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.

Streamflow overestimation during and after the drought affects both high and low flows, driving down model performance. With respect to biases in the peak- 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 ($-7.57\,\%$ to $-0.47\,\%$, on average), but overestimate it during and after the drought ($18.2\,\%$ to $75.4\,\%$ and $20.0\,\%$ to $27.4\,\%$, on average respectively), see Figure S2. 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 (114 to 122), resulting in values of r_c between -0.677 (GR4J) and -0.774 (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. The performance degradation in terms of volume estimates of the low-flow portion of the FDC is driven by the same mechanisms, i.e. overestimation of flow volumes. Here the initial values of pre-MD bias are more varied from model to model ($-20.1\,\%$ to $18.8\,\%$) and the increase in percentage overestimation are much higher: on average higher than $150\,\%$ 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. It is worth noting that the distributions of *Qlo** are heavily skewed compared to those of *Qhi** (see difference between medians and averages in Fig. S2). This is caused by the fact that reduction of flow during and after the drought caused large increases in the number of zero or near-zero flows of many the catchments in the dataset, bringing for these catchments the denominator of Eq. S5 very close to zero and inflating values of *Qlo**. This highlights the value of assessing model degradation using ranks of differences. The median values of *Qhi** and *Qlo**, instead, are comparable for all models and periods, and so are mostly their r_c values, with the only exception of Sacramento for which *Qlo** degraded significantly less than *Qhi**.

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 95 to 115 (67 to 109) 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 44.3 % of catchment-model pairs; this value reduces to 32.0 % 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 three out of five models, with average pre-MD values of *pc0** ranging from GR4J's −29.4 % to HBV's 61.9 %. *pc0** is on average underestimated both during and after the drought in all models except for IHACRES, MD (value of *pc0** ranging from −72.1 % to 8.21 %, GR4J, post-MD and IHACRES, 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 35 (22 to 33) of the 56 catchments across which these metrics are calculated which causes r_c for this metric to not be significantly below the zero for four out of five models. 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. This resulted in a total of 16 815 regressions after removing the regressions that did not meet the quality criteria outlined in §2.5.2. 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 coefficient of variation and the baseflow index. These metrics are representative of models' ability to reproduce the overall hydrograph, its volumes, variability and shape respectively and based on the results from the r_c analysis above showed different levels of degradation. Results from the remainder of the metrics can be seen in Figure S5. Additionally, in Figure S6 we show the median and interquartile ranges of values of annual model performance in each evaluation year.

In Figures 4 and S5, the catchments represented in the red bars had the least-squares fitting on the linear model in eq. 3 resulting 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 $\beta_2 = 0$. Average coefficients of determination for the linear regressions are quite low, ranging from 0.10 ($cv*$) to 0.33 (OF). This indicates that rainfall anomaly and period together explain only a small portion of the variance of annual model performance. The significance of the change in intercept associated with the change in period studied here, however, is not affected by the predictive power (or lack thereof) of the linear model chosen; as the significance of the $t-$test on $\hat{\beta_2}$ in itself takes into account the variance of the data: it is actually harder to observe significant shifts in the intercept with very noisy data having low coefficient of determination. This adds to the relevance of the cases where a significant shift is indeed observed.

During the drought, the change of KGE-to-anomaly relationship is individually significant and negative in between 42.2 % (IHACRES) and 46.7 % (SimHyd) of catchments for most models. GR4J is the only model for which the change is significant in a majority of catchments 53.9 %. After the drought, these percentages increase, with all models surpassing the threshold of half of the catchments with significantly degraded annual performance for a given P anomaly. Conversely, the number of catchments where the relationship changes significantly for the better (i.e. annual KGE is higher during or af-

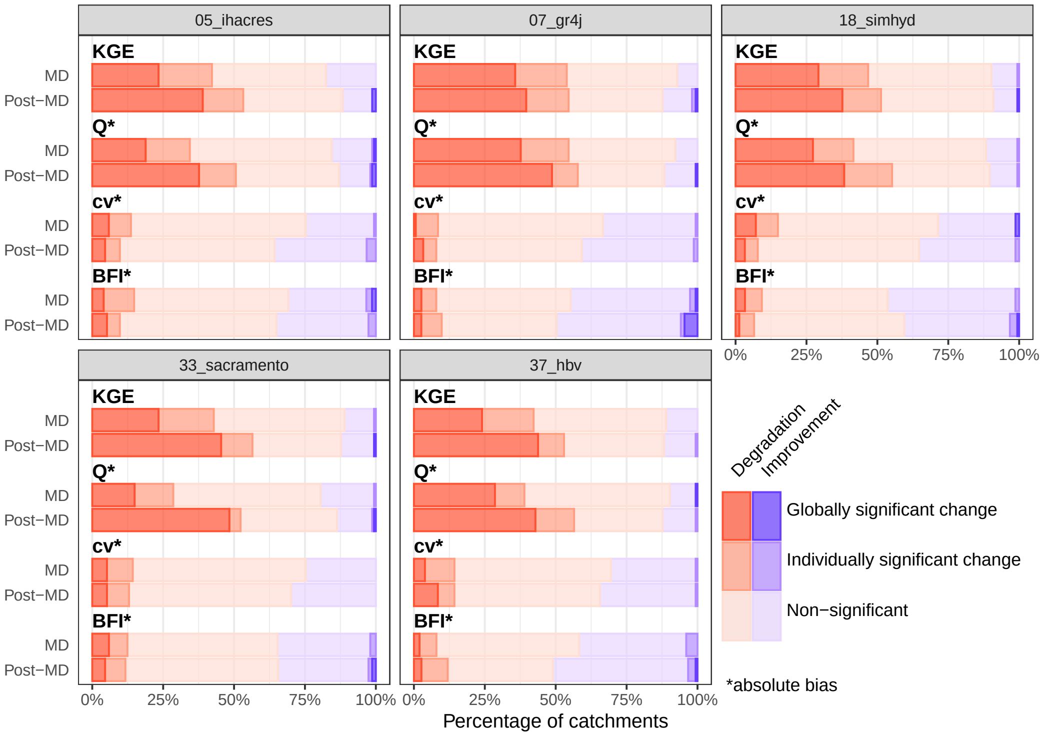

Figure 4. Changes in the relationship between annual model performance and annual rainfall anomaly, showing percentages of catchments in each class of statistical significance. Statistical significance is assessed with a t-test on the least-squares fitting of the period-specific intercept β_2 of the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

ter 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 four out of five models, amongst the catchments where the shift is significant, it is overwhelmingly towards a degradation: at least 98.5 % of catchments with significant shifts during the drought and at least 96.6 % 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 28.6 % to 41.6 % of catchments for most model, with again GR4J being the outlier with 54.5 %. Similarly as with the KGE, these percentages increase to at least 50.0 % after the drought for all models. Amongst the catchments where the change in Q-to-anomaly relationship is significant, again the change is overwhelmingly towards a degradation: always at least 96.3 % of these catchments.

Finally, with respect to the ability of models to estimate the shape and variability of the hydrograph, as measured by the biases in the coefficient of variation and the baseflow index, the results show that in the majority of catchments these were the same during and after the drought than they were in pre-MD years of similar rainfall anomaly. Here the results of the change analysis are non-significant in at least 83.7 % (SimHyd, MD) and up to 90.8 % (SimHyd, Post-MD) of catchments for *cv** and in 81.9 % to 90.2 % (IHACRES, MD and SimHyd, Post-MD respectively) of catchments for *BFI**. Nevertheless, comparing the number of catchments within the same level of significance, we see again that the catchments where a degradation in performance occurs almost always outnumber, albeit sometimes marginally, those where the performance is improved, especially with regards to *cv**.

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. The analysis of performance in individual years and its relationship with annual rainfall anomaly shows that performance degradation in representing the volumes of flow cannot alone be attributed to drier conditions in individual years. This indicates that 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. Conversely, replication of the shapes of the hydrograph and the flow duration curve is much more resilient to the drier climate. This is also confirmed by the fact that the quality of estimation of peak and low volumes were shown to degrade in a similar way. Additionally, for these performance parameters, no accumulation of error is observed as models perform during the multi-annual drought equally to during individual dry years in the pre-drought record. We also show that levels and patterns of degradation in the post-drought period are equivalent to those observed during the drought, in most cases.

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. As mentioned in the introduction, the observation of a biased response in the evaluation period of models subjected to DSST is not uncommon (e.g. Merz et al., 2011; Seiller et al., 2012; Brigode et al., 2013; Mathevet et al., 2020) and with this regards our finding is in line with findings from previous studies on model performance during the Millennium drought (e.g. Vaze et al., 2010; Coron et al., 2012; Saft et al., 2016). However, most of those studies only assessed model performance based on one or two overall-goodness-of-fit metric, typically the NSE (e.g. Vaze et al., 2010; Merz et al., 2011; Brigode et al., 2013; Saft et al., 2016; Duethmann et al., 2020) and/or its version with square-root (e.g. Seiller et al., 2012) or cube-root (e.g. Broderick et al., 2016) transformed flows, and volumetric bias (used by all of the studies mentioned).

A few studies, however, do adopt additional metrics to evaluate model performance, for example Merz et al. (2011) and Brigode et al. (2013) also look at the bias in the fifth and ninety-fifth percentiles of flow (Q_5 and Q_{95}) to assess model errors in estimating high- and low-flows. The scope of these studies, however, is not to compare model degradation in the DSST across the different metrics, as such the methodology is not designed to provide an unbiased and fair estimate of such comparison. Consider, for example, the results presented by Merz et al. (2011), they state that the error in the low flows as measured by the bias in Q_5 in evaluation is on average 35 % across their set of catchments, while that in the high flows (bias in Q_{95}) is 15 % (Merz et al., 2011). Without comparing these values to a benchmark and without considering that the bias in the low flows is probably more skewed than bias in the high flow (as mentioned in our results and as seen in Figure S2), it would be unfair to conclude that low flow estimation degrade over twice as much than high flow estimation. In our results, we show that $Qlo*$ has already generally worse values than $Qhi*$ both in calibration and in the pre-MD evaluation benchmark (see Fig. 2), and that if the MD and post-MD evaluation values are compared to these benchmark values using a metric based on ranks (i.e. which removes differences in the sensitivities of the metrics, in this case, the skewness of their distribution), the amount of degradation from each of the two metrics is actually comparable for most models.

Another possible approach to asses performance degradation across different metrics and without resorting to ranked statistics is the one presented by Mathevet et al. (2020). The authors use values of linear correlation between performance metrics in calibration and DSST evaluation to assess performance degradation, with the rationale that where linear correlation is high there is little difference between performance in calibration and in evaluation across their set of catchments and the metric is therefore more robust to changes in forcing data (Mathevet et al., 2020). This is also the only study, to our knowledge, that specifically looks at models' ability to reproduce hydrograph shape, using as metrics the KGE and its three components. The results of the degradation as described above are only reported for volumetric bias and Pearson's correlation (i.e. $Q*$ and r in this study), but they support the results shows here, with evaluation bias largely uncorrelated to calibration bias (high degradation) and evaluation r highly correlated to the calibration values (Mathevet et al., 2020). Also Mathevet et al. (2020), however, did not compare the performance in evaluation to an evaluation benchmark, but directly to the calibration performance.

The only DSST study which we could find where a period with similar climate to the calibration period is used as a “control” to compare evaluation performance to is by Broderick et al. (2016). The authors measured model performance with NSE, cube-root NSE and bias and interestingly also made use of ranked values of performance, however not to compare changes in performance across metrics, but to compare values of the same performance metric between different models and limit the effect of extreme values (Broderick et al., 2016; Seiller et al., 2012, also uses the same approach). Despite reporting, as it is common, that DSST performance from a more humid calibration period to a drier evaluation period is worse than the opposite, the authors indicate that when comparing the estimations of volumetric bias in evaluation to the “control” period, they often found the changes to be limited (Broderick et al., 2016). This is possibly because of the small variability in climate available in the dataset used, less than 20 % (Broderick et al., 2016), but, based on the results of the analysis presented here, could also be related to the fact that they used individual drier or wetter years for their DSST set-up and we showed that bias degradation over multiple, subsequent dry years is often significantly larger than during individual dry years with a similar climate. This highlights the importance of distinguishing the effects of multi-annual drought from those of individual dry years.

4.2 Implications for process understanding and process representation

Many of the catchments in this dataset experienced significant changes in their annual rainfall-runoff relationship (Saft et al., 2015; Peterson et al., 2021), 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. In general, one should not infer catchment behaviour from model behaviour because that implies confidence that models actually represent internal catchment processes well (Beven, 2019). However, the fact that models of different complexities and structures all responded in very similar ways gives us some confidence that the errors shown by the models are actually somehow representatives what changes are occurring in the catchments during the MD which causes these observed shifts in rainfall-runoff relationship.

In particular, the fact that we see similar levels of performance degradation in both peak and low flows indicates that flow during the drought decreased across the whole spectrum proportionally. This suggests that many catchments processes are concurrently responsible for the shift in behaviour across high and low flow periods. The fact that the literature hasn't yet been able to pinpoint the one specific culprit of changes in rainfall-runoff relationships observed during the MD also supports this point (see Fowler et al., 2022, for a review of hypotheses for process understanding of shifts during the MD). Additionally, the fact that metrics associated with the responsiveness of catchments (BFI^*) and the timing (r) of flows are less degraded during the drought indicates that despite the models discharging too much flow, the rates of change and timing of discharge are simulated better. This could indicate that corresponding aspects of catchment behaviour are less affected by the drought.

With regards to model structures, these results point to unrealistic representations of the mechanisms to remove water before it reaches the stream. In particular, evapotranspiration (ET) is commonly the only mechanisms for models to remove water from the system (this is true for all of the models tested here except for GR4J). Modelled ET fluxes are often proportional to storage and so is streamflow. In order to reduce the streamflow while maintaining ratios of high-to-low flows (i.e. like it would be needed to correct the errors described in this analysis), one can increase ET, proportionally, so that the rate of depletion of storage is similar to before, but less flow is available for streamflow. Note that Stephens et al. (2020) and Duethmann et al. (2020) both point to poor representation of vegetation dynamics as the culprit for poor DSST performance. Fowler et al. (2022) also suggests that hydrological shifts are driven by depletion of storage caused by out-fluxes not reducing during the drought as much as the influxes.

As mentioned, 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). Despite this, GR4J's performance here was shown to be in line with that of other models. This is because 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 actually makes GR4J more susceptible to errors due to accumulation of moisture deficits over multiple annual periods (see Fig. 4).

Finally, the results of the annual analysis 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 can also be seen in Figure S6 and 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). Of the models tested here, only IHACRES contains a deficit (i.e. bottomless) moisture accounting system. This is likely responsible for substantially reducing the amount of year-to-year accumulation of error for both KGE and bias for this model compared to the others (see Fig. 4).

4.3 Post-drought recovery

This is the first study, to our knowledge, which studies model performance in the decade after the end of the Millennium drought. We showed that, according to most performance metrics, model performance does not recover after the end of the drought (Fig. 3). 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). It is worth noting that all metrics are calculated from a single model run, therefore it is possible that PostMD performance and its degradation patterns suffer from carryover of accumulated error from the MD years.

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 (Fig. 3) 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.

With regards to the annual regression analysis, we note that limitations to our methodology arise from the limitations of the regression model chosen (i.e. eq. 3). The choice of a simple dual-intercept model was motivated by ease of interpretability of the results and to limit the number of degrees of freedom of the model. The tests performed on the regression data and the regression residuals support the applicability of the model chosen, but the low values of the coefficient of determination indicate that this simple model only explains a small portion of the variance of annual model performance in most cases. Whereas the significance of the change in intercept is not conditioned on the predictive power of the model, low coefficients of determination require particular attention in the interpretation of results in individual cases. For this reason, in this study, we only comment on aggregated results from a qualitative perspective, looking at percentages of catchments where a shift is observed and observing average patterns of behaviour rather than trying to quantify them in each individual catchment. The relationship between model reliability and climate is a matter of increasing interest in the hydrological sciences, the approach taken here to study this relationship could certainly be expanded to allow for a more thorough investigation of this relationship and how it is affected by long-term climate anomalies such as the MD.

Similarly, future research should focus on understanding the roles of catchment and climate characteristics in determining model reliability during persistent changes in climate such as the MD. Such an analysis would be complementary to the one presented here, intended to compare changes in performance between metrics rather than between catchments. The identification of local drivers of model performance degradation could shed additional light on how hydrological systems reacted to the changes in climate brought in by the MD as well as on models' responses to them. Additionally, application of the methodology described here to additional datasets and/or 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 Fowler et al., 2022), 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. Our results show extensive degradation over the years of the drought, as well as after the drought, particularly driven by overestimation of flow volumes. Metrics associated with hydrograph shape and timing are not only much more resilient to the drought in absolute term, but are in most cases not significantly more degraded during the multiple subsequent years of the drought than the are on isolated dry years before the onset of the drought. Conversely, the overestimation of streamflow mentioned above suffers from error accumulation from year to year and significantly worsen during multi-annual dry spells. We also demonstrate that models' performance deficits persist, almost always identically, in the decade after the end of the drought.

Much of the novelty of this studies lies in its methodology, which is designed to (1) provide an unbiased estimate of model performance degradation by carefully partitioning the available data to enable benchmarking of model performance values; (2) fairly compare amounts of degradation across many different metrics, by using non-parametric statistics based on ranks rather than absolute values; and (3) distinguish the effects of individual drier years from those of the multi-annual event, by analysis annual performance values before, during and after the drought. Additionally, we for the first time study the performance of models in the post-drought recovery period. Thanks to such rigorous methodology and the use of a wide set of performance metrics able to capture various aspects of catchments' regime, the analysis presented here provides substantial more detail than more traditional DSST studies, in particular in terms of interpretation of results in the context of process explanation and representation.

In particular, we identify a deficiency of models in reproducing mechanisms to delay and/or remove water from the system before it reaches the stream as well as in keeping track of moisture deficits over multiple subsequent dry seasons. With regards to increasing our understanding of hydrological processes during long-term drought, we show that retention times and rates of storage depletion are not significantly affected by the drought, while streamflow amounts are. This suggests an imbalance between the decrease of influxes during the drought (precipitation) and that of out-fluxes (ET and inter/intra-basin exchanges). 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 and after extended drought.

Open Research

Model input data is described by Peterson et al. (2021) 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 S4 is currently stored at `https://cloudstor.aarnet.edu.au/plus/s/m1HjUbNPutHOw07`. These are temporary locations for the purpose of peer review, both datasets will be uploaded to an appropriate repository and shared via a DOI before acceptance and publication of this article. The version of MARRMoT used for this study is described by Trotter et al. (2022) and stored with DOI 10.5281/zenodo.6484372 (Trotter & Knoben, 2022).

Acknowledgments

This study received support from the Australian Research Council via Linkage Project LP180100796 *Observed streamflow generation changes: better understanding and mod-*

elling, the Victorian Government Department of Environment, Land, Water and Planning, and Melbourne Water. KF also acknowledges support from LP170100598. The authors also acknowledge the contributions of the Editor, Charles Luce, the Associate Editor and three anonymous reviewers thanks to whose comments the article is much improved.

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

Ben-Shachar, M. S., Ldecke, D., & Makowski, D. (2020). effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software*, *5*(56), 2815. Retrieved from `https://doi.org/10.21105/joss.02815` doi: 10.21105/joss.02815

Beven, K. (2019). How to make advances in hydrological modelling. *Hydrology Research*, *50*(6), 1481–1494. (Publisher: IWA Publishing) doi: 10.2166/nh.2019.134

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

Brigode, P., Oudin, L., & Perrin, C. (2013, January). Hydrological model parameter instability: A source of additional uncertainty in estimating the hydrological impacts of climate change? *Journal of Hydrology*, *476*, 410–425. Retrieved 2022-05-19, from `https://linkinghub.elsevier.com/retrieve/pii/S002216941200964X` doi: 10.1016/j.jhydrol.2012.11.012

Broderick, C., Matthews, T., Wilby, R. L., Bastola, S., & Murphy, C. (2016). Transferability of hydrological models and ensemble averaging methods between contrasting climatic periods. *Water Resources Research*, *52*(10), 8343–8373. Retrieved 2022-05-20, from `https://onlinelibrary.wiley.com/doi/abs/10.1002/2016WR018850` (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/2016WR018850) doi: 10.1002/2016WR018850

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.

Cureton, E. E. (1956). Rank-biserial Correlation. *Psychometrika*, *21*(3), 287–290.

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 rainfallrunoff 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.

Duethmann, D., Blschl, G., & Parajka, J. (2020, July). Why does a conceptual hydrological model fail to correctly predict discharge changes in response to climate change? *Hydrology and Earth System Sciences*, *24*(7), 3493–3511. Retrieved 2022-05-19, from `https://hess.copernicus.org/articles/24/3493/2020/` (Publisher: Copernicus GmbH) doi: 10.5194/hess-24-3493-2020

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 rainfallrunoff models lack long, slow dynamics: implications for runoff projections. *Water Resources Research*, *56*(5). doi: 10.1029/2019wr025286

Fowler, K. J. A., Peel, M., Saft, M., Peterson, T., Western, A., Band, L., … Nathan, R. (2022). Explaining changes in rainfall-runoff relationships during and after australia's millennium drought: a community perspective. *Hydrology and Earth System Sciences Discussions*, *2022*, 1–56. Retrieved from `https://hess.copernicus.org/preprints/hess-2022-147/` doi: 10.5194/hess-2022-147

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 (20102018): 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

Hanel, M., Rakovec, O., Markonis, Y., Mca, P., Samaniego, L., Kysel, J., & Kumar, R. (2018, December). Revisiting the recent European droughts from a long-term perspective. *Scientific Reports*, *8*(1), 9499. Retrieved 2022-04-21, from `http://www.nature.com/articles/s41598-018-27464-4` doi: 10.1038/s41598-018-27464-4

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`

Kleme, V. (1986). Operational testing of hydrological simulation models. *Hydrological Sciences Journal*, *31*(1), 13–24. doi: 10.1080/02626668609491024

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

Koffler, D., Gauster, T., & Laaha, G. (2016). lfstat: Calculation of low flow statistics for daily stream flow data [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=lfstat` (R package version 0.9.4)

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

Massari, C., Avanzi, F., Bruno, G., Gabellani, S., Penna, D., & Camici, S. (2022, March). Evaporation enhancement drives the European water-budget deficit during multi-year droughts. *Hydrology and Earth System Sciences*, *26*(6), 1527–1543. Retrieved 2022-04-21, from `https://hess.copernicus.org/articles/26/1527/2022/` doi: 10.5194/hess-26-1527-2022

Mathevet, T., Gupta, H., Perrin, C., Andrassian, V., & Le Moine, N. (2020, June). Assessing the performance and robustness of two conceptual rainfall-runoff models on a worldwide sample of watersheds. *Journal of Hydrology*, *585*, 124698. Retrieved 2022-05-20, from `https://linkinghub.elsevier.com/retrieve/pii/S002216942030158X` doi: 10.1016/j.jhydrol.2020.124698

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

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

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

Moeck, C., von Freyberg, J., & Schirmer, M. (2018, May). Groundwater recharge predictions in contrasted climate: The effect of model complexity and calibration period on recharge rates. *Environmental Modelling & Software*, *103*, 74–89. Retrieved 2022-05-19, from https://linkinghub.elsevier.com/retrieve/pii/S1364815216309458 doi: 10.1016/j.envsoft.2018.02.005

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

Motavita, D. F., Chow, R., Guthke, A., & Nowak, W. (2019). The comprehensive differential split-sample test: A stress-test for hydrological model robustness under climate variability. *Journal of Hydrology*, *573*(March), 501–515. Retrieved from https://doi.org/10.1016/j.jhydrol.2019.03.054 (Publisher: Elsevier) doi: 10.1016/j.jhydrol.2019.03.054

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

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., … Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929. Retrieved 2022-04-21, from `https://onlinelibrary.wiley.com/doi/abs/10.1111/ecog.02881` (_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecog.02881) doi: 10.1111/ecog.02881

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

Royer-Gaspard, P., Andrassian, V., & Thirel, G. (2021, November). Technical note: PMR a proxy metric to assess hydrological model robustness in a changing climate. *Hydrology and Earth System Sciences*, *25*(11), 5703–5716. Retrieved 2022-04-21, from `https://hess.copernicus.org/articles/25/5703/2021/` (Publisher: Copernicus GmbH) doi: 10.5194/hess-25-5703-2021

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

Seiller, G., Anctil, F., & Perrin, C. (2012, April). Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions. *Hydrology and Earth System Sciences*, *16*(4), 1171–1189. Retrieved 2022-05-19, from https://hess.copernicus.org/articles/16/1171/2012/ doi: 10.5194/hess-16-1171-2012

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

Smith, K. A., Barker, L. J., Tanguy, M., Parry, S., Harrigan, S., Legg, T. P., ... Hannaford, J. (2019, August). A multi-objective ensemble approach to hydrological modelling in the UK: an application to historic drought reconstruction. *Hydrology and Earth System Sciences*, *23*(8), 3247–3268. Retrieved 2022-04-21, from https://hess.copernicus.org/articles/23/3247/2019/ doi: 10.5194/hess-23-3247-2019

Stephens, C. M., Marshall, L. A., Johnson, F. M., Lin, L., Band, L. E., & Ajami, H. (2020). Is Past Variability a Suitable Proxy for Future Change? A Virtual Catchment Experiment. *Water Resources Research*, *56*(2), 1–25. doi: 10.1029/2019WR026275

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.

Thirel, G., Andrassian, V., Perrin, C., Audouy, J. N., Berthet, L., Edwards, P., ...

Vaze, J. (2015, August). Hydrology under change: an evaluation protocol to investigate how hydrological models deal with changing catchments. *Hydrological Sciences Journal*, *60*(7-8), 1184–1199. (Publisher: Taylor and Francis Ltd.) doi: 10.1080/02626667.2014.967248

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. (2022). *MARRMoT v2.1.* Zenodo. doi: 10.5281/zenodo.6484372

Trotter, L., Knoben, W. J. M., Fowler, K. J. A., Saft, M., & Peel, M. C. (2022). Modular Assessment of RainfallRunoff Models Toolbox (MARRMoT) v2.1: an object-oriented implementation of 47 established hydrological models for improved speed and readability. *Geoscientific Model Development*, *15*(16), 6359–6369. doi: 10.5194/gmd-15-6359-2022

Trotter, L., Saft, M., Peel, M. C., & Fowler, K. J. A. (2021). "Naïve" inclusion of diverse climates in calibration is not sufficient to improve model reliability under future climate uncertainty. In *MODSIM2021, 24th International Congress on Modelling and Simulation.* (pp. 588–594). Modelling and Simulation Society of Australia and New Zealand. doi: 10.36334/modsim.2021.J8.trotter

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 / modsim 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