Hydrological Sciences Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/thsj20

Exploring parameter (dis)agreement due to calibration metric selection in conceptual rainfall–runoff models

Eduardo Muñoz-Castro, Pablo A. Mendoza, Nicolás Vásquez & Ximena Vargas

To cite this article: Eduardo Muñoz-Castro, Pablo A. Mendoza, Nicolás Vásquez & Ximena Vargas (2023): Exploring parameter (dis)agreement due to calibration metric selection in conceptual rainfall–runoff models, Hydrological Sciences Journal, DOI: 10.1080/02626667.2023.2231434

To link to this article: https://doi.org/10.1080/02626667.2023.2231434

Published online: 04 Aug 2023.

HYDROLOGICAL SCIENCES JOURNAL
https://doi.org/10.1080/02626667.2023.2231434

Exploring parameter (dis)agreement due to calibration metric selection in conceptual rainfall–runoff models

Eduardo Muñoz-Castro [a], Pablo A. Mendoza [a,b], Nicolás Vásquez [a] and Ximena Vargas [a]

[a]Department of Civil Engineer, Universidad de Chile, Santiago, Chile; [b]Advanced Mining Technology Center (AMTC), Universidad de Chile, Santiago, Chile

ABSTRACT

We examine the extent to which the parameters of different types of catchments are sensitive to calibration criteria selection (i.e. parameter agreement), and explore possible connections with overall model performance and model complexity. To this end, we calibrate the lumped GR4J, GR5J and GR6J hydrological models – coupled with the CemaNeige snow module – in 95 catchments spanning a myriad of hydroclimatic and physiographic characteristics across Chile, using 12 streamflow-oriented objective functions. The results show that (i) the choice of objective function has smaller effects on parameter values in catchments with low aridity index and high mean annual runoff ratio, in contrast to drier climates; and (ii) catchments with better parameter agreement also provide better performance across model structures and simulation periods. More generally, this work provides insights on the type of catchments where it is more challenging to find sub-domains in the parameter space that satisfy multiple streamflow criteria.

ARTICLE HISTORY
Received 28 September 2022
Accepted 30 May 2023

EDITOR
S. Archfield

ASSOCIATE EDITOR
(not assigned)

KEYWORDS
conceptual hydrological models; calibration objective functions; large-sample hydrology; catchment attributes; parameter agreement

1 Introduction

Hydrological models are useful tools that support decision making in water resources applications, including flood design (Boughton and Droop 2003, Newman *et al.* 2021), hydrological forecasting (Mendoza *et al.* 2012, Rakovec *et al.* 2015, Wanders *et al.* 2019), and climate change impacts on water resources (Driessen *et al.* 2010, Addor *et al.* 2014, Chegwidden *et al.* 2019). In particular, conceptual rainfall–runoff models have been widely used because of their lower data requirements and computational cost compared to more complex alternatives (Knoben *et al.* 2019). The application of these models has typically relied on the adjustment (i.e. calibration) of "free" parameters (Yapo *et al.* 1998, Vrugt *et al.* 2003b), a problem that has been challenging the hydrology community for decades. Nevertheless, in recent years we have seen tremendous advances towards new strategies advocating for more realistic process representations (Yilmaz *et al.* 2008, Shafii and Tolson 2015, Nijzink *et al.* 2018, Konapala *et al.* 2020).

The identification of an adequate set of hydrological model parameters involves several methodological choices, whose subjective nature and implications have been widely recognized (e.g. Diskin and Simon 1977, Green and Stephenson 1986, Oudin *et al.* 2006, Melsen *et al.* 2019). Among these decisions, the selection of the calibration objective function(s) is critical, since it defines the target variables and/or processes that need to be well reproduced (Pushpalatha *et al.* 2012, Pool *et al.* 2017, Nemri and Kinnard 2020, Sepúlveda *et al.* 2022).

For the case of streamflow, several calibration metrics have been formulated to achieve good performance along the time domain and/or the signature domain (Fenicia *et al.* 2018). Time-domain metrics provide a direct contrast between time series of simulations and observations, and their mathematical formulation determines which parts of the hydrograph (i.e. which processes) are given more weight (Yapo *et al.* 1998, Boyle *et al.* 2000, Vrugt *et al.* 2003a). Among time-domain metrics, least squares formulas like the Nash-Sutcliffe efficiency (NSE; Nash and Sutcliffe 1970) and variants (e.g. Kling-Gupta efficiency, KGE; Gupta *et al.* 2009) are popular choices for which variations based on transformations have been also proposed (Chiew *et al.* 1993, Santos *et al.* 2018). On the other hand, signature-domain performance metrics result from contrasting hydrological signatures – i.e. metrics that quantify streamflow properties (e.g. runoff ratio, the slope of the flow duration curve, streamflow elasticity; see review by McMillan 2021) – from observed and modelled streamflow data. The last few years have seen a proliferation of studies proposing calibration metrics that combine hydrological signatures (Shafii and Tolson 2015, Beck *et al.* 2016, Pool *et al.* 2017, Yang *et al.* 2019) to obtain parameter sets that ensure the "hydrological consistency" (Martinez and Gupta 2011) of model simulations.

Despite the large number of calibration metrics proposed and the development of multi-objective optimization algorithms for hydrological modelling (Yapo *et al.* 1998, Boyle *et al.* 2000, Vrugt *et al.* 2003a, Vrugt and Robinson 2007), the selection of objective function(s) is still a topic of active research and debate, partly because water managers typically seek to achieve model accuracy in specific streamflow properties (Mizukami *et al.* 2019). Hence, a large body of work has investigated the implications of calibration criteria selection on different aspects, including parameter identifiability (e.g. Pechlivanidis *et al.* 2014), simulation of flood hydrographs (e.g. Servat and Dezetter 1991), drought

CONTACT Pablo A. Mendoza pamendoz@uchile.cl Department of Civil Engineer, Universidad de Chile, Av. Blanco Encalada 2002, Santiago 8370449, Chile

characteristics (e.g. Melsen *et al.* 2019), annual peak flow biases (e.g. Mizukami *et al.* 2019), streamflow characteristics (e.g. Pool *et al.* 2017, 2018), spatial patterns in model states and fluxes (e.g. Dembélé *et al.* 2020), and projected hydrological changes (e.g. Najafi *et al.* 2011, Mendoza *et al.* 2016, Seiller *et al.* 2017).

In spite of the awareness that the parameters of conceptual rainfall–runoff models are inherently related to hydrological processes (Guse *et al.* 2017), only a few studies have investigated the variability in parameter values due to calibration criteria selection. Diskin and Simon (1977) showed, for the Ekron watershed in Israel, that the optimal parameters of a four-parameter conceptual hydrological model may have considerable variations depending on the choice of calibration objective function. Abdulla *et al.* (1999) compared parameter values in the ARNO model obtained with four calibration metrics for baseflow errors, using data from 24 basins; they reported large variations in the maximum soil moisture and the moisture content threshold parameter. Gupta *et al.* (2009) compared optimal parameter values of the HBV model obtained from calibrations with NSE and KGE in 49 catchments in Austria, finding that, in only a few basins, two or more parameters (out of six) suffered large variations. Muleta (2012) assessed the ability of nine calibration objective functions to produce robust streamflow simulations across five gauges in the Little River experimental watershed (USA). From their calibration experiments with the Soil and Water Assessment Tool (SWAT) model and the dynamically dimensioned search algorithm (DDS; Tolson and Shoemaker 2007), they found that the objective function affected the optimal parameter values, even when streamflow simulated with a set of known parameters was used as the "truth." Wu and Liu (2014) also reported large variations in six parameters within the SWAT model in one case study basin in China, contrasting results from two objective functions based on square errors, and two calibration metrics based on absolute errors. Garcia *et al.* (2017) compared GR4J parameter values obtained from two different objective functions for the streamflow Q – KGE ($Q^{0.5}$) vs. one combining KGE(Q) with KGE(1/Q) – across 691 catchments in France, finding large differences in X1 (capacity of production store) and X2 (groundwater exchange coefficient). They also found large differences in the X4 (unit hydrograph time base) parameter across catchments where groundwater contributions to total runoff are more important. More recently, Song *et al.* (2019) showed that the choice of objective function – NSE($Q^{0.5}$) vs. KGE($Q^{0.5}$) – produced large variations in some parameters contained in two conceptual rainfall–runoff models applied in 41 catchments across South Korea.

Although some of the aforementioned studies (Abdulla *et al.* 1999, Gupta *et al.* 2009, Garcia *et al.* 2017, Song *et al.* 2019) used a large number (>20) of catchments, several questions remain unanswered about the (dis)agreement of optimal parameter values given the choice of calibration metric (hereafter, parameter agreement), including:

(1) Are there connections between parameter agreement and catchment characteristics?
(2) What is the role of increasing model complexity in parameter agreement?
(3) Is hydrological model performance related to parameter agreement?

To address the above questions, we calibrate the parameters of three GRXJ rainfall–runoff models (Perrin *et al.* 2003, Pushpalatha *et al.* 2011) – where "X" denotes the number of parameters involved – coupled to the snowmelt and accumulation model CemaNeige (Valéry *et al.* 2014a, 2014b), using a global optimization algorithm and 12 different objective functions. The calibration results are used to compute a parameter agreement index that quantifies parameter variability arising from the choice of calibration metric. To explore possible connections between parameter agreement and climatic or physiographic characteristics, we conduct calibration experiments in 95 catchments along continental Chile (17–57°S) and perform a clustering step based on agreement results from each model structure to identify groups with good or bad parameter agreement indices. Finally, we examine whether high parameter agreement relates to overall model performance during the calibration period and two independent verification periods. Our study contributes to the existing literature by (1) providing detailed insights about the impacts of calibration metric selection on parameter values of conceptual hydrological models when the calibration problem is addressed from a single-objective perspective, and (2) providing guidance on the catchments where it is more challenging to find regions within the parameter space that satisfy multiple criteria (i.e. to find parameter sets that can be used for a myriad water resources applications).

2 Study domain and datasets

2.1 Catchment selection

Our study domain includes several catchments located in continental Chile (Fig. 1), which spans a diverse range of physiographic (i.e. topography, geology, soil types, land cover) and hydroclimatic characteristics. In particular, we select a suite of basins that meet the following requirements: (i) a low human intervention degree index (i.e. < 0.05), which is defined as the ratio between annual flow of surface water rights (consumptive permanent continuous), and the mean annual runoff measured at the catchment outlet (Alvarez-Garreton *et al.* 2018); (ii) the absence of large reservoirs and non-consumptive water withdrawals, unless their restitution is located upstream of the streamflow gauge; (iii) at least 40% of days with streamflow observations during the period 1985–2005 (which is used for model calibration; see details in Section 3); and (iv) at least 20% of daily streamflow observations in the evaluation periods 2005–2010 and 2010–2017. Importantly, the fractional area covered by impervious surfaces – which are typically associated with urbanized areas – ranges between 0 and 1.4% (obtained from Alvarez-Garreton *et al.* 2018) across our sample of basins.

The resulting 95 near-natural catchments (Fig. 1) reflect the spatial variability of hydroclimatic conditions across continental Chile. For example, we note a transition from negligible precipitation amounts in the Far North, increasing towards Southern Chile, and lower precipitation in Patagonia. Additionally, there is a marked precipitation

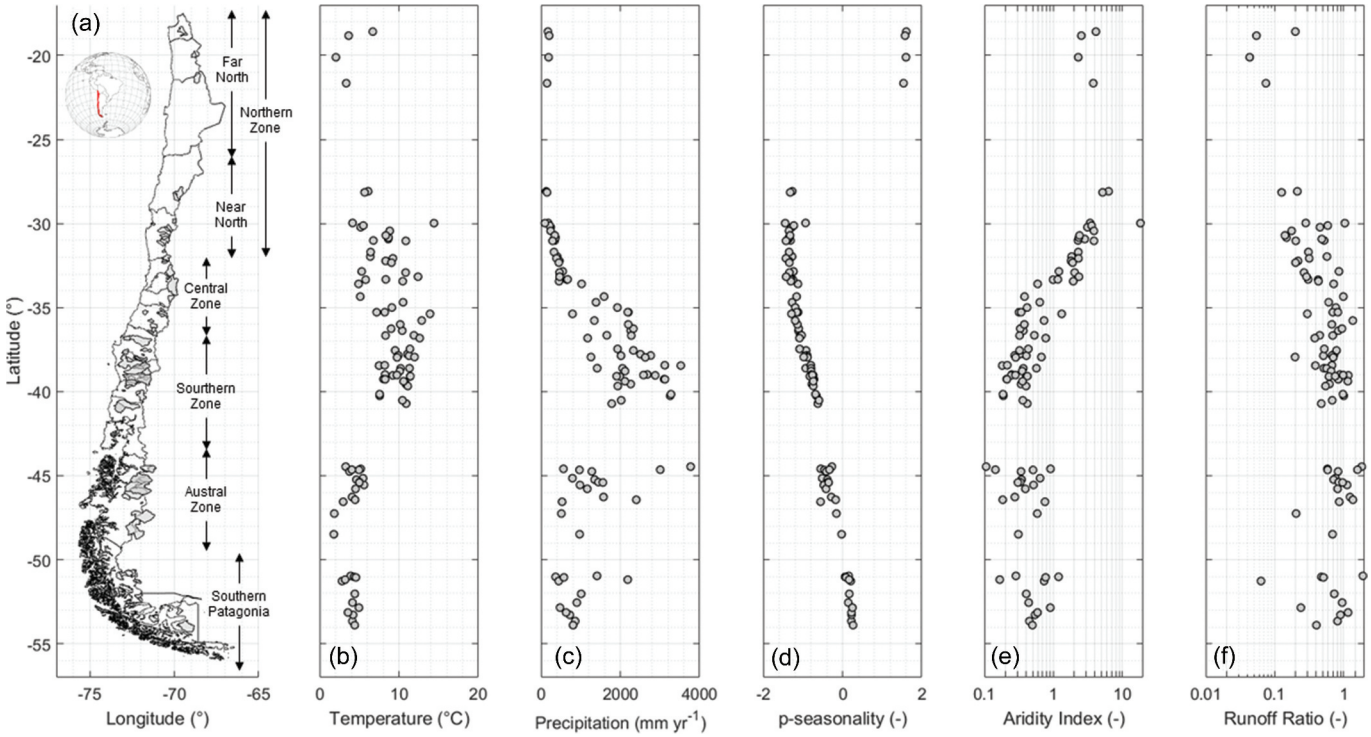


Figure 1. (a) Study domain and mean hydroclimatic characteristics over the period 1985–2015, including (b) mean annual temperature, (c) mean annual precipitation, (d) precipitation seasonality, obtained from CAMELS-CL, (e) aridity index, and (f) mean annual runoff ratio.

seasonality index (i.e. seasonality and timing of precipitation; see p-seasonality description in Table 1) in Northern and Central Chile, with peaks during Southern Hemisphere summer (January–March) and winter (July–September), respectively, unlike the Austral Zone and Southern Patagonia where the distribution of precipitation is uniform across the year (p-seasonality close to zero); a transition from very large to small aridity indices from North to South; and increasing mean annual runoff ratios towards the south (for more detailed discussion, the reader is referred to Alvarez-Garreton *et al.* 2018). Such diversity yields large spatial heterogeneities in terms of hydrological responses (Vásquez *et al.* 2021), making our sample of catchments suitable for detailed examination of possible dependencies between hydrological modelling results and catchment descriptors.

Table 1. Physiographic and climatic catchment attributes used in this study (derived from Alvarez-Garreton *et al.* 2018).

Attribute class	Attribute name	Unit	Description
Location and topography	Latitude	°	Gauge latitude (based on DGA records).
	Area	km^2	Catchment area.
	Mean elevation	m a.s.l.	Catchment mean elevation.
	Slope	m/km	Catchment mean slope.
	Elevation range	m	Difference between catchment maximum and minimum elevation.
Land cover characteristics	Forest fraction	%	Percentage of the catchment covered by forest, including native forest and forest plantations.
	Barren soil	%	Percentage of the catchment covered by barren lands.
Climatic indices (computed for 1 April 1985–31 March 2015)	Precipitation seasonality* (p-seasonality)	-	Seasonality and timing of precipitation, estimated using sine curves to represent the annual temperature and precipitation cycles; positive (negative) values indicate precipitations peaks in summer (winter); values close to 0 indicate uniform precipitation throughout the year.
	Precipitation	mm/y	Average annual precipitation at catchment scale derived from CR2Met v. 2.0.
	Temperature	℃	Average annual mean temperature at catchment scale derived from CR2Met v. 2.0.
	Potential evapotranspiration	mm/y	Average annual potential evapotranspiration (PET) at catchment scale computed using Oudin's method.
	Aridity index	-	Aridity, calculated as the ratio of average annual PET to average annual precipitation in the catchment.
Hydrological indices (computed for 1 April 1985–31 March 2015)	Runoff	mm/y	Average annual runoff at the catchment outlet (based on DGA records).
	Runoff ratio	-	Calculated as the ratio of average annual runoff and average annual precipitation in the catchment.

*p-seasonality is retrieved directly from the CAMELS-CL dataset (i.e. not re-computed for the period 1985–2015).

2.2 Hydrometeorological time series

Streamflow time series are acquired from stations maintained by the Chilean Water Directorate (DGA). This information is public and free, and can be retrieved directly from the DGA's web platform (https://dga.mop.gob.cl/), from databases such as the Chilean Climate Explorer (http://explorador.cr2.cl/) or the Catchment Attributes and Meteorology for Large-Sample Studies, Chile (CAMELS-CL) explorer (http://camels.cr2.cl/), maintained by the Center for Climate and Resilience Research (CR2). Daily time series of catchment-scale precipitation and air temperature are derived from the gridded observational product CR2Met (DGA 2017, Boisier *et al.* 2018), whose most recent version (v. 2.0) covers continental Chile for the period 1979–2020 at a 0.05° × 0.05° horizontal resolution. CR2Met precipitation estimates are produced through a statistical modelling framework that uses topographic descriptors and large-scale variables – such as water vapor fluxes and moisture fluxes – from European Centre for Medium-Range Weather Forecasts's (ECMWF) atmospheric reanalysis-Interim (Dee *et al.* 2011) in previous versions and ERA5 (C3S 2017) in the latest – as predictors, and daily precipitation data from stations as predictands. A similar approach is used to generate daily maximum and minimum temperature time series, including additional predictors from Moderate Resolution Imaging Spectroradiometer (MODIS) land-surface products to account for spatial heterogeneities (e.g. differences in land cover types). The reader is referred to DGA (2017, 2018, 2019) for more details on the development of CR2Met.

2.3 Ancillary data

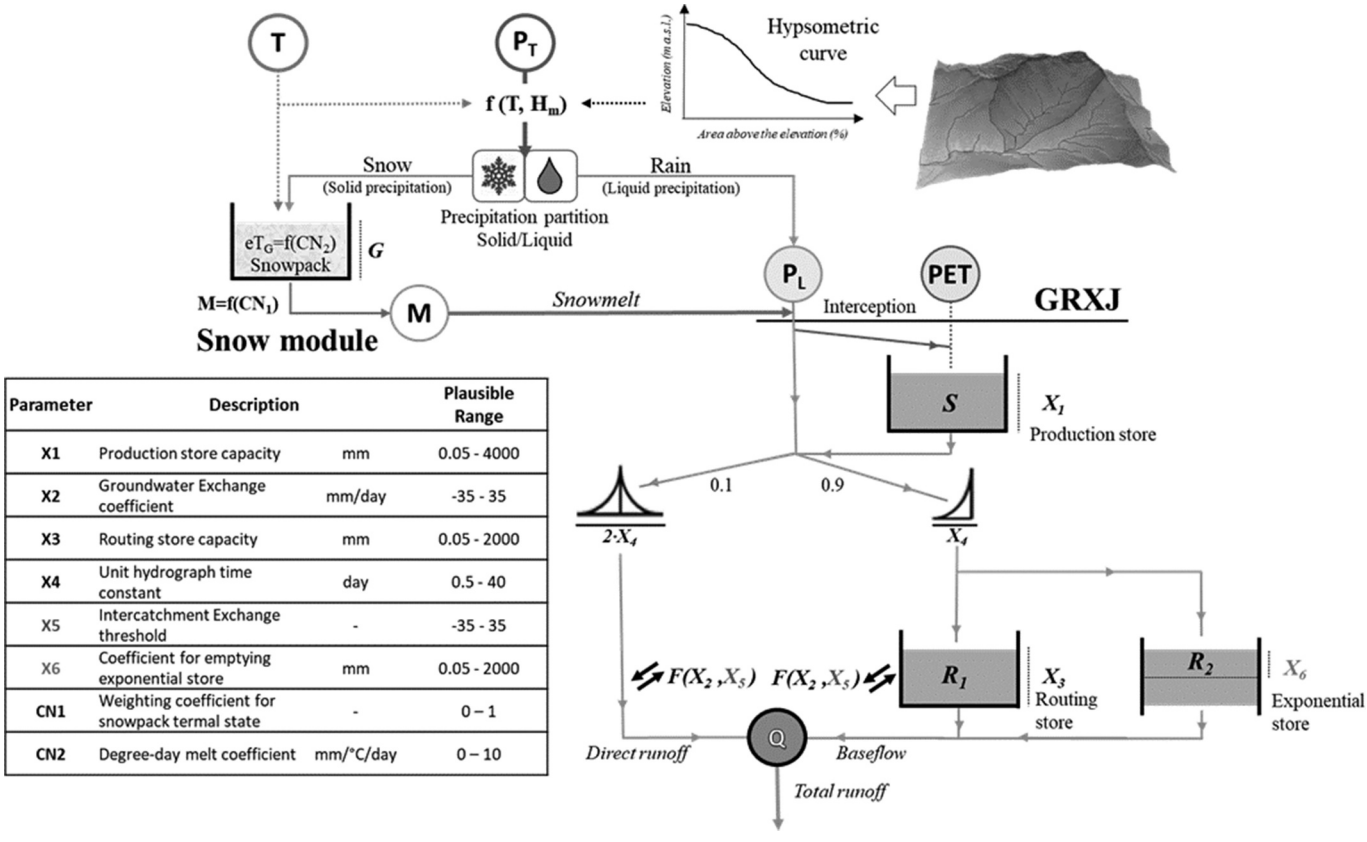
We acquired and processed digital elevation models (DEMs) from the Shuttle Radar Topography Mission (SRTM; Rabus *et al.* 2003) at a 3 arc-second horizontal resolution (approximately 90 m) to obtain hypsometric curves, which are used to configure the representation for each basin in the snow module through elevation bands (see section 3.1). Additionally, catchment boundaries and a suite of physiographic and climatic attributes (Table 1) were obtained from the CAMELS-CL dataset (Alvarez-Garreton *et al.* 2018).

3 Approach

3.1 Hydrological models

We use the GR4J (Perrin *et al.* 2003), GR5J and GR6J (Pushpalatha *et al.* 2011) lumped hydrological models, coupled to the snow accumulation and ablation module CemaNeige (Valéry *et al.* 2014a, 2014b). For simplicity, we refer to these models as GRXJCN (i.e. GRXJ + CemaNeige; Fig. 2). All these models and other utilities are available through the airGR package (Coron *et al.* 2017, Laurent 2020). GRXJ models are conceptual, bucket-style precipitation-runoff models, which provide simplified representations of hydrological processes at the catchment scale, and only require precipitation (P), potential evapotranspiration (PET) and mean temperature (T) daily time series to run. In this study, we estimate PET using the formulae proposed by Oudin *et al.* (2005).

The CemaNeige snow module simulates the snowpack accumulation and melting processes through a two-parameter



Parameter	Description		Plausible Range
X1	Production store capacity	mm	0.05 - 4000
X2	Groundwater Exchange coefficient	mm/day	-35 - 35
X3	Routing store capacity	mm	0.05 - 2000
X4	Unit hydrograph time constant	day	0.5 - 40
X5	Intercatchment Exchange threshold	-	-35 - 35
X6	Coefficient for emptying exponential store	mm	0.05 - 2000
CN1	Weighting coefficient for snowpack termal state	-	0 – 1
CN2	Degree-day melt coefficient	mm/°C/day	0 – 10

Figure 2. Scheme with the model structures, description of parameters and plausible calibration ranges.

degree-day factor approach (DDF; DeWalle and Rango 2008), and also requires catchment-scale daily time series of precipitation and temperature. Additionally, the snow module offers the option to discretize each catchment into elevation bands of equal area, based on the hypsometric curve. If elevation bands are used, precipitation and air temperature are internally extrapolated using orographic gradients defined by Valéry *et al.* (2010), and the partitioning of precipitation into rainfall and snowfall is also internally estimated as a function of band-averaged temperature and elevation following L'hôte *et al.* (2005).

3.2 Individual basin calibration

Hydrological model parameters are calibrated using the Shuffled Complex Evolution (SCE-UA; Duan *et al.* 1992) global optimization algorithm, implemented in the R package hydromad (Andrews *et al.* 2011). We define a 5-year warm-up period (1 January 1980–31 March 1985) before computing the objective function with data available for the period 1 April 1985–31 March 2005 (30 years). For each basin and GRXJCN model, we conduct 12 calibrations using the objective functions listed in Table 2, computed with daily observed and simulated runoff time series. We select these objective functions because they belong to different families of metrics and have been widely used for various modelling purposes in the hydrology community. For instance, the NSE with flows in log space (log-NSE) has been used to enhance baseflow simulations, while the recently proposed split-KGE (Fowler *et al.* 2018) aims to provide good performance in terms of streamflow timing, volume and variability under contrasting climatic conditions.

To account for parameter equifinality, we follow a similar strategy to Nemri and Kinnard (2020): all the iterations made by SCE-UA until convergence are saved, and an ensemble of parameter sets is selected based on the difference between the associated objective function value and the global optimum. If such a difference is smaller than 0.001 (arbitrarily defined), the parameter set is saved for subsequent analyses. For example, if the optimal KGE value obtained from a KGE-based calibration with SCE-UA is 0.678, only the parameter sets with KGE ≥ 0.677 are retained. The process is repeated for each objective function in order to obtain a large ensemble of parameter sets (arising from equifinal solutions for several calibration metrics). If identical parameter sets that belong to different iterations are selected, these are merged to discard redundant information.

3.3 Parameter agreement analyses

3.3.1 Parameter agreement index

To quantify parameter agreement arising from the choice of calibration metric, we use the large ensemble of parameter sets to compute a modified version of the metric proposed by Zink *et al.* (2018) for each parameter and catchment:

$$R^j_{\theta_i} = 1 - \frac{\theta^j_{i,P95} - \theta^j_{i,P5}}{\theta_{i,max} - \theta_{i,min}} \quad (1)$$

where $R^j_{\theta_i}$ span values between 0 to 1, $\theta^j_{i,P5}$ and $\theta^j_{i,P95}$ are the 5th and 95th percentiles of parameter θ_i for catchment "j" (computed from the large ensemble of parameter sets), and $\theta_{i,min}$ and $\theta_{i,max}$ indicate the plausible range for parameter θ_i (Fig. 2). Therefore, (lower) higher values of $R^j_{\theta_i}$ indicate (dis)agreement in the values of θ_i parameter in catchment j.

Additionally, to obtain a summary metric, we estimate an overall parameter agreement index for each basin j (R^j_{OA}) as follows:

$$R^j_{OA} = \frac{\sqrt{\sum_{i=1}^{N_{par}} R^{j^2}_{\theta_i}}}{\sqrt{N_{par}}} \quad (2)$$

where $R^j_{\theta_i}$ is the agreement index for parameter θ_i in catchment j, and N_{par} is the number of model parameters, which varies between six (GR4JCN) and eight (GR6JCN). The interpretation of R^j_{OA} values is analogous to $R^j_{\theta_i}$ (i.e. 0/1 represent the lowest/highest parameter agreement).

To explore controls on individual parameter agreement across our study domain, we compute the Spearman's rank correlation coefficient (ρ) between parameter agreement indices ($R^j_{\theta_i}$) obtained from each model structure, and the physiographic and hydroclimatic catchment descriptors listed in Table 1. We select these attributes regardless of possible inter-dependencies, since they are widely used to characterize different aspects of geomorphology, land cover, climate and hydrology.

Table 2. Objective functions used to calibrate the GRXJCN hydrological models (based on Fowler *et al.* 2018).

Class	Class description and reason for application	Objective function (OF) selected to calibrate
Common approach	Functions where the main goal is to minimize the sum of squares between observations and simulations at each time step.	[1] Kling-Gupta efficiency (Gupta *et al.* 2009), [2] Nash-Sutcliffe efficiency (Nash and Sutcliffe 1970)
Transformations	Logarithmic, exponential (<1) or other transformations (e.g. Box-Cox) are applied to observed and simulated runoff to emphasize the weight of the comparison between higher and lower values, stabilizing the variance of the error.	[3] NSE-Log, [4] NSE($Q^{0.5}$), [5] KGE($Q^{0.5}$)
Absolute error	By not squaring the errors, with the goal of minimizing the sum of absolute errors, the analysis is emphasized in the middle and low values.	[6] Refined index of agreement (dt; Willmott *et al.* 2012)
Time-based meta-objective	A function to assess the model performance is applied in different sub-periods, and the results are subsequently combined (e.g. averaged, weighted) into a meta-objective function, reducing the inter-annual variability of model performance due to temporal instabilities in parameter values.	[7] Split KGE (Fowler *et al.* 2018), [8] Split KGE($Q^{0.5}$).
Meta-objective	Linear combination of different functions into a meta-objective function (i.e. implicit multi-objective). Each metric or index provides information about the model performance in different components (statistical or hydrological). The assumption is that more information could improve the inference of parameters.	[9] Zhang Efficiency (Zhang *et al.* 2008), [10] Aggregate Objective Function (Beck *et al.* 2016), [11] KGE+NSE-Log, [12] KGE($Q^{0.5}$)+ NSE($Q^{0.5}$)

3.3.2 Catchment grouping

To examine spatial patterns in parameter agreement across our study domain, the catchments are grouped in quartiles based on the overall parameter agreement indices (R^j_{OA}). To this end, the 25th, 50th and 75th percentiles of R^j_{OA} are used to define limits and group membership, with group 1 (group 4) containing the catchments with the highest (lowest) R^j_{OA} values – i.e. overall best (worst) parameter agreement. The process is repeated for each GRXJCN model, and the results are used to analyse possible differences between parameter agreement groups in terms of (i) hydroclimatic catchment characteristics and (ii) individual parameter agreement indices.

3.3.3 Connections between parameter agreement and simulated catchment response

We examine possible connections between parameter agreement groups and hydrological model performance – quantified by the metrics listed in Table 2. All performance metrics are computed for the calibration period and two independent verification periods: (i) 1 April 2005–31 March 2010, characterized by average conditions; and (ii) 1 April 2010–31 March 2017, with unprecedented dry conditions (Garreaud *et al.* 2017).

4 Results

4.1 Illustrating parameter (dis)agreement and simulated streamflow response

Figure 3 shows the streamflow response simulated with the large ensemble of parameter sets in three case study basins with different R_{OA} values. Figure 3(a) displays results for a rainfall-dominated catchment with high parameter agreement (except for snow parameters CN1 and CN2), where similar runoff time series, seasonal runoff and daily flow duration curves are retrieved. Conversely, Fig. 3(b) and (C) display results for a rainfall-dominated basin and a semi-arid snowmelt-driven basin, respectively, with higher parameter disagreement arising from the choice of calibration metric and equifinality. Interestingly, similar simulated responses are obtained for the Tolten River (Fig. 3(b)), a clear example of parameter equifinality, whereas large discrepancies are obtained for the Derecho Creek, especially in runoff seasonality. Even more, in the latter case the choice of calibration metric may yield a mismatch between simulated and observed hydrological regimes. However, the daily flow duration curves are reasonably well represented, and very high KGE values (>0.9) can be achieved in all cases.

These results suggest that the extent to which the choice of calibration metric affects model outputs may depend on (i) specific catchment characteristics, including dominant hydrological processes; and (ii) the specific modelling purpose(s). In the following subsections, we expand on these ideas to explore possible connections between parameter (dis)agreement, catchment characteristics, incremental model complexity and model performance.

4.2 Individual parameter agreement

Figure 4 displays the latitudinal distribution of parameter agreement indices across the study domain. Overall, the parameters related to snow processes (CN1 and CN2) show the poorest agreement (i.e. values close to zero), and R_{θ_i} values for parameters related to storages (i.e. X1, X3 and X6) span the entire plausible range (i.e. from 0 to 1). The highest R_{θ_i} values are obtained for parameters associated with groundwater exchange (X2 and X5). Figure 4 also shows that R_{X2} improves when X5 is added to the GR4JCN model structure, suggesting that when X5 is

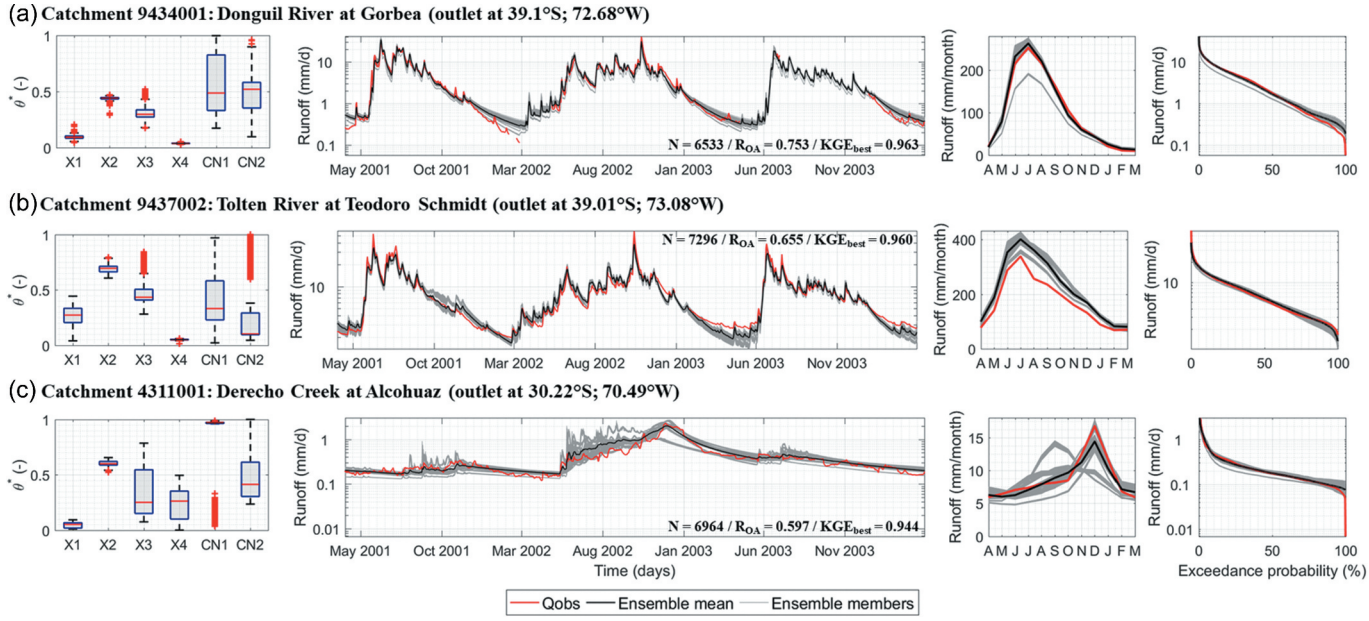


Figure 3. Illustration of parameter (dis)agreement arising from calibration metric selection and equifinality, along with associated streamflow responses in terms of daily runoff (in log space) time series, runoff seasonality and flow duration curves for the period 1985–2015) for three case study basins: (a) Donguil River at Gorbea, (b) Tolten River at Teodoro Schmidt, and (c) Derecho Creek at Alcohuaz. In the first column, each box plot contains normalized parameter values from the large ensemble (with size N), computed as $\theta_i^* = (\theta_i - \theta_{i,min})/(\theta_{i,max} - \theta_{i,min})$. The results were obtained with the GR4JCN model.

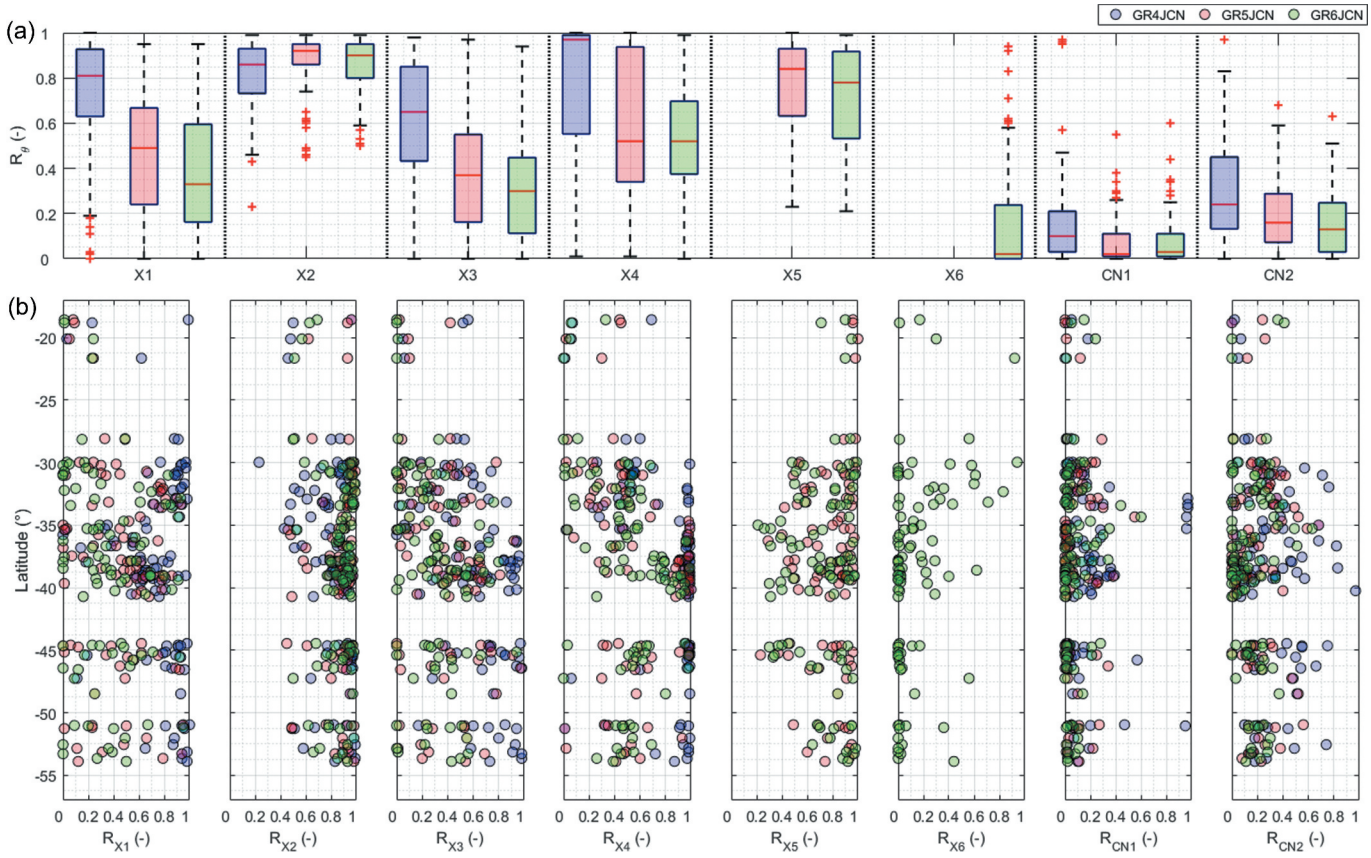


Figure 4. Effects of increased model complexity on parameter agreement in GR models. (a) Each box plot comprises agreement indices from 95 catchments, for a specific combination of parameter and model structure. (b) Variation of parameter agreement indices with latitude, where each point represents a catchment.

not included, X2 compensates for the absence (or deficiency) in the missing process in the model structure. Conversely, R_{X1} and R_{X3} progressively decrease when parameters X5 and X6 are successively added to the GR4JCN model structure. Regarding the parameter X4 (unit hydrograph time constant), the best agreement indices are achieved for GR4JCN but, in general, all GRXJCN models span the plausible range. No clear latitudinal gradients are found for R_{θ_i} values (bottom panels in Fig. 4).

The large dispersion in $R^j_{\theta_i}$ (Fig. 4) suggests that the effects of calibration metric selection and equifinality on parameter values may be related to specific catchment attributes. To explore this idea, Fig. 5 displays the Spearman's rank correlation coefficient (ρ) between R_{θ_i} and a suite of physiographic and hydroclimatic catchment descriptors (see details in Table 1). Statistical significance is indicated by bold circle outlines. Two important features are revealed: (i) statistically significant correlations exist between parameter agreement indices and some catchment descriptors, and (ii) modifications in model structure may switch the sign of ρ for specific combinations of parameters and catchment attributes. For example, $\rho(R_{X1}, P) = -0.35$ with GR4JCN, where P is the basin-averaged mean annual precipitation, suggesting that R_{X1} (i.e. agreement in X1 values) improves in drier basins when using GR4JCN. Conversely, $\rho(R_{X1}, P)$ reaches values of 0.19 and 0.30 with GR5JCN and GR6JCN, respectively (similar results for R_{X1} and runoff), which implies better agreement in X1 values (i.e. smaller effects of the choice of calibration metric and equifinality) across catchments with larger runoff production. Likewise, ρ (R_{X2}, P) = 0.24 with GR4JCN, switching to −0.43 and 0.03 with GR5JCN and GR6JCN, respectively, with a similar behaviour for R_{X2} and Runoff. In general, the type of dependence (i.e. the sign of ρ) between R_{X3}, R_{X4} and R_{X5} and catchment attributes does not change with model structure. Note that when X6 is added to GR5JCN, the correlation between R_{X2} and some catchment attributes decreases – e.g. mean annual runoff (from −0.43 to −0.02) and aridity index (from 0.43 to 0.01). Figure 5 also shows that more than half of the catchment descriptors (e.g. aridity index, mean annual runoff, mean annual precipitation, forest fraction) yield $|\rho| > 0.25$ when correlated against R_{X4}.

In terms of snow parameters, Fig. 5 shows that a large fraction of catchment attributes (85–100%, depending on the model structure) is weakly correlated (i.e. $|\rho| < 0.25$) with R_{CN1} (agreement index of the weighting coefficient for snowpack thermal state, CN1). On the other hand, R_{CN2} (agreement index for degree-day melt factor, CN2) correlates well with physiographic attributes like elevation (mean and range), slope, and barren soil fraction, and with climatic attributes like mean temperature and PET. In particular, R_{CN2} increases in snow-influenced catchments (see correlations with mean elevation and temperatures in Fig. 5). This is somehow expected, since snow module parameters become irrelevant in rainfall-dominated basins.

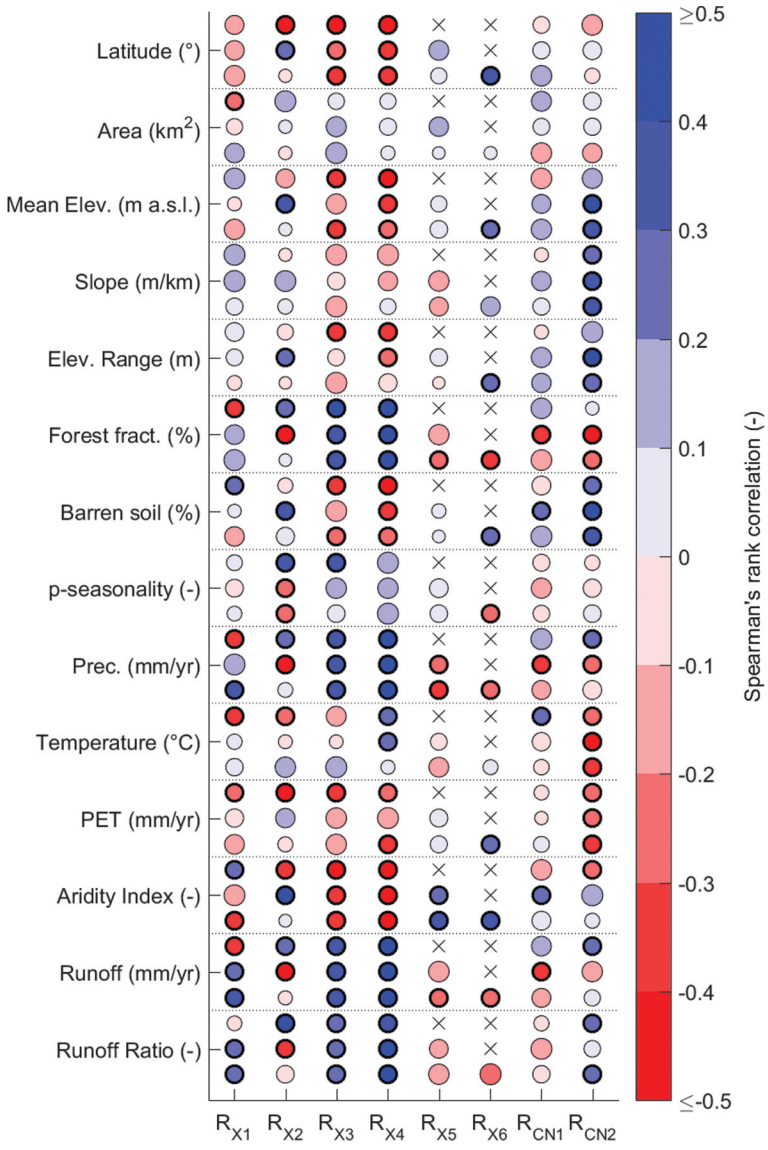


Figure 5. Spearman's rank correlation coefficient between catchment attributes (rows) and parameter agreement indices (columns). Each subpanel displays results for a specific catchment descriptor and the three coupled models from top to bottom: GR4JCN, GR5JCN and GR6JCN. The circles with thick outlines indicate statistically significant correlation coefficients at a 5% level.

4.3 Catchment grouping based on parameter agreement

Now we examine to what extent the choice of calibration objective function affects the agreement of parameter sets (quantified with R_{OA}) across the study domain. Figure 6 shows that parameter agreement groups are not clustered in space. Additionally, variations in model structure affect the membership of several catchments to parameter agreement groups, moving towards a better-ranking (e.g. 33.7% from GR4JCN to GR5JCN) or worse-ranking (e.g. 28.4% from GR4JCN to GR6JCN) group. For example, the northernmost basin in the domain (the San Jose River at Ausipar – BNA 1310002; outlet at 1245 m a.s.l. – 18.58°S, 69.81°O) is assigned to group No. 3 when the GR4JCN and GR5JCN model configurations are used, and to cluster No. 4 if GR6JCN is used. In agreement with the results presented in Fig. 5, Fig. 6 shows that R_{OA} generally improves in basins with larger annual precipitation amounts, lower aridity index, and higher runoff ratio if GR4JCN is used (i.e. parameter sets are less sensitive to calibration metric selection in humid regions); nevertheless, such a relationship is less clear as model complexity (i.e. number of parameters) increases, with reduced R_{OA} values.

Figure 7 displays, for each model structure (columns), individual parameter agreement indices R_{θ_i} (rows) stratified by R_{OA}-based catchment groups (displayed as box plots). Here, we seek to disentangle (1) whether our catchment grouping strategy is useful to discriminate agreement indices for individual parameter values – i.e. how does a specific parameter agreement index change from a "good" (i.e. 1) to "bad" (i.e. 4) cluster – and (2) how does increasing model complexity affect relative differences among groups. The results show that, in general, catchments classified as having "good" ("bad") agreement in parameter sets also hold high (low) values of R_{X1}, R_{X2}, R_{X3} and R_{X4}. Further, R_{θ_i} values do not necessarily improve when moving to groups with better ranking (e.g. X5, X6 and CN1 in all models). Finally, Fig. 7 also shows that

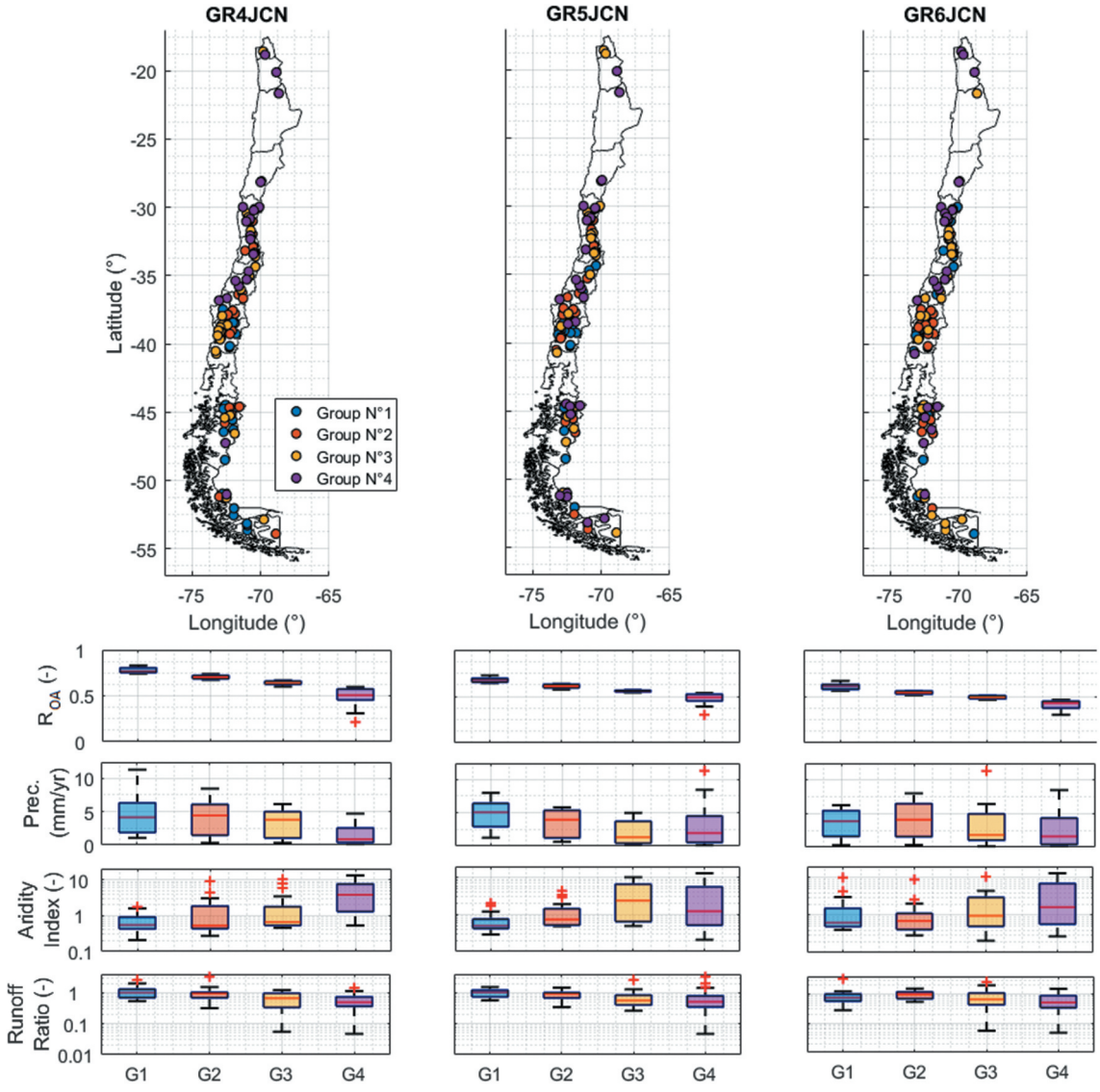


Figure 6. Agreement groups for GRXJCN models and catchment hydroclimatic attributes. Group 1 (G1) denotes the group of catchments with the best overall parameter agreement index, while Group 4 (G4) indicates the group of catchments with the overall worst parameter agreement.

increasing model complexity can degrade parameter agreement; for example, R_{X1} for group 1 spans 0.7–1.0 when using GR4JCN, with median of ~0.9, but it ranges from 0.25 to 1.0 in GR6JCN, with a median of ~0.7.

4.4 Parameter agreement and performance metrics

Each box plot in Fig. 8 displays performance metrics for the N_k catchments within group k (with k = 1,2,3,4). The metric value (e.g. KGE) for each basin is obtained as the median from the simulations conducted with the large ensemble of parameter sets (see section 3.2). The results show that, for most performance metrics, better results are obtained with GR4JCN in groups with good parameter agreement compared to GR5JCN and GR6JCN. A notable result from Fig. 8 is that, in contrast to other metrics, similar dt values can be obtained in catchments with different overall agreement (R_{OA}) levels. Although analogous efficiency variations among groups are observed with GR5JCN and GR6JCN, increasing model complexity may degrade model performance in catchments where parameter agreement is poor (e.g. KGE, KGE($Q^{0.5}$)).

5 Discussion

5.1 Parameter (dis)agreement and catchment characteristics

The results displayed in Fig. 4 partly agree with previous work on parameter identification in GR models. Garcia *et al.* (2017) used a large sample of catchments (691) in France and two objective functions in their analyses, concluding that the choice of calibration metric for the GR4JCN model structure has smaller effects on parameters X3 and X4 compared to the rest (excepting catchments with high baseflow index). Nemri and Kinnard (2020) tested different calibration strategies – including search algorithm and objective criteria – using the GR4JCN model in 12 snow-dominated basins in Canada, finding a better agreement in parameter values for X2 and X4. Our study shows better parameter agreement in X2 across

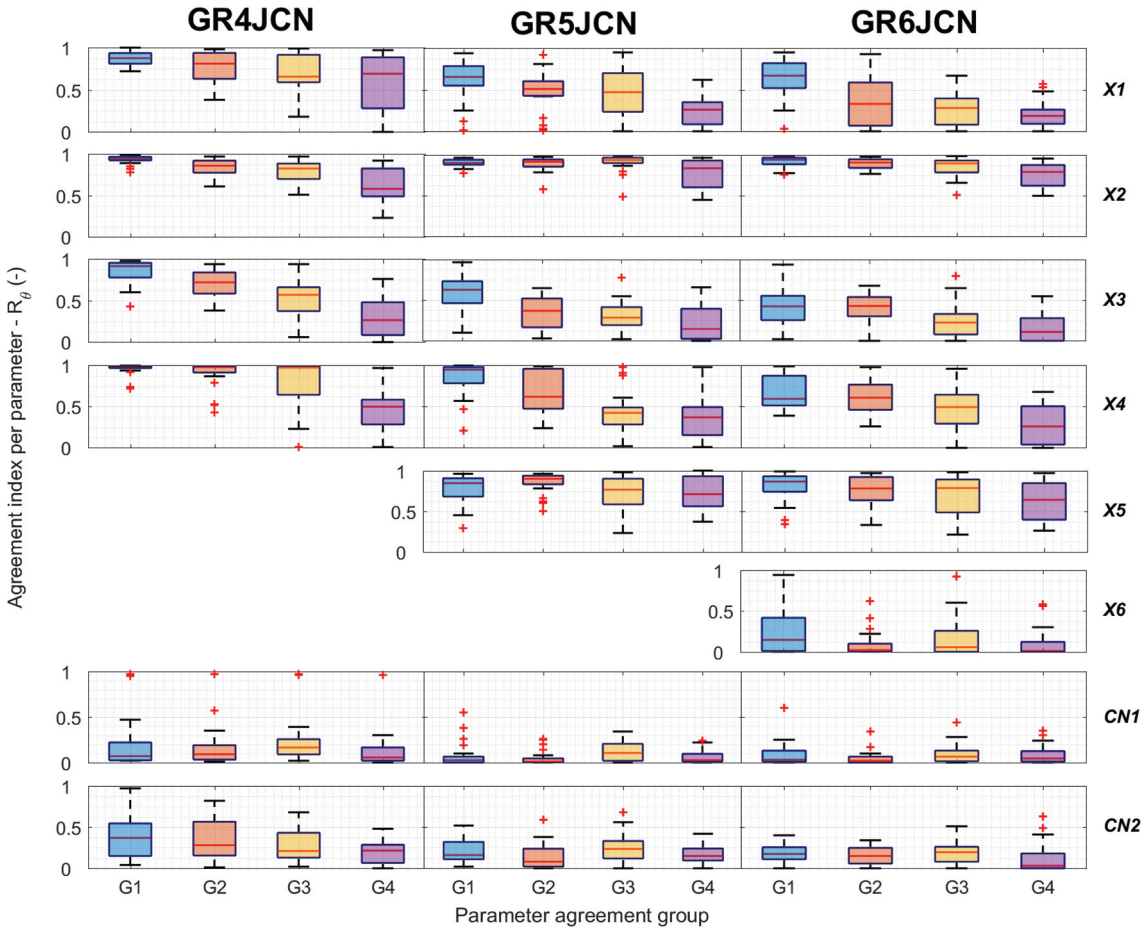


Figure 7. Individual agreement indices per group (rows) and GRXJCN models (columns). Agreement indices close to 1 indicate that the choice of calibration objective and equifinality have little effect on the parameter values.

Chilean catchments. Hence, these investigations suggest that the choice of calibration metric affects parameter values differently depending on the suite of objective functions used and catchment descriptors. In particular, the correlations between catchment attributes and parameter agreement indices reported here (Fig. 5) – and a missing piece from previous studies (Abdulla *et al.* 1999, Gupta *et al.* 2009, Garcia *et al.* 2017, Song *et al.* 2019) – suggest possible links between hydroclimatic characteristics and the potential to find multi-purpose parameter sets in specific catchments. More generally, parameter agreement indices are more influenced by annual hydroclimatology (e.g. precipitation amounts, aridity index), rather than the seasonality of runoff (not shown).

Additionally, the correlation analyses presented here demonstrate that strong associations (either positive or negative) exist between agreement indices of some parameters (e.g. X4, X5 and X6) and catchment-scale hydroclimatic attributes such as mean annual precipitation and aridity index (Fig. 5). In particular, the results show that the choice of calibration metric yields a larger disagreement in parameter values and simulated hydrological variables in (semi-)arid domains. Considering that most climate models project a warmer and drier future for continental Chile (Vicuña *et al.* 2021), the choice of appropriate parameter sets will be more challenging if the goal is to achieve hydrologically consistent model simulations.

As expected, a large disagreement in snow parameters was obtained in basins where the influence of snowmelt on runoff generation is negligible. Interestingly, such disagreement was also observed in snowy catchments (Fig. 3(c)), reflecting the lack of constraints – besides runoff – for process representations. Nemri and Kinnard (2020) showed trade-offs between calibration strategies aimed to simulate snow water equivalent (SWE) and runoff with GR4JCN, finding that when snow parameters were calibrated independently, the quality of streamflow simulations decreased considerably due to overfitting of snow parameters on SWE observations. This stresses the need to constrain the parameter search in conceptual rainfall–runoff models – using, for example, multivariate strategies (Nijzink *et al.* 2018, Széles *et al.* 2020) – to find parameter sets that provide more realistic representations of hydrological processes.

Our results show that the influence of catchment attributes on the overall parameter agreement (R_{OA}) is weaker than that obtained for individual parameter agreement (R_{θ_i}), and gets diminished with increasing model complexity. Even more, increasing complexity (reflected by a larger number of parameters) affects agreement indices of the baseline model (GR4JCN). For example, the addition of threshold-type parameters like X5 decreases the agreement index for X1 (Fig. 5), reinforcing the idea that model complexity may augment parameter identification problems (Doherty and Hunt 2009, Pushpalatha *et al.* 2011).

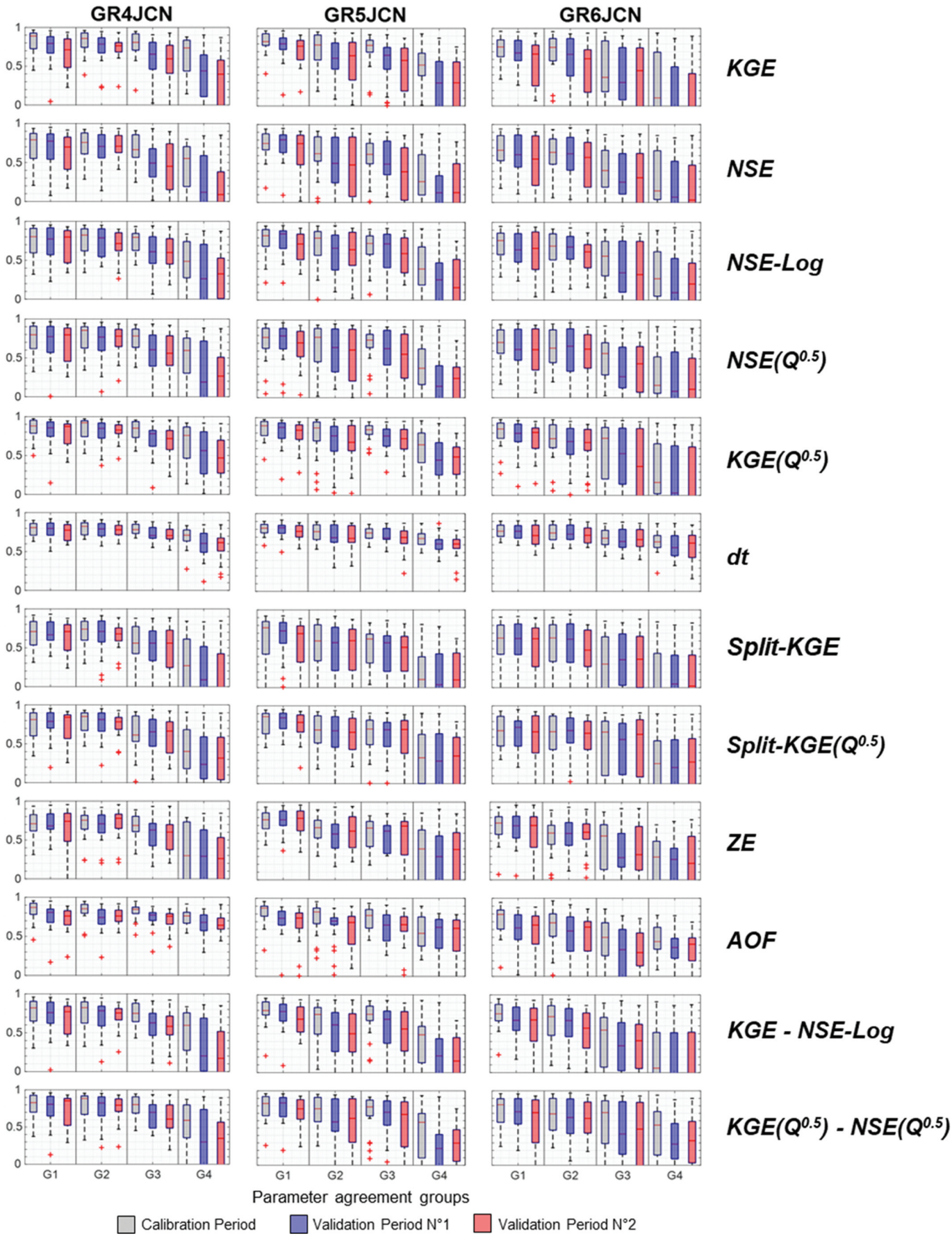


Figure 8. Performance measures for basins grouped by parameter agreement groups (shown in Fig. 5), for the three GRXNCN hydrological model configurations. All metrics are computed with data available for the calibration period (1 April 1985–31 March 2005), validation period No. 1 (1 April 2005–31 March 2010), and validation period No. 2 (1 April 2010–31 March 2017).

5.2 Limitations and future work

The concept of an "optimal" parameter set given a hydrological model structure, meteorological forcings, streamflow observations and a calibration objective function mimics a generalized practice adopted by most water managers, who seek to maximize accuracy for specific streamflow characteristics (Pool *et al.* 2017, Mizukami *et al.* 2019) in spite of the awareness that no single metric can capture all aspects observed in the hydrological system of interest (Jackson *et al.* 2019). Additionally, all the results presented here depend on the choice of the global optimization algorithm. Previous work suggests that SCE-UA yields stable results for models with a low-dimensional parameter space (Arsenault *et al.* 2014), and that is less sensitive to the seed number and the choice of parameter bounds (Abdulla *et al.* 1999) than other algorithms are. Nevertheless, other studies have reported difficulties in finding a unique optimal solution (Demirel *et al.* 2018, Nemri and Kinnard 2020), which motivated the incorporation of equifinal parameter sets in our formulation of parameter agreement.

It should be noted that, in addition to calibration metric selection, many more methodological decisions challenge the identification of parameter values, including the choice of input forcing dataset (Elsner *et al.* 2014), calibration period (Merz *et al.* 2011, Coron *et al.* 2012, Osuch *et al.* 2015), the parameters included in the calibration process (Newman *et al.* 2017), the parameter search strategy (Sorooshian and Gupta 1983, Abdulla *et al.* 1999, Nemri and Kinnard 2020), data errors (e.g. Coxon *et al.* 2015), the quantity of data (Anctil *et al.* 2004), and model structural deficiencies in the equation structure of conceptual models (Sorooshian and Gupta 1983). In particular, the hydrological model structure, input forcings, and error properties in the model and observations have been identified as the main contributors to parameter non-uniqueness (Guillaume *et al.* 2019). The parameter identifiability problem has been approached by many authors over the past several decades (Sorooshian and Gupta 1983, Wagener *et al.* 2003, Doherty and Hunt 2009, Guse *et al.* 2020) and it is not our intention to conduct formal identifiability analyses; rather, we intend to characterize the implications of calibration metric selection for parameter values across a large sample of catchments.

We decided to use the family of GR hydrological models because of their simplicity and flexibility to conduct controlled experiments with increasing complexity, although other flexible platforms with many more modelling alternatives could be explored (Clark *et al.* 2008, Fenicia *et al.* 2011, Knoben *et al.* 2019). Additionally, we did not characterize variations in parameter agreement under calibration periods with different hydroclimatic conditions (Merz *et al.* 2011). Future work could also examine to what extent the choice of calibration metric translates into temporally stable hydrological consistency, which could be assessed through hydrological signatures (Addor *et al.* 2018), flux mapping (Khatami *et al.* 2019), and/or satellite products (Nijzink *et al.* 2018) under contrasting climatic conditions.

We hypothesize that similar results would be obtained if more complex modular platforms are used, either with conceptual, bucket-style models (Clark *et al.* 2008, Knoben *et al.* 2019) or with physically-based models (Niu *et al.* 2011, Clark *et al.* 2015).

6 Conclusions

We have explored the implications of calibration metric selection on the dispersion of parameter values and simulated hydrological responses in three conceptual rainfall–runoff models. To this end, we configured and calibrated three hydrological model structures using 12 different objective functions and computed a parameter agreement index that quantifies the degree of dispersion arising from different calibration metrics, considering equifinality effects. The calibration experiments were conducted in 95 near-natural catchments across continental Chile, which span a diverse range of physiographic and hydroclimatic characteristics. Possible relationships between parameter agreement and catchment descriptors are explored through correlation analysis, and a clustering exercise is performed to examine whether common characteristics exist among catchments that exhibit high (or low) agreement in parameter sets. The results demonstrate that, for the selected model structures applied in this study, the impacts of calibration metric selection on the (dis)agreement of parameter values depend on physiographic and hydroclimatic catchment attributes. Specifically:

- Individual parameter agreement is significantly correlated with some climatic (e.g. aridity index, precipitation) and physiographic (e.g. mean elevation, forest, and barren soil fraction) catchment descriptors.
- Slight modifications in model structure yield changes in spatial patterns of agreement groups and may produce variations in the correlation between individual parameter agreement and catchment attributes.
- Higher parameter agreement is obtained in wet catchments (i.e. with low aridity index and high mean annual runoff ratio) compared to dry or semi-arid basins.
- Catchments with high (low) parameter agreement generally yield an overall better (worse) model performance for the metrics analysed here, in both the calibration period and two independent validation periods.

The results obtained in this study suggest that wet climates with little precipitation seasonality and low probability of snowfall favour conditions for parameter agreement, which means that, in these catchments, it is more likely to find subdomains in the parameter space that reproduce multiple streamflow characteristics and are therefore suitable for multiple water resources applications. Conversely, arid and semi-arid catchments where snow is a key component are generally more challenging, mainly because parameters in degree-day models are more difficult to identify. More generally, the results presented here suggest that the identification of multi-purpose parameter sets in conceptual, bucket-style rainfall–runoff models is more challenging under dry climatic conditions. However, further work with more complex model structures and even more calibration metrics – that aim at evaluating models for other fluxes and/or states than streamflow – is needed to test such hypotheses.

Finally, our study provides a benchmark for hydrological characterizations in near-natural catchments across continental Chile, enabling the assessment of (i) additional sources of

information (e.g. MODIS, Landsat, Soil Moisture Active Passive (SMAP), Soil Moisture and Ocean Salinity (SMOS)) to constrain the parameter space; (ii) alternative parameter estimation strategies; and (iii) using more complex models to improve hydrological consistency.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Pablo A. Mendoza and Nicolás Vásquez were funded by the Fondecyt Project N°11200142. Nicolás Vásquez also received support from the ANID Doctorado Nacional Scholarship N° 21230289 (Chile). Pablo Mendoza acknowledges additional support from ANID-PIA Project AFB220002 (AMTC).

ORCID

Eduardo Muñoz-Castro http://orcid.org/0000-0002-0314-3563
Pablo A. Mendoza http://orcid.org/0000-0002-0263-9698
Nicolás Vásquez http://orcid.org/0000-0002-4651-8935

Data availability

All the data and models used to produce the results included in this paper here are publicly available at Zenodo (Muñoz-Castro *et al.* 2023, https://doi.org/10.5281/zenodo.7822140).

References

Abdulla, F.A., Lettenmaier, D.P., and Liang, X., 1999. Estimation of the ARNO model baseflow parameters using daily streamflow data. *Journal of Hydrology*, 222 (1–4), 37–54. doi:10.1016/S0022-1694(99)00096-7

Addor, N., *et al.*, 2014. Robust changes and sources of uncertainty in the projected hydrological regimes of Swiss catchments. *Water Resources Research*, 50 (10), 7541–7562. doi:10.1002/2014WR015549

Addor, N., *et al.*, 2018. A ranking of hydrological signatures based on their predictability in space. *Water Resources Research*, 54 (11), 8792–8812. doi:10.1029/2018WR022606

Alvarez-Garreton, C., *et al.*, 2018. The CAMELS-CL dataset: catchment attributes and meteorology for large sample studies – Chile dataset. *Hydrology and Earth System Sciences*, 22 (11), 5817–5846. doi:10.5194/hess-22-5817-2018

Anctil, F., Perrin, C., and Andréassian, V., 2004. Impact of the length of observed records on the performance of ANN and of conceptual parsimonious rainfall-runoff forecasting models. *Environmental Modelling & Software*, 19 (4), 357–368. doi:10.1016/S1364-8152(03)00135-X

Andrews, F.T., Croke, B.F.W., and Jakeman, A.J., 2011. An open software environment for hydrological model assessment and development. *Environmental Modelling & Software*, 26 (10), 1171–1185. Elsevier Ltd. doi:10.1016/j.envsoft.2011.04.006

Arsenault, R., *et al.*, 2014. Comparison of stochastic optimization algorithms in hydrological model calibration. *Journal of Hydrologic Engineering*, 19 (7), 1374–1384. doi:10.1061/(asce)he.1943-5584.0000938

Beck, H.E., *et al.*, 2016. Global-scale regionalization of hydrologic model parameters. *Water Resources Research*, 52 (5), 3599–3622. doi:10.1002/2015WR018247

Boisier, J.P., *et al.*, 2018. CR2MET: A high-resolution precipitation and temperature dataset for hydroclimatic research in Chile. In *EGU general assembly conference abstracts* (p. 19739).

Boughton, W. and Droop, O., 2003. Continuous simulation for design flood estimation - a review. *Environmental Modelling & Software*, 18 (4), 309–318. doi:10.1016/S1364-8152(03)00004-5

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

C3S (Copernicus Climate Change Service), 2017. *ERA5: fifth generation of ECMWF atmospheric reanalyses of the global climate*. C3S. Available from: https://cds.climate.copernicus.eu/cdsapp#!/home [Accessed 20 January 2018].

Chegwidden, O.S.S., *et al.*, 2019. How do modeling decisions affect the spread among hydrologic climate change projections? Exploring a large ensemble of simulations across a diversity of hydroclimates. *Earth's Future*, 7 (6), 623–637. doi:10.1029/2018EF001047

Chiew, F.H.S., Stewardson, M.J., and McMahon, T.A., 1993. Comparison of six rainfall-runoff modelling approaches. *Journal of Hydrology*, 147 (1–4), 1–36. doi:10.1016/0022-1694(93)90073-I

Clark, M.P., *et al.*, 2008. Framework for Understanding Structural Errors (FUSE): a modular framework to diagnose differences between hydrological models. *Water Resources Research*, 44 (12), W00B02. doi:10.1029/2007WR006735

Clark, M.P., *et al.*, 2015. A unified approach for process-based hydrologic modeling: 1. Modeling concept. *Water Resources Research*. doi:10.1002/2015WR017198

Coron, L., *et al.*, 2012. Crash testing hydrological models in contrasted climate conditions: an experiment on 216 Australian catchments. *Water Resources Research*, 48 (5), W05552. doi:10.1029/2011WR011721

Coron, L., *et al.*, 2017. The suite of lumped GR hydrological models in an R package. *Environmental Modelling & Software*, 94, 166–171. doi:10.1016/j.envsoft.2017.05.002

Coron, L., *et al.*, 2020. *airGR: suite of GR hydrological models for precipitation-runoff modelling. R package* version 1.7.4. https://doi.org/10.15454/EX11NA, Recherche Data Gouv, V1.

Coxon, G., *et al.*, 2015. A novel framework for discharge uncertainty quantification applied to 500 gauging stations. *Water Resources Research*, 51 (7), 5531–5546. doi:10.1002/2014WR016532

Dee, D.P., *et al.*, 2011. The ERA-interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137 (656), 553–597. doi:10.1002/qj.828

Dembélé, M., *et al.*, 2020. Potential of satellite and reanalysis evaporation datasets for hydrological modelling under various model calibration strategies. *Advances in Water Resources*, 143, 103667. doi:10.1016/j.advwatres.2020.103667

Demirel, M.C., *et al.*, 2018. Combining satellite data and appropriate objective functions for improved spatial pattern performance of a distributed hydrologic model. *Hydrology and Earth System Sciences*, 22 (2), 1299–1315. doi:10.5194/hess-22-1299-2018

DeWalle, D.R. and Rango, A., 2008. *Principles of snow hydrology*. Cambridge University Press.

DGA, 2017. *Actualización del Balance Hídrico Nacional, SIT N° 417, Ministerio de Obras Públicas, Dirección General de Aguas, División de Estudios y Planificación*. Santiago, Chile. Realizado por: Universidad de Chile & Pontificia Universidad Católica de Chile.

DGA, 2018. Aplicación de la metodología de actualización del balance hídrico nacional a las macrozonas Norte y Centro, SIT N° 435. *Dir. Gen. Aguas, Div. Estud. y Planif. Elabor. por UNTEC en UTP con la P. U. Católica Chile*.

DGA, 2019. Aplicación de la metodología de actualización del balance hídrico nacional en la macrozona sur y parte norte de la macrozona Austral, SIT N°, 441.

Diskin, M.H. and Simon, E., 1977. A procedure for the selection of objective functions for hydrologic simulation models. *Journal of Hydrology*, 34 (1–2), 129–149. doi:10.1016/0022-1694(77)90066-X

Doherty, J. and Hunt, R.J., 2009. Two statistics for evaluating parameter identifiability and error reduction. *Journal of Hydrology*, 366 (1–4), 119–127. Elsevier B.V. doi:10.1016/j.jhydrol.2008.12.018

Driessen, T.L.A., *et al.*, 2010. The hydrological response of the ourthe catchment to climate change as modelled by the HBV model.

Hydrology and Earth System Sciences, 14 (4), 651–665. doi:10.5194/hess-14-651-2010

Duan, Q., Gupta, V., and Gupta, V., 1992. Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resources Research*, 28 (4), 1015–1031. doi:10.1029/91WR02985

Elsner, M.M., *et al*., 2014. How does the choice of distributed meteorological data affect hydrologic model calibration and streamflow simulations? *Journal of Hydrometeorology*, 15 (4), 1384–1403. doi:10.1175/JHM-D-13-083.1

Fenicia, F., *et al*., 2018. Signature-domain calibration of hydrological models using approximate Bayesian computation: empirical analysis of fundamental properties. *Water Resources Research*, 54 (6), 3958–3987. doi:10.1002/2017WR021616

Fenicia, F., Kavetski, D., and Savenije, H.H.G., 2011. Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development. *Water Resources Research*, 47 (11), W11510. doi:10.1029/2010WR010174

Fowler, K., *et al*., 2018. Improved rainfall-runoff calibration for drying climate: choice of objective function. *Water Resources Research*, 54 (5), 3392–3408. doi:10.1029/2017WR022466

Garcia, F., Folton, N., and Oudin, L., 2017. Which objective function to calibrate rainfall–runoff models for low-flow index simulations? *Hydrological Sciences Journal*, 62 (7), 1149–1166. Taylor & Francis. doi:10.1080/02626667.2017.1308511

Garreaud, R., *et al*., 2017. The 2010-2015 megadrought in central Chile: impacts on regional hydroclimate and vegetation. *Hydrology and Earth System Sciences*, 21 (12), 6307–6327. doi:10.5194/hess-21-6307-2017

Green, I.R.A. and Stephenson, D., 1986. Criteria for comparison of single event models. *Hydrological Sciences Journal*, 31 (3), 395–411. doi:10.1080/02626668609491056

Guillaume, J.H.A., *et al*., 2019. Introductory overview of identifiability analysis: a guide to evaluating whether you have the right type of data for your modeling purpose. *Environmental Modelling & Software*, 119 (April), 418–432. doi:10.1016/j.envsoft.2019.07.007

Gupta, H.V., *et al*., 2009. Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. *Journal of Hydrology*, 377 (1–2), 80–91. Elsevier B.V. doi:10.1016/j.jhydrol.2009.08.003

Guse, B., *et al*., 2017. Identifying the connective strength between model parameters and performance criteria. *Hydrology and Earth System Sciences*, 21 (11), 5663–5679. doi:10.5194/hess-21-5663-2017

Guse, B., *et al*., 2020. Assessing parameter identifiability for multiple performance criteria to constrain model parameters. *Hydrological Sciences Journal*, 65 (7), 1158–1172. doi:10.1080/02626667.2020.1734204

Jackson, E.K., *et al*., 2019. Introductory overview: error metrics for hydrologic modelling – a review of common practices and an open source library to facilitate use and adoption. *Environmental Modelling & Software*, 119 (May), 32–48. doi:10.1016/j.envsoft.2019.05.001

Khatami, S., *et al*., 2019. Equifinality and flux mapping: a new approach to model evaluation and process representation under uncertainty. *Water Resources Research*, 55 (11), 8922–8941. doi:10.1029/2018WR023750

Knoben, W.J.M., *et al*., 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

Konapala, G., Kao, S.C., and Addor, N., 2020. Exploring hydrologic model process connectivity at the continental scale through an information theory approach. *Water Resources Research*, 56 (10). doi:10.1029/2020WR027340

L'hôte, Y., *et al*., 2005. Relationship between precipitation phase and air temperature: comparison between the Bolivian Andes and the Swiss Alps/Relation entre phase de précipitation et température de l'air: comparaison entre les Andes Boliviennes et les Alpes Suisses. *Hydrological Sciences Journal*, 50 (6). doi:10.1623/hysj.2005.50.6.989

Martinez, G.F. and Gupta, H.V., 2011. Hydrologic consistency as a basis for assessing complexity of monthly water balance models for the continental United States. *Water Resources Research*, 47 (12), W12540. doi:10.1029/2011WR011229

McMillan, H.K., 2021. A review of hydrologic signatures and their applications. *WIREs Water*, 8 (1), 1–23. doi:10.1002/wat2.1499

Melsen, L., *et al*., 2019. Subjective modeling decisions can significantly impact the simulation of flood and drought events. *Journal of Hydrology*, 568 (November 2018), 1093–1104. doi:10.1016/j.jhydrol.2018.11.046

Mendoza, P.A., *et al*., 2016. How do hydrologic modeling decisions affect the portrayal of climate change impacts? *Hydrological Processes*, 30 (7), 1071–1095. doi:10.1002/hyp.10684

Mendoza, P.A., McPhee, J., and Vargas, X., 2012. Uncertainty in flood forecasting: a distributed modeling approach in a sparse data catchment. *Water Resources Research*, 48 (9), W09532. doi:10.1029/2011WR011089

Merz, R., Parajka, J., and Blöschl, G., 2011. Time stability of catchment model parameters: implications for climate impact analyses. *Water Resources Research*, 47 (2), W02531. doi:10.1029/2010WR009505

Mizukami, N., *et al*., 2019. On the choice of calibration metrics for "high-flow" estimation using hydrologic models. *Hydrology and Earth System Sciences*, 23 (6), 2601–2614. doi:10.5194/hess-23-2601-2019

Muleta, M.K., 2012. Model performance sensitivity to objective function during automated calibrations. *Journal of Hydrologic Engineering*, 17 (6), 756–767. doi:10.1061/(ASCE)HE.1943-5584.0000497

Muñoz-Castro, E., *et al*., 2023. Implementation of GR hydrological models in 95 near-natural catchments across Chile. *Zenodo*. doi:10.5281/zenodo.7822140

Najafi, M.R., Moradkhani, H., and Jung, I.W., 2011. Assessing the uncertainties of hydrologic model selection in climate change impact studies. *Hydrological Processes*, 25 (18), 2814–2826. doi:10.1002/hyp.8043

Nash, J. and Sutcliffe, J., 1970. River flow forecasting through conceptual models part I - A discussion of principles. *Journal of Hydrology*, 10 (3), 282–290. Elsevier. doi:10.1016/0022-1694(70)90255-6

Nemri, S. and Kinnard, C., 2020. Comparing calibration strategies of a conceptual snow hydrology model and their impact on model performance and parameter identifiability. *Journal of Hydrology*, 582 (December 2019), 124474. doi:10.1016/j.jhydrol.2019.124474

Newman, A.J., *et al*., 2017. Benchmarking of a physically based hydrologic model. *Journal of Hydrometeorology*, 18 (8), 2215–2225. doi:10.1175/JHM-D-16-0284.1

Newman, A.J., *et al*., 2021. Identifying sensitivities in flood frequency analyses using a stochastic hydrologic modeling system. *Hydrology and Earth System Sciences*, 25 (10), 5603–5621. doi:10.5194/hess-25-5603-2021

Nijzink, R.C., *et al*., 2018. Constraining conceptual hydrological models with multiple information sources. *Water Resources Research*, 54 (10), 8332–8362. doi:10.1029/2017WR021895

Niu, G.-Y., *et al*., 2011. The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *Journal of Geophysical Research*, 116 (D12), D12109. doi:10.1029/2010JD015139

Osuch, M., Romanowicz, R.J., and Booij, M.J., 2015. The influence of parametric uncertainty on the relationships between HBV model parameters and climatic characteristics. *Hydrological Sciences Journal*, 60 (7–8), 1299–1316. Taylor & Francis. doi:10.1080/02626667.2014.967694

Oudin, L., *et al*., 2005. Which potential evapotranspiration input for a lumped rainfall–runoff model? *Journal of Hydrology*, 303 (1–4), 290–306. doi:10.1016/j.jhydrol.2004.08.026

Oudin, L., *et al*., 2006. Dynamic averaging of rainfall-runoff model simulations from complementary model parameterizations. *Water Resources Research*, 42 (7), 1–10. doi:10.1029/2005WR004636

Pechlivanidis, I.G., *et al*., 2014. Use of an entropy-based metric in multiobjective calibration to improve model performance. *Water Resources Research*, 50 (10), 8066–8083. doi:10.1002/2013WR014537

Perrin, C., Michel, C., and Andréassian, V., 2003. Improvement of a parsimonious model for streamflow simulation. *Journal of Hydrology*, 279 (1–4), 275–289. doi:10.1016/S0022-1694(03)00225-7

Pool, S., *et al*., 2017. Streamflow characteristics from modeled runoff time series - importance of calibration criteria selection. *Hydrology and Earth System Sciences*, 21 (11), 5443–5457. doi:10.5194/hess-21-5443-2017

Pool, S., Vis, M., and Seibert, J., 2018. Evaluating model performance: towards a non-parametric variant of the Kling-Gupta efficiency.

Hydrological Sciences Journal, 63 (13–14), 1941–1953. doi:10.1080/02626667.2018.1552002

Pushpalatha, R., *et al*., 2011. A downward structural sensitivity analysis of hydrological models to improve low-flow simulation. *Journal of Hydrology*, 411 (1–2), 66–76. doi:10.1016/j.jhydrol.2011.09.034

Pushpalatha, R., *et al*., 2012. A review of efficiency criteria suitable for evaluating low-flow simulations. *Journal of Hydrology*, 420–421, 171–182. doi:10.1016/j.jhydrol.2011.11.055

Rabus, B., *et al*., 2003. The shuttle radar topography mission - a new class of digital elevation models acquired by spaceborne radar. *ISPRS Journal of Photogrammetry and Remote Sensing*, 57 (4), 241–262. doi:10.1016/S0924-2716(02)00124-7

Rakovec, O., *et al*., 2015. Operational aspects of asynchronous filtering for flood forecasting. *Hydrology and Earth System Sciences*, 19 (6), 2911–2924. doi:10.5194/hess-19-2911-2015

Santos, L., Thirel, G., and Perrin, C., 2018. Technical note: pitfalls in using log-transformed flows within the KGE criterion. *Hydrology and Earth System Sciences*, 22 (8), 4583–4591. doi:10.5194/hess-22-4583-2018

Seiller, G., Roy, R., and Anctil, F., 2017. Influence of three common calibration metrics on the diagnosis of climate change impacts on water resources. *Journal of Hydrology*, 547, 280–295. Elsevier B.V. doi:10.1016/j.jhydrol.2017.02.004

Sepúlveda, U.M., *et al*., 2022. Revisiting parameter sensitivities in the variable infiltration capacity model across a hydroclimatic gradient. *Hydrology and Earth System Sciences*, 26 (13), 3419–3445. doi:10.5194/hess-26-3419-2022

Servat, E. and Dezetter, A., 1991. Selection of calibration objective functions in the context of rainfall-runoff modelling in a Sudanese savannah area. *Hydrological Sciences Journal*, 36 (4), 307–330.

Shafii, M. and Tolson, B.A., 2015. Optimizing hydrological consistency by incorporating hydrological signatures into model calibration objectives. *Water Resources Research*, 51 (5), 3796–3814. doi:10.1002/2014WR016520

Song, J.H., *et al*., 2019. Exploring parsimonious daily rainfall-runoff model structure using the hyperbolic tangent function and tank model. *Journal of Hydrology*, 574 (April), 574–587. doi:10.1016/j.jhydrol.2019.04.054

Sorooshian, S. and Gupta, V.K., 1983. Automatic calibration of conceptual rainfall-runoff models: the question of parameter observability and uniqueness. *Water Resources Research*, 19 (1), 260–268. doi:10.1029/WR019i001p00260

Széles, B., *et al*., 2020. The added value of different data types for calibrating and testing a hydrologic model in a small catchment. *Water Resources Research*, 56 (10). doi:10.1029/2019WR026153

Tolson, B.A. and Shoemaker, C.A., 2007. Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resources Research*, 43 (1), 1–16. doi:10.1029/2005WR004723

Valéry, A., Andréassian, V., and Perrin, C., 2010. Regionalization of precipitation and air temperature over high-altitude catchments – learning from outliers. *Hydrological Sciences Journal*, 55 (6), 928–940. doi:10.1080/02626667.2010.504676

Valéry, A., Andréassian, V., and Perrin, C., 2014a. 'As simple as possible but not simpler': what is useful in a temperature-based snow-accounting routine? Part 1 - comparison of six snow accounting routines on 380 catchments. *Journal of Hydrology*, 517, 1166–1175. doi:10.1016/j.jhydrol.2014.04.059

Valéry, A., Andréassian, V., and Perrin, C., 2014b. 'As simple as possible but not simpler': what is useful in a temperature-based snow-accounting routine? Part 2 – sensitivity analysis of the cemaneige snow accounting routine on 380 catchments. *Journal of Hydrology*, 517 (Supplement C), 1176–1187. doi:10.1016/j.jhydrol.2014.04.058

Vásquez, N., *et al*., 2021. Catchment-scale natural water balance in Chile. *Water Resources of Chile*, 189–208. doi:10.1007/978-3-030-56901-3_9

Vicuña, S., *et al*., 2021. Impacts of climate change on water resources in Chile. *Water Resources of Chile*, 13, 347–363. doi:10.1007/978-3-030-56901-3_19

Vrugt, J.A., *et al*., 2003a. Effective and efficient algorithm for multiobjective optimization of hydrologic models. *Water Resources Research*, 39 (8), 1214. doi:10.1029/2002WR001746

Vrugt, J.A., *et al*., 2003b. A shuffled complex evolution metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. *Water Resources Research*, 39 (8), 1201. doi:10.1029/2002WR001642

Vrugt, J.A. and Robinson, B.A., 2007. Improved evolutionary optimization from genetically adaptive multimethod search. *Proceedings of the National Academy of Sciences of the United States of America*, 104 (3), 708–711. doi:10.1073/pnas.0610471104

Wagener, T., *et al*., 2003. Towards reduced uncertainty in conceptual rainfall-runoff modelling: dynamic identifiability analysis. *Hydrological Processes*, 17 (2), 455–476. doi:10.1002/hyp.1135

Wanders, N., *et al*., 2019. Development and evaluation of a pan-European multimodel seasonal hydrological forecasting system. *Journal of Hydrometeorology*, 20 (1), 99–115. doi:10.1175/JHM-D-18-0040.1

Willmott, C.J., Robeson, S.M., and Matsuura, K., 2012. A refined index of model performance. *International Journal of Climatology*, 32 (13), 2088–2094. doi:10.1002/joc.2419

Wu, Y. and Liu, S., 2014. A suggestion for computing objective function in model calibration. *Ecological Informatics*, 24, 107–111. Elsevier B.V. doi:10.1016/j.ecoinf.2014.08.002

Yang, Y., *et al*., 2019. In quest of calibration density and consistency in hydrologic modeling: distributed parameter calibration against streamflow characteristics. *Water Resources Research*, 55 (9), 7784–7803. doi:10.1029/2018WR024178

Yapo, P.O., Gupta, H.V., and Sorooshian, S., 1998. Multi-objective global optimization for hydrologic models. *Journal of Hydrology*, 204 (1–4), 83–97. doi:10.1016/S0022-1694(97)00107-8

Yilmaz, K.K., Gupta, H.V., and Wagener, T., 2008. A process-based diagnostic approach to model evaluation: application to the NWS distributed hydrologic model. *Water Resources Research*, 44 (9), W09417. doi:10.1029/2007WR006716

Zhang, L., *et al*., 2008. Water balance modeling over variable time scales based on the Budyko framework – model development and testing. *Journal of Hydrology*, 360 (1–4), 117–131. doi:10.1016/j.jhydrol.2008.07.021

Zink, M., *et al*., 2018. Conditioning a hydrologic model using patterns of remotely sensed land surface temperature. *Water Resources Research*, 54 (4), 2976–2998. doi:10.1002/2017WR021346